UNIVERSITY OF SOUTHAMPTON

Faculty of Social Sciences
School of Mathematical Sciences

Multiple Imputation of a Derived Variable in a Survival Analysis Context

by

Lily Clements
ORCiD: 0000-0001-8864-0552

A thesis for the degree of
Doctor of Philosophy

September 19, 2022

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Philosophy

Multiple Imputation of a Derived Variable in a Survival Analysis Context

by Lily Clements

A data set contains variables that are directly measured, and can be expanded by non-trivial transformations of the measured variable; e.g., dichotomising a continuous variable. Additionally, a new variable can be constructed from several measured variables; e.g., body mass index (BMI) is the ratio of weight and height-squared. The transformed or constructed variable is a derived variable, and the measured variable(s) that build the derived variable are constituents.

A complication in a derived variable arises if at least one value in the constituents is not stored, that is, the derived variable is incomplete. Incomplete variables are a common problem when analysing data and can lead to incorrect inferences in the analysis if mishandled. One approach to deal with them is multiple imputation (MI). In MI, each missing value is replaced several times, yielding several complete multiply imputed data sets. Each data set is analysed, with the results subsequently combined. Two approaches to impute an incomplete derived variable are active and passive imputation. In active imputation, the derived variable is directly imputed, so the functional relationship with the constituents is ignored. In passive imputation, the constituents are imputed and the derived variable is later constructed.

Previous literature finds that the performance of active and passive MI can depend on the model fitted to the multiply imputed data. One gap in the literature is in the performance of active and passive MI in a survival analysis context.

In this thesis, a simulation study is run to investigate the performance of active and passive imputation for three functional forms in a survival analysis context: ratio, additive, and index. In an additive form, the derived variable is a weighted sum of the constituents. In an index form, a numerical variable is categorised as a factor. Conditions investigated include how the missingness is imposed, and the number of predictors to impute the missing values. A special case of passive imputation outperforms active imputation for a ratio and additive functional form. Active imputation outperforms passive imputation for an index functional form.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:.. Date:..................

Acknowledgements

First I would like to acknowledge my supervisors, Dr. Alan Kimber, Dr. Stefanie Biedermann, and Professor Dankmar Boehning who I am very grateful to for their insightful comments and suggestions, invaluable advice, and patience throughout my PhD study.

I would also like to thank my parents, sisters, and brother for their constant support and encouragement, and for being a continual source of inspiration. In addition I would like to extend my thanks to my beautiful springer spaniel, Spencer, for providing companionship and keeping me active throughout my studies - particularly during the pandemic. And I would also like to thank my wonderful bearded collie, Harry, who is sadly not with us anymore.

I also would like to thank George, Sam, Tom, Matt, Ruth, and Laura for their support, handing me a pint when things went pear-shaped, and creating an entertaining and enjoyable atmosphere throughout my PhD.

Lastly, I would like to thank EPSRC for the studentship that allowed me to conduct this thesis.

To Harry and Spencer

Chapter 1

Introduction

If at least one data point in a variable is not stored, the variable is said to be incomplete and the non-stored values are said to be missing. Incomplete variables are a relatively common problem faced by analysts. If handled inappropriately, missing values can result in invalid inferences. These invalid inferences can arise for a number of reasons; for example, mishandling an incomplete variable can affect the relationship between the incomplete variable and other variables in the data set, introducing bias in the parameter estimates when fitting a model (Kang, 2013).

An incomplete data set is a data set which contains at least one incomplete variable. There are several approaches to handle an incomplete data set. A widely used method is Complete Case (CC) analysis, in which an incomplete data set is subsetted to contain only observations with complete values across all variables. The reduced data set is then subject to statistical analysis (Molenberghs and Kenward, 2007). However, CC can introduce problems in the analysis; for example, the cause of the missingness is not addressed, but if larger values of the incomplete variable tend to be missing then the reduced data set is not representative of the data had all values been stored. In addition, omitting the observations with missing values in at least one of their variables results in a reduced data set, and hence a reduced statistical power. As a result of the reduced statistical power, there is a reduced probability of detecting a significant effect when the null hypothesis should be rejected (Kang, 2013). Additionally, a statistical test is more likely to be influenced by random or systematic errors with a smaller data set, increasing the chance of concluding a significant effect when the null hypothesis should not be rejected (Button et al., 2013). An alternative approach to CC is to use imputation. In imputation, each missing value is replaced with a notional value, where these notional values are generally different from each other. There are several approaches to impute incomplete observations, but one widely used approach is Multiple Imputation (MI), introduced in Rubin (1978). MI is split into three phases: imputation, analysis, and pooling. In the first phase, the incomplete values are imputed by fitting a model called an imputation model. In an imputation model, the missing values across all variables in a data set are replaced with placeholder values. These placeholder values are set back as missing for one variable. A model is then fitted to the complete-case data where the outcome variable is the incomplete variable. The fitted model is then used to impute predicted values for the missing values in the incomplete variable. This procedure is then repeated for all variables which initially contained placeholder values. This is the first cycle. The procedure is then repeated for T cycles, where the imputed values are used instead of placeholder values after the first cycle. One complete data set is formed after T cycles. This procedure is repeated M times, yielding M complete data sets. In the analysis phase, a model, called a substantive model, is fitted to each of the M multiply imputed data sets. Finally, in the pooling phase, the estimates resulting from the M substantive models are pooled using a set of rules called "Rubin's Rules". The pooling phase results in a single set of estimates (Rubin, 1978). MI is discussed in more detail in Section 2.2.

One example of an incomplete data set, supplied by NHS Blood and Transplant, contains the survival times of 7732 patients following a kidney transplant. The kidney transplants took place between January 2001 and December 2008, with survival times followed until mid-2013. Including survival time, there are 32 variables. These variables detail characteristic information on donors and recipients, the transplant itself, and the outcome of the transplant. A list of all variables is given in Appendix A.

The transplant data set mostly comprises variables that can be measured directly. The exception to this are the identification variables and two body mass index (BMI) variables. BMI is the ratio of an individual's weight in kilograms and height in metres, squared (CDC, 2015)

$$BMI = weight(kg)/height(m)^2. \tag{1.1}$$

A BMI value between 18.5 and 24.9 indicates normal weight. A value below 18.5 suggests that the individual is underweight, and above 24.9 indicates overweight. BMI does not account for whether high weight is a result of muscle, fat, or something else, and therefore should be used with caution (CDC, 2015). Despite this shortcoming, BMI is commonly used in many application areas. Since BMI is a function of two measured variables, BMI is an example of a derived variable with a ratio functional form. The measured variables, height and weight, are called constituents. A derived variable may take a different functional form and can be constructed from more than two measured variables. For example, a data set may comprise a set of variables that are each a binary response to a question in a survey. The binary responses can be summarised into a single variable by adding the variables together. This is an additive functional form where the derived variable is a weighted sum of the constituent variables. In addition, a derived variable can be a non-trivial transformation of one of the measured variables in a data set; for example, dichotomising a continuous variable. Here, the continuous variable is the constituent, and the new dichotomised variable is the derived variable.

In Pankhurst et al. (2020), BMI is imputed like any other variable. However, BMI is a derived variable (1.1). Two methods to impute a derived variable are active imputation and passive imputation. In active imputation, the derived variable is treated as simply another variable in the imputation process, and the functional relationship with the constituents is ignored (Van Buuren, 2018). This is the approach applied by Pankhurst et al. (2020). In passive imputation, the constituents of the derived variable are first imputed, and then the derived variable is constructed using the functional relationship (Van Buuren, 2018). This led to the initial research question that this thesis aimed to investigate: How does the performance of active imputation compare to the performance of passive imputation for an incomplete derived variable?

Both active and passive imputation have their merits. One complication of passive imputation is incompatibility, which occurs when the variables in the substantive model are not present in the imputation model for the derived variable. As a result, the relationship between the outcome variable in the substantive model and the derived variable may be underestimated, and hence the estimated coefficient may be attenuated, and therefore biased (von Hippel, 2009). However, passive imputation preserves the functional relationship between the constituents and the derived variable, unlike active imputation. One approach to combat incompatibility in passive imputation is Substantive Model Compatible Fully Conditional Specification (SMCFCS) (Bartlett and Morris, 2015). In SMCFCS, the relationship between the incomplete derived variable and the outcome variable is preserved since the imputation model for the incomplete variable is its conditional distribution given the other incomplete and observed variables, together with the substantive model. SMCFCS is discussed in more detail in Section 6.1.

Pankhurst et al. (2020) concluded that MI outperforms CC for the kidney transplant data set, but did not explore the performance of passive imputation, leading to the initial research question: How does passive imputation perform when imputing the kidney transplant data set? As outlined in Section 3, previous research on the performance of active and passive imputation has resulted in different conclusions depending on the type of outcome variable in the substantive model. The treatment of derived variables in a survival analysis context has not been widely explored in the literature, so an aim of this thesis is to investigate the properties of active and passive imputation for a derived variable in the context of survival analysis. The performance of active and passive imputation is investigated in a simulation study. In the simulation study, variations on active and passive imputation models are compared under different conditions. In previous studies in which MI is investigated to impute a derived variable, the performance of active and passive imputation is also affected by the functional form of the derived variable. Therefore another condition accounted for in the simulation study is the functional form of the derived variable.

In Chapter 2 the statistical background for the approaches applied in the simulation study is outlined, beginning with a formal definition of the different structures that missingness can take, followed by a selection of approaches commonly used to handle incomplete observations. The three phases of MI are then discussed in detail along with approaches useful for assessing the fit of the imputation model. An overview of derived variables in data sets is then given, followed by a brief overview of survival analysis methods. A review of the literature is given in Chapter 3, where the literature on multiply imputing a derived variable is discussed. In Chapter 4, methods and results from a preliminary analysis are given, where the motivating kidney transplant data set is imputed using both active and passive imputation. Additional real-world data sets containing variables with additive and index functional forms are also analysed. In Chapter 5, the design of a simulation study is given with the aim to investigate the performance of active and passive imputation for three functional forms. Then the methods and results of the simulation study are given. In addition, some conditions are altered in the simulation study to investigate how they affect the performance of active and passive imputation; for example, the structure of the missing values in the derived variables. In Chapter 6, the results after applying CC and SMCFCS to the data generated in the simulation study are given to compare both CC and MI in the simulation study, as well as to compare SMCFCS to the standard procedure performed in the simulation study. A post-hoc analysis is then carried out in Chapter 7 to review the sensitivity of the results in the simulation study. In Chapter 8, an illustrative analysis is performed to one of the motivating data sets, given the results found in the simulation study. Finally, in Chapter 9 is a conclusion, detailing the results found in the analyses, as well as further work for future analyses, and a critical review of the thesis.

Chapter 2

Statistical Background

In this chapter, details of the statistical methodology for multiply imputing an explanatory variable are outlined, namely, when the explanatory variable is a derived variable. There are different mechanisms to categorise missingness in an incomplete variable, outlined in Section 2.1, along with different approaches to handling missing values. MI is discussed in Section 2.2, starting with the procedure and followed by diagnostic approaches to evaluate the imputation procedure. Derived variables, and adapting MI in the context of a derived variable, is outlined in Section 2.3. Finally, general methods for applying survival analysis to the data are given in Section 2.4. This is followed by modifications required when survival data is coupled with MI.

2.1 Missing Data

Incomplete variables are variables that are missing at least one value. The cause of missing values affects how an incomplete variable should be treated. This cause is called the missingness mechanism and is discussed first in this section. Methods for handling missing data are then described to help justify the use of MI.

2.1.1 Missingness Mechanisms

The cause of incomplete values can be categorised into one of three missingness mechanisms (Rubin, 1976), but first some notation must be established. Denote by $\boldsymbol{X}$ a data set of size $n \times p$ with n units and p covariates. Then $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ where $\boldsymbol{x}_j = (x_{1j}, ..., x_{nj})'$ is a vector of values for covariate j, $j = 1, ..., p$. For each unit, i, $i = 1, ..., n$, and each covariate, j, denote by r_{ij} a missingness indicator for x_{ij}: $r_{ij} = 1$ if x_{ij} is observed, and $r_{ij} = 0$ otherwise. Denote by $\boldsymbol{R}$ a missingness matrix where $\boldsymbol{R} = (\boldsymbol{r}_1, ..., \boldsymbol{r}_p)$. Then a binary vector indicating whether an individual's value is missing or present for variable j is given by $\boldsymbol{r}_j = (r_{1j}, ..., r_{nj})^{'}$. The observed values in the data set of explanatory variables can then be denoted by $\boldsymbol{X}_O = \{x_{ij}|r_{ij} = 1\}$ and the unobserved values by $\boldsymbol{X}_M = \{x_{ij}|r_{ij} = 0\}$. The missingness mechanism is then given by $P(\boldsymbol{R}|\boldsymbol{X})$. The three mechanisms are as follows:

Missing Completely at Random (MCAR), also referred to as "Not Data Dependent" (Hand, 2020), is the simplest form to deal with. The missingness does not depend on the data, $P(\boldsymbol{R}|\boldsymbol{X}) = P(\boldsymbol{R})$. Hence, the distribution of the unobserved values for a variable is the same as that of the observed values for the same variable.

Missing at Random (MAR), also referred to as "Seen Data Dependent" (Hand, 2020), occurs when the missingness does not depend on the unobserved data, given the observed data, $P(\boldsymbol{R}|\boldsymbol{X}) = P(\boldsymbol{R}|\boldsymbol{X}_O)$. A variable is MAR when the probability of a missing value in an incomplete variable is attributable to the observed value of at least one other variable. For example, consider a data set with a complete binary variable, sex, and an incomplete variable, weight. If weight is more likely to be observed for one sex than for another, then the weight variable is MAR.

Missing not at Random (MNAR), also referred to as "Unseen Data Dependent" (Hand, 2020), occurs when the missingness depends on other unobserved data $P(\boldsymbol{R}|\boldsymbol{X}) \neq P(\boldsymbol{R}|\boldsymbol{X}_O)$. A variable is MNAR when the probability of a missing value in an incomplete variable is attributable to the value of that variable itself. For example, income may be less likely to be disclosed for individuals in a higher income bracket. The income variable here is MNAR since the reason for its missingness is the value itself.

Diagnosing the missingness mechanism would allow the unobserved values to be appropriately imputed. However, in most studies there is no complete certainty that the diagnosis is correct since the cause of the missingness is usually unknown. Methods for handling data with a MCAR or MAR missingness mechanism are discussed next in Section 2.1.2. Approaches to handle data with a MNAR missingness mechanism can be found in Van Buuren (2018).

2.1.2 Approaches for Handling Missing Values

The most common approach to handle an incomplete covariate is complete case analysis (CC) (Kang, 2013), in which any observations containing at least one incomplete covariate are removed. CC results in a subsetted data set to analyse (Molenberghs and Kenward, 2007). An advantage of CC is that it is quick to perform using standard statistical software and simple to understand, but there are drawbacks. An individual with an incomplete covariate for one variable may have observed values for other variables. Since data can be expensive to collect, removing this row can induce a loss of resources, and can alter the distribution of fully observed variables. In addition, removing observed data can cause a loss of information, thereby increasing the standard errors when fitting a model to the data. Finally, CC can induce sampling bias when an incomplete variable does not follow a MCAR missingness mechanism since the complete case data may not follow the same distribution as the underlying data (Little and Rubin, 2002; Molenberghs and Kenward, 2007). For example, consider a data set with a partially observed variable, age, with MCAR values, and two fully observed variables, sex and income (Table 2.1). In CC, the distribution in the CC data is not substantially changed since age is MCAR (Figure 2.1), but the information observed in sex and income is removed. Wayman (2003) suggests that CC should be avoided despite its ease of use.

TABLE 2.1: Example data set with three variables: age, sex, and income. Values are MCAR in the age variable.

Age	Sex	Income
57	Male	43714
NA	Female	30191
NA	Female	33454
32	Female	25010
NA	Male	47494
51	Male	40000
41	Male	32976
NA	Female	44187
NA	Male	50955
64	Male	16588

An alternative approach to handle missing values is to replace each missing value with a notional value. This is imputation. One method of imputation that can be adopted is to replace each missing value in a variable with a point estimate, such as the mean of the observed values in the incomplete variable (Little and Rubin, 2002). This is called Single Value Imputation and is a simple and quick approach which additionally preserves the sample size. However, the variability of the values for the imputed variable is reduced, both because all missing values are imputed with the same value and because the imputed value is often in the centre of the distribution for the variable. For example, replace the missing values in the data set given in Table 2.1 by the mean of the observed age values. The distribution of the variable has changed compared to the actual values (Figure 2.1). As a result, single-value imputation can weaken the relationship between the imputed variable and other covariates due to a change in the correlation between variables (Wayman, 2003; UCLA, 2017).

Gmel (2001) recommends Hot Deck imputation over single-value imputation. In hot-deck imputation, each unobserved value in a variable is replaced by an observed

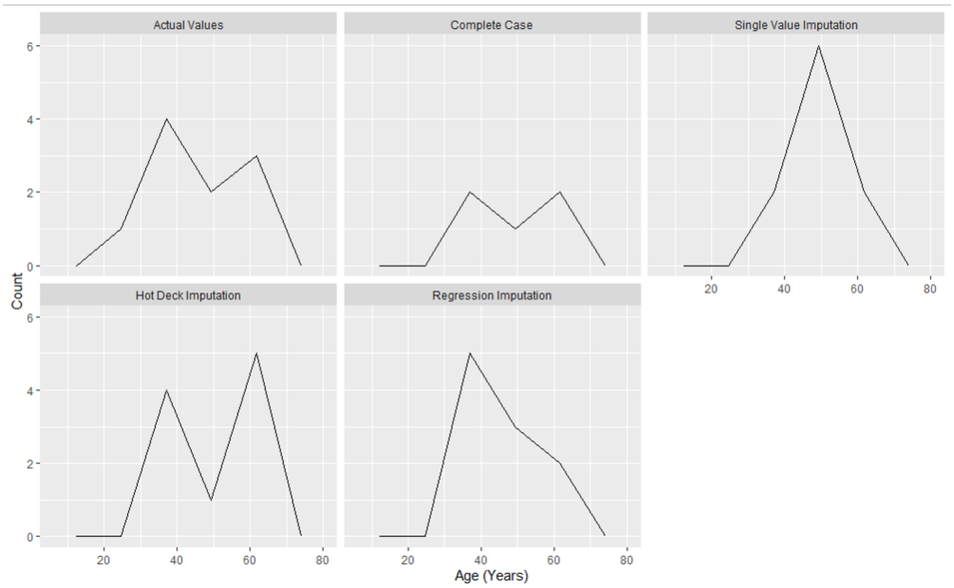


FIGURE 2.1: Frequency polygon of the age variable from Table 2.1 for the actual age, as well as different approaches to impute age.

value selected from an individual with similar characteristics (Little and Rubin, 2002). In hot deck imputation, reasonable values are imputed with variability accounted for (Little and Rubin, 2002). In addition, the imputed data is confined to the range of the observed data. However, restricting the range of the imputed data may not be appealing for variables that are MAR or MNAR since under these missingness structures, the observed portion of the variable can follow a distribution different to that of the underlying variable. Additionally, a large data set is required for hot deck imputation to be effective if there is a large proportion of incomplete variables (Andridge and Little, 2010). For example, applying hot deck imputation to the data in Table 2.1 may not be a good solution to impute female individuals since there are fewer observed females in the data set than males.

Another approach is Regression Imputation. In regression imputation, a model is fitted with the partially unobserved covariate as a response variable. The covariates can be a selection of the remaining explanatory variables in the data set and the original response variable. For example, using the data set in Table 2.1, a linear regression model can be fitted with the partially observed variable as the outcome variable: $age \sim sex + income$. This model can be used to then predict missing values in the age variable. The missing values are then imputed by replacing them with predicted values from the regression model (Little and Rubin, 2002), allowing flexibility in imputing different data types. For example, a logistic regression model can be fitted to a partially observed binary variable. Further, regression imputation preserves the relationship between the imputed variable and other variables (Zhang, 2016). However, UCLA (2017) find that regression imputation magnifies the relationship between the imputed covariate and other variables, since the imputed values are highly correlated. A main flaw in regression imputation is that the resulting estimates do not reflect the uncertainty of the imputed values. To combat uncertainty, the imputation regression can be repeated multiple times, producing several data sets (Wayman, 2003). This is multiple imputation and is explained in more detail in Section 2.2.

Complete case analysis and imputation are not the limit of approaches when handling missing data. Likelihood approaches, such as the expectation maximisation (EM) algorithm, are found to be an effective solution to handle incomplete variables (Enders, 2010). However, estimating the standard errors is a challenge in the EM algorithm. Additionally, a data set with multiple incomplete variables is handled better by MI than by EM (Enders, 2010). This is important since a derived variable may be built from multiple variables. Additionally, if an incomplete variable is formed from item-based questionnaires, such as a derived variable which is weighted sum of several constituents, Enders (2010) states that MI is more flexible and straight-forward than EM. Furthermore, MI has been shown to perform well when there is a high proportion of incomplete covariates (Wayman, 2003). Additionally, MI is readily available in many computer programs, with the output as a complete data set fit for analysis (McCleary, 2002) allowing for a simple, straight-forward approach for the user. It is for these reasons that MI is investigated when comparing the performance of active to passive imputation in this thesis. The procedure to implement MI is next discussed in more detail.

2.2 Multiple Imputation

MI was introduced by Rubin (1978), and involves repeatedly imputing each missing value to account for the uncertainty in the imputed values. MI involves three phases: Impute, Analyse, Pool (Carpenter and Kenward, 2012), outlined in Figure 2.2.

Impute In the first phase, the incomplete observations are imputed M times using an *Imputation Model*. In an imputation model, the incomplete variable subject to imputation is the outcome variable. Each imputation yields a complete data set.

Analyse Each of the M complete data sets are analysed by fitting a *Substantive Model* (or an *Analysis Model*) to them. This model does not necessarily contain all p variables in the data set (see Section 2.2.3) so Figure 2.2 has q covariates, where $q \leq p$.

Pool The results from the M substantive models are combined.

These three phases are expanded next in this section. Following this, model selection, diagnostics, and assessment are discussed for both the imputation models and substantive models.

2.2.1 Impute

A widely used approach to impute missing data is to use Multiply Imputed Chained Equations (MICE), also known as Fully Conditional Specification (FCS). The procedure for MICE, as outlined in Azur et al. (2011), is illustrated in Figure 2.3. The MICE process is repeated M times resulting in M imputed values for each missing value.

Choosing the value of M is important. It was initially standard for M to range between three and five (Enders, 2010), but in more recent years a larger value of M is more common due to an increased computing power (UCLA, 2017). Furthermore, Graham et al. (2007) finds that more than five imputations are required to stabilise the standard errors for the parameter estimates. As a result, Enders (2010) suggests that M should be at least 20, but this minimum bound increases when a large percentage of a variable is incomplete.

Step 3 of MICE involves fitting a regression model to the j^{th} variable to impute the missing values. Some specific approaches for applying the regression model are Bayesian Linear Regression (BLR) and Predictive Mean Matching (PMM).

Additionally, the regression model to impute missing values can be chosen depending on the form of the incomplete covariate; for example, a logistic regression imputation model (logit) can be fitted to an incomplete binary variable. More available approaches are given in the `MICE` package in `R` (Van Buuren, 2018). BLR, PMM and logit are discussed next in this section.

Bayesian Linear Regression Denote the observed values in the j^{th} variable by $x_{j_{obs}}$, and the missing values by $x_{j_{mis}}$. $x_{j_{obs}} \sim N(\mu, \sigma^2)$ and $x_{j_{mis}} \sim N(\mu, \sigma^2)$. The BLR method, as outlined in Van Buuren (2018, Chapter 3.1-3.2) and Schenker and Taylor (1996, pp 428-429), is as follows.

Run through the MICE procedure outlined in Figure 2.3. In step three, a regression imputation model is fitted to the observed values in the j^{th} variable. This yields a set of estimated coefficients for the $q-1$ predictors in the model, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, ..., \hat{\theta}_q)$, and the estimated variance, $\hat{\sigma}^2$. If a different set of values in the j^{th} variable were incomplete, a different set of observed data would be the basis to estimate the parameters. As a result, there is some uncertainty in the parameter estimates. To account for this uncertainty, estimate new parameters $\dot{\boldsymbol{\theta}}$ and $\dot{\sigma}^2$ by drawing random samples from their posterior distribution. White et al. (2011) give the draw for $\dot{\sigma}$ as

$$\dot{\sigma}^2 = \hat{\sigma}^2 (n_{obs} - q)/g$$

where $g \sim \chi^2_{n_{obs}-q}$ for q estimated coefficients in the regression model and n_{obs} observed values in the j^{th} variable. The numerator for $\dot{\sigma}^2$ is therefore the residual sum of squares from the initial regression model fitted to the observed data.

The draw for $\dot{\boldsymbol{\theta}}$ is given by White et al. (2011) as

$$\dot{\boldsymbol{\theta}} \sim N(\hat{\boldsymbol{\theta}}, \dot{\sigma}^2(\boldsymbol{X}_O^{'}\boldsymbol{X}_O)^{-1})$$

where $\boldsymbol{X}_O$ is the portion of the data set which contains observed x_{ij} values. In the `MICE` package in R, standard non-informative priors are used, $P(\theta, \sigma^2) \propto 1/\sigma^2$.

Impute each missing value, $x_{ij} \sim N(\boldsymbol{X}_M\dot{\boldsymbol{\theta}}, \dot{\sigma}^2)$ where $\boldsymbol{X}_M$ is the portion of the data set with incomplete x_{ij} values. Upon imputing x_j, continue from step 4 in the MICE procedure.

Predictive Mean Matching PMM was introduced by Little (1988) and involves imputing missing values in a variable by sampling from the observed values in that variable. PMM is suitable for different variable types. Additionally, under PMM the distribution of the imputed data follows that of the observed data, so PMM may be appropriate when an incomplete variable violates the assumption of normality. However, Marshall et al. (2010) find that PMM results in biased parameter estimates in the pooled substantive model if 75% of a variable is incomplete. Furthermore, PMM is not recommended with a small sample size (Marshall et al., 2010). Finally, since PMM ensures the imputed values are confined in the range of the observed data (Van Buuren, 2018), PMM is not recommended when the missing values should be imputed beyond the region set by the observed data (White et al., 2011). The method, as outlined by (Schenker and Taylor, 1996, p. 429-430), is as follows:

Follow the approach for BLR until the estimates of $\dot{\boldsymbol{\theta}}$ and $\dot{\sigma}^2$ are randomly drawn. For unobserved values of x_{ij}, calculate $\dot{x}_{ij} \sim N(\boldsymbol{X_M}\dot{\boldsymbol{\theta}}, \dot{\sigma}^2)$. For observed values of x_{ij}, calculate $\hat{x}_{ij} \sim N(\boldsymbol{X_O}\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$.

For each $\dot{x}_{ij}$, determine the closest d values from the observed data by calculating the absolute difference between the $\dot{x}_{ij}$ value, and all $\hat{x}_{ij}$ values: $|\dot{x}_{ij} - \hat{x}_{ij}|$. This can give a pool of d $\hat{x}_{ij}$ values close to the predicted $\dot{x}_{ij}$ value. Replace each $\hat{x}_{ij}$ in this pool with their observed value, x_{ij}, and replace $\dot{x}_{ij}$ by randomly sampling from the pool of observed x_{ij} values. As in BLR, PMM is repeated for all variables with incomplete covariates, resulting in one of the M data sets.

Note that in the `MICE` package, $d = 5$ by default and is recommended by Schenker and Taylor (1996) for a data set with $n > 100$. A low value of d can cause an issue of duplicated imputations, and large d can cause the imputed value to be further from the true value since a larger pool is considered (Van Buuren, 2018).

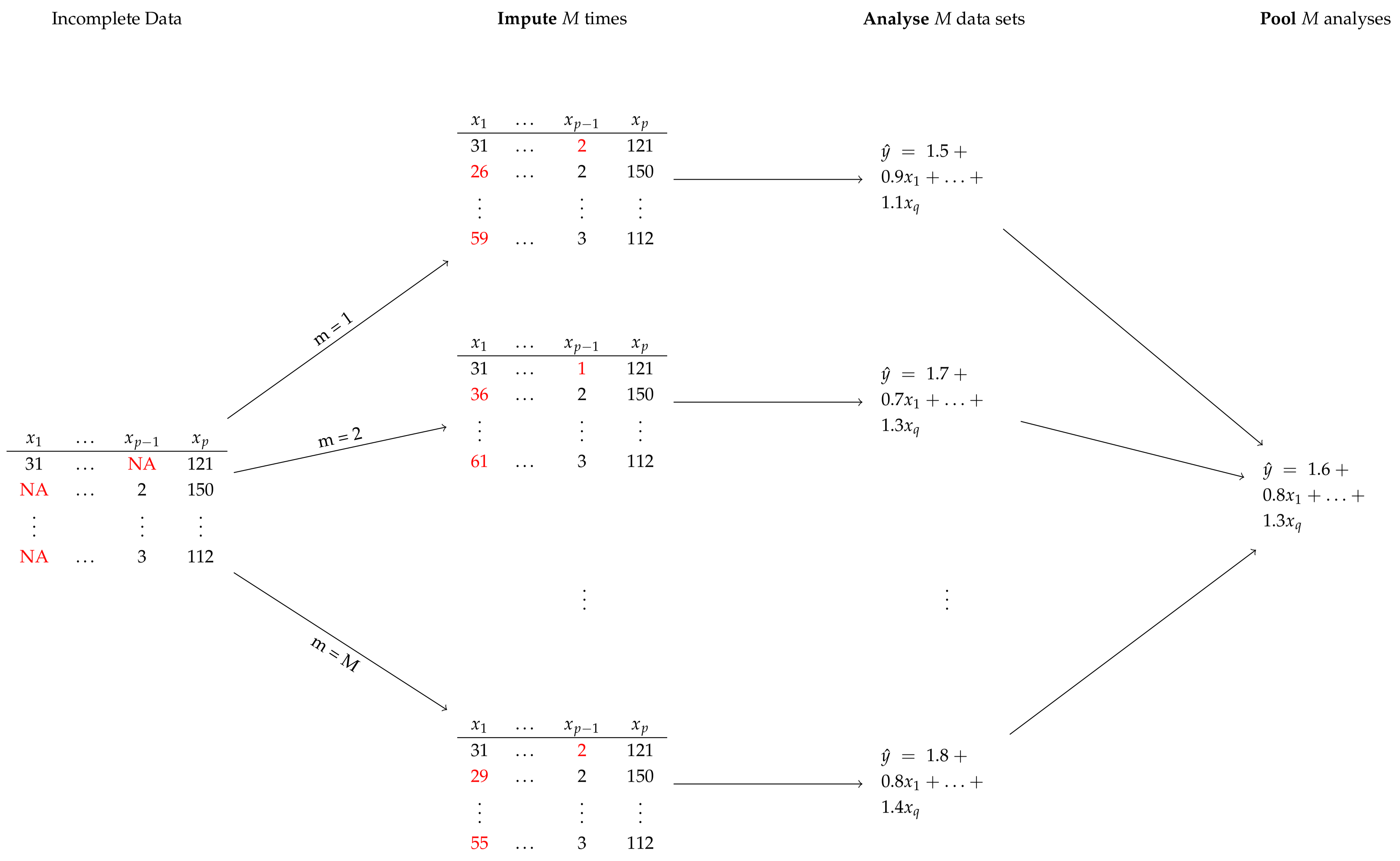


FIGURE 2.2: The three phases of Multiple Imputation: Impute, Analyse, Pool given for a hypothetical data set with a linear regression model fitted as the substantive model.

Bayesian Logistic Regression An incomplete binary variable can be imputed through Bayesian logistic regression (Van Buuren, 2018, Chapter 3.6). In step three of the MICE procedure (Figure 2.3), fit a regression imputation model to the observed values in the j^{th} variable. This yields a set of estimated coefficients for the $q-1$ predictors in the model, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, ..., \hat{\theta}_q)'$, and an estimated covariance matrix of $\hat{\boldsymbol{\theta}}$, V.

As with BLR and PMM, calculate $\dot{\boldsymbol{\theta}}$ to account for the uncertainty in the parameter estimates where

$$\dot{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + \dot{z}_1 V^{1/2}.$$

Here, $\dot{z}_1 \sim N(0,1)$ and $V^{1/2}$ is the square-root of V calculated by Cholesky decomposition. Define by n_{mis} the number of partially observed variables in $\boldsymbol{X_M}$. Estimate n_{mis} probabilities, $\dot{p}$:

$$\dot{p} = \frac{1}{1+\exp(-\boldsymbol{X_M}\dot{\boldsymbol{\theta}})}.$$

For the j^{th} partially observed variable, draw $\boldsymbol{u}_j \sim Uniform(0,1)$ distribution n_{mis} times. In $\boldsymbol{u}_j$. When $\boldsymbol{u}_j < \dot{\boldsymbol{p}}_j$, set the incomplete value at 1, otherwise set the incomplete value to 0 for $j = 1, ..., n_{mis}$.

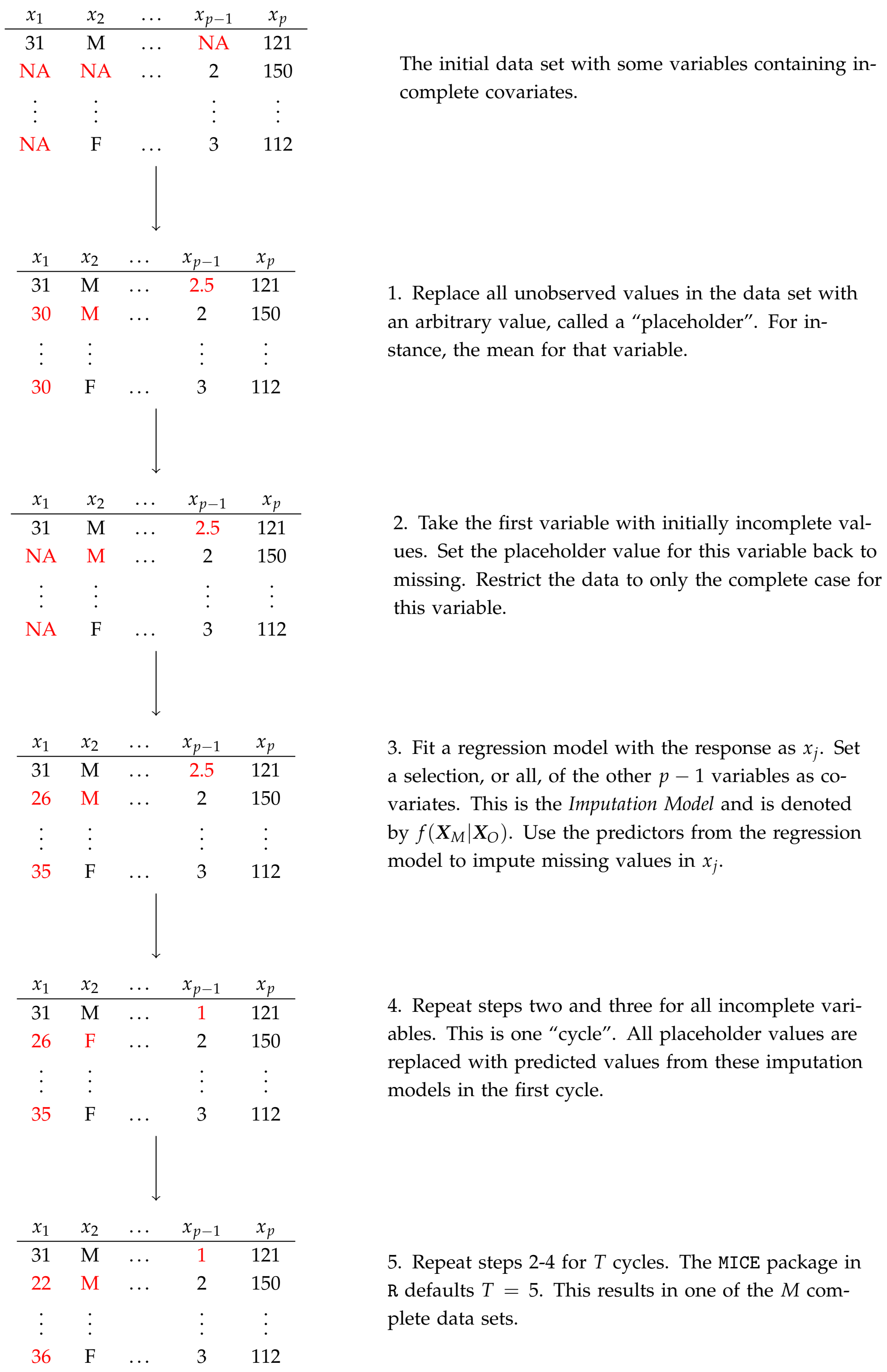
2.2.2 Analyse and Pool

A substantive model is fitted to each multiply imputed data set. Each substantive model fitted to the multiply imputed data has a set of parameter estimates. Rubin (1987) defines a set of rules to pool each estimated parameter. Define by Q the parameter of interest. Hence $\hat{Q}_m$, $m = 1, ..., M$, denotes a point estimate for the parameter of interest in the m^{th} multiply imputed data set. Similarly, $\hat{V}_{W_m}$ denotes the estimated variance for the estimated parameter of interest in the m^{th} multiply imputed data set.

Point estimates of the parameter of interest can be pooled to calculate an average of the M estimates:

$$\bar{Q}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{Q}_m.$$

The corresponding total variance for an estimate of Q is built from two components: the within-imputation variance and the between-imputation variance. The within-imputation variance is the variability within the M point estimates and is defined by calculating an average of the M estimated variances:



x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	NA	121
NA	NA	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
NA	F	$\dots$	3	112

The initial data set with some variables containing incomplete covariates.

x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	2.5	121
30	M	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
30	F	$\dots$	3	112

1. Replace all unobserved values in the data set with an arbitrary value, called a “placeholder”. For instance, the mean for that variable.

x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	2.5	121
NA	M	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
NA	F	$\dots$	3	112

2. Take the first variable with initially incomplete values. Set the placeholder value for this variable back to missing. Restrict the data to only the complete case for this variable.

x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	2.5	121
26	M	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
35	F	$\dots$	3	112

3. Fit a regression model with the response as x_j. Set a selection, or all, of the other $p-1$ variables as covariates. This is the *Imputation Model* and is denoted by $f(\boldsymbol{X}_M|\boldsymbol{X}_O)$. Use the predictors from the regression model to impute missing values in x_j.

x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	1	121
26	F	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
35	F	$\dots$	3	112

4. Repeat steps two and three for all incomplete variables. This is one “cycle”. All placeholder values are replaced with predicted values from these imputation models in the first cycle.

x_1	x_2	$\dots$	x_{p-1}	x_p
31	M	$\dots$	1	121
22	M	$\dots$	2	150
$\vdots$	$\vdots$		$\vdots$	$\vdots$
36	F	$\dots$	3	112

5. Repeat steps 2-4 for T cycles. The `MICE` package in `R` defaults $T = 5$. This results in one of the M complete data sets.

FIGURE 2.3: Diagram of the method of MICE

$$\bar{V}_W = \frac{1}{M}\sum_{m=1}^{M} \hat{V}_{W_m}.$$

The between-imputation variance is the variability between the M point estimates and is defined by

$$\hat{V}_B = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{Q}_m - \bar{Q}_M)^2.$$

The total variance is given by $\hat{V}_T = \bar{V}_W + \hat{V}_B + \frac{\hat{V}_B}{M}$, where $\frac{\hat{V}_B}{M}$ is a penalty to account for the uncertainty of the imputed values in the multiply imputed data sets. As M increases, the penalty decreases. When a considerable proportion of the total variance derives from the between-imputation variance, an increase in M reduces the total variance.

2.2.3 Variable Selection for the Models

Some adjustments are made for variable selection when constructing the imputation and substantive models. The reason for the adjustments, as well as the modifications themselves, are discussed in this section.

Imputation Model Variable selection is important to avoid bias in both the imputation and analysis, but can be a contentious point for the imputation model.

Incompatibility occurs when a variable that is present in the substantive model is not present in the imputation model. In an incompatible model, there is no relationship between the imputed value and any observed values in variables that are only present in the substantive model. The relationship between the imputed variable and any variables that are not predictors in the imputation model is then underestimated, with the resulting parameter estimate in the substantive model attenuated (von Hippel, 2009). It is of particular importance to include the response variable from the substantive model as a predictor in the imputation model to avoid underestimating this relationship (White et al., 2011).

Although variables in the substantive model should be in the imputation model, the converse does not hold. Variables excluded from subsequent analyses but in the imputation model are "auxiliary variables" and still contribute to the imputation procedure. First, auxiliary variables can improve the precision of the imputation if they are strongly correlated with the incomplete variable (Nguyen et al., 2017), and further they allow information in the imputed variables to contribute to the overall picture without being included in any further analyses (Enders, 2010). Additionally, under a MAR assumption the inclusion of auxiliary variables allows for the observed variables related to the cause of the missingness mechanism impute the unobserved values. Therefore, it is important to include auxiliary variables under a MAR assumption since imputation methods such as MI are sensitive to deviations from an assumption of MAR (Clark and Altman, 2003; White et al., 2011).

In general, White et al. (2011) suggest that predictors in an imputation model should be either present in the substantive model, strongly associated with the incomplete covariate, or associated with the missingness mechanism in the incomplete covariate. Limiting the number of covariates means that the imputation is less computationally intensive and therefore quicker to perform. However, Enders (2010) suggests that all variables in the data set should be predictors in the imputation model.

Substantive Model An appealing feature of MI is that the output takes the form of a data set ready for analysis, allowing for a smooth transition between imputing and analysing the data. However, some modifications are required when performing variable selection and hypothesis testing.

In data sets with a large number of variables, a selection of significant variables are chosen as explanatory variables when modelling for parsimony. Because the analyses of the M data sets are pooled, the variables selected across the M substantive models should be the same (White et al., 2011). One way to combat this is to stack the M imputed data sets into a data set of dimension $(n \times M) \times p$, where $n \times p$ is the dimension in the incomplete initial data set. Standard variable selection can then be applied to the stacked data set. To account for the inflated number of rows, White et al. (2011) suggests assigning a weight $(1 - FMI_x)/M$ to each row. FMI is formally defined in Section 2.2.4, but is approximately the proportion of missing values in an incomplete variable, x. Once the variable selection is complete, a substantive model can be fitted to each of the M data sets using the set of covariates arising from the variable selection process (White et al., 2011).

2.2.4 Model Diagnostics and Assessment

An ill-specified imputation model can result in poor imputations, and hence incorrect inferences can be drawn from the imputed data. Therefore it is important to assess the fit of an imputation model. Decisions made when building an imputation model include, but are not limited to, selecting the MI generation method itself (for example, BLR and PMM), determining predictor variables in the imputation model, and selecting the number of imputations to perform. Guidance for making these decisions have been outlined previously in Section 2.2. In this section, approaches for diagnosing issues in, and assessing the performance of, the imputation model are set out in the context of a simulation study. Methods to evaluate an imputation model outside of a simulation study are given in Appendix B.

A simulation study where the aim is to investigate the performance of an imputation models generally involves a fully-observed data set. Missing values are first generated in a variable(s) in the fully-observed data set, and then subsequently imputed. Comparing the imputed data set to the true data set can help to assess the imputation procedure.

A simulation study may be useful to evaluate the performance of an imputation model, or otherwise justify an imputation procedure. For example, Pankhurst et al. (2020) compare the use of MI against CC through a simulation study with an incomplete data set. The incomplete data set is first reduced to the complete data set only. The simulation then uses the single fully observed data set. A missingness structure is imposed to the fully observed data set such that the missingness follows that of the initial incomplete data. Missing values are subsequently imputed by MICE.

In a simulation setting, there are several approaches to assess the fit of the imputation model. For example, the imputed values themselves can be compared to the true values through standard descriptive statistics, such as comparing ranges, extreme values, and means of the imputed values to those of the true or observed values.

Another approach to assess the performance of an imputation model is by means of model selection, as demonstrated by Pankhurst et al. (2020). Before performing the simulation, Pankhurst et al. (2020) fit a substantive model to all variables in the complete data. They perform stepwise selection to determine a set of significant explanatory variables. This approach provides a set of covariates that should be significant in the substantive model in the simulation study. In each replication, for each imputation model, Pankhurst et al. (2020) stack the M multiply imputed data sets to give a data set of length $n \times M$. Stepwise selection is then performed to the stacked data set, with the increased length accounted for by weighting the observations. The performance of the imputation models is determined by comparing the variables selected for the substantive model in each replication with those selected in the complete data. If instead the data were generated from a model then the variables in the substantive model could be compared to the true model. First, the proportion of correctly chosen covariates can be calculated. A proportion closer to one indicates a superior model since the correct variables are selected. Similarly, the proportion of incorrectly chosen covariates is considered. A proportion closer to zero implies that an imputation model performs better than if the proportion is further from zero.

The analysis and pooling phases of MI are additionally useful to evaluate the imputation procedure. In Section 2.2.2, the use of a quantity of interest, Q, and Rubin's Rules to combine estimates of Q across the M data sets into a single value, $\bar{Q}_M$, is discussed. Van Buuren (2018) outlines approaches to evaluate the performance of the imputation procedure. An estimate, $\hat{Q}$, of the population parameter, Q, can be calculated from the multiply imputed data. Since each replication repeats the procedure of "impute, analyse, pool", illustrated in Figure 2.2, each replication and imputation model results in a pooled estimate of the population parameter, $\bar{Q}$. An overall mean of the pooled estimates in all replications is given by $\mathbb{E}[\bar{Q}]$ where $\mathbb{E}$ denotes that the mean is calculated. The performance of the imputation model can then be evaluated by considering the raw bias (RB) in the estimated parameter of interest compared to the complete data estimate, $RB = \mathbb{E}[\bar{Q}_M] - Q$. If $\bar{Q}$ is a parameter estimate, a pooled standard error is associated with each $\bar{Q}$ value. The Estimated Average Standard Error (EASE) is the mean of the pooled standard errors across all replications (Lee and Carlin, 2012). For accurate inference, EASE needs to approximately equal the empirical variance of the $\hat{Q}$ values. An imputation model has outperformed another approach if the RB and EASE values are closer to zero than the RB and EASE values of another approach.

The coverage rate (CR) is the proportion of replications where the true value of Q is in a 95% pooled confidence interval for $\hat{Q}$. In the MICE package in R a Wald confidence interval is used. Let $\hat{p}$ denote a point estimate of a proportion, n denote the sample size, and z denote the threshold value for a given confidence level. Then the Wald confidence interval for p is given by

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Undercoverage occurs when the CR is less than 95%. If there is undercoverage, there is an indication that the estimated confidence intervals are too narrow, resulting in a conclusion that the coefficient estimate significantly affects the outcome variable when it does not. When there is overcoverage, when the coverage rate is above 95%, there is an indication that the estimated confidence intervals are too wide resulting in a conclusion that there is insufficient evidence that the coefficient estimate significantly affects the outcome of interest, when there is sufficient evidence to conclude an affect. Overcoverage is preferable to undercoverage (Van Buuren, 2018).

