Ludwig-Maximilians-University Munich
Department of Statistics

bachelor's thesis

A Comparative Simulation Study of Imputation Methods

Author:	Rui Yang
Supervisor:	Prof. Dr. Christian Heumann
Work group:	Methods for Missing Data, Model Selection and Model Averaging
Date:	February 04, 2019

Abstract

The field of data science often faces the problem of missing data, especially for large-scale data. If missing data is not handled properly, to a certain degree this has a negative impact on the validity of statistical research results. Missing data imputation is an option to deal with this problem. This thesis conducts a simulation study in order to quantitatively analyze the performance of different imputation methods applied to a data set with missing values under a variety of different missing rates and missing data mechanisms.

The imputation methods compared in this simulation study are mean substitution, which is a single imputation method, and the multiple imputation method, with the help of three powerful R-packages: MICE, Amelia II, and missForest. To enable comparison, the predicted residual error sum of squares (PRESS) statistic is selected as the evaluation criterion, and is calculated based on selected models after conducting variable selection. The comparison results are presented in the form of boxplots comprising the log-transformed PRESS statistic values of the four imputation methods.

According to the comparison results three main conclusions can be drawn. First, missForest always exhibits the best performance, regardless of the missing rate and the missing data mechanism. Second, the performances of MICE and Amelia II do not show a fixed pattern. Third, mean substitution performs better than both MICE and Amelia II in certain situations.

Keywords: Missing data, imputation, mean substitution, MICE, Amelia II, missForest, variable selection, PRESS statistic

Abbreviations and Notations

Abbreviations:

MCAR	Missing completely at random
MAR	Missing at random
NMAR	Not missing at random
PRESS	Predicted residual error sum of squares
EM	Expectation Maximization
MI	Multiple Imputation

Notations:

$Y = (y_{ij})$	Complete data
$M = (M_{ij})$	Misssing-data indicator matrix
ϕ	Unknown parameters
Y_{obs}	Observed components
Y_{mis}	Missing components
Y_{obs}	Observed components
p	Probability

Contents

List of Figures

List of Tables

1 Introduction and Overview

Since almost all statistical analyses are based on data, statistical forecasts with a lack of high-quality data are prone to inaccuracy. When the probability of missingness is extremely small, the missing values may be omitted from the data set in certain situations or processed manually. However, the proportions of missing values are generally large for specific variables in actual data. In this case it is inefficient and time-consuming to process manually, and also tends to produce errors. Specifically, when the quantity of missing data is relatively large (greater than 10%) the results of subsequent statistical analysis may be biased (Derrick A. Bennett (2009)). In general, if the negative influence caused by missing data is not considered during the analyzing process, the results of the statistical forecasts will be biased and may even lead to erroneous conclusions. Therefore, it is necessary to choose an appropriate method to handle the missing data.

In practice, data may be missing due to many different factors, such as the loss of questionnaires in a survey or the reluctance of respondents to answer. To handle the remaining data correctly, it is crucial to understand the forms of missingness and the possible reasons that lead to them. According to (Roderick J. A. Little, Donald B. Rubin (2002)) it is well known that standard statistical methods have been developed to analyze rectangular data sets. Rows of data represent units, which can also be called cases or observations depending on the context, and columns represent the variables measured for each unit. Based on this prerequisite, the form of missingness can be classified into two categories as listed below.

1.**Unit missing**, also called unit nonresponse. This refers to the missingness situation whereby an interviewee does not provide sufficient information for the response to be considered of use, or even provides no information at all. For example, an epidemiological survey of lung cancer and smoking habits conducted on 1,000 smokers was carried out using a questionnaire. After recycling, the number of effective questionnaires is 500, indicating that the effective questionnaire recycling rate is 50%. Possible reasons for this rate are that the respondents are not familiar with the questionnaire or did not want to answer the questions.

