Selecting Methods for Multiple Imputation of Missing Data

Micha Philipp Fischer

A dissertation submitted in partial fulfillment
of the requirements of the degree of
Doctor of Philosophy
(Survey and Data Science)
in the University of Michigan
2023

Doctoral Committee:

Research Professor Roderick Little, Co-Chair
Research Professor Brady T. West, Co-Chair
Professor Trent Buskirk, Bowling Green State University
Research Professor Michael Elliott
Research Professor Trivellore Raghunathan
Professor Stef van Buuren, University of Utrecht

Micha Philipp Fischer

michaf@umich.edu

ORCID iD: 0000-0003-4730-8718

Acknowledgements

This endeavor would not have been possible without my advisers and dissertation co-chairs, Brady West and Roderick Little. I am extremely grateful to Rod for his critical view of my work and showing interest in my initial ideas. For many years, Brady always made time for our weekly meetings, provided feedback on my writing, and translated Rod's ideas to me. I could have not asked for more excellent mentoring to navigate my Ph.D. program.

I am also thankful for my other committee members – Michael Elliott, Trivellore Raghunathan, Stef van Buuren, and Trent Buskirk – for their guidance.

I am further grateful to Nicholas Henriksen and Lorenzo García-Amaya, not only for their financial support, but also for their mentorship and emotional support during the hardest times of this journey.

A special thank you goes to Felicitas Mittereder, who encouraged me to apply for the Ph.D. program, discussed research ideas with me, supported me in my first two years in the program, and also shared her Rmarkdown formatting template for this dissertation with me.

I would also like to thank all other students in the Ph.D. and Masters programs and our visiting students, especially Wolf and Hani for board game and movie nights; and Yanna, Ai Rene, and Fernanda for distractions, a sympathetic ear, and fun at conferences.

Many thanks go to Christine A Feak and Judith Dyer for supporting me with the writing of this dissertation.

In particular, I would like to thank Nicholas Hinkle-DeGroot for his technical support on the Likert-cluster and for solving my layer 8 errors.

Thanks should also go to Anna, Anna, Anna, and Anna for their support in various ways.

Special thanks go to Anna, who, among other supporting behavior, prevented me from quitting a year a go.

I also want to thank Torben, Elton, Marion, Michi, Lukas, Andi, Becka, Hanz, and many others for, well, everything.

Finally, I would like to thank my Mother for emphasizing the importance of education.

Table of Contents

List of Tables

List of Figures

Abstract

Most data sets from sample surveys contain incomplete observations for various reasons, such as a respondent's refusal to answer questions. Unfortunately, most analysis tools assume complete data sets. When applying such tools to incomplete data, researchers are limited to using either complete observations or complete variables, which can have problematic consequences: biased and inefficient estimates, and decreased power in statistical tests. However, often, the challenges of missing data can be circumvented through sequential imputation (SI), an iterative procedure that imputes missing values variable by variable, conditioning on observed or previously imputed values of other variables. SI generates a complete data set that can be analyzed using standard analytical tools. Multiple imputation, which generates multiple data sets with different draws of the missing values, can be used to improve efficiency and provide inferences that take into account imputation uncertainty.

Various procedures have been proposed for SI, and each procedure involves a choice of options, which can lead to subjectivity in the imputation process. Further, data are mainly analyzed with a substantive question in mind and missing data imputation might not be the primary focus of an analyst. To address these issues, previous studies compared different procedures to find the best way to apply SI. However, they often rely on one assessment strategy, e.g., simulated data only, and often compare only a small number of procedures. These shortcomings lead to findings with low generalizability. This dissertation tries to close this gap by comparing multiple parametric and non-parametric procedures for multiple imputation within the SI framework and to automate and reduce sensitivity in the SI process.

Study One compares several parametric and non-parametric procedures for SI. The evaluation uses a simulation approach, analyzing data from 1) parametric models, 2) non-parametric models, and 3) a real survey data set, a publicly available version of the National Health and Nutrition Examination Survey (NHANES) data. The procedures to be compared include parametric and tree-based procedures. The first study finds that there is no overall best performing method. However, we provide guidance for

practice based on the simulation, taking into account the data situation and required modelling effort.

Study Two proposes a modified SI procedure in which the assessment of different procedures is automated. The study develops criteria for binary, nominal, and continuous incomplete variables to assess imputation methods within SI in an automated and objective fashion. The modified SI process is assessed via a simulation study using data from the NHANES. This study provides methodology for a more automated SI procedure with included plausibility checks for a potential application to high-dimensional data sets with missing values, where specifying models via a human imputer is inefficient.

Study Three investigates the use and implications of incorporating response indicators (RIs) for covariates in the imputation process. This approach leads to imputation under a missing-not-at-random (MNAR) model. A literature review provides insights into how to include RIs for predictors into models with different analysis goals. Furthermore, a targeted simulation study suggests data situations and analysis goals where this approach is sensible. The simulation shows that, under MAR, methods including RIs perform as well as those without them. In MNAR scenarios, methods including RIs can improve performance.

Chapter 1

Introduction

Missing values are increasingly present in survey data sets. Two of multiple reasons for incomplete observations are item nonresponse (Groves, 2004) and unit nonresponse (Lessler & Kalsbeek, 1992). Additional causes are partial responses like survey breakoffs (Peytchev, 2009) and panel attrition (Freedman et al., 1980; Kalton, 1986). Related to this, refusal to participate in complementary data collection, such as the collection of biomarkers (Sakshaug et al., 2010; Schonlau et al., 2010) also results in missing data. Another cause is associated with the latest attempts to link survey data to other data sources, such as administrative records. This occurs in two ways: first, respondents do not consent to link the data (Sakshaug & Kreuter, 2012), and second, linking those data sources can fail (Sayers et al., 2015). Furthermore, high numbers of missing values can also result from survey data linked to automated measures, such as sensor data (Bähr et al., 2022).

Most analysis tools assume complete data sets. Therefore, when these tools are applied to complete cases of a data set only, the reduced sample size implies a loss of information resulting in a loss of power, especially when many incomplete variables are included in the analysis. Generally, complete case analysis also assumes that missing values are missing completely at random (MCAR) (Little & Rubin, 2019, Chapter 1.3), which is often unrealistic, because this means that the probability of observing a value does not depend on either observed or unobserved variables. If this strong assumption does not hold, the results of the analysis may be biased.

In order to analyze all available observations, whether complete or not, with standard analysis tools, imputation procedures have been developed (Rubin, 1978, 1996). One

such imputation procedure is multiple imputation (MI), where multiple complete data sets are produced and then analyzed independently, after which the results are combined for inference purposes (Rubin, 1987, Chapter 3). MI of the missing data can be applied under the weaker missing at random (MAR) assumption (Little & Rubin, 2019, Chapter 1.3) prior to the actual data analysis. This approach also incorporates the uncertainty about the process of predicting the missing values (Little & Rubin, 2019, Chapter 5.4). Importantly, this MI procedure allows all available information to be utilized in a subsequent analysis and, if MAR holds, produces unbiased estimates.

One common way to perform MI is sequential imputation (SI) (e.g., Raghunathan et al., 2001), an iterative procedure that imputes missing values variable by variable. One limitation, however, is that the SI framework needs a well-specified model for each incomplete variable in the data set to provide reasonable predictions of the missing values. In this process, model specification can become a complicated task for an imputer, particularly for data sets with many (incomplete) variables. Thus, a more automated and less burdensome model specification procedure may be desirable given the current trend towards larger (survey) data sets, which more frequently are including data from automated measures, such as sensor data with potentially high numbers of missing values (Bähr et al., 2022), and other data sources (Callegaro & Yang, 2018). Such a procedure would also be desirable for other high-dimensional data, such as when information added to a study creates a missing data problem (Gu et al., 2019).

In theory, SI can be performed using many different procedures, such as regression models (Raghunathan et al., 2001; Van Buuren & Groothuis-Oudshoorn, 2011) or supervised learning procedures (e.g., tree-based (Burgette & Reiter, 2010; Doove et al., 2014; Loh, Eltinge, et al., 2019; Shah et al., 2014; Xu et al., 2016); neural networks (Nordbotten, 1996); support vector machines (Aydilek & Arslan, 2013; Sivapriya et al., 2012)). SI further allows each procedure to then be applied in many different ways (i.e. different specification and parameterization). These many options, however, lead to subjectivity in the imputation process, which threatens reproducible science (Munafò et al., 2017); different researchers might reach different conclusions based on the same data set. In addition to this problem, data are typically analyzed with a substantive question in mind and missing data imputation (and other data preparation steps) might not be the primary focus or interest of an analyst. To address these issues, many studies have proposed and compared different procedures and model types within SI in order to find the best way to apply SI and guide practitioners. Some of the most cited

examples are Burgette and Reiter (2010), who compare classification and regression trees (CART) and linear regression and Shah et al. (2014), who compare random forests (RF) with linear models and predictive mean matching. Although these studies have advanced the field, they often rely on one assessment strategy, e.g., simulated data only, and often compare only a small number of procedures and model types, i.e. the current standard and the new procedure. These shortcomings lead to findings with low generalizability. This dissertation tries to address these limitations by comparing multiple parametric and non-parametric procedures for MI within the SI framework in different ways, and also tries to further automate and reduce subjectivity in the imputation model selection process.

The first study in this dissertation (Chapter 2) focuses on comparing several parametric and non-parametric models/procedures within SI for incomplete continuous variables. The evaluation uses a simulation approach, analyzing data from 1) parametric models, 2) non-parametric models, and 3) a real survey data set, namely, the publicly available version of the 2015-2016 National Health and Nutrition Examination Survey (NHANES). The procedures to be compared include Bayesian linear models and regularized linear models on the parametric side, and, on the non-parametric side, tree-based methods, namely regression trees, random forests, and Bayesian additive regression trees, and predictive mean matching. The methods are assessed using the quantitative properties (Bias, RMSE, confidence interval coverage) of estimated coefficients of a regression model fitted to the multiply imputed data set. This study provides an in-depth assessment of parametric and non-parametric imputation procedures based on simulated data and real data.

In the second study (Chapter 3), we also compare different procedures for imputation. However, the assessment does not solely evaluate the MI process after completion, but proposes a modified SI process where the assessment of different procedures is automated. Two different criteria are developed to assess several different imputation methods within SI in an automated and objective fashion. The criteria are developed for continuous incomplete variables and adapted for both binary and nominal cases. In a case study, we compare the enhanced SI process against one of the state-of-the-art procedures based on the current literature (Doove et al., 2014; Shah et al., 2014), imputation via random forest within the software package MICE, using a simulation process based on the NHANES data. The evaluation focuses on quantitative properties, similar to the first study, and also reports run time assessing applicability. This study

provides theory for a more automated SI procedure with plausibility checks for an application to high-dimensional data sets with missing values, where specifying models via a human imputer is inefficient.

The third study (Chapter 4) investigates the overall use and implications of incorporating response indicators in covariates in the imputation process, which leads to imputation under a missing not at random (MNAR) model (Little & Rubin, 2019, pp. 11–19). This practice is advocated by Loh et al. (2019), Ding and Simonoff (2010), and Twala et al. (2008). However, it has been criticized by Little (2020) who shows theoretically that in the simplest (3 variable) case it leads to implausible assumptions. The third study reviews the literature on including response indicators as predictors in models and algorithms. The study also includes a simulation that investigates MAR and MNAR data situations and compares methods that include and exclude response indicators in the imputation model. The methods are assessed in two different ways. The first set of assessment criteria are the quantitative properties (empirical bias, RMSE, confidence interval coverage) of estimated coefficients of a regression model fitted to the multiply imputed data set. This set of criteria focuses on model-based inference after imputation, incorporating uncertainty about the predicted values. The second criterion consists of the mean of the squared difference between imputed and observed ("true") values on multiply imputed data, which focuses purely on prediction accuracy.

Chapter 2

Parametric Models vs. Trees for Missing Data Imputation

Abstract

Chained equation multiple imputation (CEMI) is a common way to deal with missing values in survey data. Many different CEMI procedures have been proposed over the past two decades. This study compares a set of parametric models (Bayesian [regularized] linear models, predictive mean matching) with several tree-based approaches (classification and regression trees, random forest, Bayesian additive regression trees) for CEMI of missing data. Since different methods may be more suitable for different data situations, the different methods are assessed in three ways: a comparison based on data simulated from 1) parametric models; 2) non-parametric models; and 3) a real data set. We find that there is no overall best method. However, we provide guidance for practice based on the simulation, taking into account the data situation and required modelling effort.

2.1 Introduction

Unit and item nonresponse are increasing in survey data sets (e.g., Groves (2004) and Lessler & Kalsbeek (1992)). Additional causes are partial responses like survey breakoffs (Peytchev, 2009) and panel attrition (Freedman et al., 1980; Kalton, 1986). Related to this, refusal to participate in complementary data collection, such as the

collection of biomarkers (Sakshaug et al., 2010; Schonlau et al., 2010) also results in missing data. The latest attempts to link survey data to other data sources, such as administrative records, can also cause missing values. This occurs in two ways: first, respondents do not consent to link the data (Sakshaug & Kreuter, 2012), and second, linking those data sources can fail (Sayers et al., 2015). Furthermore, high numbers of missing values can also result from survey data linked to automated measures, such as sensor data (Bähr et al., 2022).

Most analysis tools assume complete data sets. When applying these tools to incomplete data, researchers are limited to using either complete observations or complete variables. These restrictions can be problematic, yielding biased estimates and decreased power in statistical tests due to reduced sample sizes. Chained equation multiple imputation (CEMI) (see Raghunathan et al. (2001); Van Buuren et al. (2006)) is a popular method for addressing those issues. Developments of CEMI procedures include Finkbeiner (1979), Raymond & Roberts (1987), Jinn & Sedransk (1989), and Gold & Bentler (2000). The CEMI process repeatedly imputes missing values in a data set variable-by-variable conditional on all other available variables; generates several completed data sets that can be analyzed using standard analysis tools; and provides inference based on simple multiple imputation combining rules (Rubin, 1987). Unlike the restricted complete observations/variable analysis, CEMI can lead to valid point and variance estimates under weaker assumptions on the data (Little & Rubin, 2019, Chapter 1.3).

One weaker assumption is called missing at random (MAR), which assumes that the distribution of missing data is related to the observed (measured) variables only. The weakest assumption on missing values is missing not at random (MNAR) (Little & Rubin, 2002, pp. 11–19). MNAR allows missingness to depend on missing variables after conditioning on variables that are observed.

The CEMI process usually assumes MAR. The procedure first fits a model on the most complete variable and imputes all missing values of this variable based on the posterior predictive distribution of the model. These imputed values are then used in conjunction with the observed values to fit the imputation model for the next variable, and so on. This process continues iteratively over all incomplete variables substituting the imputed values with the new predictions in each iteration (Raghunathan et al., 2001). Many different procedures for CEMI have been proposed. For instance, when applying Bayesian regression models in the CEMI process, the model parameters are

drawn from their posterior distributions; after which, plausible values for the missing values in one outcome variable are drawn (Rubin, 1987, pp. 166–167).

While CEMI with Bayesian regression models fulfill the criteria for proper imputation methods (Rubin, 2004, pp. 116–131), they have potential drawbacks in practice. First, every regression model needs to be specified in advance, which can lead to high modeling effort when the number of incomplete variables is high and the risk of misspecification (Van Buuren et al., 2006). Second, fitting Bayesian models in CEMI can be computationally intensive, given the iterative nature of the procedure (Raghunathan et al., 2001; Van Buuren et al., 2006).

To overcome the first drawback, Zhao and Long (2016) and Deng et al. (2016) use regularized regression procedures. Specifically, Deng et al. (2016) compares LASSO, adaptive LASSO, and elastic net regularization in two different ways. First, the regularized model is fit to complete cases of a bootstrap sample. The imputed values are then randomly drawn from the predictive distribution of this model. Second, regularized regression identifies a subset of the most important covariates, followed by a standard regression model for imputation using the identified subset. The authors find elastic net regularization to be superior in most tested scenarios. Although Zhao and Long (2016) and Deng et al. (2016) show regularization as a promising approach to avoid misspecification, the reported studies also have shortcomings. They simulated main effects only and the missingness mechanism is also only based on main effects of covariates. Further, the simulated data follow (multivariate) normal distributions and other continuous distributions or categorical variables are not investigated.

Another practical approach to CEMI is predictive mean matching (PM). In the case of one incomplete variable, PM predicts the values of the missing observations based on a model (with the incomplete variable as the outcome) fit to the completely observed portion of the data. Next, for each of these predicted values, the procedure selects a set of potential donor observations from the complete cases. A randomly drawn observation from the set of potential donors is used to replace a missing value with the observed value of its donor (Little, 1988). A main advantage of PM is that the model for the distribution of missing values is used only to provide a metric for matching; thus, PM is less prone to misspecification compared to CEMI with (regularized) regression models (Little & Rubin, 2019, Chapter 4.3.2). Another advantage of PM is that the imputed values are always realistic (in terms of scale, range, and shape of distribution), because in PM the imputed values are actually observed in the data (Van Buuren, 2018,

Chapter 3.4). Despite these advantages, the use of actual observed values can also lead to problematic behavior. For instance, the same donor value might be used multiple times, which is especially likely for small samples or when the missing rate is high. Additionally, omitting important interactions in the regression model might degrade performance (Morris et al., 2014).

To alleviate these misspecification problems, researchers have been studying the use of supervised machine learning techniques as a substitute for parametric regression models in CEMI. Burgette and Reiter (2010) propose multiple imputation of continuous variables using sequential classification and regression trees (CT) (Breiman et al., 2017). In this process the actual values for imputation are sampled through a Bayesian bootstrap within the leaves of the CT. The authors find that CT is able to capture complex interactions without high modelling effort. CT also has advantages when interactions among covariates influence categorical outcome variables (Doove et al., 2014). However, CT internally categorizes continuous variables, which potentially results in inadequate approximations of smooth relationships among variables. One study, focusing on categorical variables only (Akande et al., 2017), compares generalized linear models with CT and a Dirichlet process mixture of products of multinomial distributions (Si & Reiter, 2013). These two methods outperform the parametric models in their simulation study.

An even more accurate prediction of the missing values can be obtained by substituting CT with a random forest (RF) algorithm (Stekhoven & Bühlmann, 2012). This procedure has been criticized by Shah et al. (2014), however, because imputing predictive means leads to the understatement of variability of the true values. They propose that missing values should be "imputed by random draws from independent normal distributions centered on conditional means predicted using random forest" (Shah et al., 2014, p. 765). They compare this procedure to parametric imputation models implemented in the R software (R Core Team, 2019) package "Multivariate Imputation by Chained Equations" (MICE) (Van Buuren & Groothuis-Oudshoorn, 2011) and find an advantage of the non-parametric RF over the default parametric models in MICE. Using RF for imputation comes with additional computational effort, because multiple CT procedures are run for each incomplete variable in each iteration.

A Bayesian version of CT (Chipman et al., 1998; Denison et al., 1998) and a further development, Bayesian additive regression trees (BA) (Chipman et al., 2010), is proposed in Xu et al. (2016) for sequentially imputing missing values. The authors

compare sequential BA with parametric CEMI, the MICE default (PM and logistic regression), and MICE CT via simulation studies and find comparable performance for linear relationships among the variables, but a better performance for BA in more complex data situations.

Another recent study compares a high number of non-parametric and machine learning methods for imputation (Dagdoug et al., 2023). Their evaluation via simulation focuses on the assessment of imputation methods in terms of finite population totals. The study finds that the Cubist algorithm (Quinlan, 1993), BA, and XGBoost (Chen & Guestrin, 2016) perform best. However, the study lacks an assessment of standard error estimates, and, relatedly, the application of only single imputation. Further, there is no assessment focusing on the relationships among variables, e.g., the quantitative properties of regression coefficients (see below in this section).

Several tree and forest algorithms have been compared recently using the software packages AMELIA (Honaker et al., 2011) and MICE by applying parametric models to impute simulated missing income values in public-use survey data (Loh, Eltinge, et al., 2019). Loh et al. (2019) find that non-parametric approaches outperform the default MICE. However, there are several limitations (Little, 2020). First, the evaluation of the methods is based only on bias and MSE of mean income estimates. The relationships among variables are not evaluated. Second, Loh et al. (2019) criticize simulation studies using parametric models for data generation (e.g. Burgette & Reiter (2010)), because this approach would favor parametric models. Yet, their own simulation study data are based on one of these competing non-parametric methods, also leading to an unfair comparison.

Non-parametric approaches have both advantages and disadvantages. The major downside is that they are generally black box methods, i.e., the interpretation after fitting is not straight-forward (Breiman, 2001). Even so, model interpretation within CEMI might not be of primary interest. Further, tree-based methods categorize continuous variables (Friedman et al., 2001), so smooth relationships among variables are harder to capture. Thus, if the analysis after imputation focuses on relationships among variables with parametric models (which is often the case in empirical sciences), tree-based methods might be inferior. On the other hand, parametric imputation methods have their own limitations, such as a higher risk for misspecification (White & Carlin, 2010), and specifying parametric models becomes harder with an increasing number of covariates in the data. Furthermore, multicollinearity among variables is more likely

in high-dimensional data situations.

For non-parametric methods, there is no need to explicitly specify a model formula; thus, the strong assumptions in parametric models can be avoided, leading to more flexibility. This flexibility often also leads to a high predictive power. Additionally, high numbers of covariates can be incorporated in tree-based procedures easily and, relatedly, they also work in situations where the number of variables is greater than the number of observations. However, the complexity of non-parametric methods might be inefficient, because accounting for main effects may be more important than accounting for minor complex interactions, depending on the application.

In order to obtain a holistic picture about imputation methods' strengths and weaknesses, the procedures introduced above have to be assessed in different data situations. Following Rubin (1987), the assessment of CEMI procedures should focus on the quantitative properties of β-coefficients (i.e., empirical bias, RMSE, confidence interval coverage rate) of a regression model fit to the multiply imputed data. This assessment model should fit the data, i.e., it should be compatible with the data generating model (e.g., when the data is generated with a squared term, the analysis model should also include a squared term).

The remaining part of this chapter is structured as follows. We first introduce the framework of the simulation and the kinds of models used in this study. We then present the results, followed by a simulation based on the NHANES 2015/2016 data set. We conclude with a discussion of the results.

2.2 Framework

This study compares different imputation methods within the CEMI framework, mainly focusing on parametric models and tree-based methods, for different data situations and different analysis models. We first introduce the general process of data analysis with missing values and then provide a more detailed overview of the compared imputation methods.

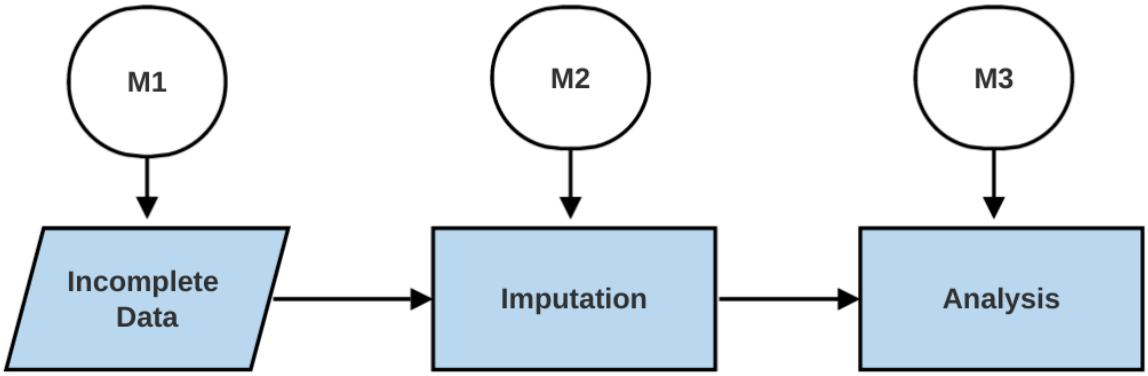

Figure 2.1: Process of missing data imputation and analysis. M1: data generating model, M2: imputation model, M3: analysis model.

Figure 2.1 shows the general process of imputing missing data. We start with incomplete data (left box), which is assumed to be generated by a data generating model (henceforth called M1). Next, the incomplete data is imputed by a CEMI model (henceforth called M2). After the imputation process is finished, the resulting data is analyzed using an analysis model (henceforth called M3).

Missing data imputation via CEMI is an iterative procedure resulting in multiple completed data sets. For a data matrix consisting of fully observed columns $\mathbf{Y}$ and incomplete columns $\mathbf{X} = (X_1, \ldots, X_K)$ ordered by ascending missing rate, CEMI imputes missing values in X_1 using a model regressing the observed elements of X_1 on $(\mathbf{Y}, \mathbf{X}_{-1})$ with $\mathbf{X}_{-1}$ representing $\mathbf{X}$ without the first column. Then moving to the next variable - imputing missings in X_2 with a model regressing X_2 on $(\mathbf{Y}, \mathbf{X}_{-2})$, and so on. After iteratively refitting the models for $\mathbf{X}$, and updating the missing values in $\mathbf{X}$,

one imputed data set is obtained. Repeatedly applying this iterative procedure results in multiple imputed data sets. These data sets can be used to apply complete data analysis tools (here M3) on each set. Rubin's rule (Rubin, 1987, Chapter 3) can be applied to compute point and variance estimates for inference from the multiple resulting estimates.

2.2.1 M1 - Data Generating Model

We generate data following three different data structures combined with three different missingness mechanisms. In each case, three variables are generated, with Y being the completely observed outcome and X_1 and X_2 being the incomplete covariates of a hypothetical analysis model M3.

We first draw values of X_1 from a uniform distribution:

$$X_1 \sim Unif(0, 2). \tag{2.1}$$

Second, we draw values of X_2 given X_1 as

$$X_2 \sim N(\alpha_0 + \alpha_1 X_1, \sigma^2_{X_2}), \tag{2.2}$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. We then draw Y given X_1, X_2 as

$$Y \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2, \sigma^2_Y). \tag{2.3}$$

After generating the values for the three variables, we introduce missing values in (X_1, X_2). We first compute response probabilities for both variables.

$$p_{X_1} = logit^{-1}(\delta_0^{X_1} + \delta_1^{X_1} X_1 + \delta_2^{X_1} X_2 + \delta_3^{X_1} Y) \tag{2.4}$$

$$p_{X_2} = logit^{-1}(\delta_0^{X_2} + \delta_1^{X_2} X_1 + \delta_2^{X_2} X_2) \tag{2.5}$$

The response indicators R_{X_1} and R_{X_2} are drawn as

$$R_{X_1} = \begin{cases} 1 \text{ for } p_{X_1} \geq u_{X_1} \\ 0 \text{ for } p_{X_1} < u_{X_1} \end{cases} \qquad R_{X_2} = \begin{cases} 1 \text{ for } p_{X_2} \geq u_{X_2} \\ 0 \text{ for } p_{X_2} < u_{X_2} \end{cases} \tag{2.6}$$

where u_{X_1} and u_{X_2} are independent $\sim Unif(0,1)$. The chosen parameter values are shown in Table 2.1. Setting $\beta_1 = 0.5$ and $\beta_3 = 0$ leads to the linear cases (L). The parameter combination $\beta_1 = 0$ and $\beta_3 = 1$ adds a quadratic term and defines the quadratic cases (Q).

Table 2.1: Overview of fixed and varying parameters for parametric cases. $logit(a) = log(a/(1-a)), a \in (0,1)$.

Equation	Parameter values
2.2	$\alpha_0 = 0$, $\alpha_1 = 0.25$, $\sigma_{X_2} = 0.3$
2.3	$\beta_0 = 1$, $\beta_2 = 0.5$, $\sigma_Y = 1$, $\beta_1 \in \{0, 0.5\}$, $\beta_3 \in \{0, 1\}$
2.4	$\delta_2^{X_1} = 1.5$, $logit(\delta_0^{X_1}) \in \{0.05, 0.15, 0.5\}$, $\delta_1^{X_1} \in \{0, 2\}$, $\delta_3^{X_1} \in \{0, 1.5\}$
2.5	$logit(\delta_0^{X_2}) = 0.15$, $\delta_1^{X_2} = 2$, $\delta_2^{X_2} \in \{0, 2\}$

For the non-parametric case (N), we replace Equation 2.3 with

$$Y|X_1, X_2 : Y = \beta_0 + f(X_1) + \beta_2 X_2 + \epsilon_Y, \ \epsilon_Y \sim N(0, \sigma_Y^2), \tag{2.7}$$

with $f(X_1) = sin(5X_1)/5$ following the wave-shaped function plotted in Figure 2.2. Second, we change the values of three fixed parameters: $\alpha_1 = 1$, $\sigma_Y = 0.05$, and $\sigma_{X_2} = 0.1$. These changes result in both a stronger relationship between X_1 and X_2 and reduced noise.

We investigate three different missingness mechanisms: first, we define the probability of missingness in X_1 and X_2 as independent of Y ($\delta_3^{X_1} = 0$) and MAR ($\delta_1^{X_1} = 0$, $\delta_2^{X_2} = 0$), henceforth called MX; second, we allow that missingness also depends on Y ($\delta_3^{X_1} = 1.5$), henceforth called MXY; and third, we simulate a MNAR mechanism with missingness depending on X_1 and X_2 ($\delta_1^{X_1} = 2$, $\delta_2^{X_2} = 2$, and $\delta_3^{X_1} = 0$), henceforth called NX. We also vary the baseline numbers of missing values, $\delta_0^{X_1}$. Table 2.2 summarizes the investigated scenarios.

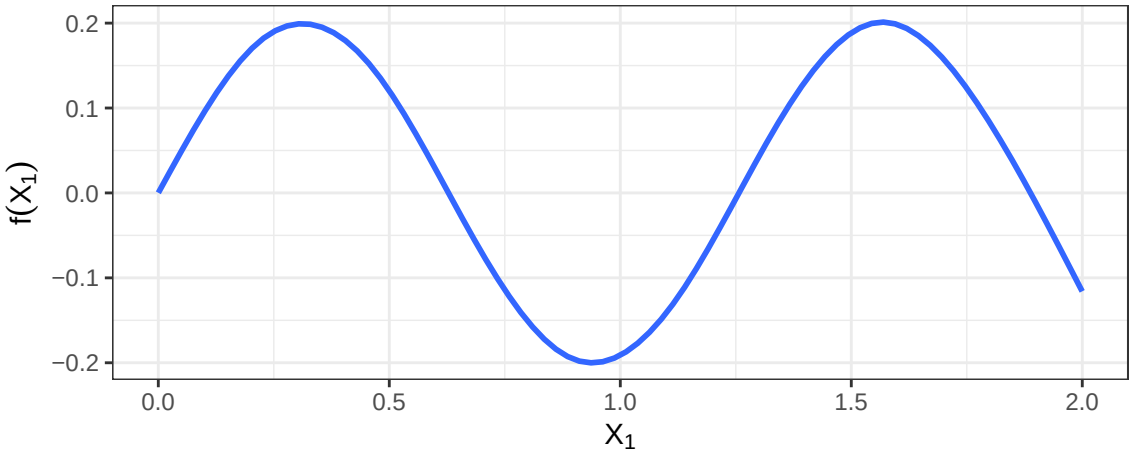

Figure 2.2: Shape of $f(X_1)$.

Table 2.2: Overview of generated data.

		Missingness Mechanism		
		MX ($\delta_1^{X_1}=0$, $\delta_2^{X_2}=0$, $\delta_3^{X_1}=0$)	MXY ($\delta_1^{X_1}=0$, $\delta_2^{X_2}=0$, $\delta_3^{X_1}=1.5$)	NX ($\delta_1^{X_1}=2$, $\delta_2^{X_2}=2$, $\delta_3^{X_1}=0$)
Data Structure	linear (L)	{L,MX}	{L,MXY}	{L,NX}
	quadratic (Q)	{Q,MX}	{Q,MXY}	{Q,NX}
	non-parametric (N)	{N,MX}	{N,MXY}	{N,NX}

2.2.2 M2 - Imputation Model

Based on the general CEMI introduction in Section 2.2, we now describe all applied M2s for $\mathbf{X} = (X_1, X_2)$ and Y. Below, X_k^{obs} ($k \in \{1,2\}$) denotes the observed part of X_k with length n_k^{obs}. If not stated differently, $\mathbf{Z}_k^{obs} = (1, Y^{obs}, X_{-k}^{obs})$ represents the covariates corresponding to X_k^{obs}. X_k^{mis} denotes the unobserved part of X_k with length n_k^{mis}, $\mathbf{Z}_k^{mis} = (1, Y^{mis}, X_{-k}^{mis})$ represents the covariates corresponding to X_k^{mis}. Let further $\gamma_k = (\gamma_{0,k}, \gamma_{1,k}, \gamma_{2,k})$ denote the parameter vector of length q of the parametric M2s with X_k as the outcome. Before starting the first CEMI iteration, the process initially imputes means of the observed values in the respective columns. The software package MICE (Van Buuren & Groothuis-Oudshoorn, 2011, version 3.14.0) is used to apply the following methods, if not stated differently.

2.2.2.1 Bayesian linear models

Bayesian linear models (BMs) were proposed for imputation by Rubin (1987, p. 167) and are specified here as follows:

$$f(X_k|Y, X_{-k}, \gamma_k, log(\sigma_k)) = \gamma_{0,k} + \gamma_{1,k}Y + \gamma_{2,k}X_{-k} + \epsilon, \text{ with } \epsilon \sim N(0, \sigma_k^2). \quad (2.8)$$

We further assume improper prior distributions for the parameters, $P(\gamma_k, log(\sigma_k)) \propto const$. The imputation process in one CEMI step is as follows.

1. Compute the matrix $\mathbf{S}_k = \mathbf{Z}'^{obs}_k \mathbf{Z}^{obs}_k$.
2. Compute $\mathbf{V}_k = (\mathbf{S}_k + diag(\mathbf{S}_k)\kappa)^{-1}$, with κ describing a small ridge parameter.
3. Compute $\hat{\gamma}_k = \mathbf{V}_k \mathbf{Z}'^{obs}_k X^{obs}_k$.
4. Draw $g \sim \chi^2_{n^{obs}_k - q}$.
5. Compute $\sigma_k^2 = (X^{obs}_k - \mathbf{Z}^{obs}_k \hat{\gamma}_k)'(X^{obs}_k - \mathbf{Z}^{obs}_k \hat{\gamma}_k)/g$.
6. Draw q i.i.d. $\mathbf{w}_1 \sim N(\mathbf{0}, \mathbf{1})$.
7. Compute $\mathbf{V}_k^{1/2}$ using Cholesky decomposition.
8. Compute $\dot{\gamma}_k = \hat{\gamma}_k + \sigma_k \mathbf{w}_1 \mathbf{V}_k^{1/2}$.
9. Draw n^{mis}_k i.i.d. $\mathbf{w}_2 \sim N(\mathbf{0}, \mathbf{1})$.
10. Compute n^{mis}_k values of $X^{imp}_k = \mathbf{Z}^{mis}_k \dot{\gamma}_k + \mathbf{w}_2 \sigma_k$.

The procedure predicts n^{mis}_k missing values (10.) accounting for uncertainty in regression parameters (8.). Adding $\dot{\mathbf{w}}_2 \dot{\sigma}_k$ in 10. incorporates additional noise in the predicted values, which is assumed to be normally distributed. Thus, this approach yields proper imputation (Rubin, 2004, pp. 116–131). For the purpose of this study, we use the BM implementation provided in the MICE function mice.impute.norm(), which is equal to Schafer's NORM procedure (Schafer, 1997) when used on all incomplete variables.

2.2.2.2 Bayesian regularized linear models

Here, in each CEMI step, we apply elastic net, the best performing regularization technique in Deng et al. (2016), before imputing via a Bayesian linear model. This indirect use of regularization works as follows.

1. Fit elastic net regularized regression model using X_k^{obs} as the outcome, $\mathbf{Z}_k^{obs}$ as the predictor variables, and the parameters γ_k with the following loss function

$$L(\hat{\gamma}_k) = \frac{\sum_{i=1}^{n_k^{obs}} (X_{i,k}^{obs} - \mathbf{Z}_{i,k}^{obs}\hat{\gamma}_k)^2}{2n_k^{obs}} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{q} \hat{\gamma}_{j,k}^2 + \alpha \sum_{j=1}^{q} |\hat{\gamma}_{j,k}| \right),$$

   with λ describing the regularization parameter and $\alpha \in [0, 1]$ describing the elastic net mixing parameter.
2. Use the model in 1. to identify the active set of covariates $\mathbf{Z}_{A,k}^{obs}$.
3. Carry out the BM procedure (2.2.2.1) replacing $\mathbf{Z}_k^{obs}$ with $\mathbf{Z}_{A,k}^{obs}$.

For further information on regularization via elastic net see e.g., Zou and Hastie (2005). The parameter α is determined via 5-fold cross-validation. We implement this procedure, henceforth E1, based on the R packages glmnet (Simon et al., 2011) (for the regularized model) and rstan (Stan Development Team, 2019) (for the Bayesian model). Further, we apply a modified E1 version on basis-expanded covariates (Hastie et al., 2009, pp. 139–190), i.e., the original covariates in addition to their interactions and higher-order terms serve as covariates in the model (henceforth called E2). For E2, we basis-extend the covariates to $\mathbf{Z}_k^{obs} = (1, Y^{obs}, X_{-k}^{obs}, (Y^{obs})^2, (X_{-k}^{obs})^2, (Y^{obs} * X_{-k}^{obs}))$ and the parameters to $\gamma_k = (\gamma_{0,k}, \gamma_{1,k}, \gamma_{2,k}, \gamma_{3,k}, \gamma_{4,k}, \gamma_{5,k})$

2.2.2.3 Predictive mean matching

We apply predictive mean matching (PM, Little, 1988) in the CEMI framework as described in van Buuren (Van Buuren, 2018, Chapter 3.4). The underlying model is the same as in BM, described in Equation 2.8. However, this model is used in PM to compute the distance metrics $\eta(i, j)$ as described below. The procedure is as follows for one CEMI step.

1. Compute the matrix $\mathbf{S}_k = \mathbf{Z}'^{obs}_k \mathbf{Z}_k^{obs}$.
2. Compute $\mathbf{V}_k = (\mathbf{S}_k + diag(\mathbf{S}_k)\kappa)^{-1}$, with κ describing a small ridge parameter.
3. Compute $\hat{\gamma}_k = \mathbf{V}_k \mathbf{Z}'^{obs}_k \mathbf{X}_k^{obs}$.
4. Draw q i.i.d. $\mathbf{w}_1 \sim N(\mathbf{0}, \mathbf{1})$.
5. Compute $\mathbf{V}_k^{1/2}$ using Cholesky decomposition.
6. Compute $\dot{\gamma}_k = \hat{\gamma}_k + \sigma_k \mathbf{w}_1 \mathbf{V}_k^{1/2}$.

7. Compute distances $\eta(i,j) = |\mathbf{Z}^{obs}_{[i],k}\hat{\gamma}_k - \mathbf{Z}^{mis}_{[i],k}\dot{\hat{\gamma}}_k|$, with $i = 1,\ldots,n^{obs}_k$ and $j = 1,\ldots,n^{mis}_k$.
8. For all observations in $\mathbf{X}^{mis}_k$:

    - select d potential donors from $\mathbf{X}^{obs}_k$ with $\sum_d \eta(i,j)$ being minimum for all $j = 1,\ldots,n^{mis}_k$.
    - randomly draw one donor from d potential donors.
    - impute the value in $\mathbf{X}^{obs}_k$ from that donor.

For this study, we use the implementation in the MICE function mice.impute.pmm() with the default parameter settings ($d = 5$). PM only imputes values that are observed in the data.

2.2.2.4 Classification and regression trees

In this study, we use classification and regression trees (CT) as proposed for imputation by Doove et al. (2014). The imputation process in one CEMI step works as follows.

1. Apply CT on outcome X^{obs}_k and covariates $\mathbf{Z}^{obs}_k = (Y^{obs}, X^{obs}_{-k})$ using recursive partitioning.
2. For each observation in X^{mis}_k,

    - identify its corresponding terminal node in the fit CT (each terminal node includes a subset of X^{obs}_k).
    - randomly draw one observation from the observations (donors) in the identified terminal node.
    - impute the observed value from that donor.

The described CT procedure is implemented in MICE, within the function mice.impute.cart(). CT categorizes continuous variables and only imputes values that are observed in the data. Further, CT is not proper in the sense of Rubin (2004), but CT has been shown to perform well in terms of quantitative properties when interactions are present in the data (Doove et al., 2014).

2.2.2.5 Random forest

Another tree-based method is the random forest (RF), which consists of multiple CTs. We apply RF in the CEMI framework as used by Doove et al. (2014). The procedure works as follows.

1. Draw b bootstrap samples from $(X_k^{obs}, \mathbf{Z}_k^{obs})$.
2. Apply one CT on each bootstrap sample.
3. For each observation in X_k^{mis}:
    - identify its corresponding terminal nodes in all b CTs.
    - randomly draw one donor from the pooled donors in all b terminal nodes.
    - impute the observed value from that donor.

The procedure is implemented in the MICE function mice.impute.rf() with the parameter settings $b = 10$ trees and minimum of five observations in terminal nodes. Like CT, RF only imputes values observed in the data.

2.2.2.6 Bayesian additive regression trees

BA for CEMI was proposed in Xu et al. (2016), the model consists of a sum of trees with estimation based on a Bayesian probability model.

For an outcome vector X_k and a covariate matrix $\mathbf{Z}_k = (Y^{obs}, X_{-k}^{obs})$, the BA model is defined as:

$$X_k = f(\mathbf{Z}_k) + \epsilon \approx \sum_{j=1}^{g} T_j^M(\mathbf{Z}_k) + \epsilon,$$

with $\epsilon \sim N(0, \sigma_k^2)$ denoting a vector of error terms. T_j represents a single tree structure, with its parameters in the terminal nodes M. BA consists of a sum of g trees. Prior distributions are assigned to T, M, and σ_k^2. Draws from the posterior distribution $P(T_1^M, \ldots, T_g^M, \sigma_k^2 | X_k)$ are generated via Gibbs sampling (Geman & Geman, 1984), where the jth tree is fit iteratively. See Kapelner and Bleich (2016) for further details.

In this study, imputation using BA in one CEMI step works as follows.

1. Fit BA on $(X_k^{obs}, \mathbf{Z}_k^{obs})$.

2. Generate draws from posterior distribution of $\hat{P}(\mathbf{Z}_k^{mis})$.
3. Impute each observation in X_k^{mis} with the corresponding draw from $\hat{P}(\mathbf{Z}_k^{mis})$.

In this study, we use an implementation based on the R package bartMachine (Kapelner & Bleich, 2016, version 1.2.6) with the $g = 50$ trees (henceforth BA).

2.2.2.7 Complete case analysis

In addition to the compared imputation models, we also apply complete case analysis (CC), where no imputation is carried out, but only the observations without missing values are used in the analysis.

All compared approaches in this study perform CEMI with five iterations and produce five multiply imputed data sets. In this chapter, we compare the different M2s based on their out-of-the-box functionality. Given this focus, we do not initially assess model fits nor do we check for convergence of the MCMC chains. While this approach is generally not recommended, it may be applied in practice, especially when a high number of incomplete variables make it difficult to implement model assessment within CEMI.

2.2.3 M3 - Analysis model

M3, the analysis model, describes the substantive model fit to the data generated by M1 after CEMI with M2 is applied. In this study, the M3s are regression models compatible with M1, i.e., when the data is generated by, e.g., Q, M3 also includes an quadratic effect. In the remaining part of the paper, we denote missingness mechanisms combined with M1s/M2s using their codes in curly brackets {M1/M2,'missingness mechanism',}. For example, {N,MXY} refers to data generated including the non-parametric term and the missingness mechanism is MAR with missing values depending on both the X and Y variables. The assessment is based on the quantitative properties of M3, i.e., empirical bias (EB) (Equation 2.9), ratio of mean estimated variance to the empirical variance (RV) (Equation 2.13), RMSE (Equation 2.10), and the confidence interval coverage rate (CICR) (Equation 2.14) of estimated β coefficients of M3 in the parametric cases.

The following equations define these quantities, starting with EB:

$$EB(\beta_i^{M2}) = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \widehat{\beta}_{i,k}^{M2} - \beta_i, \tag{2.9}$$

where β_i, $i \in \{0, 1, 2, 3\}$, describes the regression coefficient of interest; $n_{rep} = 100$ is the number of replications; β_i is the true i-th coefficient; and $\widehat{\beta}_{i,k}^{M2}$ the corresponding estimated coefficient, estimated from data multiply imputed using an imputation model $M2$. The RMSE is defined as follows.

$$RMSE(\beta_i^{M2}) = \sqrt{\frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} (\widehat{\beta}_{i,k}^{M2} - \beta_i)^2} \tag{2.10}$$

The empirical variance is defined as follows.

$$EV(\beta_i^{M2}) = (RMSE(\beta_i^{M2}))^2 - (EB(\beta_i^{M2}))^2 \tag{2.11}$$

The mean of the estimated variance for a β_i^{M2} is defined as

$$MV(\beta_i^{M2}) = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} \widehat{V(\beta_{i,k}^{M2})}, \tag{2.12}$$

with $\widehat{V(\beta_{i,k}^{M2})}$ being estimated via Rubin's combining rule (Rubin, 1987, Chapter 3). We define the ratio of the mean estimated variance (Equation 2.12) to the empirical variance (Equation 2.11) (RV) as

$$RV(\beta_i^{M2}) = \frac{MV(\beta_i^{M2})}{EV(\beta_i^{M2})}. \tag{2.13}$$

The following equations define CICR.

$$CICR(\beta_i^{M2}) = \frac{1}{n_{rep}} \sum_{k=1}^{n_{rep}} I_{\{\beta_i \in \widehat{CI(\beta_{i,k}^{M2})}\}} \tag{2.14}$$

In Equation 2.14, $I_{\{\beta_i \in \widehat{CI(\beta_{i,k}^{M2})}\}}$ describes the indicator function with the decision rule

$$I_{\{\beta_i \in \widehat{CI(\beta_{i,k}^{M2})}\}} = \begin{cases} 1 \text{ for } \{\beta_i \in \widehat{CI(\beta_{i,k}^{M2})}\}, \\ 0 \text{ for } \{\beta_i \notin \widehat{CI(\beta_{i,k}^{M2})}\} \end{cases} \tag{2.15}$$

and $\widehat{CI(\beta_{i,k}^{M2})}$ describes the estimated 95% CI of $\beta_{i,k}$ using variance estimation via Rubin's rule (1987, Chapter 3), estimated after multiple imputation via M2.

In the non-parametric case, $f(X_1)$ replaces the β coefficient for X_1 in M1, and an additive M3 with a spline function ($s(X_1)$) is fit on the data ({N,.} case). We propose a procedure that uses marginal predictions on 28 evaluation points of the X_1 axis. For each point, we carry out the following procedure. After fitting the additive M3 on all multiply imputed data sets, we obtain estimates for EB, RV, RMSE, and CICRs (analogously to Equations 2.9, 2.13, 2.10, and 2.14) for all M2s. We then average EB, RV, RMSE, and CICR values over all evaluation points; for EB, the absolute values are averaged. These measures are denoted by S-EB, S-RV, S-RMSE, and S-CICR.

2.2.4 Expected results

Following the framework introduced above, we can predict certain expected outcomes of the simulation study. From Meng (1994) we know that a M2 should be congenial to the M1, to allow for an M3 to be as complex as the M1. Similarly, Bartlett et al. (2015) show via simulation that consistent estimates are achieved when M2s are compatible with the M3s. Therefore, we expect that, in the L case, all imputation procedures would work similarly, because M1, M2, and M3 are congenial. However, in the Q and N cases, we expect that the performance of all M2s will depend on the degree to which the imputation model is compatible with M1 and M3. Thus, we expect that flexible non-parametric methods like CT, RF, and BA will still result in consistent estimates in M3 when the underlying data comes from a similar non-parametric model (M3). In line with Little (1992), we expect that CC is consistent in regression if the probability of missing values does not depend on the outcome variable of the regression model. Further, we expect that CC will lead to unbiased estimates in regression coefficients when missing values in covariates depend on the covariates themselves (Little & Zhang, 2011). However, if the information in the incomplete cases is substantial, we expect CC to be rather inefficient.