Another measure that can be calculated from the substantive model is the average width of the confidence interval (AW). In all replications, the mean AW of the estimated confidence interval for $\hat{Q}$ is calculated. AW should be used to distinguish the performance between imputation models which have a small bias and good coverage rate (Van Buuren, 2018). A smaller AW implies that the imputation model performs better than an imputation model with a larger AW.

Assessing the Imputation Procedure The between-imputation variances and within-imputation variances are used to construct measures to evaluate each estimated coefficient in the pooled substantive model. One measure is the Fraction of

Missing Information (FMI) (Enders, 2010), the proportion of total variance for an estimated parameter of interest that is caused by missing values. The FMI of a parameter estimate is defined by

$$FMI = \frac{\hat{V}_B + \frac{\hat{V}_B}{M}}{\hat{V}_T}.$$

The FMI value ranges from zero to one, where $FMI = 0 \implies \hat{V}_B = 0 \implies \hat{V}_T = \bar{V}_W$. That is, the missing values in the data do not add any additional variation in the parameter estimate. Similarly, $FMI = 1 \implies \hat{V}_T = \hat{V}_B + \hat{V}_B/M$, that is that the variation is entirely due to the missing data in the corresponding variable. UCLA (2017) suggests that the highest FMI percentage for the parameter estimates should equal the number of imputations, M.

FMI can be rearranged to give the Relative Increase in Variance (RIV). The RIV is the proportional increase in variance for an estimated parameter that is caused by missing values in the variable associated with the estimated parameter. It is defined by

$$RIV = \frac{FMI}{1 - FMI} = \frac{\hat{V}_B + \frac{\hat{V}_B}{M}}{\bar{V}_W}.$$

A large RIV or FMI value implies that either $\hat{V}_B$ is large, or $\bar{V}_W$ is small. Therefore, a larger RIV or FMI value implies that the increase in the variance is due to missing values in the data. Therefore, a large RIV or FMI value indicates either that there are poor predictors in the imputation model for the associated variable causing a high between-imputation variance, or that a large proportion of the associated variable is missing and hence imputed (UCLA, 2017).

2.3 Derived Variables

A data set initially comprises variables that can be measured directly. This data set can then be expanded by making a non-trivial transformation of one of the measured variables; for example, dichotomising a continuous variable. Additionally, the data set can be expanded by constructing a new variable by combining two or more of the measured variables in the initial data set. The new constructed or transformed variable is a derived variable, and the measured variables that are used to construct a derived variable are called constituents. Derived variables can take on different functional forms and are prominent in many different fields. In this section, examples of derived variables are given where the derived variables take on different functional forms, are useful for different purposes, or are prominent in different fields.

One functional form that a derived variable can take is a ratio. One example of this is with BMI in Pankhurst et al. (2020) where BMI is a function of two measured variables, weight and height:

$$BMI = weight/height^2$$

for weight in kilograms and height in metres. BMI is a measurement that can indicate an individual's body fat levels. By accounting for the height of the individual, BMI can be preferable to using just a weight variable. Other ratio functional forms have been used in the literature. For example, patients with acute respiratory distress syndrome (ARDS) were monitored using an invasive approach to determine their physiologic shunt fraction. Covelli et al. (1983) found a strong correlation between the invasive approach where the variable of interest is directly obtained, and a noninvasive approach where the variable of interest is constructed from the ratio of two other variables (arterial oxygen to inspired oxygen concentration). Covelli et al. (1983) concludes that the ratio of arterial oxygen to inspired oxygen concentration accurately reflects the sp/t value, and is simple to calculate. Hence, due to its non-invasive nature, is recommended as a measurement for individuals with ARDS.

Variables with a ratio functional form are present in many fields; for example, a variable with a ratio functional form is prominent in physics or engineering (Kalla, 2011), such as to calculate mass, duration, or energy. For example:

$$\text{velocity} = \frac{\text{distance}}{\text{time}}.$$

Furthermore, variables with a ratio functional form are present in other fields, such as economics. For example, if a company wished to compare the performance between two time periods, they can calculate the profit margin for each time period:

$$\text{profit margin} = \frac{\text{revenue - expenses}}{\text{revenue}}.$$

The ratios can then be compared to one another (Segal, 2021).

Variables may also be derived from a set of other variables because they cannot otherwise be calculated. For example, in a psychology study by Lent et al. (1987), self-efficacy, indecision thinking, and interest congruence are three derived variables used to predict career and academic behaviour of an individual. These three derived variables are measured using questionnaires with the results summarised to give a single value for each participant for each variable. In Lent et al. (1987), self-efficacy is measured using a questionnaire where an individual rates their own confidence on a scale of 1-10 that they would qualify in different fields in science and engineering. The questions are repeated for 15 different fields, yielding 15 different scores per individual. The mean value of these scores is calculated to give the individual's self-efficacy level, so the functional form used for measuring self-efficacy is the mean, an additive functional form. The same 15 fields are used in a questionnaire to determine an individual's levels of indecision. In this questionnaire, the individual states, on a scale of 1-10, the extent to which they have seriously considered pursuing a career in the field of interest. The sum of these scores gives the individual's indecisive level; therefore, the functional form used for measuring indecision is an additive form. Interest congruence is how compatible an individual's interests are with their occupation or subjects they study. In the study by Lent et al. (1987), interest congruence is measured using results from two different variables: how realistic the individual is, and how investigative the individual is. If an individual scores highly on both variables, they are said to be congruent. If they score highly on one and not the other, the interest congruent variable is given a moderate congruence indicator. If they do not score highly on either variable, they are given an incongruent indicator. The derived variables are present in other studies, such as Schaefers et al. (1997).

Similar summary variables are found elsewhere in the literature, including in other fields, as derived variables. Fish et al. (2010) uses both BMI and annual income as derived variables, where annual income is calculated by summing the earnings of the household members. Lagona and Zhang (2010) estimate a survival analysis model. A covariate in the survival analysis model is a factor variable that has been constructed by categorising a continuous variable. A mini-mental state examination (MMSE) is an approach used to measure an individual's cognitive ability, typically for elder individuals. The individual is asked to complete a series of tasks, such as to copy a drawing, or answer questions, such as what month it is that day. The MMSE score is the number of tasks or questions that the individual correctly performs. This score is then categorised into a multi-level index to indicate the individual's cognitive ability. The categorised score is given in a variable called MMSE Index. For example, in the 30-point questionnaire, Tombaugh and McIntyre (1992) suggests the following category cutoffs for the MMSE Index:

$$\text{MMSE Index} = \begin{cases} \text{severe impairment,} & \text{if } MMSE \leq 9 \\ \text{moderate impairment,} & \text{if } 10 \leq MMSE \leq 18 \\ \text{mild impairment,} & \text{if } 19 \leq MMSE \leq 23 \\ \text{normal,} & \text{otherwise.} \end{cases}$$

A derived variable may be suitable to aid in modelling. In Lambert and Royston (2009) cubic splines are used when fitting a model to survival data. Lambert and Royston (2009) find splines advantageous over a Cox Proportional-Hazards model for multiple reasons; for example, splines are beneficial to model more complicated time-dependent effects, and aid in investigating the effects. Hence, derived variables are used when constructing cubic splines.

A derived variable may be used if a transformation is required in modelling. A variable in its raw form may violate an assumption when fitting a statistical model to

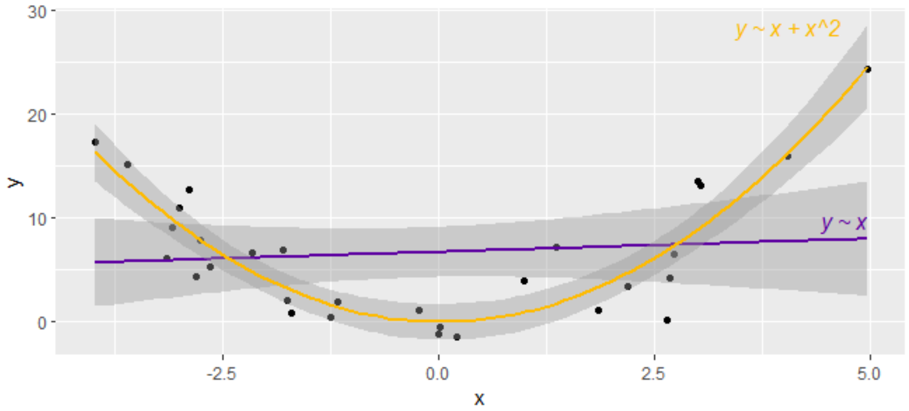


FIGURE 2.4: Fitting a $y \sim x$ and $y \sim x + x^2$ model to a non-linear relationship.

some data (van Holm, 2021). For example, if a linear model is fitted such that $y \sim x$ and the relationship between a predictor and the outcome variable is non-linear, then a transformation may be required, such as squaring the predictor: $y \sim x + x^2$. An example of fitting x in its raw form and its transformed form to a linear model is given in Figure 2.4. Alternatively, a variable may be transformed if the assumption of homoscedasticity is violated. The assumption of homoscedasticity is that the variance of the residuals is constant for all estimated values of the outcome variable, y, in the data set. If this assumption is violated, the variance would increase in the residuals as the fitted values change. One way to handle homoscedasticity is by square-rooting the outcome variable. Hence, one derived variable in a data set could be the square root of another variable. Another derived variable may occur in a data set if the effect of one variable, x_1, on the outcome variable, y, alters for different values of another variables, x_2. In this case, an interaction, x_1x_2, may be required when modelling the data. This interaction variable is another example of a functional form.

Two approaches to impute an incomplete derived variable are active and passive imputation. In active imputation, also known as "just another variable" (JAV) (White et al., 2011), a derived variable is imputed by means of an imputation model, like any other variable. As a result, the functional relationship between the constituents and the derived variable is lost for imputed values. However, active imputation does not suffer from the issue of incompatibility that was outlined in Section 2.2.3 (von Hippel, 2009).

In passive imputation, the constituents are imputed by an imputation model, and then the derived variable is constructed. Incompatibility can arise with passive imputation because the derived variable is not imputed using the variables in the substantive model (von Hippel, 2009). As a result, the estimated coefficients in the substantive model can be attenuated for the derived variable. However, passive imputation has the benefit that the functional relationship between the derived variable and its constituents is preserved. Literature comparing active and passive imputation will be discussed in Chapter 3.

2.4 Survival Analysis

In this section, some basic survival analysis methods are described. Survival analysis is then discussed in the context of MI, with modifications to the basic MI procedure highlighted.

2.4.1 Survival Analysis Methods

Survival analysis, or "time-to-event" analysis, is a set of statistical procedures that is used to analyse the time between entering a study and a predefined event of interest occurring (Collett, 2015). In survival analysis, the time variable is either the age of the participant when they experience the event, or the time duration between when an individual enters the study and experience the event. The event of interest could be, but is not limited to, recovery, relapse, or death. Furthermore, survival analysis is not limited to a medical setting; for example, the event of interest may be mechanical failure.

In survival analysis, the individual may not experience the event of interest in the study period. In this case, the survival time is censored. If an individual leaves the study without experiencing the event of interest then their survival time is said to be right censored. Right-censoring may occur, for example, if the individual drops out of the study. Let T_s denote the survival time for an individual, and T_c denote the censored time for an individual for random variables, $T_s, T_c \geq 0$. Define a random variable of an individual's survival time by T, $T \geq 0$, where

$$T = \min(T_s, T_c).$$

Then define by C a random variable that indicates whether an individual experienced the event in the study such that,

$$C = \begin{cases} 0, & \text{if } T = T_c \\ 1, & \text{if } T = T_s. \end{cases}$$

If an individual does not experience the event by the end of the study period, $C = 0$ and the individual's duration time at the end of the study is set as their survival time. A type-one right censored design is when all censored observations are at the end of the study period, that is, there are no drop outs during the study.

The actual survival time of an individual, T, is assumed to be independent of any mechanism that causes censoring for an individual. This assumption is known as non-informative censoring, and hence the censored observations in the survival time variable are essentially MCAR. The survival status for an individual is given by (T, C).

Several functions useful in survival analysis are discussed below, along with two approaches to modelling survival data. The material in this section is based on Collett (2015), Klein et al. (2014), and Lawless (2011).

Survivor Function Let T be a continuous random variable on $(0, \infty)$, then t is an observed value of T, $t > 0$. T has a probability density function (PDF), $f(t)$. The cumulative distribution function of T is defined as

$$F(t) = P(T < t) = \int_0^t f(u)du. \tag{2.1}$$

$F(t)$ is the probability that an individual experiences the event of interest before time t. The survivor function, $S(t)$, is the probability that an individual does not experience the event of interest before time t, $t \geq 0$. Hence, $S(t)$ is defined as:

$$S(t) = P(T \geq t) = 1 - F(t). \tag{2.2}$$

At the beginning of the study, when $t = 0$, $S(t) = S(0) = 1$. Hence, the probability of not experiencing the event at the start of the study is unity.

Hazard Function The hazard function, $h(t)$, the instantaneous risk that an individual experiences the event between time t and $t + \delta t$, given that they have not previously experienced the event. It is defined as

$$h(t) = lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t}, \tag{2.3}$$

where δt is the width of the time interval. A standard result in probability theory is that conditional probability is the probability that event A occurs, given that event B has already occurred, given by $P(A|B) = \frac{P(A \cap B)}{P(B)}$. $P(A \cap B)$ is the joint probability of A and B occurring, and $P(B)$ is the probability that event B occurs. Hence, (2.3) can be rearranged to

$$lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t)}{\delta t \times P(T \geq t)}$$

$$= lim_{\delta t \to 0} \frac{P(T < t + \delta t) - P(T < t)}{\delta t \times P(T \geq t)}.$$

Using (2.1) and (2.2),

$$= lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)}$$

$$= \frac{dF(t)}{dt} \frac{1}{S(t)}$$

$$= \frac{f(t)}{S(t)}$$

Hence,

$$h(t) = f(t)/S(t). \tag{2.4}$$

Cumulative Hazard Function The risk of an event occurring can be accumulated over time to give the cumulative hazard function, $H(t)$. It is defined as

$$H(t) = \int_0^t h(u)du. \tag{2.5}$$

By (2.4) and (2.2),

$$H(t) = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{f(u)}{1-F(u)}du$$
$$= -\log(1-F(t))$$

Hence, the cumulative hazard function can be shown to satisfy

$$H(t) = -\log(S(t)). \tag{2.6}$$

Additionally,

$$S(t) = \exp(-H(t)).$$

Non-Parametric Estimators The survivor, hazard, and cumulative hazard functions are usually unknown in practice, but can be estimated.

One estimate of $S(t)$ is the Kaplan-Meier estimate, $\hat{S}(t)$ (Collett, 2015). Let there be n individuals with r distinct event times, note that $r \leq n$ because multiple events may occur at the same time point, or may be censored. Arrange the r event times in ascending order to give r ordered event times: $t_1 < t_2 < ... < t_r$. Denote by t_j the j^{th} ordered event time, $j = 1, ..., r$. Then define by d_j the number of events that occur at t_j. Denote by n_j the number of individuals who have not experienced the event just before t_j, at $t_j - \delta$, for an infinitesimal time interval, δ. Hence, n_j is the sum of d_j and the number of individuals who have not experienced the event at t_j. n_j includes any censored individuals at t_j.

The estimated probability that an individual does not experience the event in the time interval $t_j - \delta$ to t_j is $\frac{n_j - d_j}{n_j}$. The estimated probability that an individual does not experience the event in the time interval t_j to $t_{j+1} - \delta$ is one since $d_j = 0$. Hence, the joint probability that an individual does not experience the event in the time interval $t_j - \delta$ to $t_{j+1} - \delta$ is estimated by $\frac{n_j - d_j}{n_j}$.

Construct r time intervals, such that (t_k, t_{k+1}) is the k^{th} constructed time interval, $k = 1, ..., r$, where $t_{r+1} = \infty$. The estimated probability that an individual does not experience the event before t_{k+1} is the probability that the individual does not experience the event in the time interval (t_k, t_{k+1}) or in the intervals leading up to t_k. This leads to the Kaplan-Meier estimate of the survivor function, given by

$$\hat{S}(t) = \prod_{j=1}^{k} \Big(\frac{n_j - d_j}{n_j}\Big)$$

for $t_k \leq t < t_{k+1}, k = 1, ..., r$. If $t < t_1$, $\hat{S}(t) = 1$.

The Kaplan-Meier estimate of a cumulative hazard function can be derived using (2.6). Hence,

$$\hat{H}(t) = -\sum_{j=1}^{k} \log \Big(\frac{n_j - d_j}{n_j}\Big)$$

for $t_k \leq t < t_{k+1}, k = 1, ..., r$.

The Kaplan-Meier estimate is derived to estimate the survivor function. Another estimate of the survivor function is the Nelson-Aalen estimate. The Nelson-Aalen estimate of the survivor function is defined as

$$\tilde{S}(t) = \prod_{j=1}^{k} \exp \Big(-\frac{d_j}{n_j}\Big)$$

for $t_k \leq t < t_{k+1}, k = 1, ..., r$. By (2.6), the estimated cumulative hazard function is given by

$$\tilde{H}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j}$$

for $t_k \leq t < t_{k+1}, k = 1, ..., r$.

A Kaplan-Meier estimate approximates a Nelson-Aalen estimate. By definition,

$$\exp(-x) = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} + ...,$$

then if x is small, $\exp(-x) \approx 1 - x$. Therefore,

$$\exp(-d_j/n_j) \approx 1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j},$$

provided that d_j/n_j is small. Hence, $\tilde{S}(t) \approx \hat{S}(t)$. Since $\exp(-x) > 1 - x$, the Nelson-Aalen estimate will be larger than the Kaplan-Meier estimate . However, both the Nelson-Aalen estimate and Kaplan-Meier estimate will be very similar.

Cox Proportional-Hazards Model The hazard of the event occurring for an individual can be modelled by a Cox Proportional-Hazards model, first introduced by Cox (1972). To build the Cox Proportional-Hazards model, first assume proportional hazards. That is,

$$\lambda_j = c_j \lambda_0$$

for an individual, j, where λ_0 is an unspecified hazard function and c_j is a constant. To model the effect of the covariates on the hazard, set

$$c_j = \exp(\boldsymbol{\beta}' \boldsymbol{x}_j),$$

where $\boldsymbol{\beta}' = (\beta_1, ..., \beta_q)$ is a vector of coefficients for q explanatory variables, and $\boldsymbol{x}_j = (x_{1j}, ..., x_{qj})$ is a vector of covariate values for the j^{th} individual. This value of c_j then gives rise to the Cox Proportional-Hazards model, defined by

$$h_j(t) = h_0(t) \exp(\boldsymbol{\beta}' \boldsymbol{x}_j). \tag{2.7}$$

The Cox Proportional-Hazards model factorises into two components. The baseline hazard function, $h_0(t)$, is the hazard when all covariates are set to zero. The second component accounts for the impact of the covariates on the survival time.

A hazard ratio, ψ, can be calculated from the Cox Proportional-Hazards model to give the effect of a variable on the survival time. Let x_k denote a factor with two levels, 0 and 1, then the hazard ratio is given by

$$\psi_k = \frac{h(t, x_k = 1)}{h(t, x_k = 0)} = \frac{h_0(t) e^{\beta_1 x_1 + ... + \beta_k \times 1 + ... + \beta_q x_q}}{h_0(t) e^{\beta_1 x_1 + ... + \beta_k \times 0 + ... + \beta_q x_q}} = e^{\beta_k}.$$

Given that all other variables are held constant, an individual with $x_k = 1$ has $\psi_k = e^{\beta_k}$ times the hazard of experiencing the event than an individual with $x_k = 0$ in a given time period. Hence if $\psi_k > 1$ then $h_j(t, x_k = 1) > h_j(t, x_k = 0)$. That is, the hazard of experiencing an event at t is larger for an individual with $x_k = 1$ relative to an individual with $x_k = 0$. Conversely, if $\psi_k < 1$ then the hazard of experiencing an event at t is smaller for an individual with $x_k = 1$ relative to an individual with $x_k = 0$. Alternatively, if x_k were a continuous variable, ψ_k is the change in the hazard of the event occurring per unit increase; for example, if x_k were age given in years then ψ_k is the change in the risk of the event for each year increase, given all other variables are held constant.

Parametric Proportional Hazards Model In the Cox Proportional-Hazards model, the baseline hazard, $h_0(t)$, is non-parametric. However, a parametric proportional hazards model can be specified.

If a parametric model has been specified with a PDF, then the survivor, hazard, and cumulative hazard functions can be derived using the relationships in (2.2), (2.4), (2.6). The derived hazard function, $h(t)$, can then be substituted into the proportional hazards model given in (2.7) to give a parametric proportional hazards model.

For example, a Weibull distribution has shape, γ, and scale, λ. The PDF of a Weibull distribution is

$$f(t) = \lambda \gamma t^{\gamma - 1} \exp(-\lambda t^{\gamma}),$$

$t \geq 0$. When $\gamma = 1$, the Weibull distribution is an exponential distribution. Hence, the PDF of an exponential distribution is

$$f(t) = \lambda \exp(-\lambda t),$$

$t \geq 0$. Using (2.1),

$$\begin{aligned} F(t) &= \int_0^t f(u)du \\ &= \int_0^t \lambda \exp(-\lambda u)du \\ &= \Big[-\frac{1}{\lambda}\lambda \exp(-\lambda u)\Big]_0^t \\ &= 1 - \exp(-\lambda t). \end{aligned}$$

Then, by (2.2),

$$S(t) = 1 - F(t) = \exp(-\lambda t). \tag{2.8}$$

By (2.4),

$$h(t) = f(t)/S(t) = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda. \tag{2.9}$$

Finally, by (2.6),

$$H(t) = -\log(S(t)) = -\log(\exp(-\lambda t)) = \lambda t, \tag{2.10}$$

$t \geq 0$. Therefore, the survivor, hazard, and cumulative hazard functions for an exponential distribution are given by (2.8 - 2.10).

Similarly, the survivor, hazard, and cumulative hazard function for a Weibull distribution are

$$S(t) = \exp(-\lambda t^{\gamma}) \tag{2.11}$$

$$h(t) = \lambda \gamma t^{\gamma - 1} \tag{2.12}$$

$$H(t) = \lambda t^{\gamma}. \tag{2.13}$$

Substituting the derived hazard function for a Weibull distribution in (2.12) into the proportional hazards model given in (2.7) yields a Weibull parametric proportional hazards model:

$$h_j(t) = \lambda \gamma t^{\gamma - 1} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j). \tag{2.14}$$

Accelerated Failure Time Model An alternative approach to model survival data is to use an Accelerated Failure Time (AFT) Model. An AFT model is defined by

$$h_j(t) = e^{-\eta_j} h_0(t/e^{-\eta_j}).$$

where $\eta_j = \boldsymbol{\alpha}' \boldsymbol{x}_j$ for individual j, $j = 1, ..., n$, and $\boldsymbol{x}_j$ is a vector of covariate values for the j^{th} individual. $\boldsymbol{\alpha}' = (\alpha_1, ..., \alpha_q)$, where $\alpha_1, ..., \alpha_q$ are the coefficient values of the q

covariates. In an AFT model, the covariates are assumed to act multiplicatively on time to the event occurring. Therefore, the effect of the covariates is to accelerate or decelerate the time until an individual experiences the event.

Define the survival time of individual, j, in a data set by random variable T_j. Then the AFT model can be rearranged to take a linear form:

$$\log(T_j) = \mu + \alpha_1 x_{1j} + ... + \alpha_q x_{qj} + \sigma\epsilon_j = \mu + \boldsymbol{\alpha}^{'}\boldsymbol{x}_j + \sigma\epsilon_j \tag{2.15}$$

where μ denotes the intercept, σ is a scale term, and ϵ_j is an error term. $\boldsymbol{\alpha}^{'}$ is a vector of the q coefficient values. If $\alpha_i > 0$, then the survival time of an individual increases with increasing values of the corresponding x_{ij} variable.

The distribution of ϵ_j determines the distribution of T_j; for example, if ϵ_j is Gumbel distributed then T_j is Weibull-distributed. A Gumbel distribution is an extreme value distribution with a location parameter, μ, and scale parameter, $\delta > 0$. A standard Gumbel distribution has $\mu = 0, \delta = 1$. If ϵ_j follows a standard Gumbel distribution, the survivor function of ϵ_j is,

$$S_{\epsilon_j}(\epsilon) = \exp(-\exp(\epsilon_j)). \tag{2.16}$$

Suppose that a random variable, T_j, follows a Weibull distribution,

$$T_j = \exp(\mu + \boldsymbol{\alpha}^{'}\boldsymbol{x}_j + \sigma\epsilon_j).$$

Rearranging the AFT model in (2.15) and substituting it into the survivor function of ϵ_j given in (2.16) results in the survivor function of T_i:

$$\begin{aligned} S_j(t) &= \exp\Big(-\exp\Big(\frac{\log(t_j) - \mu - \alpha_1 x_{1j} - ... - \alpha_q x_{qj}}{\sigma}\Big)\Big) \qquad (2.17) \\ &= \exp(-\mu_j t^{1/\sigma}) \end{aligned}$$

where $\mu_j = \exp(-(\mu + \alpha_1 x_{1j} + ... + \alpha_q x_{qj}\ \sigma))$.

Using (2.3),

$$\begin{aligned} h_j(t) &= \frac{1}{\sigma t}\exp\Big(\frac{\log(t_j) - \mu - \alpha_1 x_{1j} - ... - \alpha_q x_{qj}}{\sigma}\Big) \\ &= \mu_j \sigma^{-1} t^{\sigma^{-1}-1}. \qquad (2.18) \end{aligned}$$

Using (2.5)

$$\begin{aligned} H_j(t) &= -\log(S_j(t)) \\ &= \exp\left(\frac{\log(t_j) - \mu - \alpha_1 x_{1j} - ... - \alpha_q x_{qj}}{\sigma}\right) \end{aligned}$$

$$= \mu_j t^{1/\sigma}. \tag{2.19}$$

(2.17), (2.18), (2.19) are the survivor, hazard, and cumulative hazard functions respectively of a Weibull distribution with scale, μ_j, and shape, σ^{-1}. By comparing the survivor, hazard, and cumulative hazard function in the Weibull proportional hazards model (2.11-2.13) to that of the Weibull AFT model (2.17-2.19), it is observed that the parameters in the Weibull proportional hazards model in (2.14), λ, γ, β_i, have equivalents in the AFT model, μ, σ, α_i:

$$\lambda = \exp(\mu/\sigma),$$

$$\gamma = 1/\sigma,$$

$$\beta_i = -\alpha_i/\sigma.$$

If additionally $\sigma = 1$ then T_j follows an exponential distribution. The survivor, hazard, and cumulative hazard function of an exponential distribution respectively, with scale μ_j, are:

$$S_j(t) = \exp(-\mu_j t),$$

$$h_j(t) = \mu_j.$$

$$H_j(t) = \mu_j t,$$

Hence the parameters in the exponential proportional hazards model, λ, β_i, have equivalents in the AFT model, μ, α_i

Therefore, the Weibull distribution (and the special case of an exponential distribution if $\sigma = 1$) can be parameterised to be both a proportional hazards model and an AFT model.

2.4.2 Specifications for Multiple Imputation

In survival data, the outcome comprises two variables: survival time and a status indicator. Omitting the outcome variable in the imputation model can result in biased estimated coefficients in the substantive model (See Section 2.2.3). In this section, the literature on the inclusion of survival time and the status indicator in the imputation model is explored.

HippisleyCox et al. (2007) model the risk of cardiovascular disease in which some covariates are imputed by MI. One incomplete variable is the cholesterol ratio. In the initial publication, the status indicator was omitted from the imputation models, resulting in a biased estimated coefficient for the imputed cholesterol variable in the substantive model ($\psi = 1.001$). After subsequent research suggested a strong relationship between the effect of cholesterol on the incidence of cardiovascular disease, these findings were flagged. The imputation model was subsequently modified to include the status variable as a predictor, and the effect of cholesterol on survival time in the substantive model was in accordance with findings elsewhere in the literature (for women, $\psi = 1.170$, for men $\psi = 1.195$) (Hippisley-Cox et al., 2007).

Further to this, high variability in the denominator can additionally cause issues in the MI procedure (HippisleyCox et al., 2007). These issues occur because the imputed values can be close to zero, resulting in an unstable imputation of the derived variable and essentially removing the association with survival time. Morris et al. (2014) explored multiple imputation with a derived variable as cholesterol ratio, and suggests using active imputation if the coefficient of variation of the denominator of the ratio is larger than 0.1.

In addition to recommending active imputation if the coefficient of variation of the denominator of a ratio is larger than 0.1, Morris et al. (2014) additionally recommend active imputation if all the constituents that construct the derived variable are missing when the derived variable is missing. Finally, Morris et al. (2014) find that passive imputation performs well for a ratio if it has been first log-transformed. This is because the imputation model and substantive model are both linear: Let $x = \frac{\gamma_1}{\gamma_2}$ for constituents γ_1, γ_2. Then,

$$(\log(\gamma_1), \log(\gamma_2)|y) \sim N$$

Since $\log(x) = \log(\gamma_1) - \log(\gamma_2)$,

$$(\log(x)|y) \sim N$$

Therefore, the imputation mean function is

$$\log(x) = \alpha_0 + \alpha_1 y$$

and the substantive model mean function is

$$y = \beta_0 + \beta_1 x.$$

That is, the imputation model is in the correct form when the constituents are first log-transformed.

A simulation study by White and Royston (2009) finds that omitting the survival time from the imputation model results in very biased estimated coefficients in the substantive model. Additionally, omitting survival time as a predictor in the imputation model can result in undercoverage for the estimated coefficients for the imputed variable. Of the different imputation models considered, White and Royston (2009) find that the substantive model results in the smallest bias when both the status variable and a Nelson-Aalen estimate of the cumulative baseline hazard are present in the imputation model.

Clark and Altman (2003) and HippisleyCox et al. (2007) use a logarithmic transformation of survival time as a predictor in the imputation model. Van Buuren et al. (1999) include survival time both in its raw form and transformed logarithmically, together with status, to account for any “multiplicative relations” in later analyses, while Pankhurst et al. (2020) include time and status as covariates without transformation.

Clark and Altman (2003), HippisleyCox et al. (2007), Van Buuren et al. (1999), and Pankhurst et al. (2020) use different approaches to include the response variables in the imputation model under survival analysis. However, the consistent message is that the status and some function of the survival time should be included as predictors in the imputation model.

Chapter 3

Literature Review

In previous simulation studies, the performance of active and passive imputation has been investigated. These previous studies are reviewed in this chapter. The results from the simulation studies reviewed in this chapter are additionally summarised in Table 3.1.

In this chapter, the methods undertaken by previous researchers are evaluated. This is firstly to inform the procedures and methods undertaken in the simulation study performed in Section 5.1, and to determine where uncertainties lie in previous research that can be addressed in the simulation study.

In the first section, the study designs for the simulation studies reviewed in this chapter are discussed. The performance of active and passive imputation in the literature is then reviewed. Following this, simulation studies where modifications have been applied to enhance the performance of active and passive imputation models are outlined. Finally, the research questions that will be addressed by the simulation study are given.

3.1 Design of other Simulation Studies

In the literature, the performance of active and passive imputation is often investigated using a simulation study. Across the simulation studies discussed in this chapter, M generally ranges between five and 50, but M is high as 500 in studies with unknown actual values of a derived variable (White et al., 2011). In most simulation studies discussed in this chapter, $M = 5$.

The number of replications ranges from 100 to 5000 (Morris et al., 2014), with 1000 being the most common number of replicates. The smallest data set has 150 observations (Eekhout et al., 2018), and the largest has 5000 observations (Mitani et al., 2015).

Desai et al. (2016) note in their simulation study that the availability of auxiliary variables may change the behaviour of passive and active imputation. The presence of auxiliary variables may favour a certain environment; for example Wagstaff et al. (2009) compare active to passive imputation for a ratio functional form, BMI. One of the three auxiliary variables, waist measurement, is highly correlated with BMI and weight, but not height. One could argue that active imputation outperforms passive imputation when waist measurement is present, but not otherwise. Therefore, in the simulation study performed in this thesis, it is of interest to investigate how an auxiliary variable affects the performance of both active and passive imputation.

Auxiliary variables are frequently omitted in some simulation studies with generated data; see, for example, Seaman et al. (2012), Morris et al. (2014), Tilling et al. (2016). However, in a real-world setting auxiliary variables can be present. For example, the CLHLS data set outlined in Section 1 has 5077 variables. Some of these variables may be useful predictors for the partially observed variables, but most of them would not be in a substantive model. Hence, some variables may be suitable auxiliary variables. In addition to this, Hardt et al. (2012) show that the auxiliary variables in an imputation model both decrease the bias in the estimated coefficients and increase the precision, unless the auxiliary variables are not strongly correlated with the partially observed variable, or, when there are too many auxiliary variables present (in their data set, this is 48 auxiliary variables). Jochen et al. (2013) support the notion that too many auxiliary variables can be detrimental to the imputation procedure, noting that the relationships between variables result in a negative bias. Auxiliary variables which are good predictors of the missing values should be included. If an auxiliary variable also predicts missingness, then they may reduce bias if they also predict the actual missing values. As a result, including an auxiliary variable is one aspect that can be explored in future simulation studies such as the simulation performed in this thesis.

The simulation studies discussed in this section usually evaluate the performance of active and passive imputation by considering the bias in the estimated coefficients in the substantive model and the coverage rate. Additional metrics to evaluate the imputation procedure include the Monte Carlo standard errors (Seaman et al., 2012), mean squared errors in (Pankhurst et al., 2020; Mitani et al., 2015), and relative efficiency (Morris et al., 2014).

3.2 Active and Passive Imputation

von Hippel (2009) investigates the performance of active and passive imputation by calculating the covariance matrix with the help of the MathStatica package. Through these calculations, von Hippel (2009) demonstrate that the sample mean and covariance matrix for the data before missing values are generated are the same as an actively imputed data set under a MCAR and MAR missingness mechanism. However, the estimated covariance matrix calculated from a passively imputed data set is different from the covariance matrix of the complete data set. This difference in the underlying covariance structures results in a substantive model that is more biased when the data have been passively imputed rather than actively imputed. Seaman et al. (2012) extend on von Hippel (2009) to show through a mathematical approach that there is some bias in the estimated pooled coefficients under a MAR structure if a square or interaction variable is actively imputed. However, Seaman et al. (2012) conclude that passive imputation still results in more biased coefficient estimates than active imputation for both functional forms. A simulation study by White et al. (2011) supports the notion that active imputation can result in biased estimated coefficients in the substantive model for a squared functional form under a MAR structure. However, active imputation still outperforms standard linear passive imputation in the simulation study conducted by White et al. (2011).

The coefficient estimates after active imputation can be less biased than passive imputation for other functional forms when MICE is performed. For example, in Grobler and Lee (2020) the derived variable is dichotomised from a continuous variable. The coefficient estimates are biased, and there is severe undercoverage when passive imputation under MICE is performed. For both the MCAR and MAR structures, active imputation outperforms passive imputation under MICE. Additionally, Desai et al. (2016) run a simulation study under a MAR structure. The derived variable is the rate of change across repeated measurements, and the substantive model is a two-stage linear regression model. Desai et al. (2016) recommend active imputation since the coefficient estimates are less biased, and because the approach of active imputation is computationally simpler and widely offered across many statistical packages, unlike passive imputation.

However, there are instances in the literature where active imputation does not outperform passive imputation under functional forms not yet discussed. Eekhout et al. (2018) perform a simulation study in which the derived variable takes an additive functional form and has a MAR missingness mechanism. They consider bias and precision in the coefficient estimates from the pooled substantive model to recommend passive imputation. Ling et al. (2019) perform a simulation study where the derived variable is a propensity score. The simulation is repeated for a MCAR, MAR, and a MNAR structure. Through evaluating the bias, variance, MSE, and coverage rate of the coefficient estimates, they recommend the use of passive imputation. Furthermore, Van Buuren (2018) performs a simulation study for a derived variable with a ratio functional form that is MCAR. A linear regression model is fitted to the imputed data. They recommend passive imputation despite some bias in the coefficient estimates.

The form of the substantive model might affect the performance of active imputation and passive imputation. For example, Morris et al. (2014) fit a Cox Proportional-Hazards substantive model after imputing a ratio functional form by active and passive imputation. After calculating the coefficient estimates and width of the resulting confidence intervals for the imputed variable in the substantive model, active imputation outperforms standard passive imputation. Additionally, Wagstaff et al. (2009) investigate a ratio functional form for both MCAR and MAR missingness structures. Through evaluating the means of the imputed values under active and under passive imputation, they conclude that active imputation outperforms passive under MAR, but the two perform comparably under MCAR. Hence, overall for a ratio functional form, passive imputation outperforms active imputation when the substantive model is a linear regression model, but active imputation outperforms passive imputation when considering estimates under a Cox Proportional-Hazards substantive model.

The observation that the substantive model may affect the performance of the imputation model is not limited to a ratio functional form. In a simulation study conducted by Seaman et al. (2012) where the derived variable takes a square functional form, active imputation outperforms passive under a substantive linear regression model. However, active imputation does not outperform passive imputation for a substantive logistic regression model. Additionally, Jochen et al. (2013) find that passive imputation outperforms active imputation if the substantive model follows logistic regression and the functional form is an interaction. An overall summary of the different simulation studies is given in Table 3.1, where different functional forms are given in the columns and different substantive models in the rows. Additionally, the imputation model that resulted in the best performance is stated, alongside the paper and missingness structure. Overall, one of active and passive imputation may outperform the other depending on the substantive model.

3.3 Modifications to Imputation Models

In some simulation studies, modifications to the imputation models have improved the performance of both active and passive imputation models. Morris et al. (2014) consider a ratio functional form for a Cox Proportional-Hazards substantive model. A variant of passive imputation is used whereby the constituents are log-transformed before imputation, resulting in an additive functional form. The estimated coefficients after applying a log-transformed passive imputation model are less biased and the confidence intervals have a smaller average width than a non-transformed passive imputation model to the extent that Morris et al. (2014) recommends the use of a log-transformed passive imputation model. Morris et al. (2014) further run a simulation study comparing active and passive imputation for both MCAR and MAR structures. Overall, passive imputation after a log-transform has the best coverage rate of the imputation models, and has less biased estimated coefficients than a standard passive imputation model.

The performance of passive imputation is found to be enhanced by using PMM for a ratio form with a Cox Proportional-Hazards substantive model (Morris et al., 2014), a square form with both a logistic regression substantive model (Seaman et al., 2012) and a linear regression substantive model (Seaman et al., 2012; White et al., 2011), and an interaction functional form with a linear regression substantive model (Seaman et al., 2012). The simulation studies where it is concluded that PMM enhances passive imputation have data in MCAR or MAR structures in Seaman et al. (2012) and just a MAR structure in White et al. (2011). However, while applying PMM reduces bias in coefficient estimates after passive imputation, it does not eradicate it (Seaman et al., 2012; White et al., 2011). Seaman et al. (2012) find that the coefficient estimates after applying active imputation are less biased than those of passive imputation under PMM when the derived variable has a square functional form and is MAR. Note that this is only the case for a linear regression substantive model. In a logistic regression model, the coefficient estimates after applying active imputation are more biased than those of passive imputation under PMM. Furthermore, White et al. (2011) find that passive imputation under PMM performs comparably with active imputation under a MAR structure.

Further enhancements to passive imputation frequently occur when the derived variable has an interaction functional form. Since the actual interaction in the substantive model cannot be a predictor in the imputation model, there can be an issue of incompatibility. To reduce bias due to incompatibility, a new interaction is included in the imputation model between one of the main effects and the outcome variable. This is called an "improved passive" model and is investigated by White et al. (2011), Seaman et al. (2012), and Mitani et al. (2015). White et al. (2011) show the improved passive model reduces the bias in the coefficient estimates further than standard passive imputation, but the coefficient estimates from active imputation is still less biased than the improved passive models. Seaman et al. (2012) find that under the improved passive imputation model, the coefficient estimates are less biased and the coverage rate is closer to 95% than both regular passive imputation, and passive imputation under PMM. Mitani et al. (2015) use a MAR structure for both a binary and continuous outcome in two separate simulation studies where the functional form of the derived variable is an interaction. They consider the improved passive imputation model when the main effects are categorical variables, and overall conclude that the bias in the coefficient estimates is smaller when an improved passive imputation model is applied instead of either a standard passive imputation model or an active imputation model.

Another variation when handling interactions is “grouped passive” imputation (White et al., 2011; von Hippel, 2009; Tilling et al., 2016). This modification of passive imputation is applied when at least one constituent is a discrete complete variable. Let x_1 have L strata. Impute missing values in other covariates where $x_1 = 1$, then repeat where $x_1 = 2, ..., x_1 = L$. White et al. (2011) find under a MCAR structure that this modification reduces bias in the coefficient estimates more than the “improved passive” imputation model, allowing the performance of the “grouped passive” imputation model to be comparable to active imputation. Tilling et al. (2016) investigate applying a “grouped passive” imputation model to a continuous complete variable. As a result, Tilling et al. (2016) split x_1 at the mean to form two strata. However, this approach results in undercoverage of the coverage rate.

Slade and Naylor (2020) apply other modifications of passive imputation when handling an interaction term under a MAR structure. Consider an outcome variable, y, two main effects, x_1 and x_2, and an interaction, $x_1 * x_2$. In an “improved passive” model, there are three interactions passively imputed in the imputation phase: $x_1 * x_2$, $x_1 * y$, and $x_2 * y$. The coefficient estimates are less biased when the additional interaction terms are included in the imputation model when compared to standard passive imputation. This is in accordance with the findings of Tilling et al. (2016) who find that the modified version of the “improved passive” imputation model performs comparably to the “grouped passive” imputation model. However, the coefficient estimates are still more biased than active imputation.

Substantive Model Compatible Fully Conditional Specification (SMCFCS) is a modification to MICE where the passive imputation model is not incompatible to the substantive model. Van Buuren (2018) conclude that SMCFCS results in the least biased coefficient estimates in their simulation study. Similarly, Grobler and Lee (2020) find that passive imputation under MICE results in biased estimated coefficients and severe undercoverage. However, applying SMCFCS to passive imputation improves the imputation procedure to the extent that passive imputation under SMCFCS outperforms active imputation under MICE.

Variations of active imputation have been investigated in the literature. Morris et al. (2014) impute the derived variable with the constituents as predictors, whereas in Wagstaff et al. (2009) the constituents are not predictors for active imputation. In Grobler and Lee (2020), both of these approaches are investigated in separate imputation models. In Mitani et al. (2015) the constituents are present when imputing the derived variable, and the derived variable is a predictor in the imputation models for the constituents. Slade and Naylor (2020) consider an interaction functional form and investigate different active imputation models: where the derived variable is present for imputing the constituents; where the derived variable is not present for imputing the constituents; and two additional variations where there are interactions in the imputation models for the constituents and derived variables. The variants of active imputation models where there are not interactions as predictors in the imputation model outperform variants that include interactions (Slade and Naylor, 2020): the coefficient estimates are less biased and coverage rates are closer to 95%.

3.4 Conditions that affect Imputation Models

Variations of passive and active imputation can alleviate the bias of the estimated coefficients in the substantive model or improve the coverage rate. Passive imputation ensures plausible values are imputed for the derived variable, but can have the issue of incompatibility between the imputation model and substantive model. There are approaches to lessen or eradicate the problems that arise from imputing by passive imputation; for example, applying PMM, using SMCFCS instead of MICE, or transforming the constituents where appropriate. While active imputation does not have the issue of incompatibility, the imputed derived variable loses the functional relationship.

Further, for a given functional form the performance of the imputation model often changes depending on the substantive model. For example, active imputation outperforms passive imputation under linear regression but not under logistic regression (Seaman et al., 2012). Similarly, conclusions about imputing a ratio differ when the substantive model is a linear regression model against when it is a Cox Proportional-Hazards model. There is little research on the use of different functional forms with survival analysis. Hence, the topic of imputation for functional forms of derived variables in survival analysis is investigated in a simulation study in Section 5.

Finally, the performance of active and passive imputation differs depending on the functional form of the derived variable. Therefore, different functional forms are considered in the simulation study.

TABLE 3.1: Summary of the preferred imputation model from the literature, with columns displaying the functional form considered, and rows the substantive model applied.

		Functional Form				
		Square	Interaction	Additive	Ratio	Other functional forms
Substantive Model	Linear	**Active** (MCAR/MAR von Hippel (2009); Seaman et al. (2012)). **Either** (MCAR, modified passive or **active** White et al. (2011)).	**Active** (MCAR/MAR von Hippel (2009); Seaman et al. (2012)). **Passive** (MAR, modified passive or **either** Mitani et al. (2015)). **Either** (MCAR, modified passive or **active** White et al. (2011)). **Active** (MAR. With both modified Slade and Naylor (2020)).	**Passive** (MAR, Eekhout et al. (2018)).	**Passive** (MCAR, Van Buuren (2018)).	**Active** (Rate of change, MAR, Desai et al. (2016)). **Passive** (Propensity score, MCAR/MAR, Ling et al. (2019)). **Passive** **(SMCFCS)** (Dichotomising a continuous variable, MCAR/MAR, Grobler and Lee (2020).
	Logistic	**Passive** (MCAR, modified passive or **active** Seaman et al. (2012)). **Passive** (MAR. Seaman et al. (2012)).	**Passive** (MCAR Jochen et al. (2013)). **Passive** (MAR, IMP otherwise **either** Mitani et al. (2015)).			
	Cox PH				**Passive** (MAR/MCAR, modified passive or **active** Morris et al. (2014)).	
	No model				**Active** (MAR. Wagstaff et al. (2009)). **Either** (MCAR. Wagstaff et al. (2009)).	

Chapter 4

Preliminary Study

A preliminary analysis is carried out to investigate the performance of active and passive imputation. Three functional forms are investigated: ratio, additive, and index. The kidney transplant motivating data set is first outlined in detail in this chapter, followed by other data sets used to investigate a ratio, additive, and index functional form. The study design, methods, and results of the preliminary analysis are given in this chapter, followed by a conclusion of the results and how these results will inform the simulation study.

4.1 Motivating Data Sets

In the preliminary analysis, actively and passively imputing a ratio, additive, and index functional form are investigated in real world data sets. In this section, three real world data sets imputed in the preliminary analysis are described.

4.1.1 Transplant Data Sets

The performance of active and passive imputation for a ratio functional form is investigated using two data sets supplied by NHS Blood and Transplant. One data set investigated is the kidney transplant data introduced in Section 1.