2.**Item missing**, also called item nonresponse. This refers to the missing situation whereby answers to certain questions are absent after the interviewee has agreed to take part in the survey (Ting Yan, Richard Curtin (2010)). For example, in order to test different types of drugs used to treat high blood pressure, the blood pressure of each participant was recorded at times 0, 1, 2, 3, and 4 weeks after the start of the experiment. However, a common missing data problem arose after 2 weeks when some participants quit before the end of the study and did not return. This problem is especially noticeable for longitudinal data. Furthermore, the pattern of missing values is an example of monotone missing data, as presented in Figure 1.1 b).

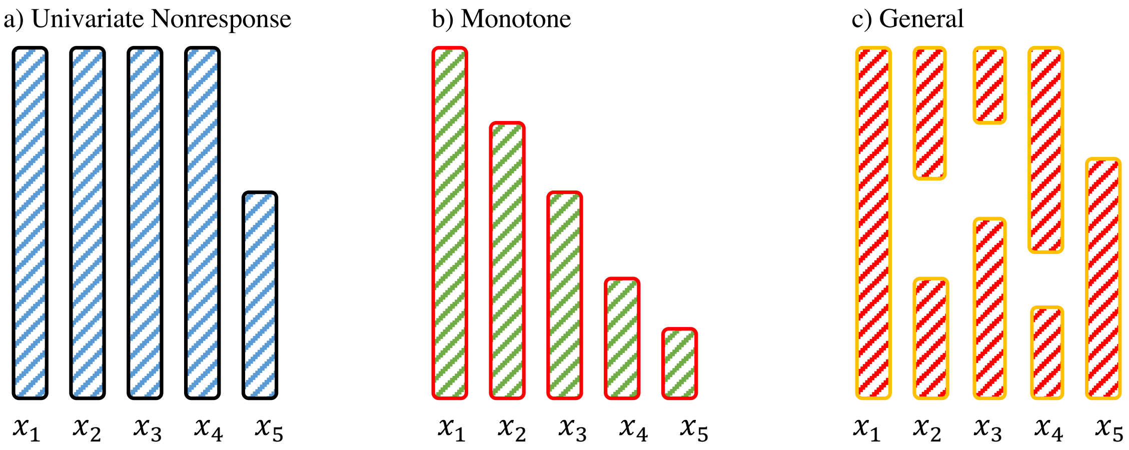


Figure 1.1: Example of missing-data patterns.

In addition to monotone missing data, other missing data patterns can be identified. For instance, Figure 1.1 a) indicates univariate missing data, whereby a single variable has missing values. In reality, the pattern of missing data is always neither monotone nor univariate nonresponse. The most common missing data pattern is the general missing data pattern shown in Figure 1.1 c), where multiple variables have missing values simultaneously with random missingness for each variable. Accordingly, this bachelor's thesis concentrates on the general missing data pattern.

Regardless of whether the form of missingness is unit nonresponse or item nonresponse, the missing data mechanism can be further divided into three types: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR) (S. Fielding, P. M. Fayers and C. R. Ramsay (2009)). Each of the three missing data mechanisms implies a relationship between the missingness rate and values of both the missing and the observed data. Regarding the missing data mechanisms, explanations and mathematical definitions are discussed in detail in section 3.

Besides the identification of the missing data mechanism, variable selection should also be conducted. Otherwise, a large number of redundant variables will remain in the regression model. Without variable selection, these redundant variables will introduce irrelevant information ("noise") into the model, which is one of the main causes of overfitting. Therefore, variable selection is an important component of this thesis. Two variable selection methods are used in the simulation study: backward elimination and forward selection. In order to compare the results produced by these two methods, they are applied to the same data set. The detailed process of variable selection is described in section 4. In addition, the value of the predicted residual error sum of squares (PRESS) is calculated based on the selected models, the details of which are provided in section 5.3.

1. INTRODUCTION AND OVERVIEW

2 Simulations and Examples

2.1 Introduction to Simulation

This section presents the motivation behind carrying out a simulation study. Simulation studies play an important role in statistical research. A simulation is an imitation of the operation of a real-world process or system (J. Banks; J. Carson; B. Nelson; D. Nicol (2001)). This definition implies that a simulation is constructed such that the product is identical to the reality. In this thesis it is advantageous to conduct a simulation because it is an efficient way to compare different imputation methods under various conditions. In addition, it is rarely possible to identify the missing data pattern of an actual data set with missing values. However, the desired missing data pattern can be simulated with the help of simulation studies.