General hypotheses:

- H1) We expect bigger differences among M2s for higher rates of missing values.
- H2) Congeniality - We expect M2s that are as complex as M1 and M3 perform better in terms of EB, EV, RMSE, and CICR, compared to less complex M2s.
- H3) We expect lower EV and RMSE values in methods imputing existing values (PM, CT, RF).

Based on the introduced scenarios in Table 2.2 and the introduced M2s, we can state the following specific hypotheses.

Scenario-specific hypotheses:

- H4) For {., MX}, we expect CC will result in low EBs and an increase in EV and RMSE.
- H5) For {., MXY}, we expect CC will lead to increased EB, EV, and RMSE values.
- H6) For {L, MXY}, we expect a low EB in all M2s, except for CC.
- H7) For {Q, MXY}, we expect an increased EB in BM, E1, PM.
- H8) For {N, MXY}, we expect BA, CT, RF to perform best in terms of EB.
- H9) For {., NX}, we expect CC will result in low EB and high EV and RMSE values due to reduced sample size. We further expect all other M2s showing increased EBs.

2.3 Results

The simulation described above results in 27 scenarios (three data structures with three missingness mechanisms - see Table 2.2, each with three levels of $logit(\delta_0^{X_1})$), each carried out with $1,000$ observations and replicated 100 times. The results of the simulation are displayed in 31 plots available in the Appendix (Section 2.8). In this section, we present summary tables (Tables 2.3, 2.4, 2.5, and 2.6) showing the resulting values of EB, RV, RMSE, and CICR for all imputation methods (M2), for all $logit(\delta_0^{X_1}) = 0.05$ scenarios.

Tables 2.3, 2.4, and 2.5 are structured as follows. The column "M1, M3" contains the model structures (L, Q, N) and column "M2" shows the different imputation methods. The remaining 16 columns are combined in sets of four, each set presents the resulting values for EB, RV, RMSE, and CICR for all four possible β coefficients.

Table 2.3: Overview of the performance of imputation methods for $logit(\delta_0^{X_1}) = 0.05$ scenarios with MX missingness mechanism. EB, RV, and RMSE values multiplied by $1,000$. CICR values multiplied by 100.

		True values				EB				RV				RMSE				CICR			
M1, M3	M2	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
L																					
	BM	1	0.5	0.5		-46	25	-120		973	972	945		212	249	324		92	93	95	
	E1	1	0.5	0.5		92	-120	-111		667	557	889		306	337	335		75	66	76	
	E2	1	0.5	0.5		98	-132	-81		786	707	1185		284	310	284		84	75	88	
	PM	1	0.5	0.5		-40	-11	-28		1075	1005	936		199	223	242		93	94	97	
	CT	1	0.5	0.5		-12	-39	-29		679	632	763		204	230	254		89	88	91	
	RF	1	0.5	0.5		-41	-28	35		470	410	556		165	180	222		84	82	88	
	BA	1	0.5	0.5		-940	973	-733		893	933	712		1161	1240	1185		62	63	82	
	CC	1	0.5	0.5		-114	74	-17		1002	951	810		467	360	587		91	94	93	
Q																					
	BM	1	0.0	0.5	1	-612	1657	-262	-769	1281	1971	741	3638	640	1671	561	771	29	1	89	0
	E1	1	0.0	0.5	1	-634	1631	-50	-778	1069	1899	855	3525	660	1644	437	780	21	0	82	0
	E2	1	0.0	0.5	1	-606	1585	-82	-761	1187	2301	704	2913	641	1599	452	765	24	5	85	0
	PM	1	0.0	0.5	1	-99	121	-142	-19	571	545	739	520	304	705	411	324	87	85	91	86
	CT	1	0.0	0.5	1	-12	-115	-42	63	561	576	511	522	332	758	367	353	82	82	90	77
	RF	1	0.0	0.5	1	76	-69	121	-62	678	591	802	599	277	647	291	309	91	88	92	89
	BA	1	0.0	0.5	1	-1144	1722	-1239	-274	1723	2055	847	1902	1212	1894	1504	446	53	78	58	92
	CC	1	0.0	0.5	1	-183	235	-27	-70	1226	1207	809	1164	816	1449	596	601	93	96	93	96
N																					
	BM	0		0.5		121		-133		1314		1361		149		155		83		77	
	E1	0		0.5		138		-149		1485		1528		150		158		72		55	
	E2	0		0.5		151		-161		1606		1638		175		180		76		74	
	PM	0		0.5		355		-352		1403		1415		361		358		16		11	
	CT	0		0.5		261		-264		2043		2042		268		269		33		25	
	RF	0		0.5		368		-364		1430		1557		370		365		0		0	
	BA	0		0.5		57		-71		1326		1329		111		113		96		95	
	CC	0		0.5		70		-47		991		1038		94		62		78		76	

Description: evaluation of different imputation procedures (M2s) in terms of empirical bias (EB), ratio of estimated variance to empirical variance (RV), root mean squared error (RMSE), and confidence interval coverage rate (CICR) of regression coefficients (β_0, β_1, β_2, and β_3) from the analysis model (M3) fit on multiply imputed data with $logit(\delta_0^{X_1}) = 0.05$ and MX missingness mechanism. Investigated are three scenarios of data generating and analysis model ('M1, M3'): linear (L), quadratic (Q), and non-parametric (N). Method codes: BM = Bayesian linear model, E1 = elastic net regularization followed by Bayesian linear model, E2 = elastic net regularization on basis-expanded covariates followed by Bayesian linear model, PM = predictive mean matching, CT = classification and regression tree, RF = random forest, BA = Bayesian additive regression trees, CC = complete case analysis.

Table 2.3 shows results for the MX missingness mechanism. For {L,MX}, we find that CT and PM perform best in terms of EB overall, BA performs worst. In terms of RV, RF shows the lowest values over all β coefficients, PM and BM return the highest and most balanced values. RF and PM return the lowest RMSE values, while BA shows the highest values. Regarding CICR, we find that BA returns the lowest values, PM and BM show the best coverage values.

Regarding {Q,MX}, BA shows the largest EB values, followed by BM. CT and RF return the smallest EB values. For RV, BM, E1, and E2 show overall high values, CT returns the lowest RV values, and CC shows the most balanced RVs. RF and PM perform best in terms of RMSE, while BA shows the highest values. Looking at CICR, we see that CC returns the best result, followed by RF; E1 shows the lowest values.

Focusing on {N,MX}, we see that CC and BA return the smallest EB values; RF and PM return the highest values. CT returns the highest RVs, CC shows the lowest and most balanced values. For RMSE, CC and BA produce the lowest values, while RF and PM return the highest. The best performing methods for CICR is BA; here RF and PM perform worst.

Table 2.4: Overview of the performance of imputation methods for $logit(\delta_0^{X_1}) = 0.05$ scenarios with MXY missingness mechanism. EB, RV, and RMSE values multiplied by $1,000$. CICR values multiplied by 100.

		True values				EB				RV				RMSE				CICR			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
L																					
	BM	1	0.5	0.5		3	-2	-134		856	1052	1334		102	110	210		92	95	90	
	E1	1	0.5	0.5		59	-24	-240		946	842	1101		204	193	305		95	96	67	
	E2	1	0.5	0.5		69	-58	-175		830	665	1049		155	179	264		86	83	89	
	PM	1	0.5	0.5		7	-5	-133		945	1043	1345		101	115	206		90	94	90	
	CT	1	0.5	0.5		55	-66	-99		725	607	807		135	165	216		88	83	89	
	RF	1	0.5	0.5		42	-62	-68		536	453	800		127	154	190		88	79	89	
	BA	1	0.5	0.5		-175	141	46		709	735	944		262	288	336		65	89	92	
	CC	1	0.5	0.5		744	-104	-386		971	999	1151		760	154	418		1	85	40	
Q																					
	BM	1	0.0	0.5	1	-312	726	-108	-310	1404	2170	1051	1917	323	746	193	322	11	22	91	30
	E1	1	0.0	0.5	1	-315	726	-100	-310	1144	1844	762	1598	328	749	234	324	15	21	81	39
	E2	1	0.0	0.5	1	-314	784	-173	-333	1310	1713	857	1575	326	812	261	350	13	23	76	31
	PM	1	0.0	0.5	1	29	-137	-47	73	657	625	940	614	196	514	179	238	89	87	92	85
	CT	1	0.0	0.5	1	129	-376	-57	173	898	889	853	867	231	608	205	273	94	91	93	87
	RF	1	0.0	0.5	1	154	-449	2	191	701	637	834	616	228	613	182	271	80	74	90	75
	BA	1	0.0	0.5	1	-117	158	-30	-14	1458	1492	923	1506	248	559	309	228	96	100	94	100
	CC	1	0.0	0.5	1	877	-502	-221	57	1028	1033	1156	1018	927	712	265	210	16	86	72	93
N																					
	BM	0		0.5		145		-156		1648		1728		156		164		55		49	
	E1	0		0.5		147		-156		1629		1752		154		161		37		16	
	E2	0		0.5		183		-190		2270		2334		196		200		63		52	
	PM	0		0.5		327		-327		1679		1669		333		331		13		12	
	CT	0		0.5		244		-246		2018		2060		251		253		46		38	
	RF	0		0.5		354		-350		1560		1627		355		352		0		0	
	BA	0		0.5		17		-31		1638		1661		72		72		97		97	
	CC	0		0.5		92		-54		972		1009		103		63		57		64	

Description: evaluation of different imputation procedures (M2s) in terms of empirical bias (EB), ratio of estimated variance to empirical variance (RV), root mean squared error (RMSE), and confidence interval coverage rate (CICR) of regression coefficients (β_0, β_1, β_2, and β_3) from the analysis model (M3) fit on multiply imputed data with $logit(\delta_0^{X_1}) = 0.05$ and MXY missingness mechanism. Investigated are three scenarios of data generating and analysis model ('M1, M3'): linear (L), quadratic (Q), and non-parametric (N). Method codes: BM = Bayesian linear model, E1 = elastic net regularization followed by Bayesian linear model, E2 = elastic net regularization on basis-expanded covariates followed by Bayesian linear model, PM = predictive mean matching, CT = classification and regression tree, RF = random forest, BA = Bayesian additive regression trees, CC = complete case analysis.

Regarding {L,MXY}, we find that BM and PM return smallest EB values, CC and BA result in the highest absolute values. PM and BM return overall highest RV values, RF shows the smallest values, CC shows the most balanced RVs. In terms of RMSE, BM and PM perform best , CC and BA perform worst. BM and PM also perform best in terms of CICR. CC and BA yield the lowest CICR values.

In the {Q,MXY} case, BA and PM show the lowest EB values, while CC and E2 return the highest values. For RV, we find that PM and RF return the lowest values, while BM results in overall highest values, and CC shows most balanced RVs. For RMSE, PM, RF, and CT perform best, CC and E2 perform worst. BA show highest CICR values, with some over-coverage; E2 and BM result in the lowest values.

For {N,MXY}, looking at EB, we find that BA and CC perform best, while RF and PM perform worst. In terms of RV, we see that CC results in the lowest and most balanced values, E2 shows the highest values. For RMSE, BA shows the best values; RF and PM perform worst. For CICR, we find again that BA performs best overall (with small over-coverage); RF and PM perform worst.

Table 2.5: Overview of the performance of imputation methods for $logit(\delta_0^{X_1}) = 0.05$ scenarios with NX missingness mechanism. EB, RV, and RMSE values multiplied by $1,000$. CICR values multiplied by 100.

		True values				EB				RV				RMSE				CICR			
		β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
L																					
	BM	1	0.5	0.5		-260	34	41		1235	1296	1173		291	107	151		64	95	95	
	E1	1	0.5	0.5		-29	-96	-153		658	679	919		388	280	307		60	75	73	
	E2	1	0.5	0.5		-49	-88	-127		565	596	769		374	269	299		66	80	76	
	PM	1	0.5	0.5		-248	27	31		1118	1189	1170		284	111	146		65	95	95	
	CT	1	0.5	0.5		-247	24	27		1003	952	791		281	110	168		52	91	92	
	RF	1	0.5	0.5		-248	22	36		850	847	822		280	102	154		46	92	88	
	BA	1	0.5	0.5		-1022	580	198		894	969	814		1048	609	301		9	25	82	
	CC	1	0.5	0.5		-44	27	-11		1026	1121	917		220	137	203		94	92	93	
Q																					
	BM	1	0.0	0.5	1	-849	809	62	-277	1888	2094	1166	2235	862	852	195	299	0	48	96	67
	E1	1	0.0	0.5	1	-879	805	143	-279	1944	1872	1133	1971	891	852	231	304	1	50	82	67
	E2	1	0.0	0.5	1	-892	864	90	-300	2229	2254	1083	2426	902	903	215	319	0	47	87	69
	PM	1	0.0	0.5	1	-278	-341	44	207	614	509	1002	505	371	619	188	306	64	79	94	69
	CT	1	0.0	0.5	1	-195	-495	33	266	738	661	681	686	323	730	223	349	79	77	90	66
	RF	1	0.0	0.5	1	-131	-608	79	293	913	785	808	787	257	765	210	355	85	72	92	65
	BA	1	0.0	0.5	1	-1173	878	-60	-74	1171	1310	408	1322	1215	1041	410	240	15	66	80	92
	CC	1	0.0	0.5	1	-73	76	-12	-18	1149	1157	916	1132	450	702	203	269	97	96	93	96
N																					
	BM	0		0.5		55		-86		1667		1701		70		94		93		66	
	E1	0		0.5		83		-107		1614		1619		90		112		62		29	
	E2	0		0.5		30		-60		1544		1622		95		102		91		88	
	PM	0		0.5		66		-96		1816		1712		82		107		88		76	
	CT	0		0.5		16		-43		1772		1871		42		56		95		90	
	RF	0		0.5		175		-192		2454		2641		178		194		12		8	
	BA	0		0.5		-86		56		1186		1263		97		70		66		81	
	CC	0		0.5		-3		-4		922		913		32		23		93		91	

Description: evaluation of different imputation procedures (M2s) in terms of empirical bias (EB), ratio of estimated variance to empirical variance (RV), root mean squared error (RMSE), and confidence interval coverage rate (CICR) of regression coefficients (β_0, β_1, β_2, and β_3) from the analysis model (M3) fit on multiply imputed data with $logit(\delta_0^{X_1}) = 0.05$ and NX missingness mechanism. Investigated are three scenarios of data generating and analysis model ('M1, M3'): linear (L), quadratic (Q), and non-parametric (N). Method codes: BM = Bayesian linear model, E1 = elastic net regularization followed by Bayesian linear model, E2 = elastic net regularization on basis-expanded covariates followed by Bayesian linear model, PM = predictive mean matching, CT = classification and regression tree, RF = random forest, BA = Bayesian additive regression trees, CC = complete case analysis.

Table 2.5 shows results for the NX missingness mechanism. For {L,NX}, we see that CC and E2 perform best in terms of EB; BA performs worst. E2 shows lowest RV values, BM returns highest values, and CC returns the most balanced RV values, overall. Here, BM, PM, CT, and RF return the lowest RMSE values, while BA shows the highest values. For CICR, CC shows the best results, followed by BM; BA performs worst.

For {Q,NX}, CC returns the lowest absolute EB values, while BA returns the highest values. PM and CT show lowest RV values, E2 shows highest RVs, and CC leads to the most balanced RVs. In terms of RMSE, PM and RF perform best; BA performs worst. For CICR, CC and RF show best values. E1 and E2 return the lowest values.

Looking at {N,NX}, we see that CC and CT lead to the lowest absolute EB values; RF and E1 show the highest values. RF returns highest, CC the lowest and most balanced RV values. For RMSE, CC and CT perform best, while RF returns the highest values. CT and CC show highest CICR values; RF results the lowest values.

Table 2.6: Overview of the performance of imputation methods in terms of spline evaluation for all N and scenarios with $logit(\delta_0^{X_1}) = 0.05$. S-EB, S-RV, and S-RMSE values multiplied by $1,000$. S-CICR values multiplied by 100.

	MX				MXY				NX			
M2	S-EB	S-RV	S-RMSE	S-CICR	S-EB	S-RV	S-RMSE	S-CICR	S-EB	S-RV	S-RMSE	S-CICR
BM	143	1279	174	59	112	2471	121	55	67	2140	76	69
E1	141	5463	144	42	112	2572	120	49	78	2060	84	54
E2	112	7958	223	67	110	2572	124	64	57	1590	89	85
PM	116	2659	134	66	188	1863	196	32	69	2111	80	71
CT	29	2044	43	92	124	1965	134	47	30	1699	46	86
RF	13	1031	29	91	209	1629	211	6	133	2141	136	12
BA	23	1824	63	95	22	1716	60	92	40	1223	58	80
CC	40	382	81	89	32	344	65	88	30	298	54	86

Description: spline evaluation of different imputation procedures (M2s) in terms of empirical bias (S-EB), ratio of estimated variance to empirical variance (S-RV), root mean squared error (S-RMSE), and confidence interval coverage rate (S-CICR) for M3s including a spline function (N case). S-EB, S-RV, S-RMSE, and S-CICR are computed based on marginal predictions and averaged over 28 evaluation points of X_1 from the analysis model (M3) fit on multiply imputed data with $logit(\delta_0^{X_1}) = 0.05$ for all three missingness mechanisms (MX, MXY, NX). Method codes: BM = Bayesian linear model, E1 = elastic net regularization followed by Bayesian linear model, E2 = elastic net regularization on basis-expanded covariates followed by Bayesian linear model, PM = predictive mean matching, CT = classification and regression tree, RF = random forest, BA = Bayesian additive regression trees, CC = complete case analysis.

Table 2.6 shows the spline evaluation for the non-parametric case (N) of M1 and M3 models and is structured as follows. The first column, "M2", contains the different imputation methods. The following 12 columns present S-EB, S-RV, S-RMSE, and S-CICR for all three investigated missingness mechanisms (MX, MXY, NX).

The spline evaluation in {N,MX} shows that for S-EB, RF and BA perform best, while BM and E1 show the highest values. CC returns the lowest S-RV, E2 shows the highest S-RV value, and RF results in the most balanced S-RV value. For S-RMSE, we find that RF and CT perform best, E2 and BM perform worst. We further find that BA and CT perform best in terms of S-CICR, while E1 and BM perform worst.

In the {N,MXY} case, we see that BA and CC perform best in terms of S-EB, while RF and PM perform worst. CC returns the lowest S-RV, E1 and E2 show the highest values, and RF returns the most balanced S-RV. For S-RMSE, BA and CC show the best results; RF and PM show highest values. BA and CC return the highest S-CICR values. RF and PM show lowest S-CICR values.

For the non-parametric part of the {N,NX} scenario, CT and CC perform best in terms of S-EB, while RF and E1 perform worst. BM and RF return the highest RV, CC shows the lowest RV values, and BA results in the most balanced S-RV. CT and CC show the best results for S-RMSE; RF and E2 show the highest values. For S-CICR, CT and CC again return the best results. RF and E1 return the lowest values.

To summarize, we find that, as expected, CC shows good performance in all MX and NX cases. In the MXY scenarios, we see that, overall, PM performs well in L and Q, while BA yield the best results in the N scenario. The parametric methods, BM, E1, and E2, show similar patterns in most of the investigated scenarios. CT works well in all MX scenarios. We find that BA generally performs well in N cases. However, BA often shows the worst performance in L and Q cases, which is an unexpected finding (see Section 2.3.0.1).

We now compare the stated hypotheses (Section 2.2.4) to the findings, starting with the general hypotheses. Overall, we observe bigger differences among the imputation procedures for higher rates of missing values, as stated in H1. Further, there is partial support for H2: in {L,MXY} all imputation methods show low EB, supporting H2. However, we find increased EB in some of the applied methods in the {Q,MXY} scenarios. Also, some of the non-parametric methods lead to high EB values in {N,MXY}. Finally, looking at H3, we do not find support for lower RMSE values after imputing

with methods that impute only observed values.

Focusing on the scenario-specific hypotheses, for the parametric cases, we observe that CC performs well in MX, as stated in H4. Further, we see increased RMSE values for CC in {L,MX}, but not in {N,MX}, and only a small increased RMSE in {Q,MX}. In {.,MXY}, CC generally leads to an increased EB, and increased RMSE in L and Q scenarios, mostly supporting H5. We find support for H6; all imputation methods show a low EB in {L,MXY}, compared to CC, except BA. While BA performs generally well in the non-parametric case (stated in H8), the general poor performance of BA in the parametric part of the simulation does not support H6. In {Q,MXY}, BM, E1, and E2 show high EB; PM performs better in terms of EB. This finding does not support H7. In the NX scenarios, CC results in empirically unbiased estimates. We further find increased RMSE values in L and Q scenarios. Both findings are mostly in line with H9.

2.3.0.1 Investigating BA behavior

From the BA results in the previous simulation we overall see that BA shows extremely underwhelming performance in the {L,MX} case, while BA performs well in the {N,MX} case. Besides the different relationships in the data in {L,MX} and {N,MX}, the scenarios differ in terms of the magnitude of variance in the data generating process. For the parametric part of the simulation (L, Q), the variance is generally higher (Equations 2.2 and 2.3) than the variance used for the N-case (Equation 2.7). We applied higher variances in L and Q to reduce collinearity in M3.

For this additional investigation of BA's behavior, we compare different BA implementations and CT on {L,MX} data. We vary the data in terms of variance levels ($\sigma_Y \in \{0.05, 1\}$, $\sigma_{X_2} \in \{0.1, 0.3\}$) and numbers of observations ($n \in \{1000, 3000, 5000\}$) and focus on EB in one simulated data set. One applied BA implementation includes cross-validation for an optimized hyperparameter selection (henceforth BA-CV). The grid of choices for BA-CV consists of: number of trees $\in \{50, 200\}$, parameter $k_{BA} \in \{1, 2, 3, 5\}$, and the parameter pair $(\nu, q) \in \{(10, 0.75), (3, 0.90), (3, 0.99)\}$. The parameters k_{BA} and (ν, q) influence the prior choices for BA. These choices are based on recommendations in Chipman et al. (2010), the implementation is based on the R package 'bartMachine' (Kapelner & Bleich, 2016, version 1.2.6). Additionally, we apply BA as implemented in the R package 'BART' (Sparapani et al., 2021, version 2.9)

(henceforth BA-P). The low performance of BA could result from BA being a stepwise function applied to continuous data. For that reason, we also investigate the performance of a BA procedure designed for smooth functions (A. Linero & Yang, 2018) implemented in the R package 'SoftBart' (A. R. Linero, 2022) (henceforth BA-S). For all additional BA implementation, the changes only apply to fitting BA, the remaining CEMI procedure for BA remains as described in Section 2.2.2.6.

Table 2.7: Overview of BA performances in terms of EB for a L scenario with different levels of variability and different numbers of observations in the data with $logit(\delta_0^{X_1}) = 0.05$. All values are multiplied by $1,000$.

				n = 1,000			n = 3,000			n = 5,000		
	σ_Y	σ_{X_2}	M2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
	0.05	0.1	BA-CV	35	-64	-45	59	-121	87	44	-79	-37
	0.05	0.1	BA-P	9	-13	31	1	-18	74	-2	-13	61
	0.05	0.1	BA-S	8	-10	23	-2	-15	71	-1	-15	68
	0.05	0.1	BA	4	-9	25	0	-14	62	-1	-15	69
	0.05	0.1	CT	12	12	-85	10	1	-25	5	0	-17
	0.05	0.3	BA-CV	109	-110	-107	60	-66	-94	35	-61	-35
	0.05	0.3	BA-P	9	-1	-18	1	2	2	-5	3	8
	0.05	0.3	BA-S	-1	10	-26	1	0	5	-7	5	8
	0.05	0.3	BA	0	-1	4	0	-2	12	-3	-1	14
	0.05	0.3	CT	42	-15	-51	25	-9	-22	23	-11	-13
	1.00	0.1	BA-CV	104	113	-1030	136	-80	-262	47	-42	-164
	1.00	0.1	BA-P	-162	1051	-3804	-63	247	-828	-116	264	-770
	1.00	0.1	BA-S	-220	1189	-4159	-63	250	-831	-80	197	-646
	1.00	0.1	BA	-249	1047	-3529	-128	368	-1079	-176	327	-813
	1.00	0.1	CT	-6	202	-924	19	28	-200	20	-31	-51
	1.00	0.3	BA-CV	109	-39	-399	102	-74	-138	62	-87	-61
	1.00	0.3	BA-P	-235	397	-784	-111	111	30	-163	137	77
	1.00	0.3	BA-S	-177	308	-707	-95	86	74	-161	160	-39
	1.00	0.3	BA	-200	329	-604	-79	60	144	-165	140	138
	1.00	0.3	CT	23	107	-517	67	-16	-130	36	-27	-65

Description: evaluation of different imputation procedures (M2s) in terms of empirical bias for all three coefficients of M3. Method codes: BA = Bayesian additive regression trees, BA-CV = BA with cross-validated hyperparameter tuning, BA-P = BA implementation from R package 'BART', BA-S = BA implementation for smooth functions, CT = classification and regression tree.

Table 2.7 shows the results for different BA approaches in the investigated scenario. For $\sigma_Y = 0.05$, $\sigma_{X_2} = 0.1$, we only find small differences among the BA implementations, with BA-CV overall showing the strongest EB. The EB of CT is of similar magnitude. In the $\sigma_Y = 0.05$, $\sigma_{X_2} = 0.3$ case, BA-CV returns the highest absolute EB, followed by CT. The remaining BA procedures show only small biases. Looking at $\sigma_Y = 1$, $\sigma_{X_2} = 0.1$, we see the strongest differences among the procedures. For 1,000 observations, BA, BA-P, and BA-S return strong EBs, especially in β_2. BA-CV and CT return lower absolute EBs in all three parameters. Increasing the number of observations overall reduces EBs for all compared procedures. For 5,000 observations we find CT performing best, followed by BA-CV. For $\sigma_Y = 1$, $\sigma_{X_2} = 0.3$ we see a similar pattern to the $\sigma_Y = 1$, $\sigma_{X_2} = 0.1$ case, but overall less extreme EB values. Overall, we find that increased variance in the outcome of M3 (σ_Y) in the L cases (compared to the N case) leads to low performance in terms of EB in BA.

We can offer guidance to practitioners based on the findings of this simulation. When missing data occur in practice, the underlying data generating model, including the missingness mechanism, is most likely unknown. If the analysis model is known by the imputer, we can base the selection of the imputation model on the complexity of the analysis model. The imputation model should then be at least as complex as the analysis model. However, when the analysis model is unknown (e.g., in imputed public-use data files), we recommend choosing an imputation model that can automatically detect important interactions in the data, like RF or CT. In each case, a careful inspection of the imputation process is necessary. We recommend at least comparing the distributions of imputed and observed values in all incomplete variables. Differences between observed and missing values can result from a bad imputation model or a non-ignorable missingness mechanism.

2.4 Simulation using Real Data

This section describes the process of evaluating the M2s with real data. As seen in the previous section, the performance of imputation procedures can be sensitive to the data generating process. We therefore generate data sets with missing values from a publicly available version of the NHANES (National Health and Nutrition Examination Survey) data. The process of evaluation follows Ezzati-Rice et al. (1995) and Schafer

et al. (1996). A detailed description can be found in the following section 2.4.1. We use the 2015-2016 NHANES data, consisting of about 12,000 respondents, as a synthetic population. The data set consists of data from different modes of data collection: all respondents completed an initial questionnaire, complemented by a physical examination taking place in a mobile examination center and two days of recording a nutrition diary. For missing data patterns, we find values of variables for the physical examination completely missing for some observations, because some respondents refused to participate in this part of data collection. There are also incomplete nutrition diaries, notably on the second day, in addition to item-missing data for sensitive questions in the questionnaire.

2.4.1 Assessment Process

Figure 2.3 depicts the process used to evaluate the different imputation procedures, which assumes an incomplete data set with observations in rows and variables in columns. We begin by defining the variables of interest (VOI), indicated in blue, that are later used to compare the different imputation procedures. After the VOI are defined, the data set (on the left side of the figure) is further divided into two subsets of variables: one that is fully observed (lefthand side, 31 variables), and one (right-hand side, 355 variables) with incomplete observations (indicated by dark gray fields) including the VOI (three variables). The 5,474 observations are then divided into two sets, one with fully observed VOI, and another where missing values are present in the VOI.

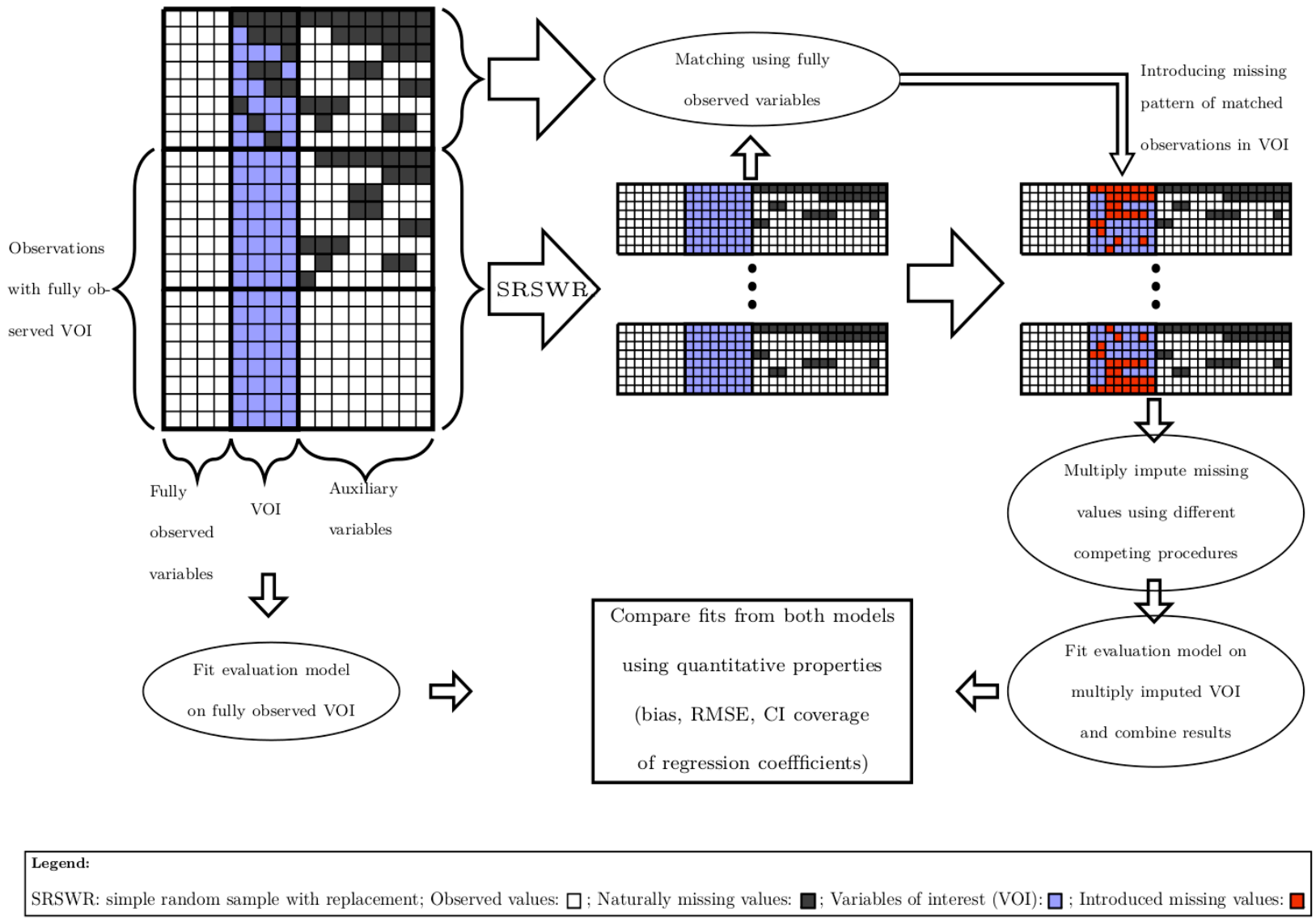

Figure 2.3: Structure of the evaluation process using NHANES data.

The evaluation process continues with repeatedly drawing simple random samples with replacement (SRSWR) of 500 observations from the subset of cases with completely observed VOI. These SRSWR-sub-data sets are the foundation of the evaluation, and consist of complete cases for all VOI, and the aforementioned subsets of variables. In the next step, missing value patterns observed in the data are introduced into the observations of VOI (indicated in red). These introduced missing values are donated from the observations of incomplete VOI matched to the observations in the subsampled data sets. In this study, only the VOI are imputed, the complete variables are used as covariates.

The matching of complete observations in VOI in each SRSWR with those observations including missing values in VOI in the original data set is based on 26 of the 31 completely observed variables. To perform the matching, we use the package StatMatch (D'Orazio, 2019) in the statistical software R (version 3.6.1) (R Core Team, 2019). Specifically, random hot deck (see Andridge & Little (2010)), a procedure that randomly selects a donor (i.e., the matched observation including missing values in the VOI) for each observation in the SRSWR from an appropriate subset of all donors is carried out. The subset is built from the 26 completely observed variables. In order to use the variables in the random hot deck procedure, it is necessary to categorize them. All continuous variables with more than 10 unique values are categorized using deciles, resulting in a data set with maximal 10 categories of similar size per variable.

The matching is performed within donation classes defined by the variables AGE, RACE and GENDER. The following parameters were set in the matching process: $dist.fun =$ "$exact$", $cut.don =$ "$span$", and $k = 0.15$. This setup allows 15% of the closest donors to be considered using exact matching. Since only observations with at least one missing value in the VOIs are considered for donating the missing data pattern, the resulting data set consists of observations without complete cases in the VOI. After finding a match for each observation in the subsampled data set, the missing pattern in the VOI of this matched observation is introduced. This procedure results in an ignorable missingness mechanism, conditional on the fully observed variables; see Ezzati-Rice et al. (1995) and Schafer et al. (1996) for further details.

This process has several advantages. For instance, the missingness pattern introduced in the VOI is actually observed in the data set. That is, missing patterns are neither artificial nor unrealistic. Further, this approach avoids the need to specify a probabilistic model for introducing missing values, an often used approach when evaluating impu-

tation procedures using simulation. Together with this advantage, no specified model is needed for the data generating process; both the data sets and the relationships among variables are observed. This leads to an evaluation of missing data imputation procedures that is closer to real world missing data problems.

2.4.2 Variables of Interest (VOI)

There are several selection criteria for the VOI in place:

1. Relationship: the selected variables should follow an approximately linear relationship, because a linear model should be fit as M3.
2. Missing values: in order to introduce missing data patterns following Ezzati-Rice et al. (1995), the VOI should have a relatively high number of missing values (i.e., $20 - 40\%$). Next, for the incomplete cases to provide substantial information, missing values should be in predictors rather than in the outcome (Little, 1992). Additionally, in order to observe different missing data patterns in the VOIs, they should also be collected from different modes of data collection.
3. Population: The variables should target the whole population of NHANES (i.e., no sub-populations like "smokers"), to avoid "not applicable" cases which would reduce the number of observations used.
4. Time period: the variables should be measured in NHANES data collection 2015/2016.

For the selection of VOI, it would have been ideal to base the choice on a substantive paper that used NHANES data. However, a search for recently published studies using NHANES 2015/2016 revealed that, if regression is used as an analysis tool, covariates mostly consist of socio-demographic variables, which are (almost) completely observed. With this in mind, we decided to select variables fulfilling the four criteria from the NHANES data and use them in a hypothetical substantive regression model. For the outcome variable, we use the log-transformed BMI (body mass index) computed from height and weight ($BMI = (\text{weight in kg})/(\text{height in cm})^2$) measured in the physical examination. The covariates are the daily kilo-calories intake (KCAL), calculated from the complete nutrition diary of the second day, and the number of days having 4 to 5 alcoholic drinks in the past 12 months (ALC) (treated as a continuous variable), which

was collected using the questionnaire. The three variables are included in the following analysis model for the i-th observation:

$$BMI_i = \beta_0 + \beta_{KCAL} KCAL_i + \beta_{ALC} ALC_i + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma^2)$.

Figure 2.4 shows the missing data pattern for the selected VOIs in the NHANES 2015/2016 data, with blue indicating 'observed', and gray meaning 'missing'. The left side of the figure shows the proportion of missing values for each variable. The right side of the figure displays the most frequent missing patterns appearing in the data (omitting patterns of frequencies smaller than 0.01), including the frequencies of the displayed missing patterns on the very right. As seen on the bar chart on the left, ALC has the highest proportion of missing values (approximately 0.41), followed by KCAL (approximately 0.23) and BMI (approximately 0.02). Further, from the graph on the right the proportion of complete cases (in VOI) is 0.49, followed by the case where only ALC is missing (0.27), and the case where ALC and KCAL are both missing (0.12). In 10% of the cases, we find only KCAL missing.

2.4.3 Results

In this simulation, we do not apply CC, because there are no complete cases in the VOI. E2 was not applied either, because using basis-expanded covariates was not feasible on this data set. We compute EB, RMSE, and CICR based on Equations 2.9, 2.10, and 2.14.

2.4.3.1 Results - Bias

Figure 2.5 shows that CT, RF, and BA perform best and result in about the same EB in β_0. The parametric methods, BM and E1, show the highest absolute EB.

Similarly, Figure 2.6 shows that CT and RF perform best with approximately the same EB in β_{KCAL}. Again, BM and E1 lead to the highest EB.

Figure 2.7 shows that PM and E1 result in the smallest EB, while BA returns the biggest absolute EB.

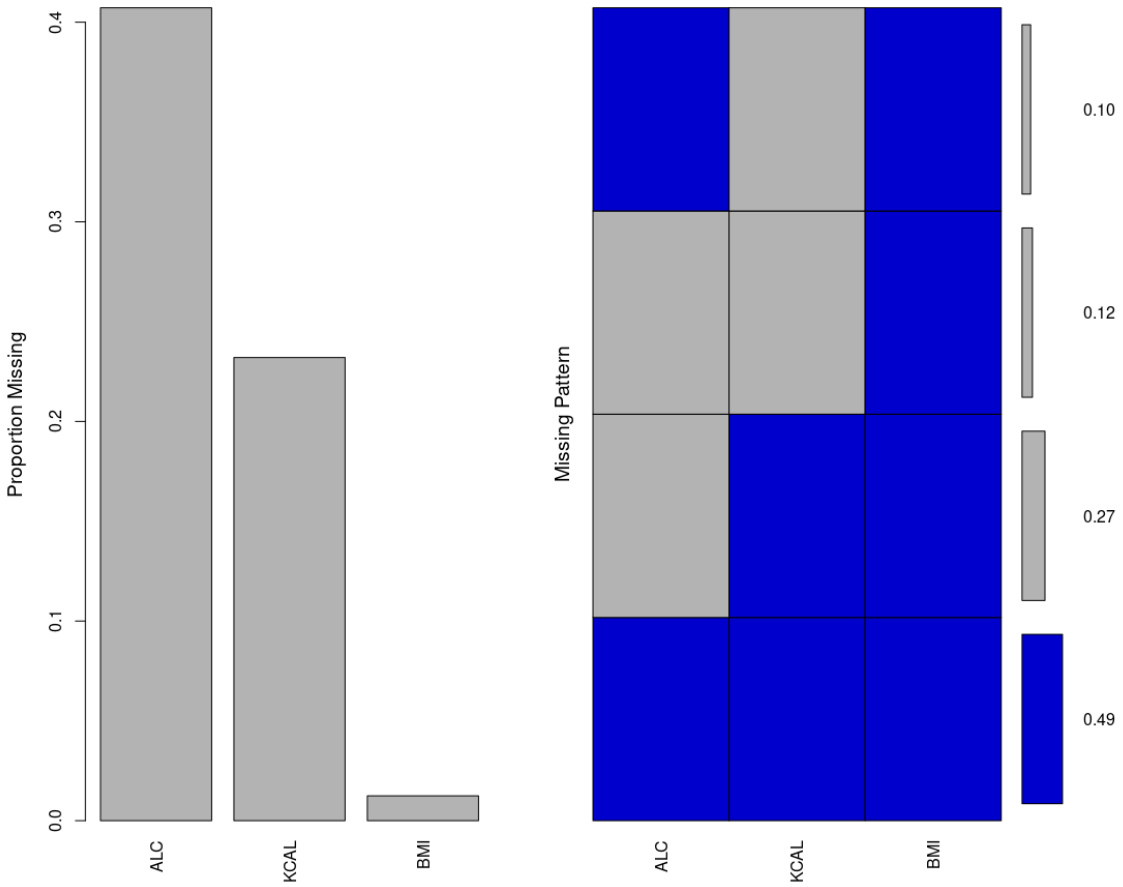

Figure 2.4: Missing data pattern of VOIs found in the NHANES 2015/2016 data, excluding missing patterns of frequencies smaller than 0.01. Blue indicates observed cases, and gray shows missing values.

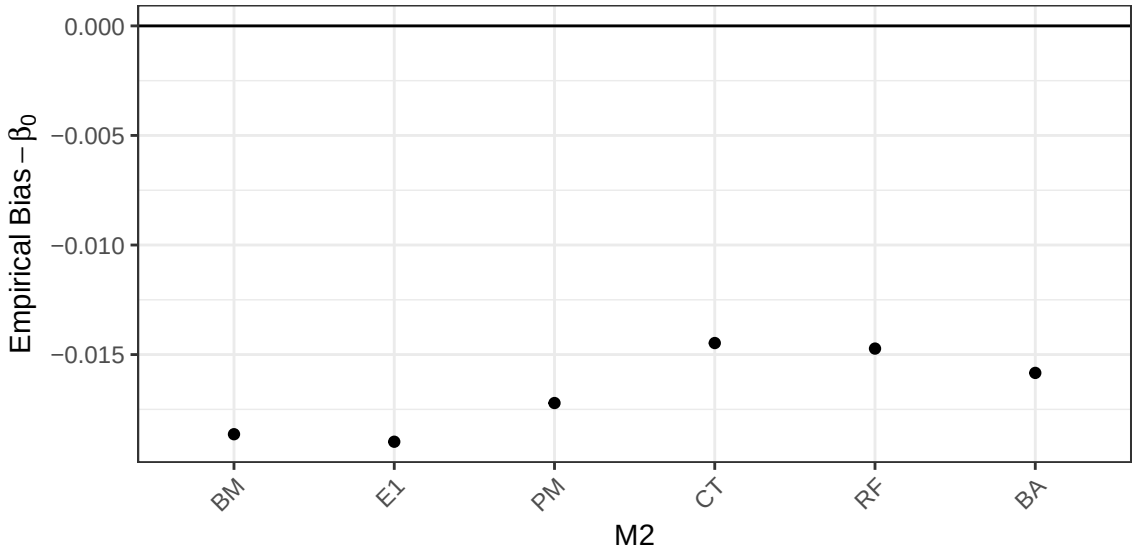

Figure 2.5: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_0 in M3 based on NHANES data. The solid black line indicates zero empirical bias.

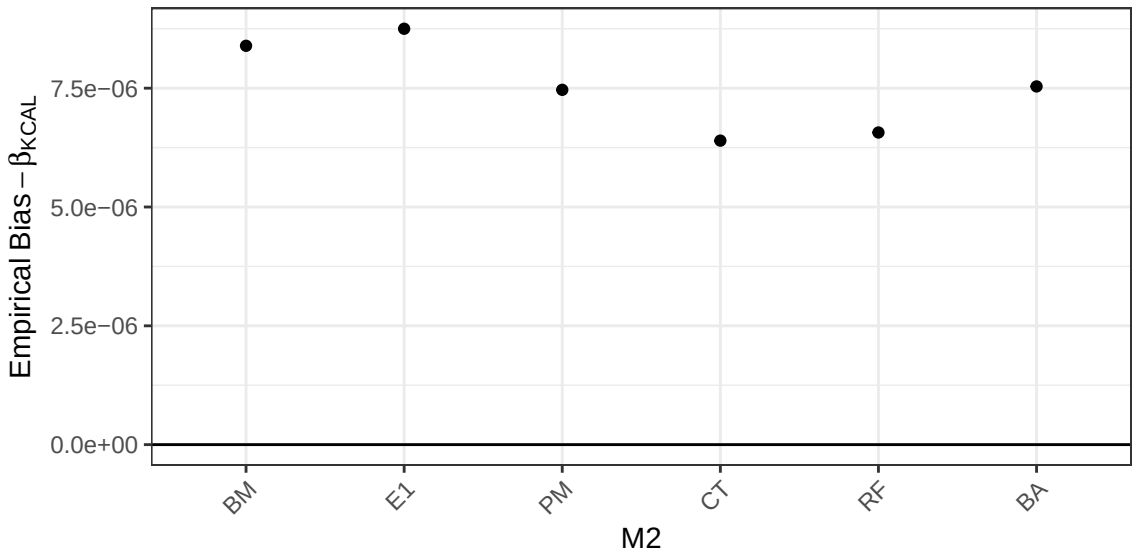

Figure 2.6: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_{KCAL} in M3 based on NHANES data. The solid black line indicates zero empirical bias.

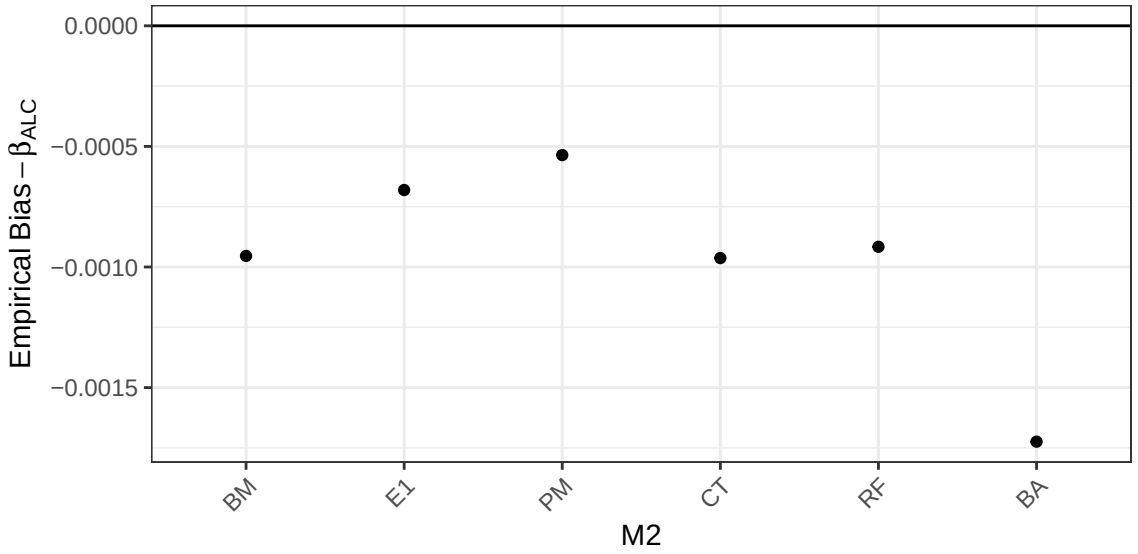

Figure 2.7: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_{ALC} in M3 based on NHANES data. The solid black line indicates zero empirical bias.

2.4.3.2 Results - RMSE

Figure 2.8 shows RMSE values for β_0 estimates for different M2s. Here, E1 performs best and yields the lowest RMSE. All other M2s show similar RMSE values.

A similar pattern is displayed in Figure 2.9 showing results for β_{KCAL}. Here, E1 returns the lowest, BA the highest RMSE values.

Figure 2.10 displays RMSE values for β_{ALC}, again, with E1 resulting in the lowest RMSE value and BA yielding the highest one.

2.4.3.3 Results - CICR

Figure 2.11 shows that BM produces the lowest CICR for β_0 (90%), while BA produces the highest result (94%), which is also closest to 95%.

For β_{KCAL} (Figure 2.12), we find that RF and E1 result in highest CICR (both at 99%). CT (94%), BM (94%), and BA (96%) are closest to 95%.

Figure 2.13 displays CICR for β_{KCAL}. Here, we find that E1 and RF result again in highest CICR (100%, 99%). BM results in the lowest coverage rate (91%). PM and BA (both at 94%) are closest to 95%.