A second data set is investigated to explore the differences in the preliminary analysis for two data sets, and additionally support, or otherwise explain, results in the preliminary analysis. This second set concerns survival times following a cardiothoracic transplant. These transplants took place between 1995 and 2019 with survival times followed to mid-2019. The data contains information on 5520 patients and 28 variables, with the variables given in more detail in Appendix C. In addition, the derived variables in the cardiothoracic transplant data are donor BMI and recipient BMI, just as the kidney transplant data set, where BMI is the ratio between an individual's weight in kilograms and height in metres, squared:

$$\text{BMI} = \frac{\text{weight}}{\text{height}^2}.$$

Both transplant data sets are used to investigate active and passive imputation for a ratio functional form.

4.1.2 CLHLS Data Sets

The Chinese Longitudinal Healthy Longevity Study (CLHLS) are a series of surveys that monitors the quality of life and health of elderly individuals (CLHLS, 2018). The results from these surveys are split into different data sets. One such data set, 'ICPSR 36692', contains 9093 participants and 5077 variables. The participants responded to a baseline survey in 1998 with questions on their demographics and characteristics, family, lifestyle, and health, amongst other items. Most of these questions were repeated in follow up surveys every two to three years until 2014.

The variables selected for analysis in the preliminary study are the variables used in a CLHLS study analysed by Lagona and Zhang (2010) (ICPSR 3891). These variables, outlined further in Appendix C, are ID, age, gender, residence, exercise level, limitations in activities in daily living (ADL), cognitive ability (MMSE), survival time, and a censoring indicator. The survival time and censoring indicator variables are taken from the most recent survey in 2014. ADL and MMSE are both derived variables, so the constituents for ADL and MMSE are additionally selected. Detail on the derived variables are given next. Only the complete cases are considered, resulting in 8900 observations.

ADL A set of questions in the study help measure the participant's limitations in their daily activity. This measurement is resolved from the participant's ability to perform six tasks: bathe themselves ("bathing"), dress themselves ("dressing"), use the toilet without help ("toileting"), transfer around indoors without help ("transferring"), whether they are continent ("continence"), and feed themselves ("feeding"). Each task constitutes a binary variable, where a zero denotes an inability to perform it without help, and a one is given otherwise. From this a derived variable with an additive functional form, ADL, is formed. ADL takes a value ranging from zero to six:

$$\text{ADL} = \text{bathing} + \text{dressing} + \text{toileting} + \text{transferring} + \text{continence} + \text{feeding}.$$

ADL can additionally be categorised into three levels: no limitations, one limitation, or two or more limitations, producing a derived variable, ADL Index:

$$ADL\_Index_i = \begin{cases} 0, & \text{if } ADL_i = 0 \\ 1, & \text{if } ADL_i = 1 \\ 2+, & \text{otherwise} \end{cases}$$

The boundaries applied for ADL Index in this report are the boundaries used in Lagona and Zhang (2010), though alternative definitions of ADL Index are considered elsewhere; see, for example, Yi and Vaupel (2002).

MMSE Each participant is given a series of questions called a Mini-Mental State Examination (MMSE), a well-established method to determine cognitive ability (Pangman et al., 2000). There are 23 questions to determine the individual's cognitive ability with each question scoring one point if answered correctly and zero otherwise. Hence, each question has a binary variable associated with it. The binary variables can be categorised into five groups: orientation, registration, attention and calculation, recall, and language. The total value from these questions provides a MMSE score out of 23. MMSE score is one example of a derived variable with an additive functional form. In addition, MMSE can be categorised into levels to give the cognitive ability level of the participant, yielding another derived variable. This variable is called MMSE Index. Calculating MMSE Index from MMSE varies for different sources for individual, i. In Lagona and Zhang (2010), MMSE is split into a binary variable to form MMSE Index, where the cut-off point, d, varies in the study between 10 and 23. In Tombaugh and McIntyre (1992), MMSE is a 30-point questionnaire, so is scored out of 30. Re-calibrating[1] the cut-offs to suit a 23-point questionnaire yields the cut-off points at:

$$MMSEIndex = \begin{cases} severe, & \text{if } MMSE \leq 7 \\ moderate, & \text{if } 8 \leq MMSE \leq 15 \\ mild, & \text{if } 16 \leq MMSE \leq 20 \\ normal, & \text{otherwise.} \end{cases}$$

In MMSE Index, 12% of participants are in the "severe" category, 28% "moderate", 20% are "mild", and 40% are "normal" (Figure 4.1).

MMSE and MMSE Index are just two further examples of a derived variable.

4.2 Design of the Preliminary Analysis

The motivating data sets discussed in Section 4.1 contain derived variables with a ratio, additive, and index functional form. The derived variables are imputed by

[1] These are calculated such that a similar percentage score for the MMSE value is obtained for each category of both the 23 and 30 item questionnaires.

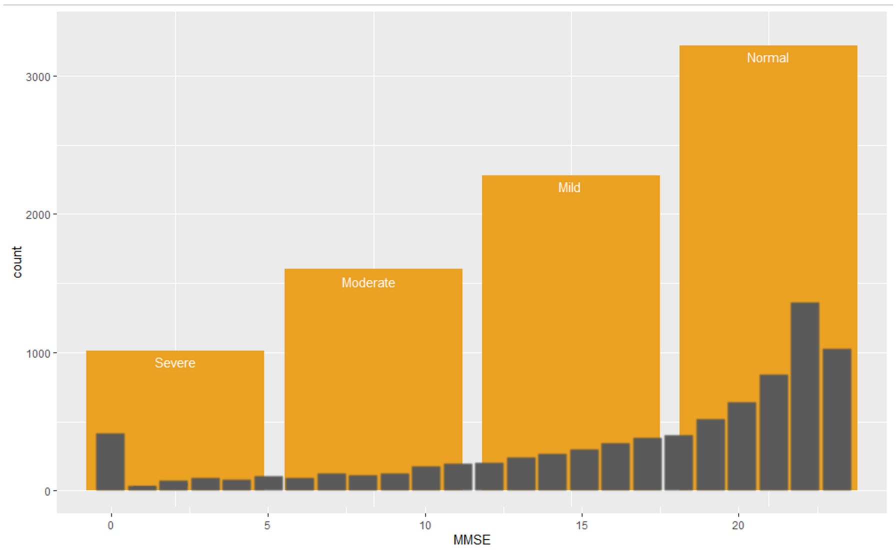


FIGURE 4.1: Distribution of the MMSE and MMSE Index variable. MMSE Index are given in the four levels, and MMSE is given along the x-axis.

multiple imputation to investigate the performance of active and passive imputation. The overarching design in the preliminary analysis is as follows:

1. Prepare the motivating data set
2. Generate missing values in the data
3. Use MICE to impute, analyse, and pool the incomplete data set.
4. Investigate the performance of the different imputation models.

These steps are given in further detail next. Following this in Sections 4.2.1-4.2.3 are specifications under the different functional forms.

Prepare the motivating data set The data sets addressed in Section 4.1 are reduced to their complete-case form. This is to ensure that the imputed data sets can be compared to a complete data set in the preliminary analysis.

Using the motivating data sets to investigate the performance of MI means that the investigation is based on realistic data. However, the results on the investigation are not as generalisable, and are less flexible when investigating certain properties of MI, such as the use of an auxiliary variable. As a result, a simulation study is designed and performed with data generated from a model in Chapter 5

Generating Missing Values A proportion, r, of values in a derived variable are generated as missing. To achieve this, a Bernoulli distribution dummy variable, W, is randomly generated with $P(w = 1) = r$. When $W = 1$ for an individual, the derived variable is set as missing, otherwise, the derived variable remains observed. Hence, each value of the derived variable is set to missing independently with probability, r, resulting in a MCAR missingness mechanism. When the derived variable is missing, at least one constituent is too.

The proportion, r, and approach to generate missing values in the constituents differ for the different functional forms. These are given in Sections 4.2.1-4.2.3 for a ratio, additive, and index functional form respectively.

MICE MICE, as outlined in Section 2.2.1, is performed with $M = 30$. Each incomplete data set is imputed by either an active or passive imputation model. Four imputation models are investigated for all functional forms: three variations of active imputation, and one passive imputation model. In all imputation models, the raw value of survival time and the censoring variable are predictors to avoid incompatibility issues. Note that the imputation models in this preliminary analysis can be optimised further by using a Nelson-Aalen estimate of the cumulative baseline hazard in an imputation model instead of the raw survival time (White and Royston, 2009). In addition, an error term, ε, is present in the imputation model such that $\varepsilon \sim N(0, \sigma^2)$. One imputation model is active imputation when the constituents are not predictors of the derived variable (**AWO**). In an AWO imputation model, all variables in the data set are predictors for the derived variable except the constituents.

APA denotes active imputation when all variables in the data set are predictors when imputing the derived variable, including the constituents. Additionally, all variables in the data set (including other constituents) are predictors when imputing the constituents, with the exception of the derived variable. The derived variable is not a predictor in the imputation model for the constituents to avoid circularity (Vink and Buuren (2017)).

APA2 denotes active imputation when all variables in the data set are predictors when imputing the derived variable, including the constituents. Additionally, all variables in the data set (including other constituents) are predictors when imputing the constituents. APA2 is performed despite the potential of circularity (Vink and Buuren, 2017) because the active imputation model investigated in Mitani et al. (2015) follows the same structure as APA2.

PNP denotes the passive imputation model. In passive imputation, the constituents are first imputed with the derived variable constructed later. To avoid circularity, the derived variable is not a predictor of the constituents for this model (Vink and Buuren, 2017).

The imputation models are defined in further detail for a ratio, additive, and index functional form in Sections 4.2.1-4.2.3 respectively.

The substantive model fitted to each multiply imputed data set is a Cox Proportional-Hazards model. The substantive model is then pooled via Rubin's Rules where the parameter of interest, Q, is the coefficient for the derived variable.

To accommodate the additional conditions, step 3 of the procedure given in Section 4.2 is repeated for each imputation model. Brand et al. (2003) suggests replicating the procedure between 200 and 1000 times. The simulation is repeated for 250 replications.

4.2.1 Ratio Functional Form

In this section, specifications for the preliminary study under a ratio functional form are discussed.

Modifications to the Data Two data sets are used to investigate the performance of active and passive imputation for a ratio functional form. Each data set additionally contains two variables with a ratio functional form: donor BMI and recipient BMI. Hence there are four cases explored:

- Kidney Donor BMI
- Kidney Recipient BMI
- Cardiothoracic Donor BMI
- Cardiothoracic Recipient BMI.

Both the kidney and cardiothoracic transplant data sets are first to reduced their complete case form in the preliminary analysis. In addition, individuals below the age of 20 are removed because BMI alone is not a good indicator of an individual's health for individuals under 20 years old (CDC, 2015). This includes removing individuals under the age of two years old, since BMI is not a suitable measure. Finally, one individual in the cardiothoracic data set with an observed weight of 6.1kg is removed since this outlier is considered to be an incorrect value. As a result, the kidney transplant data has 1938 rows and the cardiothoracic data set has 4380 rows.

Figure 4.2 is a set of plots to show the distribution for the weight, height, and BMI variables for recipients in the complete-case kidney transplant data set. From Figure 4.2 it is observed that BMI and weight variables are skewed for both donors and recipients. In addition, in Table 4.1 the minimum, maximum, and mean values are given for height, weight, and BMI for donors and recipients in both data sets. The distribution for height suggests that figures are frequently rounded to multiples of five. Some outliers are additionally visible; for example, there are values for recipient height in the kidney transplant data of only 100cm and 101cm.

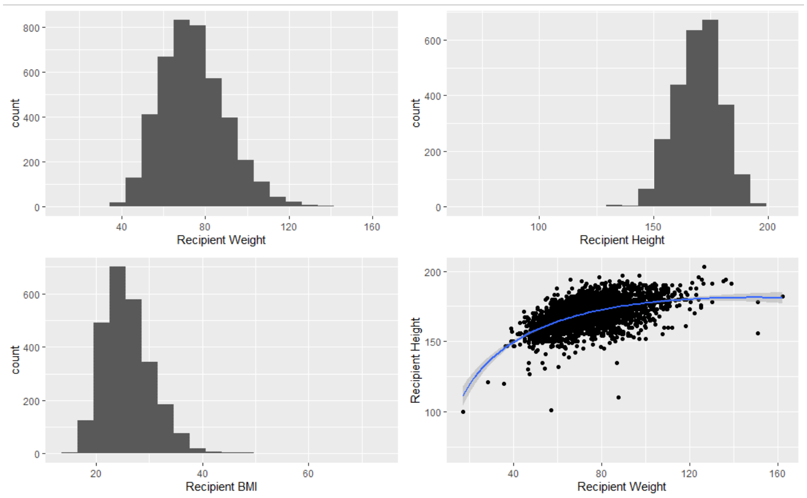


FIGURE 4.2: Histograms of recipient BMI, height, and weight for the kidney transplant data set. Similar distributions are observed for cardiothoracic transplant data set, and for donors. Additionally, a plot of recipient height against recipient weight.

TABLE 4.1: Statistics for BMI and its constituents in the complete transplant data sets. Weight is given in kilograms, and height in centimetres.

	Kidney			**Cardiothoracic**		
	Minimum	**Mean**	**Maximum**	**Minimum**	**Mean**	**Maximum**
Donor BMI	12.5	26.4	79.9	12.0	25.4	60.0
Donor Height	105.0	170.4	199.0	137.0	172.4	208.0
Donor Weight	36.0	76.6	197.0	35.5	75.6	180.0
Recipient BMI	15.4	26.0	62.0	14.0	24.5	46.0
Recipient Height	100.0	169.7	203.0	140.0	170.3	205.0
Recipient Weight	17.0	75.2	162.2	35.0	71.3	145.1

Generating Missing Values In a ratio functional form, there is a distinction between whether the numerator or the denominator is missing. Additionally, in both data sets investigated in the preliminary analysis there is an additional complication because there are two derived variables which are predictors of one another in the imputation models.

Missing values in the BMI variables are generated to follow a structure similar to the missing values in the initial motivating data set on kidney transplants. 66% of rows contain a missing value in either recipient BMI or donor BMI in the kidney transplant data set. Of this 66%:

- recipient BMI is missing and donor BMI is observed in 92% of rows,
- recipient BMI is observed and donor BMI is missing in 2% of rows,
- both recipient and donor BMI is missing in 6% of rows.

When recipient BMI is missing:

- recipient weight is missing and recipient height is observed in 1% of rows,
- recipient weight is observed and recipient height is missing in 37% of rows,
- both recipient weight and height is missing in 62% of rows,

When donor BMI is missing:

- donor weight is missing and donor height is observed in 1% of rows,
- donor weight is observed and donor height is missing in 50% of rows,
- both donor weight and height is missing in 49% of rows,

A missingness structure is generated in both data sets in the preliminary analysis that emulates a structure similar to the missing values in the BMI variables in the kidney transplant data.

Recall from Section 4.2 that a dummy variable, W, is randomly generated with $P(W = 1) = r$. When $W = 1$ for an individual, the derived variable is set as missing, otherwise, the derived variable remains observed. For a ratio functional form, $r = 0.66$ since 66% of rows are missing in the kidney transplant data set.

When $w = 1$, at least one of donor or recipient BMI is missing, otherwise both BMI variables are observed. A second dummy variable is then generated, W_1, which takes values $0, 1, 2, 3$ where if

- $W_1 = 0$ both BMI variables are observed,
- $W_1 = 1$ recipient BMI is missing, and donor BMI is observed,
- $W_1 = 2$ recipient BMI is observed, and donor BMI is missing,
- $W_1 = 3$ both BMI variables are missing.

Hence,

$$\begin{cases} P(W_1 = 0) = 1, & \text{if } W = 0, \\ P(W_1 = 1) = 0.92, & \text{if } W = 1, \\ P(W_1 = 2) = 0.02, & \text{if } W = 1, \\ P(W_1 = 3) = 0.06, & \text{if } W = 1. \end{cases}$$

A third dummy variable is then generated, W_2. W_2 denotes which constituents of recipient BMI are missing and takes values $0, 1, 2, 3$:

- $W_2 = 0$ both constituents for recipient BMI are observed,
- $W_2 = 1$ recipient weight is missing, and recipient height is observed,
- $W_2 = 2$ recipient weight is observed, and recipient height is missing,
- $W_2 = 3$ both constituents for recipient BMI are missing.

Hence,

$$\begin{cases} P(W_2 = 0) = 1, & \text{if } W_1 = 0 \text{ or } W_1 = 2, \\ P(W_2 = 1) = 0.01, & \text{if } W_1 = 1 \text{ or } W_1 = 3, \\ P(W_2 = 2) = 0.37, & \text{if } W_1 = 1 \text{ or } W_1 = 3, \\ P(W_2 = 3) = 0.62, & \text{if } W_1 = 1 \text{ or } W_1 = 3. \end{cases}$$

Similarly, a fourth dummy variable is then generated, W_3. W_3 denotes which constituents of donor BMI are missing and takes values $0, 1, 2, 3$:

- $W_2 = 0$ both constituents for donor BMI are observed,
- $W_2 = 1$ donor weight is missing, and donor height is observed,
- $W_2 = 2$ donor weight is observed, and donor height is missing,
- $W_2 = 3$ both constituents for donor BMI are missing.

Hence,

$$\begin{cases} P(W_3 = 0) = 1, & \text{if } W_1 = 0 \text{ or } W_1 = 3, \\ P(W_3 = 1) = 0.01, & \text{if } W_1 = 1 \text{ or } W_1 = 2, \\ P(W_3 = 2) = 0.50, & \text{if } W_1 = 1 \text{ or } W_1 = 2, \\ P(W_3 = 3) = 0.49, & \text{if } W_1 = 1 \text{ or } W_1 = 2. \end{cases}$$

MICE In the motivating paper by Pankhurst et al. (2020), MICE is performed with BLR. Pankhurst et al. (2020) propose that further research can investigate active and passive imputation. In Section 4.2.1, the performance of active and passive imputation are investigated for the data set applied in Pankhurst et al. (2020). Hence, for consistency with Pankhurst et al. (2020), BLR is applied in the preliminary analysis. In addition to BLR, PMM is performed in a separate set of preliminary analyses since BLR assumes that the incomplete covariate is normally distributed. However, the derived variable and its constituents display a skewed distribution (Figure 4.2).

The variables in the substantive model for the kidney transplant data set are chosen based on the significant variables in Pankhurst et al. (2020). That is, they are the significant variables following stepwise selection to the full data. Recipient BMI is additionally included as a predictor since recipient BMI is one of the derived variables. As a result, the substantive model in the preliminary analysis for the kidney transplant data set is the following Cox Proportional-Hazards Model:

$$h(t) = h_0(t)\exp(\beta_1 dage + \beta_2 dbmi + \beta_3 dcmv + \beta_4 rage + \beta_5 rsex + \beta_6 rethnic +$$

$$\beta_7 rbmi + \beta_8 prd + \beta_9 serum3 + \beta_{10} local).$$

The variables are explained in further detail in Appendix A.

For consistency, a stepwise selection model is fitted to the full cardiothoracic transplant data to determine predictors for the substantive model. Both donor and recipient BMI are included in the model resulting in the following Cox Proportional-Hazards Model:

$$h(t) = h_0(t)\exp(\beta_1 tx\_yr + \beta_2 het + \beta_3 dage + \beta_4 dcmv + \beta_5 dpast\_smoker + \beta_6 rethnic +$$

$$\beta_7 dbmi + \beta_8 rcod + \beta_9 rbmi)$$

The variables are explained in further detail in Appendix C.

The four imputation models introduced in Section 4.2 are investigated for a ratio functional form.

In an AWO imputation model, all variables in the data set are predictors for donor BMI, with the exception of both donor and recipient height and weight variables. Similarly, all variables in the data set are predictors for recipient BMI, with the exception of both donor and recipient height and weight variables. For the kidney transplant data set when imputing donor BMI, the AWO imputation model is

$$\text{dbmi} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local} +$$

$$\text{time} + \text{status} + \varepsilon.$$

In an APA imputation model, all variables in the data set are predictors for donor BMI, and for recipient BMI. In addition, all variables in the data set are predictors for donor height, with the exception of donor BMI. Similarly, all variables in the data set are predictors for donor weight, with the exception of donor BMI. An equivalent imputation model is fitted for recipient BMI. For the kidney transplant data set when imputing donor BMI, the set of APA imputation models is

$$\text{dweight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local} +$$

$$\text{time} + \text{status} + \text{dheight} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dheight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local} +$$

$$\text{time} + \text{status} + \text{dweight} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dbmi} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dweight} + \text{dheight} + \text{rweight} + \text{rheight} + \varepsilon.$$

In an APA2 imputation model, all variables in the data set are predictors for BMI, height and weight for both donors and recipients. For the kidney transplant data set when imputing donor BMI, the set of APA2 imputation models is

$$\text{dweight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dheight} + \text{dbmi} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dheight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dweight} + \text{dbmi} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dbmi} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dweight} + \text{dheight} + \text{rweight} + \text{rheight} + \varepsilon.$$

For a PNP imputation model, the set of imputation models are the same as with APA for the constituents. BMI is then constructed using the functional form. For the kidney transplant data set when imputing donor BMI, the set of PNP imputation models is

$$\text{dweight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dheight} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dheight} = \text{dage} + \text{dcmv} + \text{rage} + \text{rsex} + \text{rethnic} + \text{rbmi} + \text{prd} + \text{serum3} + \text{local}+$$

$$\text{time} + \text{status} + \text{dweight} + \text{rweight} + \text{rheight} + \varepsilon,$$

$$\text{dbmi} = \frac{\text{dweight}}{(\text{dheight}/100)^2}$$

An additional passive imputation model is investigated where the constituents are first log-transformed before imputation takes place (**LNP**), $\text{weight}^* = \log(\text{weight}); \text{height}^* = \log(\text{height})$. In LNP, the set of imputation models the same as for passive imputation, except that the functional form is altered for BMI to account for the log-transformation:

$$BMI = \exp(\text{weight}^* - 2 * \text{height}^*).$$

4.2.2 Additive Functional Form

Specifications for the preliminary analysis under an additive functional form are discussed in this section.

Modifications to the Data Two derived variables with an additive form are introduced in Section 4.1.2 from the CLHLS data set, ADL and MMSE. In addition to the two derived variables, there are three covariates: age, gender, and residence, and two outcome variables: survival time and a censored indicator. There are also K constituents. Four cases are considered with varying values of K:

- $K = 2$. The derived variable is ADL with two binary constituents: bathing and dressing.
- $K = 6$. The derived variable is ADL with all binary constituents present (Section 4.1.2).
- $K = 13$. The derived variable is MMSE with 13 binary constituents present.
- $K = 5$. The derived variable is MMSE with five non-binary constituents present. The five constituents are the five groups that categorise the survey questions: orientation, registration, attention and calculation, recall, and language. These groups take values out of five, three, six, three, and six respectively.

These four cases are applied separate studies in the preliminary analysis when investigating the performance of active and passive imputation.

Generating Missing Values The derived variable is generated as missing for 30% of rows. This is achieved by the approach outlined in Section 4.2. If the derived variable is missing, then at least one constituent needs to be generated as missing. To set which constituents are missing a dummy variable, W_1, is randomly generated for each row of the data set. When $w = 0$ (that is, if the derived variable is observed), $P(w_1 = 0) = 1$.

For an additive functional form, all constituents are weighted equally to construct the derived variable. As a result, it is structurally irrelevant which constituent(s) is missing. Hence, when $w = 1$, $P(w_1 = 1) = ... = P(w_1 = K) = 1/K$ for K constituents in the derived variable. If $w_1 = 1$ for an individual, the value in one random constituent is set as 'NA' for that row. When $w_1 = 2$, values from two random constituents are set as NA for the row. This continues until $w_1 = K$ when the values from all constituents are set as NA for the individual.

MICE A ratio functional form is imputed under MICE by BLR and PMM. For consistency when considering other functional forms, both BLR and PMM are performed for an additive functional form under MICE. Both are investigated for the reasons outlined in Section 4.2. In addition, PMM is appropriate for an additive functional form because the constituents and derived variable are binary or categorical variabes.

A Cox Proportional-Hazards substantive model is fitted to each imputed data set where

$$\text{Surv(time, status)} = \text{Age} + \text{Gender} + \text{Residence} + \text{ADL} + \text{MMSE} + \epsilon$$

The substantive model is additionally fitted to the complete-case CLHLS data to estimate the coefficient of the derived variable. In cases one and two, the derived variable is ADL.

The imputation models introduced in Section 4.2 are investigated when the derived variable takes an additive functional form. When ADL is the derived variable, the imputation model under AWO contains all variables in the substantive model:

$$\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{survival time} + \text{status} + \varepsilon.$$

When ADL is the derived variable, the imputation model under APA contains all variables in the substantive model and the constituents. The imputation models for the constituents contain all variables in the data set, except for the derived variable. For example, in the case with two constituents:

$$\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Bathing} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon,$$

$$\text{Bathing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon,$$

$$\text{Dressing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon.$$

When ADL is the derived variable, the imputation model under APA2 contains all variables in the substantive model and the constituents. The imputation models for the constituents contain all variables in the data set. For example, in the case with two constituents:

$$\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Bathing} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon,$$

$$\text{Bathing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{ADL} + \text{survival time} + \text{status} + \varepsilon,$$

$$\text{Dressing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{ADL} + \text{survival time} + \text{status} + \varepsilon.$$

Under passive imputation when ADL is the derived variable, the derived variable is constructed using the functional form. The constituents contain all variables in the data set, except for the derived variable. For example, in the case with two constituents:

$$\text{Bathing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon.$$

$$\text{Dressing} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon.$$

$$\text{ADL} = \text{Bathing} + \text{Dressing}.$$

In case 3 and 4, MMSE is the derived variable. The above models for AWO, APA, APA2, and PNP are fitted with MMSE as the outcome variable, and ADL as a predictor. In addition, the constituents for MMSE replace the constituents for ADL.

4.2.3 Index Functional Form

Specifications for the simulation design under an index functional form are discussed in this section.

Modifications to Data The performance of active and passive imputation under an index functional form is investigated using the CLHLS data set outlined in Section 4.1.2. Two index functional forms are investigated, ADL Index and MMSE Index. The data set used in the preliminary analysis additionally has three covariates: age, gender, and residence, and two outcome variables: survival time and a censoring indicator. The constituent for the derived variable is also present in the data set. Three cases are investigated:

1. ADL Index is the derived variable, constructed from ADL. ADL Index is categorised into three groups:

$$ADL\_Index = \begin{cases} 0, & \text{if } ADL = 0 \\ 1, & \text{if } ADL = 1 \\ 2+, & \text{otherwise} \end{cases}$$

2. MMSE Index is the derived variable, constructed from a 13-point MMSE score. MMSE is dichotomised into four levels to give the derived variable, MMSE Index. Readjusting the levels for a 23-point MMSE Index score, outlined in Section 4.1.2, gives:

$$MMSE\_Index = \begin{cases} severe, & \text{if } MMSE \leq 4 \\ moderate, & \text{if } 5 \leq MMSE \leq 8 \\ mild, & \text{if } 9 \leq MMSE \leq 10 \\ normal, & \text{otherwise} \end{cases}$$

3. MMSE Index is the derived variable, constructed from a 23-point MMSE score. MMSE is categorised into four levels to give the derived variable, MMSE Index:

$$MMSE\_Index = \begin{cases} severe, & \text{if } MMSE \leq 7 \\ moderate, & \text{if } 8 \leq MMSE \leq 14 \\ mild, & \text{if } 15 \leq MMSE \leq 18 \\ normal, & \text{otherwise} \end{cases}$$

These index derived variables have different structures. ADL Index is more skewed than MMSE Index due to its structure from its constituent since values 5/7 of the values ADL take group into the same level in ADL Index.

Generating Missing Values The data sets generated to have missing data for the relevant additive derived variable are used to generate missing data for the corresponding index derived variable. When the additive derived variable is missing, so is the index derived variable.

MICE Step 3 in the simulation procedure is performed using PMM to multiply impute missing values.

A Cox Proportional-Hazards substantive model is fitted to each imputed data set where

$$\text{Surv(time, status)} = \text{Age} + \text{Gender} + \text{Residence} + \text{ADL Index} + \text{MMSE Index} + \epsilon.$$

The substantive model is additionally fitted to the complete-case CLHLS data to estimate the coefficient of the derived variable. In case one, the derived variable is ADL Index.

The imputation models AWO, APA, APA2, and PNP introduced in Section 4.2 are investigated for an index functional form. When ADL Index is the derived variable, the imputation model under AWO contains all variables in the substantive model:

$$\text{ADL Index} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{survival time} + \text{status} + \varepsilon.$$

When ADL Index is the derived variable, the imputation model under APA contains all variables in the substantive model and the constituent, ADL. The imputation models for the constituent contain all variables in the data set, except for the derived variable:

$$\text{ADL Index} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{ADL} + \text{survival time} + \text{status} + \varepsilon,$$

$$\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon.$$

When ADL Index is the derived variable, the imputation model under APA2 contains all variables in the substantive model and the constituent, ADL. In addition, the imputation models for the constituent contain all variables in the data set:

$$\text{ADL Index} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{ADL} + \text{survival time} + \text{status} + \varepsilon,$$

$$\begin{aligned}\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{ADL Index} + \text{Dressing}\\ + \text{survival time} + \text{status} + \varepsilon.\end{aligned}$$

Under PNP, the constituents contain all variables in the data set, except for the derived variable. When ADL Index is the derived variable, the set of imputation models is

$$\text{ADL} \sim \text{Age} + \text{Gender} + \text{Residence} + \text{MMSE Index} + \text{Dressing} + \text{survival time} + \text{status} + \varepsilon.$$

$$ADL\_Index = \begin{cases} 0, & \text{if } ADL = 0 \\ 1, & \text{if } ADL = 1 \\ 2+, & \text{otherwise.} \end{cases}$$

In cases two and three, MMSE Index is the derived variable. The above models for AWO, APA, APA2, and PNP are fitted with MMSE Index as the outcome variable, and ADL Index as a predictor. In addition, MMSE replaces ADL.

4.3 Methods to Compare Imputation Models

In this section, the approaches used in the preliminary analysis to evaluate the performance of the imputation models are given.

4.3.1 Trace plots

Trace plots can check for convergence in the imputation of the imputed variable. In a trace plot, the standard deviation or mean of the imputed values in a variable can be plotted against the iteration number for the T cycles. This is then repeated for each of the M multiply imputed data sets. A pattern in the trace plot indicates that the imputation for the imputed variable has not converged (Van Buuren, 2018).

4.3.2 Estimates from the Substantive Model

Some approaches discussed in Chapter 2 are applied to compare the different imputation models, namely Raw Bias (RB), Coverage Rate (CR), and Average Width

(AW). The parameter of interest, Q, denotes the estimated coefficient of the derived variable in the Cox Proportional-Hazards model fitted to the complete data set, before missingness is imposed. The estimated parameter of interest, $\bar{Q}_M$ denotes the pooled value of a parameter estimate in the substantive model, Q, over M multiply imputed data sets.

Recall in Chapter 2 that,

$$RB = \mathbb{E}[\bar{Q}_M] - Q$$

where $\mathbb{E}[\bar{Q}_M]$ denotes the mean of all $\bar{Q}_M$ values across all replications. The coverage rate (CR) is the proportion of replications where Q is in a 95% estimated pooled confidence interval for $\bar{Q}_M$. The AW is the average width of the estimated confidence interval for $\bar{Q}_M$ across all replications.

For cases 1 and 2 with an additive functional form, the RB, CR, and AW are calculated by comparing the pooled estimated coefficient for ADL in the multiply imputed data to the estimated coefficient for ADL in the complete case CLHLS data. Similarly, in cases three and four the derived variable is MMSE. Hence, the RB, CR, and AW are evaluated by comparing the pooled estimated coefficient for MMSE in the multiply imputed data to the estimated coefficient for MMSE in the complete case CLHLS data.

For case 1 under an index functional form, the RB, CR, and AW are calculated by comparing the pooled estimated coefficient for ADL Index in the multiply imputed data to the estimated coefficient for ADL Index in the complete case CLHLS data. Similarly, in cases two and three the derived variable is MMSE Index. Hence, the RB, CR, and AW are evaluated by comparing the pooled estimated coefficient for MMSE Index in the multiply imputed data to the estimated coefficient for MMSE Index in the complete case CLHLS data.

Additionally the imputation procedure is evaluated by calculating the FMI and RIV.

4.3.3 Formalised Testing of Estimated Values

Formalised testing can be used to compare estimated values from the different imputation models. As a result, the formal tests can indicate if one imputation model results in estimates that are different to another.

As described in Johnson and Wichern (2007), a t-test to determine whether a specific value, μ_0, is a plausible value for the population mean, μ, would have the hypotheses

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0,$$

where H_0 is the null hypothesis, and H_1 is a two-sided alternative hypothesis.

Denote by $X_1, ..., X_n$ a random sample such that $X_j \sim N(\mu, \sigma^2)$ for $j = 1, ..., n$. The mean, $\bar{X}$, and variance, s^2, are given by

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_i - \bar{X})^2.$$

To test the hypothesis that μ_0 is a plausible value for the population mean, μ, the test statistic is

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}},$$

where t has a student's t-distribution with $n-1$ degrees of freedom. t can be squared and rearranged to

$$t^2 = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0).$$

H_0 is rejected at a significance level, α, and hence there is significant evidence that μ_0 is not a plausible value of μ, if

$$t^2 = n(\bar{X} - \mu_0)(s^2)^{-1}(\bar{X} - \mu_0) > t^2_{n-1}(\alpha/2),$$

where $t_{n-1}(\alpha/2)$ is the upper $100(\alpha/2)^{th}$ percentile of the t-distribution with $n-1$ degrees of freedom.

A t-test is performed when establishing the plausibility of a univariate mean. However, a t-test can be extended to consider whether a $p \times 1$ vector, $\boldsymbol{\mu_0}$ is a plausible value for the mean of a multivariate normal distribution. To test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu_0}$ against a general alternative, first denote by $\boldsymbol{X_1}, ..., \boldsymbol{X_n}$ a random sample with p parameters, $j = 1, ..., n$, such that $\boldsymbol{X_j} \sim \text{MVN}_p(\boldsymbol{\mu}, \Sigma)$. The test statistic from the univariate case is generalised to a test statistic for the multivariate case, T^2, where T^2 is called Hotelling's T^2:

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}),$$

where

$$\bar{\boldsymbol{X}} = \frac{1}{n}\sum_{j=1}^{n}\boldsymbol{X_j},$$

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X_i} - \bar{\boldsymbol{X}})(\boldsymbol{X_i} - \bar{\boldsymbol{X}})'.$$

$$T^2 \approx \frac{(n-1)p}{n-p}F_{p,n-p}.$$

Hence, H_0 is rejected in favour of H_1 at a $(100 \times (\alpha/2))\%$ significance level if

$$T^2 > \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha).$$

If $n - p$ is large then H_0 is rejected in favour of H_1 at a $(100 \times \alpha/2)\%$ significance level if

$$T^2 > \chi^2_p(\alpha).$$

In the preliminary analysis performed in this thesis, there are multivariate means. A t-test can be generalised to a multivariate case. To test $H_0 : \mu_1 = \mu_2 = ... = \mu_p$ for $\boldsymbol{\mu}' = (\mu_1, ..., \mu_p)$ against a general alternative, we first rearrange H_0 such that

$$\mu_1 = ... = \mu_p \iff \mu_2 - \mu_1 = 0 = ... = \mu_p - \mu_1 \tag{4.1}$$

(Note that this can be rearranged in many other ways).

If $\boldsymbol{X} \sim MVN(\boldsymbol{\mu}, \Sigma)$, then $A\boldsymbol{X} \sim MVN(A\boldsymbol{\mu}, A\Sigma A')$ (Johnson and Wichern, 2007) for some matrix, A, with size $p \times (p-1)$. With the hypotheses given in 4.1, let

$$A = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ -1 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ -1 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Let $Y = AX$ and denote $\boldsymbol{\mu}_y = A\boldsymbol{\mu}$ and $\Sigma_y = A\Sigma A'$. Then $\boldsymbol{Y} \sim MVN_{p-1}(\boldsymbol{\mu}_y, \Sigma_y)$, and $T^2 = n\bar{\boldsymbol{Y}}(S_y^{-1})\bar{\boldsymbol{Y}}'$ is the test statistic to test the null hypothesis.

Hence, to test whether there is a difference in the mean estimated coefficients, or the calculated AW values, between $p = 4$ imputation models with $n = 250$ replications,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 : \mu_1 \neq \mu_2, \text{ or } \mu_1 \neq \mu_3, \text{ or } \mu_1 \neq \mu_4, \text{ or } \mu_2 \neq \mu_3, \text{ or } \mu_2 \neq \mu_4, \text{ or } \mu_3 \neq \mu_4.$$

$$A = \begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix}.$$

$\bar{\boldsymbol{X}}$ is a vector of the mean pooled estimated coefficients, or the mean AW values, for each imputation model.

The performance of active and passive imputation is investigated for several conditions and functional forms. A Bonferroni correction is calculated by dividing α by the number of different conditions. For example, a ratio functional form is repeated for BLR and PMM, for two BMI variables, and for two data sets. Hence, the Bonferroni correction is calculated by dividing α by $2^3 = 8$.

4.3.4 Formalised Testing of Coverage Rates

The performance of the imputation models can be assessed by investigating whether the observed coverage rates (CR) are in line with a true coverage of 95%. In Section 2.2.4, overcoverage is defined when the $CR > 95\%$, and undercoverage is when the $CR < 95\%$. However, some variability around the 95% interval should be accounted for. In this section, a boundary that defines a reasonable CR is calculated.

The preliminary analysis is repeated for 250 replications and a suitable coverage rate is 0.95. Hence, a binomial distribution can be fitted with $\pi = 0.95, n = 250$. Since π is close to 1, the binomial distribution is negatively skewed. As a result, the Wilson Score is recommended to construct the confidence interval instead of a Wald confidence interval (Brown et al., 2001). The Wilson Score is additionally recommended irrespective of skewed observations since it results in accurate, robust estimated confidence intervals (Brown et al., 2001). A $100(1 - \alpha/2)\%$ confidence interval for Wilson Score is defined by

$$\frac{\hat{\pi} + z^2/2n}{1 + z^2/n} \pm \frac{z}{1 + z^2/n}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2}{4n^2}} \tag{4.2}$$

where $z = 1.96$ for a 95% confidence interval. Additionally, $n = 250$. When $\hat{\pi} < 0.923$, the upper limit of the Wilson Score confidence interval is less than 0.95. Hence, if a calculated 95% confidence interval is less than 92.3%, undercoverage is indicated. When $\hat{\pi} > 0.977$, the lower limit of the Wilson Score confidence interval is greater than 0.95. Hence, if a calculated 95% confidence interval is greater than 97.7%, overcoverage is indicated.

Another confidence interval useful in handling skewed binomial observations is the Agressi-Coull confidence interval. The Agressi-Coull confidence interval is more appropriate for large n (Brown et al., 2001). Agressi-Coull is defined by

$$\tilde{\pi} \pm z\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n}} \tag{4.3}$$

where

$$\tilde{\pi} = \frac{\hat{\pi} + z^2/2n}{1 + z^2/n}.$$

When $\hat{\pi} < 0.922$, the upper limit of the Wilson Score confidence interval is less than 0.95. When $\hat{\pi} > 0.978$, the lower limit of the Wilson Score confidence interval is greater than 0.95. Hence a 95% confidence interval less than 92.2% indicates undercoverage, and a confidence interval greater than 97.8% indicates overcoverage.

A reasonable CR is implied for an imputation model under all functional forms if it is in either the Wilson Score interval (92.3, 97.7) or the Agressi-Coull interval (92.2, 97.8).

4.4 Results when Comparing Imputation Models

Results from the preliminary analysis are given in this section, starting with the ratio functional form, and then the additive and index functional forms. Trace plots are first used to check for convergence in the imputation models. Following this, the RB, CR, and AW are investigated and given in a summary table. Finally, the mean and standard deviation FMI values are discussed. The imputation models are compared to each other through the different cases.

4.4.1 Ratio Functional Form

Trace plots when the kidney and cardiothoracic data set are imputed by PMM are given in Figures 4.3 and 4.4 respectively. The trace plot for APA2 does not imply convergence in the imputation of the derived variable. This non-convergence in APA2 arises because of circularity. That is, the constituents are predictors when imputing the derived variable and vice versa (Vink and Buuren, 2017). APA2 is known to perform poorly elsewhere in the literature due to circularity. However, APA2 is still investigated since the active imputation model considered in Mitani et al. (2015) follows a similar structure. As a result, APA2 is omitted from future analyses when investigating a ratio functional form.

The RB, CR, and AW after fitting a substantive model are given in Table 4.2. The RB, CR, and AW are calculated for both donor and recipient BMI, and both transplant data sets.

The performance of the active and passive imputation models is discussed next in this section for donor BMI, and then for recipient BMI.

There is consistently overcoverage for the imputation models under all conditions when the derived variable is donor BMI. This overcoverage occurs for donor BMI and not for recipient BMI because a smaller proportion of donor BMI values are generated as missing than that of recipient BMI values. There is only one source of variability in the preliminary analysis, but, the Wald confidence interval may be such that it assumed a higher variability hence the variability assumption is incorrect. Hence, there is less variability for the imputed donor BMI values than the imputed recipient BMI values.

A hypothesis test is performed to test for the equality of means in the pooled estimated coefficients of donor BMI. The null hypothesis is that the mean pooled estimated coefficient of donor BMI is equal across all imputation models; the alternative hypothesis is that at least one estimated coefficient for donor BMI resulting from an imputation model is not equal to another estimated coefficient for donor BMI

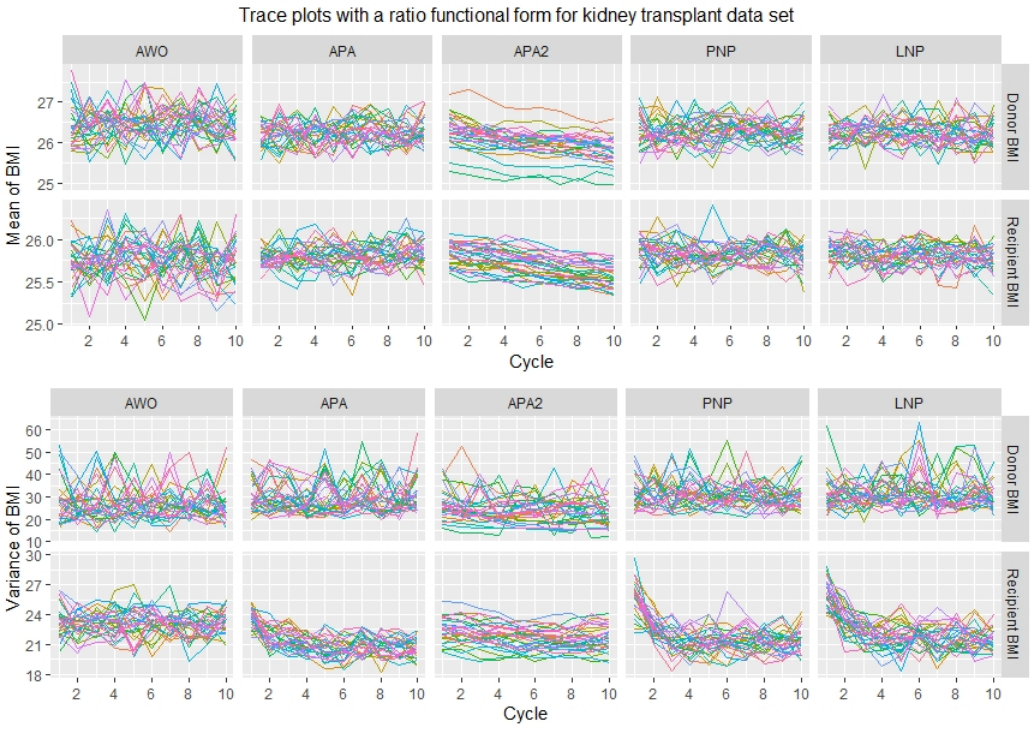


FIGURE 4.3: Trace Plots for each imputation model when PMM is applied to the kidney transplant data set. Trace plots are given for one replication. Other replications and conditions display a similar trend. The t cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of donor BMI and recipient BMI in each of the M imputed data sets; the bottom plots displays the variance of both donor and recipient BMIs in each of the M imputed data sets.

resulting from another imputation model. This test is performed for both BLR and PMM. A similar test is additionally performed for AW, and for the estimated coefficient for recipient BMI.

The performance of imputation models for recipient and donor BMI differ slightly.

The hypothesis tests performed for each case yield insufficient evidence to reject the equality of means in the estimated coefficient for donor BMI. For donor BMI, there is not a clear distinction for the estimated coefficients of the derived variable, as observed from the RB values in Table 4.2. When considering the AW for donor BMI, there is significant evidence that the AW resulting from at least one imputation model is larger than the AW resulting from another imputation model. On inspection of Table 4.2, the AW values are very similar for APA, PNP, and LNP imputation models but larger for AWO. Hence, there is some suggestion that APA, PNP, LNP outperform AWO.

The RB and AW values are more distinct between imputation models for recipient BMI. For each case, there is statistically significant evidence that the mean estimated coefficients of recipient BMI are not equal between one imputation model to at least one other. By inspection from the RB values in Table 4.2, APA is consistently among

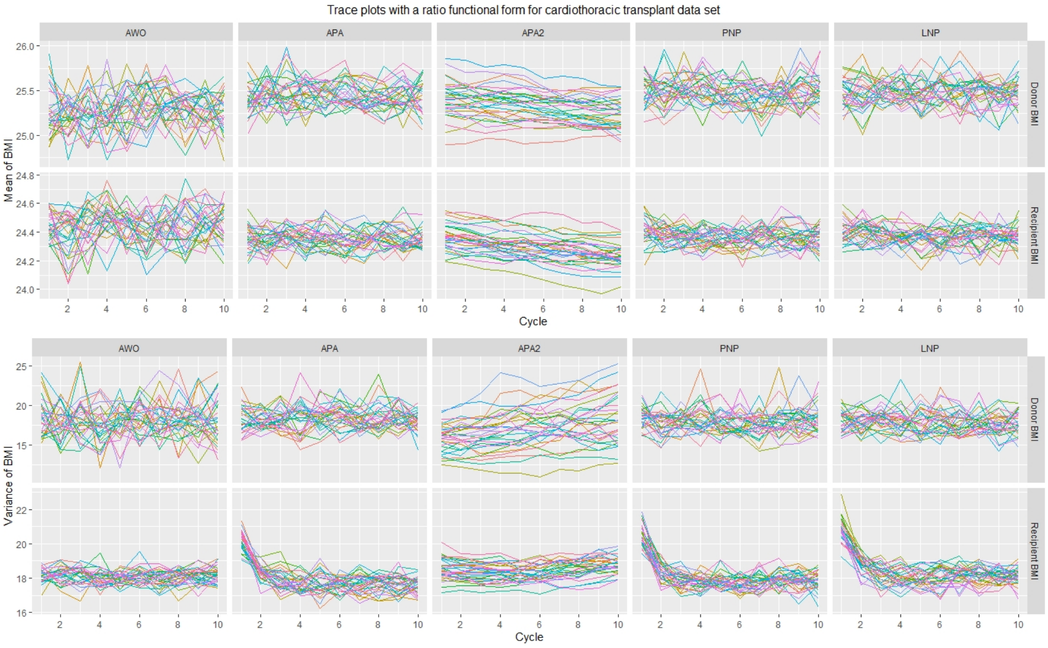


FIGURE 4.4: Trace Plots for each imputation model when PMM is applied to the cardiothoracic transplant data set. Trace plots are given for one replication. Other replications and conditions display a similar trend. The t cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of donor BMI and recipient BMI in each of the M imputed data sets; the bottom plots displays the variance of both donor and recipient BMIs in each of the M imputed data sets.

the least biased for all methods. In addition, for all cases there is significant evidence that the mean AW is not equal between the imputation models. Upon inspection of Table 4.2, the AW values are largest for AWO. There is some overcoverage when investigating recipient BMI, but not as extreme as that of donor BMI.