Algorithm

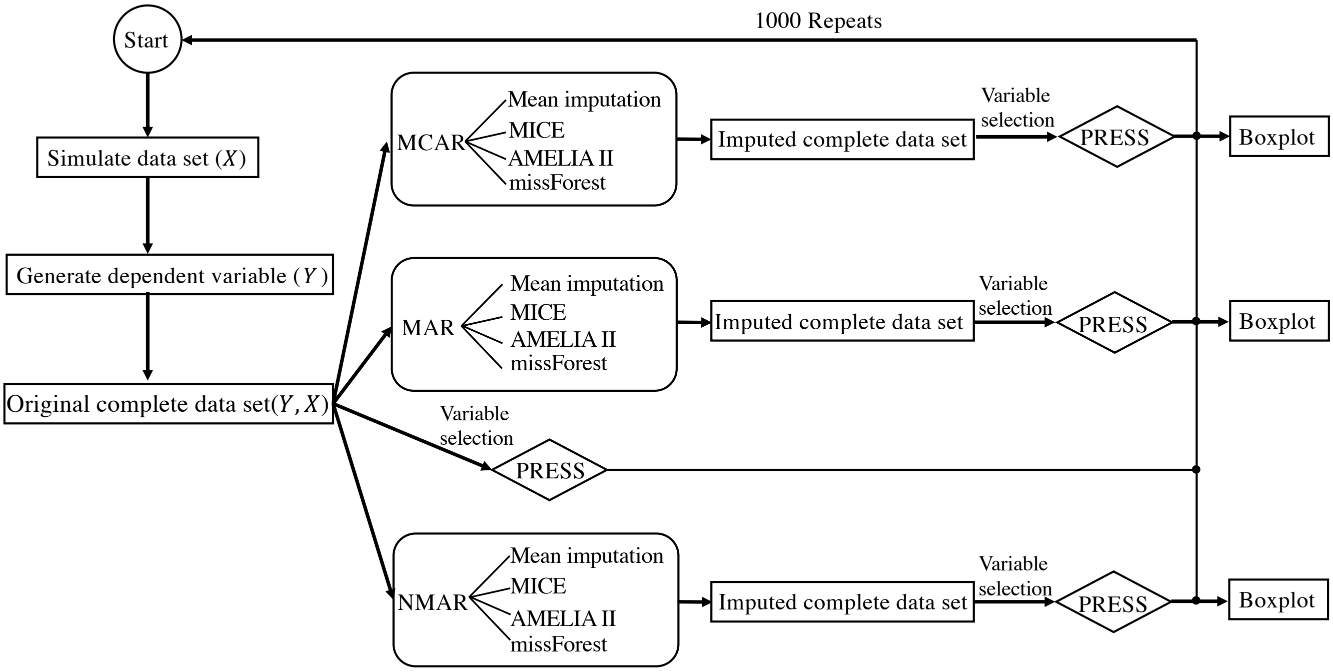


Figure 2.1: The "6 steps": an algorithm for comparison of imputation methods for simulated data.

Figure 2.1 presents an unambiguous algorithm to determine the PRESS statistic in order to compare different imputation methods. The algorithm includes a series of steps that each perform a particular computation or task, and generally runs with six steps:

1. Simulate a data set (X) with different types of variables,

2. Generate a dependent variable (Y) from a Poisson distribution,

3. Simulate three types of missing data mechanisms,

4. Use different methods to impute missing values,

5. Perform variable selection for the original complete data set and imputed complete data set,

6. Calculate the PRESS statistic.

First, the six steps of the algorithm are defined. These six steps are then run 1,000 times and all repeat loop outputs are stored in a matrix, after which boxplots based on this matrix are created. According to the results displayed in the boxplots, the different imputation methods can then be compared visually.

In the next section, the original complete data set containing different types of variables is explained. Multiple types of missing data are then simulated and analyzed in order to compare a range of imputation methods under certain conditions.