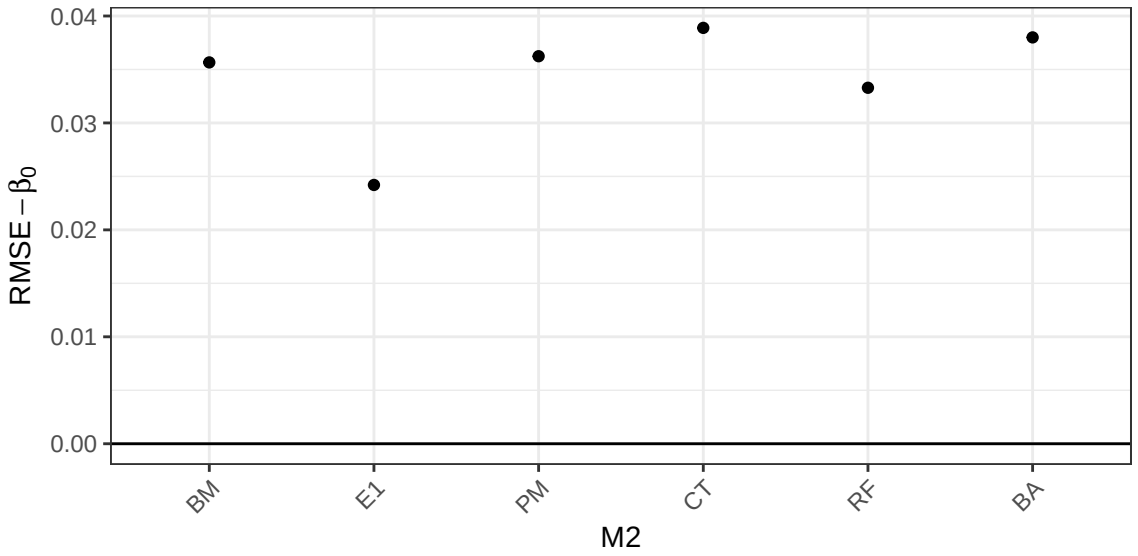

Figure 2.8: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_0 in M3 based on NHANES data. The solid black line indicates zero RMSE.

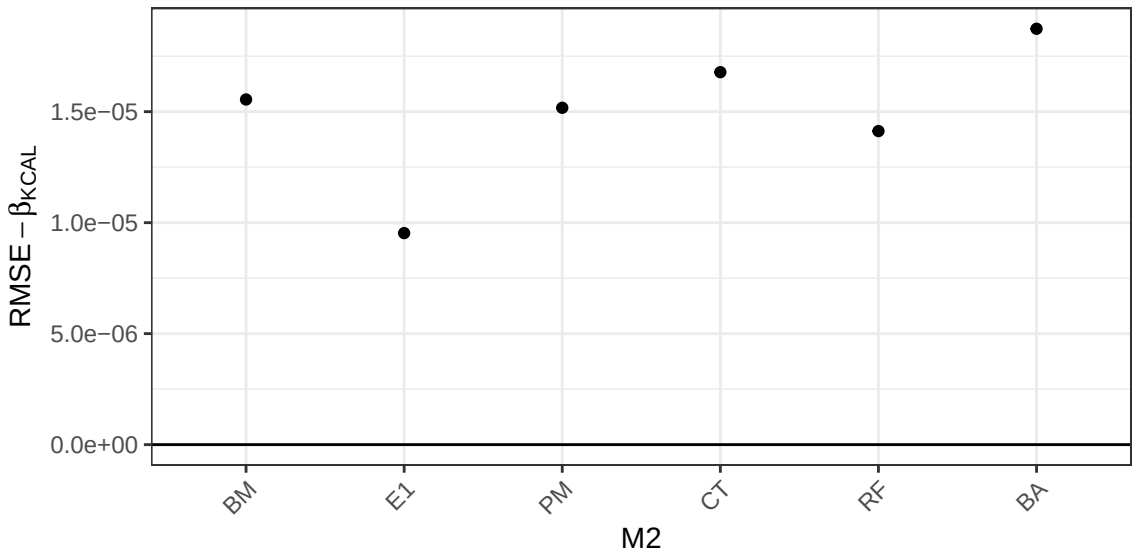

Figure 2.9: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_{KCAL} in M3 based on NHANES data. The solid black line indicates zero RMSE.

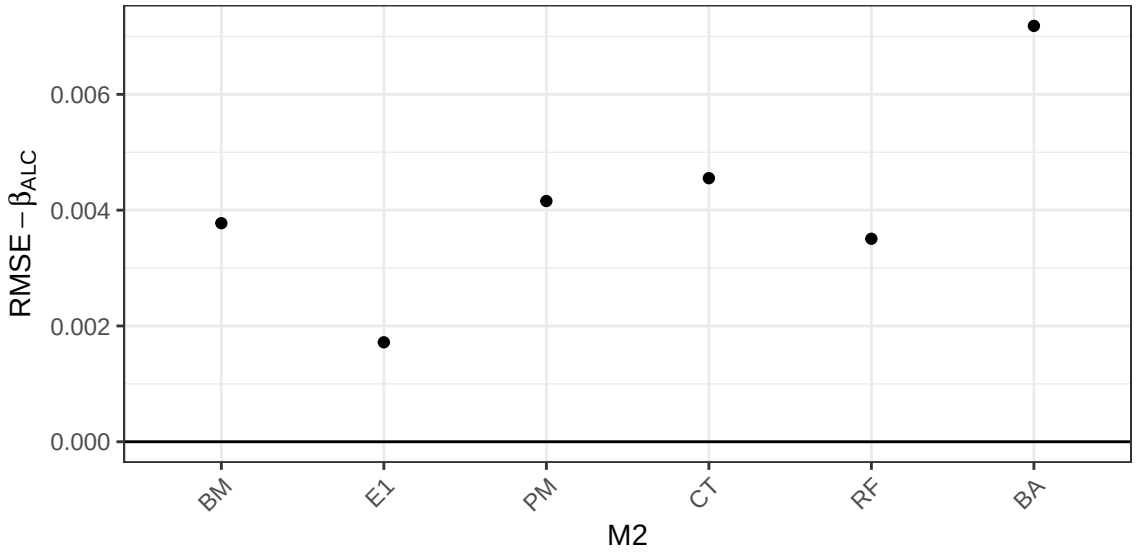

Figure 2.10: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_{ALC} in M3 based on NHANES data. The solid black line indicates zero RMSE.

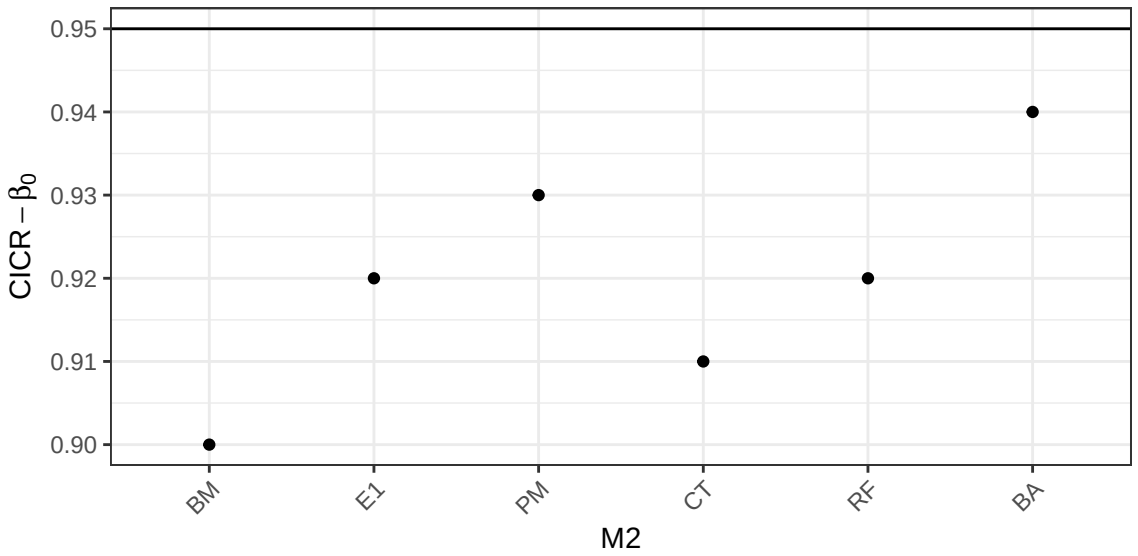

Figure 2.11: Different M2s compared in terms of the resulting confidence interval coverage rates (CICR) in the estimated regression coefficient β_0 in M3 based on NHANES data.

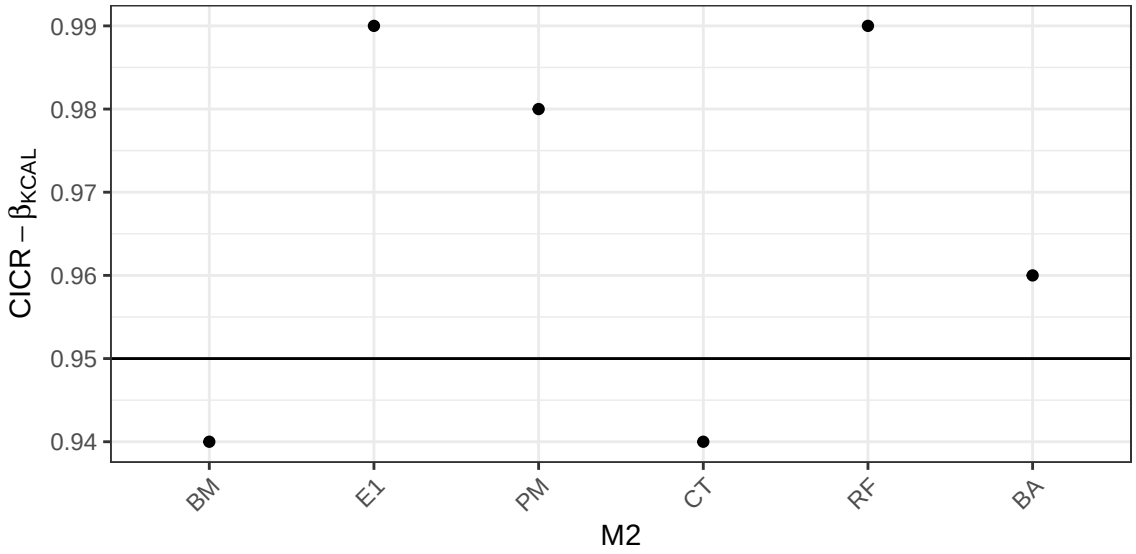

Figure 2.12: Different M2s compared in terms of the resulting confidence interval coverage rates (CICR) in the estimated regression coefficient β_{KCAL} in M3 based on NHANES data.

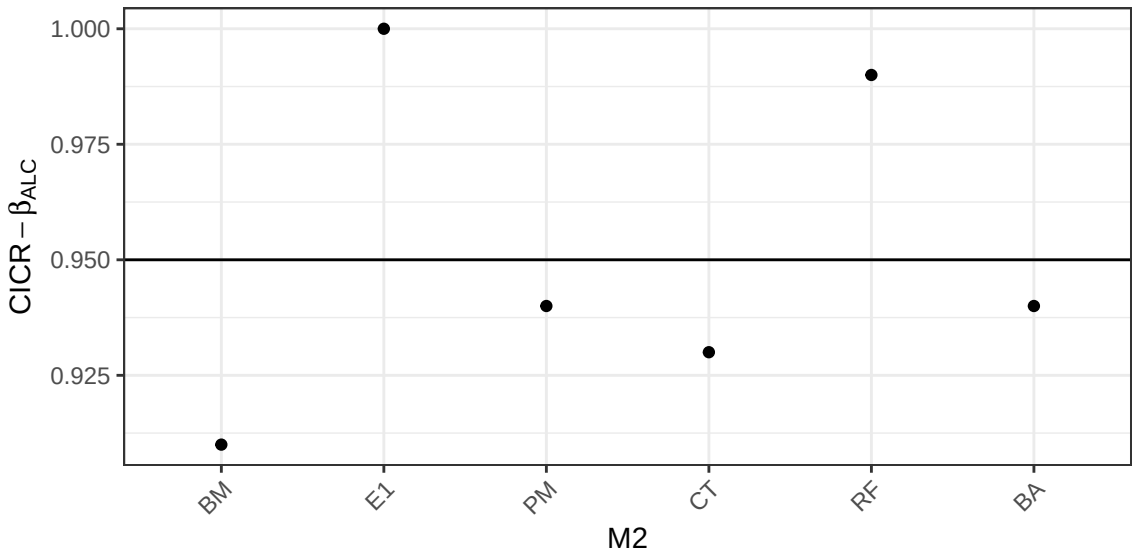

Figure 2.13: Different M2s compared in terms of the resulting confidence interval coverage rates (CICR) in the estimated regression coefficient β_{ALC} in M3 based on NHANES data.

We find no overall best procedure in terms of EB. However, we find that E1 performs best in terms of RMSE and BA results in the best CICR, over all three β coefficients.

In this simulation, we applied M2s on data with many (complete) covariates. In this case, we recommend procedures that include some sort of variable selection to remove unnecessary covariates and address potential collinearity. All applied imputation methods except BM and PM have such a feature.

2.5 Discussion

This study reveals how different imputation procedures perform under different data scenarios and how their performance depends on the underlying data situation. In this discussion section, we first restate the main findings of the simulation, followed by the limitations of the study and directions for future research.

Summarizing the findings of the simulation, we see that, as expected, CC shows good performance in all scenarios where the probability of missingness only depends on the covariates. When missingness also depends on the outcome, we see that, overall, PM performs well in the parametric scenarios, while BA yield the best results in the non-parametric cases. The parametric methods (BM, E1, and E2) show similar patterns in most of the investigated scenarios. CT works well in all MAR scenarios where missingness is independent of the outcome. While we find that BA generally performs well in non-parametric cases, BA often shows the worst performance in parametric cases.

The investigation of the unexpected behavior of BA reveals the following. Imputation using BA on data with increased variance leads to higher absolute empirical bias, compared to imputation using CT. The tested BA implementations from different packages show overall similar patterns in terms of empirical bias, but selecting hyperparameters and prior-settings via cross-validation helps to reduce strong biases. Increasing the sample size leads to reduced empirical bias for all compared procedures, but the overall patterns remain the same.

Regarding the simulation using real data, CT and RF perform well in terms of EB, but E1 shows the best performance for RMSE in all three parameters. This finding suggests that applying regularization can remove unnecessary "noisy" terms in settings

with more covariates present, as is the case in the NHANES data. For CICR, BA shows overall the best coverage results.

The results from the simulations reveal no best overall imputation method. Therefore, missing data imputation in practice requires a careful consideration of the best imputation method for the given situation. We now provide recommendations for practice based on the given data set and other aspects. If the analysis model is a regression with a known outcome and covariates, not imputing the missing values can be considered; CC leads to consistent regression coefficients if missingness does not depend on the outcome. As seen in the simulation, compared to the imputation methods, CC works well in many of the investigated scenarios. In general, utilizing information from the incomplete observations requires effort in terms of diagnostics of the imputation process.

When the number of variables is small, modeling effort is likely manageable and parametric models like BM can be properly modified. If the analyst can not put in the effort to select appropriate imputation models, the simulations suggest that PM without modifications can be used as a robust alternative to BM. In case of a high number of (incomplete) variables, we recommend the use of automated imputation methods as the sequential imputation with integrated model selection (SIIMS) procedure introduced in Chapter 3 of this dissertation. Another recently published automated imputation procedure is multiple imputation by super learning (MISL) (Carpenito & Manjourides, 2022). An advantage of both SIIMS and MISL is that they use several imputation models in each SI step, which eliminates the need for a preceding model selection.

The current simulation investigates scenarios with congenial M1s and M3s. Thus, in data where the probability of missingness only depends on the covariates, we find CC performing well. However, for uncongenial M1 and M3, this finding might not hold. The missing values might have a stronger effect on parameter estimates in a simplified M3 (compared to M1), while imputation with a M2 congenial to M1 can restore the joint distribution of the outcome and covariates. Thus, a simplified M3 can still result in consistent estimates for uncongenial M1s and M3s when imputation is performed.

Machine learning (ML) procedures like tree-based methods perform well in prediction tasks and their default parameter settings were often determined from studies with assessment based on straight forward prediction. While missing data imputation is a

form of prediction, there are some unique aspects to imputation. CEMI is an iterative procedure and additional parameter tuning in each CEMI step can lead to increased runtime. For instance, we experienced strongly increased runtime when applying BA with parameter tuning via cross-validation within CEMI (Section 2.3.0.1). On the other side, RF and CT default values often lead to good performance, in contrast to neural networks (Nordbotten, 1996) or support vector machines (Y. Zhang & Liu, 2009), where parameter tuning is highly recommended. The resulting time advantage of RF and CT is the main reason for mostly focusing on tree-based ML methods in this study. However, generally, optimal parameters for machine learning procedures within CEMI lack investigation. For instance, Shah et al. (2014) compare CEMI using RF with different parameter settings and find, for example, that there is no further gain in performance when increasing the number of trees above ten. However, such studies are relatively rare. Future studies can investigate the sensitivity to non-optimal parameters and the effect of parameter tuning within CEMI. These studies can examine parameter tuning in CEMI within time constraints (Z. Wang et al., 2018) to determine feasibility for application in larger data sets.

Despite the variety of multiple imputation methods compared, this study is limited in several aspects. First, the simulated scenarios are kept simple and focus only on continuous variables. We set up these scenarios so that relationships among variables are obvious and can be adjusted as needed in order to learn about the behavior of different imputation procedures. Reality is, of course, complex, and analysis models with only three continuous variables are rare. Future research can build on this current study and investigate the performance of imputation procedures in scenarios with a higher number of variables, including binary and nominal variables.

Second, in the simulation, we find the parametric methods (BM, E1, and E2) mostly perform in similar ways, for instance, they result in similar EB values. Care should be taken in generalizing this conclusion because these methods have different underlying model structures. However, it is possible that the regularized methods identify active sets of covariates that result in models close or equal to the BM. A similar performance of BM and the regularized methods in the {Q,.} cases supports this explanation. Further, in M1, the data generating model, we specify an incomplete variable uniformly distributed, thus, all parametric methods are misspecified to some degree. Varying the strength of the quadratic relationship in M1 can help to compare the performances among the parametric imputation procedures.

Third, this study assumes simple random sampling of the data, ignoring complex sampling features like weights and clusters. Work by Zhou, Elliott, and Raghunathan (e.g., Zhou et al., 2016b, 2016a) provides a two-step approach for incorporating these complex sampling features of survey data in multiple imputation. In their approach, the first step is based on the finite population Bayesian bootstrap, and the second step consists of imputing missing values using a parametric model. Future studies could evaluate how substituting this parametric model with other procedures compared in this study affects performance.

Fourth, the setup of the data generating process could lead to unintended MNAR situations when missingness of an incomplete variable depends on missing values of another incomplete variable. An alternative to sequentially generating missing value indicators is introducing missing values based on missing patterns as in the MICE function "ampute()" (Schouten et al., 2018). This procedure uses weighted scores of the variables and assures the defined missingness mechanism.

2.6 Appendix 1 - table of acronyms

Table 2.8: Table of acronyms used in Chapter 2, ordered by as they appear in the text.

Acronym	Description
CEMI	Chained equation multiple imputation
MAR	Missing at random
MNAR	Missing not at random
PM	Predictive mean matching
CT	Classification and regression trees
RF	Random forest
MICE	Multivariate Imputation by Chained Equations
BA	Bayesian additive regression trees
RMSE	Root mean squared error
NHANES	National Health and Nutrition Examination Survey
M1	Data generating model
M2	Imputation model
M3	Analysis model
L	Linear case
Q	Quadratic case
N	Non-parametric case
MX	MAR, missingness independent of outcome of M1/M3
MXY	MAR, missingness dependent of outcome of M1/M3
NX	MNAR, missingness independent of outcome of M1/M3
BM	Bayesian linear model
E1	Bayesian regularized linear model using elastic net
E2	E1 with basis-expanded covariates
CC	Complete case analysis
MCMC	Markov chain Monte Carlo
EB	Empirical bias
RV	Ratio of mean estimated variance to the empirical variance
CI	Confidence interval
CICR	CI coverage rate
S-EB	EB measure for spline in N case
S-RV	RV measure for spline in N case
S-RMSE	RMSE measure for spline in N case
S-CICR	CICR measure for spline in N case
BA-CV	BA including cross-validation
BA-P	BA as implemented in the R package "BART"
BA-S	BA designed for smooth functions
VOI	Variables of interest
SRSWR	Simple random sample with replacement
BMI	Body mass index
KCAL	Kilo-calories intake
ALC	Number of days having 4 to 5 alcoholic drinks in the past 12 months
SIIMS	Sequential imputation with integrated model selection
MISL	Multiple imputation by super learning

2.7 Appendix 2 - design table

Table 2.9: Design table for Chapter 2.

Method	Parameter	Description	Levels	Choices	Tuning
NORM	-	No tuning parameters	-	-	None
E1, E2	α	Elastic net mixing parameter	$[0, 1]$	Interval of all possible values.	5-fold cross-validation
PMM	-	No tuning parameters	-	-	None
CT	minbucket	The minimum number of observations in any terminal node used.	5	Default of MICE package version 3.14.7	None
CT	cp	Complexity parameter	1e-04	Default of MICE package version 3.14.7	None
RF	b	Number of trees	10	Provided by Shah et al. (2014)	None
RF	minbucket	The minimum number of observations in any terminal node used.	5	Default of MICE package version 3.14.7	None
BA	mem_cache_for_speed	Speed enhancement that caches the predictors and the split values that are available at each node for selecting new rules.	TRUE, FALSE	Recommened for large number of predictors, set 'FALSE' in simulation with three variables (Section 2.2.1), set 'TRUE' in simulation using real data (Section 2.4).	None
BA	use_missing_data	If TRUE, incomplete observations are included.	FALSE	Only complete observations are used in each CEMI step.	None
CC	-	No tuning parameters	-	-	None
BA-CV	m	Number of trees	50, 200	Recommended in Chipman et al. (2010)	5-fold cross-validation
BA-CV	k_{BA}	Hyperparameter influencing the effect of a single tree component.	$\{1, 2, 3, 5\}$	Recommended in Chipman et al. (2010).	5-fold cross-validation
BA-CV	(u, q)	Pair of prior degrees of freedom and quantile, see Chipman et al. (2010) for detail.	$\{(10, 0.75), (3, 0.90), (3, 0.99)\}$	Recommended in Chipman et al. (2010).	5-fold cross-validation
BA-P	-	No tuning parameters	-	-	None
BA-S	-	No tuning parameters	-	-	None

All other parameters of the presented imputation procedures are specified as in the corresponding software packages.

2.8 Appendix 3 - detailed results

The plots in this appendix show results of the described simulation study and follow the same basic organizational scheme. The y-axes show the quantitative properties (i.e., either EB, RMSE, or CICR); the x-axes display the different M2s. The figures' columns show the results for the different parameter values of $logit(\delta_0^{X_1}) \in \{0.05, 0.15, 0.5\}$; the rows are separated by the different missingness mechanisms. The dots display the simulation results and compare M2s.

Parametric Data

EB - {L,.}

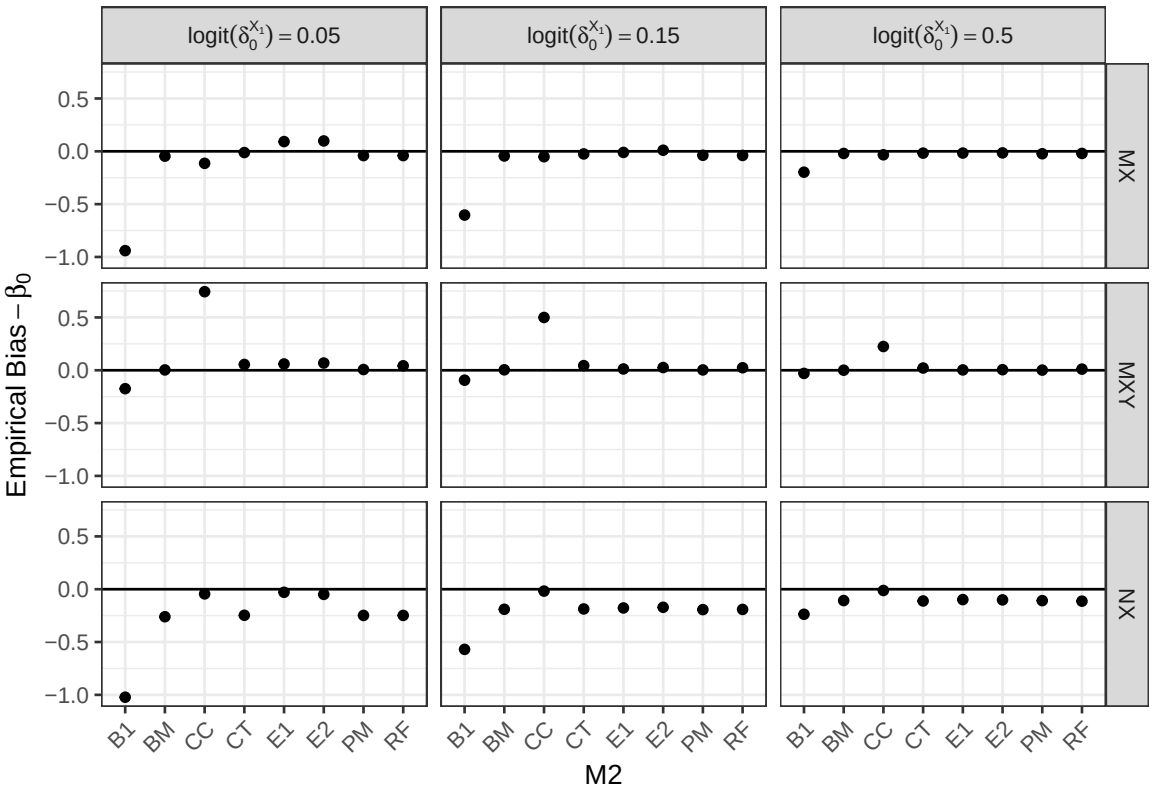

Figure 2.14: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_0 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero empirical bias.

We first focus on EB in the {L,MX} case. Figures 2.14, 2.15, and 2.16 show that the imputation by BA results in the largest EB in all three M3 coefficients. E1 and E2

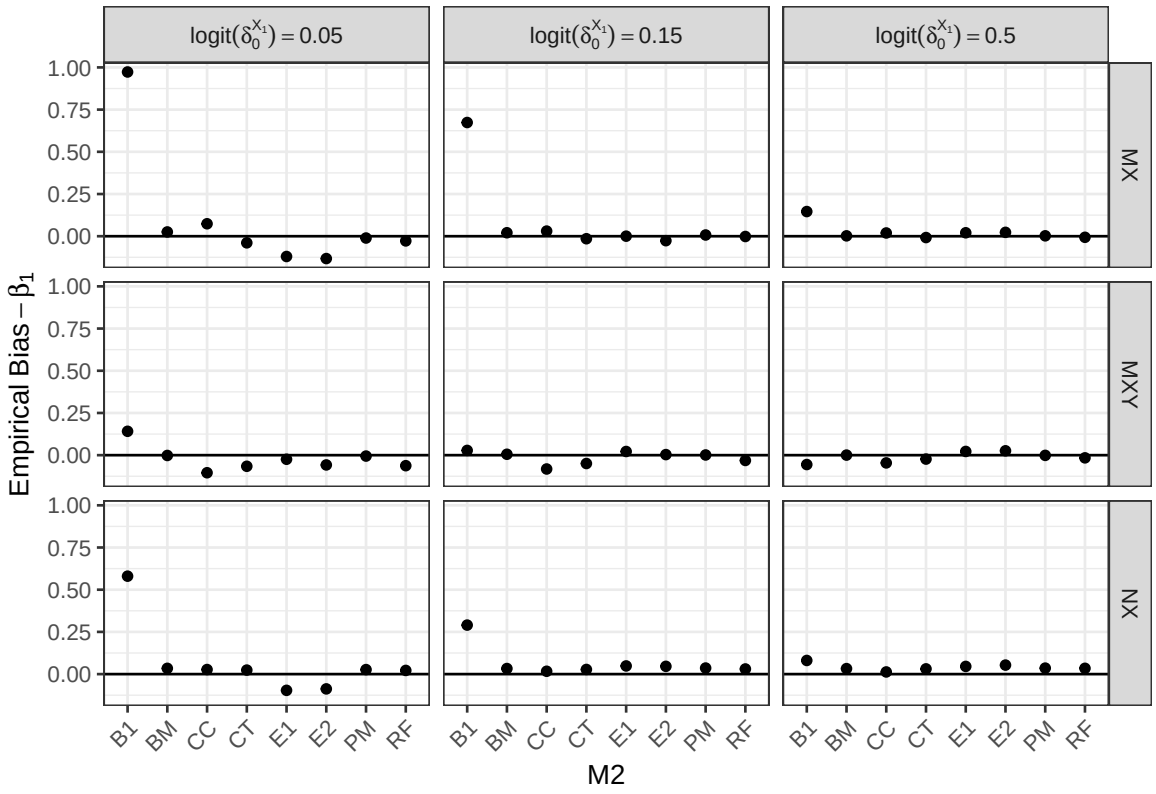

Figure 2.15: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_1 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero empirical bias.

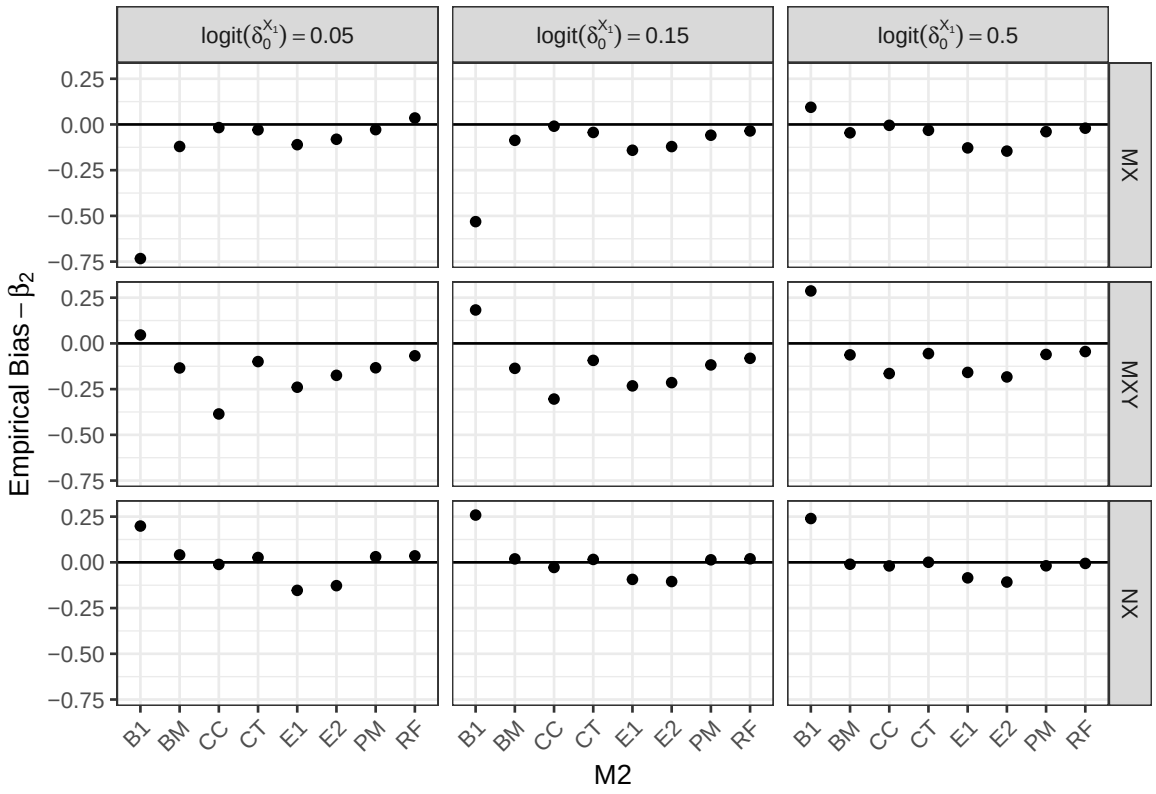

Figure 2.16: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_2 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero empirical bias.

show similar performances with a small EB in all three coefficients. CC, CT, BM, PM, and RF also perform on a similar level with mostly small EB.

For EB in the {L,MXY} case, we see differences in EB for the three coefficients of M3. For the intercept (Figure 2.14), CC has the largest EB, followed by BA. All other imputation models perform similarly well with only small EB values. For β_1 (Figure 2.15), BA results in a small EB; all other M2s show no or a very small EB. For β_2 (Figure 2.16), we see the largest EB in CC, followed by E1; all other M2s showing small EB values.

For EB in {L,NX}, the imputation by BA results in the largest EB in β_0 and β_1. E1 and E2 perform similarly well with a small EB in all three coefficients. CT, BM, PM, and RF are mostly empirically unbiased. CC is empirically unbiased in this scenario. For β_2, the compared methods are mostly empirically unbiased; BA, E1, and E2 show small EBs.

The described patterns appear to be the same for different values of $logit(\delta_0^{X_1})$, but tend to be strongest for $logit(\delta_0^{X_1}) = 0.05$ and weakest for $logit(\delta_0^{X_1}) = 0.5$. In other words, the patterns are strongest for the highest number of missing values.

EB - {Q,.}

For {Q,MX}, we find that CC results in only small EB in parameter estimates for M3. For β_0, β_1, and β_3 (Figures 2.17, 2.18, and 2.20), we find similar magnitudes of EBs in estimates for BM, BA E1, and E2 (β_0 and β_3 negatively, β_1 positively biased), except for BA, which shows a less strong EB in the β_3 case. For β_2 (Figure 2.19), we find that BA results in the strongest EB; all remaining M2s are mostly empirically unbiased.

For {Q,MXY}, we find that BM, E1, and E2 results in EB for all parameters, except for β_2, where no method shows a large EB. All other M2s result in either no or small EB. One exception of this pattern is the large EB for CC in the β_0 estimate.

For {Q,NX}, we find that CC results in small EB in parameter estimates for M3. For the β_0, β_1, and β_3, we find similar magnitudes of EB in estimates for BM, BA, E1, and E2 (β_0 and β_3 negatively, β_1 positively biased), except for BA, which shows a less strong EB in the β_3 case. In β_3, we also find positive EB in CT, PM, and RF. For β_2, all M2s are mostly empirically unbiased.

RMSE - {L,.}

The RMSE for {L,MX} is highest for BA in all estimated coefficients (cf. Figures 2.21,

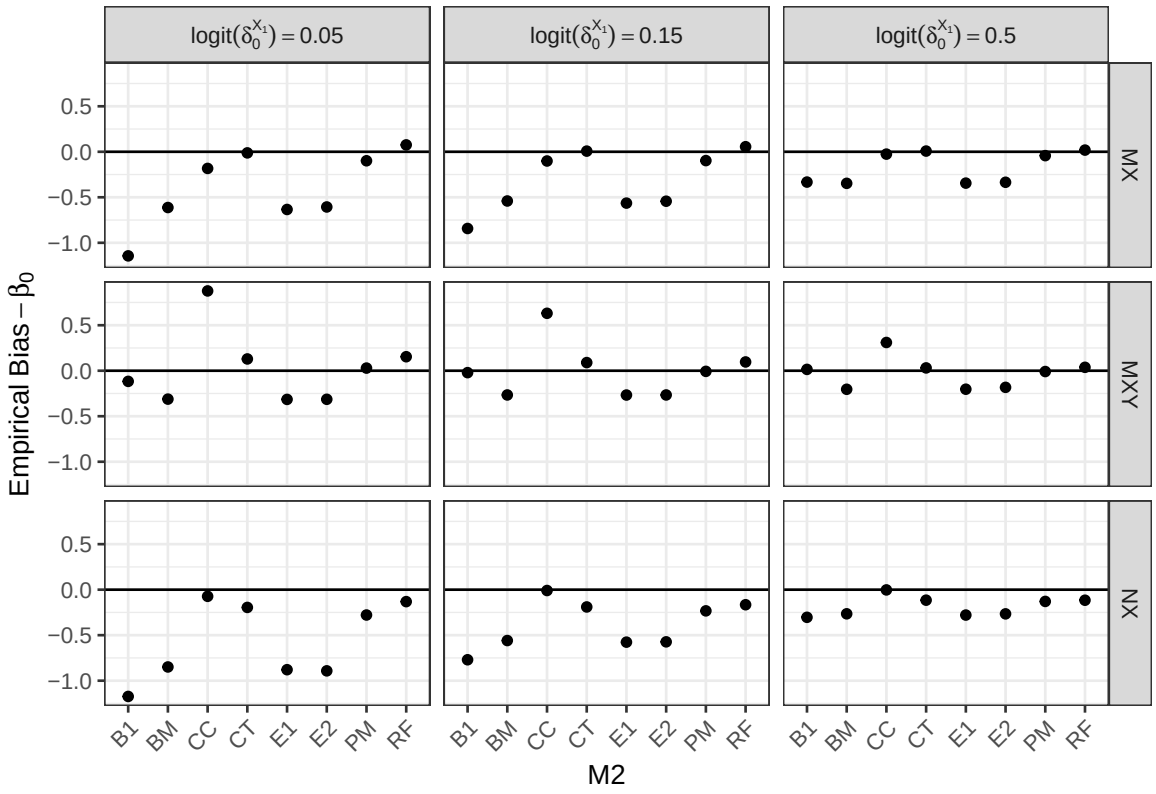

Figure 2.17: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero empirical bias.

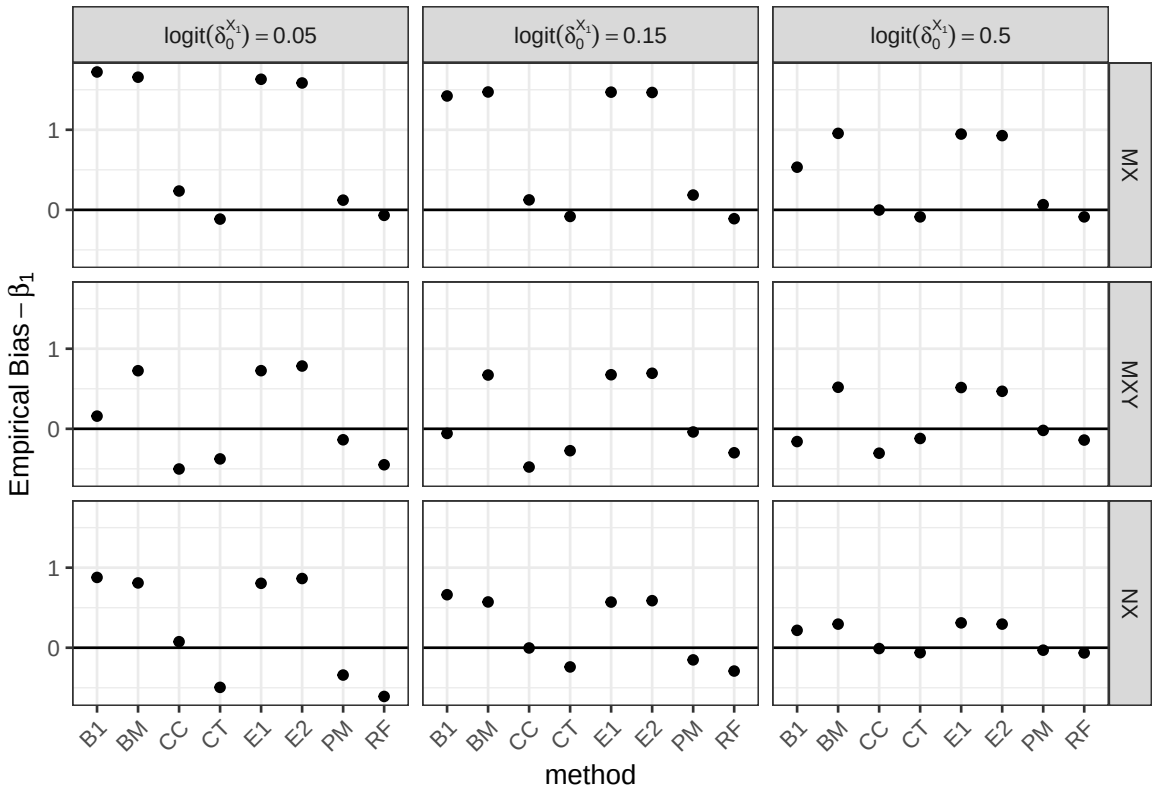

Figure 2.18: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_1 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero empirical bias.

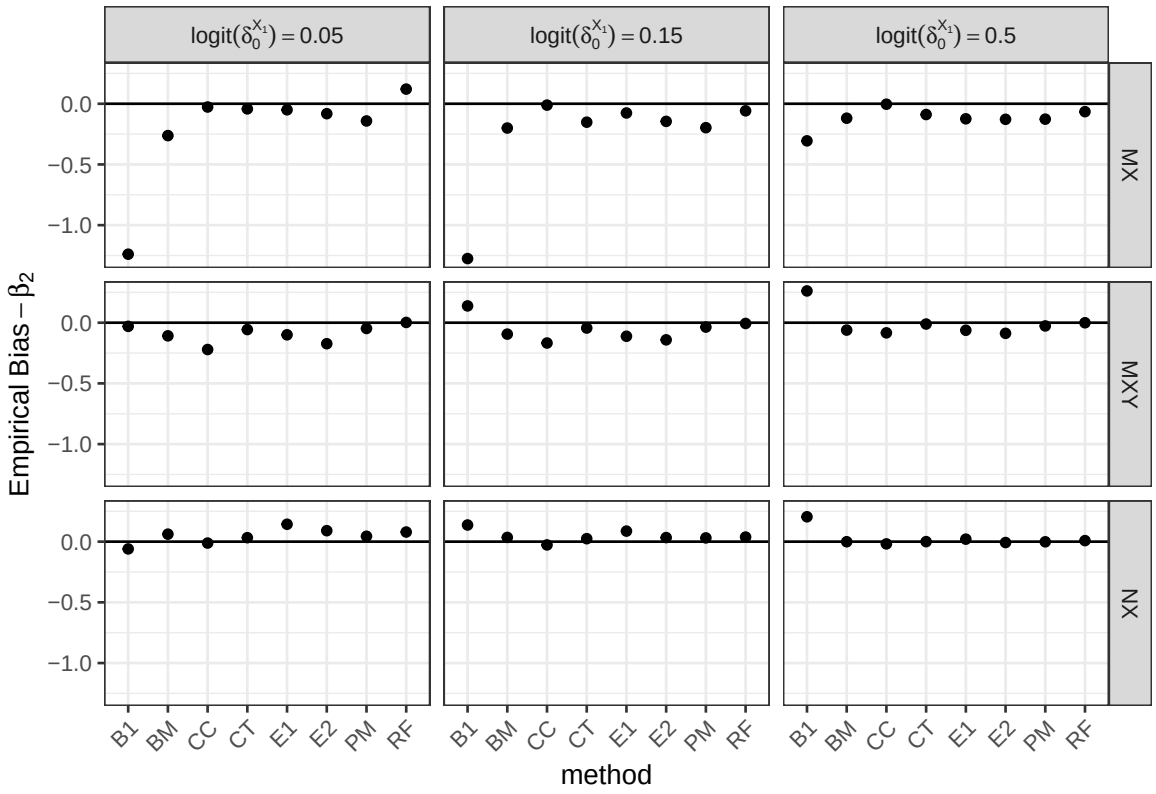

Figure 2.19: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_2 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero empirical bias.

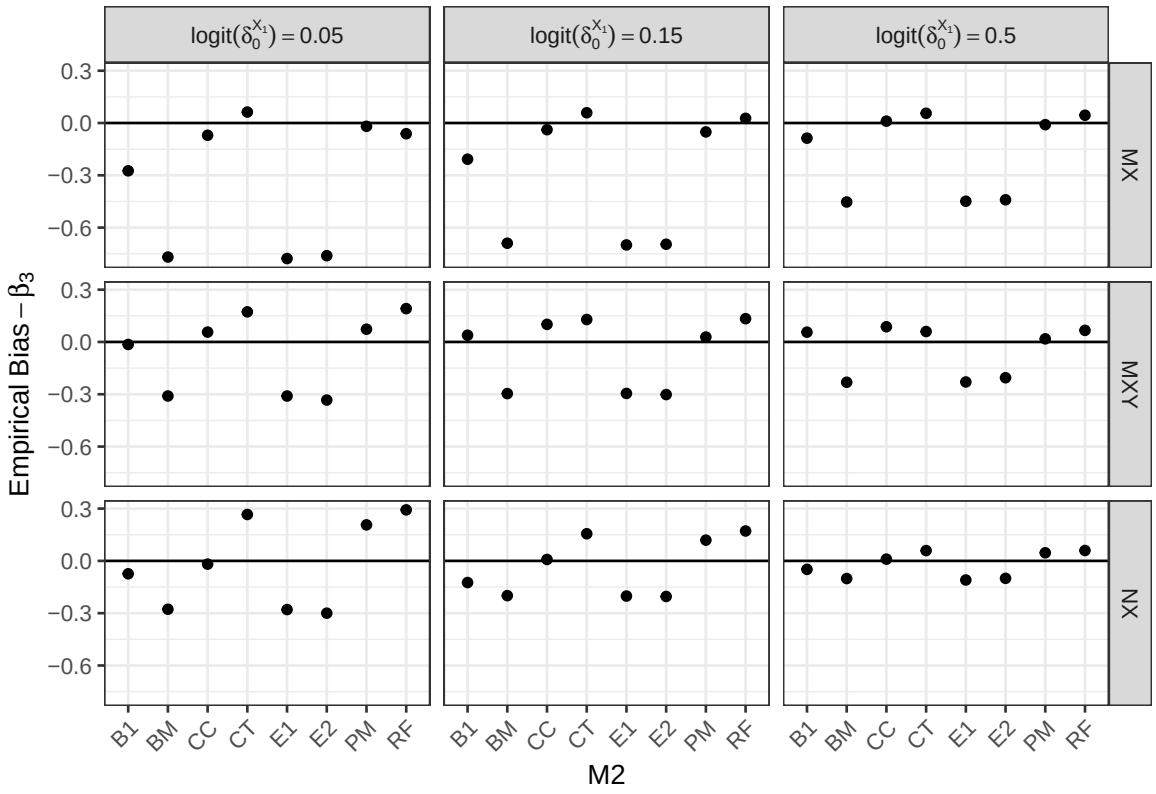

Figure 2.20: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_3 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero empirical bias.