FMI values calculated from the substantive model are given in Table 4.3. A higher proportion of the total sampling variance is attributable to the missing data in the BMI variable when the imputation model is AWO compared to other imputation models. The same generated data sets are imputed by the different imputation models so the varying FMI values do not indicate a difference in the proportion of imputed values for imputation models. Instead a high FMI value indicates that the multiply imputed BMI value is not strongly associated with other variables in the imputation model. Therefore the mean FMI values are larger under AWO because there are fewer predictors for the BMI variables under an AWO imputation model. Hence, judging by the mean FMI values, more multiply imputed data sets are required under AWO than for other imputation models.

Overall, the performance of the imputation models for donor BMI differ to the performance of the imputation models for recipient BMI. This is evident from the RB, CR, AW, and FMI values, and a large reason for this stems from the difference in the proportion of variables that are missing. Only 7.5% of donor BMI values are generated

TABLE 4.2: The RB, CR, and AW values for a ratio functional form in the kidney transplant data set and cardiothoracic transplant data set.

			Kidney Transplant Data			**Cardiothoracic Transplant Data**		
			RB	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
Donor	BLR	AWO	0.00019	100.0	0.0376	0.00067	100.0	0.0244
		APA	0.00020	100.0	0.0363	0.00053	100.0	0.0237
		PNP	0.00010	100.0	0.0363	0.00050	100.0	0.0238
		LNP	0.00035	100.0	0.0362	0.00059	100.0	0.0238
	PMM	AWO	0.00033	100.0	0.0376	0.00084	100.0	0.0242
		APA	0.00022	100.0	0.0362	0.00059	100.0	0.0238
		PNP	0.00040	100.0	0.0362	0.00054	100.0	0.0238
		LNP	0.00013	100.0	0.0363	0.00062	100.0	0.0238
Recipients	BLR	AWO	-0.00126	94.4	0.0948	-0.00733	92.4	0.0447
		APA	-0.00046	97.2	0.0689	-0.00450	97.6	0.0318
		PNP	-0.00030	98.4	0.0702	-0.00488	97.2	0.0325
		LNP	-0.00308	98.4	0.0720	-0.00555	95.6	0.0327
	PMM	AWO	-0.00062	95.2	0.1061	-0.00838	90.4	0.0430
		APA	0.00029	96.8	0.0716	-0.00431	97.6	0.0327
		PNP	0.00080	98.0	0.0714	-0.00462	96.0	0.0323
		LNP	-0.00161	96.8	0.0698	-0.00561	95.2	0.0314

* denotes that the CR is not in the 95% confidence interval.

as missing in the preliminary analysis, whereas 66% of recipient BMI values are missing. For example, the AW is larger when there is a higher proportion of missing values. The AW is larger for recipient BMI because there is a greater uncertainty in the estimated coefficient for recipient BMI when fitting the substantive model since a higher proportion of recipient BMI is imputed than for donor BMI. Due to this, there is more variability in the denominator for recipient BMI than for donor BMI. High variability in the denominator can result in issues in the MI procedure under passive imputation, as illustrated in HippisleyCox et al. (2007). As given in Section 2.4.2, the high variability in the study performed by HippisleyCox et al. (2007) results in a number of imputed values close to zero. These small imputed values cause a very unstable imputed ratio variable, and hence the association with survival were essentially removed. In addition, FMI values are larger for recipient BMI than donor BMI.

Recipient BMI has a stronger association with other variables in the cardiothoracic data set than in the kidney transplant data set. As a result, the FMI values are smaller in the cardiothoracic transplant data set than in the kidney transplant data set.

TABLE 4.3: The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated across the 250 replications. These values are given for each of the imputation models and for each case under an ratio functional form.

			Kidney Transplant Data	**Cardiothoracic Transplant Data**
Donors	BLR	AWO	0.1129 (0.0714)	0.0110 (0.0743)
		APA	0.0535 (0.0370)	0.0566 (0.0411)
		PNP	0.0530 (0.0389)	0.0568 (0.0410)
		LNP	0.0479 (0.0329)	0.0536 (0.0386)
	PMM	AWO	0.1279 (0.0906)	0.0905 (0.0561)
		APA	0.0495 (0.0341)	0.0622 (0.0473)
		PNP	0.0602 (0.0465)	0.0584 (0.0395)
		LNP	0.0571 (0.0436)	0.0595 (0.0455)
Recipients	BLR	AWO	0.7494 (0.1508)	0.6660 (0.1651)
		APA	0.5625 (0.1927)	0.4435 (0.1709)
		PNP	0.5740 (0.1870)	0.4691 (0.1873)
		LNP	0.5829 (0.1709)	0.4796 (0.1905)
	PMM	AWO	0.7674 (0.1527)	0.6407 (0.1782)
		APA	0.5636 (0.1814)	0.4568 (0.1841)
		PNP	0.5807 (0.1812)	0.4604 (0.1718)
		LNP	0.5533 (0.1876)	0.4353 (0.1822)

An appeal of passive imputation is that the imputed derived variable does not lose its functional form. However, passive imputation can result in unrealistic values under a ratio functional form. In Table 4.4 are the minimum and maximum imputed values for PNP and LNP over the preliminary analysis. BMI values less than ten are generally implausible, and occur due to small imputed weight values.

TABLE 4.4: Range for imputed BMI values for the ratio form under passive imputation models.

			Donor BMI		**Recipient BMI**	
	Data	**IM**	**Min.**	**Max.**	**Min.**	**Max.**
PMM	Cardiothoracic	Passive	10.04	77.91	10.45	56.68
		lnPassive	10.15	73.03	11.30	56.68
	Kidney	Passive	4.49	88.56	5.08	62.50
		lnPassive	10.74	105.22	6.10	81.00
BLR	Cardiothoracic	Passive	4.91	61.89	5.18	50.25
		lnPassive	10.78	63.98	9.94	61.61
	Kidney	Passive	4.49	88.56	5.08	62.50
		lnPassive	10.17	87.99	5.78	68.26

For instance, an individual in the kidney transplant data set has an observed weight value of 17kg. When the weight of this male is observed in the preliminary analysis, their height can be imputed as an incompatible value yielding an unrealistically small BMI imputation. Furthermore, under PMM, other individuals missing weight can be imputed as 17kg, resulting in BMI values as small as 5.08 (Table 4.4).

PNP under PMM has 636 instances of recipient BMI less than ten. Of this 636, 635 individuals have a weight of 17kg (the final individual has a weight imputed as 28.6kg). Furthermore, 4319 imputed values are less than ten under PNP for BLR. Of these 4319, 580 have an observed weight of 17kg, and the remaining 3747 imputations have a small imputed weight - ranging between 15 and 38kg. The problem occurs slightly less often for LNP with 509 imputed BMI values less than ten under PMM, and 483 times for BLR. All extreme values under LNP occur when the weight is 17kg. Hence, despite not losing the functional form, passive imputation can result in extreme or implausible values for a ratio functional form.

4.4.2 Additive Functional Form

Trace plots resulting from an additive functional form are shown in Figure 4.5. Note that this is only shown for the second case (ADL), but other cases display a similar trend. As with a ratio functional form, APA2 has a circularity problem. The mean estimates for ADL remain constant throughout all cycles because a linear model is imposed to impute a variable of an additive form. The constituents are among the predictors to impute ADL. To impute each constituent, the remaining constituents and derived variable are among the predictors. In essence the relations are effectively all fixed after the first imputation for each imputed data set. A similar trace plot would occur for a ratio functional form if passive imputation were used to construct the constituents and the derived variable (Vink and Buuren, 2017), that is,

$$\begin{aligned}\gamma_1 &= X_3 \times (\gamma_2/100)^2\\ \gamma_2 &= \sqrt{\gamma_1/X_3} \times 100\\ X_3 &= \frac{\gamma_1}{(\gamma_2/100)^2}.\end{aligned}$$

As a result, the metrics for an APA2 imputation model are omitted from future analyses in this sub-section.

For a similar reason, APA and PNP perform comparably; for APA, ADL is imputed by a linear model containing a sum of the constituents, whereas PNP is constructed by summing the value of the constituents together. Hence, both imputation models have a similar structure and result in similar estimated coefficients. This is supported by the similar values for RB, CR, and AW given in Table 4.5.

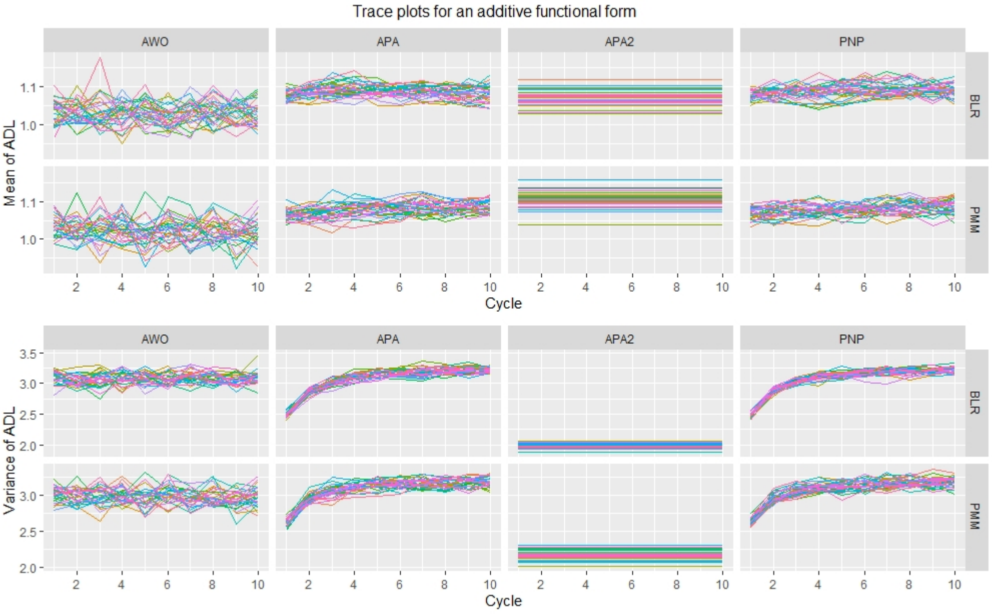


FIGURE 4.5: Trace Plots when BLR and PMM is applied under an additive functional form. Trace plots are given for one replication. Other replications and conditions display a similar trend. The cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of ADL in each of the M imputed data sets; the bottom plots displays the variance of ADL in each of the M imputed data sets.

As observed with a ratio functional form, there is an issue of overcoverage, potentially for the same reason as for a ratio functional form.

Within each case and for each BLR and PMM, hypothesis tests are performed to test for the equality of means between the RB in each imputation model. The null hypothesis is rejected for each test for an additive functional form, yielding sufficient evidence to reject the equality of means. On inspection, the mean pooled estimated coefficients after imposing AWO consistently result in a larger bias than the APA and PNP imputation models.

Hypothesis tests are also performed to test for the equality of means between the AW after imposing each imputation model. The null hypothesis is rejected for each test for an additive functional form, yielding sufficient evidence to reject the equality of AW values. On inspection, the AW is consistently larger after imposing AWO than the AW in the APA and PNP imputation models. In Table 4.6 the mean and standard deviation FMI diagnostic values after fitting a substantive model are given.

The FMI after applying an AWO imputation model is large relative to other imputation models, implying that AWO requires more multiply imputed data sets than the other imputation models (Table 4.6).

APA and PNP result in smaller FMI values than AWO. A small FMI value indicates that the imputed value of the derived variable is not strongly associated with other variables in the imputation model. The FMI is smaller for APA and PNP because the

TABLE 4.5: The RB, CR, and AW values for an additive functional form for the four cases in the CLHLS data set.

			RB	CR (%)	AW
Two constituents	BLR	AWO	-0.01293	98.80	0.0939
		APA	-0.00644	100.0	0.0801
		PNP	-0.00712	100.0	0.0800
	PMM	AWO	-0.02813	86.75	0.0966
		APA	-0.01415	100.0	0.0812
		PNP	-0.01243	100.0	0.0812
Six constituents	BLR	AWO	-0.00683	97.20	0.0524
		APA	-0.00444	100.0	0.0432
		PNP	-0.00396	100.0	0.0433
	PMM	AWO	-0.01588	69.60	0.0564
		APA	-0.00601	100.0	0.0441
		PNP	-0.00724	100.0	0.0442
13 constituents	BLR	AWO	0.00224	100.0	0.0151
		APA	0.00012	100.0	0.0102
		PNP	0.00042	100.0	0.0102
	PMM	AWO	0.00342	99.60	0.0182
		APA	0.00133	100.0	0.0109
		PNP	0.00130	100.0	0.0109
Five constituents	BLR	AWO	0.00092	99.60	0.0122
		APA	0.00012	100.0	0.0092
		PNP	0.00015	100.0	0.0091
	PMM	AWO	-0.00222	99.60	0.0152
		APA	-0.00942	100.0	0.0104
		PNP	-0.00943	100.0	0.0105

constituents are predictors of the derived variable in the APA and PNP imputation models, but not in the AWO imputation model. Hence, the APA and PNP imputation models have better predictors than the AWO imputation model. Additionally, the FMI value decreases for APA and PNP as the number of constituents increases due to an increase in the number of predictors with which the derived variable is associated. FMI is similar regardless of the number of constituents for an AWO imputation model since the constituents are not predictors for the derived variable.

Overall, APA and PNP outperform AWO. APA and PNP are less biased and result in a smaller AW than the estimated coefficients in an AWO imputation model.

In addition, BLR outperforms PMM for an additive functional form. The estimated coefficients are more biased under PMM than BLR (Table 4.5), and the FMI value is generally smaller for BLR than for PMM. A smaller FMI value indicates that the

TABLE 4.6: The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated across the 250 replications. These values are given for each of the imputation models and for each case under an additive functional form.

		Additive			
	IM	**Two**	**Six**	**13**	**Five**
BLR	AWO	0.101 (0.0771)	0.117 (0.0789)	0.552 (0.156)	0.289 (0.202)
	APA	0.052 (0.0416)	0.042 (0.0419)	0.130 (0.0772)	0.098 (0.081)
	PNP	0.042 (0.0190)	0.032 (0.0161)	0.200 (0.101)	0.108 (0.0735)
PMM	AWO	0.012 (0.0917)	0.096 (0.0270)	0.359 (0.177)	0.302 (0.152)
	APA	0.071 (0.0504)	0.075 (0.0588)	0.170 (0.103)	0.125 (0.121)
	PNP	0.047 (0.0164)	0.036 (0.0223)	0.119 (0.0658)	0.133 (0.0765)

procedure is more efficient. Hence there is a suggestion that BLR outperforms PMM for an additive functional form.

4.4.3 Index Functional Form

Trace plots resulting from an index functional form are shown in Figure 4.6. Note that this is displayed for the first case (ADL Index) only, but other cases have a similar trend. As with other functional forms, there is an indication of circularity in the APA2 imputation model, and hence APA2 is omitted from future analyses in this section.

The RB, CR, and AW are given for the three imputation models under each case in Table 4.7. In each case, there is sufficient evidence to reject the null hypothesis that the mean estimated coefficient of the derived variable is equal for the imputation models. On inspection of the RB values in Table 4.7, the estimated coefficient of the derived variable under AWO is more biased than that of PNP or APA. In addition, there is significant evidence that the mean AW values are not equal for the three imputation models, for each case. Upon inspection of the AW values in Table 4.7, AWO generally results in a wider confidence interval in the substantive model. Overcoverage is present for all imputation models.

The mean and standard deviation FMI values after fitting the different imputation models to an index functional form are given in Table 4.8. As observed in Section 4.4.2 for an additive functional form, the FMI values are larger for AWO than APA or PNP.

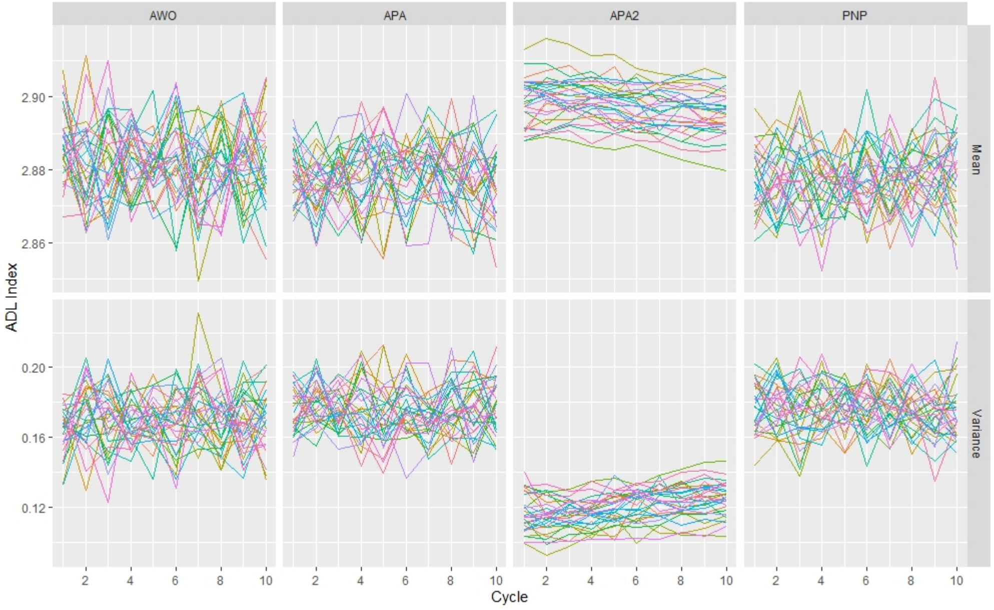


FIGURE 4.6: Trace Plots when PMM is applied for an index functional form. Trace plots are given for one replication and for the ADL Index case. Other cases and replications display a similar trend. The cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of ADL Index in each imputed data set; the bottom plots displays the variance of ADL Index in each imputed data set.

This is because the AWO imputation model has fewer predictors for the derived variable than the APA or PNP imputation model. Taking into account the RB, AW, and FMI values, APA and PNP outperform AWO for an index functional form.

4.5 Conclusion

There is overcoverage present in the CR for the various imputation models under different functional forms. The overcoverage may occur because there is only one source of variability in the preliminary analysis: only missing values are generated. As a result, in a more extensive simulation the data can be simulated from a model to ensure that there is more variability in the approach to investigate the performance of imputation models.

A larger simulation study is conducted, with details given in Section 5. In the simulation study, the data sets are generated, resulting in a higher variability in the approach to investigate the performance of the imputation models. In addition, generating the data sets allows for more consistency across the three functional forms than using motivating data sets; for example, the relationship between the derived variable and other variables in the data set can be the same regardless of the functional form.

TABLE 4.7: The RB, CR, and AW values for an index functional form for the three cases in the CLHLS data set.

		Level	RB	CR (%)	AW
Six constituents	AWO	1	-0.021	100.0	0.0781
		2+	-0.039	96.4	0.1582
	APA	1	-0.006	100.0	0.0839
		2+	-0.018	100.0	0.0154
	PNP	1	-0.012	100.0	0.0712
		2+	-0.019	100.0	0.1480
13 constituents	AWO	moderate	0.010	100.0	0.0780
		mild	0.015	100.0	0.2223
		normal	0.030	99.6	0.1863
	APA	moderate	-0.008	100.0	0.0839
		mild	-0.013	100.0	0.2005
		normal	0.010	100.0	0.1573
	PNP	moderate	0.005	100.0	0.0712
		mild	-0.012	100.0	0.1984
		normal	0.009	100.0	0.1558
Five constituents	AWO	moderate	0.012	100.0	0.0781
		mild	0.027	100.0	0.2036
		normal	0.032	99.60	0.1960
	APA	moderate	-0.008	100.0	0.0839
		mild	0.003	100.0	0.1850
		normal	0.009	100.0	0.1758
	PNP	moderate	-0.007	100.0	0.0712
		mild	0.005	100.0	0.1874
		normal	0.009	100.0	0.1756

Going forward, APA2 is not a good approach and hence will not be investigated further. AWO often results in the highest RB of the imputation models, and largest AW. In addition, AWO requires more multiply imputed data sets due to the lack of predictors in the imputation procedure, evident by the large FMI values for AWO across all imputation models. APA, however, performs well, often resulting in a smaller RB and AW. Going forward, APA and passive-type imputation models are promising approaches.

Under a ratio functional form there is not much distinction in the performance of BLR to PMM. However, under an additive functional form, BLR outperforms PMM, implying that BLR outperforms PMM when the conditional model is defined correctly. However, there are many other conditions that could be the cause of the difference in performance. This is investigated in the simulation study.

TABLE 4.8: The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated across the 250 replications. These values are given for each of the imputation models and for each case under an index functional form.

	Index							
	Six		**13**			**Five**		
IM	**1**	**2+**	**mod.**	**mild**	**normal**	**mod.**	**mild**	**normal**
AWO	0.274	0.309	0.487	0.411	0.423	0.285	0.279	0.313
	(0.188)	(0.194)	(0.233)	(0.144)	(0.172)	(0.201)	(0.192)	(0.210)
APA	0.209	0.180	0.220	0.151	0.176	0.175	0.198	0.215
	(0.126)	(0.090)	(0.130)	(0.061)	(0.129)	(0.138)	(0.190)	(0.085)
PNP	0.214	0.167	0.185	0.253	0.187	0.181	0.227	0.167
	(0.211)	(0.190)	(0.121)	(0.109)	(0.053)	(0.099)	(0.132)	(0.123)

Chapter 5

Simulation Study

A simulation study allows for a controlled environment to compare imputation models in so that some discrepancies in the performance of imputation models can be investigated. In this thesis, a simulation study is designed to investigate the performance of active and passive imputation. The simulation study follows a similar structure to the preliminary analysis, but the data sets are generated from models. By generating the data sets, the parameter estimates in the substantive model can be compared to the true parameter, giving clearer results. Additionally, the generated data and parameter values can follow a similar structure for different functional forms; for example, the ratio and additive functional form can have a similar strength of relationship for the derived variable and covariates. The procedure to generate the data sets, alongside justification for the decisions made, are given in this chapter. Controlling for these external factors can help indicate that the differences for the performance of imputation models is due to functional form.

A ratio, additive, and index functional form are investigated in the simulation study, with additional conditions altered to investigate the impact on the performance of active and passive imputation. In Section 4, only a MCAR missingness structure was investigated. However, studies in the literature have found that the performance of active and passive imputation is contingent on the missingness structure (Table 3.1. As a result, in the simulation study three missingness structures for the derived variable are investigated: one MCAR and two MAR structures. Additionally, the effect that the presence of an auxiliary variable has on the performance of active and passive imputation is investigated. Results for the ratio functional form have been published in Clements et al. (2022). In this chapter, the design of the simulation study is discussed. The methods used to compare active and passive imputation are then given before the results are presented.

5.1 Design of the Simulation Study

The simulation study design is similar to that of the design applied in the preliminary analysis. One key difference is that in the simulation study, the data sets are generated from models. Additional conditions are also investigated: namely, the missingness structure and the presence of an auxiliary variable. The overarching design of the simulation study is as follows:

1. Generate the data from a model,
2. Generate missing values in the data,
3. Use MICE to impute, analyse, and pool the incomplete data set.

These steps are given in further detail next. The procedure is then outlined in more detail in Sections 5.1.1-5.1.3 to specify how the procedure is tailored to the different functional forms.

Generating the Data The generated data is informed from the variables in the motivating data sets outlined in Section 4.1.2. For consistency in the different functional forms, each generated data sets has 2000 rows. In addition, the generated data sets consist of k constituents, $\gamma_1, ..., \gamma_k$, three covariates, X_1, X_2, X_3, an auxiliary variable, Z, and two response variables: survival time and status. X_3 is the derived variable, $X_3 = f(\gamma_1, ..., \gamma_k)$.

For consistency in the data generation for the three functional forms, X_1, X_2, Z are all continuous variables. In addition, the relationship between X_1, X_2, Z and X_3 is similar for the different functional forms. The variables from the preliminary data sets to base X_1, X_2, Z on are firstly chosen such that the relationship between X_1, X_2, Z and X_3 are similar for the different functional forms. A consistent relationship helps ensures that X_3 is influenced by the covariates in the imputation models equally across the different functional forms. X_1 is chosen to have some relationship with X_3 so that the influence of a covariate on the performance of both active and passive imputation is investigated. As a result, $\text{cor}(X_1, X_3) = 0.3$. X_2 is chosen to not have a relationship with X_3, so $\text{cor}(X_2, X_3) = 0$. Finally, since Z is an auxiliary variable, a variable to base Z on is chosen such that $\text{cor}(Z, X_3) = 0.5$. Secondly, the variables to base X_1, X_2, and Z are chosen such that they are significant in the survival substantive model. This significance helps to ensure that the simulated scenario can imitate a real-world scenario.

Because an AFT model can be rearranged to a linear form, data can be easily generated from an AFT model (see equation 2.15). In addition, a Weibull AFT model can be parameterised as a proportional-hazards model (Section 2.4). In the simulation study performed, Survival time is generated from a log-linear Weibull AFT model (equation 2.15). That is, if $\epsilon \sim \text{Gumbel}(0, 1)$ then survival time is generated from:

$$\text{time} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon). \tag{5.1}$$

An exponential distribution is a special instance of the Weibull distribution and occurs when $\sigma = 1$. Since an exponential distribution is simpler to handle than a Weibull distribution, $\sigma = 1$ when generating survival time.

Any generated survival time values above a certain threshold are censored, creating type one right-censoring in the simulation study (See Section 2.4.1). Type one censoring is a simplistic mechanism to investigate, and is unlike the mechanisms in the motivating data sets. However, it is investigated as a starting point to explore the properties of active and passive imputation after multiply imputing a derived variable. For each functional form, the threshold value is chosen such that approximately 15% of survival time values are censored for the generated data sets.

Generating Missing Values In the preliminary analysis, the imputation models perform similarly to one another for a ratio functional form when a small percentage of the derived variable is missing. However, the imputation models perform statistically differently to one another when a larger proportion of the derived variable is missing. PMM is advised against when a large proportion of a variable is missing. In addition, M should roughly be equivalent to the percentage of missing values in a variable, and $\sim$ 30% of missing data is often the proportion chosen in similar studies that compare imputation models; for example the Aurum data set analysed by Morris et al. (2014). Hence, a proportion of 30% missingness is investigated in this simulation study.

To generate 30% of values as missing in a variable, the approach outlined in 4.2 is performed with $r = 0.3$. That is, a Bernoulli distributed dummy variable, W, is randomly generated with $P(w = 1) = 0.3$. When $W = 1$ for an individual, X_3 is set as missing, otherwise, X_3 remains observed. Hence, each value of the derived variable is set to missing independently with probability 0.3. When the derived variable is missing, at least one constituent is too.

When X_3 is MCAR each value of the derived variable is set to missing independently with probability 0.3. Two MAR structures are additionally investigated: MAR1 and MAR2. When the derived variable is MAR1

$$P(X_3 = \text{NA}) = \begin{cases} 0.5, & \text{if } X_1 < \text{median}(X_1) \\ 0.1 & \text{otherwise.} \end{cases}$$

To achieve this,

$$\begin{cases} W \sim Bern(n_1, r = 0.5), & \text{if } X_1 < \text{median}(X_1) \\ W \sim Bern(n_2, r = 0.1) & \text{otherwise.} \end{cases}$$

where n_1 is the number of rows where $X_1 < \text{median}(X_1)$, and where n_2 is the number of rows where $X_1 \geq \text{median}(X_1)$.

When X_3 is MAR2, all X_3 values are set as missing for the smallest 30% of X_1 values. When the derived variable is missing, at least one constituent is too. The specifics on setting the constituents as missing for each functional form are given in detail in Sections 5.1.1-5.1.3

MICE The incomplete data set is imputed by either an active or passive imputation model. MICE is performed with $M = 30$. Step 3 in the simulation procedure is performed for both BLR and PMM for all functional forms. The performance of the imputation models under both BLR and PMM are investigated for the reasons outlined in Section 4.2. Logistic regression is investigated for an additive and index form when the incomplete variable is binary, outlined in more detail in Sections 5.1.2-5.1.3.

As shown in equation 5.1, survival time is generated from a log-linear exponential AFT model. The exponential AFT model fitted to the multiply imputed data sets is:

$$T_i = \exp(\mu + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon).$$

Here, $\mu, \beta_1, \beta_2, \beta_3$ are the intercept, and coefficients for X_1, X_2, X_3 respectively. $\epsilon \sim \text{Gumbel}(0, 1)$ where Gumbel is introduced in Section 2.4 and $\sigma = 1$.

The substantive model fitted to each multiply imputed data set is therefore an exponential AFT model where

$$surv(time, status) \sim X_1 + X_2 + X_3$$

After each replication, the estimated coefficients are pooled by Rubin's Rules, yielding $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ where $\hat{\beta}_0$ is the pooled estimate of the intercept, $\hat{\beta}_1$ is the pooled estimate of the X_1 covariate, and $\hat{\beta}_2$ is the pooled estimate of the X_2 covariate. The substantive model is then pooled via Rubin's Rules where the parameter of interest, Q, is β_3.

The imputation models introduced in the preliminary study (Section 4.2) are investigated in the simulation study, with the exception of APA2.

In an AWO imputation model, $X_3 \sim X_1 + X_2 + \text{time} + \text{status} + \varepsilon$ for $\varepsilon \sim N(0, \sigma^2)$ (Section 4.2) The set of imputation models under APA for $\forall \gamma_k, k = 1, ...K$ is

$$\gamma_k \sim \sum_{k=1}^{k-1} (\gamma_k) + \sum_{k=k+1}^{K} (\gamma_k) + X_1 + X_2 + \text{time} + \text{status} + \varepsilon.$$

The imputation model for the derived variable is

$$X_3 \sim \sum_{k=1}^{K} (\gamma_k) + X_1 + X_2 + \text{time} + \text{status} + \varepsilon.$$

The set of imputation models under PNP for $\forall \gamma_k, k = 1, ...K$ is

$$\gamma_k \sim \sum_{k=1}^{k-1} (\gamma_k) + \sum_{k=k+1}^{K} (\gamma_k) + X_1 + X_2 + \text{time} + \text{status} + \varepsilon.$$

The derived variable is constructed from the constituents

$$X_3 = f(\gamma_1, ..., \gamma_k).$$

When an auxiliary variable is present, Z is an extra predictor for the imputation models outlined above.

To accommodate the additional conditions, Steps 2-3 of the simulation are repeated to account for the different missingness mechanisms. For each missingness mechanism, Step 3 is again repeated to account for whether the auxiliary variable is a predictor or not, and additionally repeated for each imputation model.

The simulation is repeated for 1000 replications in line with the sample size calculation given in Burton et al. (2006), outlined next. The parameter of interest, Q, is the true coefficient of the derived variable in the substantive model. Assume that Q is normally distributed, then the estimate of the parameter of interest, $\hat{Q} \sim N(Q, \sigma^2/n)$ for variance, σ^2 and number of replications, n. The $100(1-\alpha)\%$ confidence interval for this estimate is

$$\hat{Q} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Let δ be the level of accuracy. Hence, an estimate at a 5% accuracy of Q is $\delta = 0.05 \times \hat{Q}$. Then

$$\delta = z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Rearranging for n gives

$$n = \left(\frac{z_{1-\alpha/2}\sigma}{\delta}\right)^2.$$

To produce an estimate within a 5% accuracy of Q depends on the value of Q and σ.

$$n = \Big(\frac{1.96 \times \sigma}{0.05 \times Q}\Big)^2.$$

Q varies depending on the functional form (See sections 5.1.1-5.1.3), along with σ. In Table 5.1 are the values of Q and σ in the motivating data sets where the variables that X_1, X_2, and X_3 are based on are fitted in the model in the complete-case motivating data sets. In addition, the Q and σ values in the generated data are given. The selected values of Q and σ in the generated data and given in detail in Sections 5.1.1-5.1.3. The sample sizes vary between at least 5 and at least 694 replications are required to produce an estimate $\hat{Q}$ within a 5% accuracy of Q. The number of replications is set to 1000 to account for this fluctuation. In addition, recall from Section 4.3.4 that the calculated confidence intervals in the preliminary analysis were fairly wide. Hence, the simulation study is repeated for 1000 replications. This results in a narrower confidence interval for both the Wilson Score and Agressi-Coull of (93.7%, 96.3%).

TABLE 5.1: Sample size for different values of Q and σ. The variables that X_1, X_2, and X_3 are based on are fitted in the model in the complete-case motivating data sets to get values of Q and σ. Additionally, the Q and σ values in the generated data are given. The selected values of Q and σ in the generated data and given in detail in sections 5.1.1-5.1.3.

Functional Form	X_3 **variable**	Q	σ	n
Ratio	Kidney-RBMI	0.015	0.009	568.0
Ratio	Cardiothoracic-RBMI	0.009	0.006	693.8
Additive	Orientation	0.129	0.075	61.5
Index	MMSE Index	0.450	0.043	4.70
Ratio	Generated	0.050	0.010	61.50
Additive	Generated	0.010	0.050	384.2
Index	Generated	0.450	0.050	19.00

5.1.1 Ratio Functional Form

Specifications for the simulation design under a ratio functional form are discussed in this section.

Generating the Data The generated variables follow a structure similar to that of the two motivating data sets discussed in Section 4.1. The generated variables primarily follow a structure similar to the kidney transplant data set because it is the data set that initiated the research. However, information from the cardiothoracic data set is additionally consulted.

The two constituents, γ_1, γ_2, are generated to follow a structure similar to that of weight in kg and height in cm, respectively. Recall from Section 2.4 that Gumbel(μ, δ) with a location, μ, and scale, δ. To account for the skewed distribution in the weight variable, $\gamma_1 \sim$ Gumbel$(64, 14)$. γ_2 is generated from a linear regression model with γ_1 and $\log(\gamma_1)$ as the explanatory variables to account for the non-linear relationship for height and weight (Figure 4.2). To reflect the distribution of the height variables, $\epsilon \sim N(0, 8.6^2)$:

$$\gamma_2 = -36.0 - 0.36\gamma_1 + 54.0 \log(\gamma_1) + \epsilon.$$

The coefficient values, $-36.0, -0.36,$ and 54.0 are the estimated coefficients in the linear model, height $=$ weight $+ \log($weight$)$ calculated from the recipients data in the kidney transplant data set. The generated constituents follow a structure based on the recipient height and weight instead of the donor height and weight because the relationship for suitable X_1, X_2 variables and X_3 have a correlation of approximately 0.3 and 0, respectively. These correlation values are ideal since there is then consistency when generating data in other functional forms. X_3, based on a structure similar to BMI, is subsequently calculated by

$$X_3 = \frac{\gamma_1}{(\gamma_2/100)^2}.$$

The two covariates, X_1 and X_2 are next generated. X_1 is generated to follow a structure similar to recipient age. X_1 is generated from a linear regression model with both constituents, and the derived variable as predictors to maintain a relationship for X_1 and γ_1, γ_2, X_3. An error term is additionally included such that $\epsilon \sim N(0, 13^2)$ to reflect the roughly normal distribution in recipient age. Additionally, the standard deviation of recipient age in the motivating data set is 13, so $\sigma = 13$:

$$X_1 = 3.2 - 0.12\gamma_1 + 0.14\gamma_2 + 1.18X_3 + \epsilon.$$

The coefficient values, $3.2, -0.12, 0.14, 1.18$, are the estimated coefficients calculated when fitting age $=$ weight $+$ height $+$ BMI for recipients in the motivating data set. This model can result in age values less than 20. The CDC (2015) recommends that BMI is not calculated for individuals under 20 years old. As a result, if $X_1 < 20$, then X_1 is re-generated by $X_1 \sim U(20, 100)$. As a result, $\text{cor}(X_1, X_3) \approx 0.3$.

X_2 is generated to follow a structure similar to donor age, so $X_2 \sim N(40, 10^2)$. The parameter values are chosen so that X_2 reflects donor age in the kidney transplant data set. If $X_2 < 20$, X_2 is re-generated by $X_2 \sim U(20, 45)$ to ensure the age is in a reasonable range. X_2 has no relationship with X_3, so $\text{cor}(X_2, X_3) \approx 0$.

An auxiliary variable, Z, is simulated to be associated with BMI. Z is based on 'waist measurement' from the US National Health and Nutrition Examination Survey (NHANES). Z is based on 'waist measurement' from same data set investigated in

Wagstaff et al. (2009) which can be found at CDC (2000).

$$Z = -8.8 + 0.21\gamma_2 + 2X_3 + \epsilon,$$

where $\epsilon \sim N(0, 16^2)$. σ is slightly inflated from the standard deviation of waist measurement in the motivating data set to decrease the correlation between X_3 and Z to approximately 0.5. This approach can result in small values, so if $Z < 40$, Z is re-generated by $Z \sim U(40, 150)$ resulting in $\text{cor}(Z, X_3) \approx 0.5$.

Finally, survival time and a censoring indicator are generated. Survival time is calculated by

$$\text{time} = \exp(6 - 0.02X_1 - 0.02X_2 + 0.05X_3 + \epsilon).$$

The coefficient values are based on the coefficients when fitting the model

$$Surv(time, censored) \sim \text{recipient age} + \text{donor age} + \text{recipient BMI}$$

in the kidney transplant data set. However, the coefficient for X_3 is slightly inflated to have a clearer effect, and the intercept is slightly reduced to avoid too large values. $\epsilon \sim \text{Gumbel}(0, 1)$, so the survival time is exponentially distributed as discussed in Section 5.1. The survival time is censored at 500 for any observation with a calculated survival time greater than 500. In total, approximately 15% of observations are censored similar to in the motivating data set.

Generating Missing Values As outlined in Section 5.1, X_3 is missing for 30% of rows. If X_3 is missing, then at least one constituent needs to be generated as missing. To set which constituents are missing a dummy variable, W_1, is randomly generated for each row of the data set. When $w = 0$ (that is, if the derived variable is observed), $P(w_1 = 0) = 1$. When $w = 1$, $P(w_1 = 1) = P(w_1 = 2) = P(w_1 = 3) = 1/3$. When $w_1 = 0$ for an individual, both constituents are observed. When $w_1 = 1$ for an individual, weight is set as missing for that individual. When $w_1 = 2$, the individual's corresponding height is set as missing. When $w_1 = 3$, both constituents are missing. This account for the scenarios where only height is observed, only weight is observed, and when both constituents are missing.

MICE As mentioned in Section 5.1, the performance of active and passive imputation is investigated under both BLR and PMM for a ratio functional form. Additional approaches to impute a ratio functional form are investigated. The procedure and results are given in Appendix E.

The four imputation models discussed in 5.1 are investigated with two constituents. Hence, for a ratio functional form, $k = 2$. Hence, the set of imputation models under APA is

$$\gamma_1 \sim \gamma_2 + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$\gamma_2 \sim \gamma_1 + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$X_3 \sim \gamma_1 + \gamma_2 + X_1 + X_2 + \text{time} + \text{status} + \varepsilon.$$

The set of imputation models under PNP is

$$\gamma_1 \sim \gamma_2 + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$\gamma_2 \sim \gamma_1 + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$X_3 = \frac{\gamma_1}{(\gamma_2/100)^2}.$$

Recall in Section 4.2 that the LNP imputation model is introduced. The set of imputation models under LNP is investigated in the simulation study, and the set of imputation models is

$$\gamma_1^* \sim \gamma_2^* + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$\gamma_2^* \sim \gamma_1^* + X_1 + X_2 + \text{time} + \text{status} + \varepsilon$$

$$X_3 = \exp(\gamma_1^* - 2 * (\gamma_2^* - \log(100)))$$

where $\gamma_1^* = \log(\gamma_1); \gamma_2^* = \log(\gamma_2)$.

For all imputation models, $\varepsilon \sim N(0, 1)$.

For a ratio functional form, the substantive model is

$$y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

where survival time and censoring is denoted by y. Hence, rearranging the substantive model gives:

$$X_3 \sim \alpha_0 y - \alpha_1 X_1 - \alpha_2 X_2$$

This rearrangement is similar to the active-type imputation models defined for X_3. As a result, when X_3 is imputed actively, the substantive model and the imputation model are compatible due to this linearity. Using that $X_3 = \frac{\gamma_1}{\gamma_2^2}$, the rearranged substantive model give:

$$\gamma_1 \sim \alpha_0 \gamma_2^2 y - \alpha_1 \gamma_2^2 X_1 - \alpha_2 \gamma_2^2 X_2 \tag{5.2}$$

$$\gamma_2^2 \sim \alpha_0 \frac{1}{\gamma_1 y} - \alpha_1 \frac{1}{\gamma_1 X_1} - \alpha_2 \frac{1}{\gamma_1 X_2}. \tag{5.3}$$

These implied conditional distributions of γ_1 and γ_2 display a non-linear relationship between the constituent and outcome variable. This lack of linearity is not reflected in the imputation models when imputing γ_1 and γ_2. Hence, it may be concluded that the coefficient estimates are attenuated under a PNP model. Morris et al. (2014) found a LNP imputation model, however, retains the linearity, as given in Section 2.4.2. Hence, LNP should be an improvement on PNP in the simulation study.

5.1.2 Additive Functional Form

Specifications for the simulation design under an additive functional form are discussed in this section.

Generating the Data The generated variables follow a structure similar to that of the CLHLS data set discussed in Section 4.1.2. In the preliminary analysis, active imputation performed relatively similarly to passive imputation for each case. Therefore, to investigate an additive functional form in this simulation study, the derived variable is constructed from five binary variables, since it is less computationally intensive than using too many variables.

X_3 is based on the five questions in the MMSE study that construct the "orientation" score. Recall that orientation score is discussed in Section 4.2.2. X_3 has five binary constituents, $k = 1, ..., 5$, so $X_3 = \sum_{k=1}^{5} \gamma_k$. If one question is answered correctly, $\gamma_k = 1$. Otherwise, $\gamma_k = 0$. In Figure 5.1 the relative frequencies are displayed for the constituents of the orientation variable. If one question is answered correctly, it is likely the others are also answered correctly. Similarly, if one question is answered incorrectly, it is likely that the others are also answered incorrectly (Figure 5.1).

There are a 5077 variables in the CLHLS data set (ICPSR 36692). There are certain criteria when selecting a suitable variable to base X_1, X_2, Z on: the variable should be numeric and have a correlation with orientation of approximately $0.3, 0, 0.5$ respectively for consistency with a ratio functional form. In addition, variables commonly present in other data sets, such as demographic variables, are considered since they are more representative of real-world data sets. Suitable variables for X_1, X_2, Z are weight, diastolic blood pressure, and age respectively $(\text{cor}(\text{weight}, \text{orientation}) = 0.24, \text{cor}(\text{diastolic}, \text{orientation}) = 0.10, \text{cor}(\text{age}, \text{orientation}) = -0.37)$.

Summary statistics of the weight, diastolic blood pressure, and age variables are given in Table 5.2. Weight and diastolic blood pressure are positively skewed. The small kurtosis value for age indicates there are few values in the tails for the age variable. In addition, there is not much skewness in the age variable.

The constituents are generated based on the orientation variable in the MMSE data. In the CLHLS data, denote the first constituent in the orientation variable by $C11$. Then, in the CLHLS data, 13.4% of the $C11$ variable is 0, and 86.6% is a 1. Hence, γ_1 is randomly generated such that $P(\gamma_1 = 0) = 0.1; P(\gamma_1 = 1) = 0.9$. For $k > 1$, values of

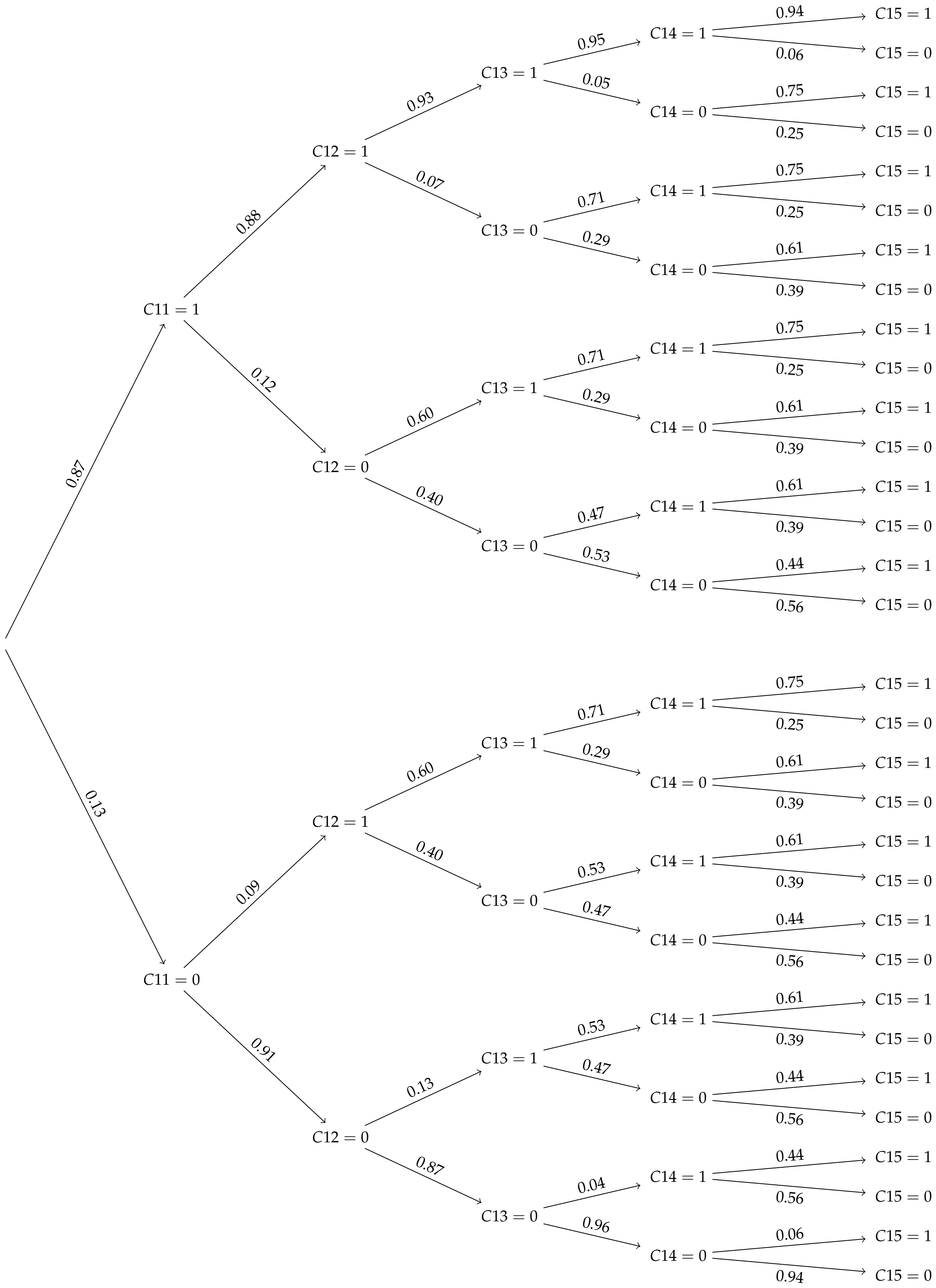


FIGURE 5.1: Relative frequencies for the constituents of the orientation variable (C11-C15) in the CLHLS data set.