2.2 Generating Simulated Data Set

A variety of methods can be used to impute missing data, the effects of which depend largely on the simulated data set. Many factors can significantly affect the result of the comparison of different imputation methods. These include different types of variables, such as continuous variables and categorical variables; the size of the data set; and the missing rate. Therefore, in order to obtain a more convincing result from the comparison it is necessary to introduce the simulated data set in detail. In this section the original complete data sets are presented, based on which missing data are generated. Four different methods are then applied to impute these missing data, and the performances of the methods are evaluated and compared.

2.2.1 Types of Variables

In this simulation study two kinds of variables are simulated: continuous variables and categorical variables.

A continuous variable is one of two types of numerical variables which takes on infinite and uncountable values and is always collected in the form of numbers, despite the fact that other types of data also appear in the form of numbers. Examples of continuous variables include the number of gallons of milk that a cow produces, or the length of time taken for a train to travel from one city to another. In contrast to continuous variables, a discrete variable can only take on a certain number of values, meaning that a discrete variable is numerical and countable. In other words, if a set of items can be counted, then it is a discrete variable. Examples of discrete variables are the number of applicants who apply for a vacant position at a company, or the number of students who enroll in a university at the start of a semester.

Categorical variables are another type of variable and differ from numerical variables. A categorical variable is a type of statistical variable that can take on one of a finite and usually fixed number of possible values. Examples of categorical variables include the breed of a cat (e.g. Ragamuffin, American Shorthair, Scottish Fold) or the brand of a pair of shoes. Based on previously known qualitative properties, this kind of variable assigns each individual or other single unit of observed objects to a specific group or nominal category (Daren S. Starnes(2012)). This simulation study includes three categorical variables, two of which are binary variables: gender and smoker status. The two possible outcomes of the gender variable are "Male" and "Female," whereas "Yes" and "No" are the possible outcomes of the smoker status variable. Another simulated variable is occupation class, which is a multi-way variable. Multi-way variables have more than two possible outcomes; in this simulation study occupation class has four possible outcomes, which are "A", "B", "C", and "D".

In the field of life and health reinsurance, smoker status and occupation class are two of the most significant risk factors that influence the price of an insurance premium for an insurance policy. Gender is also a rather important characteristic of the person being insured. In disability and mortality studies of reinsurance companies, the consideration of gender is shown to improve the accuracy of insurance product pricing. As mentioned above, gender, smoker status, and occupation class are simulated in this study.

Different types of variables indicate various kinds of distributions. Based on generated data, the theories of various distributions are briefly explained in the following sections. Graphs are also included to provide further detail for this simulation study.

Binomial distribution

The binomial distribution is a common discrete probability distribution used in statistics. Here the possible outcome of a single trial takes one of two independent values having a specified set of parameters and assumptions. The parameters are established as n and p, where n represents the number of trials and p represents the probability of success in each trial. More specifically, for a single trial (where n is equal to 1) the binomial distribution can be classified as a Bernoulli distribution. An example is the result of a university exam which may be either "pass" or "fail". If a random variable X has the Bernoulli distribution, then it can be presented as:

$$\Pr(X=1) = p = 1 - \Pr(X=0) = 1 - q \tag{1}$$

The probability mass function of this distribution with possible outcomes k is written as:

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1-p & \text{if } k = 0. \end{cases} \tag{2}$$

In general, the Bernoulli distribution can simply be written as $X \sim \mathrm{B}(1,p)$ or $X \sim \mathrm{Bernoulli}(p)$.

Generally, the binomial distribution is the sum of multiple Bernoulli trials. Remarkably, there are three assumptions of the binomial distribution, which are listed as follows.

- There is only one outcome for each trial.
- Each trial is mutually exclusive or indepent.
- Each trial has the same probability of success.

A typical example of the binomial distribution would be the results of flipping a coin for multiple times, which are either "head" or "tail". The probability mass function of this distribution is written as:

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, n \in \mathbb{N}, p \in [0,1] \tag{3}$$

for $k = 0, 1, 2, ..., n$, where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

In comparison to the Bernoulli distribution, the binomial distribution can simply be written as $X \sim \mathrm{B}(n,p)$. This simulation study includes two variables that follow the binomial distribution, namely gender and smoker status.