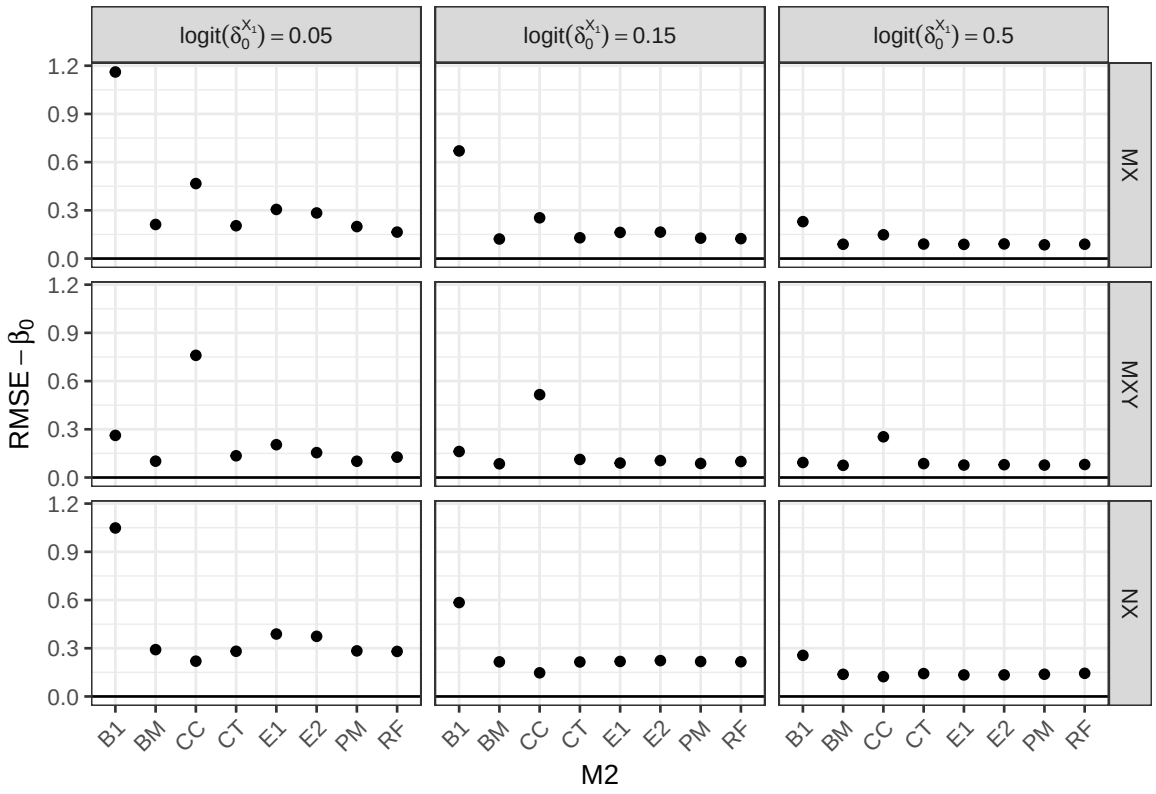

Figure 2.21: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_0 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero RMSE.

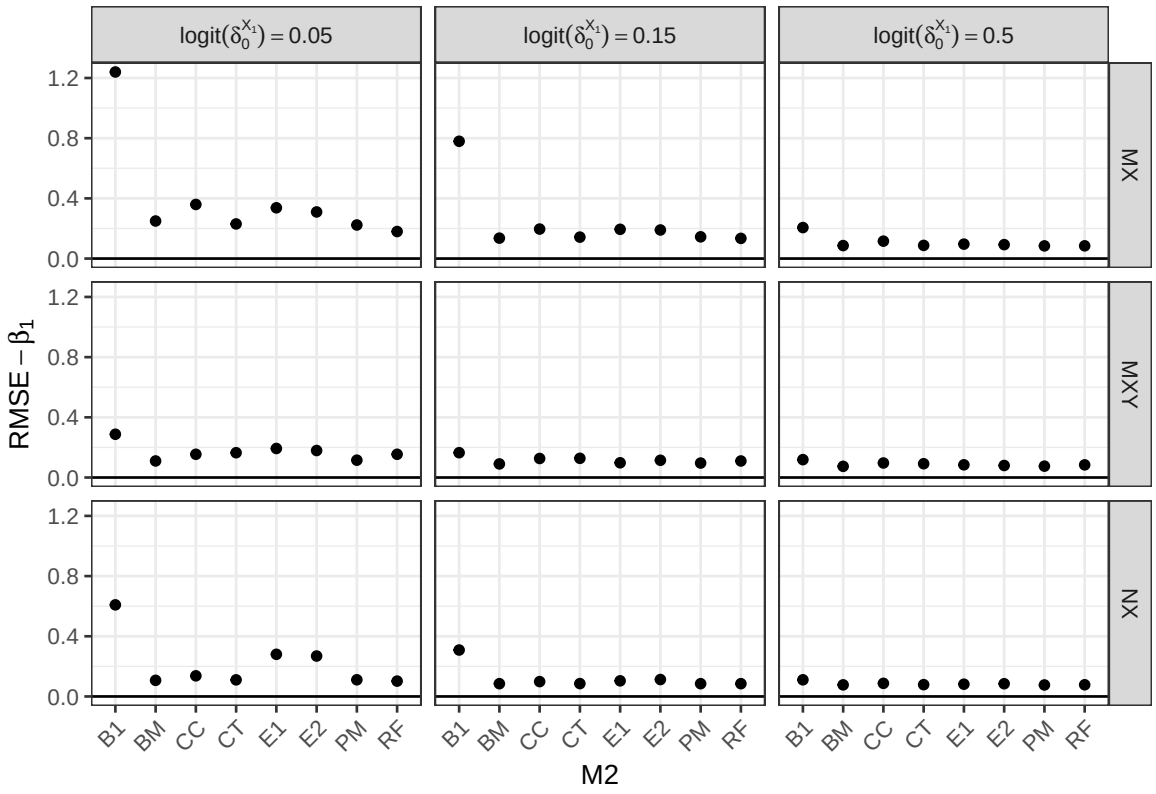

Figure 2.22: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_1 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero RMSE.

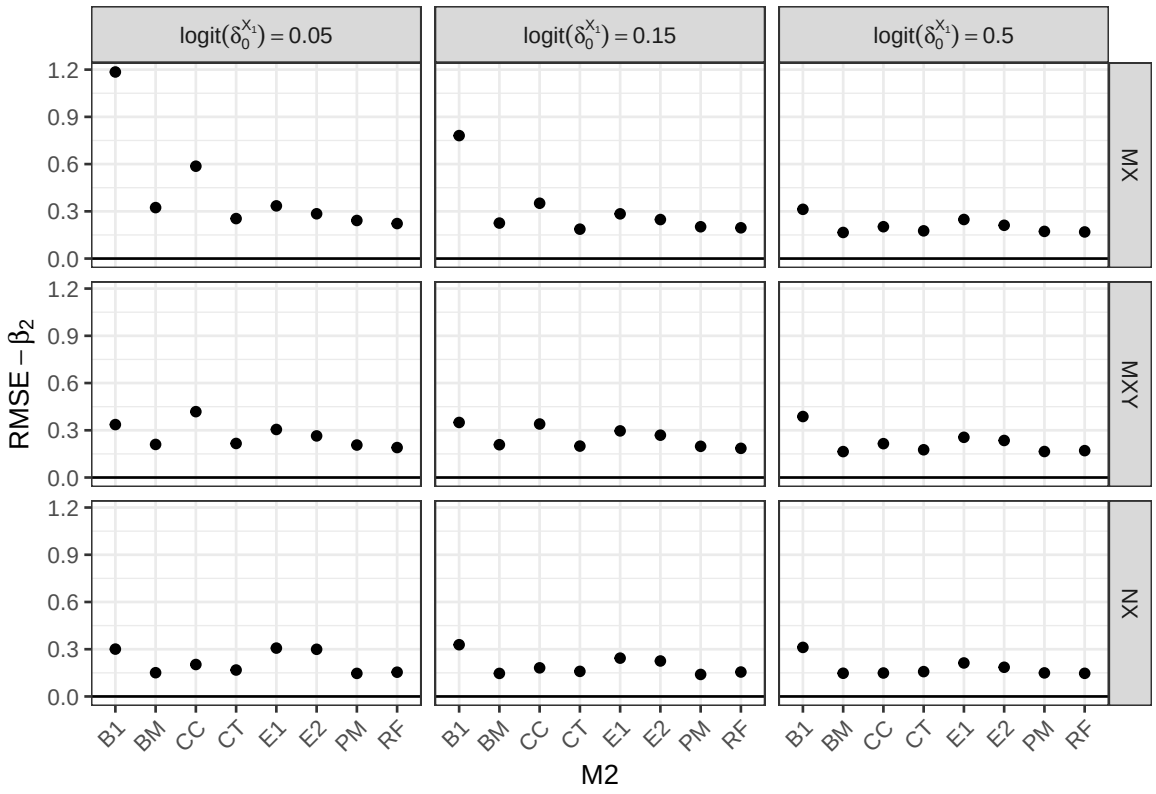

Figure 2.23: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_2 in M3 for nine different data scenarios with linear relationships. The solid black line indicates zero RMSE.

2.22, 2.23), followed by CC, and then by all other M2s. Overall, RF performs best in terms of RMSE in this scenario.

For {L,MXY}, we generally see lower RMSE, compared to {L,MX}. Here, CC results in a noticeable higher RMSE value for β_0; otherwise, the values are similar.

For {L,NX}, RMSE is highest for BA in β_0 and β_1. For β_2, only small differences occur, with E1, E2, and BA resulting in highest RMSEs, followed by all other methods.

RMSE - {Q,.}

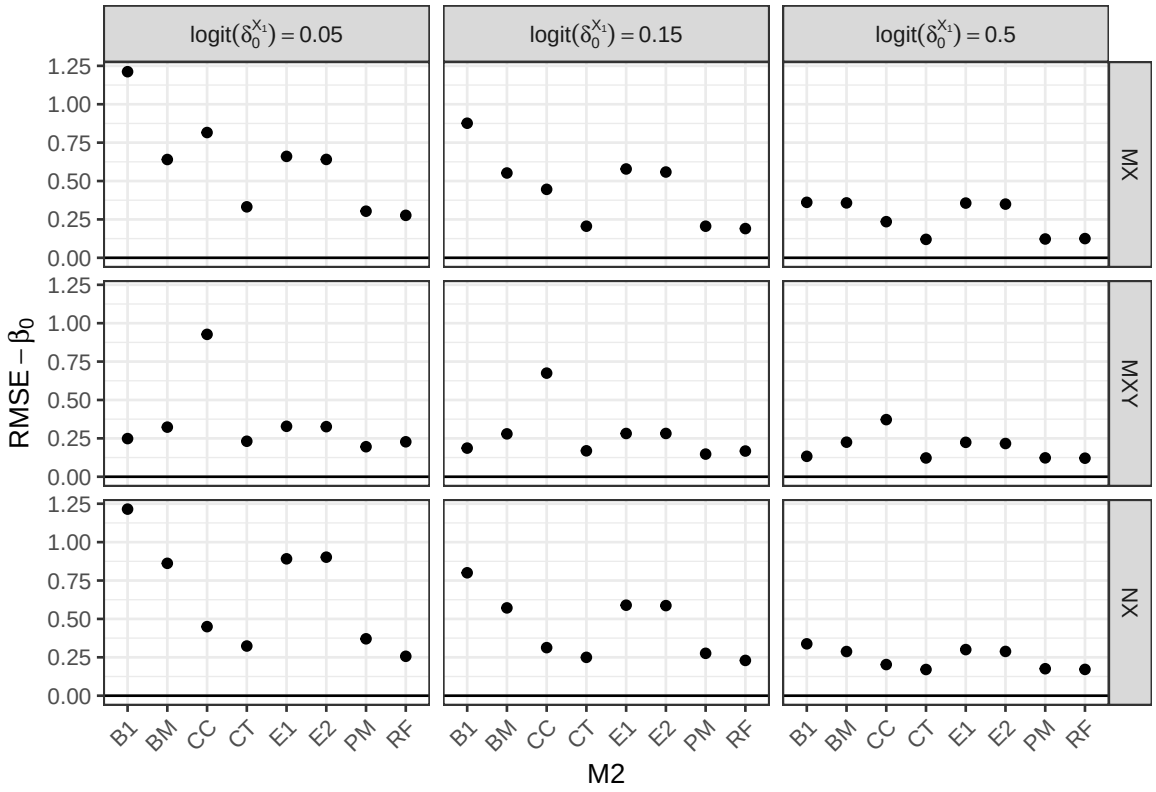

Figure 2.24: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero RMSE.

For {Q,MX}, BA results in the highest RMSE in β_0 (Figures 2.24). In β_1 (Figures 2.25), CT, PM, and RF return the lowest RMSE values. In β_2 (Figures 2.26), only BA yields a noticeably higher RMSE values compared to the other imputation methods. For β_3 (Figure 2.27), CT, PM, and RF result in similar low RMSE values compared to all other M2s.

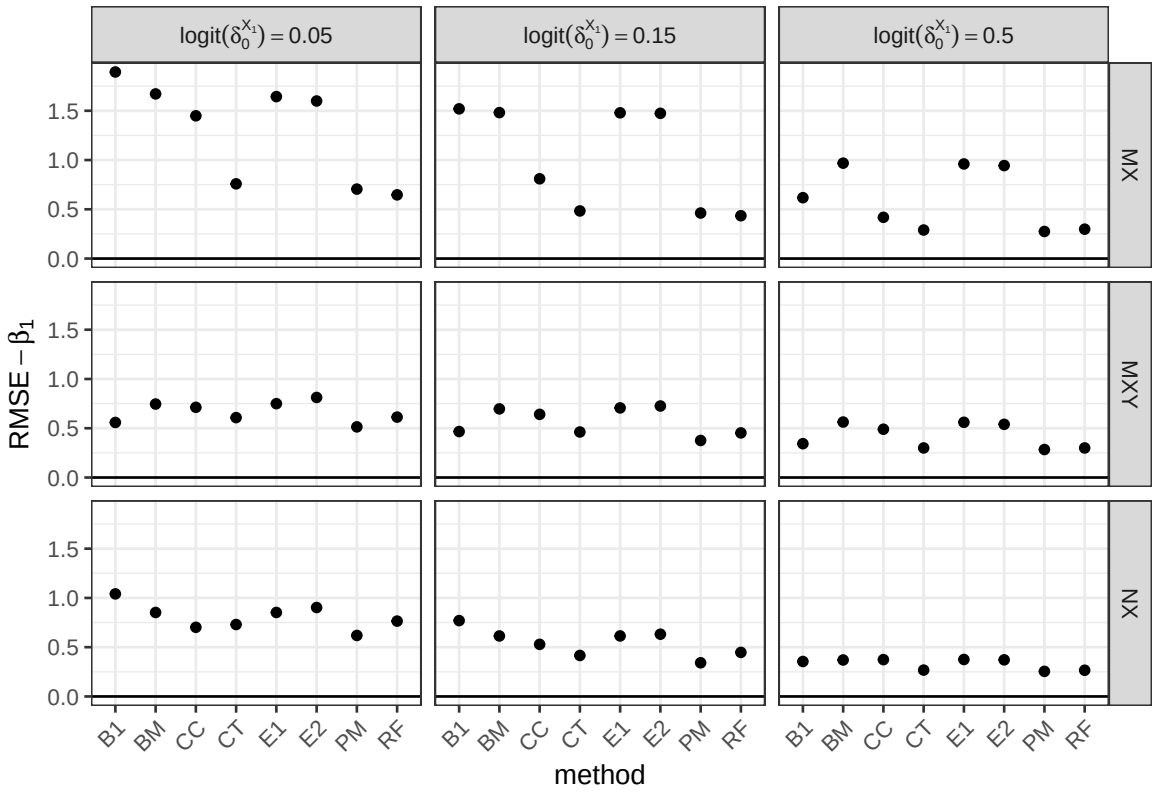

Figure 2.25: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_1 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero RMSE.

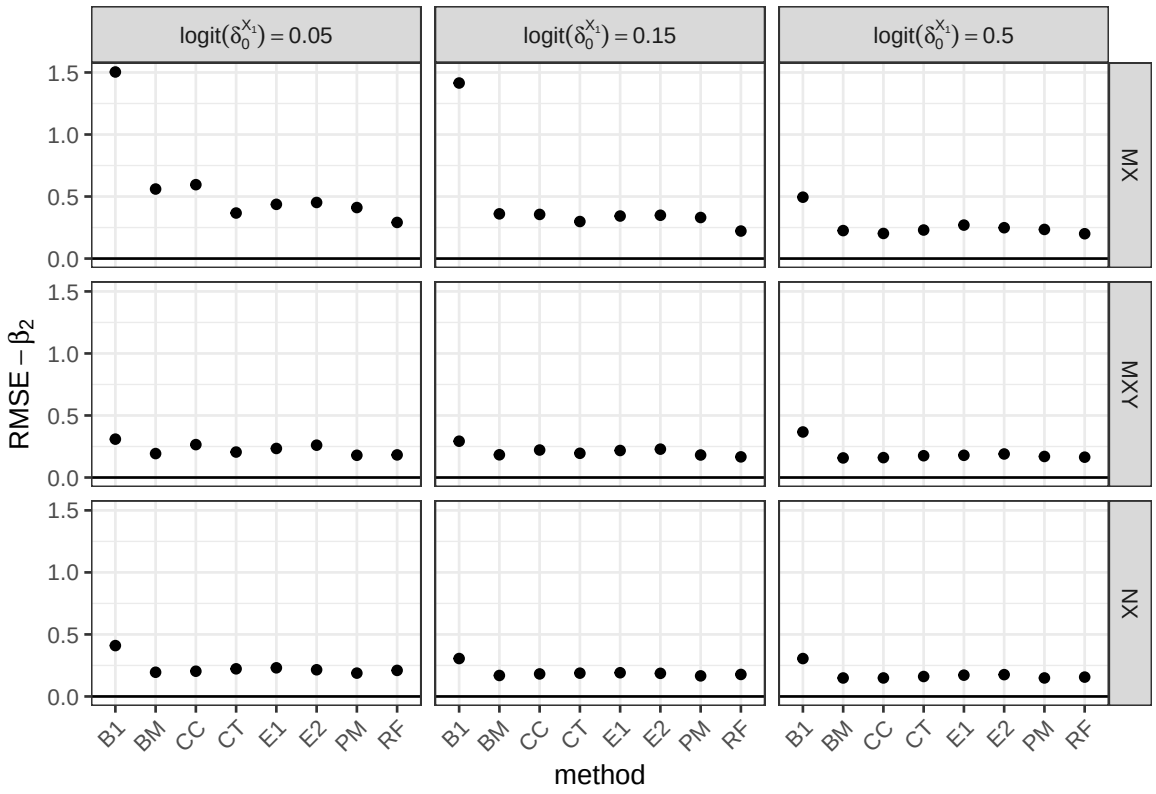

Figure 2.26: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_2 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero RMSE.

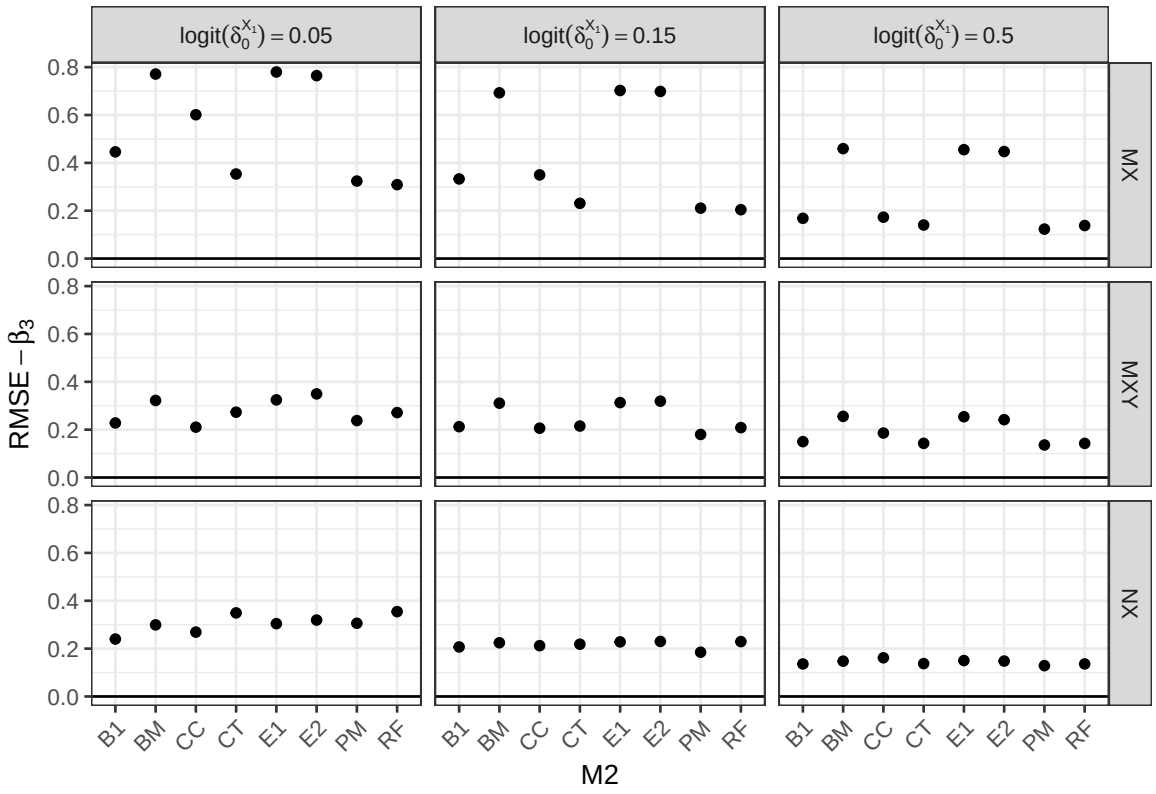

Figure 2.27: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_3 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates zero RMSE.

For {Q,MXY}, we again find lower RMSE values, compared to {Q,MX}. All M2s are generally at the same performance level, except for β_0, where CC yields higher values.

In the {Q,NX} case, the biggest differences in M2s occur in β_0. BA shows highest RMSEs, followed by BM, E1, and E2. CC, CT, PM, and RF show lowest RMSEs, all three being at a similar level. For β_1 and β_2, we find only minor differences in resulting RMSE values, with BA showing the highest values. The RMSE values in β_3 are on a similar level.

CICR - {L,.}

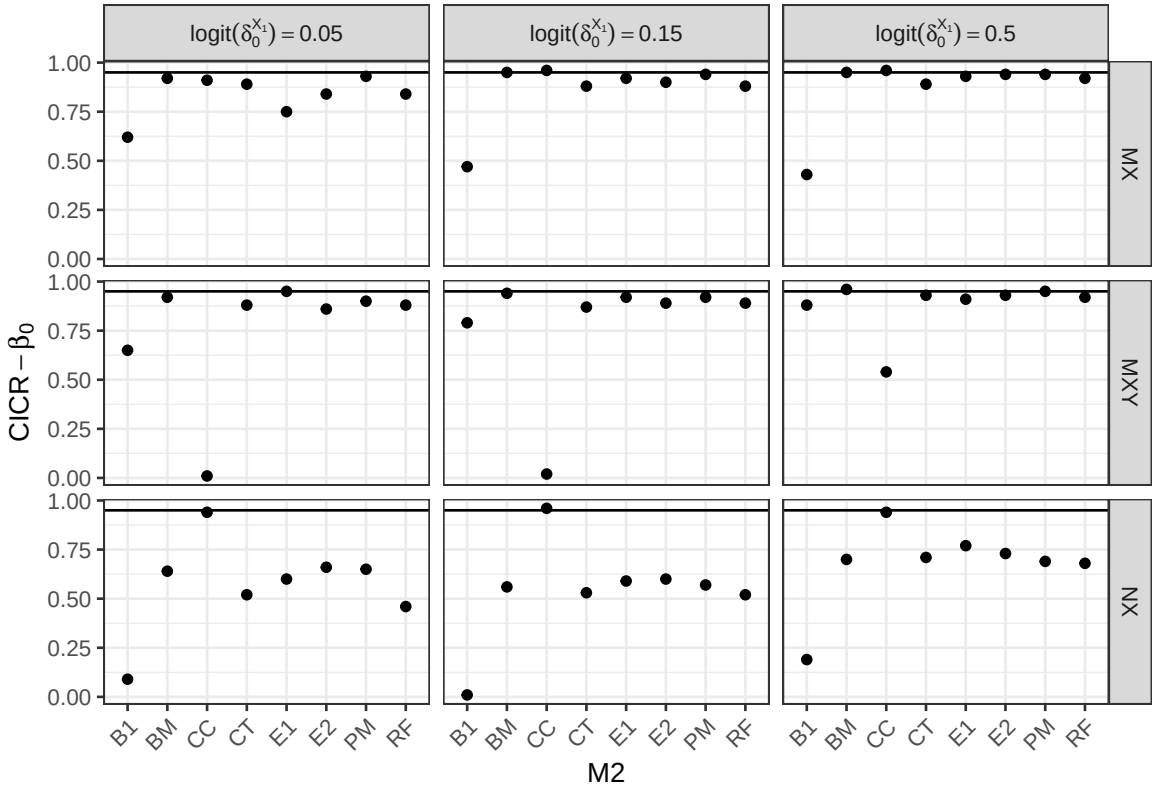

Figure 2.28: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_0 in M3 for nine different data scenarios with linear relationships. The solid black line indicates 95% coverage rate.

Regarding CICR, for {L,MX}, $logit(\delta_0^{X_1}) = 0.05$, and the coverage rates of β_0 estimates (Figure 2.28), BA shows the lowest coverage rate (62%), followed by B2 and E1 (both at 75%), E2 and RF (both at 80%). All remaining M2s approach the 95% mark. This pattern emerges with minor numerical differences for estimates of β_1 and β_2 as well

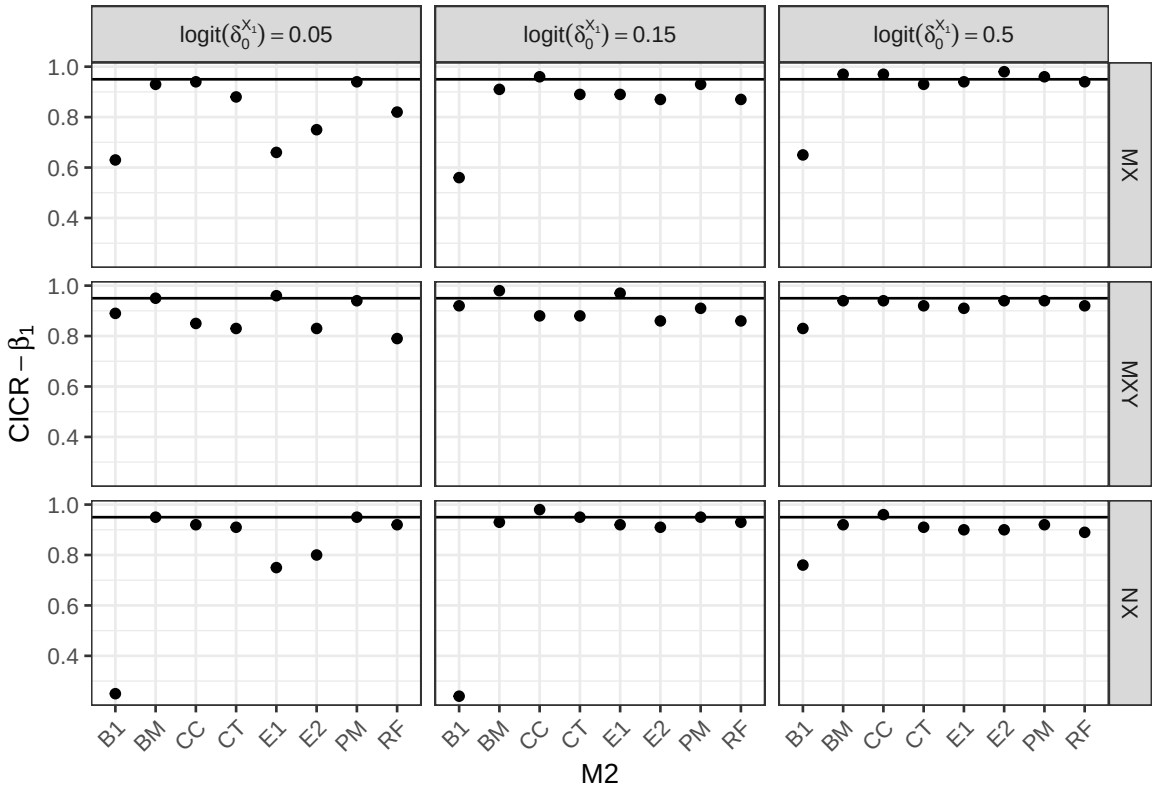

Figure 2.29: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_1 in M3 for nine different data scenarios with linear relationships. The solid black line indicates 95% coverage rate.

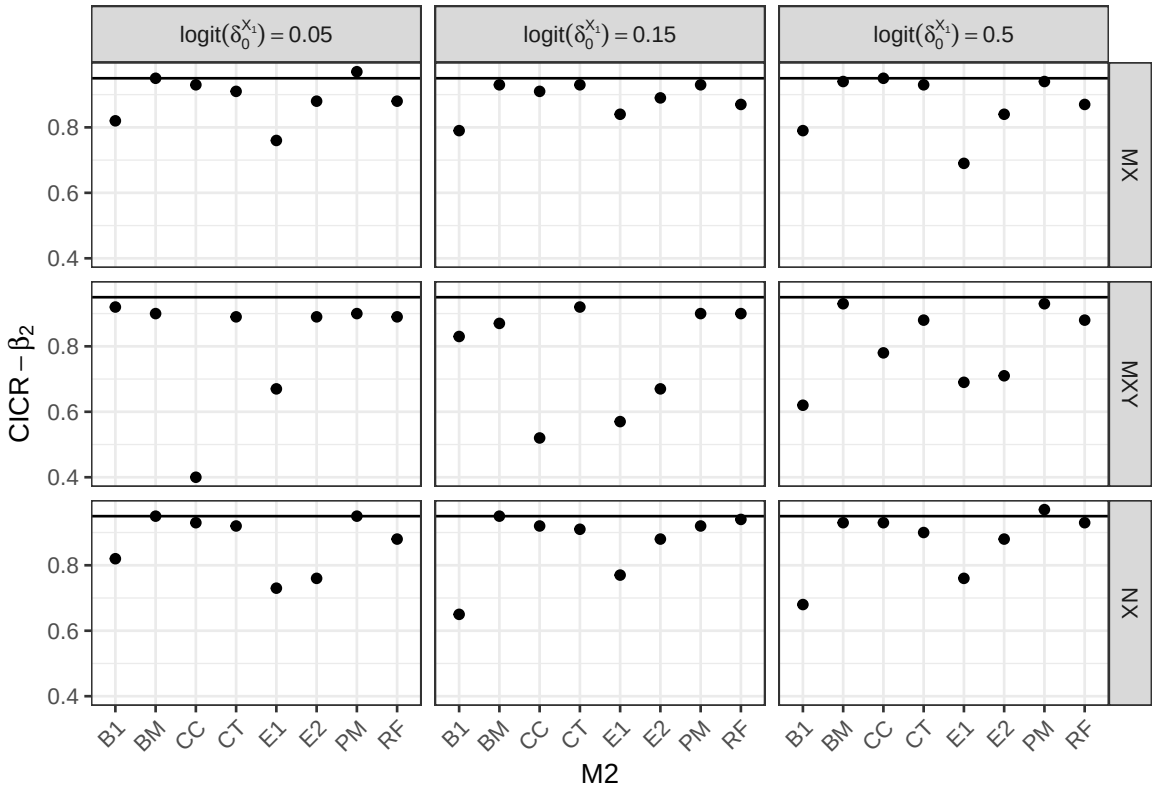

Figure 2.30: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_2 in M3 for nine different data scenarios with linear relationships. The solid black line indicates 95% coverage rate.

(see Figures 2.29, and 2.30). For increasing $logit(\delta_0^{X_1})$, we find all M2 CICR increase as well for β_0 and β_1, except for BA, which remains at approximately the same level. For β_2, coverage rates do not increase for increasing $logit(\delta_0^{X_1})$ values.

For {L,MXY}, β_0 and $logit(\delta_0^{X_1}) = 0.05$, CC yields the lowest coverage rate (0%), next is BA (65%). For increasing $logit(\delta_0^{X_1})$, the coverage rates of CC and BA increase; for BA, CICRs approach 95%, and CC increases to 55%. For β_1, BM, PM, and E1 result in a CICR close to 95% for $logit(\delta_0^{X_1}) = 0.05$. The remaining M2s stay below 95%, but all methods move closer to 95% with increasing $logit(\delta_0^{X_1})$. For β_2, we find that CC and E1 lead to low coverage (40% and 65%) for $logit(\delta_0^{X_1}) = 0.05$. In the case of increasing $logit(\delta_0^{X_1})$, CC's coverage increases, while other BA, E1, and E2 leads to lower coverage.

For {L,NX}, β_0 and $logit(\delta_0^{X_1}) = 0.05$, we find that CC yields best coverage rate at 95% followed by E1, E2, BM, and PM (all around 64%); BA shows lowest rates close to 0%. For increasing $logit(\delta_0^{X_1})$, the coverage rates of all methods increase; the coverage rate of BA approach 25%, while the remaining methods increase to approximately 75%. For β_1, BA results in the lowest coverage (12%, 25%, and 25%) for $logit(\delta_0^{X_1}) = 0.05$. The remaining M2s approach 95% and all methods move closer to 95% with increasing $logit(\delta_0^{X_1})$, except for BA, which only reach 75%. For β_2 and $logit(\delta_0^{X_1}) = 0.05$, the methods resulting in lowest CICIR are E1, E2, and BA (at or above 75%). In the case of increasing $logit(\delta_0^{X_1})$, all coverage rates increase, except for BA which drops to about 75%.

CICR - {Q,.}

Regarding the CICRs for {Q,MX}, and across $logit(\delta_0^{X_1})$ values, we see similar patterns for estimates of β_0, β_1, and β_3 (Figures 2.31, 2.32, and 2.34). LM, E1, and E2, perform noticeably worse than the other methods. BA also show a poor performance for β_0 and β_1. Different $logit(\delta_0^{X_1})$ values do not substantially change these patterns. Increasing $logit(\delta_0^{X_1})$ leads to lower CICR values in BA.

For {Q,MXY}, β_0, β_1, β_3, and $logit(\delta_0^{X_1}) = 0.05$, CC, BM, E1, E2 clearly perform worse than other methods. Their performance increased with increasing $logit(\delta_0^{X_1})$, but do not approach 95%. Regarding β_2, for lower values of $logit(\delta_0^{X_1})$, we find that CC results in lower coverage.

For {Q,NX}, CC performs best at 95% in all four parameters. BM, E1, and E2 show lowest CICR in β_0 and β_1; BA has the lowest CICR in β_2. In β_3, CT, BM, PM, RF,

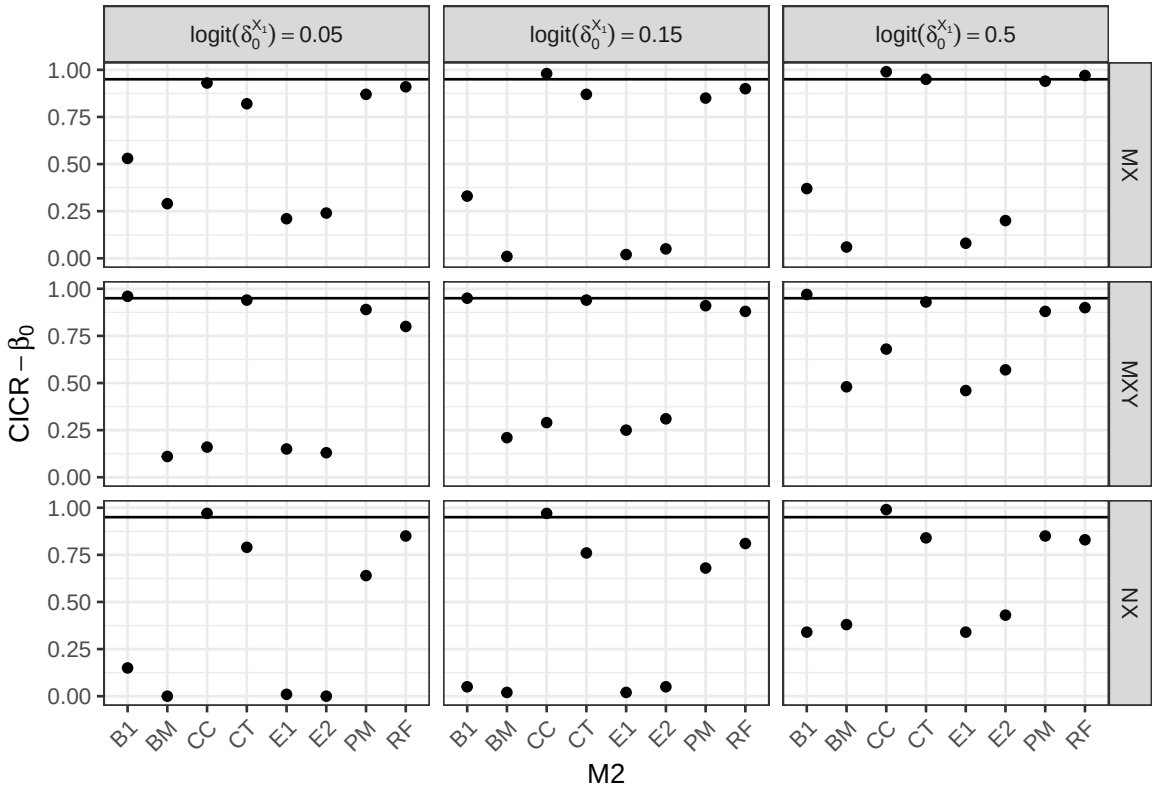

Figure 2.31: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates 95% coverage rate.

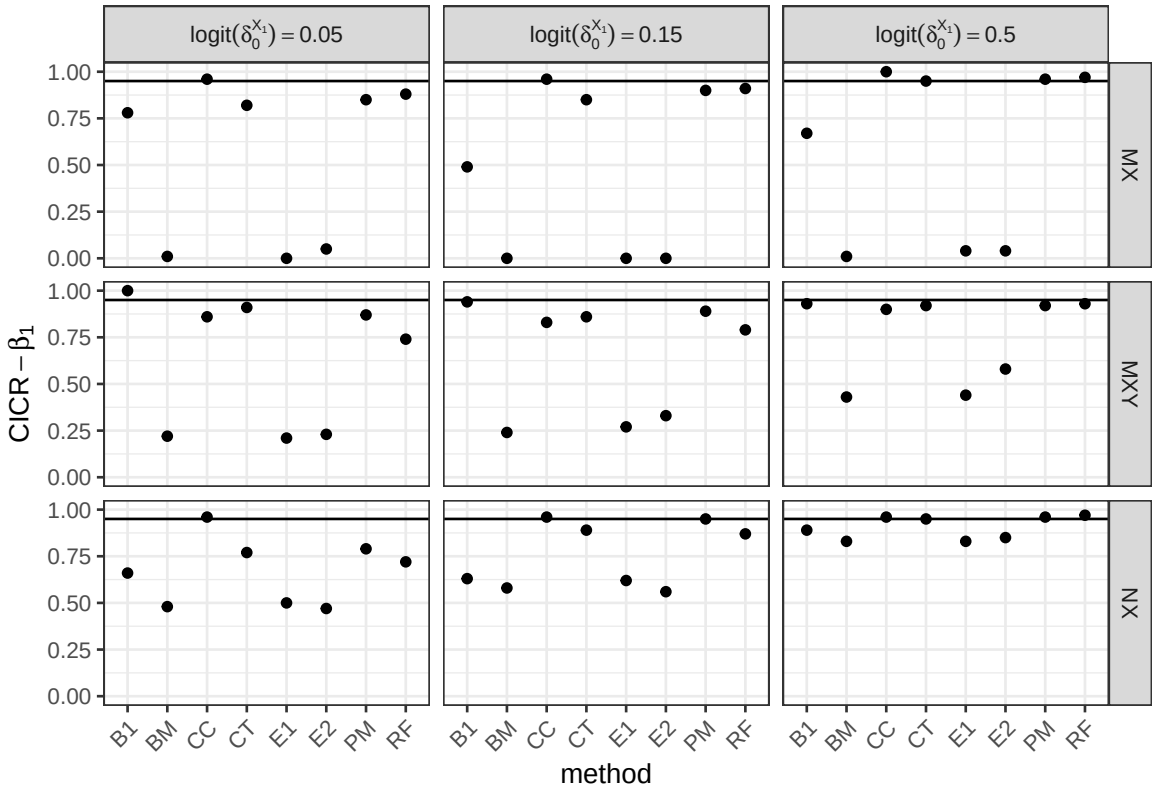

Figure 2.32: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_1 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates 95% coverage rate.

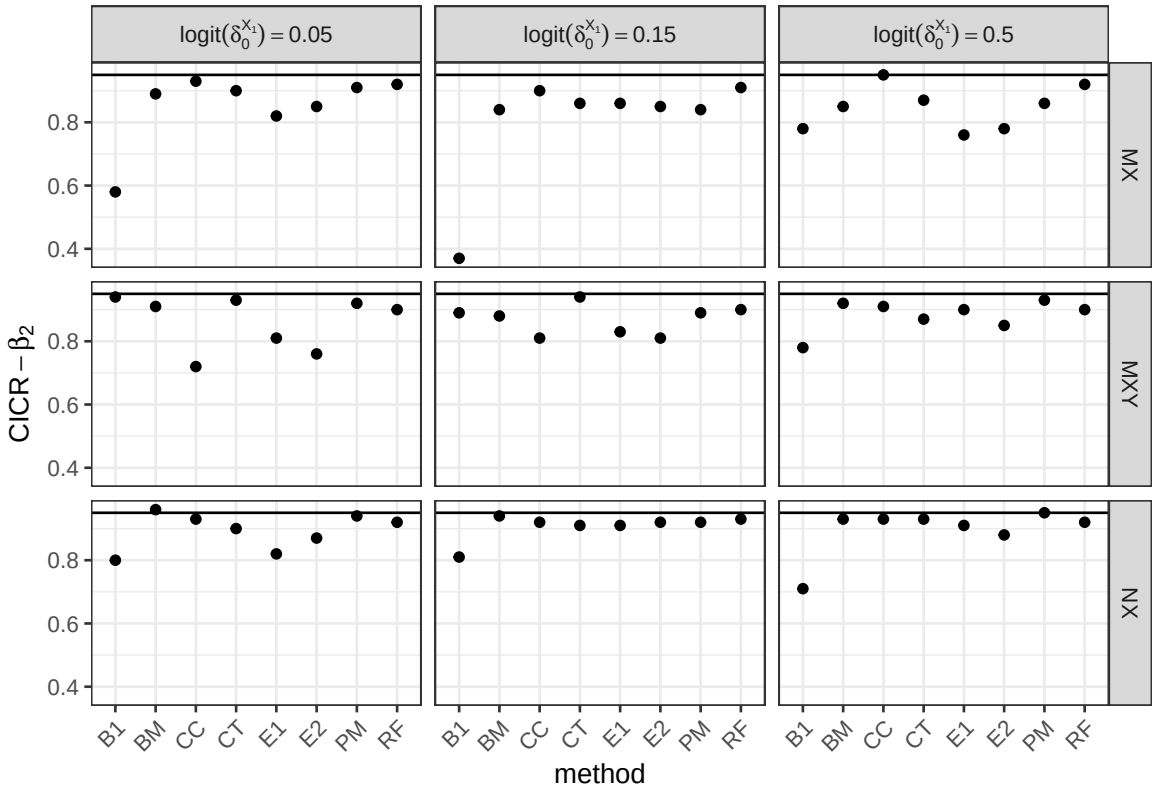

Figure 2.33: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_2 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates 95% coverage rate.

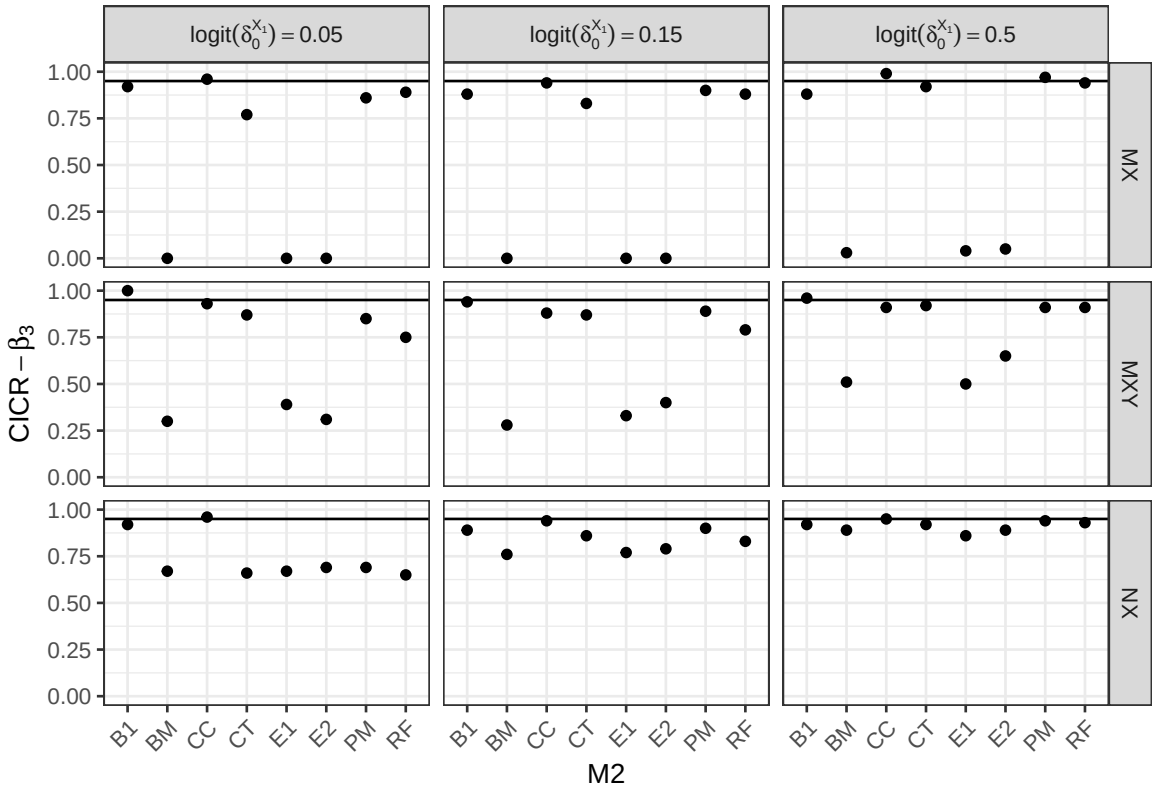

Figure 2.34: Different M2s compared in terms of the resulting confidence interval coverage rates (CICRs) in the estimated regression coefficient β_3 in M3 for nine different data scenarios including a quadratic relationship. The solid black line indicates 95% coverage rate.

E1, and E2 all show low CICR (between 63% and 72%).

Non-Parametric Data

Broadly speaking, regarding values of $logit(\delta_0^{X_1})$, we mostly see the same pattern for all three investigated values. This is true for EB and RMSE. For CICR, there are some exceptions, described below.

EB - {N,.}

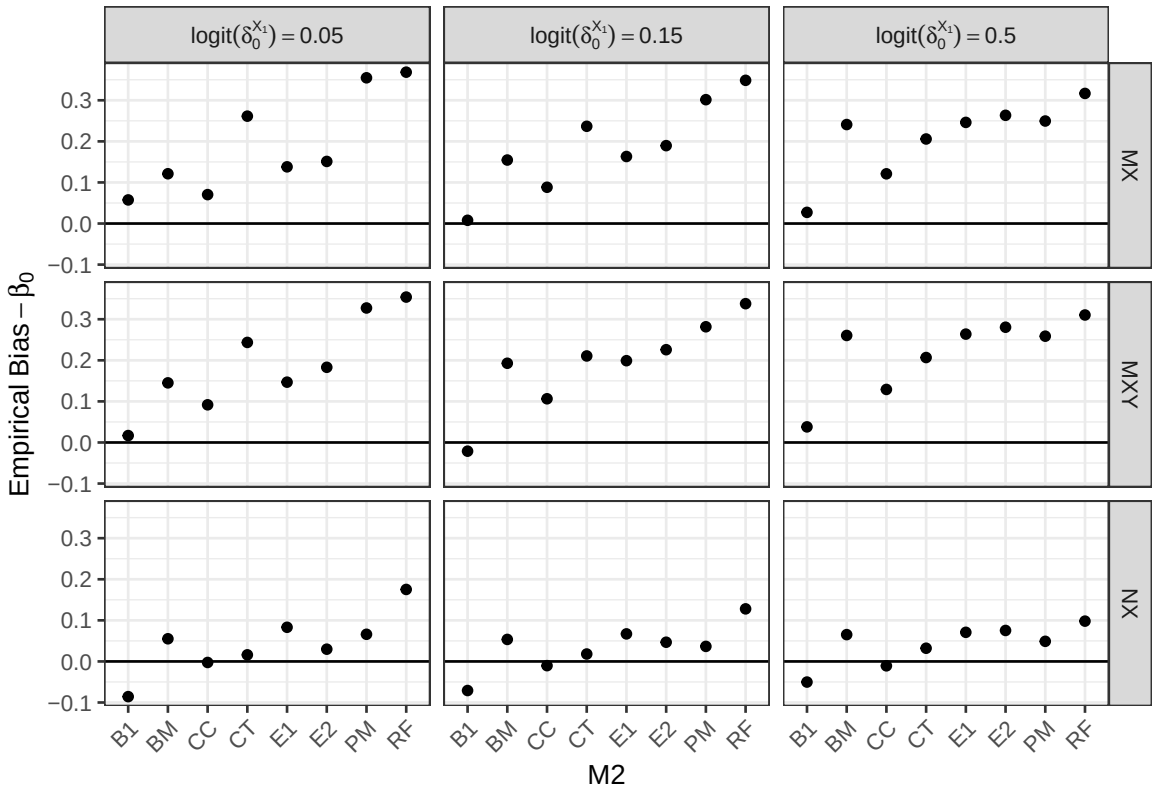

Figure 2.35: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates zero empirical bias.