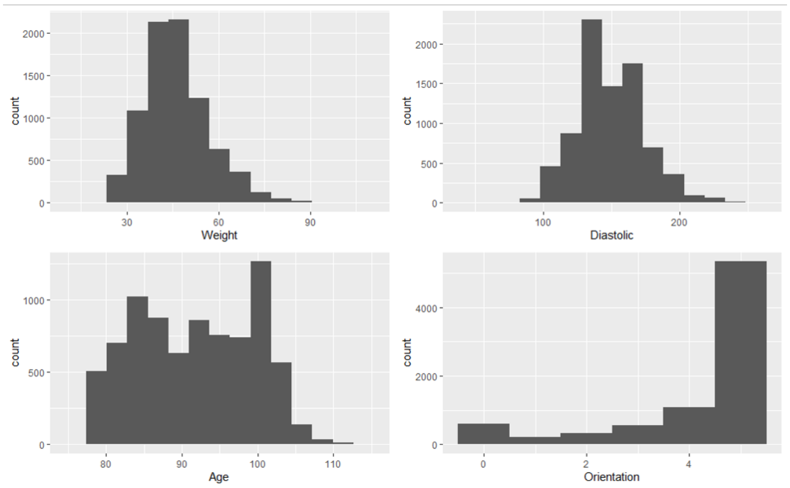


FIGURE 5.2: Distributions of the weight, diastolic blood pressure, age, and orientation variables in the CLHLS data set.

TABLE 5.2: Mean, standard deviation (SD), skewness, and kurtosis for the diastolic blood pressure, weight, and age variables in the CLHLS data set.

	Mean	**SD**	**Skewness**	**Kurtosis**
Diastolic blood pressure	48.5	9.6	0.88	4.09
Weight	149.9	27.4	0.33	3.13
Age	92.8	7.78	0.04	1.90

γ_k depend on $\gamma_1, ..., \gamma_{k-1}$ since the scores for orientation in the CLHLS data set are not independent to one another. The probabilities to generate the constituents is based on the relative frequencies given in Figure 5.1 to one decimal place, and outlined next:

If γ_1 is 1, then $P(\gamma_2 = 1) = 0.9$ and hence $P(\gamma_2 = 0) = 0.1$.
If γ_1 is 0, then $P(\gamma_2 = 1) = 0.9$, and hence $P(\gamma_2 = 1) = 0.1$.
If $\gamma_1 + \gamma_2 = 2$, then $P(\gamma_3 = 1) = 0.9$;
if $\gamma_1 + \gamma_2 = 1$, then $P(\gamma_3 = 1) = 0.5$;
if $\gamma_1 + \gamma_2 = 0$, then $P(\gamma_3 = 1) = 0.1$.
If $\gamma_1 + \gamma_2 + \gamma_3 = 3$, then $P(\gamma_4 = 1) = 0.95$;
if $\gamma_1 + \gamma_2 + \gamma_3 = 2$, then $P(\gamma_4 = 1) = 0.75$;
if $\gamma_1 + \gamma_2 + \gamma_3 = 1$, then $P(\gamma_4 = 1) = 0.50$;
if$\gamma_1 + \gamma_2 + \gamma_3 = 0$, then $P(\gamma_4 = 1) = 0.1$.
If $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 4$, then $P(\gamma_5 = 1) = 0.95$;
If $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 3$, then $P(\gamma_5 = 1) = 0.75$;
if $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 = 2$, then $P(\gamma_5 = 1) = 0.50$;
if $\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4 < 2$, then $P(\gamma_5 = 1) = 0.1$.

$X_3 = \sum_{k=1}^{5} \gamma_k$. X_1 is based on 'weight' since $\text{cor}(\text{weight}, \text{orientation}) = 0.24$. X_1 is generated from a linear model with X_3 and $\exp(X_3)$ as predictors to account for skewness in the relationship between X_3 and X_1. In addition, $\epsilon \sim \text{Gumbel}(0, 6.2)$ so that X_1 follows a structure similar to the weight variable, and to help ensure that

$\text{cor}(X_1, X_3) \approx 0.3$.

$$X_1 = 40.0 + 0.3X_3 + 0.04 \exp X_3 + \epsilon$$

Since X_1 is based on weight, any $X_1 < 30$ are regenerated such that $X_1 \sim U(30, 100)$ to ensure plausible generated values. Overall, $\text{cor}(X_1, X_3) \approx 0.3$.

To generate X_2 a combination of an independent uniform and normal distribution is chosen so that X_2 mimicks the diastolic blood pressure variable in the CLHLS data set.

$$X_2 \sim U(90, 210) + N(0, 10). \tag{5.4}$$

To generate Z,

$$Z \sim 110 - 2.6X_3 + \epsilon$$

where $\epsilon \sim U(-17, 7)$ to follow a distribution similar to age in the motivating data set, and to create a relationship between Z and X_3 similar to that of the motivating variables in the CLHLS data set.

Furthermore, survival time is generated by fitting an exponential model. As a result, the model to generate survival time is

$$time = \exp(6 + 0.007X_1 + 0.01X_2 + 0.1X_3 + \epsilon).$$

$\epsilon \sim \text{Gumbel}(0, 1)$ as explained in Section 5.1. For consistency with a ratio functional form, approximately 15% of survival times are generated to be right censored. To do this, the survival times are censored at 1900.

Generating Missing Values As outlined in Section 5.1, X_3 is generated as missing for 30% of rows. The procedure to generate missing values in the constituents in the simulation study is outlined in Section 4.2.2.

If X_3 is missing, then at least one constituent is generated as missing. To set which constituents are missing a dummy variable, W_1, is randomly generated for each row of the data set. When $w = 0$ (that is, if the derived variable is observed), $P(w_1 = 0) = 1$. When X_3 is observed ($w = 1$), $P(w_1 = 1) = ... = P(w_1 = 5) = 0.2$ since $K = 5$. If $w_1 = 1$ for an individual, a random constituent is set as 'NA' for that row. This continues until $w_1 = 5$ when all constituents are missing for the individual.

MICE Step 3 in the simulation procedure is investigated with PMM in one set of simulations, and BLR in another set of simulations. Additionally, a logistic regression model is fitted under passive imputation since the constituents are binary variables. For both BLR and PMM, the imputation models investigated are AWO, APA, and PNP, as defined in Section 5.1.1.

The imputation models defined in Section 5.1 are investigated with $K = 5$ because there are five constituents for an additive functional form.

The substantive model is

$$y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

where survival time and censoring is denoted by y. Hence, rearranging the substantive model gives:

$$X_3 \sim \alpha_0 y - \alpha_1 X_1 - \alpha_2 X_2$$

This rearrangement is similar to the active-type imputation models defined for X_3. As a result, when X_3 is imputed actively, the substantive model and the imputation model are compatible due to this linearity. Under an additive functional form, $X_3 = \gamma_1 + ... + \gamma_5$. Hence, the rearranged substantive models are still linear; for example, for γ_1:

$$\gamma_1 \sim \alpha_0 y - \alpha_1 X_1 - \alpha_2 X_2 - \alpha_3 \gamma_2 - ... - \alpha_5 \gamma_2. \tag{5.5}$$

The implied conditional distributions of the constituents are always linear, and hence follow a relationship similar to that of the defined imputation models in this section for imputing a constituent under an additive functional form.

5.1.3 Index Functional Form

Specifications for the simulation design under an index functional form are discussed in this section.

Generating the Data The generated variables for an index functional form follow a structure similar to that of the CLHLS data set discussed in Section 4.1.2. X_3 is based on MMSE Index, and the constituent is based on MMSE. In the CLHLS data, MMSE is a score out of 23, and MMSE Index has four levels: severe, moderate, mild, and normal. A histogram of both MMSE Index and MMSE is given in Figure 4.1. Overall, 12% of participants are in the severe category, 20% are in the moderate category, 28% are in the mild category, and 40% are in the normal category.

In the interest of simplicity, X_3 has two levels: 'low' if $\gamma_1 < 16$ and 'high' otherwise. A cut off value of 16 is chosen because then 'low' encapsulates the individuals who would have scored severe or moderate in MMSE Index, and 'high' encapsulates "mild" and "normal". Hence, X_3 is generated such that

$$P(X_3 = low) = 0.32$$

$$P(X_3 = high) = 0.68.$$

γ_1 is generated using values in X_3. The probabilities chosen to generate γ_1 reflect the distribution of MMSE but also avoid an overly complex set of probabilities:

Then, if $X_3 = \text{'}low\text{'}$:

$$P(\gamma_1 = 0) = 0.135$$

$$P(\gamma_1 = 1) = ... = P(\gamma_1 = 7) = 0.034;$$

$$P(\gamma_1 = 8) = P(\gamma_1 = 9) = 0.044$$

$$P(\gamma_1 = 10) = 0.069$$

$$P(\gamma_1 = 11) = P(\gamma_1 = 12) = 0.075$$

$$P(\gamma_1 = 13) = 0.094$$

$$P(\gamma_1 = 14) = 0.106$$

$$P(\gamma_1 = 15) = 0.120.$$

If $X_3 = \text{'}high\text{'}$:

$$P(\gamma_1 = 16) = 0.062$$

$$P(\gamma_1 = 17) = 0.070$$

$$P(\gamma_1 = 18) = 0.074$$

$$P(\gamma_1 = 19) = 0.095$$

$$P(\gamma_1 = 20) = 0.111$$

$$P(\gamma_1 = 21) = 0.153$$

$$P(\gamma_1 = 22) = 0.247$$

$$P(\gamma_1 = 23) = 0.188.$$

The variables that are used to base X_1, X_2, Z on in Section 5.1.2 are applied again an index functional form ($\text{cor}(\text{weight}, \text{MMSE}) = 0.31$, $\text{cor}(\text{diastolic}, \text{MMSE}) = 0.10$, $\text{cor}(\text{age}, \text{MMSE}) = -0.46$). X_2 is generated the same way as for an additive functional form (see equation 5.4). There are some minor changes in the linear models to generate X_1 and Z since the relationship in the motivating data set is now with MMSE score not orientation score. The linear model to generate X_1 is

$$X_1 = 41.7 + 1.1\gamma_1 - 3.4\sqrt{\gamma_1} + \epsilon$$

where $\epsilon \sim N(0, 7.5^2)$ to reflect the distribution of the weight variable that X_1 is based on. Because X_1 is based on weight, any $X_1 < 30$ are removed and regenerated from $X_1 \sim U(30, 100)$ to allow for reasonable values. Overall, $\text{cor}(X_1, \gamma_1) \approx 0.3$. To generate Z,

$$Z = 100 - 0.5\gamma_1 + \epsilon$$

where $\epsilon \sim U(-10, 10)$ to follow a distribution similar to age in the motivating data set. Overall, $\text{cor}(Z, \gamma_1) \approx -0.5$.

In the motivating data set, MMSE Index is transformed to a binary variable, like observed in Lagona and Zhang (2010). The two levels are 'low' and 'high'. An exponential model is then fitted to generate a 'time' variable. The coefficient values remain the same for the exponential AFT model in the CLHLS data and the generated data.

$$\text{time} = \exp(5.8 + 0.01X_1 + 0.001X_2 - 0.45X_3 + \epsilon)$$

where $\epsilon \sim \text{Gumbel}(0, 1)$ to ensure an exponential survival time. For consistency with other functional forms, approximately 15% of observations are censored. Hence, if time > 920, time is right-censored.

Generating Missing Values The approach to generating missing values in X_3 is given in Section 5.1. A derived variable with an index functional form consists of one constituent. Hence γ_1 is missing if X_3 is missing.

MICE Active and passive imputation are both investigated under PMM and BLR for an index functional form. Under active imputation for BLR, X_3 is first transformed to a numeric variable before imputation taking values '1' or '2' for 'low' or 'high' respectively. After imputation by BLR, X_3 is transformed back into a logistic variable where X_3 is 'low' for values of one or less, and 'high' otherwise.

The imputation models defined in Section 5.1 are investigated with $k = 1$ since there is one constituent.

Additionally, a logistic regression model is investigated for the active imputation models since X_3 is a binary variable. Under APA, γ_1 is imputed by PMM when X_3 is imputed by logistic regression.

5.2 Methods to Compare Imputation Models

The methods outlined in Section 4.3 are used to evaluate the performance of the imputation models.

When calculating the Raw Bias (RB), Coverage Rate (CR), and Average Width (AW), the parameter of interest, Q, denotes the true coefficient of the derived variable in the AFT model fitted when generating the data sets. The estimated parameter of interest, $\bar{Q}_M$ denotes the pooled value of a parameter estimate in the substantive model, Q, over M multiply imputed data sets.

There are some minor alterations when performing hypothesis testing in the simulation study compared to the preliminary analysis. Namely, in the simulation study $n = 1000$ since there are 1000 replications. Furthermore, the Bonferroni correction is altered to account for the additional conditions investigated in the simulation study. Hence, a Bonferroni correction is calculated by dividing α by the number of different condition combinations. For example, for a ratio functional form both BLR and PMM are investigated as well as three missingness structures and whether an auxiliary variable is present in the imputation model or not. Hence, a Bonferroni correction is calculated for a ratio functional form by dividing the significance level, α, by $2 \times 3 \times 2 = 12$.

In addition, paired comparison tests can be performed to compare two imputation models to each other (Johnson and Wichern, 2007). This is useful if there are only two imputation models to compare, such as for an additive functional form when logistic regression is applied only to the active-type imputation models. Additionally, if the null hypothesis in a Hotelling's T^2 test is rejected, a paired comparison test (PCT) can be performed to test for the equality of means of a metric between two specific imputation models.

Recall from Section 4.3.3 that a t-test is performed to determine whether a specific value, μ_0, is a plausible value for the population mean, μ. In a paired comparison test, the aim is to determine whether there is a difference in means in two populations.

Using Johnson and Wichern (2007), denote by $X_{11}, ..., X_{n1}$ a random sample such that $X_{j1} \sim N(\mu_1, \sigma^2)$ for $j = 1, ..., n$. Then X_{j1} is the response to treatment 1 for individual j. Similarly, X_{j2} is the response to treatment 2 for individual j, where $X_{j2} \sim N(\mu_2, \sigma^2)$. Let D_j be the difference in values for X_{j1} and X_{j2} for individual, j:

$$D_j = X_{j1} - X_{j2}, j = 1, ..., n,$$

$$D_j \sim N(\delta, \sigma_d^2)$$

Then the null and alternate hypothesis to test for a difference in means between the two treatments are

$$H_0 : \delta = 0 \qquad H_1 : \delta \neq 0,$$

where H_0 is the null hypothesis, and H_1 is a two-sided alternative hypothesis.

The test statistic to test the null hypothesis is

$$t = \frac{\bar{D}_j - \delta}{s_d / \sqrt{n}},$$

where $\bar{D}_j = \frac{1}{n} \sum_{j=1}^{n} D_j$ and $s_d^2 = \frac{1}{n-1} \sum_{j=1}^{n} (D_j - \bar{D})^2$. The test statistic, t, has a t-distribution with $n - 1$ degrees of freedom.

H_0 is rejected at a significance level, α, and hence there is significant evidence that there is a difference in the means of the two treatments, if

$$t = \frac{\bar{D}_j - \delta}{s_d / \sqrt{n}} > t_{n-1}(\alpha/2),$$

where $t_{n-1}(\alpha/2)$ is the upper $100(\alpha/2)^{th}$ percentile of the t-distribution with $n-1$ degrees of freedom.

Hence a t-test can be performed to compare two imputation models to one another where the treatments are the imputation models and $n = 1000$ since there are 1000 replications. To test for an equality of means between the estimated pooled coefficient for two imputation models, X_{j1} denotes the pooled coefficient value in one imputation model, and X_{j2} denotes the pooled coefficient value in another imputation model.

Finally, the Wilson Score and Agressi-Coull confidence intervals in equations 4.2-4.3 respectively are recalculated from the preliminary analysis since n has increased from to 1000 in the simulation study. For both confidence intervals, An estimate lower than 0.937 indicates undercoverage because if $\hat{pi} < 0.937$, 0.95 is not in the confidence intervals when calculating the upper limit. An estimate greater than 0.963 indicates overcoverage because if $\hat{pi} > 0.963$, 0.95 is not in the confidence intervals when calculating the lower limit. Therefore in the simulation study a reasonable CR is implied for an imputation model under all functional forms if the CR is in (93.7%, 96.3%).

5.3 Results when Comparing Imputation Models

Results after fitting an AFT model to each multiply imputed data set are discussed in this section for ratio, additive, and index functional forms.

5.3.1 Ratio Functional Form

The RB and CR are calculated in the complete generated data sets before any missingness is imposed. To perform this, an exponential AFT model is fitted to the complete generated data sets and the coefficient estimates of β_3 are compared to the true coefficient of β_3. Additionally, the CR is calculated from the estimated confidence intervals that result from the exponential AFT model. The RB in the estimated coefficients in the procedure to generate the data sets is small, (RB = 0.00021). Hence in the simulation study, a relatively large portion of the bias in the coefficient estimates in the imputation models is owing to the bias in the imputation model themselves. The CR does not indicate overcoverage (CR = 96.3%), but is on the upper bound of the

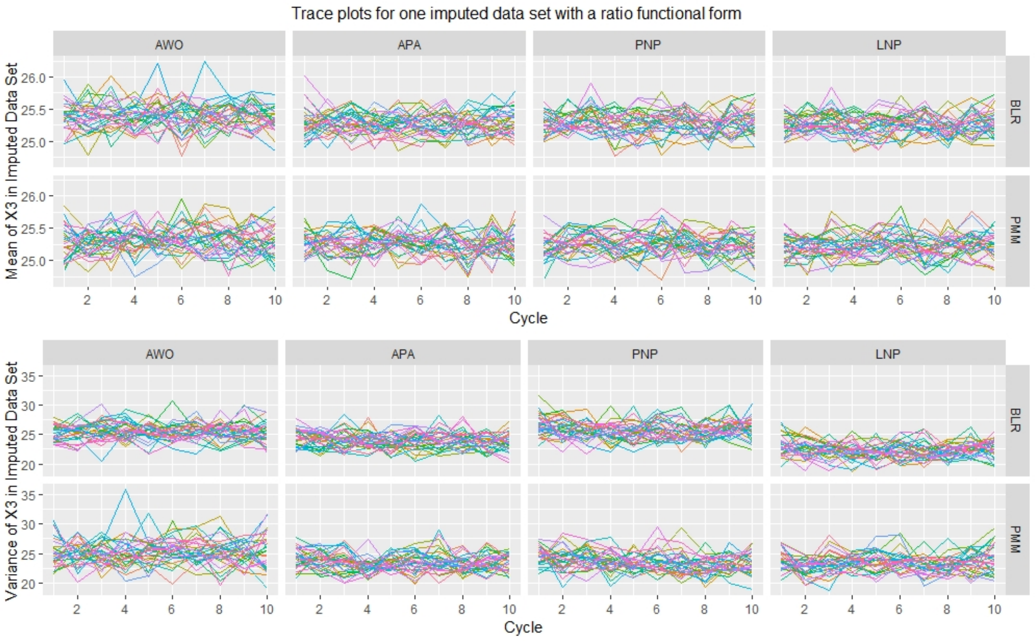


FIGURE 5.3: Trace Plots for each Imputation Model when BLR and PMM is applied for a ratio functional form. Trace plots are given for one replication under a MCAR structure with an auxiliary variable present. Other replications and conditions display a similar trend. The t cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of X_3 in each of the M imputed data sets; the bottom plots displays the variance of X_3 in each of the M imputed data sets.

suitable interval, (93.7, 96.3%). As a result, the imputation model should not be disregarded in this sub-section if there is slight overcoverage present.

Trace plots are displayed in Figure 5.3 where it is evident from the lack of trend that the imputed X_3 values in all imputation models have converged.

The RB, CR, and AW after fitting a substantive model are given in Table 5.3. Hotelling's T^2 tests are performed to test for an equality of means in the metrics after applying the imputation models. The hypothesis test is repeated for the different conditions, and is repeated to separately test an equality in the mean pooled estimated coefficients of X_3, and to test an equality in the AW values. The null hypothesis is rejected in all the tests. Hence, there is sufficient evidence to suggest that there is not an equality of means in either the estimated coefficients of X_3 or in the AW values for the imputation models.

Upon inspection of Table 5.3, under a MCAR scheme the RB is smaller for LNP than other imputation models for both BLR and PMM. There is no overcoverage indicated for LNP. In addition, APA results in small bias for both BLR and PMM, and PNP results in a small RB for PMM.

Under a MAR scheme, PMM results in a poor performance for all the imputation models compared to BLR. The RB in β_3 is larger for all imputation methods under

TABLE 5.3: Metrics for the estimated coefficients of the derived variable in a exponential AFT substantive model for a ratio functional form.

			No Auxiliary Variables			One Auxiliary Variable		
			RB	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
BLR	MCAR	AWO	-0.00072	95.1	0.0239	-0.00044	96.2	0.0231
		APA	-0.00064	96.3	0.0230	-0.00044	96.6*	0.0224
		PNP	-0.00156	95.1	0.0228	-0.00138	95.5	0.0223
		LNP	-0.00013	95.7	0.0231	0.00004	96.7*	0.0226
	MAR1	AWO	0.00083	95.1	0.0231	0.00014	95.8	0.0224
		APA	0.00048	96.2	0.0223	0.00010	96.1	0.0219
		PNP	-0.00054	96.7*	0.0222	-0.00092	95.6	0.0217
		LNP	0.00128	95.3	0.0224	0.00087	96.0	0.0221
	MAR2	AWO	0.00213	93.3*	0.0224	0.00066	94.8	0.0219
		APA	0.00128	95.1	0.0218	0.00038	94.8	0.0215
		PNP	0.00016	95.2	0.0217	-0.00067	94.5	0.0213
		LNP	0.00232	92.4*	0.0220	0.00142	94.6	0.0217
PMM	MCAR	AWO	-0.00096	94.3	0.0238	-0.00091	95.9	0.0229
		APA	0.00057	96.0	0.0232	0.00029	96.7*	0.0222
		PNP	0.00043	95.6	0.0232	0.00012	96.8*	0.0221
		LNP	-0.00030	95.8	0.0231	-0.00008	96.8*	0.0224
	MAR1	AWO	0.00123	95.0	0.0232	0.00077	96.2	0.0222
		APA	0.00240	93.1*	0.0226	0.00147	95.1	0.0217
		PNP	0.00223	93.4*	0.0226	0.00123	95.9	0.0216
		LNP	0.00145	95.3	0.0225	0.00097	96.1	0.0219
	MAR2	AWO	-0.00032	76.7*	0.0231	0.00160	93.7	0.0216
		APA	0.00366	88.4*	0.0223	0.00220	93.3*	0.0213
		PNP	0.00344	88.8*	0.0222	0.00194	92.9*	0.0212
		LNP	0.00262	91.2*	0.0221	0.00163	93.8	0.0216

* denotes that the CR is not in the 95% confidence interval.

PMM than under BLR, with the exception of AWO under MAR2 when no auxiliary variables are present, but there is severe undercoverage. PMM may perform worse than BLR under a MAR-scheme because the observed X_3 values do not follow the distribution of X_3, so under PMM the imputed X_3 values do not follow the distribution of X_3 and are instead skewed.

When BLR follows a MAR scheme, the APA imputation model performs well: the CR in the substantive model is in a good range, and the RB is small. This is also the case for AWO when there is an auxiliary variable present. Under a MAR2 scheme, PNP outperforms other imputation models when there are no auxiliary variables in the imputation model. AWO, APA, and PNP imputation models result in a small RB and CR in a good range when an auxiliary variable is a predictor.

A paired comparison test under BLR yields significant evidence that the mean estimated coefficients of X_3 are not equal for AWO and APA under a MAR2 scheme, and for a MAR1 scheme when an auxiliary variable is not present. APA results in a smaller RB than AWO, suggesting that APA outperforms AWO in these instances. Additionally the AW is smaller under APA, indicating that fitting the constituents as predictors to impute X_3 improves the performance of active-type imputation.

Regardless of missingness structure or the presence of an auxiliary variable, under BLR all PCTs yield significant evidence that the mean estimated coefficients for β_3 differ for APA to PNP. From observing RB, CR, and AW values in Table 5.3, APA generally outperforms PNP under BLR, except for MAR2 data when Z is not present.

A final set of paired comparison tests compares LNP to PNP. Regardless of missingness structure, presence of auxiliary variables, and whether BLR or PMM is applied, log-transforming the constituents significantly alters the resulting mean pooled estimated coefficients. The specifics of this are discussed later in this section.

Overall BLR outperforms PMM. Under a MCAR scheme, LNP outperforms the other imputation models used in the simulation study. When X_3 is MAR-type, APA outperforms the other imputation models (this is also true of PNP if there are no auxiliary variables and a MAR2 structure is imposed).

The effect of the missingness mechanism on the imputation procedure is discussed next in this section. Following this, results for two imputation models under specific conditions are explored since they do not fit the general trend: AWO under a MAR2 scheme for PMM, and PNP when BLR is applied. The effect of auxiliary variables is then discussed, followed by an investigation into the model diagnostics.

5.3.1.1 Missingness Structure

In this section, the affect of the missingness mechanism on the imputation models are discussed.

In Figure 5.6, a frequency density plot is displayed of the observed X_3 values for different missingness structures, and for the generated data. Under a MCAR mechanism, the missing values in a variable follow a distribution similar to that of the observed values in that variable (Figure 5.6). Hence, the distribution of the observed values of the derived variable, $X_{3,O}$, is similar to that of the values in the full generated data, $X_{3,G}$. In the particular MAR-type structure imposed in the simulation, X_3 is likely to be missing for smaller X_1 values since $\text{cor}(X_1, X_{3,G}) \approx 0.3$ and X_3 is missing for small values of X_1. Hence $E(X_{3,O}) > E(X_{3,G})$ (Figure 5.6). As a result, the relationship between observed X_1 and X_3 values weakens under a MAR-type scheme, meaning that X_1 is a less effective predictor for X_3 (mean correlations across all generated data sets: $\text{cor}(X_1, X_{3,G}) = 0.286$; $\text{cor}(X_1, X_{3,MCAR}) = 0.285$; $\text{cor}(X_1, X_{3,MAR1}) = 0.246$; $\text{cor}(X_1, X_{3,MAR2}) = 0.106$).

In addition to the weakened relationship, the distribution of the observed X_3 values is more negatively skewed under a MAR-scheme than in the generated X_3 values. These skewed observed values result in skewed imputed values for BLR and PMM. Under BLR, an incomplete variable is imputed to follow a normal distribution with a mean and variance equal to those of the observed values in that variable. For active imputation, $X_{3,M} \sim N(\text{E}(X_{3,O}), \text{var}(X_{3,O}))$ so X_3 is skewed after imputation. Under PNP, X_3 is constructed from the ratio of γ_1 and γ_2. As a result, when γ_1 is imputed under BLR, X_3 is constructed from the ratio of two normally distributed variables. When only γ_2 is imputed then X_3 is constructed from the ratio of a Gumbel and a normally-distributed variable and hence is negatively skewed. Under PMM an incomplete variable is imputed to follow the distribution of observed values in that variable. Therefore, for both active and passive imputation, X_3 is imputed to be more negatively skewed under MAR-type than under MCAR since $X_{3,O}$ is more negatively skewed. As a result, the distribution of X_3 is more skewed under a MAR2 structure than a MAR1 structure, and more skewed under a MAR1 structure than a MCAR structure.

The relationship between X_3 and survival time is given in Figure 5.7, split by whether X_3 is missing or observed. Under the MAR assumption, the conditional distribution of the partially observed variable, given the fully observed variables, is the same. As a result Figure 5.7 does not imply bias in the relationships but is to illustrate how the MAR assumption affects the relationship between X_3 and survival time. The mean correlations between X_3 and survival time across all multiply imputed data sets are given in Table 5.4. As the mean of the X_3 values increases, the relationship between survival time and the derived variable gets relatively stronger (Figure 5.7; Table 5.4). In the generated data before missingness is imposed, the mean correlation between generated X_3 and survival time is 0.150 (95% confidence interval: (0.149, 0.152)).

- When X_3 is MCAR, $\text{cor}(X_{3,O}, \textit{time}) = 0.150$, 95% CI: (0.148, 0.151).
- When X_3 is MAR1, $\text{cor}(X_{3,O}, \textit{time}) = 0.163$, 95% CI: (0.161, 0.164).
- When X_3 is MAR2, $\text{cor}(X_{3,O}, \textit{time}) = 0.190$, 95% CI: (0.188, 0.192).

The $\text{cor}(X_{3,O}, \textit{time}) \approx \text{cor}(X_{3,M}, \textit{time})$ under a MCAR scheme whereas under a MAR-scheme $\text{cor}(X_{3,O}, \textit{time}) > \text{cor}(X_{3,M}, \textit{time})$ (Figure 5.7). Because X_3 is more likely to be missing when X_3 is small, there are fewer $X_{3,O}$ values that are small but have the survival time at the maximum value of 500. Hence the correlation between time and

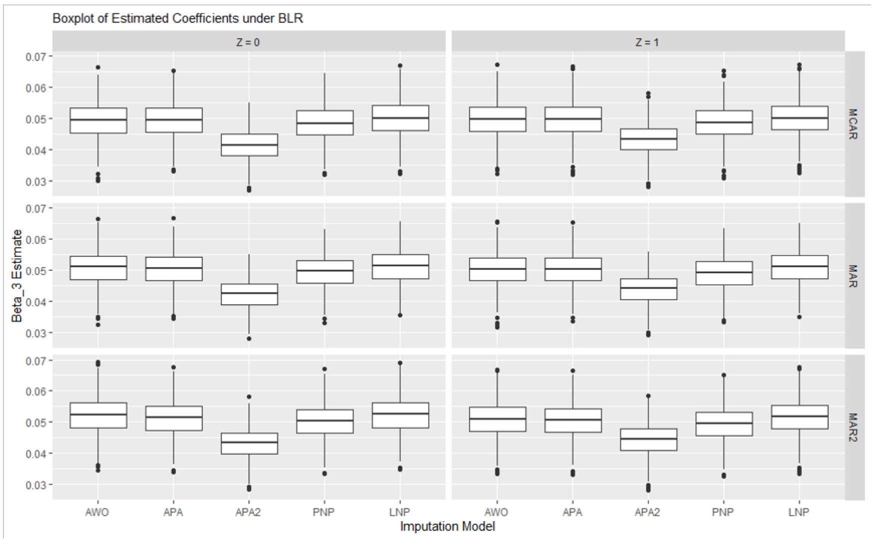


FIGURE 5.4: Boxplot of the pooled estimated coefficients of $\bar{\beta}_3$ for each imputation model when X_3 has a ratio functional form and imputed by BLR. The rows give the different in missingness structures and the columns give the number of auxiliary variables.

$X_{3,O}$ is larger under MAR2 than under MCAR. In the particular simulation performed in this Chapter, $\text{cor}(X_{3,O}, time) > \text{cor}(X_{3,G}, time)$ (Table 5.4) under MAR2, so the imputed X_3 values generally have a stronger relationship with survival time under MAR2, increasing the estimated coefficient of X_3, and generally resulting in an increased RB (Table 5.3). Note that when X_3 follows a MAR-scheme, the correlation between time and the imputed and observed X_3 values are relatively large compared to the correlation between time and the generated X_3 values before missing values are imposed.

Overall, the observed values of X_3 are more skewed under a MAR-type structure, than under a MCAR structure. Due to the skewed observed X_3 values, the relationship between the response variables and X_3 is altered to a greater extent under a MAR-type structure than under a MCAR structure. Missing values in a variable are imputed to follow the relationship of the observed values in that variable. A stronger relationship between survival time and X_3 results in a larger estimated coefficient for the derived variable, and, therefore in this generated data set, results in a higher RB. This is with the exception of PNP under BLR and AWO under PMM MAR2 with no auxiliary variables. These cases are investigated further next.

PNP under BLR For all imputation models, $E(\bar{\beta}_3)$ increases as the missingness structure gets less random. Hence, the RB decreases for PNP under BLR because $E(\bar{\beta}_3) < \beta_3$ under MCAR for PNP BLR. The performance of PNP is next compared with that of LNP.

In Figure 5.8, the distribution of a random 100,000 imputed X_3 values are given, split by whether the numerator, denominator, or both constituents are imputed. When

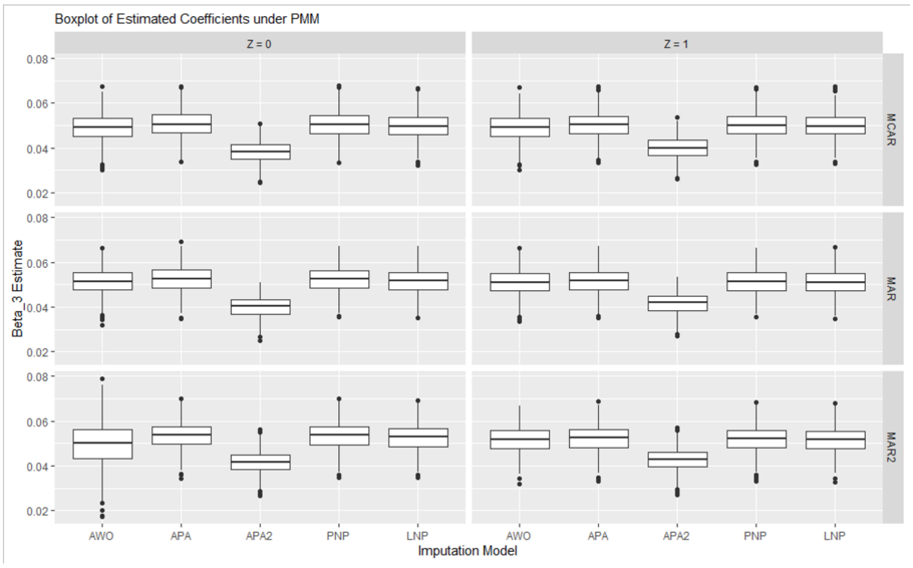


FIGURE 5.5: Boxplot of the estimated coefficients of $\bar{\beta}_3$ for each imputation model when X_3 has a ratio functional form and imputed by PMM. The rows give the different in missingness structures and the columns give the number of auxiliary variables.

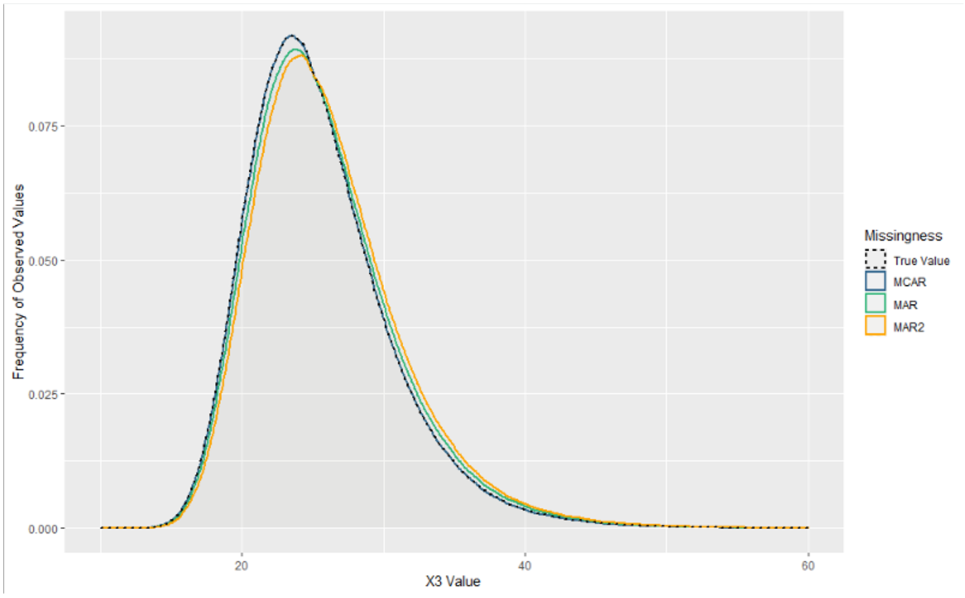


FIGURE 5.6: Frequency density plot of observed X_3 values for the different missingness structures, and for X_3 when no missing values are generated when X_3 has a ratio functional form. Note that this plot is given for X_3 values up to 60.

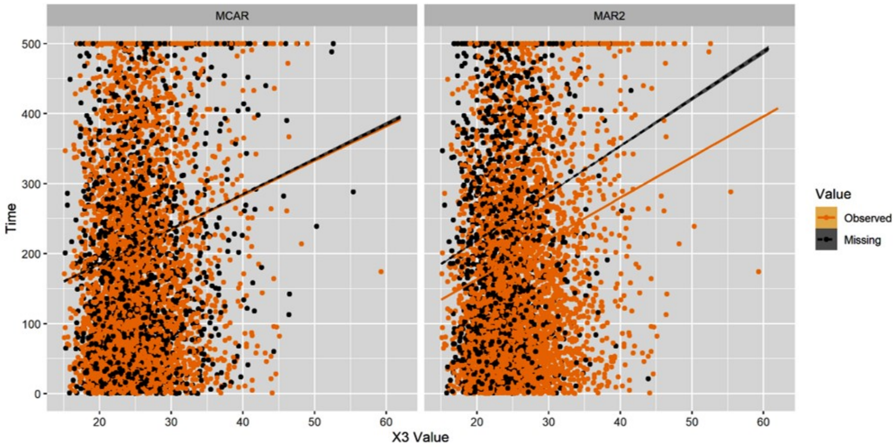


FIGURE 5.7: The relationship between X_3 and survival time for a ratio functional form, split by whether X_3 is missing or observed. These plots are displayed for a MCAR missingness structure and a MAR2 missingness structure. The values plotted are a random sample of three generated data sets, the line of best fits are given after calculating across all generated data sets.

TABLE 5.4: Mean correlation between X_3 and survival time across all multiply imputed data sets when X_3 has a ratio functional form. Note that the mean correlation between X_3 and survival time is 0.150 when no missing values are generated.

		No Auxiliary Variables			One Auxiliary Variable		
		MCAR	**MAR1**	**MAR2**	**MCAR**	**MAR1**	**MAR2**
BLR	AWO	0.150	0.164	0.173	0.150	0.160	0.166
	APA	0.149	0.160	0.166	0.150	0.158	0.162
	PNP	0.149	0.160	0.166	0.149	0.157	0.161
	LNP	0.149	0.160	0.166	0.150	0.157	0.162
PMM	AWO	0.142	0.161	0.158	0.150	0.158	0.162
	APA	0.150	0.162	0.168	0.151	0.159	0.163
	PNP	0.151	0.162	0.168	0.151	0.159	0.163
	LNP	0.149	0.160	0.166	0.150	0.158	0.162

In the generated data, $\text{cor}(time, X_3) = 0.150$ when X_3 is MCAR; $\text{cor}(time, X_3) = 0.163$ when X_3 is MAR1; $\text{cor}(time, X_3) = 0.190$ when X_3 is MAR2.

generating the data, γ_1 is Gumbel distributed and γ_2 is normally distributed. Under PNP BLR, the imputed values of γ_1 are normally distributed and hence the imputed X_3 values are constructed from the ratio of two normally distributed variables. As a result, the imputed X_3 values follow a distribution different from that of the generated X_3 values, $X_{3,G}$ (Figure 5.8).

Under LNP, the imputed log (γ_1) values follow a normal distribution, so the imputed γ_1 values are log-normally distributed. Similarly, γ_2 is log-normally distributed. Therefore, the constructed X_3 values have a distribution closer to X_3 for LNP than for PNP (Figure 5.8). Hence the X_3 values in the multiply imputed data sets follow a distribution closer to that of the generated X_3 values under LNP than PNP. The structure of PNP may explain why the estimated coefficients of X_3 are attenuated under MCAR. As a result, the relationship between X_3 and the outcome variables is weakened and hence the estimated β_3 coefficients are biased under MCAR.

Furthermore, Morris et al. (2014) concluded that LNP outperforms PNP. As outlined in Section 2.4.2, Morris et al. (2014) show that the imputation model and substantive model for a ratio functional form are both linear under LNP. If we simplify the substantive model to just one explanatory variable, X_3,

$$(\log(\gamma_1), \log(\gamma_2)|y) \sim N$$

where y is the outcome variables. Since $\log(X_3) = \log(\gamma_1) - \log(\gamma_2)$,

$$(\log(X_3)|y) \sim N$$

Therefore, the imputation mean function is

$$\log(X_3) = \alpha_0 + \alpha_1 y$$

and the substantive model mean function is

$$y = \beta_0 + \beta_1 X_3.$$

Hence, the imputation model is in the correct form when the constituents are first log-transformed. However, for PNP,

$$(\gamma_1, \gamma_2|y) \sim N$$

$$\implies (X_3|y) \sim N.$$

The imputation model mean functions are

$$\gamma_1 \sim \alpha_0 + \alpha_1 \gamma_2^2 y,$$

$$\gamma_2^2 \sim \alpha_0 + \alpha_1 \gamma_1 \frac{1}{y}.$$

Hence PNP does not have a correct conditional relationship, and LNP should be chosen over PNP because the conditional relationship is approximately correct.

PNP performs like other imputation models under PMM: the bias in the estimated coefficients increases as the missingness structure goes from MCAR to MAR1, and from MAR1 to MAR2. This is potentially because the imputed constituents follow the distribution of the observed constituents, and hence the imputed X_3 values follow the

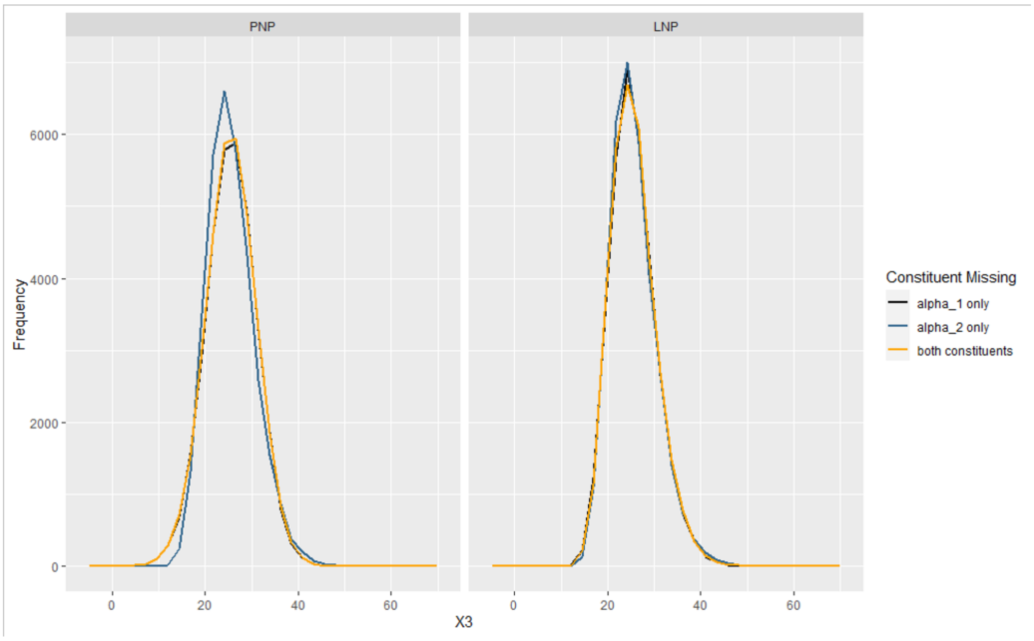


FIGURE 5.8: Distribution of the imputed X_3 values for a ratio functional form, split by which constituent is imputed. These results are based on a random sample of 100,000 imputed values.

distribution of $X_{3,O}$. However, imputed X_3 values do not follow the distribution of $X_{3,O}$ under BLR.

AWO under PMM (with MAR2, no auxiliary variables) The CR for an AWO imputation model when X_3 follows a MAR2 structure and is imputed by PMM without any auxiliary variables present is smaller than the CR values for any other imputation models and any other conditions; that is, other missingness structures, when an auxiliary variable is present, or for BLR. The estimated coefficients of X_3 for each imputation model and condition are displayed in boxplots in Figures 5.4-5.5 for BLR and PMM respectively. Under BLR in Figure 5.4, the pooled estimated coefficients are in a similar range for AWO, APA, PNP, and LNP under the considered conditions. However, it is clear from Figure 5.5 that the variance of the pooled estimates of β_3 is larger for AWO under PMM (with MAR2, no Z) than for other imputation models, and than for BLR.