Multinomial distribution

In probability theory the multinomial distribution is a generalization of the binomial distribution. In the latter, the number of possible outcomes or categories k equals two, whereas in the multinomial distribution k is larger than two and the number of trials n is larger than one. To be more specific, for a single trial (when n is equal to one) the multinomial distribution can be classified as a categorical distribution, which is an extended distribution of the Bernoulli distribution for a categorical random variable. In this case the sum of the probabilities of all possible outcomes is equal to one.

A classic example of categorical distribution is shown by the possible outcomes of rolling a dice once, which are $\{1,2,...,6\}$ with the same probability of $\frac{1}{6}$. If a random variable X has the categorical distribution, then the probability mass function f can be presented as:

$$f(x = i \mid p) = p_i, \tag{4}$$

where $p = (p_1, \ldots, p_k)$ represents the probability of the ith category and $\sum_{i=1}^{k} p_i = 1$.
According to Minka, T. (2003), a more complicated mathematical formulation is written as:

$$f(x \mid p) = \prod_{i=1}^{k} p_i^{[x=i]} \tag{5}$$

where $[x = i]$ evaluates to 1 if $x = i$, 0 otherwise.
As mentioned above, the multinomial distribution can be applied to model the probabilities of more than two possible categories over n trials. An example of this distribution is provided by the results of the German federal election, whereby several parties run for political leadership in Germany, thus implying that k is larger than 2. In this case every lawful voter supports one of many parties. As there are millions of voters, this implies that n is larger than 1.
The probability mass function of this multinomial distribution is:

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_k = x_k) \tag{6}$$

$$= \begin{cases} \dfrac{n!}{x_1! \cdots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k}, & \text{when } \sum_{i=1}^{k} x_i = n \\ \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

for non-negative integers $x_1, x_2, ..., x_k$
In this simulation study one variable follows the multinomial distribution, which is the occupation class having four possible categories.

The following figure summarizes the relationships between the four types of distributions discussed above. When k is two and n is one, the multinomial distribution is the Bernoulli distribution. When k is two and n is larger than one, it is the binomial distribution. When k is larger than two and n is one, it is the categorical distribution. When k is larger than two and n is larger than one, it is the multinomial distribution.

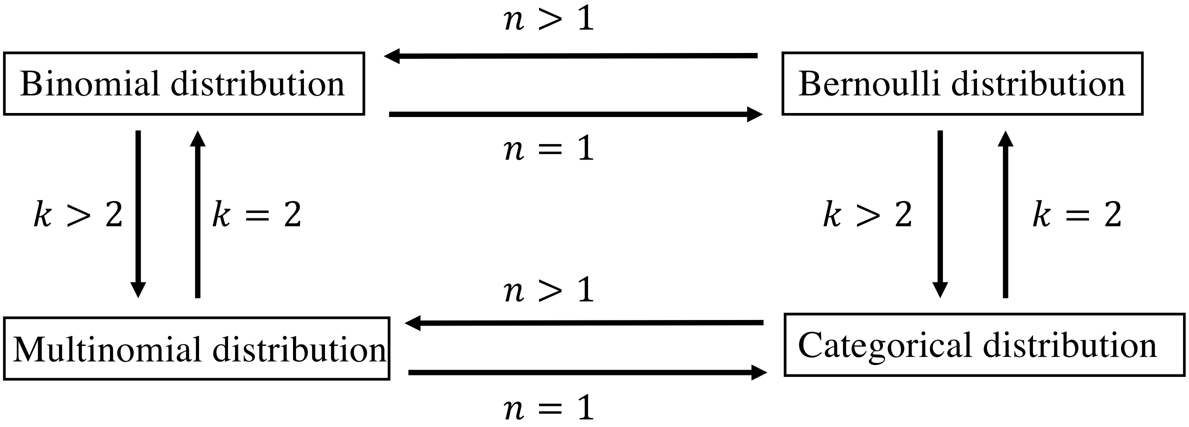


Figure 2.2: Relationships among 4 distributions for categorical variables

Multivariate normal distribution

In this data set seven variables are simulated, which altogether comprise a multivariate distribution. In probability theory, unlike a discrete probability distribution, the multivariate normal distribution is a relatively common continuous probability distribution, on the basis of which several variables are simulated in this study. The multivariate normal distribution, also termed the multivariate Gaussian distribution, is one of the most important multivariate distributions. Indeed, it is the multivariate form of the univariate (one-dimensional) normal distribution.