We first investigate EB in β_0 and β_2 estimates in {N,.}, the non-parametric data. Results for β_0 and β_1 are shown in Figures 2.35 and 2.36.

For {N,MX}, we find opposite signs for EBs in β_0 and β_2. We find the largest absolute EB in PM and BM. CC, and BA perform best.

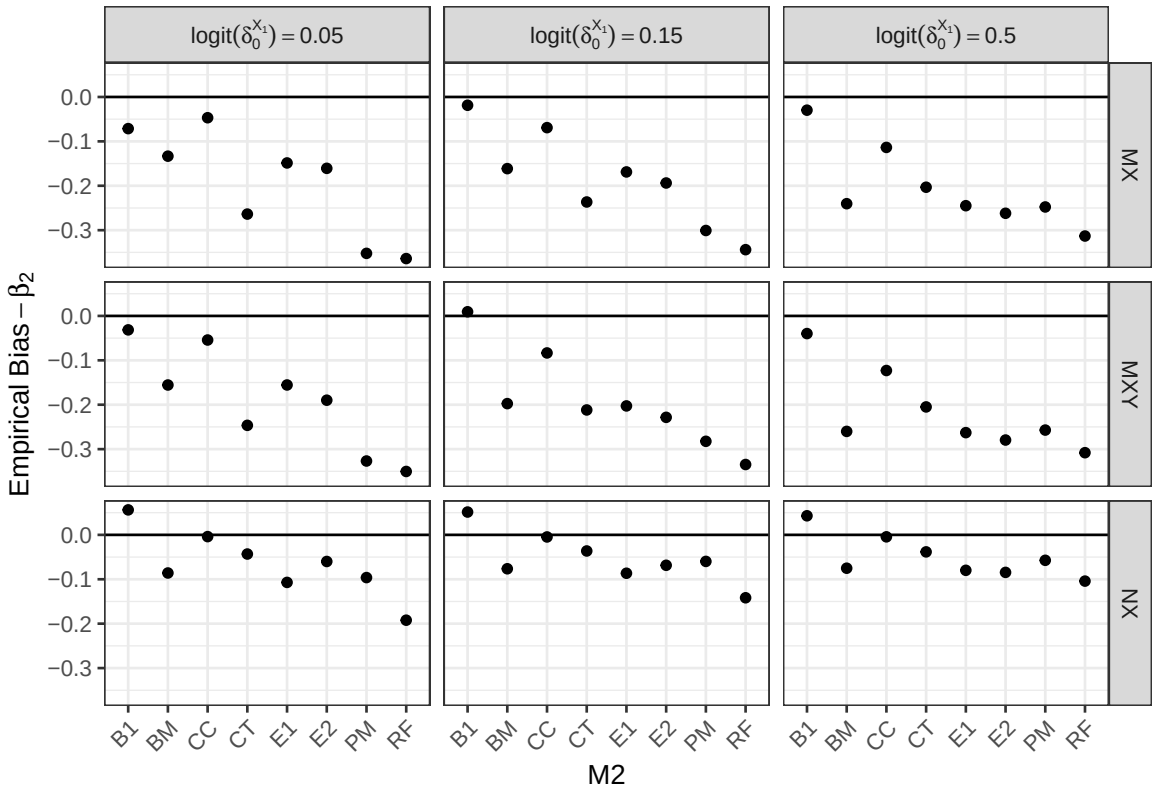

Figure 2.36: Different M2s compared in terms of the resulting empirical bias in the estimated regression coefficient β_1 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates zero RMSE.

For {N,MXY}, EBs in β_0 and β_2 have opposite signs, but are otherwise very similar. We find that BA performs best, followed by CC, and then the remaining methods that show minor differences for different $logit(\delta_0^{X_1})$ values.

For {N,NX}, CC leads to empirically unbiased estimates in both β_0 and β_2. We find the largest EBs in RF, followed by E1, PM, and BM.

RMSE - {N,.}

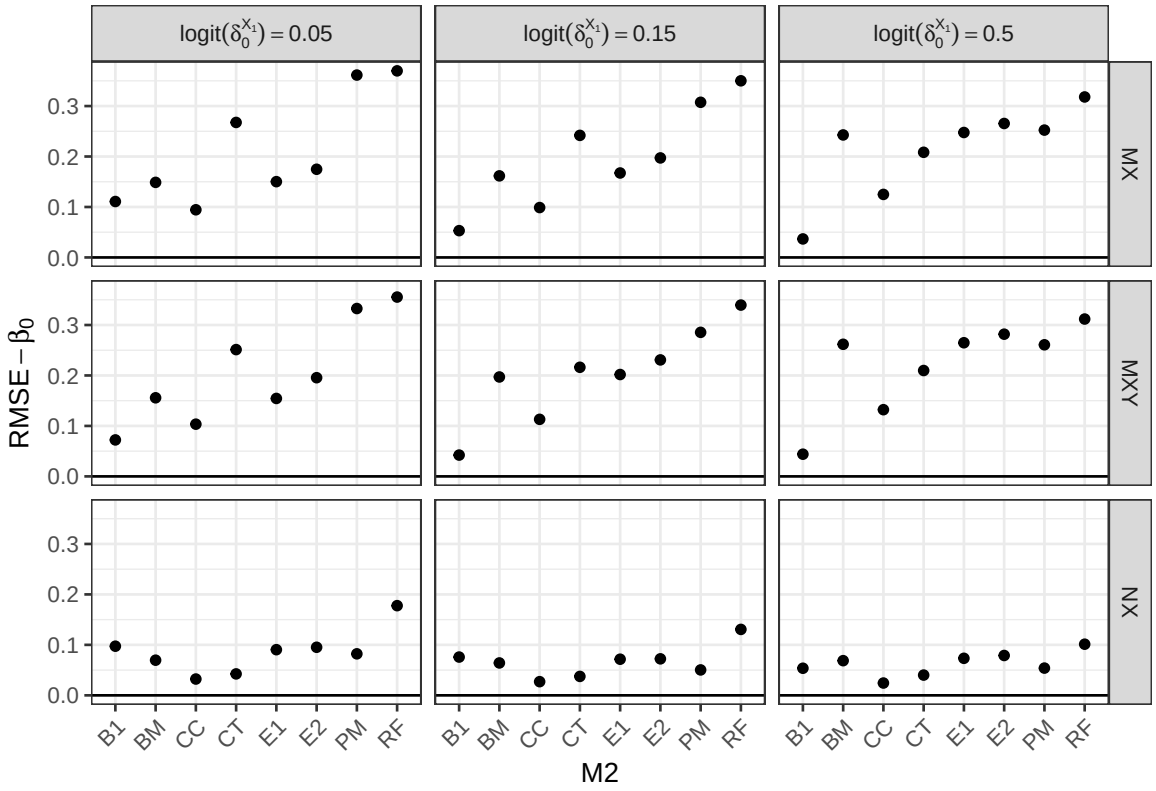

Figure 2.37: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates zero RMSE.

Regarding RMSE, we find a nearly identical pattern for β_0 and β_2 (Figures 2.37 and 2.38) for {N,MX} and {N,MXY} cases. Here, CC and BA result in the smallest RMSE values in both parameter estimates. While the RMSE of BA decline with increasing $logit(\delta_0^{X_1})$ values, the RMSE values of all other M2s increase or remain stable.

For {N,NX}, CC show the lowest RMSE values, while RF yield the highest values. The RMSE values of all methods decrease with increasing $logit(\delta_0^{X_1})$.

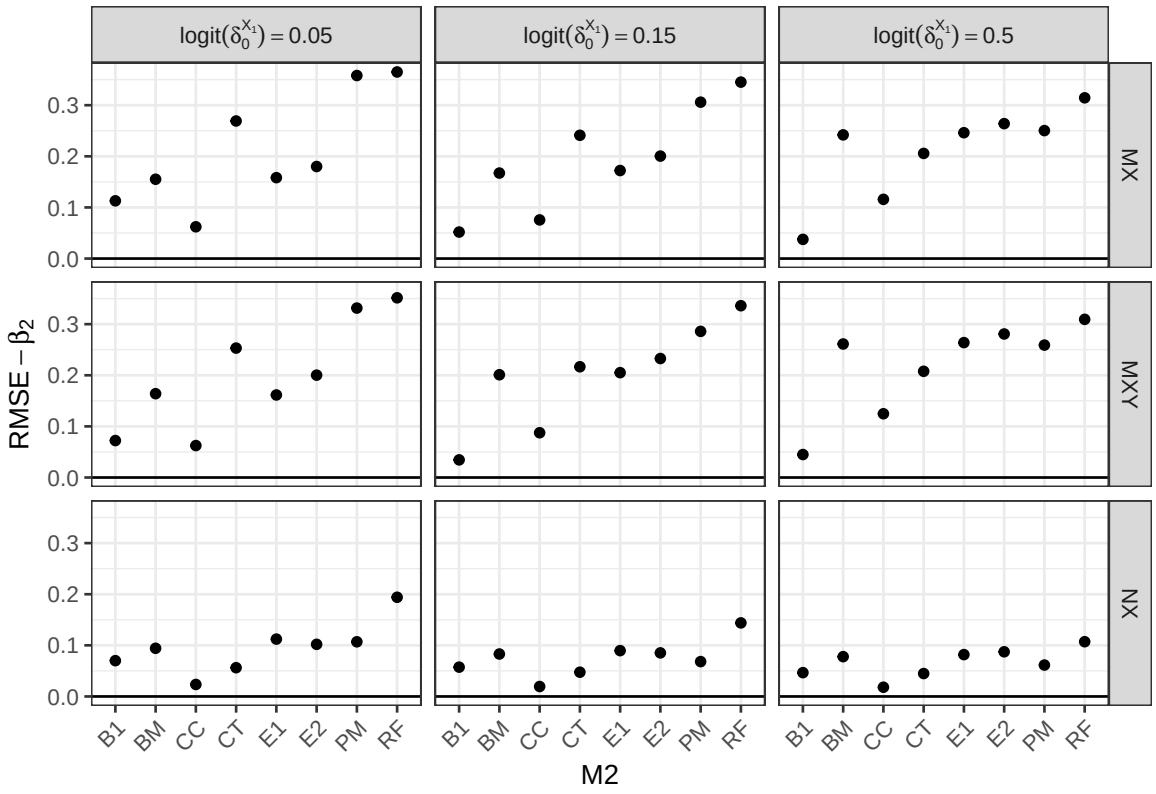

Figure 2.38: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_2 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates zero RMSE.

CICR - {N,.}

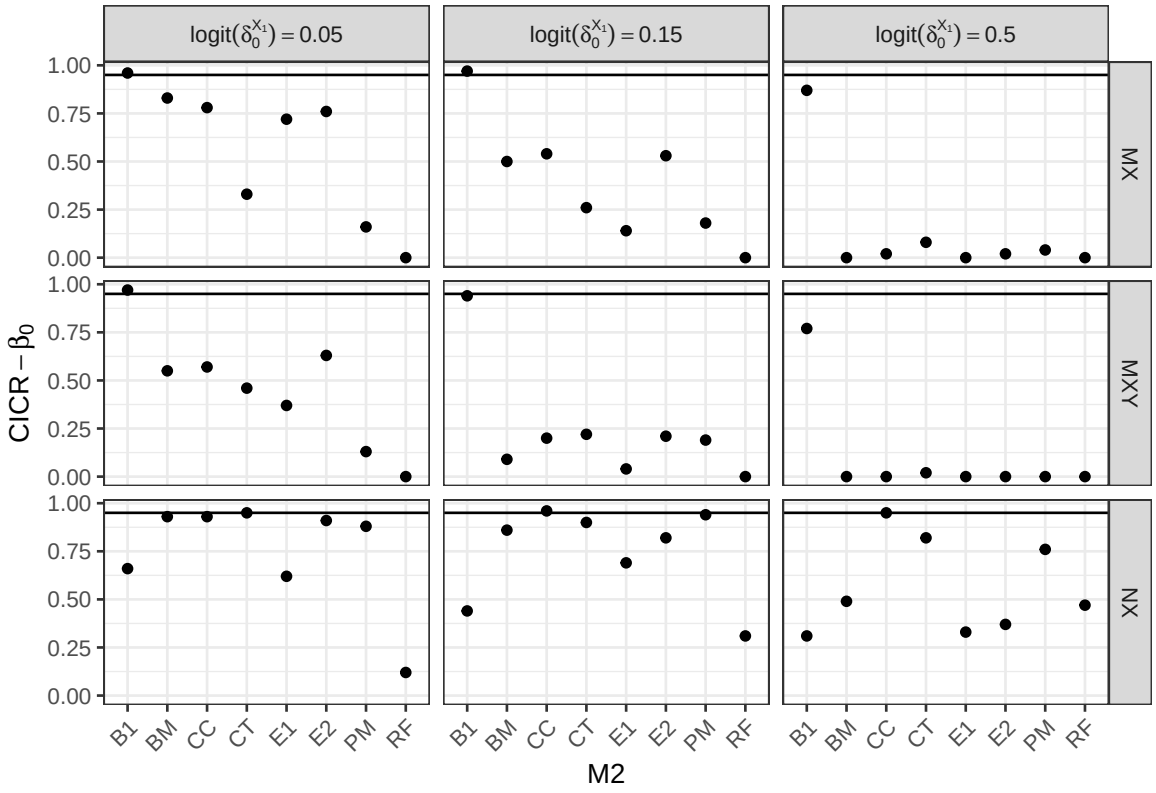

Figure 2.39: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_0 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates 95% coverage rate.

For {N,MX}, we see similar patterns in β_0 and β_2 in terms of CICR. BA performs best at approximately 95% CICR. While RF, E1, and E2 result in CICR above 50% for $logit(\delta_0^{X_1}) = 0.05$, all methods drop to (or close to) 0% coverage rate for $logit(\delta_0^{X_1}) = 0.5$.

In the {N,MXY} case, for both parameters (Figures 2.39 and 2.40), BA performs best in terms of CICR, while all other methods drop to zero for increasing $logit(\delta_0^{X_1})$ values.

For {N,NX}, we again see similar patterns for β_0 and β_2. CC always performs best, followed by CT. RF performs worst overall.

Spline evaluation - {N,.}

We now focus on the mean absolute distance between the true shape and the estimated spline (S-EB), presented in Figure 2.41. In {N,MX} scenarios, the strongest deviations

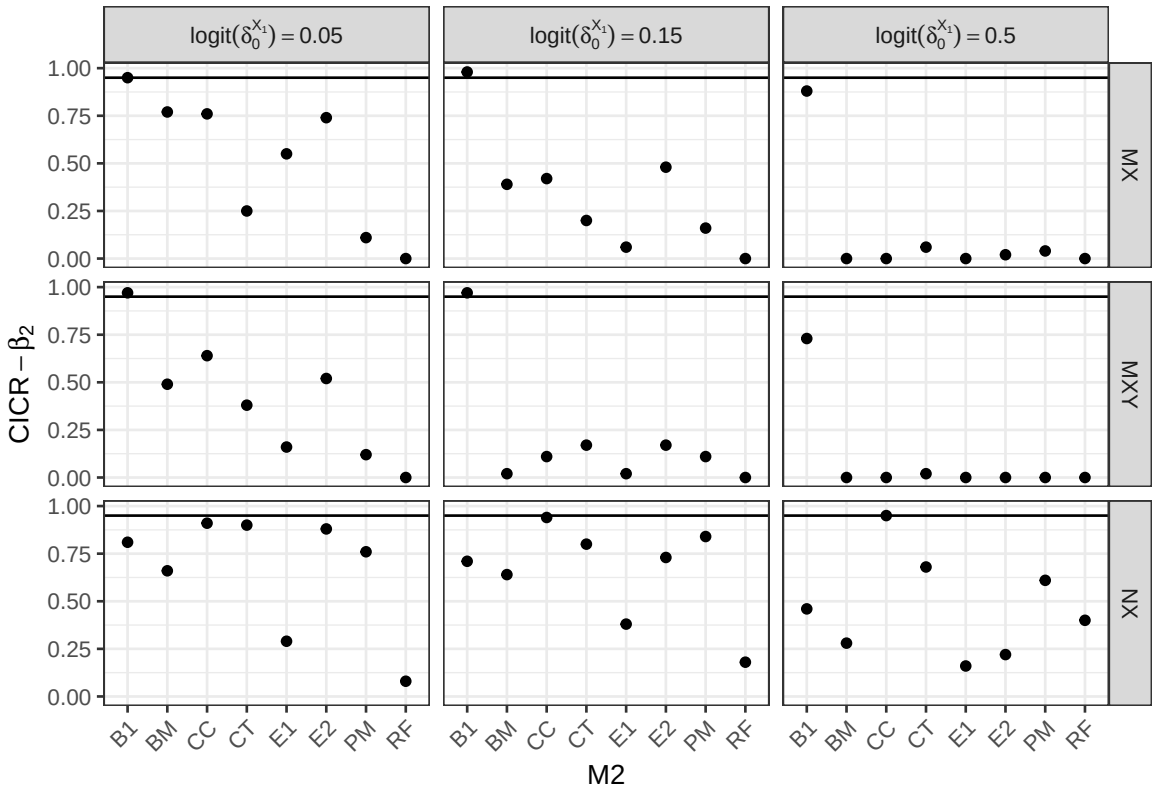

Figure 2.40: Different M2s compared in terms of the resulting RMSE in the estimated regression coefficient β_2 in M3 for nine different data scenarios including a non-parametric relationship. The solid black line indicates 95% coverage rate.

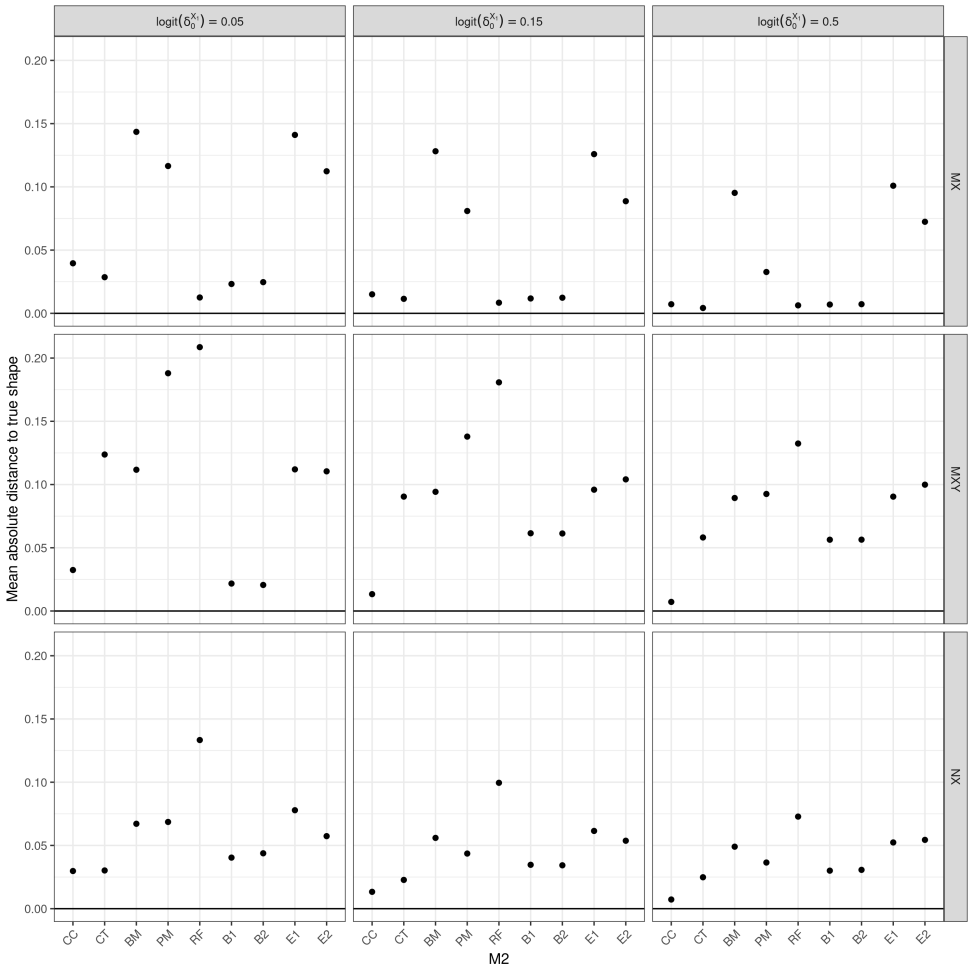

Figure 2.41: Different M2s compared in terms of mean divergence of marginal predicted means from true spline shape (S-EB) in the M3. Results presented for nine different data scenarios including a non-parametric relationship.

occur for BM, PM, E1, and E2, the best performing methods are CT, RF, and BA. The deviations are generally reduced for increasing $logit(\delta_0^{X_1})$. Focusing on {N,MXY}, we find BA performing best, PM and RF performing worst. In {N,NX} cases, CC and CT perform best, RF shows highest S-EB. For increasing $logit(\delta_0^{X_1})$, differences between the M2s diminish.

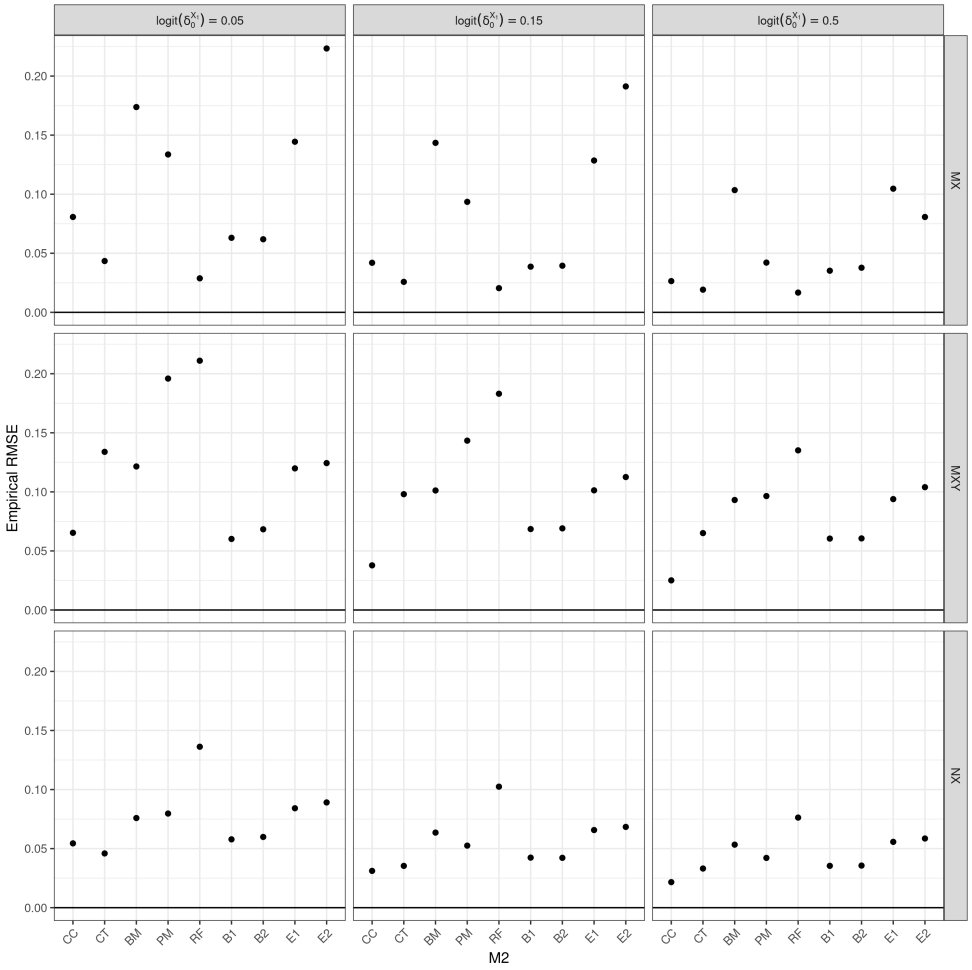

Figure 2.42: Different M2s compared in terms of mean RMSE based on marginal predictions (S-RMSE) in the M3. Results presented for nine different data scenarios including a non-parametric relationship.

We now look at S-RMSE of the estimated spline, visualized in Figure 2.42. In {N,MX} scenarios, the highest S-RMSE value occurs for E2, and the best performing methods are CT and RF. S-RMSE values are reduced for increasing $logit(\delta_0^{X_1})$ values.

Focusing on {N,MXY}, we find that CC and BA perform best in terms of S-RMSE, while PM and RF performing worst. Again higher $logit(\delta_0^{X_1})$ values lead to smaller

S-RMSE values.

For {N,NX} cases, RF and E2 show the highest S-RMSE values; CC and CT show the best performance. For increasing $logit(\delta_0^{X_1})$, differences between methods diminish.

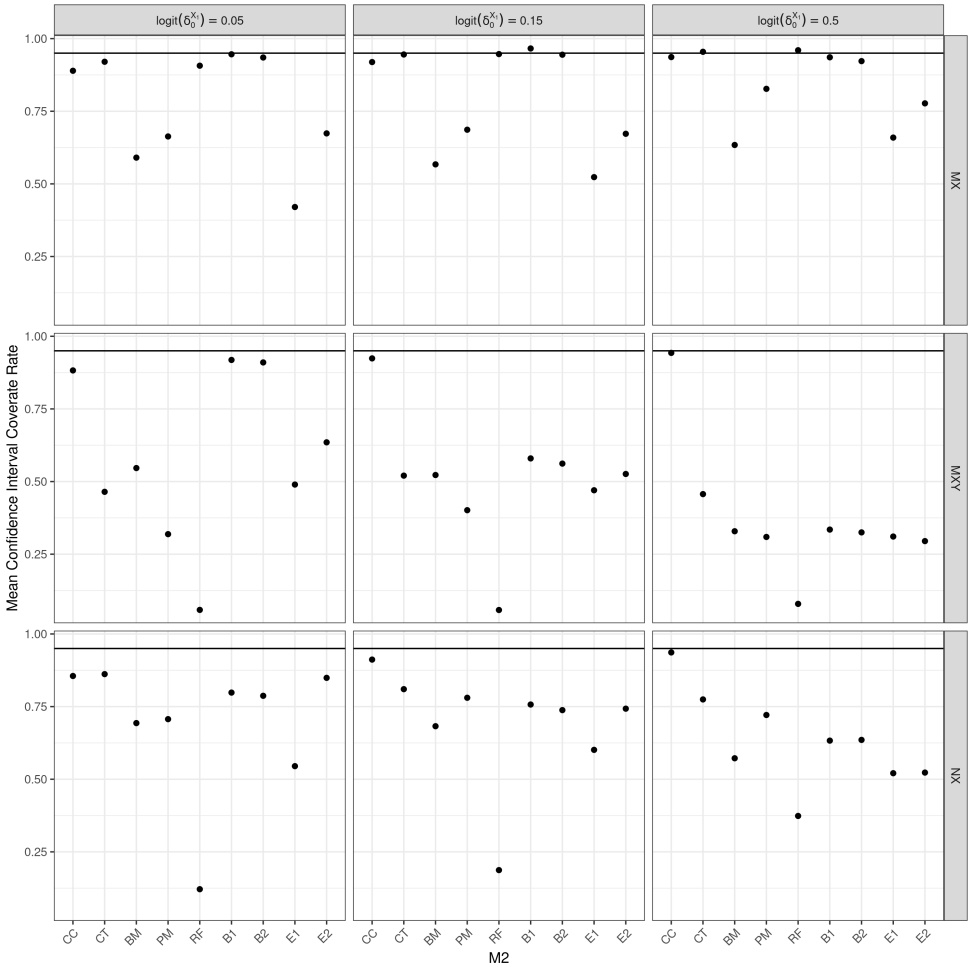

Figure 2.43: Different M2s compared in terms of mean confidence interval coverage rates (S-CICR) based on marginal predictions in the M3. Results presented for nine different data scenarios including a non-parametric relationship.

We now look at mean confidence interval coverage rates of marginal predictions (S-CICR) of the estimated spline, visualized in Figure 2.43. In {N,MX} scenarios, BA, CT, and RF result in the best S-CICR (approximately 95%); E1 achieves the lowest S-CICR ($< 50\%$), followed by BM, RF, and E2 (between 60% and 70%). Coverage generally increases for increasing $logit(\delta_0^{X_1})$ values.

Focusing on {N,MXY}, we find that CC, and BA perform best (close to a 95% coverage

rate), while RF performs worst. Here, higher $logit(\delta_0^{X_1})$ values lead to lower S-CICR values overall, except for CC.

For {N,NX} cases, CT, CC, and E2 show the highest coverage rates (at approximately 85%); RF shows the worst performance (12%). For increasing $logit(\delta_0^{X_1})$, differences between methods diminish and move towards 60%; only CC remains high in S-CICR.

Chapter 3

Sequential Imputation with Integrated Model Selection

Abstract

The issue of incomplete observations, which are attributable to item nonresponse, unit nonresponse, failure to link records, or panel attrition, is an inevitable problem in survey data sets. Sequential imputation (SI) is often used to impute those missing values. However, in data sets where many variables are affected by missing values, appropriate specifications of sequential regression models can be burdensome and time consuming, as a separate model needs to be developed by a human imputer for each incomplete variable. This task is even more complex because survey data typically consists of many different kinds of variables (i.e., continuous, binary, and nominal) with possibly non-trivial and non-linear relationships. Available software packages for imputation procedures (e.g., Multivariate Imputation by Chained Equations (MICE) or IVEware) require model specifications for each variable containing missing values. Additionally, the default models in this software can lead to bias in imputed values, for example, when variables are non-normally distributed or when important interactions are not included in the imputation models.

In this chapter we propose an enhanced SI procedure with automated model selection. The procedure takes into account data sets consisting of potentially non-normally distributed variables and potentially complex and non-linear interactions. In each step of the procedure, model selection is carried out from a pool of several parametric and non-parametric models based on two criteria. The first criterion compares models based on

their predictive power. The second criterion focuses on the similarity between imputed and observed values conditional on the response propensity score for the outcome variable. A case study based on a survey data set (the 2015-2016 National Health and Nutrition Examination Survey) illustrates the proposed procedure using state-of-the-art imputation methods. The evaluation of the proposed method focuses on differences in the quantitative properties of a hypothetical analysis regression model of interest, fit to the imputed data. We further assess runtime of the imputation process. Overall, SIIMS' performance mostly lies in between the performances of the single models.

3.1 Introduction

Missing values in data sets introduce several problems in data analysis, because most analysis tools assume complete cases. Thus, when applied to complete cases only, the reduced sample size leads to a loss of information, often resulting in loss of power. This problem is exacerbated for each additional variable included in an analysis, because these variables further reduce the number of complete cases. To address this issue, multiple imputation (MI) of the missing data can be applied under the missing at random (MAR) assumption (Little & Rubin, 2002, pp. 11–19) in order to compute unbiased estimates and also to use all available information in a subsequent analysis (i.e., not only complete cases).

First proposed by Rubin (1978, 1996), the MI procedure produces multiple complete data sets that can later be analyzed with standard data analysis tools. Combining the analysis results from each MI data set following Rubin's rule leads to valid point and variance estimates under MAR and correct model specification. This means that the method can account for both the uncertainty about the imputed missing data (Rubin, 1987, Chapter 3) and the uncertainty of the model (i.e., proper imputation (Rubin, 2004, pp. 116–131)). Once a data set is multiply imputed, different analysts can use the same data sets for their analyses.

To carry out MI, sequential imputation (SI), where missing values are imputed variable by variable, is often used in practice (see e.g. Paulin et al. (2006), Schenker et al. (2006), Stuart et al. (2009))[1]. First proposed and applied by Gleason and Staelin

[1] An alternative approach is to jointly model the data using a multivariate distribution, which is not part of this chapter. Further information about MI using joint modeling can be found in Li (1988),

(1975), several evaluations of the SI approach in simulation studies have found that it is a robust, easily applicable, and flexible way to implement MI in practice (Raghunathan et al., 2001; Van Buuren et al., 2006; Van Buuren, 2007). Further variations of SI have since been developed by Finkbeiner (1979), Raymond & Roberts (1987), Jinn & Sedransk (1989), and Gold & Bentler (2000). Classic linear regression models are often applied (Raghunathan et al., 2001) to specify the necessary conditional distributions within SI using all partially observed variables as dependent variables, and all other available variables as covariates.

MI with SI can be realized by applying Bayesian regression models, first drawing model parameters from their posterior distributions, and then drawing plausible values for the missing values (Rubin, 1987, pp. 166–167). These imputed values are then used in conjunction with the observed values to fit the imputation model to the next variable. This process continues iteratively over all incomplete variables, substituting the imputed values with the new draws in each iteration (Kennickell, 1991).

Despite its flexibility, the SI procedure has some drawbacks. First, SI can be computationally intensive, given the iterative nature of the procedure (Raghunathan et al., 2001; Van Buuren et al., 2006). Every regression model needs to be specified in advance, leading to a potentially high modeling effort and the risk of misspecification (Van Buuren et al., 2006). Each of these difficulties can be especially problematic when the number of incomplete variables is high. Predictor selection for each incomplete variable can be performed, as implemented in the software package MICE (Multivariate Imputation by Chained Equations) (Van Buuren & Groothuis-Oudshoorn, 2011).

To limit misspecification, researchers have studied supervised machine learning techniques as substitutes for parametric regression models in SI. Burgette and Reiter (2010), for example, propose MI of continuous variables using sequential classification and regression trees (CART) (Breiman et al., 2017). The authors conclude that CART is able to capture complex interactions without high modelling effort and can outperform models that omit important interactions. CART also has advantages when interactions of covariates influence categorical outcome variables (Akande et al., 2017; Doove et al., 2014). Substituting a random forest (RF) algorithm for CART leads to a more accurate prediction of the missing values and can better account for model uncertainty (Stekhoven & Bühlmann, 2012). This procedure, however, has been criticized by Shah

Rubin and Schafer (1990), and Schafer (1997).

et al. (2014) because the variance of the imputed values can potentially underestimate the variance of the actual values. Shah et al. (2014) propose that missing values should be "imputed by random draws from independent normal distributions centered on conditional means predicted using [RF]" and compare this procedure to parametric imputation models implemented in the R software (R Core Team, 2019) package MICE (Van Buuren & Groothuis-Oudshoorn, 2011). According to Shah et al. (2014), the non-parametric RF has an advantage over the default parametric models in MICE. Despite this advantage, using RF for imputation comes with additional computational expense since multiple CART procedures are applied for each incomplete variable in each iteration.

Bayesian additive regression trees (BART) (Chipman et al., 2010) were proposed in Xu et al. (2016) for sequentially imputing missing values. The authors compare sequential BART, parametric SI within MICE, and CART within MICE via simulation studies. They find that, for simple linear cases, all of the approaches perform equally well, but for more complex data situations, BART performs better than other approaches. A recent publication combines BART with a penalized splines of propensity prediction (PSPP) to create a doubly-robust (G. Zhang & Little, 2009) imputation procedure (Tan et al., 2019). Tan et al. (2019) find that PSPP carried out with BART is more robust than PSPP alone.

Missing data imputation can also be combined with regularization procedures, which can prevent overfitting in parametric models. For instance, Zhao et al. (2016) compares several regularized approaches - LASSO (Tibshirani, 1996), adaptive LASSO (Zou, 2006), elastic net (Zou & Hastie, 2005), and Bayesian LASSO (Park & Casella, 2008) - for MI of one incomplete variable in high-dimensional data sets. Their comparison via simulation suggests that Bayesian LASSO performs best in terms of empirical bias, standard error, and the coverage rate of a regression coefficient estimated from the multiply imputed data sets (i.e., in terms of quantitative properties). This approach was expanded to multiple incomplete variables in a paper by Deng et al. (2016). In their work, LASSO, adaptive LASSO, and elastic net regularization are incorporated in an SI procedure and tested on similarly simulated data sets. The authors find elastic net regularization to be superior to the other methods in most tested scenarios. Although Zhao and Long (2016) and Deng et al. (2016) conclude that regularization is a promising approach in high-dimensional data situations, the studies also have shortcomings. For one, only main effects were simulated and the missing data mechanism

is also based on main effects of covariates. For another, the simulated data follow (multivariate) normal distributions only; neither other continuous distributions nor categorical variables were investigated.

Two additional studies focus on missing data imputation in high-dimensional data. Razzak and Heumann (2019) propose a hybrid approach combining joint modeling (JM) and SI to impute high-dimensional survey data in response to the failure of MICE in their example data set. They find that the proposed hybrid approach involving JM and SI (using CART and PMM in MICE) performs better than CART in MICE alone, in terms of both quantitative properties of regression coefficients and runtime. However, this hybrid approach also has shortcomings. In a first step, JM is applied to incomplete categorical variables (ignoring incomplete continuous variables). Next, SI (using MICE) imputes the incomplete continuous variables, conditional on the JM-imputed categorical variables. This procedure potentially leads to suboptimal performance. First, JM is based only on the subset of categorical variables leading to a less plausible MAR assumption. Second, the SI portion treats imputed and observed values in the JM-imputed variables as the same.[2]

The second study (Liang et al., 2018) proposes a general algorithm for missing data imputation in high-dimensional data. The two-step procedure procedure alternates between imputation of missing values and regularized optimization. The imputation conditions on the latest parameter estimates and the observed data. The optimization estimates the parameters based on the current data. The evaluation is based on simulated data from an auto-regressive process, and precision-recall curves from different imputation approaches are compared. Although the proposed method works best in the given scenario, the missing values often follow an unrealistic missing completely at random mechanism (Little & Rubin, 2002, pp. 11–19) (i.e., they were randomly deleted). Further, the assessment is based only on precision-recall curves; quantitative properties of regression coefficients (as recommended by Rubin (2004)) are not investigated.

The studies discussed above compare different approaches after completion of MI. In contrast to these studies, in this chapter we propose a framework that assesses and compares models within SI, building on Bondarenko and Raghunathan (2016) and

[2]A potential solution would be an iterative procedure alternating between the JM and the SI step updating missing values. For the JM step, continuous variables could be categorized for use as covariates.

Su et al. (2011). Bondarenko and Raghunathan (2016) propose several (graphical) diagnostic tools for imputation. One of these tools compares residual density plots of observed and proposed values from an imputation model of the incomplete variable. Both densities are conditioned on the estimated response propensity score of observing this particular variable. In this approach, an appropriate imputation model should lead to similar densities in both cases under MAR. Still, because the plots need to be investigated visually, this idea is limited to situations where a modest number of variables are imputed. Another resource for diagnostics in SI is a software package (called mi) by Su et al. (2011). This package implements residual plots and plots comparing distributions of observed and imputed values.

In this chapter, we extend the tools presented in Bondarenko and Raghunathan (2016) and Su et al. (2011) to automatically compare competing imputation models in terms of different criteria under an MAR assumption. We present criteria for continuous, binary, and nominal incomplete variables, as well as a modified SI procedure with integrated model selection. Afterwards, we describe a case study applying some current state-of-the-art procedures for SI in a simulation based on a real data set. The procedures in this example are regularized regression, following Deng et al. (2016), CART, following Burgette and Reiter (2010), RF, following Shah et al. (2014), and BART, following Xu et al. (2016). The proposed procedure can be applied to data sets with a high number of incomplete variables with arbitrary distributions, where model specification and model diagnostics by a human imputer is not feasible. This combination of methods can be especially useful, because Chapter 2 finds no universally optimal imputation method.

The remaining part of this chapter is structured as follows. First, we introduce the different selection criteria for continuous incomplete variables and the new SI framework, as well as the modified criteria for binary and nominal variables. Then we describe the setup of the case study, followed by the results. Finally, we discuss the new approach and provide directions for future research.

3.2 Methods

3.2.1 Assessment Strategy

3.2.1.1 Missing Values in Continuous Variables

We propose two different criteria to assess imputation models, specifically an MSE criterion and one comparing densities of imputed and observed values, henceforth called the similarity criterion. The assumption about the conditional mean (i.e., the structural form) is essential to the model, and so we propose an MSE criterion assessing prediction accuracy on the observed values of the incomplete variable. For the second criterion, we build on Bondarenko and Raghunathan (2016) to assess the plausibility of imputed values. For an incomplete variable, this procedure regresses observed and imputed values on the estimated response propensity score of the variable. The resulting residuals are used to estimate the densities of the observed and imputed values. The greater the similarity between these densities, the better the imputation model under MAR. In short, the similarity criterion assesses imputation models in terms of marginal distributions. Instead of investigating the similarity of densities by hand (as proposed by Bondarenko & Raghunathan (2016)), we automate the evaluation to assess automatically competing methods.

3.2.1.1.1 MSE criterion. Let X be a continuous variable with missing values and let R denote the corresponding vector of response indicators. Also, let $X|R = 1$ be the subset of observed values of length N_R, and $X|R = 0$ be the subset of missing values for X. Also, let model $m \in \{1, \ldots, M\}$ be an imputation model in a pool of models of size M. Each model can be fit to the observed data $X|R = 1$ and $\mathbf{Z}|R = 1$ with $\mathbf{Z}$ as the fully observed covariates. For an observation $X_i|R_i = 1$, $i = 1, \ldots, N_R$, and a given model m, we compute the MSE as a measure of prediction accuracy:

$$S_{i,m} = (\bar{X}_{i,m} - X_i)^2 \tag{3.1}$$

with $\bar{X}_{i,m} = \frac{1}{B} \sum_{b=1}^{B} X_{i,m}^{(b)}$ and $X_{i,m}^{(b)}$ describing the b-th prediction for the value $X_i|R_i = 1$ produced by model m, and B representing the number of predicted values drawn. Averaging over all $S_{i,m}$ leads to the MSE criterion for model m:

$$MSE_m = \frac{1}{N_R} \sum_{i \in \{X|R=1\}} S_{i,m} \tag{3.2}$$

In Bayesian models, the necessary B draws for the i-th observation can be obtained from the posterior predictive distributions. In other procedures, the generation of draws is specific to the procedure. For example, in CART, values can be drawn from the observations in the corresponding terminal node of the tree.

3.2.1.1.2 Similarity criterion. For the second criterion, we follow the framework in Bondarenko and Raghunathan (2016), again for X being one incomplete continuous variable. Specifically, if $X|R = 0$ is MAR, the conditional distributions of X, given the covariates $\mathbf{Z}$, are the same for the observed and missing values, i.e., $f(X|\mathbf{Z}, R = 1) = f(X|\mathbf{Z}, R = 0)$. Using the response propensity score $e = P(R = 1|\mathbf{Z})$ as an aggregate for $\mathbf{Z}$ (Rosenbaum & Rubin, 1983), $f(X|e, R = 1) = f(X|e, R = 0)$ also holds under MAR in X. Since the missing values $X|R = 0$ are unobserved, they will be substituted with a set of values $X_m|R = 0$ estimated by an imputation model m. We propose the Hellinger distance (see e.g. Van der Vaart (1988), p. 211-212) to quantify the similarity of $f(X|e, R = 1)$ and $f(X_m|e, R = 0)$):

$$H_m = H_m(f(X|e, R = 1), f(X_m|e, R = 0)) = \sqrt{1 - \int \sqrt{f(X|e, R = 1) f(X_m|e, R = 0)} \mathrm{d}X} \tag{3.3}$$

The better model m is in terms of providing plausible imputed values under MAR, the more similar are $f(X|e, R = 1)$ and $f(X_m|e, R = 0)$, and the lower is H_m.

We assume the following models for $f(X_m|e, R = 0)$ and $f(X|e, R = 1)$:

$$f(X_m|e, R = 0) = s(e)_0 + \epsilon_{R=0}, \tag{3.4}$$

$$f(X|e, R = 1) = s(e)_1 + \epsilon_{R=1} \tag{3.5}$$

with $s(e)_0$ and $s(e)_1$ defining spline functions of e. We estimate $\hat{e}$ via BART based on $\mathbf{Z}$. After fitting the imputation models, we estimate $\hat{f}(X_m|\hat{e}, R = 0)$ via kernel density

estimation using the estimated residuals $\hat{\epsilon}_{R=0}$:

$$\widehat{f}(X_m|\hat{e}, R=0) = \widehat{f}(\hat{\epsilon}_{R=0}). \tag{3.6}$$

We estimate $\widehat{f}(X|\hat{e}, R=1)$ in the same way. Afterwards, $\hat{H}_m$ is estimated using both $\widehat{f}(X_m|\hat{e}, R=0)$ and $\widehat{f}(X|\hat{e}, R=1)$.

In order to finally decide on a best imputation model, we propose to combine both metrics, the MSE and the similarity criterion. To accomplish this, a transformation is necessary, because $MSE_m \geq 0$ and $H_m \in [0,1]$. Computing MSE_m and H_m for all M models and standardizing all values (subtracting estimated means and dividing these differences by estimated standard deviations) leads to measures on the same scale, denoted by $\widetilde{MSE}_m$, and $\widetilde{H}_m$. We define a final model assessment criterion (MAC) for a model m as follows:

$$MAC_m = w_1 * \widetilde{MSE}_m + w_2 * \widetilde{H}_m \tag{3.7}$$

with criteria weights $w_1 + w_2 = 1$.

For the case study, we consider weights averaging $\widetilde{MSE}_m$ and $\widetilde{H}_m$ in different ways. The two most extreme weight sets ($\{w_1 = 0, w_2 = 1\}$ and $\{w_1 = 1, w_2 = 0\}$) lead to completely relying on one criterion ($\widetilde{H}_m$ or $\widetilde{MSE}_m$). Further, we consider other sets of weights ($\{0.25, 0.75\}$, $\{0.5, 0.5\}$, $\{0.75, 0.25\}$) as possible choices in between those two extremes.

3.2.2 Sequential Imputation with Integrated Model Selection (SIIMS)

We now present the modified SI procedure, including model selection based on the presented criteria, where a different model can be selected for each incomplete variable. Let a data set consist of fully observed variables $\mathbf{Z}$ and K continuous incomplete variables $\mathbf{X} = (X_1, \ldots, X_K)$ with corresponding response indicators $\mathbf{R} = (R_1, \ldots, R_K)$. Also let $\mathbf{X}_{-k}$ denote $\mathbf{X}$ without variable X_k. We propose the following SI procedure under an MAR assumption.