It has been established in Section 5.3.1.1 that the relationship between X_3 and X_1 is weaker under a MAR2 mechanism than under another missingness mechanism. In addition, the constituents are not predictors for the AWO imputation model, so X_3 has fewer predictors under AWO than for other imputation models. Furthermore, when Z is not a predictor in the imputation model, the AWO imputation model has one fewer predictors than when an auxiliary variable is present. As a result, the lack of predictors contributes to the poor CR under AWO (when MAR2 and Z is not a predictor in the imputation model). Another factor contributing to the poor performance of AWO under MAR2 data is that when PMM is imposed, X_3 is imputed to follow the

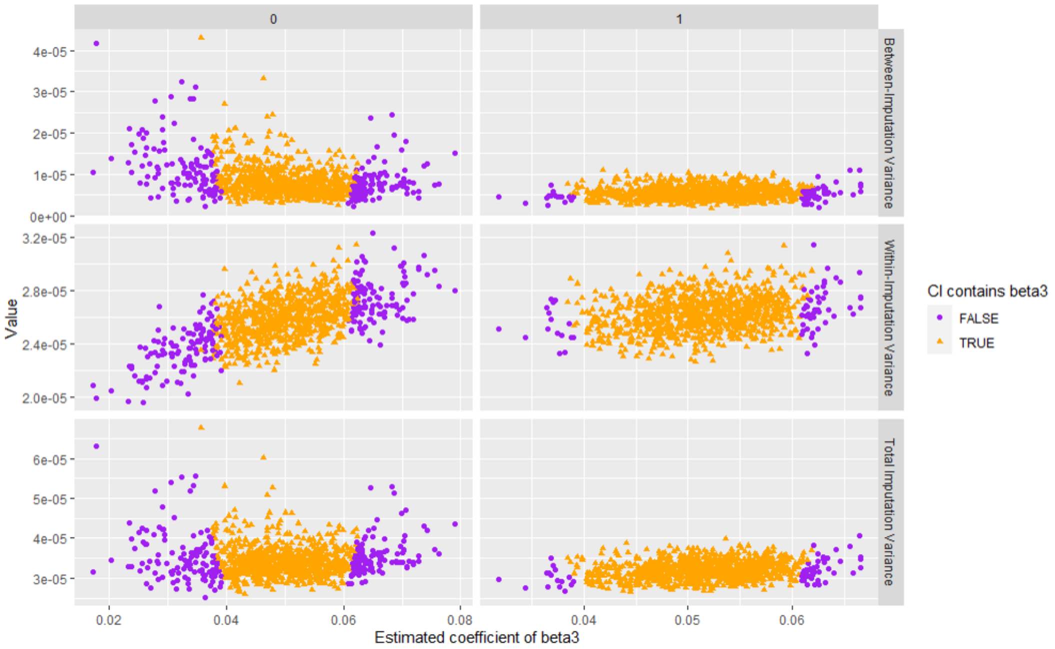


FIGURE 5.9: Total, between-imputation, and within-imputation variances against the estimated pooled coefficient for X_3 for AWO PMM MAR2 when Z is not present ($Z = 0$), and when Z is present ($Z = 1$).

distribution of observed X_3 values. Therefore X_3 values are generally imputed to be larger than for other missingness structures. As a result, X_3 is imputed poorly when AWO is the imputation model under these conditions, and so the CR is low.

A 95% confidence interval using Wald is $\bar{\beta}_3 \pm (1.96 \times SE(\bar{\beta}_3))$. However, the $\bar{\beta}_3$ estimates vary more for AWO under PMM MAR2 without an auxiliary variable than the estimates vary for other imputation models. In the simulation, $sd(\bar{\beta}_{3,AWO}) \approx 1.67 \times E(sd(\bar{\beta}_{3,APA}), sd(\bar{\beta}_{3,PNP}), sd(\bar{\beta}_{3,LNP}))$ where $\bar{\beta}_{3,.}$ denotes the mean pooled estimated coefficients across the given imputation model (under PMM, MAR2, without Z present). If the calculated β_3 estimates are multiplied by a factor of 1.67, the ratio functional form accounts for the increase in variance and the CR is 94.7%.

The $SE = \sqrt{V_T}$, for total variance, V_T. Therefore an increase in the standard error results in a wider confidence interval. A wider confidence interval then results in a larger CR than a smaller confidence interval. Therefore either the between- or the within-imputation variance is underestimated, resulting in undercoverage.

Each $\bar{\beta}_3$ value from the 1000 replicates is plotted against the between-, within-, and total variance for AWO under a ratio functional form when X_3 is MAR2, imputed by PMM, and no auxiliary variables are present. This is displayed in Figure 5.9. This process is repeated with an auxiliary variable present, and additionally given in Figure 5.9. An indicator is given to denote when the resulting confidence interval from the replication contains β_3.

The within-imputation variance tends to be smaller when the pooled coefficient estimate is smaller. As the pooled estimate increases, the variability for the M estimates increases. However, the variability does not increase for the smaller estimated coefficients (Figure 5.9), so the issue of undercoverage may lie in the within-imputation variance.

When Z is not a predictor in the imputation model, the estimated coefficients vary over a larger range than when Z is present ($0.017 \leq \hat{\beta}_3 \leq 0.079, 0.032 \leq \hat{\beta}_3 \leq 0.067$ respectively). When an auxiliary variable is not present, the estimated confidence intervals that contain β_3 (denoted by a triangle in Figure 5.9) indicate a parabola when examining the between-imputation variance, V_B. However, when the confidence intervals for the estimated coefficients do not contain β_3, the V_B values are attenuated. Hence, the undercoverage may be due to V_B values that are smaller than they should be. For example, replication 917 has a relatively small estimated value of β_3. However, there is a relatively large V_B value for β_3. As a result, the confidence interval contains β_3. Most replications with an estimated β_3 value relatively far from β_3 result in a relatively small V_B. Hence, the issue may stem from the between-imputation variance.

In Figure 5.9 the between-imputation variance is plotted against the estimated pooled coefficient value of X_3 for AWO under PMM MAR2 when Z is not a predictor. Two linear models are fitted to the data: one when the CI contains β_3, and the other when the CI does not contain β_3. If a second order polynomial is fitted to a linear model with the V_B as the outcome and estimated coefficients as the explanatory variables (Figure 5.10), the linear model predicts higher values when the CI contains β_3 than when it does not. This indicates the V_B values are too small for the cases when the CI does not contain β_3. Overall the undercoverage may stem from attenuated values of V_B.

5.3.1.2 The Number of Auxiliary Variables

The inclusion of an auxiliary variable enhances the imputation: the AW is narrower and the bias is reduced. Additionally, the CR tends to increase. Often an increase in CR causes it to be outside of the $(93.7, 96.3\%)$ boundary but, as discussed in Section 5.3.1, the CR in the exponential AFT models fitted to the complete generated data is 96.3% . Therefore slight overcoverage observed in Table 5.3 is not of concern. PNP under BLR and AWO under MAR2 BLR do not decrease in bias, but these two instances have previously been highlighted to not follow the general trend. Additionally, the AW decreases for both of these cases when an auxiliary variable is present compared to when it is not, and the CRs are consistently in a good range when an auxiliary variable is present. This is particularly of note for AWO under BLR MAR2 where there was severe undercoverage when an auxiliary variable is not present (CR is 76.7% when Z is not present, and 93.7% when Z is present).

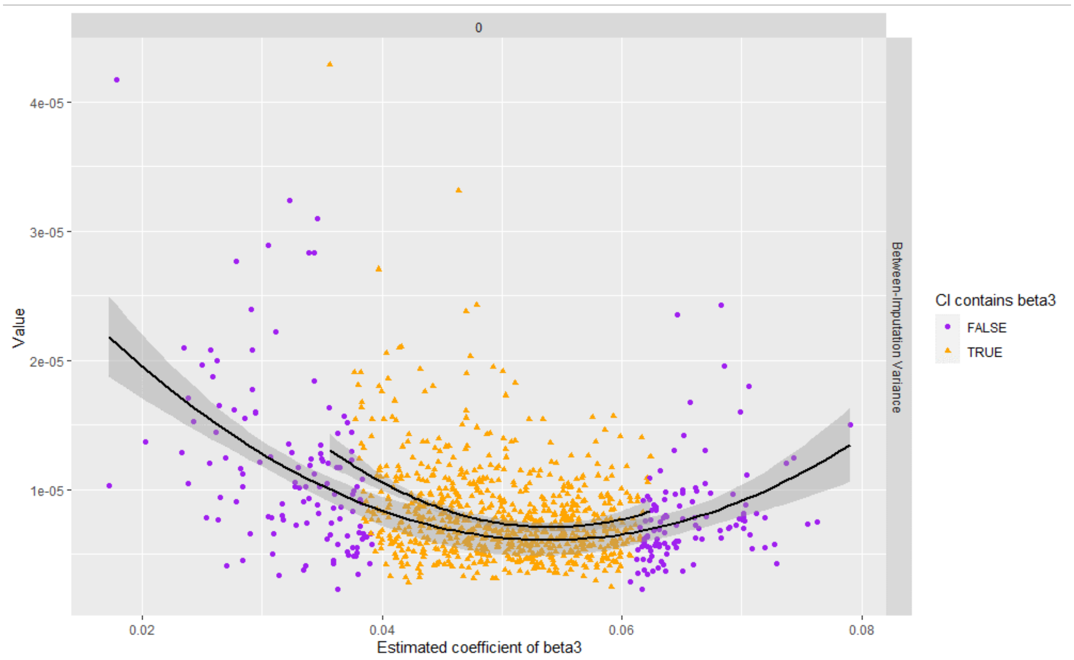


FIGURE 5.10: Between-imputation variance against the estimated pooled coefficient, β_3 for AWO PMM MAR2 no Z. Two linear models are fitted with the outcome variable as the V_B, and the explanatory variables the $\hat{\beta}_3$ as a second-order polynomial. The dotted line represents the linear model when the CI contains β_3, and solid line represents a linear model when the CI does not contain β_3.

With the exception of PNP under BLR for a MAR-type structure, the relative performances are the same for both values of Z.

5.3.1.3 Model Diagnostics

The mean and standard deviation of FMI values from the substantive models are given in Table 5.5. For each condition, a higher proportion of the total sampling variance is attributable to the missing data in X_3 when the imputation model is AWO compared to other imputation models. Hence judging by the mean FMI values, more multiply imputed data sets are required under AWO than for other imputation models. Conversely, LNP results in the smallest FMI value for each condition, implying that imputation by LNP requires fewer multiply imputed data sets than for other imputation models.

The same generated data sets are imputed by the different imputation models so the varying FMI values do not indicate a difference in the proportion of imputed values for imputation models. Instead a high FMI value indicates that the multiply imputed X_3 value is not strongly associated with other variables in the imputation model. Therefore the mean FMI values are larger when an auxiliary variable is not present than when an auxiliary variable is present because there are fewer predictors. Additionally, the mean FMI values are larger for AWO than for other imputation models since the constituents are not predictors for AWO and so there are fewer predictors for X_3.

For AWO, APA and PNP the mean FMI values are smaller under PMM than under BLR, indicating that fewer data sets are needed to impute the incomplete variables under PMM than under BLR. For LNP, the FMI values are virtually the same for BLR and PMM.

TABLE 5.5: Summary diagnostics resulting from the substantive models when X_3 has a ratio functional form. The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated using the 1000 replications.

		No Auxiliary Variables			**One Auxiliary Variable**		
		MCAR	**MAR1**	**MAR2**	**MCAR**	**MAR1**	**MAR2**
BLR	AWO	0.304	0.296	0.286	0.256	0.250	0.243
		(0.0601)	(0.0590)	(0.0564)	(0.0524)	(0.0515)	(0.0505)
	APA	0.244	0.231	0.223	0.209	0.204	0.196
		(0.0528)	(0.0511)	(0.0483)	(0.0460)	(0.0452)	(0.0441)
	PNP	0.259	0.249	0.240	0.223	0.218	0.210
		(0.0537)	(0.0524)	(0.0516)	(0.0466)	(0.0488)	(0.0460)
	LNP	0.224	0.202	0.185	0.193	0.175	0.159
		(0.0495)	(0.0440)	(0.0421)	(0.0433)	(0.0389)	(0.0379)
PMM	AWO	0.281	0.247	0.248	0.245	0.210	0.178
		(0.0557)	(0.0760)	(0.0867)	(0.0511)	(0.0470)	(0.0401)
	APA	0.229	0.208	0.188	0.197	0.179	0.161
		(0.0494)	(0.0477)	(0.0426)	(0.0431)	(0.0415)	(0.0378)
	PNP	0.230	0.207	0.188	0.200	0.180	0.161
		(0.0493)	(0.0456)	(0.0425)	(0.0450)	(0.0406)	(0.0371)
	LNP	0.225	0.199	0.178	0.195	0.174	0.154
		(0.0486)	(0.0420)	(0.0398)	(0.0433)	(0.0410)	(0.0367)

5.3.2 Additive Functional Form

The RB and CR are calculated in the generated data sets before any missingness is imposed for an additive functional form. As with a ratio functional form, this is performed by fitting an exponential AFT model to the complete generated data sets. The RB and CR are then calculated by comparing the estimates in the exponential AFT model with the true parameters. The RB in the procedure to generate the data sets is small and the CR is in the suitable interval (RB = 0.00023; CR = 95.3%). Hence in the simulation study, a relatively large portion of any bias in the coefficient estimates is attributable to the bias in the imputation model themselves.

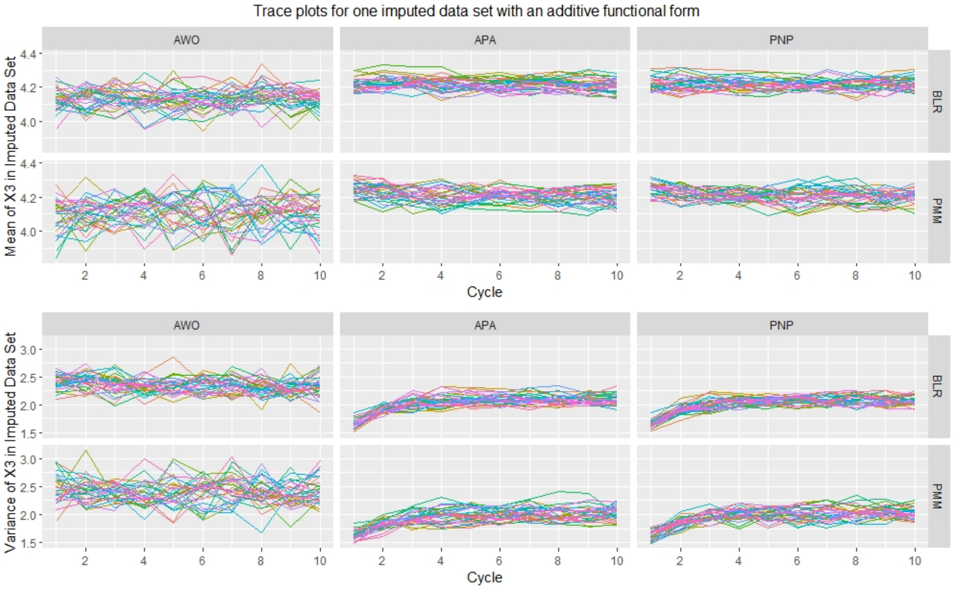


FIGURE 5.11: Trace Plots for each Imputation Model when BLR and PMM is applied under an Additive Functional Form. Trace plots are given for one replication, a MCAR structure, and Z is present. Other replications and conditions display a similar trend. The cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of X_3 in each of the M imputed data sets; the bottom plots displays the variance of X_3 in each of the M imputed data sets.

Trace plots are given in Figure 5.11. Convergence is implied for all imputation models in the trace plots.

The RB, CR, and AW from the substantive models are given in Table 5.6. Hotelling's T^2 test statistic is calculated to test for an equality of means in the estimated coefficients of the derived variable for each imputation model. This test is repeated for the different conditions. In addition, Hotelling's T^2 test statistic is calculated to test for an equality of means in the AW for the imputation models. To test for an equality of means between APA and AWO under a logistic regression MICE structure, a paired comparison test is performed, as outlined in Section 5.2. All hypothesis tests are significant, so there is sufficient evidence to suggest that there is not an equality of means in either the estimated coefficients of the derived variable, or in the AW values for the imputation models. Upon inspection of Table 5.6, the RB and AW values for APA and PNP are very similar. The RB and AW values for AWO is relatively larger than that of APA or PNP, indicating that APA and PNP outperform AWO. The metrics after applying the APA imputation model are similar to those of the PNP imputation model. This is for the same reason as found in the preliminary analysis (Section 4.4.2): in the APA imputation model, the predictors and an error are summed together to construct X_3, and the constituents are highly correlated with X_3. Hence, the imputation model is approximately $X_3 \approx \sum_{k=1}^{K} \gamma_k$. That is, the APA imputation model is performing a very similar calculation to the PNP imputation model.

The coefficient estimates resulting from BLR are less biased than the coefficient estimates after imposing PMM. As established with a ratio functional form, the mean of observed X_3 values is larger under the particular MAR-schemes imposed in this simulation study than a MCAR scheme, resulting in larger mean of imputed X_3 values.

When X_3 is MCAR, passive imputation using logistic regression outperforms all the other imputation models under PMM or BLR, provided the imputation model is defined correctly (that is, has the correct conditional relationship, see Chapter 2.4.2): the estimated coefficients are less biased, the CR is in a good range, and the AW is never larger than the other imputation methods. When X_3 is MAR-type and no auxiliary variables are present, the estimated coefficients of X_3 are less biased under BLR for APA and PNP than under all other imputation models. Additionally, the CR is in a good range. When X_3 is MAR-type and an auxiliary variable is present, the estimated coefficient of X_3 is less biased under BLR than all other imputation models. The CR is in a good range and the AW is smaller consistently for APA and PNP.

5.3.2.1 Missingness Structure

As the missingness gets less random, that is as a MCAR missingness structure moves to a MAR2 missingness structure, the AW increases for all imputation models. This increase indicates that the imputation models perform worse under a stricter missingness structure (Table 5.6). Additionally, the estimated coefficient of β_3 is more biased under a MAR2 structure than under MCAR, with the exception of AWO when an auxiliary variable is present for both PMM and BLR.

The imputation models perform worse as the missingness scheme becomes less random for the same reason under an additive form as for a ratio form: When X_3 is MAR-type, smaller X_3 values are missing so smaller γ_k values are missing. As a result, under a MAR-scheme there are fewer observed values with $\gamma_k = 0$ for an individual than under a MCAR missingness structure. Hence, γ_k is more likely to be imputed to be as one than zero due to the nature of how PMM, BLR, and logistic regression function. Additionally the relationship between X_3 and X_1 weakens due to the missingness structure, and hence X_1 is a worse predictor to impute missing values.

The estimated coefficients of X_3 get less biased for AWO under BLR when an auxiliary variable is present as the missingness becomes less random, suggesting that the bias increases in the other imputation models due to a poor imputation of the constituents. Smaller X_3 values are more likely to be missing under MAR2 than under MAR1, and under MAR1 than under MCAR due to the missingness structure imposed. Hence, larger X_3 values are observed. The missing values are imputed based on the distribution of the observed data for PMM. Further, under BLR the missing X_3 values

TABLE 5.6: Metrics for the estimated coefficients of the derived variable in a exponential AFT substantive model for an additive functional form

			No Auxiliary Variables			One Auxiliary Variable		
			RB	CR (%)	AW	RB	CR (%)	AW
BLR	MCAR	AWO	-0.00228	95.9	0.0780	-0.00180	96.3	0.0752
		APA	-0.00137	95.8	0.0683	-0.00106	95.4	0.0677
		PNP	-0.00138	95.6	0.0683	-0.00106	95.7	0.0678
	MAR1	AWO	-0.00332	95.5	0.0807	-0.00126	96.3	0.0779
		APA	-0.00183	95.5	0.0687	-0.00139	95.1	0.0682
		PNP	-0.00186	95.2	0.0687	-0.00137	94.8	0.0682
	MAR2	AWO	-0.00401	95.1	0.0908	0.00122	95.6	0.0873
		APA	-0.00289	94.7	0.0698	-0.00188	95.0	0.0692
		PNP	-0.00285	94.8	0.0697	-0.00183	95.2	0.0692
PMM	MCAR	AWO	0.00216	96.4*	0.0781	-0.00190	96.1	0.0737
		APA	-0.00404	95.4	0.0691	-0.00375	95.7	0.0679
		PNP	-0.00406	95.9	0.0691	-0.00377	95.5	0.0679
	MAR1	AWO	0.00963	94.6	0.0844	-0.00249	96.0	0.0748
		APA	-0.00442	95.3	0.0699	-0.00460	94.3	0.0683
		PNP	-0.00440	95.8	0.0699	-0.00457	94.5	0.0682
	MAR2	AWO	-0.00565	91.5*	0.1043	0.00168	94.8	0.0802
		APA	-0.00520	95.6	0.0716	-0.00618	94.8	0.0692
		PNP	-0.00528	95.5	0.0714	-0.00611	94.4	0.0693
logit	MCAR	PNP	-0.00108	95.7	0.0683	-0.00078	95.4	0.0676
	MAR1	PNP	-0.00206	94.9	0.0684	-0.00175	94.7	0.0676
	MAR2	PNP	-0.00367	94.4	0.0692	-0.00319	95.2	0.0681

* denotes that the CR is not in the 95% confidence interval.

are imputed based on the mean and variance of the observed X_3 values. As a result, under both BLR and PMM, the constituents are imputed to be larger than the values in the generated data.

5.3.2.2 The Number of Auxiliary Variables

When an auxiliary variable is a predictor in the imputation model, the AW is lower for all imputation models than when Z is not a predictor since Z positively influences the imputation. Additionally, the bias in the estimated coefficient is smaller when an auxiliary variable is present than when it is not, with the exception of APA and PNP when imputed by PMM for a MAR-scheme.

For AWO when an auxiliary variable is not present, X_3 is only associated with X_1, and, as established in Section 5.3.1.1, the relationship between X_1 and X_3 weakens as the missingness mechanism becomes less random. Hence when an auxiliary variable is not present, AWO performs poorer than APA and PNP. However, for APA and PNP there is an association between each constituent and X_1. Additionally, the constituents are all strongly associated with one another (the correlations range from 0.50 to 0.63 across the different γ_k), and associated with X_3 (each correlation is ≈ 0.8). If X_3 is missing, some constituents may be observed and hence influence the imputation of other constituents and X_3. Hence, APA and PNP outperform AWO when an auxiliary variable is not present. As a result, the presence of Z does not benefit APA or PNP as much as AWO under PMM.

5.3.2.3 Model Diagnostics

The mean and standard deviation of FMI values from the substantive models are given in Table 5.7.

Consistent to a ratio functional form, the FMI after applying an AWO imputation model is large relative to other imputation models, implying that AWO requires more multiply imputed data sets than the other imputation models (Table 5.7). APA and PNP result in much lower FMI values than AWO or than with a ratio functional form. The FMI values suggest that under BLR, $9 \leq M \leq 10$, under PMM $10 \leq M \leq 19$, and under logistic regression $10 \leq M \leq 16$.

A larger FMI value indicates that the imputed X_3 value is not strongly associated with other variables in the imputation model. As a result, FMI is smaller when an auxiliary variable is present than when it is not. In addition, there are more constituents that have a stronger relationship with X_3 for an additive functional form than for a ratio functional form (Table 5.5 for a Ratio functional form; Table 5.7 for an additive functional form). In the ratio functional form imposed in this simulation study, the denominator is not strongly correlated with X_3. However, for the additive functional form imposed in this simulation study, all constituents are strongly correlated with X_3. As a result, for the APA and PNP imputation models the RIV values are smaller for an additive functional form than for a ratio functional form.

5.3.3 Index Functional Form

The RB and CR are calculated in the complete generated data sets for an index functional form before any missingness is imposed. The RB in the coefficient estimates in the generated datasets is small and the CR is in the suitable interval (RB = -0.00313;

TABLE 5.7: Summary diagnostics resulting from the substantive models when X_3 has an additive functional form. The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated using the 1000 replications.

		No Auxiliary Variables			**One Auxiliary Variable**		
		MCAR	**MAR1**	**MAR2**	**MCAR**	**MAR1**	**MAR2**
BLR	AWO	0.296 (0.0575)	0.284 (0.0575)	0.294 (0.0577)	0.245 (0.0516)	0.237 (0.0510)	0.252 (0.0524)
	APA	0.101 (0.0248)	0.098 (0.0256)	0.103 (0.0261)	0.087 (0.0226)	0.085 (0.0226)	0.089 (0.0235)
	PNP	0.103 (0.0267)	0.097 (0.0248)	0.101 (0.0257)	0.089 (0.0230)	0.084 (0.0224)	0.088 (0.0236)
PMM	AWO	0.328 (0.0622)	0.429 (0.0741)	0.502 (0.1268)	0.237 (0.0532)	0.271 (0.0585)	0.319 (0.0670)
	APA	0.134 (0.0367)	0.157 (0.0413)	0.191 (0.0464)	0.102 (0.0276)	0.115 (0.0300)	0.138 (0.0350)
	PNP	0.134 (0.0355)	0.156 (0.0397)	0.187 (0.0461)	0.103 (0.0282)	0.114 (0.0301)	0.140 (0.0360)
logreg	PNP	0.109 (0.0294)	0.128 (0.0332)	0.160 (0.0404)	0.091 (0.0258)	0.107 (0.0297)	0.134 (0.0363)

CR = 95.2%). Hence, a large proportion of any in the coefficient estimates is attributable to the bias in the imputation model themselves.

Trace plots for each imputation model for BLR and PMM are given in Figure 5.12. The lack of trend for each imputation model implies convergence in the procedure to impute X_3.

The RB, CR, and AW after fitting a substantive model to an index functional form are given in Table 5.8. Hotelling's T^2 tests are performed to test for equality in means in the imputation models. These tests are repeated for each condition, and to test for the equality in means in the estimated coefficient of X_3. The null hypothesis is rejected in all tests performed for an index functional form, yielding sufficient evidence for each condition that the mean estimated coefficients of X_3 are not equal between at least two imputation models. The tests are repeated to test for an equality in the AW values between the imputation models, for each condition. The null hypothesis is rejected in all tests performed for an index functional form, yielding sufficient evidence for each condition that the AWs are not equal between at least two imputation models.

Upon inspection of Table 5.8, under BLR, the active-type imputation models perform very similarly to one another, whereas the passive imputation model is more biased. Despite the results for BLR, the active imputation models under logistic regression are consistently less biased. In Figure 5.13, the frequency of observed values for an index

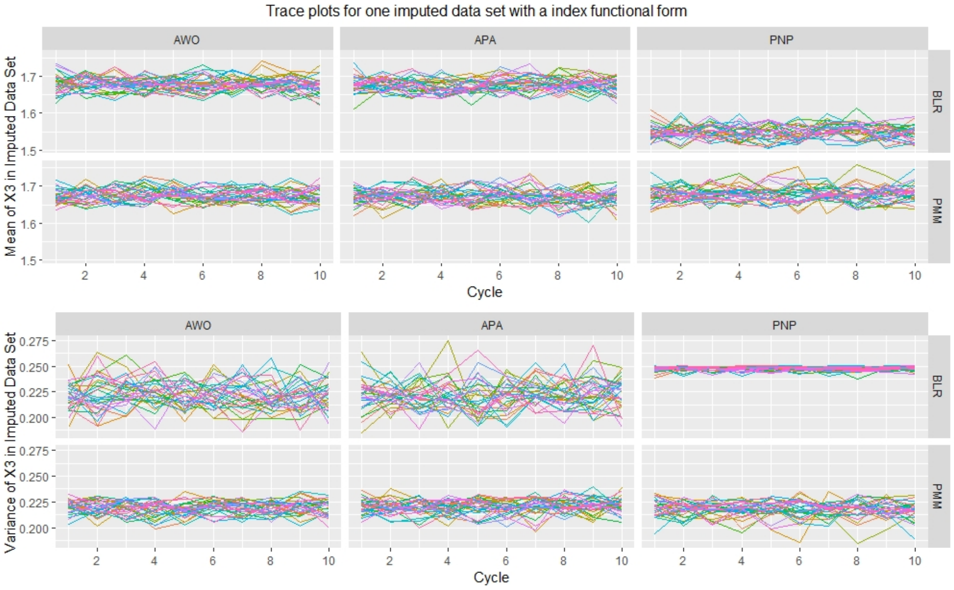


FIGURE 5.12: Trace Plots for each Imputation Model when BLR and PMM is applied under an Index Functional Form. Trace plots are given for one replication, a MCAR structure, and Z is present. Other replications and conditions display a similar trend. The cycle from one to ten is on the x-axis. On the y-axis for the top plots is the mean of X_3 in each of the M imputed data sets; the bottom plots displays the variance of X_3 in each of the M imputed data sets.

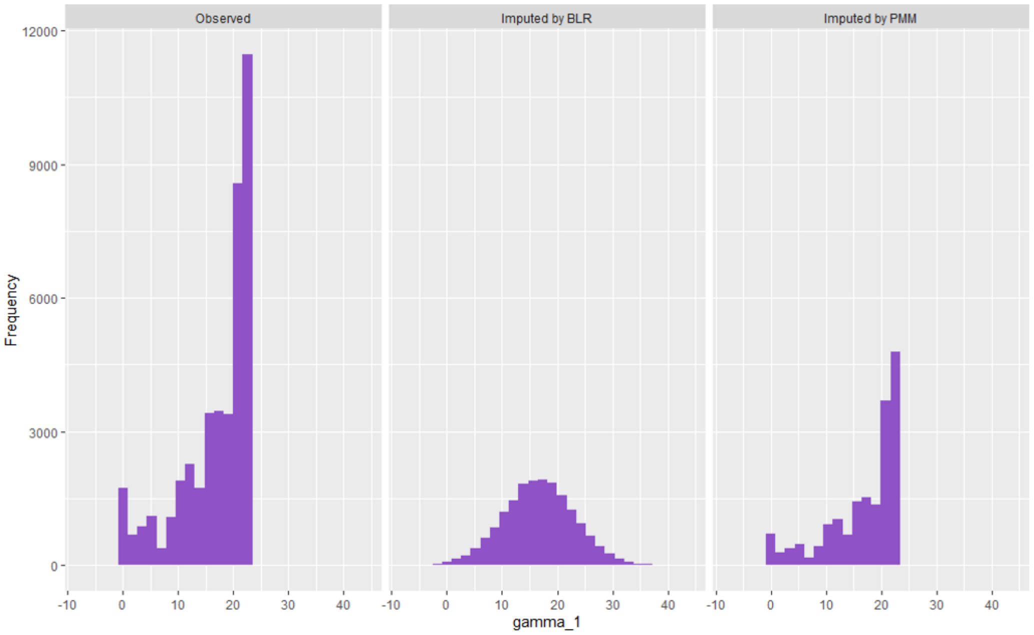


FIGURE 5.13: Frequency of observed and imputed values under passive imputation for an index functional form. This plot is given for one replication.

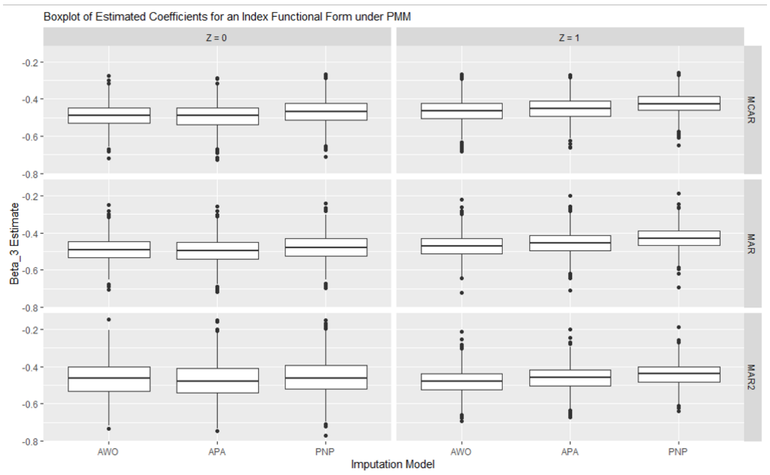


FIGURE 5.14: Boxplot of the estimated coefficients of $\bar{\beta}_3$ under PMM for each imputation model when X_3 follows an index functional form. The rows give the different in missingness structures and the columns give the number of auxiliary variables.

functional form are given, alongside the frequency of imputed values under BLR and PMM. It is evident that the observed values in the data are negatively skewed for the constituent variable. However, γ_1 is imputed to follow a normal distribution under BLR. In this simulation study, observed γ_1 values range from 0 to 23, with $X_3 = 'low'$ if $\gamma_1 \leq 15$, and $X_3 = 'high'$ otherwise. In the data generation process, 32% of γ_1 values are categorised into the ‘low’ group, and ‘68%’ in the ‘high’ group. However, under BLR, approximately half of the imputed values of γ_1 are recategorised as ‘low’ in X_3, and half as ‘high’. As a result, dichotomising γ_1 into a two-level variable may perform worse under BLR than under PMM.

As with other functional forms in the simulation study, there is a poor coverage rate for AWO under PMM, when Z is not present, and X_3 follows a MAR2 missingness structure. However, unlike the other functional forms, APA and PNP result in a CR similar to AWO under these conditions. A boxplot displaying the pooled estimated coefficients for AWO, APA, and PNP imputation models is given in Figure 5.14. The estimated coefficients range over a larger interval for all three imputation models when data is MAR2 and Z is not a predictor in the imputation model compared to any other condition. The three imputation models perform similarly because there is only one constituent present in an index functional form. As a result, there are no observed values of γ_1 to influence the imputation of missing X_3. A similar trend is seen later for a ratio functional form when all constituents are missing (Chapter 7). Additionally, when an auxiliary variable is not present, the relationship between X_1 and γ_1 is weakened under a MAR2 structure (in the generated data $\text{cor}(X_1, \gamma_1)$ = 0.29; under MAR2 $\text{cor}(X_1, \gamma_1) = 0.09$). Hence, there are no predictors strongly correlated with γ_1 or X_3 to influence the imputation and so all models perform similarly.

When no auxiliary variables are present, logistic regression is more suitable than PMM and BLR for data of this functional form: the estimated coefficients are less biased, and the CR is more often in a suitable range. When an auxiliary variable is present, logistic regression still outperforms PMM and BLR, with the exception of APA under PMM when X_3 is MCAR.

When a logistic regression model is fitted in MICE (Table 5.8) AWO outperforms APA: the estimated coefficients are significantly closer to the true value of β_3 and the AWs are significantly smaller for AWO than for APA. Hence, AWO under logistic regression outperforms all other imputation models when X_3 has a MAR-type scheme: it is the least biased, the AW is the smallest, and the CR is in the appropriate range. If an auxiliary variable is present and X_3 is MAR2, then not using the auxiliary variable as a predictor may improve the imputation procedure for an AWO logistic regression imputation model. If X_3 is MCAR and Z is not present in the imputation model, AWO under a logistic regression model outperforms all the other imputation models. If Z is a predictor in the imputation model, then AWO under a logistic regression model and APA under PMM both outperform other imputation models.

5.3.3.1 Missingness Structure

As with other functional forms, the AW increases as the missingness structure becomes less random. This is because the relationship between X_3 and its predictors in the imputation model is less representative of the relationships in the generated data when X_3 is MAR-type than when it has a MCAR missingness structure.

5.3.3.2 The Number of Auxiliary Variables

As with other functional forms, the presence of an auxiliary variable results in a smaller AW and often the estimated coefficients of β_3 are less biased. The presence of an auxiliary variable enhances the imputation process and slightly alters how different imputation models scale up to one another. AWO under logistic regression outperforms all other imputation models when X_3 has a MAR-type scheme. Further, AWO under logistic regression performs well if X_3 is MCAR, but APA PMM has a similar AW and is less biased than AWO logistic regression if Z is a predictor in the imputation model.

5.3.3.3 Model Diagnostics

The mean and standard deviation of FMI values are given in Table 5.9. These summary diagnostics from the substantive models after fitting the imputation models to an index functional form follow a similar structure to that of an additive functional

TABLE 5.8: Metrics for the estimated coefficients of the derived variable in a exponential AFT substantive model for an index functional form

			No Auxiliary Variables			One Auxiliary Variable		
			RB	CR (%)	AW	RB	CR (%)	AW
BLR	MCAR	AWO	0.0296	93.5*	0.251	0.0290	93.6*	0.248
		APA	0.0298	93.5*	0.252	0.0292	93.7	0.248
		PNP	0.0594	86.3*	0.246	0.0590	86.5*	0.240
	MAR1	AWO	0.0351	93.3*	0.258	0.0331	92.9*	0.253
		APA	0.0351	93.5*	0.257	0.0331	93.7	0.253
		PNP	0.0610	86.4*	0.251	0.0589	85.9*	0.246
	MAR2	AWO	0.0358	93.4*	0.268	0.0338	94.0	0.262
		APA	0.0359	93.5*	0.268	0.0339	94.6	0.262
		PNP	0.0604	89.1*	0.262	0.0572	89.3*	0.255
PMM	MCAR	AWO	-0.0368	93.1*	0.261	-0.0134	94.4	0.251
		APA	-0.0406	91.6*	0.263	-0.0005	95.6	0.249
		PNP	-0.0177	95.1	0.264	0.0251	93.5*	0.248
	MAR1	AWO	-0.0399	92.4*	0.273	-0.0208	93.0*	0.256
		APA	-0.0466	89.5*	0.273	-0.0040	93.8	0.253
		PNP	-0.0271	93.0*	0.275	0.0192	94.3	0.252
	MAR2	AWO	-0.0150	90.9*	0.326	-0.0319	92.9*	0.267
		APA	-0.0253	90.5*	0.328	-0.0123	94.4	0.262
		PNP	-0.0059	90.3*	0.326	0.0078	95.9	0.260
logit	MCAR	AWO	-0.0038	94.6	0.246	-0.0040	96.0	0.242
		APA	-0.0198	94.8	0.263	0.0218	93.9	0.247
	MAR1	AWO	0.0001	94.2	0.250	0.0001	93.8	0.245
		APA	-0.0293	93.2*	0.274	0.0154	94.4	0.251
	MAR2	AWO	0.0002	95.0	0.260	0.0016	94.7	0.252
		APA	-0.0075	91.4*	0.324	0.0045	95.4	0.259

* denotes that the CR is not in the 95% confidence interval.

form: FMI is larger if Z is not a predictor in the imputation model than if Z is a predictor, and FMI is larger for PMM than for BLR.

A high FMI value indicates that the multiply imputed X_3 value is not strongly associated with other variables in the imputation model, so it is no surprise that FMI is smaller when an auxiliary variable is present. Additionally, FMI is smaller under a MCAR structure than a MAR structure since there is a weaker relationship between X_1 and X_3 under a MAR-type structure.

The FMI values for BLR and PMM are similar for the imputation models for each condition. This is a contrast to a ratio and additive functional form where the FMI values were smaller when the constituents were predictors in the imputation model compared to when they were not present. The availability of a constituent does not affect the FMI value as much for an index functional form since there is only one constituent, so when X_3 is missing, so is γ_1.

FMI is smallest under AWO with logistic regression, supporting further that the AWO logistic regression imputation model outperforms other imputation models. In contrast to this, APA under logistic regression results in some of the largest values for FMI, highlighting the importance of imputing the constituents well.

TABLE 5.9: Summary diagnostics resulting from the substantive models when X_3 has an index functional form. The mean and standard deviation (SD in brackets) FMI values for $\hat{\beta}_3$ have been calculated using the 1000 replications.

		No Auxiliary Variables			**One Auxiliary Variable**		
		MCAR	**MAR1**	**MAR2**	**MCAR**	**MAR1**	**MAR2**
BLR	AWO	0.278	0.295	0.298	0.259	0.270	0.280
		(0.0547)	(0.0570)	(0.0575)	(0.0535)	(0.0542)	(0.0544)
	APA	0.280	0.294	0.296	0.257	0.270	0.282
		(0.0565)	(0.0568)	(0.0557)	(0.0520)	(0.0540)	(0.0548)
	PNP	0.283	0.292	0.310	0.247	0.259	0.279
		(0.0563)	(0.0575)	(0.0585)	(0.0517)	(0.0533)	(0.0551)
PMM	AWO	0.303	0.352	0.477	0.252	0.271	0.295
		(0.0623)	(0.0748)	(0.1470)	(0.0543)	(0.0562)	(0.0610)
	APA	0.311	0.352	0.487	0.240	0.257	0.275
		(0.0661)	(0.0735)	(0.1421)	(0.0498)	(0.0554)	(0.0581)
	PNP	0.318	0.362	0.480	0.237	0.252	0.266
		(0.0683)	(0.0767)	(0.1476)	(0.0536)	(0.0560)	(0.0565)
logit	AWO	0.222	0.242	0.249	0.196	0.212	0.217
		(0.0464)	(0.0516)	(0.0511)	(0.0415)	(0.0452)	(0.0458)
	APA	0.314	0.359	0.475	0.236	0.251	0.264
		(0.0647)	(0.0725)	(0.0147)	(0.0503)	(0.0536)	(0.0576)

5.4 Conclusion

In this chapter, a simulation study has been undertaken to investigate the performance of active and passive imputation of a derived variable in a survival analysis context. Some general results can be concluded across all three functional forms, which are outlined in this section.

Firstly, the imputation models generally perform well when the method applied is tailored towards the functional form: that is, when logistic regression is used to impute an incomplete binary variable in an additive or index functional form. Results when applying different imputation methods to an incomplete ratio functional form are given in Appendix E. In Chapter 6, two further approaches are considered when imputing a derived variable: SMCFCS and CC. These approaches are then compared to the MICE results to investigate further the relative performance of active and passive imputation under different conditions.

In the simulation study, AWO under logistic regression performs well for an index functional form potentially because there is only one constituent with an index functional form, so when the derived variable is missing, there is no observed predictor. This is investigated further in Chapter 7 where a post-hoc analysis is undertaken to review the sensitivity of the simulation study. In Chapter 7, a simulation study is performed under a ratio functional form where both constituents are missing if the derived variable is missing. Furthermore, additional changes to the simulation study are made are performed to further investigate the sensitivity of the study: the sample size, the proportion of censored data, and the strength of relationship between X_3 and X_1.

Chapter 6

Further Approaches to Handling Missing Data

In this chapter, two additional approaches to handling missing data are considered. Firstly a modification to the MICE procedure, substantive-model compatible fully conditional specification (SMCFCS), is outlined, and its performance is then investigated for the three functional forms. Second, a complete case analysis (CC) is performed to compare the MICE procedure against a simpler deletion approach that does not require imputation.

6.1 SMCFCS

As outlined in Section 2.2.3, a known problem with passive imputation is incompatibility, which occurs when the outcome variable does not directly influence the imputation of the passively imputed variable. As a result, the relationship between the passively imputed variable and the outcome can be undermined, and hence the coefficient estimates are attenuated. Bartlett and Morris (2015) introduce a modification to the standard MICE procedure called substantive-model compatible fully conditional specification (SMCFCS). SMCFCS extends MICE so that the imputation model for each partially observed variable is compatible with the substantive model, thereby reducing the bias in the estimated coefficients. The theory in this section is based on material from Bartlett and Morris (2015).

In the substantive model, denote the outcome variables by Y, the set of q incomplete covariates by X^M, and the set of $p-q$ complete covariates by X^O. Then X_j^M is the j^{th} partially observed variable, $j = 1, ..., q$, and $X_{-j}^M = (X_1, ..., X_{j-1}, X_{j+1}, ..., X_q)$. Let θ_j be the parameter in the imputation model, and β the parameters in the substantive model. Define the imputation model for each X_j^M by $f(X_j^M|X_{-j}^M, X^O, Y, \theta_j)$, and the substantive model is given by $f(Y|X^M, X^O, \beta)$.

For the imputation model of X_j^M to be compatible with the substantive model, Bartlett and Morris (2015) note that

$$
\begin{aligned}
& f(X_j^M|X_{-j}^M, X^O, Y) \\
&= \frac{f(Y, X_j^M|X_{-j}^M, X^O)}{f(Y|X_{-j}^M, X^O)} \\
&= \frac{f(Y|X_j^M, X_{-j}^M, X^O) f(X_j^M|X_{-j}^M, X^O)}{f(Y|X_{-j}^M, X^O)} \\
&\propto f(Y|X^M, X^O) f(X_j^M|X_{-j}^M, X^O) \qquad (6.1)
\end{aligned}
$$

Given β and θ_j, missing values in X_j^M are imputed from the density proportional to

$$f(Y|X^M, X^O, \beta) f(X_j^M|X_{-j}^M, X^O, \theta_j).$$

Note that $f(X_j^M|X_{-j}^M, X^O, \theta_j)$ is the sole component applied in the MICE method outlined in Section 2.2.1. Coupling this second component with the first component allows for compatibility between the imputation model and the substantive model.

In the t^{th} cycle of the MICE process outlined in Section 2.2.1, the β and θ_j parameters are drawn for the j^{th} variable by:

$$\theta^{(t,j)} \sim f(\theta) f(y|x_j^{M^{mis(t-1)}}, x_j^{M^{obs}}, x_{-j}^{M^*}, x^O, \theta)$$

$$\beta^{(t)_j} \sim f(\beta_j) f(x_j^{M^{mis(t-1)}}, x_j^{M^{obs}}|x_{-j}^{M^*}, x^O, \beta_j)$$

where $f(\theta), f(\beta_j)$ are uninformative priors. $\boldsymbol{y}$ is a vector of the outcome variable Y, and x^O is a matrix of the observed values, X^O. Additionally, $\boldsymbol{x}_j^{M^{obs}}$ is a vector of the observed values in the j^{th} partially observed variable. Similarly, $\boldsymbol{x}_j^{M^{mis}}$ is a vector of missing values in the j^{th} partially observed variable. $\boldsymbol{x}_j^{M^{mis(t)}}$ denotes the imputed values of $\boldsymbol{x}_j^{M^{mis}}$ at the t^{th} cycle for the j^{th} variable, and the imputed values at the $(t-1)^{th}$ cycle for the j^{th} variable are given by $\boldsymbol{x}_j^{M^{mis(t-1)}}$. $\boldsymbol{x}_{-j}^{M^*}$ denotes the most recent values imputed for other incomplete covariates.

Draw X_j^M from a density proportional to $f(Y|X^M, X^O, \beta) f(X_j^M|X_{-j}^M, X^O, \theta_j)$. The missing values in the j^{th} variable can be imputed using Monte Carlo Rejection Sampling.

Monte Carlo Rejection Sampling is an algorithm where data is sampled by continually drawing from a proposal density until a certain condition is met. In Rejection

Sampling for SMCFCS, the proposal density is $f(X_j^M|X_{-j}^M, X^O, \theta_j)$. Further details are given in Section 6 of Bartlett and Morris (2015).

SMCFCS is applied under the different functional forms and repeated for the various conditions. Bartlett and Morris (2015) state that more investigation is required to consider SMCFCS with an auxiliary variable present, particularly when the substantive model involves censoring. Additionally, PMM is not yet offered in the SMCFCS package in R, so SMCFCS is not applied with PMM in this simulation study.

6.1.1 Ratio Functional Form

SMCFCS is applied to both PNP and LNP, creating two imputation models denoted by SPNP and SLNP respectively. The RB, CR, and AW are given in Table 6.1.

When no auxiliary variables are present at least one of the SMCFCS imputation models results in a smaller bias in the β_3 estimate than the MICE models considered (Table 6.1 for SMCFCS; Table 5.3 for MICE). As a result, the β_3 estimate is least biased for SPNP under a MCAR structure; both SPNP and SLNP for a MAR structure; and SLNP for MAR2 structure. There is some overcoverage in the confidence intervals with SMCFCS for a MCAR and MAR structure. However, this is acceptable, as there is some overcoverage in the confidence intervals when generating the data. Furthermore, the AW is similar for the SMCFCS and MICE imputation models.