The normal distribution is a crucial probability distribution. Its two parameters, mean and variance, determine the shape of the probability density curve. The most significant characteristic of the normal distribution is symmetry; this implies that most of the observations are situated around the central peak and that the probabilities for values further from the mean decrease in both directions to the same degree. The steepness of the curve depends on the variance.

The probability density of the univariate normal distribution is

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), x \in \mathbb{R} \tag{8}$$

where μ is the mean or expectation of the distribution and σ^2 is the variance. Specifically, when a random variable X is normally distributed, the

mathematical notation can simply be written as $X \sim N(\mu, \sigma^2)$.
Just as mentioned above, the multivariate normal distribution is a generalization of the univariate normal distribution to higher dimensions. To be more specific, a random vector $\mathbb{X} = (X_1, X_2, ..., X_k)^T$ is multivariate normal if for any constants $a_1, a_2, .., a_k$ every linear combination of these random variables $X_1, X_2, ..., X_k$ has a normal distribution,

$$a_1X_1 + a_2X_2 + ... + a_kX_k$$

is normally distributed.
A k-variate normally distributed random variable $\mathbb{X} = (X_1, X_2, ..., X_k)^T$ has density function

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mu)\right)}{\sqrt{(2\pi)^k det(\Sigma)}} \qquad \mu \in \mathbb{R}^k, \quad \Sigma \in \mathbb{R}^{k\times k} \tag{9}$$

where $\mu = \mathrm{E}[\mathbf{X}] = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]^{\mathrm{T}}$ is the known k-dimensional mean vector.
If there is completely no correlation among the simulated variables, indicating that under all circumstances the covariance is equal to 0, then it contradicts with the reality because in reality there is a correlation among variables to a certain extent. Therefore, the covariance matrix among multiple random variables is generally defined as follows:
The $k \times k$ covariance matrix

$$\Sigma =: \mathrm{E}((\mathbf{X}-\mu)(\mathbf{X}-\mu)^{\mathrm{T}}) = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_k, X_1) & \mathrm{Cov}(X_k, X_2) & \cdots & \mathrm{Var}(X_k) \end{pmatrix}$$

After the concrete values of mean and variance are determined, variables that match the multivariate normal distribution could be generated. Similar to the univariate normal distribution, the mathematical notation of multivariate normal distribution can be written as $X \sim N_k(\mu, \sigma^2)$ where k components has a univariate normal distribution.

Poisson distribution

In the field of reinsurance the Poisson distribution is often applied to describe the number of losses in a portfolio. The Poisson distribution is a discrete probability distribution that presents the probability of a number of independent events occurring within a specified interval, where a known constant rate λ is given (Frank A. Haight (1967)). In the case of reinsurance the constant rate λ is the expected value of the number of losses, which is not necessarily an integer in reality. The horizontal axis usually represents the number of losses, which is a discrete random variable, while the vertical axis is the probability of losses given λ.
In the simulated data set the response or dependent variable Y is generated based on the Poisson distribution, which is a particular distribution in the exponential family. This family has a mass function or probability density function of the following form:

$$f(y_i,|\theta_i,\phi_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i,\phi_i)\right) \tag{10}$$

where

θ_i	is the natural parameter of the family
ϕ_i	is a scale or dispersion parameter and
$b(.)$ and $c(.)$	are specific function corresponding to the type of the family