For an iteration $j > 1$ the following steps are performed:

1) Repeat for all $k \in \{1, \ldots, K\}$ variables containing missing values:

   a) Estimate response propensity scores $\hat{e}_k^j = P(R_k = 1|\mathbf{Z}, \mathbf{X}_{-k}^j)$ for all values in X_k.

   b) Estimate the density of residuals for X_k regressed on $\hat{e}_k^j$ (Equation 3.4) for the observed values $\hat{f}(X_k|\hat{e}_k^j, R_k = 1)$ using kernel density estimation (Equation 3.6).

   c) Repeat for all $m \in \{1, \ldots, M\}$ assessed imputation models:

      - Fit the model m to $X_k|R_k = 1$ as the dependent variable and $\mathbf{Z}, \mathbf{X}_{-k}^j|R_k = 1$ as the independent variables.
      - Estimate $\widehat{MSE^j}_{k,m}$.
      - Predict values $X_{k,m}^j|R_k = 0$ for $X_k|R_k = 0$ using the model m.
      - Estimate $\hat{f}(X_{k,m}^j|\hat{e}_k^j, R_k = 0)$ (cf. Equations 3.5 and 3.6).
      - Estimate the Hellinger distance

$$\hat{H^j}_{k,m} = H(\hat{f}(X_k|\hat{e}_k^j, R_k = 1), \hat{f}(X_{k,m}^j|\hat{e}_k^j, R_k = 0))$$

      .

   d) Standardize the criteria: $\widehat{MSE^j}_{k,m} \Rightarrow \widehat{\widetilde{MSE^j}}_{k,m}$; $\hat{H^j}_{k,m} \Rightarrow \widehat{\tilde{H^j}}_{k,m}$:

$$\widehat{\widetilde{MSE^j}}_{k,m} = \frac{\widehat{MSE^j}_{k,m} - \overline{MSE^j}_k}{SD(\widehat{MSE^j}_k)},$$

   with $\overline{MSE^j}_k = \frac{1}{M}\sum_{m=1}^{M} \widehat{MSE^j}_{k,m}$ and

$$SD(\widehat{MSE^j}_k) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\widehat{MSE^j}_{k,m} - \overline{MSE^j}_k)^2}.$$

   We compute $\widehat{\tilde{H^j}}_{k,m}$ analogously.

   e) Calculate the weighted sum to obtain the model assessment criterion for all M models:

$$\widehat{MAC^j}_{k,m} = w_1 * \widehat{\widetilde{MSE^j}}_{k,m} + w_2 * \widehat{\tilde{H^j}}_{k,m}.$$

   f) Select the best model $m_{opt} = \min_m \widehat{MAC^j}_{k,m}$ and use $X_{k,m_{opt}}^j|R_k = 0$ to update $X_k^j|R_k = 0$.

2) Repeat step 1) J times or until convergence, i.e., $|(X_k^j|R_k = 0) - (X_k^{j-1}|R_k = 0)| < c_k, \forall_k$ with $c_k > 0$, and use the imputed values from the last iteration to obtain one imputed data set.
3) Repeat steps 1)-2) l times to obtain l multiply imputed data sets.

The observations $\mathbf{X}_{-k}^j|R_k = 0$ are missing for the first iteration ($j = 1$). Therefore, the variables $\mathbf{X}_{-k}^j$ are excluded from the estimation of the response propensity score and the imputation models. The differences in step 1) are as follows:

1) Repeat for all $k \in \{1, \ldots, K\}$ variables containing missing values:

   a) Estimate the response propensity score $\hat{e}_k^1 = P(R_k = 1|\mathbf{Z})$ for all n values in X_k.
   b) Estimate the density of residuals for X_k regressed on $\hat{e}_k^1$ (Equation 3.4) for the observed values $\hat{f}(X_k|\hat{e}_k^1, R_k = 1)$ using kernel density estimation (Equation 3.6).
   c) Repeat for all $m \in \{1, \ldots, M\}$ potential imputation models:
      - Fit the model m to $X_k|R_k = 1$ as the dependent variable and $\mathbf{Z}|R_k = 1$ as the independent variable.

The remaining part of the proposed procedure is the same for the first iteration. In the presented procedure, a different imputation model can be selected for each incomplete variable in each SI iteration. After multiply imputing and receiving l complete data sets, data analysis can be performed using analysis tools for complete data sets. The resulting l estimates can be combined with Rubin's rule (Rubin, 1987, Chapter 3) to compute valid point and variance estimates.

3.2.3 Missing Values in Categorical Variables

For categorical variables, the structure of SIIMS introduced in the previous section remains the same, only the assessment criteria ($\widehat{MSE}_{k,m}$ and $\hat{H}_{k,m}$) are replaced. The following sub-sections introduce criteria for incomplete binary and nominal variables.

3.2.3.1 Binary Variables

3.2.3.1.1 Prediction accuracy criterion The previous MSE criterion cannot be applied to binary variables with missing values ($X \in \{0, 1\}$). We thus use the absolute distance for each observation and substitute Equation 3.1 with

$$S_{i,m} = |\bar{X}_{i,m} - X_i|. \tag{3.8}$$

As in the continuous variable case, $\bar{X}_{i,m}$ is the average of repeated draws for each value $X_i|R_i = 1$. The $S_{i,m}$ are then combined as described in Equation 3.2.

3.2.3.1.2 Similarity criterion For the second measure, comparable to continuous variables, the conditional densities of observed and missing values should be similar under MAR (similarity indicated by $\sim$):

$$f(X|e, R = 1) \sim f(X|e, R = 0) \Leftrightarrow P(X = 1|e, R = 1) \sim P(X = 1|e, R = 0)$$

For a well performing imputation model, the distance between $P(X = 1|e, R = 1)$ and $P(X = 1|e, R = 0)_m$ should be small, with $P(X = 1|e, R = 0)_m$ being estimated from the potentially imputed values generated from model m. However, for binary X, a comparison based on residual densities, as done in the continuous case, is not feasible. Thus, for each model m, we propose comparing $P(X = 1|R = 1)$ and $P(X = 1|R = 0)_m$ within S subsets of similar e values. Let $(d_s, d_{s+1}]$, $s \in \{1, \ldots, S\}$, be half-open intervals of equal lengths with $d_s \in [0, 1], d_s < d_{s+1}$, and $\bigcup_{s=1}^{S-1} (d_s, d_{s+1}] = [0, 1]$. Further let

$$|\Delta|_{(d_s,d_{s+1}],m} = |P(X = 1|R = 1) - P(X = 1|R = 0)_m|_{(d_s,d_{s+1}]} \tag{3.9}$$

define the absolute difference between the probabilities of $X = 1$ for observed and predicted missing values from the fitted model m restricted to $e \in (d_s, d_{s+1}]$. The number of intervals S can be determined by the number of observations in the data. The sum of all absolute differences, $|\Delta|_m = \sum_{s=1}^{S-1} |\Delta|_{(d_s,d_{s+1}],m}$, for all intervals of e leads to an overall similarity criterion, i.e., a small $|\Delta|_m$ indicates $f(X|e, R = 1) \sim f(X|e, R = 0)_m$. All measures needed in Equation 3.9 can be estimated in a straightforward way from the observed data, or from the imputation model m.

3.2.3.2 Nominal Variables

3.2.3.2.1 Prediction accuracy criterion For multiple categories, the prediction accuracy criterion can be applied in the same manner as in the nominal case by transforming the nominal comparison to a binary comparison. After generating repeated draws for each value $X_i|R_i = 1$, the assessment is based on the draw being in the right category as opposed to the draw being in the wrong category.

3.2.3.2.2 Similarity criterion Now we expand the approach of comparing densities of observed and imputed values presented in sub-section 3.2.3.1.2 to multiple categories. The underlying principle can be applied to incomplete variables with multiple (C) categories:

$$f(X|e, R = 1) \sim f(X|e, R = 0) \Leftrightarrow P(X = c|e, R = 1) \sim P(X = c|e, R = 0), \quad (3.10)$$

$\forall c \in \{1, \ldots, C\}$. Analogous to the binary case (Equation 3.9), we can define a sum of differences over all categories within classes of response propensity score values:

$$|\Delta|_{(d_s,d_{s+1}],m} = \sum_{c=1}^{C} |P(X = c|R = 1) - P(X = c|R = 0)_m|_{(d_s,d_{s+1}]}.$$

The overall measure for similarity is defined analogously to the binary case as $|\Delta|_m = \sum_{s=1}^{S-1} |\Delta|_{(d_s,d_{s+1}],m}$.

3.3 Case Study

In this section, we first describe an implementation of the proposed SIIMS procedure, then we introduce the simulation setup, after which we report the results. For both continuous and binary incomplete variables, we incorporate four different state-of-the-art imputation models in SIIMS: Bayesian regularized linear model (BLM), CART, RF, and BART. Technical details on each procedure are provided in the following paragraphs, along with information concerning the assessment process and general notes on the implementation. We apply the procedure with balanced weights, leading

to a mean of both criteria ($w_1 = w_2 = 0.5$). We used the R software version 4.1.2 (R Core Team, 2021) for this case study.

3.3.1 Continuous Incomplete Variables

We now describe all applied imputation models for the outcome X_k and covariates $\mathbf{Z}, \mathbf{X}_{-k}$. The observed part of X_k ($X_k|R_k = 1$) has length n_k^{obs}. The missing part of X_k ($X_k|R_k = 0$) has length n_k^{mis}. Let further γ_k denote the parameter vector of the parametric imputation model with X_k as the outcome.

The first imputation model is the BLM. This procedures applies elastic net, the best performing regularization technique in Deng et al. (2016), before imputing via a Bayesian linear model (Rubin, 1987, p. 167). First, the Bayesian linear model is specified here as follows:

$$f(X_k|\mathbf{Z}, \mathbf{X}_{-k}, \gamma_k, log(\sigma_k)) = (\mathbf{Z}, \mathbf{X}_{-k})\gamma_k + \epsilon, \text{ with } \epsilon \sim N(0, \sigma_k^2). \tag{3.11}$$

We further assume improper prior distributions for the parameters, $P(\gamma_k, log(\sigma_k)) \propto const.$ The BLM fitting process in one SIIMS step is as follows.

1. Fit the elastic net regularized regression model using $X_k|R_k = 1$ as the outcome, $\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$ as the predictor variables, and the parameters γ_k with the following loss function

$$L(\hat{\gamma}_k) = \frac{\sum_{i=1}^{n_k^{obs}} ((X_k|R_k = 1) - (\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)\hat{\gamma}_k)^2}{2n_k^{obs}} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{q} \hat{\gamma}_{j,k}^2 + \alpha \sum_{j=1}^{q} |\hat{\gamma}_{j,k}| \right),$$

   with λ describing the regularization parameter and $\alpha \in [0, 1]$ describing the elastic net mixing parameter.
2. Use the model in 1. to identify the active set of covariates $(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)_{A,k}$.
3. Compute the matrix $\mathbf{S}_k = (\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)'_{A,k}(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)_{A,k}$.
4. Compute $\mathbf{V}_k = (\mathbf{S}_k + diag(\mathbf{S}_k)\kappa)^{-1}$, with κ describing a small ridge parameter.
5. Compute $\hat{\gamma}_k = \mathbf{V}_k(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)'_{A,k}(X_k|R_k = 1)$.
6. Draw $g \sim \chi^2_{n_k^{obs}-q}$.
7. Compute $\sigma_k^2 = ((X_k|R_k = 1) - (\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)_{A,k}\hat{\gamma}_k)'((X_k|R_k = 1) - (\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1)_{A,k}\hat{\gamma}_k)/g$.

8. Draw q i.i.d. $\mathbf{w}_1 \sim N(\mathbf{0}, \mathbf{1})$.
9. Compute $\mathbf{V}_k^{1/2}$ using Cholesky decomposition.
10. Compute $\dot{\gamma}_k = \hat{\gamma}_k + \sigma_k \mathbf{w}_1 \mathbf{V}_k^{1/2}$.
11. Draw n_k^{mis} i.i.d. $\mathbf{w}_2 \sim N(\mathbf{0}, \mathbf{1})$.
12. Compute n_k^{mis} values of $(X_k|R_k = 0) = (\mathbf{Z}, \mathbf{X}_{-k}|R_k = 0)_{A,k}\dot{\gamma}_k + \mathbf{w}_2\sigma_k$.

For further information on regularization via elastic net see e.g., Zou and Hastie (2005). The parameter α is determined via 5-fold cross-validation. We implement this procedure based on the R packages glmnet (Simon et al., 2011) (for the regularized model) and rstan (Stan Development Team, 2019) (for the Bayesian model). B sets of draws for $X_k|R_k = 0$ from the posterior predictive distribution (PPD) are obtained for use in the MSE criterion. Only one set of draws from the PPD is used to compare the densities of potentially imputed and observed values.

For CART, we implemented the model as proposed for imputation by Doove et al. (2014). The imputation process in one SIIMS step works as follows.

1. Apply CART on outcome $X_k|R_k = 1$ and covariates $\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$ using recursive partitioning.
2. For each observation in $X_k|R_k = 0$,
    - identify its corresponding terminal node in the fit CART (each terminal node includes a subset of $X_k|R_k = 1$).
    - randomly draw one observation from the observations (donors) in the identified terminal node.
    - impute the observed value from that donor.

The described CART procedure is based on the R package rpart (Therneau & Atkinson, 2019). In this implementation, the parameter minbucket (the minimum number of observations in any terminal node) is set to 5, and the parameter maxsurrogate (the number of competitor splits retained in the output) is set to 0. The B sets of draws for $X_k|R_k = 0$ are generated from draws from the terminal nodes for each observation. For the similarity criterion, only one value is drawn from the terminal nodes for each observation.

Another tree-based model is RF, which consists of multiple CARTs. We use RF in the SIIMS framework as used by Doove et al. (2014). The procedure works as follows.

1. Draw b bootstrap samples from $X_k, \mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$.
2. Apply one CART on each bootstrap sample with the outcome $X_k|R_k = 1$ and the covariates $\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$.
3. For each observation in $X_k|R_k = 0$:

    - identify its corresponding terminal nodes in all b CARTs.
    - randomly draw one donor from the pooled donors in all b terminal nodes.
    - impute the observed value from that donor.

For RF, we use the implementation of the randomForest package (Liaw & Wiener, 2002). Here, the minimum node size is set to 5, as in Doove et al. (2014), and number of trees is set to $b = 20$. The MSE criterion uses B sets of draws, generated from predictions of B randomly selected single trees. The set of draws for the similarity criterion is generated from normal draws with means and standard deviations estimated from all 20 trees.

All BART applications within SIIMS are based on Xu et al. (2016). The BART model consists of a sum of trees with estimation based on a Bayesian probability model. For an outcome vector $X_k|R_k = 1$ and a covariate matrix $\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$, the BART model is defined as:

$$(X_k|R_k = 1) = f(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1) + \epsilon \approx \sum_{j=1}^{g} T_j^M(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 1) + \epsilon,$$

with $\epsilon \sim N(0, \sigma_k^2)$ denoting a vector of error terms. T_j represents a single tree structure, with its parameters in the terminal nodes M. BART consists of a sum of g trees. Prior distributions are assigned to T, M, and σ_k^2. Draws from the posterior distribution $P(T_1^M, \ldots, T_g^M, \sigma_k^2|Z_k)$ are generated via Gibbs sampling (Geman & Geman, 1984), where the jth tree is fit iteratively. See Kapelner and Bleich (2016) for further details.

In this study, the imputation using BART in one SIIMS step works as follows.

1. Fit BART on $X_k, \mathbf{Z}, \mathbf{X}_{-k}|R_k = 1$.
2. Generate draws from posterior distribution of $\hat{P}(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 0)$.
3. Impute each observation in $X_k|R_k = 0$ with the corresponding draw from $\hat{P}(\mathbf{Z}, \mathbf{X}_{-k}|R_k = 0)$.

In this study, we use an implementation based on the R package bartMachine (Kapelner & Bleich, 2016, version 1.2.6) with $g = 50$ trees, because Kapelner & Bleich (2016) show that a lower number of trees can lead to low performance. The parameter $mem\_cache\_for\_speed$ is set to "FALSE" to avoid memory problems in larger data situations. For the MSE criterion, B sets of draws are generated from the PPD. The similarity criterion is based on the one set of draws from the PPD created in point 3.

3.3.2 Binary Incomplete Variables

For binary incomplete variables, all models target a binary outcome variable. The BLM consists of a logistic regression model, and the other procedures now perform classification instead of regression. The MSE criterion used is now replaced with a prediction accuracy criterion, as described in Section 3.2.3.1.1. The binary counterpart of the Hellinger distance is implemented as proposed in Section 3.2.3.1.2 with $S = 4$ subsets. The finally imputed values are the predicted classes of the selected procedure. All other settings are the same as those defined for continuous variables. However, there are differences in the implementation, as described below. For BLM, logistic regression is performed using the package rstan (Stan Development Team, 2019). For all models, draws for the prediction accuracy criterion are probabilities. For BLM and BART, the draws come from the PPD, and for CART and RF, they are estimated from the terminal nodes, respectively single trees in RF.

3.3.3 Imputation Step

After a best model is selected, we impute the missing values as follows. For Bayesian models, the values are drawn from the PPD. For RF, the imputed values are drawn from single trees, following Shah et al. (2014). CART selects imputed values from terminal nodes, following Burgette and Reiter (2010). BART draws values for imputation from the PPD, following Xu et al. (2016).

3.3.4 Further Details

Starting at the first iteration, variables are imputed ordered by completeness, starting with the variable with the fewest number of missing values, to exploit potential mono-

tone missing data patterns (see Little and Rubin p. 11-12 (2019)). We set $B = 20$ draws to estimate the MSE criterion. For the similarity criterion, the response propensity score is estimated using BART.

Several checks are implemented to detect extreme distributions in partially observed variables. Depending on the type of extreme distribution, we either remove the incomplete variable or we change the imputation process. First, for a variable with constant observed values, all missing values are imputed by that constant value. For a binary variable with a low frequency in one category ($< 4\%$), all missing values are imputed by the modal category, because some of the applied models returned problems in the fitting process. If a variable has few or no observed values (< 20 observations), it is not imputed and removed from the imputation process. In some cases, such as for a small number of missing values in the variable (10 observations), the similarity criterion cannot be computed because of unstable kernel density estimates. When this occurs, model assessment is based on the MSE criterion (prediction accuracy criterion in the binary case) only.

The current implementation of SIIMS imputes and updates missing values with a maximum number of $J = 5$ iterations. Convergence checks start after the second iteration. For binary variables, less than 5 of the imputed values are allowed to change compared to the previous iteration; for continuous variables, the relative variance of differences between currently and previously imputed values needs to be < 5. The process stops when all incomplete variables fulfill the convergence criteria or when the maximum number of iterations is reached. The procedure provides multiply imputed data sets along with information about the selected imputation model of each incomplete variable in each iteration.

3.3.5 Simulation Setup

This section describes the process of evaluating SIIMS with real data. We generate data sets with missing values from a publicly available version of the NHANES (National Health and Nutrition Examination Survey) data. The process of evaluation follows Ezzati-Rice et al. (1995) and Schafer et al. (1996). A detailed description can be found in the following Section 3.3.6. We use the 2015-2016 NHANES data, consisting of about 12,000 respondents, as a synthetic population. The data set consists of data from different modes of data collection: all respondents completed an initial questionnaire,

complemented by a physical examination taking place in a mobile examination center and two days of recording a nutrition diary. For missing data patterns, we find values of variables for the physical examination completely missing for some observations, because some respondents refused to participate in this part of data collection. There are also incomplete nutrition diaries, notably on the second day, in addition to item-missing data for sensitive questions in the questionnaire.

3.3.6 Assessment Process

Figure 3.1 depicts the process used to evaluate the different imputation procedures, which assumes an incomplete data set with observations in rows and variables in columns. We begin by defining the variables of interest (VOI), indicated in blue, that are later used to compare the different imputation procedures. After the VOI are defined, the data set (on the left side of the figure) is further divided into two subsets of variables: one that is fully observed (lefthand side, 31 variables), and one (right-hand side, 355 variables) with incomplete observations (indicated by dark gray fields) including the VOI (four variables). The 5,474 observations are then divided into two sets, one with fully observed VOI, and another where missing values are present in the VOI.

The evaluation process continues with repeatedly drawing simple random samples with replacement (SRSWR) of 500 observations from the subset of cases with completely observed VOI. These SRSWR-sub-data sets are the foundation of the evaluation, and consist of complete cases for all VOI, and the aforementioned subsets of variables. In the next step, missing value patterns observed in the data are introduced into the observations of VOI (indicated in red). These introduced missing values are donated from the observations of incomplete VOI matched to the observations in the subsampled data sets. In this study, only the VOI are imputed, the complete variables are used as covariates.

The matching of complete observations in VOI in each SRSWR with those observations including missing values in VOI in the original data set is based on 26 of the 31 completely observed variables. To perform the matching, we use the package StatMatch (D'Orazio, 2019) in the statistical software R (version 3.6.1) (R Core Team, 2019). Specifically, random hot deck (see Andridge & Little (2010)), a procedure that randomly selects a donor (i.e., the matched observation including missing values in the

VOI) for each observation in the SRSWR from an appropriate subset of all donors is carried out. The subset is built from the 26 completely observed variables. In order to use the variables in the random hot deck procedure, it is necessary to categorize them. All continuous variables with more than 10 unique values are categorized using deciles, resulting in a data set with maximal 10 categories of similar size per variable.

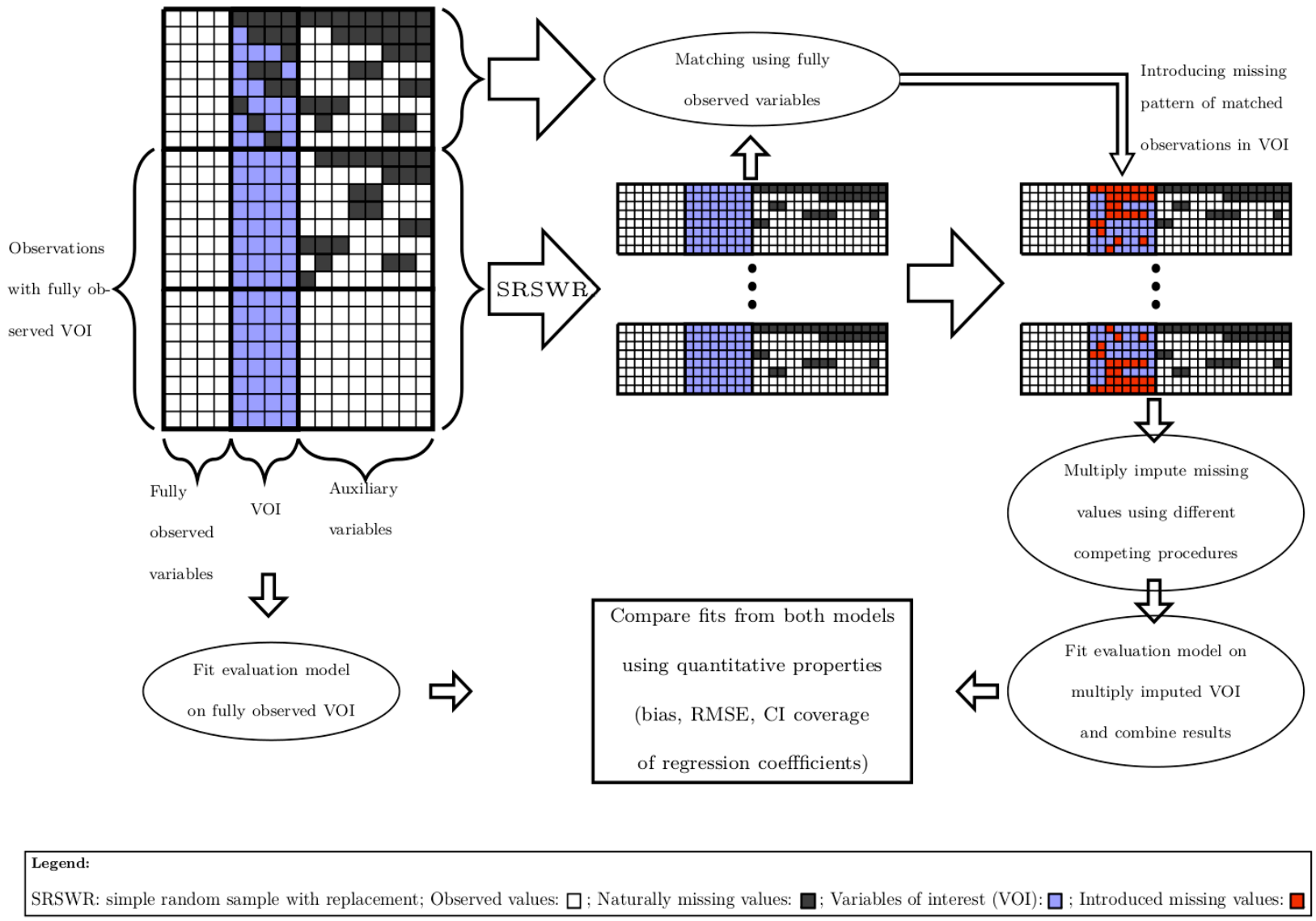

Figure 3.1: Structure of the evaluation process using NHANES data.

The matching is performed within donation classes defined by the variables AGE, RACE and GENDER. The following parameters were set in the matching process: dist.fun (the distance function used) set to "exact", cut.don (the rule to create the subset of the closest donors) set to "span", and k (the proportion of closest donors from all possible donors) set to 0.15. This setup allows 15% of the closest donors to be considered using exact matching. Since only observations with at least one missing value in the VOIs are considered for donating the missing data pattern, the resulting data set consists of observations without complete cases in the VOI. After finding a match for each observation in the subsampled data set, the missing pattern in the VOI of this matched observation is introduced. This procedure results in an ignorable missingness mechanism, conditional on the fully observed variables; see Ezzati-Rice et al. (1995) and Schafer et al. (1996) for further details.

This process has several advantages. For instance, the missingness pattern introduced in the VOI is actually observed in the data set. That is, missing patterns are neither artificial nor unrealistic. Further, this approach avoids the need to specify a probabilistic model for introducing missing values, an often used approach when evaluating imputation procedures using simulation. Together with this advantage, no specified model is needed for the data generating process; both the data sets and the relationships among variables are observed. This leads to an evaluation of missing data imputation procedures that is closer to real world missing data problems.

3.3.7 Variables of Interest (VOI)

There are several selection criteria for the VOI in place:

1. Relationship: the selected variables should follow an approximately linear relationship, because a linear model should be fit as an analysis model.
2. Missing values: in order to introduce missing data patterns following Ezzati-Rice et al. (1995), the VOI should have a relatively high number of missing values (i.e., $20 - 40\%$). Next, for the incomplete cases to provide substantial information, missing values should be present in predictors (Little, 1992). Additionally, in order to observe different missing data patterns in the VOIs, they should also be collected from different modes of data collection.

3. Population: The variables should target the whole population of NHANES (i.e., no sub-populations like "smokers"), to avoid "not applicable" cases which would reduce the number of observations used.
4. Time period: the variables should be measured in NHANES data collection 2015/2016.

For the selection of VOI, it would have been ideal to base the choice on a substantive paper that used NHANES data. However, a search for recently published studies using NHANES 2015/2016 revealed that, if regression is used as an analysis tool, covariates mostly consist of socio-demographic variables, which are (almost) completely observed. With this in mind, we decided to select variables fulfilling the four criteria from the NHANES data and use them in a hypothetical substantive regression model.

For the outcome variable, we use the "total calorie intake day 2 of the nutrition diary" (CAL). The covariates are the "loud noise indicator" (NOISE) (binary), "Sagittal Abdominal Diameter 1st" (SAD) (cm), and "carbohydrate intake day 1" (CARB). The four variables are included in the following analysis model for the i-th observation:

$$CAL_i = \beta_0 + \beta_{NOISE} NOISE_i + \beta_{SAD} SAD_i + \beta_{CARB} CARB_i + \epsilon_i, \tag{3.12}$$

with $\epsilon_i \sim N(0, \sigma^2)$. The NHANES 2015-2016 contains 3,294 observations with complete VOI. The model in 3.12 fit on these observation returns a significant effect for the coefficients for NOISE and CARB. Figure 3.2 shows the missing data pattern for the selected VOIs in the NHANES 2015/2016 data, with blue indicating 'observed', and gray meaning 'missing'. The left side of the figure shows the proportion of missing values for each variable. The right side of the figure displays the most frequent missing patterns appearing in the data (omitting patterns of frequencies smaller than 0.01), including the frequencies of the displayed missing patterns on the very right. As seen in the bar chart on the left, CAL has the highest proportion of missing values (approximately 0.23), followed by SAD (approximately 0.16), NOISE (approximately 0.09), and CARB (approximately 0.08). Further, from the graph on the right we see that the proportion of complete cases (in VOI) is 0.602.

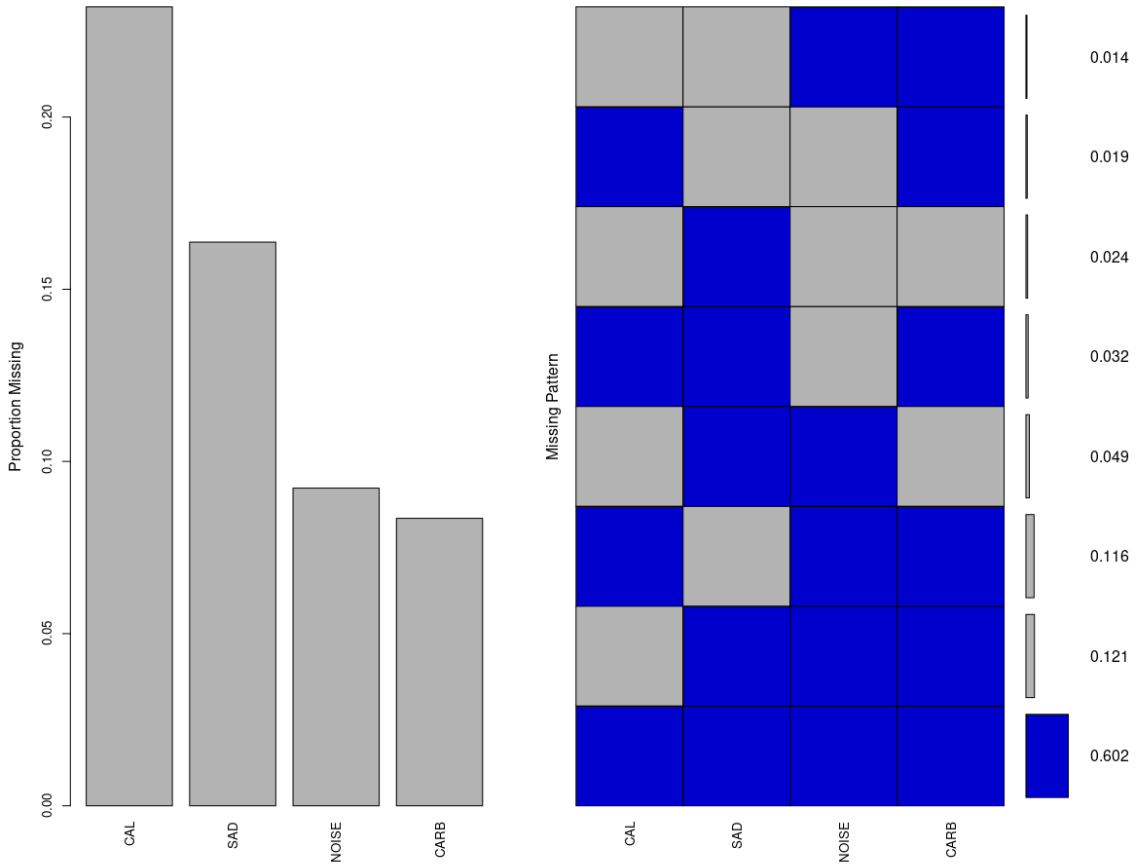

Figure 3.2: Overview of missing values found in NHANES 2015/2016 data for the variables of interest (VOI). Blue indicating 'observed', and gray meaning 'missing'. On the left, a bar chart displays proportions of missing values. The right side of the figure shows a plot with the most frequent missing patterns appearing in the VOI (omitting patterns of frequencies smaller than 0.01). On the very right, the plot displays the frequencies for each missing pattern.

3.3.8 Results

We compare SIIMS using five different sets of criteria weights w_1, w_2 ($\{0, 1\}$, $\{0.25, 0.75\}$, $\{0.5, 0.5\}$, $\{0.75, 0.25\}$, $\{1, 0\}$) with only applying its component methods (BART, CART, BLM, RF) using selected quantitative properties, namely empirical bias (EB), ratio of estimated variance to empirical variance (RV), RMSE, and confidence interval coverage rate (CICR) (Table 3.1). Additionally, we investigate the selection frequency of the models assessed within SIIMS and the runtime of the compared models. All applied methods use $J = 5$ iterations of SI and $l = 5$ multiply imputed data sets. For the simulation, a server with an E5 Xeon processor was employed. The results are averaged over 40 simple random samples with replacement with the previously described evaluation process.

Table 3.1: SIIMS with different criteria weights and the component methods compared in terms of the resulting empirical bias (EB), ratio of estimated variance to empirical variance (RV), root mean squared error (RMSE), and confidence interval coverage rate (CICR) in the estimated regression coefficients. EB, RV, and RMSE values are multiplied by 1,000. CICR values are multiplied by 100. The best value in each line is indicated in bold. The cells showing true values are highlighted in gray.

		Component Methods				SIIMS				
		BART	CART	BLM	RF	0, 1	0.25, 0.75	0.5, 0.5	0.75, 0.25	1, 0
CARB	True value	3142	3142	3142	3142	3142	3142	3142	3142	3142
CARB	EB	**-39**	-809	-2443	-591	-959	-969	-503	-1147	-1112
CARB	RV	82	53	**1504**	234	439	504	344	549	423
CARB	RMSE	1149	1489	2498	1003	1228	1207	**992**	1317	1305
CARB	CICR	42	20	3	60	48	50	**72**	49	45
SAD	True value	-1204	-1204	-1204	-1204	-1204	-1204	-1204	-1204	-1204
SAD	EB	2154	**1914**	2062	7717	4114	4817	6546	5911	5212
SAD	RV	93	166	**1138**	336	412	408	302	452	438
SAD	RMSE	25462	16832	**3672**	15869	11649	13648	17350	12442	11903
SAD	CICR	49	62	87	76	**88**	73	75	85	81
NOISE	True value	-155503	-155503	-155503	-155503	-155503	-155503	-155503	-155503	-155503
NOISE	EB	294619	87076	135370	**57139**	109317	183099	143741	163611	157876
NOISE	RV	515	576	1843	**1207**	1547	1233	1273	1628	1580
NOISE	RMSE	311341	147825	142358	**94761**	131362	195520	164196	175059	171379
NOISE	CICR	4	78	21	**84**	64	43	50	46	45

We first focus on the EB of the regression coefficients of the VOI after MI. For two of the coefficients (NOISE and CARB), we see that all SIIMS procedures return EB values that lie between the highest and lowest EB values of the component methods (BART, CART, BLM, RF), i.e., the best and worst performing methods are component methods. For CARB, we find SIIMS performing overall better in terms of EB than their components. The different SIIMS weights lead to similar EB values overall. Looking at RV, we find for CARB that the SIIMS procedures result in a better RV value (closer to 1,000) overall, compared to their component methods. For SAD, BLM shows the best result. For the NOISE coefficient, we find the SIIMS procedures performing mostly in between their component methods and above 1,000. For performance in terms of the RMSE, we find in all coefficients (CARB, SAD, NOISE) that the tested SIIMS procedures perform among their component methods. For CICR, we find that SIIMS performs overall better than the components in the coefficient for SAD. For the other two coefficients we find the tested SIIMS procedures perform among the components.

Table 3.2: SIIMS with different criteria weights and the component methods compared in terms of the resulting empirical bias (EB) in estimated means of variables of interest. All values are multiplied by 1,000. The cells showing true values are highlighted in gray.

		Component Methods				SIIMS				
		BART	CART	BLM	RF	0, 1	0.25, 0.75	0.5, 0.5	0.75, 0.25	1, 0
CAL	True value	1948098	1948098	1948098	1948098	1948098	1948098	1948098	1948098	1948098
CAL	EB	930637	151485	6700	340766	304963	327915	342470	300072	296874
CARB	True value	251996	251996	251996	251996	251996	251996	251996	251996	251996
CARB	EB	12102	153	-134	5499	4858	7051	6534	7620	5727
SAD	True value	22987	22987	22987	22987	22987	22987	22987	22987	22987
SAD	EB	109	60	78	61	82	94	76	73	74
NOISE	True value	855	855	855	855	855	855	855	855	855
NOISE	EB	-24	3	7	3	-4	-21	-17	-16	-18

We now focus on comparing SIIMS with different criteria weights and the component methods in terms of the resulting EB in the estimated means of all VOIs (Table 3.2). We find for all mean estimates that BART shows the highest EBs, while CART and BLM perform best, overall. RF shows low EBs (similar to the best performing method) for the variables SAD and NOISE. Similar to EB in the regression coefficients, we find the SIIMS procedures return EB values that lie between the highest and lowest EB values of the component methods (BART, CART, BLM, RF). We find no influence of the different criteria weights on the EB values of the investigated means.

In Chapter 2 we find low performance for imputation by BART on parametric incomplete data (L), but high performance when the data include a non-parametric relationship (N) and overall less variance. To investigate if SIIMS can select the right model in each scenario, we applied SIIMS with different criteria weights on 100 replications of both L and N scenarios with missingness depending on the outcome (MXY) and the highest base-line missing rate ($logit(\delta_0^{X_1}) = 0.05$) for 1,000 observations and 5,000 observations in the data. Figure 3.3 shows the relative frequencies of the selected models for L and N cases. In the L case, we find that SIIMS always selects RF, regardless of the used criteria weights and sample size. In the N case for $1,000$ observations, the SIIMS criteria select RF in approximately 75% of the cases, BART in approximately 24%, and BLM in about 1%, regardless of the applied criteria weights. Comparing these results with those in Table 2.4 in Chapter 2, the rejection of BART and the selection of RF in the L scenario is desirable as RF works well in terms of quantitative properties. In the N scenario, although RF is selected most of the time, it shows low performance (also cf. Table 2.6, Chapter 2). In this same scenario, however, SIIMS selects BART, the best performing model on N data, in one fourth of the cases. In the N case for $5,000$ observations, the SIIMS criteria select RF almost exclusively, regardless of the weights. One possible explanation of this unexpected finding is the different software used in Chapters 2 and 3. While in Chapter 2 we apply RF within the software package MICE (Van Buuren & Groothuis-Oudshoorn, 2011), RF in SIIMS is based on a custom implementation. In Chapter 3, the number of trees is set to $b = 20$, because the MSE criterion uses $B = 20$ sets of draws, generated from predictions of B randomly selected single trees, which forces $b \geq B$. In Chapter 2, the number of trees is set to 10, based on recommendations by Doove et al. (2014).

We now present further findings regarding the SIIMS procedure. First, we focus on how often each model is actually selected for imputation, given the criteria. The relative

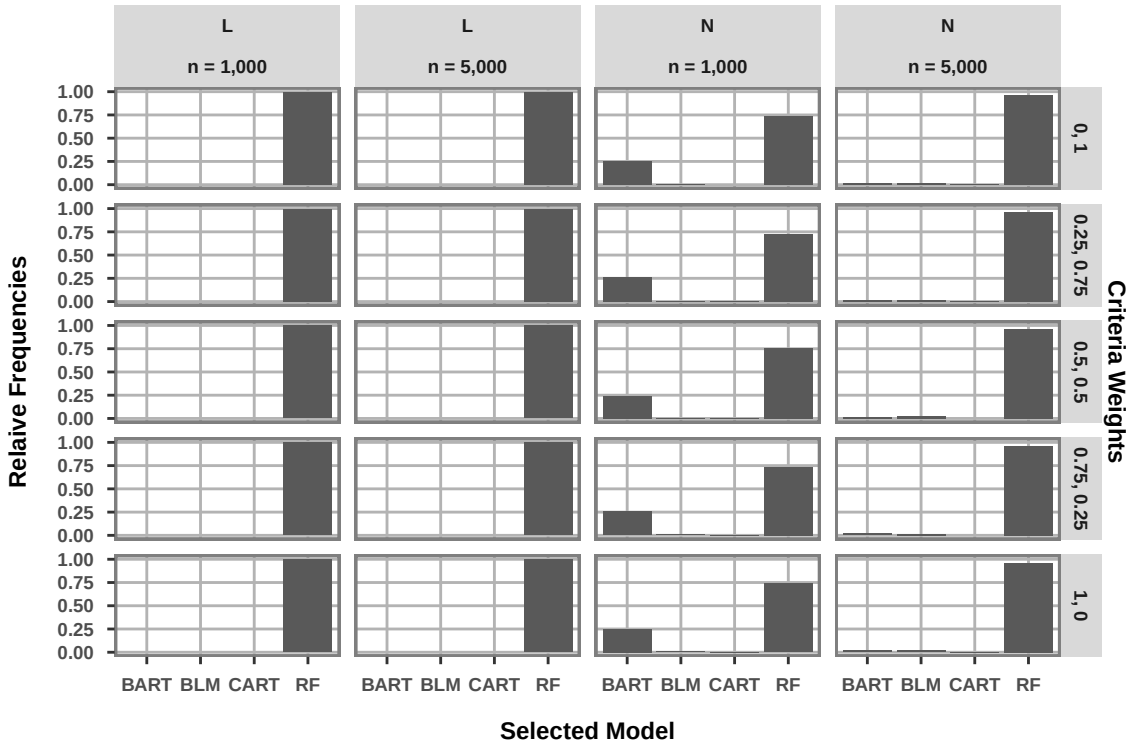

Figure 3.3: Barplot of selected models in SIIMS for L and N scenarios of Chapter 2. The rows separate different sets of criteria weights, the columns separate data scenarios (L, N) and numbers of observations in the data $(1,000,\ 5,0000)$. The y-axis displays relative frequencies.

frequencies over all $l = 5$ MIs; all $J = 5$ iterations; 280 incomplete variables; and all 40 replications is as follows. With a relative frequency slightly below 0.4, RF is the model selected most often, followed by BLM at approximately 0.29 and BART at slightly above 0.25. CART has the lowest relative frequency at approximately 0.07.

For the runtime of the applied models in SIIMS, BART, CART, and RF all require 10 seconds or less to fit and produce draws. BLM shows a median runtime of approximately 20 seconds. However, in some cases BLM requires several minutes (maximum at 6.75 minutes).

In summary, we find different criteria weighting in SIIMS lead to overall lower variation in performance in all investigated metrics. We do not find a best set of weights here. Generally, we find SIIMS with different criteria weights performing in between the component methods for both estimated regression coefficients and means.

3.4 Discussion

In this chapter we propose a modified SI procedure allowing for multiple competing models and plausibility checks during the imputation process. We further present a case study of this procedure in a simulation based on a real survey data set and compare it to the single models. The presented results suggest that SIIMS' performance mostly lies in between the performances of the single models.

We use two criteria to select imputation models within SI. The MSE criterion is computed on an observation level and then averaged over all observations to which the imputation model is fit. This criterion focuses on the predictive power of the imputation model. The similarity criterion automatically assesses the plausibility of imputed values by comparing their density to the density of the observed values of the incomplete variable to be imputed, conditional on the response propensity score. The higher the similarity between those two densities, the better the imputation model under MAR. This criterion compares the imputation models in terms of marginal distributions. The presented criteria can handle continuous, binary, and nominal variables.

While the current SIIMS implementation focuses on comparing tree-based and parametric models, the developed criteria can objectively compare a wide range of model types within SI. To be incorporated in the current SIIMS framework, an imputation

model only needs to be able to predict values of an outcome. We present current state-of-the-art imputation models; however, future imputation models can still be assessed using the same criteria proposed in this chapter. Additionally, the criteria can be used to compare models only differing in parametrization and perform parameter tuning that is specific to the imputation problem.

The SIIMS procedure shares some characteristics with multiple imputation by super learning (MISL) (Carpenito & Manjourides, 2022), which imputes missing values using weighted estimates from multiple models (see also Laqueur et al. (2021) for an implementation). Similar to SIIMS, super learning (SL) also requires a set of pre-specified models to be combined. The results of all these models are weighted by optimizing the prediction error (or another metric) via cross-validation. SL automatically selects the best combination of estimators, reducing the risk of misspecified models. While SIIMS selects one best model and uses the predicted values directly, MISL for continuous variables performs a form of PMM using the SL model to predict the observed and missing values of the variable to be imputed. Similar to SIIMS, for categorical variables MISL uses the predicted probabilities from the SL model to draw the imputed category.

Using SIIMS with the similarity criterion only also shares some characteristics with doubly-robust (DR) estimators for missing data, such as augmented inverse probability weighting (AIPW) (Robins et al., 1994) and penalized splines of propensity prediction (PSPP) (G. Zhang & Little, 2009). When either the mean model or the response propensity (RP) model is correctly specified, DR estimators produce consistent estimators of means. PSPP includes a spline of the estimated RP score in the mean model, similar to SIIMS' similarity criterion, where a spline is used to condition on the RP score. While DR estimators use one mean model, SIIMS selects from multiple mean models. A comparison between DR estimators and SIIMS under MAR reveals the following. When all imputation models in SIIMS are misspecified, the estimates will be inconsistent, regardless of the RP model's specification. Similarly, when all models compared within SIIMS are correct, the estimates will be consistent, again regardless of the RP model's specification. When at least one imputation model is correctly specified, and when the RP model is correctly specified, we expect the estimates to be consistent, because the similarity criterion can select the correctly specified model. When at least one imputation model is misspecified and the RP model is misspecified, we expect inconsistent estimates, because the criterion can no longer select correctly specified imputation models. While we already use BART as the RP model to further

increase robustness (Tan et al., 2019), the estimated RP score could also be included in the imputation models, as done in PSPP. Future studies could investigate a hybrid of SIIMS, PSPP, and MISL, such as RP scores included in all imputation models.

Regarding the results of the case study, we expected lower EB and RMSE values for SIIMS because the best model is selected according to plausible criteria, i.e., more checks for bad model fits are implemented. However, all compared procedures can perform variable selection and can capture important interactions in the fitting process. This might be the reason why SIIMS and its component models perform similarly well. Another possible explanation for a similar performance is that the criteria developed compare the models in terms of marginal distributions of the variables, while the evaluation focuses on joint distributions. Future research can base the model selection on criteria for regression models, like focusing on the distribution of residuals. Such a criterion would require a pre-specified structure of the substantive model before the imputation.

Another unexpected finding is the increased RMSE for BART in the SAD coefficient, compared to the RMSEs of the other component methods. The investigation of BART's unexpected behavior in Chapter 2 revealed a tendency for BART to produce relatively extreme imputed values in some data situations. This behavior is a possible explanation for BART resulting in such a high RMSE in this coefficient, since the same BART implementation is used in both Chapters 2 and 3.

The univariate analysis reveals a similar pattern as for the EB for regression coefficients. Specifically, the SIIMS procedures return EB values that lie between the highest and lowest EB values of the component methods. This finding suggests that SIIMS averages the component methods, rather than selecting the best. One possible explanation is that even one BART selection in SIIMS could introduce outliers that are carried forward in the SI procedures, regardless of the following selected methods. A further investigation could first examine which methods are selected, followed by the influence of a selected BART on the imputed values in the remaining iterations of SI. If BART leads to the underwhelming performance of SIIMS, future research needs to investigate the reason for the criteria selecting this method.

Due to the sequential nature of the procedure in conjunction with the assessment of multiple potential models, SIIMS is relatively slow compared to the runtime of a single model class. Removing slow models is a first step to increase runtime. For example,

while RF, BART, and CART require less than 10 seconds, BLM requires on average about twice as long to fit and produce draws. Thus, replacing BLM with a faster procedure could reduce runtime by several hours. Further, the selection frequencies of the models can be analyzed after the first iterations to identify low frequency models. The low-frequency models could then be dropped to increase process speed.

Another possibility for increasing the speed of SIIMS is to use an upstream variable selection. This selection process could function similarly to the function "quickpred" in MICE, which provides a subset of covariates for each incomplete variable to be imputed subsequently. Determining the subsets of covariates can be based on correlations with the corresponding outcome. Another option to reduce dimensionality and increase speed is principal component analysis (PCA) (Abdi & Williams, 2010). Before starting the imputation process, PCA can be performed on the complete variables of the data set. The resulting principal components can substitute the complete covariates in the imputation models. Since the complete variables are "fixed" in the imputation process, only one PCA is necessary before the imputation process starts. While PCA originally required continuous variables, the use of polychoric correlation allows PCA on mixed data as well (Kolenikov et al., 2004).

An evaluation of sets of imputed values from the same model could also be a time-saving alternative to assessing multiple models within SIIMS. These sets could be assessed by the similarity criterion, leading to a faster imputation process while keeping a plausibility check for the imputed values. See the appendix (Section 3.5) for further details on the procedure.