When an auxiliary variable is a predictor, the CR is consistently in a good range for SPNP and SLNP. Under a MCAR structure, the coefficient estimates have a smaller bias under SPNP than the results under any MICE imputation models. For a MAR-type structure, APA BLR performs at least as well as SMCFCS: the AW is small and the estimated coefficients are slightly less biased. The RB in the coefficient estimates is smaller for SMCFCS when an auxiliary variable is not present than that of MICE when an auxiliary variable is present. Hence, SMCFCS outperforms MICE even if an auxiliary variable is available.

6.1.2 Additive Functional Form

SMCFCS is applied with both logistic regression and BLR for an additive functional form. The RB, CR, and AW are given in Table 6.2. For all imputation models, the estimated coefficients consistently result in a small RB and a CR in a reasonable range.

For both logistic regression and BLR, all imputation models under SMCFCS consistently outperform all imputation models under MICE (Table 5.6), regardless of missingness structure: When no auxiliary variables are present, the estimated coefficients are less biased under SMCFCS than under the MICE imputation models.

TABLE 6.1: Raw bias (RB) and coverage rates (CR) for the pooled estimated coefficients for the derived variable when SMCFCS under BLR is applied to a ratio functional form.

		No Auxiliary Variables			One Auxiliary Variable		
Missingness	**IM**	**RB**	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
MCAR	SLNP	0.00027	96.4*	0.0230	0.00049	96.5*	0.0224
	SPNP	0.00003	96.5*	0.0228	<0.00001	96.2	0.0222
MAR	SLNP	0.00029	96.2	0.0223	0.00017	96.0	0.0218
	SPNP	0.00006	96.8*	0.0221	0.00053	96.1	0.0216
MAR2	SLNP	0.00005	95.4	0.0218	0.00040	95.1	0.0214
	SPNP	0.00037	94.5	0.0212	0.00122	94.2	0.0210

* denotes that the CR is not in the 95% confidence interval.

Additionally, the CRs are consistently in a good range, and the AW is narrower than the MICE imputation models.

When an auxiliary variable is present, the β_3 estimates are less biased under BLR SMCFCS than MICE for all missingness structures. The same observation is made when comparing a logistic regression SMCFCS imputation model to a logistic regression MICE imputation model if X_3 follows a MCAR structure.

Overall, if no auxiliary variables are present for a MAR2 structure, SMCFCS under logistic regression outperforms all other models: the coefficient estimates are least biased, the AW is narrow, and the CR is in a suitable range. Other under conditions, the coefficient estimates are least biased under BLR SMCFCS of all SMCFCS and MICE models applied. Furthermore, the AW is narrow, and the CR is in a suitable range.

TABLE 6.2: Raw bias (RB) and coverage rates (CR) for the pooled estimated coefficients for the derived variable when SMCFCS is applied to an additive functional form.

		No Auxiliary Variables			One Auxiliary Variable		
	Missingness	**RB**	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
Logit	MCAR	-0.00042	95.3	0.0682	-0.00019	95.4	0.0674
	MAR	-0.00129	95.2	0.0685	-0.00259	94.8	0.0695
	MAR2	-0.00090	95.2	0.0675	-0.00178	94.7	0.0681
BLR	MCAR	-0.00059	95.4	0.0682	-0.00030	95.5	0.0676
	MAR	-0.00087	95.3	0.0686	-0.00037	95.0	0.0681
	MAR2	-0.00117	94.8	0.0695	-0.00056	94.8	0.0692

* denotes that the CR is not in the 95% confidence interval.

6.1.3 Index Functional Form

SMCFCS is applied with BLR for an index functional form. The resulting RB, CR, and AW are given in Table 6.3 after performing SMCFCS, and in Table 5.8 after performing MICE.

AWO under a proportional odds imputation model outperforms SMCFCS for all conditions: the bias and AW are smaller. Additionally, for all missingness mechanisms and when an auxiliary variable is present, the β_3 estimates are less biased when APA under PMM (and PNP under PMM if X_3 is MAR2) is applied than imputation models under SMCFCS.

TABLE 6.3: Raw bias (RB) and coverage rates (CR) for the pooled estimated coefficients for the derived variable when SMCFCS is applied to an index functional form.

	No Auxiliary Variables			**One Auxiliary Variable**		
	RB	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
MCAR	0.0132	94.3	0.249	0.0126	95.2	0.242
MAR	0.0068	95.5	0.258	0.0054	95.1	0.248
MAR2	-0.0069	95.3	0.273	-0.0084	94.7	0.261

* denotes that the CR is not in the 95% confidence interval.

6.2 Complete Case Analysis

In Complete Case analysis (CC), any observations with missing values are removed and the reduced data set is then analysed. Often the literature finds that MI outperforms CC; for example, Janssen et al. (2010) find that the resulting analyses are less biased under MI than CC. However, White and Carlin (2010) find that the relative performance of CC and MI depends on the missing data mechanism. In this section, different functional forms are analysed using CC to investigate how the performance of CC compares to that of MI approaches.

6.2.1 Ratio Functional Form

The results when a CC analysis is performed are given in Table 6.4 and are compared to the results under MICE (Table 5.3) and under SMCFCS (Table 6.1).

The AW under CC is wider than that of MICE and SMCFCS imputation models, resulting in less certainty about the coefficient estimates from the substantive model. For each missingness structure, there is at least one imputation model with a smaller bias in $\hat{\beta}_3$ than CC. Under MICE this imputation model generally has an auxiliary variable present, but SMCFCS outperforms CC regardless of the presence or absence of an auxiliary variable. Additionally, the coefficient estimates are least biased under SMCFCS when compared to all of the methods considered.

TABLE 6.4: Metrics for the estimated coefficients of a ratio derived variable under a CC.

Missingness Type	RB	CR (%)	AW
MCAR	0.00013	95.6	0.0241
MAR	0.00035	95.9	0.0231
MAR2	0.00023	96.0	0.0224

* denotes that the CR is not in the 95% confidence interval.

6.2.2 Additive Functional Form

Results when a CC analysis is performed are given in Table 6.5, and are compared to the results under MICE (Table 5.6) and under SMCFCS (Table 6.2). The estimated coefficients for β_3 in the imputation models under MICE are more biased than those after CC, but the AWs are much smaller. However, SMCFCS outperforms CC: while the CR remains in the acceptable interval and the bias is similar for both approaches, the AW is much smaller under SMCFCS (Relative increase: 12.5% for logit MCAR; 14.9% for BLR MAR; 23.5% for BLR MAR2)

Therefore SMCFCS is a preferred method of handling the missing values: when X_3 is MCAR, this is by applying SMCFCS with a logistic regression model, otherwise this is by applying SMCFCS with BLR as given in Section 6.1.2

TABLE 6.5: Metrics for the estimated coefficients of an additive derived variable under CC.

Missingness Type	RB	CR (%)	AW
MCAR	-0.00412	95.2	0.259
MAR	-0.00474	95.0	0.266
MAR2	-0.00205	95.7	0.276

* denotes that the CR is not in the 95% confidence interval.

6.2.3 Index Functional Form

Results when a CC analysis is performed are given in Table 6.6, and are compared to results under MICE (Table 5.8) and SMCFCS (Table 6.3). Regardless of the presence of an auxiliary variable, AWO under a proportional odds model for MICE outperforms CC: the estimated coefficients in the substantive model are less biased, and the AW is smaller. If an auxiliary variable is present and X_3 is MCAR, APA under PMM additionally outperform CC since $\hat{\beta}_3$ is less biased and there is a smaller AW.

TABLE 6.6: Metrics for the estimated coefficients of an index derived variable under CC.

Missingness Type	RB	CR (%)	AW
MCAR	0.00013	95.6	0.0241
MAR	0.00035	95.9	0.0231
MAR2	0.00023	96.0	0.0224

* denotes that the CR is not in the 95% confidence interval.

6.3 Revision of Imputation Models

In this section, a guidance for applied researchers is given which aims to summarise which imputation models work best, or are acceptable, for each functional form. An overall summary is given in Figure 6.1, and expanded on next using the conditional linearity criterion.

For a ratio functional form, it was shown in equations 5.2-5.3 that the rearranged substantive model for the constituents are not linear, giving a more complex relationship between the constituents and outcome variable. Hence, when γ_1 is missing and γ_2 is observed, γ_1 should be imputed and X_3 constructed. Similarly, when γ_2 is missing and γ_1 is observed, γ_2 should be imputed and X_3 constructed. Hence, if there are a mix of missing values in the data for the constituents when X_3 is missing and a ratio functional form, SMCFCS should be applied. However, if both constituents are always missing when the derived variable is missing, then active imputation should be applied due to linearity in the active imputation model. Since no constituents are present when X_3 is present, AWO performs well as observed in Chapter 7.

Leading on from this, no constituents are present when X_3 is present for an index functional form since there is only one constituent. Hence, AWO should be applied as the imputation model under these conditions. An appropriate AWO model should be used, such as a logistic regression model or proportional odds model where appropriate.

For an additive functional form, active imputation is recommended regardless of the number of observed constituents for a missing derived variable. Active imputation is recommended because the functional form is linear, so a rearranged the substantive

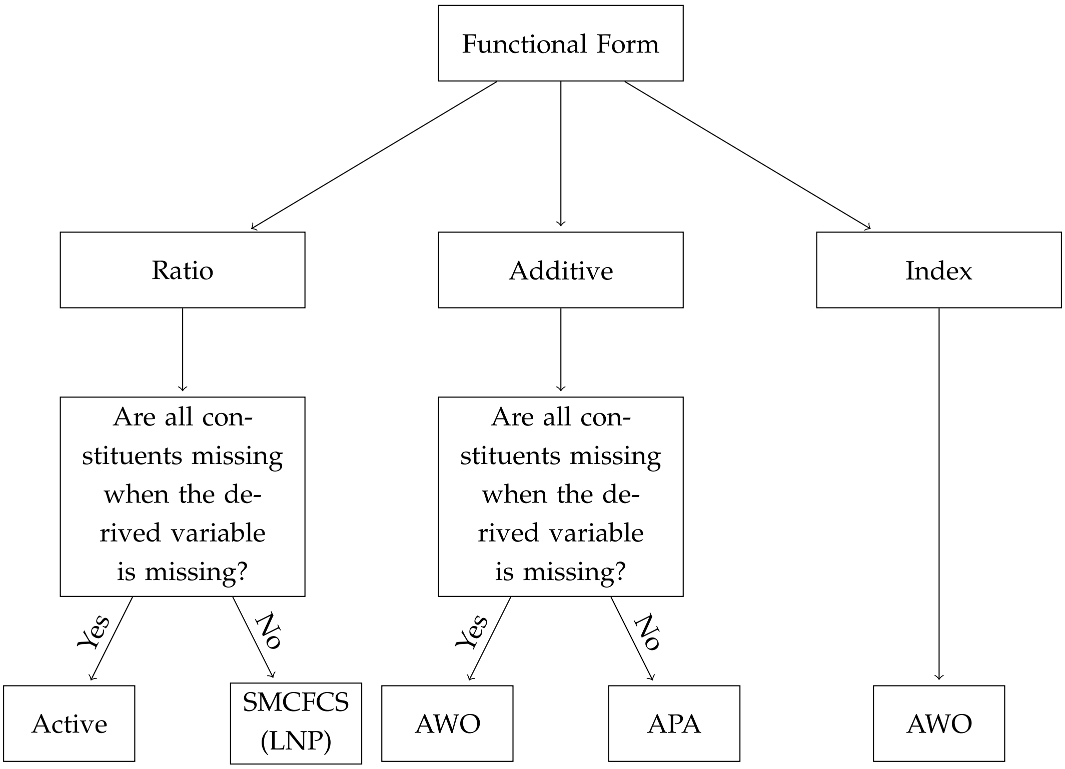

Functional Form
Ratio
Additive
Index
Are all constituents missing when the derived variable is missing?
Are all constituents missing when the derived variable is missing?
Yes
No
Yes
No
Active
SMCFCS (LNP)
AWO
APA
AWO

FIGURE 6.1: A flow chart of which imputation model to use given certain criteria.

model is linear too, as demonstrated in equation 5.5. If at least one constituent is observed for a missing value of the derived variable, APA should be applied; otherwise AWO should be applied.

From the results in the simulation study, PMM should be avoided under a MAR-type structure. Furthermore, logistic regression (or polytomous logistic regression) should be strongly considered for an index functional form.

SMCFCS takes substantially longer to perform than MICE in current R packages. In Chapter 8, an illustrative analysis of one of the motivating data sets using the different imputation models is performed. The MICE procedures varied between 48 seconds (AWO) to one minute 44 seconds (APA) with $M = 100$, whereas SMCFCS took between one hour, 26 minutes, 7 seconds (SMCFCS-PNP) and one hour, 31 minutes, 43 seconds (SMCFCS-LNP). Therefore, where SMCFCS can be avoided, active imputation is recommended.

Chapter 7

Sensitivity Analysis

A post-hoc analysis is undertaken to review the sensitivity of the simulation study when some underlying conditions are altered. Conditions changed are the proportion of missing constituents, the sample size, the proportion of survival times that are censored, and the relationship between X_3 and other predictors.

7.1 Changing Missingness

The index functional form used in the simulation study consists of one derived variable and hence γ_1 is missing when X_3 is missing. As a result, the coefficient estimates and AW are virtually identical for AWO and APA under both BLR and PMM (Table 5.8). However, AWO under a logistic regression model outperforms all other imputation models. That AWO outperforms other imputation models when all constituents are missing suggests that, if all constituents are missing when X_3 is missing, using the constituents as predictors may disadvantage the imputation procedure: the coefficient estimates may be more biased, AW wider, and CR out of the suitable range. As a result, a sensitivity analysis is undertaken to investigate the performance of the imputation models when both constituents are missing for a ratio functional form. The results of this simulation are given in Table 7.1.

There are fewer observed values in the constituents to influence the imputation of X_3 when both constituents are missing than when only one constituent is missing. As a result, the imputation models perform worse when both constituents are missing: the RB in the estimated coefficient of β_3 increases and the AW is wider than the results seen previously in Table 5.3.

The RB and AW are comparable for AWO and APA when both constituents are missing. A paired comparison test gives insufficient evidence to suggest that the estimated coefficient for X_3 after AWO is performed is not equivalent to the estimated coefficient of X_3 after APA is performed. This therefore supports the notion that when both constituents are missing, APA and AWO perform comparably to one another, as was observed for an index functional form in Section 5.8.

The FMI values for the imputation models with both constituents missing are given in Table 7.2. The FIV values for APA, PNP, and LNP when both constituents are missing are similar to that of AWO (with AWO results given in Table 5.5). This was also previously observed with an index functional form in Table 5.9. The FMI value is larger when both constituents are always missing than when at least one constituent is missing because there are fewer observed constituent values to influence the imputation in the imputation models.

TABLE 7.1: Metrics for the estimated coefficients of the derived variable in an exponential AFT substantive model for a ratio functional form with both constituents missing if the derived variable is not observed.

		No Auxiliary Variables			**One Auxiliary Variable**		
		RB	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
MCAR	APA	-0.00074	95.4	0.0239	-0.00048	95.8	0.0232
	PNP	-0.00203	94.1	0.0238	-0.00173	94.8	0.0229
	LNP	-0.00020	95.1	0.0240	0.00007	97.3*	0.0233
MAR	APA	0.00081	95.4	0.0230	0.00014	96.3	0.0224
	PNP	-0.00056	95.7	0.0229	-0.00122	94.8	0.0222
	LNP	0.00186	94.7	0.0233	0.00115	95.5	0.0227
MAR2	APA	0.00211	93.7	0.0224	0.00062	95.8	0.0219
	PNP	0.00054	95.5	0.0223	-0.00078*	95.1	0.0217
	LNP	0.00353	91.2*	0.0227	0.00198	92.9*	0.0222

* denotes that the CR is not in the 95% confidence interval.

TABLE 7.2: FMI values for BLR MICE when X_3 has a ratio functional form and both constituents are missing.

	No Auxiliary Variables			**One Auxiliary Variable**		
	MCAR	**MAR**	**MAR2**	**MCAR**	**MAR**	**MAR2**
APA	0.307	0.296	0.285	0.261	0.254	0.244
PNP	0.331	0.320	0.315	0.278	0.275	0.272
LNP	0.281	0.260	0.242	0.240	0.220	0.201

7.2 Changing the Generated Data Set

Parameters in the data generation can be altered. In this section, the number of rows, the proportion of censoring, and the strength of the relationship between a covariate and the derived variable are altered. This is performed for a ratio functional form to investigate how changing these parameters changes the overall results of active and passive imputation.

7.2.1 Changing the Sample Size and the Censoring Percentages

In Section 5.1, the design and results from a simulation study are given where 15% of observations are censored and the generated data sets have 2000 rows. In this sub-section, the sensitivity of the sample size and censoring percentages in the simulation study results are investigated. To investigate the sensitivity, further simulations are performed with different sample sizes and censoring percentages ($N = 500, 1000, 2000$; proportions of censoring $10\%, 15\%, 20\%$). The results, given in Appendix F (Tables F.1-F.8), support the overall findings of the simulation study in Section 5.1. The results are consistent regardless of the sample size or censoring proportions investigated.

7.2.2 Changing the Strength of Relationship between a Predictor and the Derived Variable

In this sub-section, the sensitivity of the relationship between X_3 and X_1 in the simulation study is investigated. To investigate the sensitivity, further simulations are performed where the relationship of X_1 and X_3 strengthened to investigate if, and how, strengthening the relationship affects the performance of active and passive imputation. X_1 is generated such that $\text{cor}(X_1, X_3) \approx 0.8$.

Recall that in a ratio functional form, $X_3 = \frac{\gamma_1}{(\gamma_2/100)^2}$. Due to the ratio aspect of this functional form, X_3 is not strongly associated with the denominator. As a result, when $\text{cor}(X_1, X_3) = 0.8$, $\text{cor}(X_3, \gamma_1) \approx 0.85$, but $\text{cor}(X_3, \gamma_2) \approx 0.06$. Therefore, changing the correlation between a covariate and the derived variable may improve the performance of active imputation more than passive imputation.

The RB, CR, and AW when $\text{cor}(X_1, X_3) = 0.8$ and Z is not a predictor in the imputation model are given in Table 7.3. Under PMM for a MAR2 structure when Z is not a predictor in the imputation model, the AWO imputation model results in a poor CR. This is consistent with findings when $\text{cor}(X_1, X_3) \approx 0.3$ (Table 5.3). In addition, when $\text{cor}(X_1, X_3) \approx 0.8$, the bias in the coefficient estimate is larger than when $\text{cor}(X_1, X_3) \approx 0.3$ for an AWO imputation model when X_3 has a MAR2 structure, Z is not present, and MICE is performed with PMM. In the MAR2 structure when $\text{cor}(X_1, X_3) \approx 0.8$, smaller X_3 values are more likely to be missing than when $\text{cor}(X_1, X_3) \approx 0.3$. As a result, the distribution of the observed X_3 values is more skewed. This may explain why PMM performs worse for this case despite a stronger relationship with the predictor, X_1.

The bias in the coefficient estimate of β_3 decreases when $\text{cor}(X_1, X_3) = 0.8$ than when $\text{cor}(X_1, X_3) = 0.3$ for the other imputation models. This is because when $\text{cor}(X_1, X_3) = 0.8$, $\text{cor}(X_1, \gamma_1)$ is larger than when $\text{cor}(X_1, X_3) = 0.3$. Hence, X_1 is a better predictor for γ_1 when $\text{cor}(X_1, X_3) = 0.8$, and as a result γ_1 is a better predictor for X_3. Hence, with the exception of AWO under PMM MAR2, all active-type imputation models are less biased and all CRs are in a good range when the relationship between X_3 and X_1 is stronger. However, the AW increases when the correlation is 0.8 rather than 0.3.

AW increases for passive-type imputation models when the relationship between X_1 and X_3 is stronger. Results after applying LNP are more consistent than other imputation models, however, with a similar bias in the coefficient estimate whether the correlation is 0.3 or 0.8.

TABLE 7.3: Metrics from a substantive model when $\text{cor}(X_1, X_3) = 0.8$ for a ratio functional form. This is given for BLR and PMM for the values of the MCAR and MAR2 X_3 values when no auxiliary variable is present.

		BLR			**PMM**		
		RB	**CR (%)**	**AW**	**RB**	**CR (%)**	**AW**
MCAR	AWO	-0.00032	96.2	0.0366	-0.00019	95.8	0.0362
	APA	-0.00029	96.1	0.0355	0.00042	94.9	0.0354
	PNP	-0.00224	95.6	0.0352	-0.00013	95.7	0.0353
	LNP	-0.00035	95.9	0.0354	-0.00028	95.5	0.0353
MAR2	AWO	-0.00012	96.2	0.0350	-0.00912	73.5*	0.0318
	APA	0.00021	95.5	0.0341	0.00188	94.0	0.0338
	PNP	-0.00235	95.2	0.0336	0.00134	94.5	0.0337
	LNP	0.00227	93.7	0.0339	0.00191	94.7	0.0337

* denotes that the CR is not in the 95% confidence interval.

7.3 Overall Sensitivity

In this section, the sensitivity of the simulation study performed in Chapter 5 is investigated by addressing the predictors in the imputation models, the number of missing constituents, and the generated data set.

In Section 7.1, all the constituents were set as missing when X_3 was missing. The imputation models performed worse, as expected. However, the relative performance in the imputation models is the same when both constituents are missing to when at least one constituent is missing. This indicates that the findings in this thesis are not contingent on the number of constituents missing for a ratio functional form.

The sensitivity of the data set generated in the simulation study is investigated by altering the sample size, censoring proportions, and the strength of the relationship between X_1 and X_3. Although there are some changes; for example, the bias in the estimated coefficients for the active imputation models is reduced when the relationship between X_1 and X_3 is stronger, the performance of the imputation models does not change in how they compare with one another. This indicates that the simulation study is not sensitive to departures from the generated values chosen in the data generated in Section 5.

Chapter 8

Illustrative Analysis

In this chapter, an illustrative analysis is performed to one of the motivating data sets. First of all, a brief exploration of the patterns in the data set is given, alongside the cause of missingness. Following this, both MICE and SMCFCS are performed to impute the missing values.

The motivating data set used in this chapter is the kidney transplant data. A ratio functional form is explored because it allows for LNP to be investigated alongside passive imputation. In addition, the kidney transplant data is selected instead of the cardiothoracic data set because it contains a higher proportion of missing values, and therefore is a more interesting data set to investigate. Finally, the initial motivation to this thesis by Pankhurst et al. (2020) analysed the kidney transplant data set. Pankhurst et al. (2020) concluded that MI outperforms CC under an AWO imputation model, but did not investigate alternative imputation models. In this section, these findings are expanded to investigate different types of imputation models. Given that Pankhurst et al. (2020) found that MI outperforms CC for this particular data set, CC is not investigated in this section.

8.1 Missingness Structure in the Data

A few minor alterations are made to the kidney transplant data set before the analysis. Firstly, only recipients and donors aged 20 or older are considered since BMI is not a reliable measurement on individuals below 20 years old (CDC, 2015). In addition, there is an observed weight of 17kg for a recipient. This observed weight is set as missing since it is assumed to be a misprint.

A plot displaying the proportion of missing values in the kidney transplant data set for each variable is given in Figure 8.1. Recipient weight and height generally are missing for individuals who had a transplant in the earlier years. This is highlighted

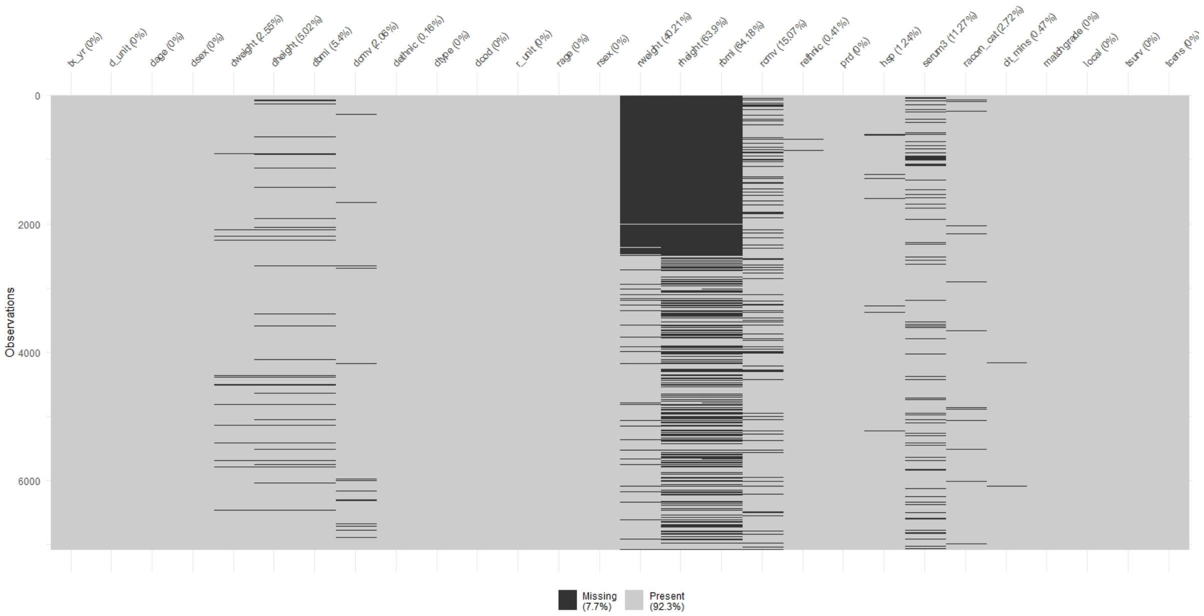


FIGURE 8.1: Plot of missing values for each variable in the Kidney Transplant data set.

in Figure 8.2 where the proportion of missing values in recipient weight and height is given, split by year of transplant and the recipients unit. In 2001 and 2002, the recipients height and weight were not recorded. In subsequent years, these measurements begin to be taken, but, for some recipient units a high proportion of the height variable is still not taken. As a result, recipient height and weight (and therefore recipient BMI) are MAR because there are missing values for certain units and years. However, the distribution of the recipients weight and height values are the same regardless of the recipient unit and year. As a result, recipient unit and year are not necessarily required as predictors in the imputation model for recipient height and weight. The simulation study found that BLR should be used for a MAR structure, so BLR is applied in the analysis in this section. This is additionally consistent with the study performed in Pankhurst et al. (2020).

8.2 Multiple Imputation

Active and passive imputation models are fitted to impute the kidney transplant data. Passive imputation is expected to perform well, particularly for recipient BMI, because there is a mixture of observed and missing values for each missing BMI value (as addressed in Section 4.2.1). The methods to multiply impute the kidney transplant data set are first given in this section, followed by the results.

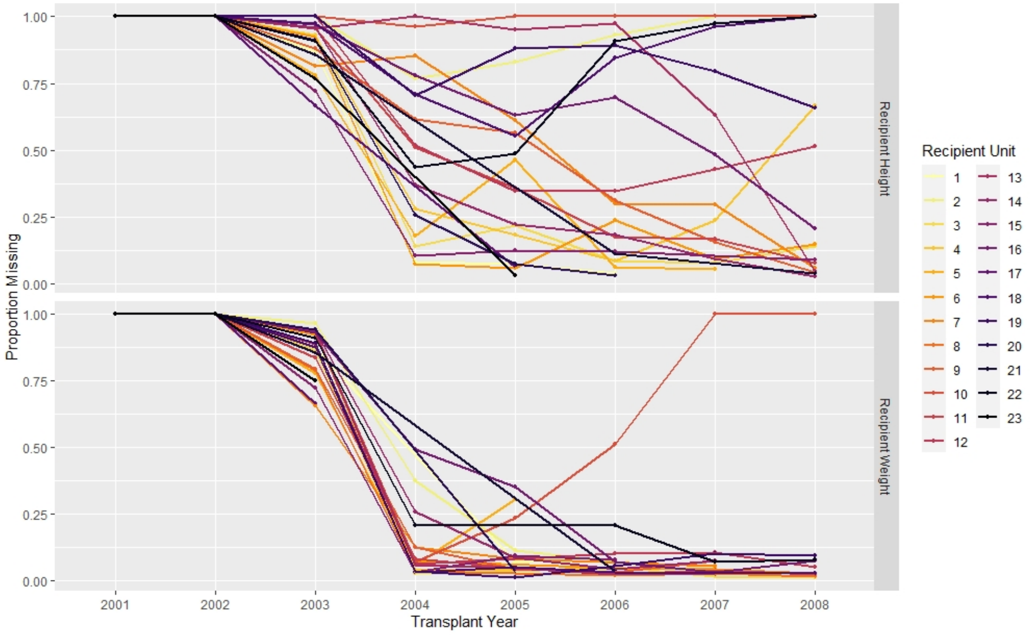


FIGURE 8.2: Proportion of missing values for recipient height and weight in the kidney transplant data set, split by recipient unit and year of transplant.

8.2.1 Methods

In the study performed by Morris et al. (2014), $M = 100$ and $M = 300$ depending on the length of the data set. Since the Kidney Transplant data set contains over 7000 rows, $M = 100$. Additionally, the minimum value of M should be equivalent to the largest percentage of missing values in a single variable (White and Royston, 2009), and 67% of recipient BMI is missing, so M needs to be greater than 67.

The performance of the four imputation models applied in Section 5.1.1 is evaluated after fitting the four imputation models to the kidney transplant data set. These models are:

- **AWO:** Active imputation without constituents present as predictors for BMI,
- **APA:** Active imputation with constituents present,
- **PNP:** Standard passive imputation,
- **LNP:** Passive imputation where the constituents are first log-transformed.

In addition, SMCFCS is investigated for the passive-type imputation models.

For consistency with Pankhurst et al. (2020), a Cox Proportional-Hazards model is fitted to the imputed data as the substantive model. Following this, Rubin's rules are applied to combine estimates and explore the performance of the imputation models.

The Cox Proportional-Hazards model fitted to the multiply imputed data sets include donor and recipient BMI to assess the imputation procedure for the different imputation models. Any other significant variables after applying backwards selection are also included in the substantive model. This results in the following Cox Proportional-Hazards model:

$$h(t) = h_0(t)\exp(\beta_1 dage + \beta_2 dbmi + \beta_3 dcmv + \beta_4 rage + \beta_5 rbmi + \beta_6 prd + \beta_7 serum).$$

All variables in the substantive model are in the imputation model when imputing recipient and donor weight, height, and BMI, with the except that donor BMI is not a predictor of recipient donor weight or height to avoid the issue of circularity. For the same reason, recipient BMI is not a predictor of either recipient weight or height.

By calculating the correlation between each variable in the Kidney transplant data set and recipient and donor weight, height, and BMI, potential auxiliary variables are determined. Firstly, recipient sex is found to be a good predictor of recipient weight (correlation of -0.381). In addition, donor sex is found to be a good predictor of donor height (correlation of -0.637). Hence, recipient sex is an auxiliary variable for recipient weight (and recipient BMI in the active imputation models), and donor sex is an auxiliary variable for donor height (and donor BMI in the active imputation models).

8.2.2 Results

MICE was substantially quicker than SMCFCS. The MICE procedure took between 48 seconds (AWO) and 1 minute 44 seconds (APA), whereas SMCFCS took between 1 hour 26 minutes 7 secs (SMCFCS-PNP) and 1 hour 31 minutes 43 seconds (SMCFCS-LNP).

The estimated coefficients for donor BMI and recipient BMI from the different imputation models are given in Figures 8.3-8.4 respectively. For donor BMI, there is little difference in the estimated coefficient and the width of the confidence intervals for all imputation models. The estimated coefficient of recipient BMI varies slightly across between imputation models, with the MICE methods slightly attenuated to zero. In addition, the confidence intervals are slightly narrower for the SMCFCS methods than MICE. Overall, there is a suggestion that SMCFCS outperforms MICE for this data.

Since the same data set is imputed for all imputation models, a high FMI value indicates that the multiply imputed value is not strongly associated with other variables in the imputation model. For both donor and recipient BMI, FMI is larger for

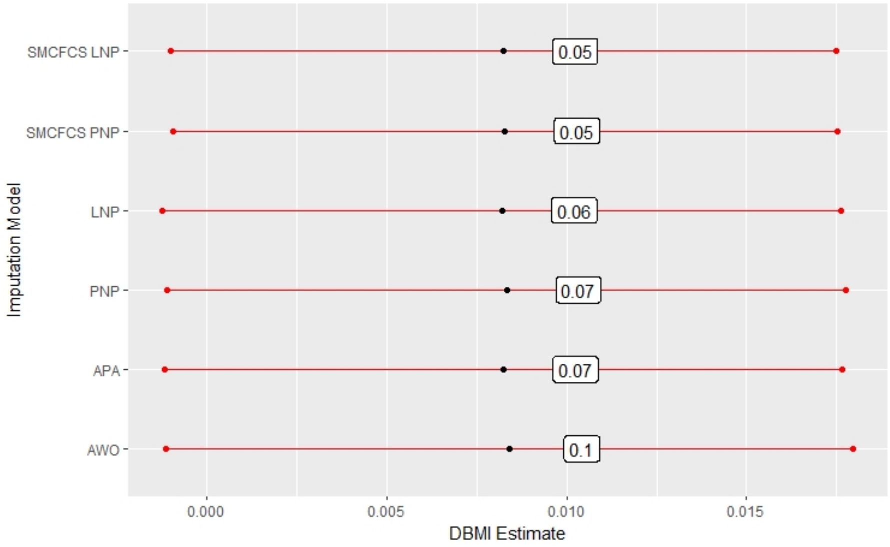


FIGURE 8.3: Estimated coefficient of donor BMI with a 95% confidence interval for the six imputation models. The FMI value is additionally given.

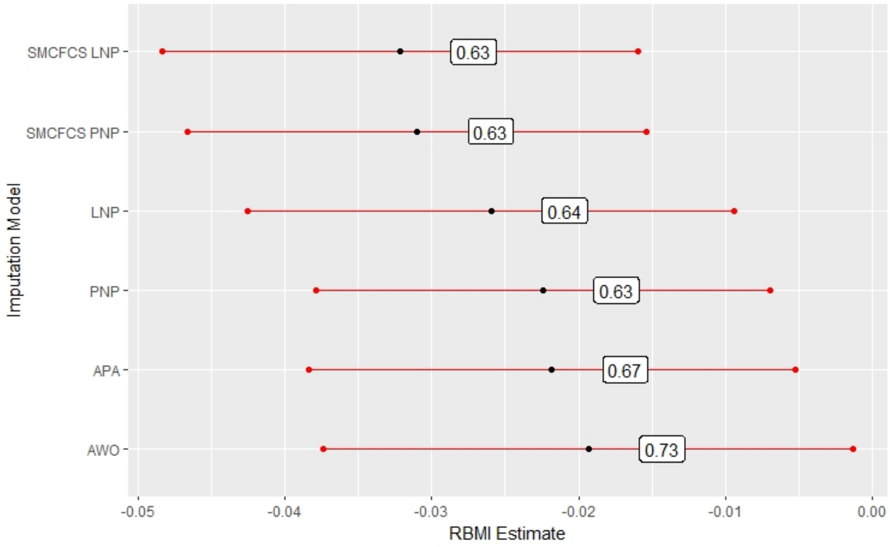


FIGURE 8.4: Estimated coefficient of recipient BMI with a 95% confidence interval for the six imputation models. The FMI value is additionally given.

AWO since there are fewer predictors in the AWO imputation model. FMI is larger under MICE than under SMCFCS, indicating that the association with other variables in the imputation model is greater under SMCFCS than under MICE.

The imputed values across all multiply imputed data set for recipient BMI are given in Figure 8.5, with the minimum and maximum values labelled. LNP under both MICE and SMCFCS, results in imputed BMI values in a more reasonable range than the other imputation models. While the imputed values are small, they are realistic under LNP. In contrast to this, there are negative values under PNP for both SMCFCS and MICE. As a result, SMCFCS under LNP is a better fit when imputing data of this form.

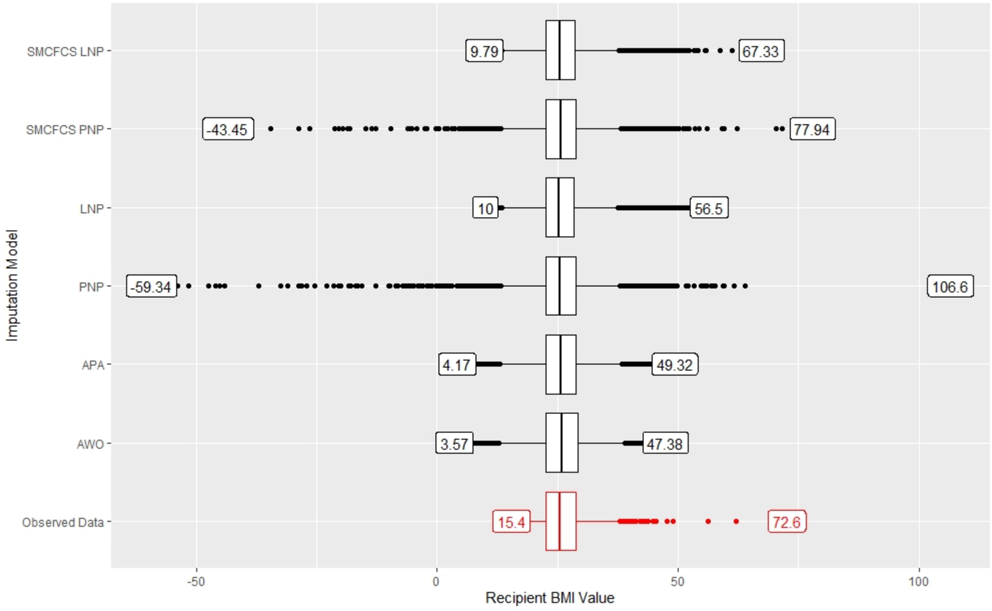


FIGURE 8.5: Boxplot of the imputed values from all multiply imputed data set for recipient BMI, and a boxplot of the observed values in recipient BMI. The labels give the minimum and maximum imputed value.

In addition, there are some impossibly small imputed recipient BMI values under active-type imputation. While a minimum boundary can be fit under the MICE procedure for MICE to help avoid such extreme values, it would result in a lot of individuals at the minimum boundary given.

8.3 Conclusion

SMCFCS was substantially slower to perform than MICE. When a small proportion of values are imputed, MICE and SMCFCS perform similarly to one another, so it is unclear if there is a benefit to applying SMCFCS when a small proportion of values are missing. However, when a larger proportion of values are missing, SMCFCS outperforms MICE. Under MICE, the confidence intervals were wider, the estimated coefficients were attenuated, and the FMI values were larger than under SMCFCS. When exploring the imputed values themselves, it was found that LNP results in realistic values. As a result, SMCFCS-LNP outperforms SMCFCS-PNP.

Chapter 9

Conclusion

In this thesis, a simulation study was carried out to investigate the performance of active and passive imputation when multiply imputing a derived variable in a survival analysis context. While the performance of active and passive MI has been investigated in previous studies, the literature is limited when investigating in a survival analysis context. A preliminary analysis and simulation study were designed to investigate the performance of active and passive imputation for three functional forms: a ratio, additive, and index functional form. Following the simulation study, an illustrative analysis was performed, resulting in consistent findings with the simulation study. In this chapter, the overall conclusions from multiply imputing a derived variable in a survival analysis context are first given. Following the overall conclusions, several notable remarks on the outcome of the simulation study are outlined. Following these remarks is a critical review of the simulation study, with a guide for further work that can be undertaken.

Under a ratio functional form there is a non-linear relationship between the outcome variable and each constituent when the substantive model is rearranged (Section 6.3). Hence, if both constituents are missing, an active imputation model should be applied. However, if one constituent is missing, and one constituent is observed, then the derived variable should be imputed by SMCFCS. Active imputation is recommended under an additive functional form because the functional form is linear, so a rearranged substantive model is linear too (equation 5.5). Finally, under an index functional form, no constituents are present when the derived variable is present. Hence, AWO should be applied as the imputation model under these conditions. An appropriate AWO model should be used; for example, a logistic regression model performed well in the simulation study for the binary derived variable.

Next, several notable remarks given the simulation study are outlined. First, in practice, it is uncommon for a data analyst to know the true missingness mechanism in an incomplete data set. Regardless of functional form, it is evident from the simulation study that the relative performance in the imputation models changes when the missingness mechanism changes. In practice, when an analyst faces an incomplete data set, they should aim to fit an imputation model which performs well across different missingness schemes. In the simulation study performed in Chapter 6, SMCFCS outperforms passive imputation regardless of the missingness mechanism or functional form. Therefore, SMCFCS should be seriously considered instead of passive imputation by MICE when imputing a derived variable. Although SMCFCS performs well under an index functional form, AWO under a logistic regression model additionally results in a small RB and AW, and a CR in the desired range for all three missingness mechanisms. As outlined in Section 6.3, active imputation outperforms passive imputation under an index functional form because there are no observed predictors when the derived variable is observed. Hence, AWO under a logistic regression model should additionally be considered when multiply imputing a variable which dichotomises a continuous variable.

Secondly, continuing from the first point, it is evident from the good performance of SMCFCS that passive imputation is consistently an effective method to impute the generated data provided incompatibility is accounted for by applying SMCFCS. In addition to this, SMCFCS helps create a simpler procedure. For instance, White and Royston (2009) find that one way to decrease the bias in MICE is by using a Nelson-Aalen estimate of the cumulative baseline hazard instead of the raw survival time in an imputation model. However, using the Nelson-Aalen estimate is only an approximation, and hence an advantage of SMCFCS is that these choices can be avoided, and the process made simpler. Furthermore, SMCFCS is not reliant on the presence of an auxiliary variable to enhance the imputation. Finally, both logistic regression and BLR under SMCFCS for an additive functional form perform well, highlighting the effectiveness of SMCFCS. However, more research is required when applying SMCFCS. It is unclear how best to incorporate an auxiliary variable (Bartlett and Morris, 2015). Furthermore, Bartlett and Morris (2015) noted that SMCFCS requires more investigation when there are censored data.

A third remark from the simulation study is that under MICE, the presence of an auxiliary variable can improve the imputation; for example, the CR is in a reasonable range under all imputation models when an auxiliary variable is present for an additive functional form, and the AW is smaller when an auxiliary variable is present than when it is not. Under some conditions, the presence of an auxiliary variable can slightly impair the performance of the imputation model; for example, there is overcoverage in a BLR-APA imputation model under a MCAR scheme for a ratio functional form if an auxiliary variable is present, but otherwise the CR is in a good range. However, in general, an auxiliary variable tends to improve the imputation procedure much more often than to impair it. Hence, it is worth considering the use of an auxiliary variable if it is present in the data set.

Fourth, the initial aim at the beginning of this thesis was to investigate the performance of active and passive imputation in a survival analysis context. From the literature review, preliminary analysis, and simulation study, it is apparent that this question is too simplistic. The relative performance of active and passive imputation alters when different conditions change in the study; for example, PNP has performed well under some conditions (MICE-BLR under a ratio functional form), and poorly under others (MICE-PMM under a ratio functional form). Likewise, active imputation has performed well under some conditions, and poorly under others (for example, AWO under logistic regression outperforms AWO under BLR when a derived variable has an index functional form with a MAR-type scheme). In future investigations and in practice, additional factors should be considered when investigating or applying MI to a derived variable.

There are several limitations in the simulation study. When the conditions change, such as the presence of an auxiliary variable, the proportion of data that is censored, or whether BLR, PMM, or another approach is applied, there are different and interesting results. One limitation of the study is that these results have only been investigated for an exponential AFT substantive model. However, it would be of interest to investigate the performance of active and passive imputation under different AFT models, or under a Cox Proportional-Hazards model. Further to this, one key limitation in the simulation study is that only Type I right-censoring is investigated. When altering the proportion of censored data in Chapter 7, there is some change in the performance of active imputation relative to that of passive imputation. For example, when $n = 500$, LNP outperforms PNP when the proportion of censored data increases under a MAR1 scheme, but not when the proportion of censored data increases under a MCAR scheme. In addition, more complex censoring mechanisms are often present in real-world data sets, including in the motivating data sets. Therefore, future investigations could investigate whether the results are generalised under different, more complex, censoring mechanisms, and how the performance of active and passive imputation is altered under different substantive models. This notion holds for SMCFCS imputation models too, supported by Bartlett and Morris (2015) who comment that more investigation is needed into applying SMCFCS with censored data.

When generating the data, the coefficient of the derived variable changes for the different functional forms to ensure that the generated data is based on the underlying data sets. Hence, the relative performance of active and passive imputation may change depending on the three functional forms because the relationship between survival time and the derived variable is altered. Furthermore, in Chapter 7, it is concluded that the relative performance of active and passive imputation changes when the relationship between X_3 and X_1 is altered. Hence, one limitation to explore further in future studies is to investigate the performance of active and passive MI when the relationship between the covariates or outcome variable and the derived variable is altered. In addition, while generating the data sets on some real world data allows for a more realistic data set, it means that the results are not as generalisable. For example, BMI is not the only ratio functional form present in a data set that can be analysed by survival analysis, and similarly so for the generated derived variables under an additive and index functional form. Future studies could investigate different types of ratio, additive, or index functional forms. For example, logistic regression under AWO outperforms APA for an index functional form potentially because all constituents are missing when the derived variable is missing. Further non-trivial transformations of a single variable could be investigated in future studies; for example, squaring a variable. Furthermore, in the simulation study only dichotomising a continuous variable is investigated for an index functional form. However, in practice, a continuous variable may be split into a factor containing several levels. In addition, there are further functional forms that have yet to be investigated when applying MI in a survival analysis context. Investigating active and passive MI for different functional forms and well as different examples of a ratio, additive, or index functional form can build a greater understanding of the procedure to impute an incomplete derived variable. In addition, future investigations can provide further clarity of the appropriate imputation models to fit for different derived variables in practice.

The missingness mechanism applied can additionally be investigated further. Firstly, when a derived variable is constructed from more than one constituent, there is a very specific missingness mechanism for the constituents in the simulation study design. This design results in a negative correlation for missingness between the constituents. In practice, a positive correlation might sometimes be expected; that is, if one constituent is missing, it is likely another is too. This may be, for example, due to limited resources when undertaking measurements, or because individuals are unable to answer questions in a survey. Given the results for an index functional form and the initial sensitivity analysis for a ratio functional form in Chapter 7, AWO performs as well as, or better than, APA when all constituents are missing if X_3 is too. Therefore, the investigation of MI when all constituents are missing should be further explored. Finally, only one type of MCAR and two types of MAR structures are investigated. The relative performance of active and passive imputation changes as the missingness changes. It would be of interest to investigate using MI when the derived variable follows a MNAR structure.

Appendix A

Variables in the Kidney Transplant Data Set

There are 32 variables in the kidney transplant data set. These are given in Table A.1.

TABLE A.1: Variables in the Kidney Data Set.