The Poisson distribution is included in the generalized linear model, which is a flexible generalization of linear regression that considers response variables whose error distribution models are not restricted to a normal distribution. The response variable Y is generated by applying the linear predictor as follows:

$$\eta_i = x_i^T\beta \tag{11}$$

where x_i^T stands for the independent variables, and β represents the regression coefficients, which are used to estimate the unknown population parameters and to describe the statistical relationship between one or more independent variables and the response variable.
In the generalized linear model the link function is a crucial element. Generally, it can be written as

$$g(\mu_i) = \eta_i = x_i^T \beta \tag{12}$$

The conditional expectation $\mu_i = E(y_i|x_i)$ is determined by

$$\mu_i = h(\eta_i) = h(x_i^T \beta) \tag{13}$$

where h is the inverse of g.
In the Poisson distribution, the link function is the *log* link function, which can be expressed as

$$g(\mu_i) = ln(\mu_i) = \eta_i = x_i^T \beta \tag{14}$$

Given X and β, the mean function is applied to specify the only parameter λ and to generate the response variable Y from the Poisson distribution. In this case, the mean function is written as

$$\mu_i = h(\eta_i) = h(x_i^T \beta) = exp(x_i^T \beta) \tag{15}$$

The simulation process of all variables mentioned above will be explained in detail in the section 2.3 "Implementation in R".

2.2.2 The Size of the Data Set

According to (Roderick J.A. Little (2002)) standard statistical methods are often applied to analyze rectangular data sets, in which $Y = (y_{ij})$ represents an $(n \times p)$ rectangular data set without missing values, and y_{ij} is the value of the jth variable associated with the ith row $y_i = (y_{i1}, ..., y_{ip})$. Generally, the columns of a data matrix represent variables measured for each unit, while the rows of the data matrix represent units, which are also known as observations or subjects depending on the context.

In addition, a data set can be described as a matrix of data which has a dimension of n-by-p, where n is the number of samples observed and p is the number of variables.

In this study 10 different variables are simulated, and exist together in the form of an $(n \times 10)$ matrix. In this case the value of p equals 10. To be more specific, of these 10 variables 7 are continuous and make up a set of numerical data. The remaining variables are categorical variables. For each variable 1,000 observations are simulated; thus here n equals 1,000. Ultimately, the size of one single data matrix is an (1000×10) rectangular data set. In total 1,000 data sets with the same size are simulated randomly in this study, although these data sets are different and irrelevant.

By applying the algorithm mentioned in section 2.1, processing one single (1000×10) matrix of data can produce a set of values. However, it is not convincing or reasonable to compare only one set of values to determine the optimal imputation method. Therefore, in order to improve the stability and validity of the comparison result, it is necessary to simulate multiple data sets under the same circumstance. Thus in this simulation study 1,000 data sets are simulated and 1,000 sets of values are generated, based on which four different missing data imputation methods are compared.

2.2.3 The Missing Rate

Missing data is a common situation and a constant challenge in actuarial statistical analyses. According to a survey by Peng et al. (2006) of 11 quantitative studies in the field of education and psychology, 36% of these studies have no missing data, 48% have missing data, and for about 16% this cannot be determined. Enders (2003) also states that missing data commonly occur in education and psychology studies, whose missing rate usually ranges from 15% to 20%. The missing rate, which indicates the proportion of missing data, has a significant influence on the quality of statistical inferences. This influence tends to vary with different degrees of the missing rate. However, approaches to handling data with different proportions of missing values remain inconsistent. For instance, Schafer (1999) states that a missing rate of 5% or less can be ignored because the missing values would barely affect the results of statistical predictive analyses. Meanwhile, Bennett (2001) asserts that statistical analysis can produce a biased result when the missing rate exceeds 10%. According to Yiran Dong et al. (2013) an acceptable percentage of missing data in a data set has not been established for valid statistical inferences.

Theoretically, if the missing rate is low then the missing data can be ignored because there is no noticeable effect on statistical inferences. Conversely, if the missing rate is relatively high then observed values for the considered variable in the data set are not representative, thus the variable should not be taken into account in the statistical analysis.

However, currently no standardized criteria have been established for the missing rate. If the missing rate is relatively low then the imputed complete data set after the application of imputation methods is relatively similar to the original complete data set. If the missing rate is particularly high then an imputed complete data set can also be generated, but it may vary relatively widely from the original complete data set.