In the presented case study, we evaluate SIIMS based on different criteria weights. However, we find that the different weight choices fail to provide insights of whether SIIMS results in valid standard errors. A further investigation of this issue is needed, ideally in a more simple situation with data generated from a probabilistic model. Related to that, optimal weighting of both criteria might depend on the kind of analysis following the imputation. For instance, in univariate analyses, the optimal weight for estimating means and their standard errors likely uses only the MSE criterion. When kernel density estimation is performed, weights should emphasize the similarity criterion to focus on the whole distribution. Future studies could further evaluate both criteria separately, as well as explore weighting that results in a mixture of both proposed SIIMS selection criteria to find optimal weights for different analysis goals. Another venue for future research is the performance of SIIMS on only categorical,

binary, or continuous incomplete variables. Further, this study also did not investigate different models within SIIMS. The performance of the procedure likely depends on the criterion used and the pool of potential models evaluated.

While the evaluation process does not require a specified data generating model, several expansions could be explored. For example, pooling several waves of NHANES data would increase the variety of the sampled observations and the diversity of the introduced missing data patterns in the data. Further, a different set of VOI and different data sets can be used to obtain a more comprehensive picture of performance differences. Ideally, an independent third person determines a model of interest and the resulting VOI, to mimic the situation where the imputer and the analyst are different individuals.

3.5 Appendix 1 - SIIMS Modification: Rejection of Samples

The proposed SIIMS procedure in this chapter is computationally intensive when the number of variables with missing values (K) is large. Thus we present a more computationally-efficient alternative. Instead of fitting M models for each variable with missing values, only one model is used in each iteration. The predicted plausible values $X^j_{m,k}|R_k = 0$ from one model can be assessed by the proposed procedure using $\hat{H}^j$. The values can be rejected or accepted, based on a threshold H_0. The proposed procedure from Section 3.2.2 can be changed in step 1) as follows:

1) Repeat for all $k \in \{1, \ldots, K\}$ variables containing missing values:

    a) Estimate response propensity scores $\hat{e}^j_k = P(R_k = 1|\mathbf{Z}, \mathbf{X}^j_{-k})$ for all values in X_k.
    b) Estimate the density of residuals for X_k regressed on $\hat{e}^j_k$ (Equation 3.4) for the observed values $\hat{f}(X_k|\hat{e}^j_k, R_k = 1)$ using kernel density estimation (Equation 3.6).
    c) Fit a model with X^{j-1}_k as the dependent variable and $\mathbf{Z}$ and $\mathbf{X}^j_{-k}$ as the independent variables.
    d) Repeat until $\hat{H}^j < H_0$:

- Draw plausible values $\hat{X}_k^j | R_k = 0$ for $X_k | R_k = 0$ using model m.
- Estimate the density of residuals for X_k regressed on $\hat{e}_k^j$ for imputed values $(\hat{f}(X_k^j | \hat{e_k^j}, R_k = 0))$ using kernel density estimation.
- Estimate the Hellinger distance $\hat{H}^j = H(\hat{f}(X_k | \hat{e}_k^j, R_k = 1), \hat{f}(X_k^j | \hat{e}_k^j, R_k = 0))$.
- Compare $\hat{H}^j$ with H_0.

If the current set of values is rejected, a new set is drawn and assessed in the same way. However, modifying the procedure can lead to higher computational effort if the model fails to produce plausible values that fulfill the criterion, i.e., if the model is not a good fit to the particular outcome variable. To avoid a long search, H_0 can be modified after several unsuccessful tries. For example, starting with a low value of $H_0 = 0.01$ and increasing this value by 0.05 every 10 unsuccessful tries. Alternatively, the best set of plausible values among a pre-specified number of drawn sets can be used.

3.6 Appendix 2 - Design Table

Table 3.3: Design table for Chapter 3.

Method	Parameter	Description	Levels	Choices	Tuning
BLM	α	Elastic net mixing parameter	$[0, 1]$	Interval of all possible values.	5-fold cross-validation
CART	minbucket	The minimum number of observations in any terminal node used.	5	Default of MICE package version 3.14.7	None
CART	max-surrogate	Number of competitor splits retaine in the output	0	Not of interest here	None
RF	b	Number of trees	20	Minimum number needed for generating draws for MSE criterion	None
RF	minbucket	The minimum number of observations in any terminal node used.	5	Default of MICE package version 3.14.7	None
BART	mem_cache_for_speed	Speed enhancement that caches the predictors and the split values that are available at each node for selecting new rules.	TRUE, FALSE	Recommened for large number of predictors, set 'FALSE' in simulation with three variables (Chapter 2 data), set 'TRUE' in simulation using real data (Section 3.3.6).	None
BART	use_missing_data	If TRUE, incomplete observations are included.	FALSE	Only complete observations are used in each CEMI step.	None

All other parameters of the presented imputation procedures are specified as in the corresponding software packages.

Chapter 4

Multiple Imputation under Missing Not at Random: Incorporating Response Indicators into Sequential Imputation

Abstract

Multiple imputation (MI) of missing values is mostly applied under the assumption of missing at random (MAR), but the alternative missing not at random (MNAR) assumption may be more plausible. MI approaches that include response indicators (RIs) for incomplete covariates in predictions of missing values assume MNAR. This chapter investigates MI under MNAR assumptions using RIs as covariates. We review literature on imputation under MNAR and prediction with incomplete covariates. We then compare the performance of different strategies for incorporating RIs in a simulation with two objectives. The first objective is analytic inference. Specifically, MI methods are assessed based on the quantitative properties of regression coefficient estimates of a model fit to multiply imputed data. The second objective is descriptive inference that focuses on predicting a missing value as accurately as possible. From the simulation, we find that under an MAR mechanism in the data, methods including RIs perform as well as those without them. In MNAR data scenarios, methods including RIs can help to improve performance for both analytic and descriptive inference.

4.1 Introduction

Most data sets from sample surveys contain incomplete observations for various reasons, such as a respondent's refusal to answer certain questions. If researchers limit themselves to analyzing only the complete observations in a data set (i.e., cases where all variables are observed), they reduce the sample size and potentially lose power for statistical inference.

An additional issue with complete case analysis is that for many purposes, this method (implicitly) assumes that the missing data are missing completely at random (MCAR) (Little & Rubin, 2002, pp. 11–19), meaning that the distribution of the missing data is unrelated to variables in the data set. Exceptions where complete case analysis does not assume MCAR can be found in Little (1992) and Little and Zhang (2011).

A weaker assumption than MCAR is missing at random (MAR), which assumes that the distribution of missing data is related only to the observed (measured) variables (Little & Rubin, 2002, pp. 11–19). To handle missing values under MAR, a commonly-used tool is multiple imputation (MI) (Rubin, 1987, Chapter 3). MI is a general approach to statistical inference with incomplete data. Following MI, the imputed data set can be analyzed using complete-case data analysis tools. Generally, MI replaces the missing values with plausible values estimated from the observed values in the incomplete data set. This process creates complete data sets, which under MAR and correct model specifications can produce consistent estimates of univariate and multivariate quantities and valid estimates of uncertainty.

Besides MCAR and MAR, missing not at random (MNAR) models (Little & Rubin, 2002, pp. 11–19) allow missingness to depend on missing variables after conditioning on values of variables that are observed. The missingness mechanism in many applications is likely MNAR (e.g. nonresponse in sample surveys); thus, the alternative MNAR assumption may be more realistic. However, since available software like MICE (Van Buuren & Groothuis-Oudshoorn, 2011) or IVEware (Raghunathan et al., 2016) imputes missing values assuming MAR as a default, practitioners generally believe that MI always assumes MAR. In reality, however, MI can also be applied under MNAR models. These models are based on selection models, pattern-mixture models (PMM), or hybrids of both (Little & Rubin, 2002, Chapter 15). Including response indicators (RIs) for predictors into MI assumes a form of MNAR and can be applied without additional knowledge about the missing values, as demonstrated recently by Beesley

et al. (2021).

Earlier literature does not recommend including RIs in an analysis model targeted at the whole population (Greenland & Finkle, 1995; Jones, 1996), because this approach can lead to biased estimates (Donders et al., 2006; Knol et al., 2010; Vach & Blettner, 1991). However, in an MI analysis, the imputation model and analysis model can differ. Thus, it is possible to base MI on a model that conditions on RIs, but then analyze the imputed data with a model not conditioning on RIs, such as a regression model for the whole population. When training data for fitting a prediction model are incomplete, Loh et al. (2019) propose a tree-based algorithm that includes RIs of predictors (specifically GUIDE (Loh, 2002)). Their results suggest that the information about a value being observed or missing can be potentially useful in a prediction task.

Here, we will further explore MI when RIs in predictors are included in the imputation models and evaluate different procedures in terms of analytic and descriptive inference. In this chapter, we focus on analytic and descriptive inference, a major part of statistical analysis. For analytic inference, we consider inference about regression coefficients estimated from a sample drawn from a well-defined population. Inference about particular values of a variable in a test data set via a prediction model is viewed as a form of descriptive inference.

The remainder of this chapter is structured as follows. We first review relevant literature on MI under MNAR and then present work on predictive models, specifically tree-based procedures, fit to incomplete data. To explore MI under MNAR, we then present a simulation focusing on two possible deviations from MAR and compare methods under two analysis goals, analytic and descriptive inference. We end with a discussion and guidance for practice.

4.2 Literature Review

4.2.1 Imputation under MNAR

For data on variables $(X_1, \ldots, X_p, Z)$, let $\mathbf{X} = (X_1, \ldots, X_p)$, and let $\mathbf{X_{-i}}$ $(i = 1, \ldots, p)$ represent $\mathbf{X}$ without the i-th variable. Let $\mathbf{R} = (R_{X_1}, \ldots, R_{X_p}, R_Z)$ be a RI matrix with $R_{X_i,j} \in \{0, 1\}$ and $R_{Z,j} = 1$, for each subject j (i.e., Z is fully observed). Similarly, define $\mathbf{R}$ and $\mathbf{R_{-i}}$. Finally let $f()$ denote the probability distribution of the argument.

Generally speaking, MNAR models are based on selection models (see Equation 4.5), PMMs, or hybrids of both (Little & Rubin, 2002, Chapter 15). For the PMM representation, the joint distribution of $\mathbf{X}$ and $\mathbf{R}$ is factorized as

$$f(\mathbf{X}, \mathbf{R}|Z, \gamma, \pi) = f(\mathbf{X}|\mathbf{R}, Z, \gamma) f(\mathbf{R}|Z, \pi) \tag{4.1}$$

with π representing the parameter vector for the missingness mechanism (Little & Rubin, 2002) and γ representing the parameters in the model for each pattern. The factorization in Equation (4.1) specifies separate models for each response pattern. Methods that include RIs of covariates in predictions of missing values are effectively based on PMMs (Little, 1993).

A simple example of a PMM is the delta-adjustment procedure (Rubin, 1977), which adds a fixed parameter (called δ) to the imputed values under MAR that are obtained from posterior predictive distribution draws. For example, with a continuous incomplete variable X_1, a fully observed variable Z, and a simple MAR imputation model with the expected value modeled as

$$E(X_1|Z, R_{X_1}) = E(X_1|Z) = \beta_0 + \beta_1 Z, \tag{4.2}$$

a complementary delta-adjustment MNAR model is given by

$$E(X_1|Z, R_{X_1}) = \beta_0 + \beta_1 Z + \delta(1 - R_{X_1}). \tag{4.3}$$

Since values for δ cannot be estimated from the given data, plausible values are provided by subject experts or other auxiliary information like past studies that typically lead to a range of plausible δ values. In order to assess the impact of potential δ values, sensitivity analysis is often carried out (see e.g., Leacy et al. (2017) and Rezvan et al. (2018)).

Another imputation approach using RIs focuses on missing values in the outcome of a regression model and is based on randomly drawing an additional set of RIs, called random indicators (Jolani et al., 2012, Chapter 4). A response propensity model, also including the incomplete outcome itself, generates these indicators. In this MNAR-imputation method, the random indicators help to adjust the regression coefficients of the analysis model.

For multiple incomplete variables, e.g., X_1, X_2, and a complete variable Z, a common

way to execute MI under MAR is sequential imputation (SI) (see e.g., Van Buuren et al. (2006)). For all incomplete variables in the data, this approach requires specified imputation models for all conditional distributions, here $f(X_i|Z, \mathbf{X}_{-\mathbf{i}}), \forall_i$. The SI process starts with filling in the missing data with some plausible values. Next, SI iteratively updates the missing values variable by variable. For each variable, SI fits the corresponding imputation model and draws the values to be updated. This process is repeated several times, resulting in one imputed data set. These steps are repeated to produce multiply imputed data.

As with delta-adjustment for one incomplete variable, RIs can be incorporated into an SI procedure. Giusti and Little (2011) perform a sensitivity analysis using the delta-adjustment in one incomplete variable. Their SI procedure imputes one variable under MNAR, while missing values in other variables are imputed under MAR. Leacy (2016) introduced a general approach, where all conditional distributions incorporate all RIs:

$$f(X_i|Z, \mathbf{X}_{-\mathbf{i}}, \mathbf{R}), \forall_i. \tag{4.4}$$

An additive model in $(\mathbf{X}, Z)$ and $\mathbf{R}$ in this representation results in p sensitivity parameters $\delta = (\delta_1, \ldots, \delta_p)$, one for each incomplete variable, in addition to other parameters that can be estimated from the data. In order to apply the procedure, values for all sensitivity parameters need to be assumed. Tompsett et al. (2018) note that, given the conditional nature of each sensitivity parameter, performing sensitivity analysis becomes much harder. For one variable, δ is a marginal parameter, but for multiple incomplete variables, each sensitivity parameter describes the (mean) differences in X_i of two groups ($R_i = 1$ and $R_i = 0$) of subjects, conditional on Z, $\mathbf{X}_{-\mathbf{i}}$, and $\mathbf{R}_{-\mathbf{i}}$. Therefore, providing useful values for sensitivity parameters is quite challenging for experts. To overcome this problem, Tompsett et al. (2018) propose fully conditional specification only including $\mathbf{R}_{-\mathbf{i}}$ in the i-th equation in 4.4. They also state that all RIs $\mathbf{R}_{-\mathbf{i}}$ should always be included in the imputation models so that any possible relationships between the $\mathbf{R}$ and $\mathbf{X}$ can be exploited.

Although Tompsett et al. (2018) provide practical guidance on how to include RIs in imputation, the paper does not provide theoretical guidance for how specific missingness mechanisms impact their proposed procedure. Beesley et al. (2021) contribute here by providing theory for binary, nominal, and continuous incomplete variables, as well as an assessment via simulation. Motivating the problem with a Bayesian Markov

Chain Monte Carlo (MCMC) approach, the authors start by factorizing the joint distribution of $\mathbf{X}$ and $\mathbf{R}$ into a selection model

$$f(\mathbf{X}, \mathbf{R}|Z) = f(\mathbf{X}|Z)f(\mathbf{R}|\mathbf{X}, Z). \tag{4.5}$$

However, in an MCMC algorithm, this approach results in drawing from unidentified distributions $f(R_i|\mathbf{X}, \mathbf{R}_{-\mathbf{i}})$, because R_i also conditions on the incomplete X_i. Thus, they state the following specific assumptions for the approaches investigated:

A1) $R_i \perp R_{i'}|\mathbf{X}, Z \; \forall_{i \neq i'}$

A2) $f(R_i|\mathbf{X}, Z) = f(R_i|\mathbf{X}_{-\mathbf{i}}, Z)$.

Importantly, A2 leads to relaxing MAR to a specific MNAR assumption because the probability of observing a missing value for one variable is now allowed to depend on other potentially incomplete variables $\mathbf{X}_{-\mathbf{i}}$. Under these assumptions, the resulting general conditional distributions are still challenging to draw from directly. Thus, Beesley et al. (2021) provide details on approximations based on further assumptions. Their simulation compares several approximation approaches to SI including $\mathbf{R}$, e.g., including $\mathbf{R}_{-\mathbf{i}}$ as main effects only, or also adding interactions of $\mathbf{R}_{-\mathbf{i}}$ with covariates $\mathbf{X}_{-\mathbf{i}}$.

Another approximation introduced by Beesley et al. (2021) is based on modeling the mode instead of approximating the mean with Taylor-series approximations. Assuming a unimodal distribution and normally distributed data for each incomplete variable in an SI step, this approach (henceforth MI-NORM-OFFSET) first estimates the probabilities of observing the incomplete variables not being updated in this step using a logistic regression model. Second, these probabilities are incorporated to adjust the estimated mean of a normal distribution from which to draw the missing values. The generated draws are then used to update the missing values.

Also included in their work are SI without $\mathbf{R}_{-\mathbf{i}}$ (assuming MAR) and imputing from the ideal distribution (i.e., without approximations) as benchmark methods. In their simulation, the probabilities of missing in a covariate are allowed to depend on other incomplete covariates and the degree of the dependency is varied to result in deviations from MAR of different strength. The authors find that when the data are generated by a limited class of MNAR mechanisms, including RIs in MI can reduce bias in estimates of an analysis model. Further, methods including RIs show the same performance under

MAR compared to default SI procedures. They also find that including interactions of RIs with other covariates generally results in increased variance, but also in reduced bias. Assuming normality for all variables, the MI-NORM-OFFSET approach works best in the investigated simulation scenarios.

4.2.2 Prediction Models fit to Incomplete Data

Wang et al. (2006) consider the inclusion of RIs for descriptive inference (under the name "orthogonal coding scheme") describing it as "(...) one of the useful input coding schemes (that) are widely used in machine learning technology such as neural networks and support vector machines.". Wang et al. (2006) state further that "(i)n recent years, it (the orthogonal coding scheme) has been successfully used in various fields in biology, such as prediction of protein secondary structure (Qian & Sejnowski, 1988), solvent accessibility (Yuan et al., 2002), etc.".

Ding and Simonoff (2010) investigate different ways of incorporating missing indicators as predictors into classification trees (RPART, C4.5, and CART). One of their approaches recodes missingness as a separate category for categorical covariates and an unobserved value in continuous covariates (here called the "separate class method"). In their work, the impact of different MCAR, MAR, and MNAR scenarios as well as incomplete test and training data sets on classifying a binary outcome are investigated via mathematical theory and simulation (see Table 1 in Ding and Simonoff (2010)). They find that, overall, the separate class method performs best in terms of prediction accuracy, especially when RIs are related to the outcome of the prediction model and when test data are incomplete.

Twala (2009) surveys several other ways of using incomplete covariates, like surrogate variable splitting (Therneau et al., 1997) and fractional cases (FC) (Quinlan, 2014) in tree algorithms. Twala (2009) finds strengths and limitations for all methods reviewed. The study reports that the predictive performance of the methods depends on the proportion of missing values, the missingness mechanism, and the type of covariates. All procedures perform worst for the MNAR case, and missingness in one variable seems less severe than missing values in multiple variables. Overall, applying MI with the EM algorithm under an MAR model (Moon, 1996) (henceforth EMMI) before training a tree performs best, while listwise deletion performs worst.

A related study by Twala et al. (2008) investigated the performance of a method called "Missing Incorporated in Attributes" (MIA), which is similar to the separate class method in Ding and Simonoff (2010). In their comparison of MIA, EMMI, and FC, Twala et al. (2008) find that while MIA performs well in many settings with different missingness mechanisms, a computationally heavy combination of EMMI and MIA performs best overall.

Not investigated in either Ding and Simonoff (2010) or Twala et al. (2008) is the GUIDE procedure used by Loh et al. (2019). The GUIDE tree algorithm uses χ^2-tests to determine whether a split between the observed values and the missing values in an incomplete covariate should be performed (Loh, 2002). GUIDE, and its forest version (Loh, 2014), are assessed on a diverse range of data sets (Loh et al., 2013; Loh, Man, et al., 2019). Loh et al. (2019) set up missing value imputation in a continuous variable (income) as a prediction task. Biases and RMSEs of mean income after imputing are smaller for GUIDE than for compared MI procedures. This work is critiqued in detail in Section 4.2.3.

A Bayesian version of tree-based algorithms is Bayesian additive regression trees (BART). While the original BART implementation of Chipman et al. (2010) requires completely observed data, a paper by Kapelner and Bleich (2015) extends BART by incorporating an MIA option for use with incomplete covariates as well. The simulation results show that incorporating RIs via MIA into BART results in equal or better predictive performance in MCAR, MAR, and MNAR situations compared to the original BART procedure.

The previously described approaches to incorporating RIs in prediction models focus on specific methods. A more general idea is presented in Fletcher et al. (2020), where separate models are trained on observations with different response patterns in covariates. Specifically, the authors propose to first split the training data by response patterns in the covariates and then fit a separate prediction model (called pattern submodel) on each sub-data set. Fletcher et al. (2020) state that no further assumptions about the missingness mechanism are necessary, because submodels are fit to specific response patterns only. One disadvantage of this approach, however, is sparse data for some response patterns, when the number of response patterns is large, as argued in Loh et al. (2020). In this case, Fletcher et al. (2020) propose simplifying the models or combining patterns. In their simulation, looking at MAR and different MNAR scenarios, they find that, in terms of total prediction error, pattern submodels perform as well

as MI which includes RIs, and are faster computationally. This advantage in speed is particularly important when predictions in new data sets are needed in real-time.

To summarize, the literature focusing on descriptive inference using incomplete data overall supports including RIs of incomplete covariates. Multiple strategies, like the pattern submodel or the orthogonal coding scheme, have been proposed for the general use of RIs in prediction models. Further, specific methods like GUIDE or BART have been developed that incorporate RIs in incomplete training data.

4.2.3 Loh-Little Debate

MI of missing data can have two different analysis goals: analytic and descriptive inference. These two objectives are discussed in Loh et al. (2019) (Section 4.2.2), Little (2020), and the rejoinder from Loh et al. (2020). This debate is summarized here because it shows the importance of clearly stating analysis goals and motivates the following simulation in this chapter. Little (2020) considers a three-variable case where Z is fully observed, and X_1 and X_2 are incomplete with 4 patterns: both X_1 and X_2 complete, one missing, or both missing.

To impute missing values in X_1, the GUIDE tree used in Loh et al. (2019) is built on all observations where X_1 is observed, i.e. $R_{X_1} = 1$. Little (2020) shows that the main assumption underlying this procedure is that the distribution of X_1 (regardless of R_{X_1}) is the same conditional on R_{X_2}:

$$f(X_1|R_{X_1} = 1, R_{X_2} = 1, X_2, Z) = f(X_1|R_{X_1} = 0, R_{X_2} = 1, X_2, Z), \tag{4.6}$$

$$f(X_1|R_{X_1} = 1, R_{X_2} = 0, Z) = f(X_1|R_{X_1} = 0, R_{X_2} = 0, Z). \tag{4.7}$$

When X_2 is imputed, the roles of X_1 and X_2 and their corresponding RIs (R_{X_1} and R_{X_2}) in these equations are reversed (additional parameters are omitted for simplicity). As Little (2020) points out, this setup corresponds with a pattern-mixture model and a specific MNAR situation. In contrast, MAR assumes the following:

$$f(X_1|R_{X_1} = 0, R_{X_2} = 1, X_2, Z) = f(X_1|X_2, Z), \tag{4.8}$$

$$f(X_2|R_{X_2} = 0, R_Z = 1, X_1, Z) = f(X_2|X_1, Z), \tag{4.9}$$

and

$$f(X_1, X_2|R_{X_1} = 0, R_{X_2} = 0, Z) = f(X_1, X_2|Z). \tag{4.10}$$

Little (2020) further states that the relative plausibility of these two assumptions (Equations 4.6 and 4.7 on the one side and Equations 4.8 to 4.10 on the other) cannot be assessed based on a given data set. Further, if there is no conditioning on X_2 for $R_{X_2} = 0$ when imputing X_1, and also if X_2 is related to both X_1 and R_{X_1}, assuming this specific MNAR mechanism can lead to bias. In this situation, MI based on MAR imputing X_1 and X_2 iteratively (via SI) would lead to unbiased results. Further, Little (2020) points out that the simulation in Loh et al. (2019) is rather narrow and favors models including RIs.

Loh et al.'s (2020) rejoinder is based on a simulation study comparing imputation via GUIDE and sequential regression imputation using the MICE software package (Van Buuren & Groothuis-Oudshoorn, 2011) in terms of analytic and descriptive inference. In the descriptive inference case, imputation via GUIDE and sequential regression imputation (via MICE) are performed before applying multiple prediction methods. Additionally, they compare the training of methods without previous imputation and method-specific defaults for missing observations in covariates (like surrogate splits in classification and regression trees (CART)). The authors find that overall, GUIDE's forest version performs best for both analytic and descriptive inference.

Loh et al. (2020) compare methods based on a prediction accuracy measure and the bias in estimated regression parameters. An important goal of MI is accounting for uncertainty in the imputed values to obtain valid standard errors, tests, and confidence intervals. Thus, the evaluation of imputation procedures should also be based on a measure of uncertainty, like the RMSE of estimated regression coefficients.

While Loh et al. (2020) compare several methods with and without previous imputation, some procedures are not assessed in their simulation. First, they do not multiply impute the whole data set and average multiply imputed values for prediction; second, they do not use tree-based methods implemented in MICE (like CART (Doove et al., 2014) or random forest (Shah et al., 2014)). We compare both approaches with and without including RIs.

4.2.4 Analysis Goals

The Loh-Little debate points out the importance of distinguishing two goals of data analysis: analytic inference, focusing on inference about regression coefficients or univariate quantities estimated from a sample, and descriptive inference, focusing on inference about values of a variable in a test data set. Doing analytic inference generally also requires variance estimates of parameters. The success of descriptive inference can be assessed via a prediction accuracy metric.

When these analysis goals are not clearly distinguished, the results can be misleading. For instance, Loh et al. (2019) actually compare procedures for MI, where missing values are generally replaced by draws, with algorithms developed for descriptive inference, like GUIDE. Although Loh et al. (2020) evaluate the different methods based on the bias of estimated regression coefficients and prediction accuracy, assessment in terms of analytic inference also requires investigating estimates of variance and confidence coverage.

Another recent study by Dagdoug et al. (2023) follows a similar assessment approach as Loh et al. (2019), focusing on a point estimate of one incomplete variable; i.e., they ignore variance estimates and do not assess relationships between variables. In their study, a high number of non-parametric and machine learning methods are compared in terms of their single imputation (prediction) abilities. Specifically, the authors find that the Cubist algorithm (Quinlan, 1993), BART, and XGBoost (Chen & Guestrin, 2016) perform best in terms of finite population totals. An analysis of all observations of one particular variable is unbiased for point estimates, but the variance is generally underestimated, because single imputed and observed values are treated the same. The results of such a study can suggest that methods designed for descriptive inference perform better than MI procedures, even in the case of imputation for analytic inference. Therefore, a clear distinction between these two analysis goals, as well as the different corresponding assessment strategies, is necessary to reach valid conclusions.

4.3 Simulation

In this chapter, we extend the simulation in Beesley et al. (2021) to include another method based on the PMM factorization, and show that in some MNAR situations,

MI based on this factorization may have advantages over MAR methods. We simulate two deviations from MAR to investigate the effect of including RIs in imputation models: a selection model factorization and a scenario with correlated RIs. Using these generated data, we compare two types of imputation methods: those including and those excluding RIs. We further add to the Loh-Little debate by evaluating the applied methods in terms of both analytic and descriptive inferential properties. For analytic inference, we assess the methods by analyzing the multiply imputed data via a hypothetical regression model. We do this in terms of empirical bias (EB), ratio of estimated variance to empirical variance (RV), RMSE, and confidence interval (CI) coverage rates of β-coefficients (see Section 4.3.2). For descriptive inference, we base our assessment on the squared error of the predicted values (see Section 4.3.3).

4.3.1 Data Generating Process

Four variables $\mathbf{X} = (X_1, X_2, X_3, X_4)$ are generated from a multivariate normal distribution,

$$\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}),$$

with

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.3 & 0.4 & 0.4 \\ 0.3 & 1 & 0.3 & 0 \\ 0.4 & 0.3 & 1 & 0 \\ 0.4 & 0 & 0 & 1 \end{pmatrix}.$$

We now introduce missing values in X_1 with the following response probabilities:

$$p_{X_1} = logit^{-1}(\delta_0^{X_1} + \delta_1^{X_1} X_1 + \delta_2^{X_1} X_2 + \delta_3^{X_1} X_3 + \delta_4^{X_1} X_4). \tag{4.11}$$

The RI vector R_{X_1} is calculated by comparing the vector of p_{X_1} with values drawn from a random variable $u_{X_1} \sim Unif(0, 1)$ using the following decision rule:

$$R_{X_1} = \begin{cases} 1 \text{ for } p_{X_1} \geq u_{X_1}, \\ 0 \text{ for } p_{X_1} < u_{X_1}. \end{cases} \tag{4.12}$$

Finally, we introduce missing values in X_2 following the same procedure as for X_1:

$$p_{X_2} = logit^{-1}(\delta_0^{X_2} + \delta_1^{X_2} X_1 + \delta_2^{X_2} X_2 + \delta_3^{X_2} X_3 + \delta_4^{X_2} X_4 + \delta_5^{X_2} R_{X_1}), \tag{4.13}$$

with R_{X_2} results from

$$R_{X_2} = \begin{cases} 1 \text{ for } p_{X_2} \geq u_{X_2}, \\ 0 \text{ for } p_{X_2} < u_{X_2}. \end{cases} \tag{4.14}$$

This data generating process results in a specific MAR mechanism in the incomplete data set (X_1, X_2, X_3) for $\delta_1^{X_2} = \delta_2^{X_1} = \delta_1^{X_1} = \delta_2^{X_2} = \delta_4^{X_1} = \delta_4^{X_2} = \delta_5^{X_2} = 0$. Other fixed parameters are: $\delta_0^{X_1} = \delta_0^{X_2} = \delta_3^{X_1} = \delta_3^{X_2} = 0.5$. These parameter values result in approximately 50% missing values in both X_1 and X_2, with approximately 25% overlapping missing values and approximately 25% complete cases. The data were simulated using the R software version 4.1.2 (R Core Team, 2021); the code is available upon request. The simulated scenarios consist of $n = 1,000$ observations, all of which were replicated 200 times. We next describe the two deviations from MAR.

4.3.1.1 MNAR Scenario 1 (MNAR1)

For the first deviation from MAR, we link $\delta_1^{X_2} = \delta_2^{X_1}$ and investigate the scenarios of both $\delta_1^{X_2}$ and $\delta_2^{X_1} \in \{0, 0.5, 1, 1.5, 2\}$, similar to the simulation in Beesley et al. (2021). We further set $\delta_4^{X_1} = \delta_4^{X_2} = 0$ and ignore X_4 in imputation and evaluation. This setup is represented by the following selection model:

$$f(X_1, X_2, R_{X_1}, R_{X_2}|X_3) = f(X_1, X_2|X_3) f(R_{X_1}, R_{X_2}|X_1, X_2, X_3), \tag{4.15}$$

with additional parameters omitted for simplicity. Since $\delta_1^{X_1} = \delta_2^{X_2} = 0$, we can further decompose the second factor:

$$\begin{aligned} f(R_{X_1}, R_{X_2}|X_1, X_2, X_3) &= f(R_{X_1}|X_1, X_2, X_3) f(R_{X_2}|X_1, X_2, X_3) \\ &= f(R_{X_1}|X_2, X_3) f(R_{X_2}|X_1, X_3). \end{aligned} \tag{4.16}$$

4.3.1.2 MNAR Scenario 2 (MNAR2)

In this scenario, we set $\delta_2^{X_1} = 0$ and link $\delta_1^{X_2} = \delta_5^{X_2}$. We investigate $\delta_1^{X_2}$, $\delta_5^{X_2} \in \{0, 0.5, 1, 1.5, 2\}$. Further, we allow for the influence of X_4 on the RIs for both X_1 and

X_2 by simulating $\delta_4^{X_1}, \delta_4^{X_2} \in \{0, 0.5, 1, 1.5, 2\}$. As in MNAR1, we exclude X_4 from the analysis and thus, in this case, R_{X_1} and R_{X_2} serve as substitutes for X_4 for $\delta_4^{X_1}, \delta_4^{X_2} \neq 0$.

4.3.2 Objective 1 - Analytic Inference with Incomplete Data

The assessment for analytic inference is based on quantitative properties (EB, RMSE, and CI coverage) of the estimated β-coefficients from a linear model

$$X_1 = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \epsilon, \tag{4.17}$$

with X_1 as the outcome, X_2 and X_3 included as main effects, and $\epsilon \sim N(0, \sigma^2)$. The quantities are computed after combining multiple imputation results using Rubin's rule (Rubin, 1987, Chapter 3). All compared approaches perform SI with 5 iterations and produce 10 multiply imputed data sets. Next, we describe the imputation methods in more detail.

CART

We apply CART, following Doove et al. (2014). In this implementation, first, a regression tree is fit via recursive partitioning. Second, the terminal node for each observation with missing outcomes is derived. Finally, imputed values are drawn randomly from observations in the corresponding terminal nodes of the trees. We use the MICE implementation in the R package Multivariate Imputation by Chained Equations (MICE) (Van Buuren & Groothuis-Oudshoorn, 2011, version 3.14.0) with a minimum of five observations in terminal nodes. This method is henceforth called MI-CART.

CART-R

This SI method includes the RIs, R_{X_1} and R_{X_2}, as covariates along with the other variables in the data set into the imputation process, with all other settings equal to MI-CART (henceforth called MI-CART-R). Since CART can automatically model important interactions present in the covariates (Doove et al., 2014), the procedure can, in theory, exploit potentially useful information in RIs in combination with the original covariates.

NORM

This SI method uses MICE (Van Buuren & Groothuis-Oudshoorn, 2011, version 3.14.0)

to apply imputation via Bayesian linear regression models (henceforth called MI-NORM) following Schafer (1997).

NORM-OFFSET

This method is described in Section 4.2.1 as one of the approximations presented in Beesley et al. (2021) (see the article for further detail). For this simulation, we adapted the code from the corresponding online appendix of the article (https://github.com/lbeesleyBIOSTAT/SRMIMI_Example_Code). This method is henceforth called MI-NORM-OFFSET.

4.3.3 Objective 2 - Descriptive Inference with Incomplete Data

The assessment for descriptive inference is based on the squared error of the predicted values (SEV) for $R_{X_1} = 0$ observations:

$$SEV = \sum_{j \in \{X_1 | R_{X_1} = 0\}} (X^t_{j,1} - X^p_{j,1})^2, \tag{4.18}$$

with $X^t_{j,1}$ representing the true (t) value and $X^p_{j,1}$ representing the predicted (p) value of the j-th element in $\{X_1 | R_{X_1} = 0\}$.

The four approaches from Objective 1 (MI-CART, MI-CART-R, MI-NORM, MI-NORM-OFFSET) are used here to perform 10-fold MI. After MI, multiply imputed values are averaged to obtain one predicted value for each missing value. We also use BART to predict missing values in X_1 directly, as described below.

BART

BART is trained on the complete data with X_1 as the outcome variable, with no previous imputation. Predicted values for the missing values in X_1 are obtained from averaged posterior predictive distribution draws. The procedure is implemented in the R package bartMachine (Kapelner & Bleich, 2016, version 1.2.6).

BART-R

Here, BART is trained on all $R_{X_1} = 1$ observations with no previous imputation. Thus, R_{X_2} can be used as a covariate in the training process, as described in Section 4.2.2.

4.3.4 Expectations

For the MAR scenarios, we expect MI-NORM to perform best in terms of quantitative properties, because the imputation model is closest to the data generated. We further expect BART-R to perform better than BART in terms of SEV, because of the increased amount of training data.

For both deviations from MAR (MNAR1 and MNAR2), we generally expect methods including RIs to perform better in terms of quantitative properties and show lower SEV compared to the standard methods, because the information in RIs can be exploited.

For MNAR1 and Objective 1, we expect that MI-NORM-OFFSET performs better than MI-NORM in terms of quantitative properties, because the simulated scenario is similar to the one in Beesley et al. (2021). We also expect that MI-CART-R performs better than MI-CART with increased deviation from MAR.

4.3.5 Results

4.3.5.1 Objective 1

4.3.5.1.1 MNAR1 Table 4.1 shows the results for MNAR1 and Objective 1. For MNAR1, we first focus on the EB. The table shows that for MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$) all methods perform similarly well, resulting in approximately no EB in all regression coefficients. In the MNAR cases ($\delta_1^{X_2} = \delta_2^{X_1} \neq 0$), MI-CART and MI-NORM show the lowest EB for β_0, but result in the highest bias for β_1 and β_2 (MI-CART being more biased than MI-NORM for both parameters). While MI-NORM-OFFSET yields the best results for β_1 in MNAR cases with almost no bias, MI-CART-R performs best for β_2. In the MNAR cases, MI-NORM-OFFSET performs best over all β-coefficients in terms of EB, followed by MI-CART-R.

Table 4.1: Effect of deviating from MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$) to MNAR1. Different imputation methods are compared in terms of the resulting empirical bias (EB), ratio of estimated variance to empirical variance (RV), root mean squared error (RMSE), and confidence interval coverage rate (CICR) in the estimated regression coefficients. EB, RV, and RMSE values multiplied by $1,000$. CICR values multiplied by 100.

		EB			RV			RMSE			CICR		
$\delta_1^{X_2}$, $\delta_2^{X_1}$	Method	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
0, 0	MI-CART	-4	-7	-3	687	711	627	46	65	53	88	90	90
0, 0	MI-CART-R	-4	-2	-4	742	639	684	46	67	53	92	90	91
0, 0	MI-NORM	-5	-1	2	951	1084	892	45	56	50	95	97	94
0, 0	MI-NORM-OFFSET	8	-19	-6	962	1205	1023	46	57	48	94	98	96
0.5, 0.5	MI-CART	16	17	-20	628	887	647	51	68	58	86	90	88
0.5, 0.5	MI-CART-R	14	-21	-15	669	553	737	54	78	57	90	84	89
0.5, 0.5	MI-NORM	10	22	-12	905	1032	853	48	60	52	92	92	95
0.5, 0.5	MI-NORM-OFFSET	36	-23	-18	958	1455	984	59	55	51	88	97	94
1, 1	MI-CART	20	104	-53	629	989	690	54	123	76	86	66	74
1, 1	MI-CART-R	32	-54	-17	578	378	723	65	103	60	83	68	88
1, 1	MI-NORM	14	82	-36	1004	1003	930	48	99	60	94	70	88
1, 1	MI-NORM-OFFSET	58	-22	-28	1108	1538	1210	74	55	54	82	98	95
1.5, 1.5	MI-CART	17	193	-87	634	1201	763	53	200	102	85	16	52
1.5, 1.5	MI-CART-R	59	-81	-23	521	335	709	90	127	66	73	57	87
1.5, 1.5	MI-NORM	16	138	-60	979	1091	914	48	147	77	94	34	74
1.5, 1.5	MI-NORM-OFFSET	73	-11	-37	1194	1620	1206	86	51	59	71	98	90
2, 2	MI-CART	12	264	-115	613	1178	766	51	269	125	86	0	34
2, 2	MI-CART-R	73	-98	-27	428	265	644	107	149	70	68	56	82
2, 2	MI-NORM	12	189	-78	1063	907	966	45	197	91	96	11	60
2, 2	MI-NORM-OFFSET	81	0	-45	1307	1579	1291	94	52	64	70	98	88

For RV, for MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$), MI-NORM and MI-NORM-OFFSET show the best RV (closest to 1,000) in all regression coefficients, both other methods show lower values. In the MNAR cases ($\delta_1^{X_2} = \delta_2^{X_1} \neq 0$), for all β- coefficients, MI-NORM remains on a similar level as in the MAR case, while MI-NORM-OFFSET's RV increases with increasing $\delta_1^{X_2}, \delta_2^{X_1}$ values. Overall, MI-CART-R shows the lowest RVs in the MNAR case. Generally, MI-CART also leads to low RV values in MNAR, but there is an increase in β_1 for increasing $\delta_1^{X_2}, \delta_2^{X_1}$.

Focusing on RMSE, for MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$), all methods again perform equally well, resulting in approximately the same RMSE in all regression coefficients. For the MNAR cases ($\delta_1^{X_2} = \delta_2^{X_1} \neq 0$), MI-CART and MI-NORM have the lowest RMSE for β_0, but MI-CART results in the highest RMSE for β_1 and β_2, followed by MI-NORM. While MI-NORM-OFFSET shows the lowest RMSE for β_1 (followed by MI-CART-R), MI-CART-R and MI-NORM-OFFSET both perform best for β_2. In the MNAR cases, MI-NORM-OFFSET performs best over all β-coefficients in terms of RMSE, followed by MI-CART-R.

Looking at CI coverage rate, for MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$), all methods perform equally well at approximately 95% coverage for all regression coefficients. When there are deviations from MAR, MI-CART and MI-NORM perform best for β_0, but result in the lowest coverage rate for β_1 and β_2 (here, MI-CART performs worst, followed by MI-NORM). For both β_1 and β_2, MI-NORM-OFFSET returns the best coverage rate (at or close to 95%). MI-CART-R performs as well as MI-NORM-OFFSET for β_2, but has a lower coverage rate for β_1. In the MNAR cases, MI-NORM-OFFSET performs best over all β-coefficients in terms of CI coverage rates, again followed by MI-CART-R.

Overall, MI-NORM-OFFSET shows the best performance for all three metrics investigated, followed by MI-CART-R.

4.3.5.1.2 MNAR2 The results for the MNAR2 scenarios are presented in Tables 4.2, 4.4, and 4.5.

We first focus on the EB. Table 4.2 shows that for MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) all methods perform similarly well in β_0 and β_2, resulting in approximately no EB in these regression coefficients. One exception, however, is the negative EB in β_0 from MI-CART-R in high $\delta_1^{X_2}, \delta_5^{X_2}$ scenarios. For β_1, we see that, for low $\delta_1^{X_2}, \delta_5^{X_2}$ values, MI-NORM-OFFSET returns the highest absolute EB. For $\delta_1^{X_2}, \delta_5^{X_2} \geq 1$, MI-CART shows the strongest

Table 4.2: Effect of deviating from MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) to MNAR2. Different imputation methods are compared in terms of the resulting empirical bias (EB) with values multiplied by $1,000$.

$\delta_4^{X_1}$, $\delta_4^{X_2}$	$\delta_5^{X_2}$, $\delta_1^{X_2}$	Method	EB - β_0	EB - β_1	EB - β_2
0, 0	0, 0	MI-CART	-4	-7	-2
0, 0	0, 0	MI-CART-R	-5	-3	-3
0, 0	0, 0	MI-NORM	-6	-1	4
0, 0	0, 0	MI-NORM-OFFSET	8	-20	-7
0, 0	1, 1	MI-CART	0	-29	6
0, 0	1, 1	MI-CART-R	-70	-11	17
0, 0	1, 1	MI-NORM	-6	-8	5
0, 0	1, 1	MI-NORM-OFFSET	9	-11	-8
0, 0	2, 2	MI-CART	1	-56	11
0, 0	2, 2	MI-CART-R	-151	-21	47
0, 0	2, 2	MI-NORM	-7	-15	8
0, 0	2, 2	MI-NORM-OFFSET	8	-6	-11
1, 1	0, 0	MI-CART	170	-7	-36
1, 1	0, 0	MI-CART-R	154	-2	-36
1, 1	0, 0	MI-NORM	165	-2	-32
1, 1	0, 0	MI-NORM-OFFSET	175	-14	-41
1, 1	1, 1	MI-CART	173	-14	-34
1, 1	1, 1	MI-CART-R	28	1	-11
1, 1	1, 1	MI-NORM	164	-3	-32
1, 1	1, 1	MI-NORM-OFFSET	176	2	-46
1, 1	2, 2	MI-CART	173	-27	-31
1, 1	2, 2	MI-CART-R	-66	12	11
1, 1	2, 2	MI-NORM	163	-4	-32
1, 1	2, 2	MI-NORM-OFFSET	175	7	-46
2, 2	0, 0	MI-CART	250	-1	-42
2, 2	0, 0	MI-CART-R	199	1	-38
2, 2	0, 0	MI-NORM	243	1	-39
2, 2	0, 0	MI-NORM-OFFSET	254	-3	-49
2, 2	1, 1	MI-CART	250	-1	-42
2, 2	1, 1	MI-CART-R	46	11	-22
2, 2	1, 1	MI-NORM	243	2	-39
2, 2	1, 1	MI-NORM-OFFSET	254	9	-53
2, 2	2, 2	MI-CART	249	-12	-40
2, 2	2, 2	MI-CART-R	-61	16	-6
2, 2	2, 2	MI-NORM	242	0	-40
2, 2	2, 2	MI-NORM-OFFSET	255	14	-55

absolute EB in β_1, all other methods result in some EB as well. In the MNAR cases ($\delta_4^{X_1} = \delta_4^{X_2} \neq 0$) and for increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values, EB in β_0 increases in all methods except for MI-CART-R, which shows the lowest EB.

For β_1 in the MNAR cases, all methods start with a negative EB and move towards no EB or a positive one with increasing $\delta_4^{X_1} = \delta_4^{X_2}$ values. MI-CART shows the strongest negative EB and the differences from the other methods decrease with increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values. For β_2 in the MNAR cases, all methods start with approximately no EB and result in negative EB with increasing $\delta_4^{X_1} = \delta_4^{X_2}$ values. One exception, however, is the EB in MI-CART-R, which starts with a higher EB that decreases with increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values. MI-CART-R performs on the same level as the other methods for $\delta_1^{X_2} = \delta_5^{X_2} = 0$, but shows stronger bias for low $\delta_4^{X_1}, \delta_4^{X_2}$ values and less bias for high $\delta_4^{X_1}, \delta_4^{X_2}$ values. In the MNAR cases, we find no clear best method over all β-coefficients in terms of EB.

Table 4.3 presents the results for RV. In the MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) condition, we find that, overall, the CART and CART-R return too low RV values, compared to both other methods. For MNAR, the CART and CART-R remain on a low RV level, while MI-NORM shows the best results (closest to 1,000) in β_1, and β_2. For β_0, MI-NORM overall returns RV values below 1,000, while MI-NORM-OFFSET shows values on or above the 1,000 mark. A change in the parameters $\delta_1^{X_2}$ and $\delta_5^{X_2}$ does not result in an obvious pattern in RV.

In the case of RMSE (Table 4.4) in β_0, for MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$), all methods perform equally well, except for MI-CART-R, which results in increasing RMSE for increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values. For MAR in β_1, we find MI-CART and MI-CART-R resulting in increasing RMSEs for increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values. For β_2, MI-CART shows increased RMSE values in scenarios with high $\delta_1^{X_2}, \delta_5^{X_2}$ values, all other methods remain mostly on the same level. For the MNAR cases ($\delta_4^{X_1} = \delta_4^{X_2} \neq 0$), MI-CART-R shows the lowest RMSE for β_0 (RMSE values increase equally for increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values in other methods). For β_1, MI-NORM and MI-NORM-OFFSET both result in the lowest RMSEs. While MI-CART's RMSE values decrease with increasing $\delta_1^{X_2}, \delta_5^{X_2}$ values, MI-CART-R remains on a high level. For β_2 and increasing $\delta_4^{X_1}$ and $\delta_4^{X_2}$ values, all methods increase in terms of RMSE, MI-NORM-OFFSET shows higher RMSE values compared to the other methods. In the MNAR cases, there is no best method over all β-coefficients in terms of RMSE.

Table 4.3: Effect of deviating from MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) to MNAR2. Different imputation methods are compared in terms of ratio of estimated variance to empirical variance (RV) with values multiplied by $1,000$.