Variable	Description
recip_id	Variable identifying the recipient. Each recipient has one graft.
donor_id	Variable identifying the donor.
tx_id	Variable identifying the transplant.
dbmi	BMI of the donor.
dheight	Height of the donor in centimetres (cm).
dweight	Weight of the donor in kilograms (kg).
rbmi	BMI of the recipient.
rheight	Height of the recipient in centimetres (cm).
rweight	Weight of the recipient in kilograms (kg).
tx_yr	Year of the transplant, ranging from 2001-2008 inclusive.
d_unit	Unit for the donor. There are 23 in total.
dage	Age of the donor at transplant in years, ranging from 20 to 82.
dsex	Sex of the donor.
dcmv	Whether the donor has the cytomegalovirus (CMV).
dethnic	Ethnicity of the donor.
dtype	Whether the donor died from brain injury (DBD), or from circulatory death (DCD).
dcod	Cause of death of the donor.
r_unit	Unit for the donor. There are 23.
rage	Age of the recipient at the start of the study, ranging from 20 to 85 years.
rsex	Sex of the recipient.
rcmv	Whether the recipient has CMV.
rethnic	Ethnicity of the recipient.
prd	Primary renal disease of the recipient.
hsp	Whether the patient is highly sensitised.
wait_time	How long the patient was waiting for a transplant, in days.
cit_mins	The cold ischemic time (CIT) in minutes, ranging between 19 and 2722 minutes.
matchgrade	Factor determining how good the organ match was between donor and recipient.
local	Indicator for whether the kidney was used by a local centre.
tsurv	Survival time of the transplant in days.
tcens	Whether the survival time is censored (0) or not (1).

Appendix B

Further Approaches to Evaluate the Imputation Model

The imputation model is integral when performing multiple imputation. It is therefore important for this model to be checked. In this section, approaches to analyse an imputation model when the true underlying values are not known is discussed.

One way to evaluate the imputation model is through standard descriptive statistics. The performance of the imputation model can be investigated through standard statistical summaries for the imputed variable, x_j. These summaries can be explored in two ways. First, summaries of just the imputed values can help to indicate problems with the imputation model by identifying any extreme or implausible values. Note that imputed values will not be out the range when PMM is applied. Second, the distributions of both the imputed and observed values can be observed by considering standard summary statistics. These distributions can be represented graphically using plots such as histograms, density plots, strip plots, and QQ plots, where the plot is split by whether the value is observed or imputed (Nguyen et al., 2017). The boxplot of the observed data can be visualised alongside M boxplots for each of the M imputed data sets for the imputed variable x_j. However, the other plots are generally displayed for each of the multiply imputed data sets. Stuart et al. (2009) therefore repeats the plot for two of the ten imputed data sets to check for consistency, rather than display plots for each of the multiply imputed data sets. Alternatively, the standard statistical summaries found can be tabulated to compare descriptive statistics such as means, ranges, and standard deviations.

Further to this, Stuart et al. (2009) suggest two thresholds to identify potential issues in the imputation models for imputed variable x_j. First, an issue in the imputation model may be indicated if the absolute difference in the mean for the observed values and imputed values is greater than two standard deviations. Additionally, a problem may be signalled if the ratio of the variance of the observed values for x_j and that of the imputed values for x_j is less than 0.5 or greater than 2.

However, alternative approaches should also be considered when investigating the performance of the imputation model. First, Rodwell et al. (2014) demonstrate through a simulation study that unbiased imputations for a continuous variable are more likely when the imputation does not impose restrictions on the range of imputed data. Hence, the range of the imputed variables does not have to be a priority. Further, under a MAR structure, an inconsistency in the distribution between observed and imputed values may be of no great importance. For example, consider the scenario with the variables weight and sex, where weight is less likely to be disclosed for female participants than for male participants. Hence, across the incomplete values for weight there are a higher proportion of missing values for women than for men. Women, on average, weigh less than men (ONS (2010) found in Britain women on average weigh 70.2kg, while men on average weigh 83.6kg). Therefore, after imputation, it should be expected for the mean of imputed weight values to be smaller than that of the observed weight values.

Standard model checking procedures can also help to investigate the performance of the imputation model. These procedures can be applied to both the imputation and substantive models. Model checking procedures with respect to the imputation model are first addressed. Subset the data to just the observed cases, and then fit the prospective imputation model to this reduced data set. Treat this as a standard analysis model. Perform standard diagnostics to explore the fit of this model to the data; for example, plot residuals against fitted values. The underlying model assumptions can be explored, and hence the imputation can be adjusted if appropriate (Marchenko and Eddings, 2011).

Similar checks can be performed after fitting the substantive model. For example, a plot of the residuals against fitted values can be run for each of the M analysed data sets to examine the fit of both the imputation model and the analysis model (White et al., 2011). If some of the imputed data sets display problems such as extreme values, the imputation model may not be a good fit. Alternatively, if these problems are consistent across most, or all, of the imputed data sets, the substantive model may not be a good fit.

The estimated coefficients and standard errors from the substantive model can be valuable when exploring how effective an imputation model is. An estimated coefficient close to zero could indicate bias, reflecting issues in the imputation model, as discussed in Section 2.2.3. A bias may also be indicated from larger estimated intercepts, since a flatter slope would cause the intercept to increase (von Hippel, 2009). Additionally, standard errors for estimated coefficients in the analysis model can be a good indication of the performance of the imputation model (White et al.,

2011) where smaller standard errors imply the imputation model is performing well. These methods are more appropriate when comparing imputation models, rather than evaluating the performance of a single imputation model. Finally, Monte Carlo error can be considered when evaluating an estimated coefficient in the substantive model for a previously imputed variable (White et al., 2011). Monte Carlo error is given by $\sqrt{V_B/M}$.

Appendix C

Variables in the Cardiothoracic Transplant Data Set

The variables in the cardiothoracic transplant data set are given in Table C.1.

TABLE C.1: Variables in the Cardiothoracic Data Set.

Variable	**Description**
recip_id	Variable identifying the recipient.
donor_id	Variable identifying the donor.
tx_id	Variable identifying the transplant.
dbmi	BMI of the donor.
dheight	Height of the donor in centimetres (cm).
dweight	Weight of the donor in kilograms (kg).
rbmi	BMI of the recipient.
rheight	Height of the recipient in centimetres (cm).
rweight	Weight of the recipient in kilograms (kg).
tx_yr	Year of transplant, split into five-year groups from 1995-1999 to 2015-2019, inclusive.
dage	Age of the donor at transplant in years, ranging from 20 to 74.
dsex	Sex of the donor.
dcmv	Whether the donor has the cytomegalovirus (CMV).
dethnic	Ethnicity of the donor.
dcod	Cause of death for the donor. Most are living donors (61.8%)
dbg	Blood group for the donor.
dpast_diabetes	Whether the donor has previously had diabetes.
dpast_drug_abuse	Whether the donor has previously abused drugs.
rage	Age of the recipient at the start of the study, ranging from 20 to 75 years.
rsex	Sex of the recipient.
rethnic	Ethnicity of the recipient.
rcod	Cause of death for the recipient. Most have not died.
reg_diabetes	Does the recipient had diabetes at registration?
reg_smoker	Has the recipient smoked 5+ a day in the six months prior to registration?
tx	Whether heart (H) or lung (L) is transplanted.
pcd	The primary cause of disease for the recipient.
tsurv	Survival time of the transplant in days.
tcens	Whether the survival time is censored (0) or not (1).

Appendix D

Some variables in the CLHLS Data

In this section the list variables and corresponding summary measures are given for the CLHLS data set. Table D.1 shows the variables in the CLHLS data set. Note that when the functional form is additive, ADL_Index and MMSE_Index are omitted. With an the index functional form, ADL and MMSE are omitted depending on the imputation model. Table D.2 shows the proportion of participants within each level of each factor variable.

TABLE D.1: Variables in the CLHLS Data Set. Note that the covariate values are taken at entry to the study

Variable	Description
ID	Variable identifying the participant.
TRUEAGE	Age (years) of participant
A1	Sex of participant.
ADL	Number of limitations in daily living. This ranges from 0 to 6.
ADL_Index	A factor variable grouping the ADL variable.
D81	Activity level of the participant.
MMSE	Cognitive ability of the participant as a score.
MMSE_Index	Cognitive ability of the participant as an index category.
tsurv	Survival time (years) of the participant.
tcens	Whether the survival time is censored (1) or not (0).

The constituents of the composite variables are omitted from some imputation models. Six binary variables make up ADL are given in Table D.3, where “0” denotes the individual can perform the task without assistance; “1” otherwise. This is given in Table D.3. The 23 binary variables that build MMSE are given in Table D.4, where “0” denotes the participant did not respond correctly; and a “1” otherwise.

TABLE D.2: Values for Factor Variables in the CLHLS Data Set.

Variable	Level	Proportion of Participants
Sex	male (1)	0.401
	female (2)	0.599
Residential status	urban (1)	0.152
	rural (2)	0.848
ADL_Index	no limitations (0)	0.633
	one limitation (1)	0.135
	two or more limitations (2)	0.232
Activity Level	active (1)	0.275
	sedentary (2)	0.725

TABLE D.3: Constituents for the ADL Variable

Variable	Description	Proportion not requiring assistance
E1	Can the participant bathe themselves?	0.684
E2	Can the participant dress themselves?	0.825
E3	Do they require help in using the toilet?	0.807
E4	Do they require help transferring indoors?	0.836
E5	Are they incontinent?	0.912
E6	Do they need help feeding?	0.882

TABLE D.4: The 23 variables that construct the MMSE score

Category	Variable	Description
Orientation	C11	What time of day is it right now?
	C12	What month is it right now?
	C13	What is the date of the mid-autumn festival?
	C14	What season is it right now?
	C15	What is the name of this county/district?
Registration	C21A	Were they able to repeat “table” correctly the first time?
	C21B	Were they able to repeat “apple” correctly the first time?
	C21C	Were they able to repeat “clothes” correctly the first time?
Calculation	C31A	What is “\$20 - \$3”?
	C31B	What is “\$20 - \$3 - \$3”?
	C31C	What is “\$20 - \$3 - \$3 - \$3”?
	C31D	What is “\$20 - \$3 - \$3 - \$3 - \$3”?
	C31E	What is “\$20 - \$3 - \$3 - \$3 - \$3 - \$3”?
	C32	Able to successfully draw the figure?
Recall	C41A	Able to repeat the word “table” a while later?
	C41B	Able to repeat the word “apple” a while later?
	C41C	Able to repeat the word “clothes” a while later?
Language	C51A	Able to name a “pen”?
	C51B	Able to name a “watch”?
	C52	Able to repeat a sentence?
	C53A	Able to take paper using right hand?
	C53B	Able to fold the paper?
	C53C	Able to put the paper on the floor?

Appendix E

Other Imputation Models for a Ratio Functional Form

A handful of approaches are investigated to impute a ratio functional form when X_3 is MCAR and no auxiliary variables are present: stochastic regression imputation, bootstrap multiple imputation, simple random sampling (SRS), and classification and regression trees (CART). The RB, CR, and AW after applying these methods are given in Table E.1.

Stochastic regression imputation and BLR.boot are variations on the method BLR. As outlined in Section 2.2.1, under BLR an incomplete variable x_{ij} is imputed from a normal distribution: $x_{ij} \sim N(\boldsymbol{X}_M\dot{\boldsymbol{\theta}}, \dot{\sigma}^2)$ where, $\boldsymbol{X}_M$ is the portion of the data set with x_{ij} missing, and $\dot{\boldsymbol{\theta}}, \dot{\sigma}^2$ are estimated by drawing random values from the posterior distribution, given the data. In bootstrap multiple imputation (BLR.boot), $\dot{\boldsymbol{\theta}}, \dot{\sigma}^2$ are estimated by calculating the least-squared errors from a bootstrap sample taken from the observed data. In stochastic regression imputation (BLR.nob), parameters estimated from a regression model fitted to the observed values in the j^{th} variable are used to impute x_{ij}: $x_{ij} \sim N(\boldsymbol{X}_M\hat{\boldsymbol{\theta}}, \hat{\sigma})$ (Van Buuren, 2018). Van Buuren (2018) recommends using BLR.nob in large data sets with a small sampling variance since it is quicker and simpler to run than other other BLR approaches. If BLR.nob or BLR.boot is applied instead of BLR, the bias in the coefficient estimates is closer to zero for all imputation models, and the CRs are all in a good range. Additionally, BLR.nob results in the smallest AW, suggesting it may outperform other BLR-type methods. Therefore, alternative variations on BLR may impute the incomplete data better than BLR.

In SRS, missing values in x_{ij} are randomly replaced with observed values from the same variable. This method does not perform well, particularly for passive imputation, because the imputed values in the constituents may be incompatible to one another, resulting in very small or large values for the derived variable. For example, if γ_1 is a small observed value, and γ_2 is randomly imputed with a large value, then X_3 will be very small. As a result, SRS is not a good approach, particularly when performing passive imputation.

The approach for CART follows that of PMM outlined in Section 2.2.1 except a tree model is calculated instead of a regression model. A bootstrapped data set is created using the observed portion of the data. The tree model is then fitted to this bootstrapped data set, and predicted values are drawn from it. Under CART, the estimated coefficients are biased for all imputation models, except AWO. For all imputation methods applied to AWO, the CART method is the least biased method. However, there is undercoverage in this instance, suggesting that CART is not an appropriate alternative to use.

TABLE E.1: RB, CR, and AW from the substantive model for the pooled estimated coefficients for a ratio functional form. Here, different methods are applied in the MICE scheme when X_3 is MCAR and Z is not present.

Method	IM	RB	CR (%)	AW
BLR.nob	AWO	-0.00065	94.9	0.0228
	APA	-0.00056	95.0	0.0225
	PNP	-0.00151	94.7	0.0223
	LNP	-0.00007	95.7	0.0226
BLR.boot	AWO	-0.00063	95.6	0.0240
	APA	-0.00057	96.3	0.0230
	PNP	-0.00151	95.2	0.0229
	LNP	-0.00007	96.3	0.0231
CART	AWO	-0.00014	92.5	0.0221
	APA	-0.00150	95.1	0.0225
	PNP	-0.00128	94.1	0.0224
	LNP	-0.00131	95.0	0.0225
SRS	Active	-0.01652	10.1*	0.0225
	PNP	-0.02058	0.1*	0.0199
	LNP	-0.02055	0.1*	0.0199

* denotes that the CR is not in the 95% confidence interval.

Appendix F

Sensitivity Analysis Results when altering the sample size and censoring percentages.

Percentage bias (PB), coverage rate (CR), and average width (AW) for the estimates coefficients of a ratio derived variable when N and the proportion of censored observations is altered ($N = 500, 1000, 2000$; proportion of censored observations $= 10\%, 15\%, 20\%$).

The Wilson Score and Agressi-Coull confidence intervals in equations 4.2-4.3 respectively are recalculated for different values of n. For both confidence intervals, when $n = 500$, a reasonable CR is implied for an imputation model if the CR is in (93.1%, 96.9%). When $n = 1000$, a reasonable CR is implied for an imputation model if the CR is in (93.7%, 96.3%). When $n = 2000$, a reasonable CR is implied for an imputation model if the CR is in (94.1%, 95.9%).

TABLE F.1: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 500 and 10% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	0.46	96.3	0.0468	0.08	96.2	0.0452
		APA	0.26	96.3	0.0449	0.24	96.0	0.0439
		PNP	2.07	96.4	0.0447	1.75	96.3	0.0438
		LNP	0.78	96.6	0.0451	1.01	96.0	0.0442
	MAR1	AWO	1.88	95.8	0.0452	0.47	95.8	0.0440
		APA	1.23	94.7	0.0436	0.44	95.1	0.0429
		PNP	0.69	95.5	0.0435	1.61	95.1	0.0426
		LNP	2.85	94.7	0.0440	1.98	95.3	0.0432
	MAR2	AWO	4.91	93.5	0.0439	2.01	94.5	0.0428
		APA	3.13	93.8	0.0427	1.44	95.0	0.0421
		PNP	0.96	94.5	0.0424	0.75	95.3	0.0417
		LNP	5.33	93.3	0.0431	3.54	94.4	0.0425
MICE-PMM	MCAR	AWO	1.47	96.4	0.0465	0.77	96.2	0.0454
		APA	2.54	96.1	0.0453	2.12	95.4	0.0443
		PNP	1.79	96.3	0.0452	1.22	95.2	0.0440
		LNP	0.64	96.3	0.0449	0.77	95.5	0.0441
	MAR1	AWO	2.67	95.5	0.0453	1.82	94.9	0.0443
		APA	5.11	93.7	0.0442	3.18	94.7	0.0434
		PNP	4.48	94.4	0.0440	2.48	94.7	0.0431
		LNP	3.22	93.9	0.0439	2.09	94.8	0.0432
	MAR2	AWO	0.38	92.4*	0.0447	3.90	93.6	0.0434
		APA	7.79	92.0*	0.0435	5.30	93.5	0.0428
		PNP	7.16	92.7*	0.0433	4.39	93.8	0.0425
		LNP	5.88	93.0*	0.0432	3.92	93.7	0.0426
SMCFCS-BLR	MCAR	PNP	1.06	94.0	0.0446	2.62	94.9	0.0439
		LNP	0.78	94.8	0.0449	1.60	94.9	0.0435
	MAR1	PNP	0.13	94.9	0.0433	1.00	94.6	0.0428
		LNP	0.58	93.6	0.0437	0.25	94.8	0.0423
	MAR2	PNP	0.76	94.4	0.0429	0.67	94.1	0.0422
		LNP	0.05	95.1	0.0426	0.81	94.7	0.0417

* denotes that the CR is not in the 95% confidence interval.

TABLE F.2: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 500 and 15% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	3.37	94.8	0.0484	2.17	95.9	0.0467
		APA	2.47	95.8	0.0462	1.88	95.9	0.0451
		PNP	4.02	95.0	0.0460	3.89	95.9	0.0450
		LNP	1.01	94.7	0.0465	0.94	95.6	0.0456
	MAR1	AWO	0.59	95.5	0.0465	0.21	94.9	0.0451
		APA	0.25	94.8	0.0449	0.51	94.6	0.0440
		PNP	1.60	95.2	0.0447	2.46	95.0	0.0438
		LNP	2.22	93.7	0.0453	1.18	94.5	0.0445
	MAR2	AWO	4.22	94.4	0.0452	1.54	95.1	0.0442
		APA	2.18	94.6	0.0440	0.82	95.1	0.0433
		PNP	0.00	94.7	0.0437	1.48	95.4	0.0430
		LNP	4.46	94.4	0.0444	2.81	94.8	0.0438
MICE-PMM	MCAR	AWO	2.77	95.1	0.0481	2.85	95.3	0.0467
		APA	0.40	94.7	0.0467	0.05	95.1	0.0456
		PNP	0.21	94.7	0.0465	0.62	95.7	0.0454
		LNP	1.55	94.9	0.0463	1.09	95.6	0.0454
	MAR1	AWO	2.36	94.5	0.0468	1.41	94.7	0.0456
		APA	4.34	93.0*	0.0456	2.61	94.0	0.0445
		PNP	3.69	93.5	0.0454	1.68	94.4	0.0444
		LNP	2.42	94.1	0.0453	1.34	94.1	0.0445
	MAR2	AWO	0.08	89.5*	0.0461	3.36	94.4	0.0447
		APA	7.02	93.2	0.0448	4.46	94.2	0.0440
		PNP	6.27	93.5	0.0446	3.68	94.8	0.0437
		LNP	4.93	93.9	0.0444	3.31	94.3	0.0439
SMCFCS-BLR	MCAR	PNP	0.45	94.3	0.0438	0.91	95.4	0.0449
		LNP	0.28	95.4	0.0435	0.25	95.6	0.0452
	MAR1	PNP	0.68	94.9	0.0443	1.52	94.7	0.0438
		LNP	0.09	94.8	0.0440	0.20	95.2	0.0443
	MAR2	PNP	0.39	94.4	0.0441	2.27	95.2	0.0426
		LNP	0.40	94.5	0.0438	0.67	95.2	0.0431

* denotes that the CR is not in the 95% confidence interval.

TABLE F.3: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 500 and 20% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	4.01	95.3	0.0502	2.86	94.7	0.0485
		APA	3.41	94.5	0.0481	2.70	94.7	0.0470
		PNP	4.97	93.8	0.0478	4.64	95.0	0.0468
		LNP	1.89	94.5	0.0483	1.51	94.8	0.0474
	MAR1	AWO	0.57	94.5	0.0485	0.98	94.2	0.0470
		APA	1.13	93.9	0.0468	1.49	94.1	0.0459
		PNP	2.91	93.9	0.0465	3.70	93.9	0.0455
		LNP	1.00	93.3	0.0472	0.21	94.2	0.0464
	MAR2	AWO	2.51	94.2	0.0470	0.32	94.9	0.0459
		APA	1.12	95.2	0.0458	0.61	95.3	0.0450
		PNP	1.17	96.0	0.0455	2.79	95.2	0.0446
		LNP	3.27	94.7	0.0461	1.53	95.2	0.0455
MICE-PMM	MCAR	AWO	3.93	93.9	0.0499	3.42	95.0	0.0486
		APA	0.65	94.5	0.0486	0.78	94.1	0.0474
		PNP	1.19	94.4	0.0485	1.64	94.7	0.0472
		LNP	2.29	93.7	0.0483	1.85	95.2	0.0472
	MAR1	AWO	0.86	94.3	0.0487	0.55	94.5	0.0474
		APA	3.18	93.4	0.0475	1.39	93.8	0.0464
		PNP	2.51	94.3	0.0474	0.64	93.9	0.0463
		LNP	1.29	93.7	0.0472	0.49	93.7	0.0463
	MAR2	AWO	2.52	90.4*	0.0479	1.45	94.7	0.0464
		APA	5.49	93.1	0.0466	2.96	94.6	0.0457
		PNP	4.92	93.2	0.0463	2.08	94.5	0.0455
		LNP	3.76	93.9	0.0462	1.92	94.9	0.0454
SMCFCS-BLR	MCAR	PNP	0.47	94.7	0.0460	0.38	93.5	0.0458
		LNP	0.97	94.0	0.0457	1.57	94.1	0.0453
	MAR1	PNP	0.59	93.6	0.0463	0.82	93.8	0.0459
		LNP	1.08	93.6	0.0461	2.14	93.5	0.0454
	MAR2	PNP	0.73	93.5	0.0462	1.20	94.4	0.0453
		LNP	1.29	93.6	0.0459	2.65	94.1	0.0448

* denotes that the CR is not in the 95% confidence interval.

TABLE F.4: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 1000 and 10% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	1.86	95.1	0.0338	0.80	95.2	0.0317
		APA	2.27	95.7	0.0314	0.32	95.0	0.0308
		PNP	4.33	95.5	0.0314	2.05	95.5	0.0307
		LNP	1.45	95.6	0.0317	0.68	96.0	0.0310
	MAR1	AWO	0.73	95.5	0.0327	0.37	95.6	0.0307
		APA	0.15	95.8	0.0306	0.53	95.3	0.0301
		PNP	2.15	96.1	0.0304	1.44	96.1	0.0299
		LNP	1.43	95.3	0.0309	2.13	95.4	0.0303
	MAR2	AWO	5.02	93.6	0.0308	2.08	94.6	0.0302
		APA	3.15	94.3	0.0300	1.44	94.0	0.0296
		PNP	0.82	94.9	0.0299	0.76	94.4	0.0293
		LNP	5.06	92.6*	0.0302	3.34	93.6	0.0298
MICE-PMM	MCAR	AWO	2.14	94.6	0.0325	1.60	95.3	0.0318
		APA	1.57	95.0	0.0317	1.16	95.3	0.0310
		PNP	1.09	95.3	0.0317	0.74	95.6	0.0309
		LNP	0.23	95.8	0.0315	0.19	95.5	0.0309
	MAR1	AWO	2.51	94.9	0.0318	1.65	95.1	0.0310
		APA	5.07	93.7	0.0311	3.25	95.3	0.0304
		PNP	4.66	94.2	0.0309	2.66	95.3	0.0303
		LNP	3.16	94.8	0.0309	2.18	95.4	0.0304
	MAR2	AWO	1.04	86.4*	0.0316	3.76	93.9	0.0305
		APA	7.89	91.1*	0.0306	4.91	93.5	0.0300
		PNP	7.33	91.6*	0.0304	4.30	93.6	0.0299
		LNP	5.69	92.4*	0.0303	3.69	93.8	0.0299
SMCFCS-BLR	MCAR	PNP	1.10	95.5	0.0315	1.20	94.4	0.0453
		LNP	0.48	95.9	0.0313	2.65	94.1	0.0448
	MAR1	PNP	1.09	95.1	0.0307	1.20	94.4	0.0453
		LNP	0.40	95.8	0.0304	2.65	94.1	0.0448
	MAR2	PNP	0.34	94.9	0.0300	0.74	94.4	0.0453
		LNP	0.35	94.8	0.0298	2.65	94.1	0.0448

* denotes that the CR is not in the 95% confidence interval.

TABLE F.5: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 1000 and 15% of observations are censored.

			No Auxiliary Variables			One Auxiliary Variable		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	1.86	95.1	0.0338	1.41	95.7	0.0328
		APA	3.02	96.1	0.0324	1.07	96.1	0.0317
		PNP	5.02	96.1	0.0324	2.90	96.2	0.0316
		LNP	2.13	96.3	0.0327	0.08	96.3	0.0320
	MAR1	AWO	0.73	95.5	0.0327	0.51	94.6	0.0318
		APA	1.41	95.1	0.0315	0.74	95.2	0.0310
		PNP	3.65	95.1	0.0314	2.75	95.1	0.0309
		LNP	0.22	95.2	0.0319	0.85	95.0	0.0313
	MAR2	AWO	3.83	93.9	0.0318	1.07	94.0	0.0310
		APA	1.27	94.4	0.0309	0.72	94.4	0.0305
		PNP	1.01	94.8	0.0307	1.53	95.0	0.0302
		LNP	3.18	93.9	0.0313	2.71	93.7	0.0308
MICE-PMM	MCAR	AWO	2.30	95.2	0.0337	2.20	95.1	0.0327
		APA	0.95	95.8	0.0329	0.60	95.5	0.0320
		PNP	0.49	95.4	0.0328	0.02	95.9	0.0319
		LNP	0.71	95.9	0.0326	0.28	96.2	0.0319
	MAR1	AWO	1.45	95.0	0.0328	0.81	94.7	0.0321
		APA	3.75	93.7	0.0321	2.03	94.7	0.0314
		PNP	3.29	94.0	0.0320	1.40	94.5	0.0312
		LNP	1.83	94.7	0.0319	0.96	95.0	0.0313
	MAR2	AWO	0.76	84.6*	0.0325	2.81	92.5*	0.0314
		APA	6.92	91.3*	0.0316	4.26	92.7*	0.0309
		PNP	6.46	91.5*	0.0314	3.61	93.7	0.0308
		LNP	4.89	93.1*	0.0313	3.11	93.6*	0.0308
SMCFCS-BLR	MCAR	PNP	0.80	94.1	0.0317	1.10	95.4	0.0344
		LNP	0.12	94.2	0.0315	2.67	93.9	0.0342
	MAR1	PNP	0.13	94.7	0.0317	0.43	94.5	0.0338
		LNP	0.47	94.5	0.0314	1.74	94.8	0.0334
	MAR2	PNP	0.14	95.2	0.0308	1.20	94.8	0.0326
		LNP	0.98	95.0	0.0306	3.27	94.1	0.0330

* denotes that the CR is not in the 95% confidence interval.

TABLE F.6: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 1000 and 20% of observations are censored.

			No Auxiliary Variables			One Auxiliary Variable		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	1.24	95.3	0.0354	1.20	94.2	0.0342
		APA	2.31	94.6	0.0339	0.86	94.4	0.0331
		PNP	4.35	94.7	0.0337	2.75	94.4	0.0329
		LNP	1.26	94.5	0.0342	0.17	94.3	0.0334
	MAR1	AWO	0.52	94.2	0.0340	0.93	94.4	0.0331
		APA	1.46	94.8	0.0329	0.86	94.6	0.0324
		PNP	3.53	95.2	0.0328	2.84	94.3	0.0321
		LNP	0.33	94.6	0.0333	0.94	94.7	0.0326
	MAR2	AWO	3.69	93.3*	0.0331	0.64	94.7	0.0323
		APA	1.30	94.5	0.0322	0.38	95.0	0.0318
		PNP	1.03	94.9	0.0319	1.72	95.1	0.0315
		LNP	3.34	93.4*	0.0325	2.55	93.7	0.0321
MICE-PMM	MCAR	AWO	1.58	95.3	0.0353	1.89	94.4	0.0343
		APA	1.74	93.8	0.0343	0.82	94.1	0.0334
		PNP	1.33	94.7	0.0343	0.32	94.8	0.0333
		LNP	0.10	94.3	0.0341	0.02	94.5	0.0333
	MAR1	AWO	1.40	93.8	0.0343	0.63	94.4	0.0335
		APA	3.84	93.5*	0.0336	1.97	93.4*	0.0328
		PNP	3.38	93.7	0.0334	1.44	94.6	0.0326
		LNP	1.95	94.3	0.0332	1.12	93.9	0.0326
	MAR2	AWO	1.64	86.1*	0.0338	2.36	93.4*	0.0326
		APA	6.86	91.7*	0.0329	3.94	93.5*	0.0323
		PNP	6.36	91.8*	0.0327	3.38	93.6*	0.0321
		LNP	4.92	92.3*	0.0326	2.96	93.5*	0.0321
SMCFCS-BLR	MCAR	PNP	1.01	94.5	0.0340	1.41	94.1	0.0337
		LNP	0.49	93.9	0.0338	0.74	93.6	0.0333
	MAR1	PNP	0.09	94.7	0.0331	0.34	94.3	0.0329
		LNP	0.69	95.0	0.0328	0.09	94.3	0.0326
	MAR2	PNP	0.10	94.6	0.0323	0.77	94.4	0.0321
		LNP	0.63	94.6	0.0320	2.56	93.9	0.0319

* denotes that the CR is not in the 95% confidence interval.

TABLE F.7: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 2000 and 10% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	1.72	95.3	0.0231	1.33	95.8	0.0224
		APA	1.36	95.9	0.0223	1.16	95.1	0.0217
		PNP	3.14	94.8	0.0222	2.91	94.9	0.0217
		LNP	0.45	95.3	0.0224	0.27	95.2	0.0219
	MAR1	AWO	0.67	95.0	0.0223	0.75	95.0	0.0217
		APA	0.18	95.1	0.0216	0.84	95.5	0.0212
		PNP	2.13	95.5	0.0215	2.74	95.3	0.0211
		LNP	1.43	94.2	0.0218	0.67	95.0	0.0214
	MAR2	AWO	4.09	93.6*	0.0217	1.19	95.1	0.0213
		APA	2.13	95.4	0.0212	0.37	95.4	0.0209
		PNP	0.00	95.3	0.0210	1.72	95.2	0.0207
		LNP	4.08	94.6	0.0214	2.32	95.3	0.0211
MICE-PMM	MCAR	AWO	2.29	95.0	0.0230	2.28	95.4	0.0224
		APA	0.79	95.1	0.0224	0.14	94.7	0.0219
		PNP	0.56	94.9	0.0224	0.10	95.3	0.0218
		LNP	0.78	95.3	0.0223	0.55	95.2	0.0218
	MAR1	AWO	1.37	95.2	0.0224	0.40	95.4	0.0220
		APA	3.48	93.6*	0.0219	1.76	94.3	0.0215
		PNP	3.24	93.6*	0.0218	1.35	94.9	0.0214
		LNP	1.67	94.2	0.0218	0.84	95.2	0.0214
	MAR2	AWO	0.63	79.2*	0.0224	2.96	94.3	0.0216
		APA	6.78	90.7*	0.0216	3.91	94.4	0.0212
		PNP	6.41	92.2*	0.0215	3.36	94.1	0.0211
		LNP	4.75	93.9*	0.0214	2.67	94.5	0.0211
SMCFCS-BLR	MCAR	PNP	0.52	95.6	0.0222	0.81	95.5	0.0217
		LNP	0.08	95.5	0.0221	0.15	95.0	0.0215
	MAR1	PNP	0.46	95.5	0.0216	0.68	95.3	0.0212
		LNP	1.15	95.5	0.0214	1.95	95.4	0.0209
	MAR2	PNP	0.61	95.3	0.0211	1.38	95.6	0.0208
		LNP	1.43	95.8	0.0210	2.98	94.7	0.0205

* denotes that the CR is not in the 95% confidence interval.

TABLE F.8: PB, CR, and AW for the estimated coefficients of the derived variable in a exponential AFT substantive model when N = 2000 and 20% of observations are censored.

			No Auxiliary Variables			**One Auxiliary Variable**		
			PB	**CR (%)**	**AW**	**PB**	**CR (%)**	**AW**
MICE-BLR	MCAR	AWO	1.83	94.6	0.0250	0.87	96.9*	0.0231
		APA	1.69	95.3	0.0240	0.87	96.4*	0.0225
		PNP	3.57	95.3	0.0238	2.75	95.0	0.0223
		LNP	0.64	95.2	0.0241	0.07	97.0*	0.0226
	MAR1	AWO	0.42	94.8	0.0241	0.69	94.4	0.0233
		APA	0.00	94.3	0.0233	0.82	94.2	0.0228
		PNP	2.07	94.8	0.0232	2.67	94.0*	0.0227
		LNP	1.78	94.1	0.0235	0.99	94.3	0.0231
	MAR2	AWO	3.23	93.5*	0.0234	0.44	94.9	0.0228
		APA	1.69	94.4	0.0228	0.03	94.6	0.0225
		PNP	0.46	95.0	0.0227	2.12	94.4	0.0222
		LNP	3.90	93.2*	0.0230	2.11	94.4	0.0227
MICE-PMM	MCAR	AWO	2.21	94.5	0.0249	2.21	94.0	0.0241
		APA	0.80	94.7	0.0243	0.15	94.5	0.0236
		PNP	0.45	94.6	0.0242	0.25	94.5	0.0236
		LNP	0.82	94.7	0.0241	0.59	94.3	0.0235
	MAR1	AWO	1.29	93.8*	0.0242	0.59	93.6*	0.0236
		APA	3.85	93.2*	0.0237	1.95	93.9*	0.0231
		PNP	3.45	93.4*	0.0236	1.55	93.8*	0.0231
		LNP	2.05	93.7*	0.0235	1.15	94.6	0.0230
	MAR2	AWO	1.74	78.5*	0.0241	2.08	93.9*	0.0231
		APA	6.48	90.3*	0.0232	3.52	93.6*	0.0228
		PNP	6.07	90.6*	0.0231	2.95	94.1	0.0227
		LNP	4.55	92.8*	0.0230	2.52	94.7	0.0227
SMCFCS-BLR	MCAR	PNP	0.09	94.2	0.0240	0.34	94.3	0.0234
		LNP	0.53	94.2	0.0238	0.67	93.7*	0.0232
	MAR1	PNP	0.26	94.6	0.0233	0.55	94.4	0.0228
		LNP	0.96	94.6	0.0231	1.86	93.8*	0.0226
	MAR2	PNP	0.45	94.5	0.0227	1.34	93.5*	0.0223
		LNP	1.21	94.0*	0.0225	2.92	93.1*	0.0221

* denotes that the CR is not in the 95% confidence interval.

References

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

Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J. (2011), 'Multiple imputation by chained equations: What is it and how does it work?', *International Journal of Methods in Psychiatric Research* **20**(1), 40–49.

Bartlett, J. W. and Morris, T. P. (2015), 'Multiple imputation of covariates by substantive-model compatible fully conditional specification', *The Stata Journal* **15**(2), 437–456.

Brand, J. P., Van Buuren, S., Groothuis-Oudshoorn, K. and Gelsema, E. S. (2003), 'A toolkit in SAS for the evaluation of multiple imputation methods', *Statistica Neerlandica* **57**(1), 36–45.

Brown, L. D., Cai, T. T. and DasGupta, A. (2001), 'Interval estimation for a binomial proportion', *Statistical Science* **16**(2), 101–133.

Burton, A., Altman, D. G., Royston, P. and Holder, R. L. (2006), 'The design of simulation studies in medical statistics', *Statistics in Medicine* **25**(24), 4279–4292.

Button, K. S., Ioannidis, J., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. and Munafò, M. R. (2013), 'Power failure: why small sample size undermines the reliability of neuroscience', *Nature Reviews Neuroscience* **14**(5), 365–376.

Carpenter, J. and Kenward, M. (2012), *Multiple Imputation and its Applications*, Wiley and Sons.

CDC (2000), 'Nhanes (1999-2000)'. Accessed: February-2019.
URL: *https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=1999*

CDC (2015), 'Body mass index (BMI)'. Accessed: August-2019.
URL: *https://www.cdc.gov/healthyweight/assessing/bmi/index.html*

Clark, T. G. and Altman, D. G. (2003), 'Developing a prognostic model in the presence of missing data: An ovarian cancer case study', *Journal of Clinical Epidemiology* **56**(1), 28–37.

Clements, L., Kimber, A. C. and Biedermann, S. (2022), 'Multiple imputation of composite covariates in survival studies', *Stats* **5**(2), 358–370.

CLHLS (2018), 'Chinese longitudinal healthy longevity survey (CLHLS) series'. Accessed: October-2019.
URL: *https://www.icpsr.umich.edu/icpsrweb/NACDA/series/487*

Collett, D. (2015), *Modelling Survival Data in Medical Research*, third edn, Chapman and Hall CRC.

Covelli, H. D., Nessan, V. J. and Tuttle 3rd, W. (1983), 'Oxygen derived variables in acute respiratory failure.', *Critical Care Medicine* **11**(8), 646–649.

Cox, D. R. (1972), 'Regression models and life-tables', *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202.

Desai, M., Mitani, A. A., Bryson, S. W. and Robinson, T. (2016), 'Multiple imputation when rate of change is the outcome of interest', *Journal of Modern Applied Statistical Methods* **15**(1), 160–192.

Eekhout, I., de Vet, H. C., de Boer, M. R., Twisk, J. W. and Heymans, M. W. (2018), 'Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales', *Statistical Methods in Medical Research* **27**(4), 1128–1140.

Enders, C. K. (2010), *Applied Missing Data Analysis*, Guilford press.

Fish, J. S., Ettner, S., Ang, A. and Brown, A. F. (2010), 'Association of perceived neighborhood safety on body mass index', *American Journal of Public Health* **100**(11), 2296–2303.

Gmel, G. (2001), 'Imputation of missing values in the case of a multiple item instrument measuring alcohol consumption', *Statistics in Medicine* **20**(15), 2369–2381.

Graham, J. W., Olchowski, A. E. and Gilreath, T. D. (2007), 'How many imputations are really needed? some practical clarifications of multiple imputation theory', *Prevention Science* **8**(3), 206–213.

Grobler, A. C. and Lee, K. (2020), 'Multiple imputation in the presence of an incomplete binary variable created from an underlying continuous variable', *Biometrical Journal* **62**(2), 467–478.

Hand, D. J. (2020), *Dark Data: Why What You Don't Know Matters*, Princeton University Press.

Hardt, J., Herke, M. and Leonhart, R. (2012), 'Auxiliary variables in multiple imputation in regression with missing x: a warning against including too many in small sample research', *BMC Medical Research Methodology* **12**(1), 184–196.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J. and Brindle, P. (2007), 'Qrisk cardiovascular disease risk prediction algorithm – comparison of the revised and the original analyses. technical supplement 1', *Q-Research* .

HippisleyCox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M. and Brindle, P. (2007), 'Derivation and validation of qrisk, a new cardiovascular disease risk score for the united kingdom: Prospective open cohort study', *BMJ* **335**(7611), 136–147.

Janssen, K. J., Donders, A. R. T., Harrell Jr, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E. and Moons, K. G. (2010), 'Missing covariate data in medical research: To impute is better than to ignore', *Journal of Clinical Epidemiology* **63**(7), 721–727.

Jochen, H., Max, H., Tamara, B. and Wilfried, L. (2013), 'Multiple imputation of missing data: A simulation study on a binary response', *Open Journal of Statistics* **3**(5), 370–378.

Johnson, R. A. and Wichern, D. W. (2007), *Applied multivariate statistical analysis*, Vol. 6, Pearson.

Kalla, S. (2011), 'Measurement scales', Explorable.com: `https://explorable.com/measurement-scales`. Accessed: April-2022.

Kang, H. (2013), 'The prevention and handling of the missing data', *Korean Journal of Anesthesiology* **64**(5), 402–406.

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G. and Scheike, T. H. (2014), *Handbook of survival analysis*, CRC Press Boca Raton, FL:.

Lagona, F. and Zhang, Z. (2010), 'A missing composite covariate in survival analysis: A case study of the chinese longitudinal health and longevity survey', *Statistics in Medicine* **29**(2), 248–261.

Lambert, P. C. and Royston, P. (2009), 'Further development of flexible parametric models for survival analysis', *The Stata Journal* **9**(2), 265–290.

Lawless, J. F. (2011), *Statistical models and methods for lifetime data*, Vol. 362, John Wiley & Sons.

Lee, K. J. and Carlin, J. B. (2012), 'Recovery of information from multiple imputation: A simulation study', *Emerging Themes in Epidemiology* **9**(1), 1–10.

Lent, R. W., Brown, S. D. and Larkin, K. C. (1987), 'Comparison of three theoretically derived variables in predicting career and academic behavior: Self-efficacy, interest congruence, and consequence thinking.', *Journal of Counseling Psychology* **34**(3), 293—-298.

Ling, A. Y., Montez-Rath, M. E., Mathur, M. B., Kapphahn, K. and Desai, M. (2019), ‘How to apply multiple imputation in propensity score matching with partially observed confounders: A simulation study and practical recommendations’, *arXiv preprint arXiv:1904.07408* .

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

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

Marchenko, Y. and Eddings, W. (2011), ‘A note on how to perform multiple-imputation diagnostics in stata’, http://www.stata.com/users/ymarchenko/midiagnote.pdf. Accessed: May-2020.

Marshall, A., Altman, D. G. and Holder, R. L. (2010), ‘Comparison of imputation methods for handling missing covariate data when fitting a Cox Proportional Hazards model: A resampling study’, *BMC Medical Research Methodology* **10**(1), 112–121.

McCleary, L. (2002), ‘Using multiple imputation for analysis of incomplete data in clinical research’, *Nursing Research* **51**(5), 339–343.

Mitani, A. A., Kurian, A. W., Das, A. K. and Desai, M. (2015), ‘Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects’, *Statistical Methodology* **27**, 82–99.

Molenberghs, G. and Kenward, M. (2007), *Missing Data in Clinical Studies*, Wiley and Sons.

Morris, T. P., White, I. R., Royston, P., Seaman, S. R. and Wood, A. M. (2014), ‘Multiple imputation for an incomplete covariate that is a ratio’, *Statistics in Medicine* **33**(1), 88–104.

Nguyen, C. D., Carlin, J. B. and Lee, K. J. (2017), ‘Model checking in multiple imputation: An overview and case study’, *Emerging Themes in Epidemiology* **14**(1), 1–12.

ONS (2010), ‘The average briton’, https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/theaveragebriton. Accessed: May-2020.

Pangman, V. C., Sloan, J. and Guse, L. (2000), ‘An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: Implications for clinical practice’, *Applied Nursing Research* **13**(4), 209–213.

Pankhurst, L., Mitra, R., Kimber, A. and Collett, D. (2020), 'Multiply imputing missing values arising by design in transplant survival data', *Biometrical Journal* **62**(5), 1192–1207.

Rodwell, L., Lee, K. J., Romaniuk, H. and Carlin, J. B. (2014), 'Comparison of methods for imputing limited-range variables: A simulation study', *BMC Medical Research Methodology* **14**(1), 57–67.

Rubin, D. (1976), 'Inference and missing data', *Biometrika* **63**(3), 581–592.

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

Rubin, D. B. (1987), *Multiple Imputation for Survey Non-Response*, Wiley and Sons.

Schaefers, K. G., Epperson, D. L. and Nauta, M. M. (1997), 'Women's career development: Can theoretically derived variables predict persistence in engineering majors?', *Journal of Counseling Psychology* **44**(2), 173–183.

Schenker, N. and Taylor, J. M. (1996), 'Partially parametric techniques for multiple imputation', *Computational Statistics & Data Analysis* **22**(4), 425–446.

Seaman, S. R., Bartlett, J. W. and White, I. R. (2012), 'Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods', *BMC Medical Research Methodology* **12**(1), 1–13.

Segal, T. (2021), 'Profit margin', https://www.investopedia.com/terms/p/profitmargin.asp. Accessed: April-2022.

Slade, E. and Naylor, M. G. (2020), 'A fair comparison of tree-based and parametric methods in multiple imputation by chained equations', *Statistics in Medicine* **39**(8), 1156–1166.

Stuart, E. A., Azur, M., Frangakis, C. and 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.

Tilling, K., Williamson, E. J., Spratt, M., Sterne, J. A. and Carpenter, J. R. (2016), 'Appropriate inclusion of interactions was needed to avoid bias in multiple imputation', *Journal of Clinical Epidemiology* **80**, 107–115.

Tombaugh, T. N. and McIntyre, N. J. (1992), 'The mini-mental state examination: a comprehensive review', *Journal of the American Geriatrics Society* **40**(9), 922–935.

UCLA (2017), 'Multiple imputation in Stata', `https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/`. Accessed: April-2020.

Van Buuren, S. (2018), *Flexible Imputation of Missing Data*, Chapman and Hall/CRC.

Van Buuren, S., Boshuizen, H. C. and Knook, D. L. (1999), 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medicine* **18**(6), 681–694.

van Holm, E. (2021), *Introduction to Research Methods*, Bookdown.

Vink, G. and Buuren, S. v. (2017), 'Mice: Passive imputation and post-processing', `https://www.gerkovink.com/miceVignettes/Passive_Post_processing/Passive_imputation_post_processing.html`. Accessed: March-2019.

von Hippel, P. (2009), 'How to impute interactions, squares, and other transformed variables', *Sociological Methodology* **39**(1), 265–291.

Wagstaff, D. A., Kranz, S. and Harel, O. (2009), 'A preliminary study of active compared with passive imputation of missing body mass index values among non-hispanic white youths', *The American Journal of Clinical Nutrition* **89**(4), 1025–1030.

Wayman, J. C. (2003), Multiple imputation for missing data: What is it and how can I use it? Paper presented at The Annual Meeting of the American Educational Research Association, Chicago, IL.

White, I. R. and 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.

White, I. R. and Royston, P. (2009), 'Imputing missing covariate values for the Cox model', *Statistics in Medicine* **28**(15), 1982–1998.

White, I. R., Royston, P. and Wood, A. M. (2011), 'Multiple imputation using chained equations: Issues and guidance for practice', *Statistics in Medicine* **30**(4), 377–399.

Yi, Z. and Vaupel, J. W. (2002), 'Functional capacity and self–evaluation of health and life of oldest old in china', *Journal of Social Issues* **58**(4), 733–748.

Zhang, Z. (2016), 'Missing data imputation: Focusing on single imputation', *Annals of Translational Medicine* **4**(1), 9–17.