$\delta_4^{X_1}$, $\delta_4^{X_2}$	$\delta_5^{X_2}$, $\delta_1^{X_2}$	Method	RV - β_0	RV - β_1	RV - β_2
0, 0	0, 0	MI-CART	667	743	654
0, 0	0, 0	MI-CART-R	751	693	640
0, 0	0, 0	MI-NORM	932	1029	893
0, 0	0, 0	MI-NORM-OFFSET	967	1297	962
0, 0	1, 1	MI-CART	697	814	636
0, 0	1, 1	MI-CART-R	670	536	659
0, 0	1, 1	MI-NORM	938	977	948
0, 0	1, 1	MI-NORM-OFFSET	1069	1267	1129
0, 0	2, 2	MI-CART	653	755	583
0, 0	2, 2	MI-CART-R	656	474	597
0, 0	2, 2	MI-NORM	975	967	918
0, 0	2, 2	MI-NORM-OFFSET	1042	1155	1078
1, 1	0, 0	MI-CART	677	662	708
1, 1	0, 0	MI-CART-R	698	637	767
1, 1	0, 0	MI-NORM	927	1031	1063
1, 1	0, 0	MI-NORM-OFFSET	992	1173	1170
1, 1	1, 1	MI-CART	671	630	687
1, 1	1, 1	MI-CART-R	618	584	655
1, 1	1, 1	MI-NORM	906	1006	1066
1, 1	1, 1	MI-NORM-OFFSET	1064	1173	1311
1, 1	2, 2	MI-CART	660	660	666
1, 1	2, 2	MI-CART-R	628	549	595
1, 1	2, 2	MI-NORM	963	950	1017
1, 1	2, 2	MI-NORM-OFFSET	1184	1219	1269
2, 2	0, 0	MI-CART	693	629	751
2, 2	0, 0	MI-CART-R	661	691	696
2, 2	0, 0	MI-NORM	868	1045	979
2, 2	0, 0	MI-NORM-OFFSET	998	1141	1101
2, 2	1, 1	MI-CART	711	687	710
2, 2	1, 1	MI-CART-R	571	604	614
2, 2	1, 1	MI-NORM	947	991	1074
2, 2	1, 1	MI-NORM-OFFSET	1051	1145	1210
2, 2	2, 2	MI-CART	698	680	710
2, 2	2, 2	MI-CART-R	655	550	661
2, 2	2, 2	MI-NORM	1008	992	987
2, 2	2, 2	MI-NORM-OFFSET	1016	1132	1263

Table 4.4: Effect of deviating from MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) to MNAR2. Different imputation methods are compared in terms of the resulting root mean squared error (RMSE) with values multiplied by $1,000$.

$\delta_4^{X_1}$, $\delta_4^{X_2}$	$\delta_5^{X_2}$, $\delta_1^{X_2}$	Method	RMSE - β_0	RMSE - β_1	RMSE - β_2
0, 0	0, 0	MI-CART	47	65	53
0, 0	0, 0	MI-CART-R	46	66	55
0, 0	0, 0	MI-NORM	45	57	49
0, 0	0, 0	MI-NORM-OFFSET	46	57	49
0, 0	1, 1	MI-CART	45	68	53
0, 0	1, 1	MI-CART-R	83	71	55
0, 0	1, 1	MI-NORM	46	55	48
0, 0	1, 1	MI-NORM-OFFSET	46	50	47
0, 0	2, 2	MI-CART	46	91	57
0, 0	2, 2	MI-CART-R	158	93	73
0, 0	2, 2	MI-NORM	44	56	49
0, 0	2, 2	MI-NORM-OFFSET	47	52	51
1, 1	0, 0	MI-CART	176	61	60
1, 1	0, 0	MI-CART-R	161	62	60
1, 1	0, 0	MI-NORM	171	53	54
1, 1	0, 0	MI-NORM-OFFSET	181	52	59
1, 1	1, 1	MI-CART	179	65	60
1, 1	1, 1	MI-CART-R	55	70	53
1, 1	1, 1	MI-NORM	169	52	54
1, 1	1, 1	MI-NORM-OFFSET	182	49	62
1, 1	2, 2	MI-CART	179	74	59
1, 1	2, 2	MI-CART-R	82	84	58
1, 1	2, 2	MI-NORM	169	52	54
1, 1	2, 2	MI-NORM-OFFSET	180	49	63
2, 2	0, 0	MI-CART	254	55	61
2, 2	0, 0	MI-CART-R	204	58	62
2, 2	0, 0	MI-NORM	247	48	58
2, 2	0, 0	MI-NORM-OFFSET	257	47	64
2, 2	1, 1	MI-CART	254	53	62
2, 2	1, 1	MI-CART-R	69	70	59
2, 2	1, 1	MI-NORM	247	49	58
2, 2	1, 1	MI-NORM-OFFSET	258	47	67
2, 2	2, 2	MI-CART	253	57	60
2, 2	2, 2	MI-CART-R	78	85	55
2, 2	2, 2	MI-NORM	246	49	59
2, 2	2, 2	MI-NORM-OFFSET	258	48	69

Table 4.5: Effect of deviating from MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) to MNAR2. Different imputation methods are compared in terms of the resulting confidence interval coverage rate (CICR) with values multiplied by 100.

$\delta_4^{X_1}$, $\delta_4^{X_2}$	$\delta_5^{X_2}$, $\delta_1^{X_2}$	Method	CICR - β_0	CICR - β_1	CICR - β_2
0, 0	0, 0	MI-CART	90	91	90
0, 0	0, 0	MI-CART-R	91	91	89
0, 0	0, 0	MI-NORM	96	97	94
0, 0	0, 0	MI-NORM-OFFSET	94	97	94
0, 0	1, 1	MI-CART	90	91	86
0, 0	1, 1	MI-CART-R	52	84	85
0, 0	1, 1	MI-NORM	94	94	96
0, 0	1, 1	MI-NORM-OFFSET	93	97	98
0, 0	2, 2	MI-CART	90	83	86
0, 0	2, 2	MI-CART-R	6	79	70
0, 0	2, 2	MI-NORM	94	94	95
0, 0	2, 2	MI-NORM-OFFSET	96	96	96
1, 1	0, 0	MI-CART	0	90	84
1, 1	0, 0	MI-CART-R	5	92	85
1, 1	0, 0	MI-NORM	4	96	91
1, 1	0, 0	MI-NORM-OFFSET	2	96	86
1, 1	1, 1	MI-CART	0	90	83
1, 1	1, 1	MI-CART-R	82	86	90
1, 1	1, 1	MI-NORM	6	96	91
1, 1	1, 1	MI-NORM-OFFSET	3	96	87
1, 1	2, 2	MI-CART	1	85	84
1, 1	2, 2	MI-CART-R	58	86	86
1, 1	2, 2	MI-NORM	4	94	90
1, 1	2, 2	MI-NORM-OFFSET	4	98	90
2, 2	0, 0	MI-CART	0	88	80
2, 2	0, 0	MI-CART-R	0	90	78
2, 2	0, 0	MI-NORM	0	96	86
2, 2	0, 0	MI-NORM-OFFSET	0	96	82
2, 2	1, 1	MI-CART	0	90	80
2, 2	1, 1	MI-CART-R	71	86	86
2, 2	1, 1	MI-NORM	0	94	88
2, 2	1, 1	MI-NORM-OFFSET	0	97	84
2, 2	2, 2	MI-CART	0	88	80
2, 2	2, 2	MI-CART-R	63	82	90
2, 2	2, 2	MI-NORM	0	95	86
2, 2	2, 2	MI-NORM-OFFSET	0	95	84

In terms of the CI coverage rate (Table 4.5), for MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$), all methods perform equally well at approximately 95% coverage for all regression coefficients. One exception is the reduced coverage rate in β_0 for MI-CART-R for $\delta_1^{X_2}, \delta_5^{X_2} \geq 1$. When there are deviations from MAR ($\delta_4^{X_1} = \delta_4^{X_2} \geq 0.5$ scenarios), coverage rates for β_0 decrease in a similar way and approach 0 in all investigated methods. Only MI-CART-R shows values notably above zero for high $\delta_1^{X_2}, \delta_5^{X_2}$ values. For β_1, MI-NORM and MI-NORM-OFFSET remain mostly at the 95% level; MI-CART and MI-CART-R show reduced coverage (between 75% and 90%) for all $\delta_1^{X_2}, \delta_5^{X_2}$ values. For β_2, the MNAR scenarios lead to reduced coverage rates (between 75% and 95%) in all methods. In the MNAR cases, no method stands out over all β-coefficients in terms of CI coverage rate.

In summary, in MNAR1, the MI-NORM-OFFSET procedure performs best overall, followed by MI-CART-R. MI-NORM generally performs better than MI-CART. In the MNAR2 scenarios, we see that including RIs in the SI process can improve the results for some regression parameters, and at the same time achieve similar results as standard procedures in the MAR case. However, in MNAR2 we do not find a clear best imputation method.

4.3.5.2 Objective 2

The figures in this sub-section present box plots for each method investigated. The SEV (Equation 4.18) is shown on the y-axis.

4.3.5.2.1 MNAR1 In Figure 4.1 the x-axis displays the $\delta_1^{X_2}$ and $\delta_2^{X_1}$ values. Under MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$), MI-CART and MI-CART-R perform worst, followed by MI-NORM and MI-NORM-OFFSET. BART and BART-R perform best here. For increasing $\delta_1^{X_2}$ and $\delta_2^{X_1}$, BART-R results in the lowest SEV, followed by MI-CART-R, and then MI-CART, MI-NORM, and MI-NORM-OFFSET, which perform similarly well. BART clearly performs worst in the MNAR cases.

4.3.5.2.2 MNAR2 For MNAR2, the different columns in Figure 4.2 show results for different $\delta_1^{X_2}, \delta_5^{X_2}$ values; the x-axes correspond to the values of $\delta_4^{X_1}$ and $\delta_4^{X_2}$, but only display $0, 1, 2$ for readability. Under MAR ($\delta_4^{X_1} = \delta_4^{X_2} = 0$) and $\delta_1^{X_2} = \delta_5^{X_2} = 0$, we find BART-R performing best, and MI-CART and MI-CART-R performing worst. For

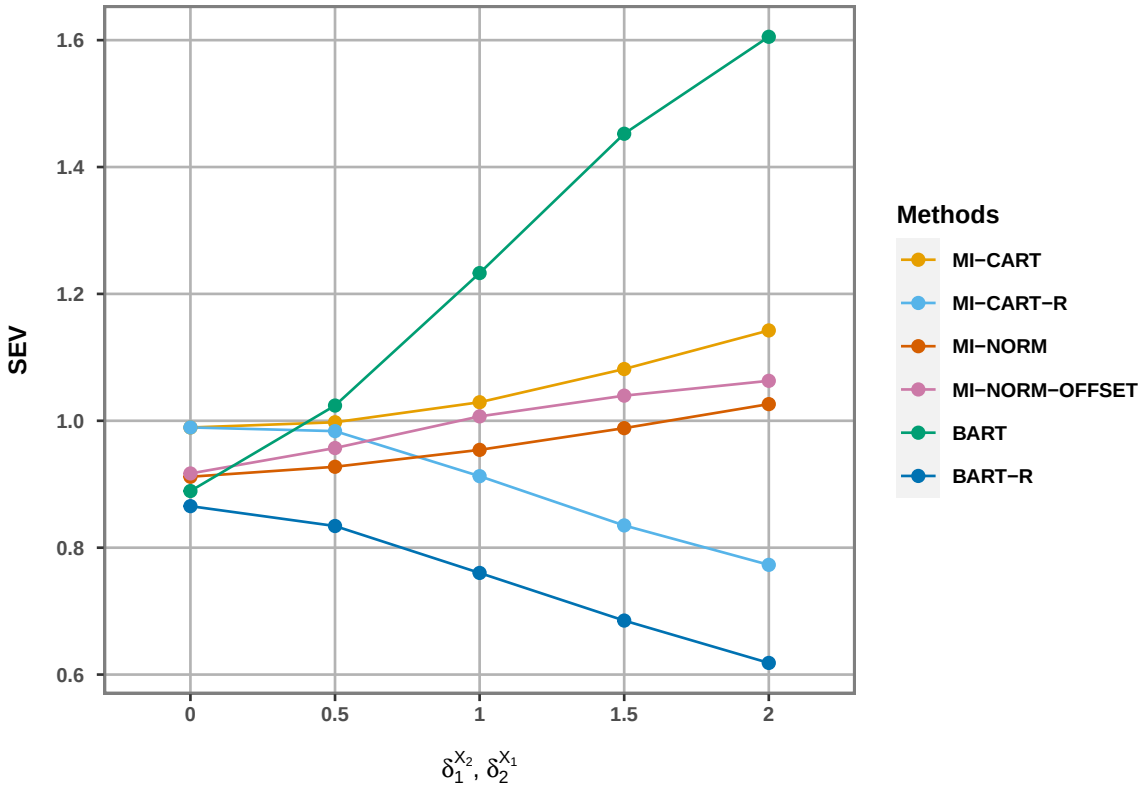

Figure 4.1: Effect of deviating from MAR ($\delta_1^{X_2} = \delta_2^{X_1} = 0$) to MNAR1. Different imputation methods are compared in terms of the resulting squared error of predicted values (SEV). The solid black line indicates zero SEV.

$\delta_1^{X_2} = \delta_5^{X_2} = 0$ and increasing $\delta_4^{X_1}$ and $\delta_4^{X_2}$, BART-R shows the lowest and BART shows the highest SEV. For $\delta_4^{X_1} = \delta_4^{X_2} = 0$, when $\delta_1^{X_2}, \delta_5^{X_2}$ values increase, BART-R results in the lowest SEV, followed by MI-CART-R, and then MI-NORM, MI-NORM-OFFSET, and MI-CART, all of which perform similarly well. BART performs worst here. This pattern remains the same with greater differences among the methods for increasing $\delta_4^{X_1}$, $\delta_4^{X_2}$, and $\delta_1^{X_2}, \delta_5^{X_2}$ values.

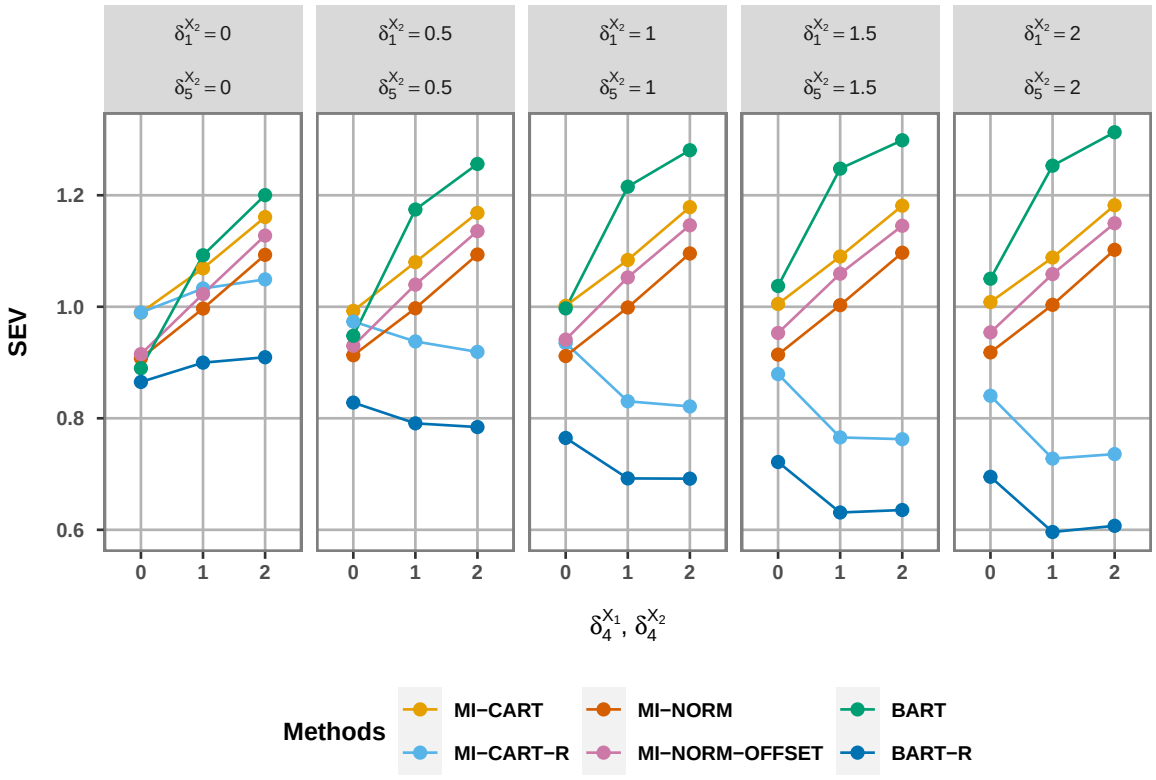

Figure 4.2: Effect of deviating from MAR ($\delta_4^{X_2} = \delta_4^{X_1} = 0$) to MNAR2. Different imputation methods are compared in terms of the resulting squared error of predicted values (SEV). The solid black line indicates zero SEV.

The findings for Objective 2 imply that, first, under MAR, the inclusion of RIs does not lead to decreased predictive performance. Second, under MAR, averaging the values of multiply imputed data can result in a performance that is similar to when the outcome is predicted directly. Third, in the simulated MNAR cases, RIs can improve predictive power. However, performance differs among the different approaches incorporating RI information. Specifically, including RIs as additional covariates in MI-CART-R results in lower SEVs than MI-NORM-OFFSET. Finally, including RIs directly in a prediction

method (as in BART-R) leads to lower SEVs than including RIs in an MI process (as in MI-CART-R).

4.4 Discussion

For analytic inference, including RIs in the model can help to improve parameter estimates in specific MNAR situations. For parametric models, including RIs as covariates in the model is generally not recommended (Donders et al., 2006; Greenland & Finkle, 1995; Jones, 1996; Knol et al., 2010; Vach & Blettner, 1991). Rather, models should be fit to multiply imputed data. We can further learn that studies comparing methods can produce misleading results when analysis goals are not stated clearly, e.g., when draws from MI procedures are assessed only in terms of their predictive power, as described in Section 4.2.4.

The simulation aiming at analytic inference (Objective 1) shows that under MAR, all MI procedures perform mostly equally well in terms of quantitative properties and have similar SEV values. For MNAR cases, MI-NORM-OFFSET performs best in terms of the evaluated quantitative properties when the MNAR mechanism follows a selection model (MNAR1), but no clear best method can be found in the MNAR scenario allowing for a correlation between RIs (MNAR2). The slightly inferior performance of MI-CART-R in MNAR1 can be explained by the properties of the compared methods. While incorporating RIs into MI-CART-R as additional covariates is a simple way to access the potential information in RIs, MI-NORM-OFFSET might better approximate the underlying selection model of the data. Further, the underlying data generating process is parametric, which likely favors parametric imputation models.

The simulation addressing descriptive inference shows that, in the MAR case, the performance of the methods for direct prediction (BART and BART-R) is slightly better than that of the MI methods. For MNAR scenarios, we see a clear advantage of the methods that include RIs directly (BART-R, MI-CART-R). These findings suggest that MI procedures can also perform well in prediction tasks if the mean of multiply imputed values is used as the predicted value. However, fitting a model including RIs on unimputed data seems to be the best choice for descriptive inference.

In the case of descriptive inference, the behavior of MI-NORM-OFFSET is similar to that of methods that do not include RIs, but MI-CART-R remains on the same

level of performance in the MNAR situations compared to the MAR situation. These performance differences likely occur because MI-NORM-OFFSET follows a selection model factorization, while MI-CART-R is based on a PMM. A direct comparison of the selection model representative, MI-NORM-OFFSET, and the PMM representative, MI-CART-R, suggests the following: MI-NORM-OFFSET performs best in terms of quantitative properties when the underlying data are generated from a selection model; for data allowing for correlated RIs, there is no best performing model overall. When assessing prediction accuracy, MI-CART-R clearly performs best among the imputation methods in both MNAR scenarios.

Table 4.6 provides guidance for practice when RIs are considered for inclusion in statistical models. The table summarizes the conclusions drawn from the literature review and the simulation study. We distinguish between two different analysis goals, analytic and descriptive inference (see Objectives 1 and 2 in Sections 4.3.2 and 4.3.3), and between two different missingness mechanisms, MAR and MNAR.

		Analysis Goal	
		Analytic Inference	Descriptive Inference
Missingness Mechanism	MAR	RIs not useful, but ... (1)	RIs not useful, but ... (2)
	MNAR	RIs potentially useful (3)	RIs potentially useful (4)

Table 4.6: Summary: usefulness of response indicators (RIs) in statistical modeling by analysis goals and missingness mechanisms.

(1) For analytic inference under MAR, RIs do not provide information and thus are unnecessary in imputation models. Methods including RIs can show decreased performance in MAR scenarios, as seen in the MNAR1 scenario, where MI-NORM-OFFSET leads to increased EB in the intercept parameter (cf. Table 4.1)

(2) For descriptive inference under MAR, RIs do not provide information and thus are unnecessary in imputation models. However, the additional training data (observations with incomplete covariates rather than complete cases only) can lead to better performance in terms of prediction accuracy. In the simulation for Objective 2 we find including RIs under MAR does not lead to worse performance.

(3) For analytic inference under MNAR, the simulation suggests that RIs can help to improve MI and the estimates of subsequent analyses. However, the simulated

scenarios are narrow and their findings might not generalize to other missing data patterns and other MNAR mechanisms. There are likely MNAR mechanisms that lead to worse performance when RIs are included in the MI process, compared to MI under MAR.

(4) For descriptive inference under MNAR, the simulation suggests that RIs can help to improve performance. However, how RIs are used in models is key. In the investigated scenarios, an imputation method based on the PMM is preferred, as well as fitting predictive models without a previous imputation of missing values. Again, these suggestions are based on a limited number of simulation scenarios and might not hold in other data situations.

In practice, given a data set with missing values, the decision of whether to include RIs in a statistical model can be based on the following reasoning. First, the analysis goal needs to be clear, i.e., whether the goal is analytic or descriptive inference. When performing descriptive inference, RIs can be included, and methods with and without RIs can also be compared in order to investigate whether including RIs increases performance. For example, tree-based methods like BART and GUIDE provide specific features to incorporate RIs. For a low number of response patterns, the pattern submodel provides a more general approach to adding RIs in models and can be combined with many different procedures; a completely different model class for each response pattern is possible.

When the analysis goal is analytic inference, the more plausible missingness mechanism can be decided based on the number of variables and the strength of their associations. The higher the number of variables and the stronger their relationships, the more plausible an MAR mechanism becomes. Although the missingness mechanism in many applications is likely MNAR (e.g. nonresponse in sample surveys), what matters is "how far away" from MAR it is, i.e., how the results change when assuming MAR as opposed to MNAR. Comparing analysis results after performing MI without RIs (under MAR) vs. with RIs (assuming a specific MNAR mechanism) can provide information about how sensitive the results are to the MAR assumption.

While distinguishing MAR and MNAR solely based on a given data set is not possible, there are two views based on the analysis goals. On the one hand, Loh et al. (2020), who mostly focus on descriptive inference, state that "(...) it is hard to justify that the hundreds of covariates are missing at random (MAR)". Rubin, on the other hand,

argues that MI under MAR is plausible in the presence of useful covariates, and that assuming MAR is a good starting point for the analysis (Rubin et al., 1995). This approach is similar to other statistical analyses focusing on analytic inference that start by assuming, for example, a normal distribution and then evaluate the impact of potential deviation from this assumption. These two views are described in Breiman's famous article about the two cultures of statistical modeling (Breiman, 2001). The article describes the use of algorithmic and more flexible methods vs. classic statistical modeling (i.e. parametric models). While Breiman's article favors algorithmic methods, we argue that a distinction between analytic and descriptive inference is crucial and that the modeling approach should be selected based on the analysis goal.

The research presented here is limited in several ways. First, the literature review is not based on a systematic literature search, but rather on an unstructured search. Although we think that we have included the most important papers, it is still possible that some relevant literature has been overlooked. While we covered important parts of statistical modeling, we did not investigate the full range of statistical methods like unsupervised learning / clustering or causal inference. Second, the presented simulation investigates only one MAR scenario based on continuous variables with relatively low associations and only two out of many possible deviations from this MAR scenario. As stated earlier, there are likely MNAR scenarios where procedures under MAR perform better than procedures including RIs. Further, the present study presents only a limited number of methods mentioned in the preceding literature review. We focus on those methods because they are examples of parametric and non-parametric methods available in standard software and thus are likely used in practice. However, multiple additional methods, e.g., SI via BART, BART-R, and GUIDE, are not investigated here, suggesting avenues for future work.

While here we focus on a literature review and a simulation, future theoretical work can further clarify the underlying assumptions of models including RIs. One possibility is to focus on two categorical incomplete variables, because this setup requires only minimal distributional assumptions on the variables. The assumptions of an MAR model can be compared with two different MNAR models: conditioning on one or both RIs. Further, connections with other MNAR models for incomplete categorical variables (Little, 1985; Little & Rubin, 2019, Chapter 13) can be explored.

RIs can be classified into a nominal variable if additional information about the missing values is available. Unpublished work by Kamphuis et al. (2015) extends the penalized

spline of propensity prediction method (G. Zhang & Little, 2009) by utilizing codes for missing data like "don't know" or "refusal" (or other indicators like for "top-coded" values). These codes, often available in survey data (e.g. in the U.S. Consumer Expenditure Survey), could also be incorporated as covariates in MI. As mentioned by Loh et al. (2020), GUIDE now also allows for splits between RIs with more than two categories.

Another situation where the RI approach can be useful is when filter questions in survey questionnaires are used to give a subgroup of respondents (henceforth G1) additional questions. This common practice leads to "not applicable" (NA) values for the subgroup that does not receive the additional questions (henceforth G2). NA values are not missing values because true values do not exist. In a regression context, the interactions between the filter question and additional questions are comparable to using RIs with interactions, but more plausible, because a slope can be estimated for G1, while the G2 estimate for the variable is constant.

4.5 Appendix - Design Table

Table 4.7: Design table for Chapter 4.

Method	Parameter	Description	Levels	Choices	Tuning
CART, CART-R	minbucket	The minimum number of observations in any terminal node used	5	Default of MICE package version 3.14.7	None
CART, CART-R	cp	Complexity parameter	1e-04	Default of MICE package version 3.14.7	None
NORM	-	No tuning parameters	-	-	None
NORM-OFFSET	-	No tuning parameters	-	-	None
BART, BART-R	mem_cache_for_speed	Speed enhancement that caches the predictors and the split values that are available at each node for selecting new rules.	FALSE	Recommened for large number of predictors, not the case here	None
BART, BART-R	use_missing_data	If TRUE, additional RIs are included.	TRUE	Difference between BART and BART-R achieved via different data input.	None

Any other parameters of imputation via CART and CART-R are as specified in function rpart::rpart.control(), package version 4.1.16. All other parameters of BART or BART-R are set as specified in function bartMachine::bartMachine(), package version 1.2.7

Chapter 5

Conclusion

This dissertation aims to produce research in the domain of method selection for multiple imputation (MI) of missing data. Specifically, the three studies provide a comparison of methods and guidance for practitioners (Study One), present a framework for automated method selection within sequential imputation (SI) (Study Two), and investigate the use of response indicators (RIs) in imputation models (Study Three). In this chapter, we first restate the main findings of the three studies, followed by the limitations and directions for future research.

Study One (Chapter 2) reveals how different imputation procedures perform under different data scenarios. We find that, in general, all parametric procedures (Bayesian linear models and regularized Bayesian linear models) perform similarly. While Bayesian additive regression trees (BART) performs generally well in the non-parametric case, overall it performs poorly in the parametric part of the simulation. For practical applications, we find random forest and classification and regression trees work best in most of the investigated scenarios. Further, in the parametric cases, complete case analysis (CC) performs well when missingness only depends on covariates, and shows reduced performance in MAR scenarios where missingness depends on both the covariates and the outcome. Both of these findings can be explained following Little (1992). In MNAR scenarios where missingness depends on covariates, CC results in empirically unbiased estimates, as explained in Little and Zhang (2011). However, CC generally has lower performance in terms of RMSE, compared to other methods, because all incomplete observations are excluded from the analysis.

Study Two (Chapter 3) proposed a modified SI procedure, called SI with integrated

method selection (SIIMS). SIIMS allows for multiple competing methods and automated plausibility checks during the imputation process. The plausibility checks in SIIMS are carried out via two criteria. One criterion is an MSE-like measure including a variance and a bias component; it is computed at the level of observations and then averaged over all of the observations that the imputation model is fit on. The assessment here focuses on the predictive power of the imputation model. The other criterion automatically assesses the plausibility of imputed values by comparing their density to the density of the observed values of the incomplete variable to be imputed, conditional on the response propensity score. The higher the similarity between those two densities, the better the imputation model under MAR. This criterion is applied to compare the imputation models in terms of marginal distributions. The study developed both criteria for continuous, binary, and nominal variables with missing values. The developed criteria are broad in the sense that they can be used to compare compare any methods that can predict values of an outcome. While we illustrate the proposed SIIMS framework using some current state of the art imputation methods, the proposed criteria can still assess new, potential better, methods. The presented case study suggest that SIIMS' performance is among the performance of the component methods applied separately.

Study Three (Chapter 4) reviews the literature on the use of RIs in statistical models and presents a simulation study of missing data imputation under MNAR, specifically when RIs are included in the imputation models. For descriptive inference, when the focus is on predicting a missing value as accurately as possible, including RIs in the model can improve predictive power when response patterns provide information. For non-informative response patterns, the algorithm ignores the RIs. For analytic inference, i.e., multiple imputation assessed on the basis of the quantitative properties of a regression model fit on multiply imputed data, including RIs in the model can help to improve parameter estimates in specific MNAR situations. For parametric models fit on incomplete data, including RIs in the model is generally not recommended. Further, misleading results can occur when analysis goals are not stated clearly, e.g., when draws from MI procedures are assessed only in terms of their predictive power.

The simulation investigates MNAR scenarios where RIs can help to improve MI procedures. The simulation focused on analytic inference shows that under MAR, all MI procedures perform equally well. For MNAR cases, the methods incorporating RIs of covariates perform better in terms of quantitative properties. The simulation focusing

on descriptive inference shows that in the MAR case the performance of the method doing prediction without previous imputation is slightly better than that of the MI methods. For MNAR scenarios, we see a clear advantage of the methods that directly include RIs. These findings suggest that MI procedures can also perform well in prediction tasks, if the mean of multiply imputed values is used as the predicted value. However, fitting a model including RIs on unimputed data seems to be the best choice for descriptive inference.

Although Study One investigates a high number of imputation methods and many scenarios, the study is limited in several aspects. First, the simulation contains only three continuous variables, and analysis models with only three variables are rare. Future research can build on this current study and investigate scenarios with a higher number of variables that are binary and nominal. Second, this study did not consider complex sampling feature like weights and clusters. Work by Zhou, Elliott, and Raghunathan (e.g., Zhou et al., 2016b, 2016a) provides a two-step approach for incorporating these complex sampling features of survey data in multiple imputation using finite population Bayesian bootstrap. Future studies could evaluate how performance is affected by substituting the parametric model used in the second step with the procedures compared in this study.

In Study Two, the criteria developed compare the imputation methods in terms of marginal distributions of the variables. However, the case study presents an evaluation focusing on joint distributions. Future research can explore the role of BART on the SIIMS process to examine if this method introduces extreme imputed values that are carried forward in the imputation process. Further, several incomplete variables are removed in the SIIMS procedure due to extreme distributions, which cause problems when computing the assessment criteria (e.g., a low number of missing observations can lead to an unstable kernel density estimation). Further modifications can incorporate these variables into the imputation process, potentially leading to increased performance. Related to Study Three, those variables could be excluded from SI, but serve as incomplete covariates together with their RIs in the imputation models.

Another major aspect of future SIIMS research should focus on runtime. SIIMS is relatively slow compared to other methods, because SI is an iterative process, multiple potential methods need to be fit and evaluated in each step, and the potential applications are high-dimensional data situations with many incomplete variables. An increase in speed could be achieved, first, by aggregating the complete variables via

principal component analysis, and using only a low number of principal components as covariates in SI. Another option is to provide only a subset of covariates for each incomplete variable based on correlations. In fact, removing slow methods from the process might be the simplest way to increase speed. Compared to the other procedures, Bayesian linear models often result in runtime outliers. Removing or replacing this method with a faster procedure would certainly speed up the whole process.

The presented case study serves only as an example of an application for SIIMS. Future studies can further evaluate both criteria separately, along with performance on only categorical, binary, and continuous incomplete variables. While the evaluation process does not require a specified data generating model, there are several possible expansions. First, pooling several waves of NHANES data would further increase the synthetic population from which samples are drawn. Second, a different set of VOI, given by an independent third person, and different data sets can be used to receive a more comprehensive picture of performance differences.

Study Three is also limited in several ways. First, the literature review is not based on a systematic literature search; thus, it is still possible that some relevant literature has been overlooked, although most important papers are included. While the literature review covers important parts of statistical modeling, excluded are topics like unsupervised learning / clustering. Second, the presented simulation investigates only one MAR scenario and only two possible deviations. For example, not investigated are MNAR scenarios where procedures under MAR perform better than procedures including RIs. Further, the present study applies only a limited number of methods. For instance, SI via BART or GUIDE are not investigated here, which suggests avenues for future work.

Future studies can further clarify the underlying assumptions of models including RIs. In two categorical incomplete variables only minimal distributional assumptions on the variables are necessary. The assumptions of a MAR model can be compared with two different MNAR models: conditioning on one or both RIs. Further, connections with other MNAR models for incomplete categorical variables (Little, 1985; Little & Rubin, 2019, Chapter 13) can be further explored.

A simple extension of Study Three would be to further sub-classify RIs into multinomial variables utilizing codes for missing data like “don’t know” or “refusal” (or other indicators like for “top-coded” values). These codes can be incorporated and thus

utilized in MI in order to further enhance imputation of missing values.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459.

Akande, O., Li, F., & Reiter, J. (2017). An empirical comparison of multiple imputation methods for categorical data. *The American Statistician*, *71*(2), 162–170.

Andridge, R. R., & Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64.

Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences*, *233*, 25–35.

Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2022). Missing data and other measurement quality issues in mobile geolocation sensor data. *Social Science Computer Review*, *40*(1), 212–235.

Bartlett, J. W., Seaman, S. R., White, I. R., Carpenter, J. R., & Alzheimer's Disease Neuroimaging Initiative*. (2015). Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model. *Statistical Methods in Medical Research*, *24*(4), 462–487.

Beesley, L. J., Bondarenko, I., Elliot, M. R., Kurian, A. W., Katz, S. J., & Taylor, J. M. (2021). Multiple imputation with missing data indicators. *Statistical Methods in Medical Research*, *30*(12), 2685–2700.

Bondarenko, I., & Raghunathan, T. (2016). Graphical and numerical diagnostic tools to assess suitability of multiple imputations and imputation models. *Statistics in Medicine*, *35*(17), 3007–3020.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Burgette, L. F., & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, *172*(9), 1070–1076.

Callegaro, M., & Yang, Y. (2018). The role of surveys in the era of "big data." *The Palgrave Handbook of Survey Research*, 175–192.

Carpenito, T., & Manjourides, J. (2022). MISL: Multiple imputation by super learning. *Statistical Methods in Medical Research*, *31*(10), 1904–1915.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chipman, H. A., George, E. I., & McCulloch, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association*, *93*(443), 935–948.

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

D'Orazio, M. (2019). *StatMatch: Statistical matching or data fusion*. https://CRAN.R-project.org/package=StatMatch

Dagdoug, M., Goga, C., & Haziza, D. (2023). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. *Journal of Survey Statistics and Methodology*, *11*(1), 141–188.

Deng, Y., Chang, C., Ido, M. S., & Long, Q. (2016). Multiple imputation for general missing data patterns in the presence of high-dimensional data. *Scientific Reports*, *6*, 21689.

Denison, D. G., Mallick, B. K., & Smith, A. F. (1998). A Bayesian CART algorithm. *Biometrika*, *85*(2), 363–377.

Ding, Y., & Simonoff, J. S. (2010). An investigation of missing data methods for classification trees applied to binary response data. *Journal of Machine Learning Research*, *11*(Jan), 131–170.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, *59*(10), 1087–1091.

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

Ezzati-Rice, T. M., Johnson, W., Khare, M., Little, R., Rubin, D. B., & Schafer, J. L. (1995). A simulation study to evaluate the performance of model-based multi-

ple imputations in NCHS health examination surveys. *Proceedings of the Annual Research Conference*, 257–266.

Finkbeiner, C. (1979). Estimation for the multiple factor model when data are missing. *Psychometrika*, *44*(4), 409–420.

Fletcher Mercaldo, S., & Blume, J. D. (2020). Missing data and prediction: The pattern submodel. *Biostatistics*, *21*(2), 236–252.

Freedman, D. S., Thornton, A., & Camburn, D. (1980). Maintaining response rates in longitudinal studies. *Sociological Methods & Research*, *9*(1), 87–98.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Giusti, C., & Little, R. J. A. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics*, *27*, 211–229.

Gleason, T. C., & Staelin, R. (1975). A proposal for handling missing data. *Psychometrika*, *40*(2), 229–252.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A monte carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling*, *7*(3), 319–355.

Greenland, S., & Finkle, W. D. (1995). A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264. https://doi.org/10.1093/oxfordjournals.aje.a117592

Groves, R. M. (2004). *Survey errors and survey costs*. John Wiley & Sons.

Gu, T., Taylor, J. M., Cheng, W., & Mukherjee, B. (2019). Synthetic data method to incorporate external information into a current study. *Canadian Journal of Statistics*, *47*(4), 580–603.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer.

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

Jinn, J.-H., & Sedransk, J. (1989). Effect on secondary data analysis of common

imputation methods. *Sociological Methodology*, *19*, 213–241.

Jolani, S. et al. (2012). *Dual imputation strategies for analyzing incomplete data*. Utrecht University.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, *91*(433), 222–230.

Kalton, G. (1986). Handling wave nonresponse in panel surveys. *Journal of Official Statistics*, *2*(3), 303–314.

Kamphuis, R., Jolani, S., Lugtig, P., & Lugtig, P. (2015). *The blocked imputation approach for missing data* [Master's thesis].

Kapelner, A., & Bleich, J. (2015). Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics*, *43*(2), 224–239.

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine learning with Bayesian additive regression trees. *Journal of Statistical Software*, *70*(4), 1–40. https://doi.org/10.18637/jss.v070.i04

Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, *1*.

Knol, M. J., Janssen, K. J., Donders, A. R. T., Egberts, A. C., Heerdink, E. R., Grobbee, D. E., Moons, K. G., & Geerlings, M. I. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: An empirical example. *Journal of Clinical Epidemiology*, *63*(7), 728–736.

Kolenikov, S., Angeles, G., et al. (2004). The use of discrete data in PCA: Theory, simulations, and applications to socioeconomic indices. *Chapel Hill: Carolina Population Center, University of North Carolina*, *20*, 1–59.

Laqueur, H. S., Shev, A. B., & Kagawa, R. M. C. (2021). SuperMICE: An Ensemble Machine Learning Approach to Multiple Imputation by Chained Equations. *American Journal of Epidemiology*, *191*(3), 516–525. https://doi.org/10.1093/aje/kwab271

Leacy, F. (2016). *Multiple imputation under missing not at random assumptions via fully conditional specification* [PhD thesis]. University of Cambridge, Cambridge, UK: MRC Biostatistical unit.

Leacy, F. P., Floyd, S., Yates, T. A., & White, I. R. (2017). Analyses of sensitivity to the missing-at-random assumption using multiple imputation with delta ad-

justment: Application to a tuberculosis/HIV prevalence survey with incomplete HIV-status data. *American Journal of Epidemiology*, *185*(4), 304–315.

Lessler, J., & Kalsbeek, W. (1992). *Nonresponse errors in surveys*. John Wiley & Sons.

Li, K. (1988). Imputation using markov chains. *Journal of Statistical Computation and Simulation*, *30*(1), 57–79.

Liang, F., Jia, B., Xue, J., Li, Q., & Luo, Y. (2018). An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *80*(5), 899–926.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22. https://CRAN.R-project.org/doc/Rnews/

Linero, A. R. (2022). SoftBart: Soft Bayesian additive regression trees. *arXiv Preprint arXiv:2210.16375*.

Linero, A., & Yang, Y. (2018). Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *80*(5), 1087–1110.

Little, R. (1985). Nonresponse adjustments in longitudinal surveys: Models for categorical data. *Bulletin of International Statistical Institute*, *15*, 1–15.

Little, R. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Little, R. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, *88*(421), 125–134.

Little, R. (2020). On algorithmic and modeling approaches to imputation in large data sets. *Statistica Sinica*, *30*(4), 1685–1696.

Little, R., & Rubin, D. B. (2002). *Statistical analysis with missing data*. John Wiley & Sons.

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

Little, R., & Zhang, N. (2011). Subsample ignorable likelihood for regression analysis with missing data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *60*(4), 591–605.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, *12*, 361–386.

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*(3), 329–348.

Loh, W.-Y., Eltinge, J., Cho, M. J., & Li, Y. (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, *29*(1), 431–453.

Loh, W.-Y., Man, M., & Wang, S. (2019). Subgroups from regression trees with adjustment for prognostic effects and postselection inference. *Statistics in Medicine*, *38*(4), 545–557.

Loh, W.-Y., Zhang, Q., Zhang, W., & Zhou, P. (2020). Missing data, imputation and regression trees. *Statistica Sinica*, *30*(4), 1697–1722.

Loh, W.-Y., Zheng, W., et al. (2013). Regression trees for longitudinal and multiresponse data. *The Annals of Applied Statistics*, *7*(1), 495–522.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *9*(4), 538–558.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, *13*(6), 47–60.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, *14*(1), 75.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9.

Nordbotten, S. (1996). Neural network imputation applied to the Norwegian 1990 population census data. *Journal of Official Statistics-Stockholm-*, *12*, 385–402.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Paulin, F., Geoffrey, & Reyes-Morales, S. (2006). Multiple imputation manual: Supplement to 2004 consumer expenditure interview survey public use microdata documentation. In *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. US Department of Labor, Bureau of Labor Statistics, Division of Consumer Expenditure Surveys.

Peytchev, A. (2009). Survey breakoff. *Public Opinion Quarterly*, *73*(1), 74–97.

Qian, N., & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, *202*(4), 865–884.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. *Proceedings of the Tenth International Conference on Machine Learning*, 236–243.

Quinlan, J. R. (2014). *C4. 5: Programs for machine learning*. Elsevier.

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–96.

Raghunathan, T., Solenberger, P., Berglund, P., & Hoewyk, J. (2016). IVEware: Imputation and variance estimation software (version 0.3). *URL Http://Www. Src. Isr. Umich. Edu/Wp-Content/Uploads/IVEware-Version-0.3-User-Guide-Linked.pdf. Survey Methodology Program, Institute for Social Research-University of Michigan*.

Raymond, M. R., & Roberts, D. M. (1987). A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, *47*(1), 13–26.

Razzak, H., & Heumann, C. (2019). Hybrid multiple imputation in a large scale complex survey. *Statistics in Transition New Series*, *20*(4), 33–58.

Rezvan, P. H., Lee, K. J., & Simpson, J. A. (2018). Sensitivity analysis within multiple imputation framework using delta-adjustment: Application to longitudinal study of australian children. *Longitudinal and Life Course Studies*, *9*(3), 259–278.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, *1*, 20–34.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*(359), 538–543.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B., & Schafer, J. L. (1990). Efficiently creating multiple imputations for incomplete multivariate normal data. *Proceedings of the Statistical Computing Section of the American Statistical Association*, *83*, 88.

Rubin, D. B., Stern, H. S., & Vehovar, V. (1995). Handling "don't know" survey responses: The case of the Slovenian plebiscite. *Journal of the American Statistical Association*, *90*(431), 822–828.

Sakshaug, J. W., Couper, M. P., & Ofstedal, M. B. (2010). Characteristics of physical measurement consent in a population-based survey of older adults. *Medical Care*, *48*(1), 64.

Sakshaug, J. W., & Kreuter, F. (2012). Assessing the magnitude of non-consent biases in linked survey and administrative data. *Survey Research Methods*, *6*, 113–122.

Sayers, A., Ben-Shlomo, Y., Blom, A. W., & Steele, F. (2015). Probabilistic record linkage. *International Journal of Epidemiology*, *45*(3), 954–964.

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

Schafer, J. L., Ezzati-Rice, T., Johnson, W., Khare, M., Little, R., & Rubin, D. B. (1996). The NHANES III multiple imputation project. *Race/Ethnicity*, *60*(21.2), 15–15.

Schenker, N., Raghunathan, T. E., Chiu, P.-L., Makuc, D. M., Zhang, G., & Cohen, A. J. (2006). Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association*, *101*(475), 924–933.

Schonlau, M., Reuter, M., Schupp, J., Montag, C., Weber, B., Dohmen, T., Siegel, N. A., Sunde, U., Wagner, G. G., & Falk, A. (2010). Collecting genetic samples in population wide (panel) surveys: Feasibility, nonresponse and selectivity. *Survey Research Methods*, *4*, 121–126.

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: A multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909–2930.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *American Journal of Epidemiology*, *179*(6), 764–774.

Si, Y., & Reiter, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics*, *38*(5), 499–521.

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5), 1–13. https://doi.org/10.18637/jss.v039.i05

Sivapriya, T., Kamal, A., & Thavavel, V. (2012). Imputation and classification of missing data using least square support vector machines–a new approach in dementia diagnosis. *International Journal of Advanced Research in Artificial Intelligence*, *1*(4), 29–33.

Sparapani, R., Spanbauer, C., & McCulloch, R. (2021). Nonparametric machine learning and efficient computation with Bayesian additive regression trees: The BART R package. *Journal of Statistical Software*, *97*(1), 1–66. https://doi.org/10.18637/jss.v097.i01

Stan Development Team. (2019). *RStan: The R interface to Stan*. http://mc-stan.org/

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118.

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology*, *169*(9), 1133–1139.

Su, Y.-S., Gelman, A., Hill, J., & Yajima, M. (2011). Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *Journal of Statistical Software*, *45*, 1–31.

Tan, Y. V., Flannagan, C. A., & Elliott, M. R. (2019). "Robust-squared" imputation models using bart. *Journal of Survey Statistics and Methodology*, *7*(4), 465–497.

Therneau, T. M., & Atkinson, B. (2019). *Rpart: Recursive partitioning and regression*

trees. https://CRAN.R-project.org/package=rpart

Therneau, T. M., Atkinson, E. J., et al. (1997). *An introduction to recursive partitioning using the RPART routines*. Technical report Mayo Foundation.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J., & White, I. R. (2018). On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in Medicine*, *37*(15), 2338–2353.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, *23*(5), 373–405.

Twala, B., Jones, M., & Hand, D. J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, *29*(7), 950–956.

Vach, W., & Blettner, M. (1991). Biased estimation of the odds ratio in case-control studies due to the use of ad hoc methods of correcting for missing values for confounding variables. *American Journal of Epidemiology*, *134*(8), 895–907.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219–242.

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

Van Buuren, S., Brand, J. P., Groothuis-Oudshoorn, C. G., & Rubin, D. B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, *76*(12), 1049–1064.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. http://www.jstatsoft.org/v45/i03/

Van der Vaart, A. W. (1988). *Asymptotic statistics*. Cambridge University Press.

Wang, X., Li, A., Jiang, Z., & Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, *7*(1), 32.

Wang, Z., Agung, M., Egawa, R., Suda, R., & Takizawa, H. (2018). Automatic hyperparameter tuning of machine learning models under time constraints. *2018 IEEE International Conference on Big Data (Big Data)*, 4967–4973.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, *29*(28), 2920–2931.

Xu, D., Daniels, M. J., & Winterstein, A. G. (2016). Sequential BART for imputation of missing covariates. *Biostatistics*, *17*(3), 589–602.

Yuan, Z., Burrage, K., & Mattick, J. S. (2002). Prediction of protein solvent accessibility using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, *48*(3), 566–570.

Zhang, G., & Little, R. (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics*, *65*(3), 911–918.

Zhang, Y., & Liu, Y. (2009). Data imputation using least squares support vector machines in urban arterial streets. *IEEE Signal Processing Letters*, *16*(5), 414–417.

Zhao, Y., & Long, Q. (2016). Multiple imputation in the presence of high-dimensional data. *Statistical Methods in Medical Research*, *25*(5), 2021–2035.

Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2016a). A two-step semiparametric method to accommodate sampling weights in multiple imputation. *Biometrics*, *72*(1), 242–252.

Zhou, H., Elliott, M. R., & Raghunathan, T. E. (2016b). Multiple imputation in two-stage cluster samples using the weighted finite population Bayesian bootstrap. *Journal of Survey Statistics and Methodology*, *4*(2), 139–170.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.