Therefore, this bachelor's thesis conducts a simulation study in order to quantitatively study and analyze the differences between an imputed complete data set after applying imputation methods, and the original complete data set under the circumstances of different missing rates.

In this thesis 1,000 original complete data matrices are simulated by applying R. In other words, there are no missing values in these 1,000 original complete data matrices. Because the purpose of this study is to compare the benefits and disadvantages of various imputation methods and determine the most appropriate approach under different circumstances, it is necessary to generate a number of missing values. These should be based on three different missing data mechanisms and should also be conducted with a proper missing rate.

Two typical examples of the significance of the missing rate in the field of life and health reinsurance are now introduced.

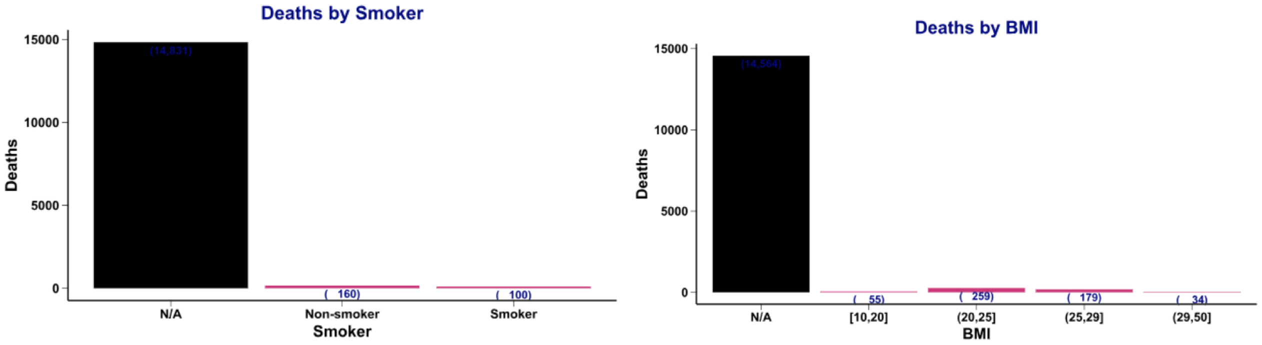


Figure 2.3: The number of missing values in mortality analysis

As shown in Figure 2.3, there are two common risk factors in mortality analysis, namely smoker status and body mass index. These two risk factors exhibit a large number of missing values in the collected data, which are denoted by "N/A", a common abbreviation for the phrase "not available" or "no answer." In this circumstance, the proper distributions of these two variables cannot be estimated based on the observed values. Hence, it is inappropriate to apply imputation methods to impute these missing values.

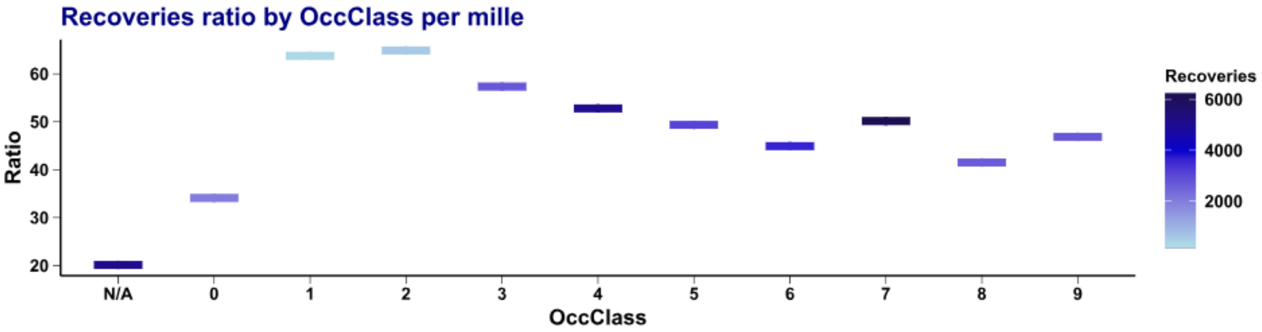


Figure 2.4: Distribution of recoveries ratio according to occupation class in disability analysis

Figure 2.4 presents the distribution of recoveries ratio by occupation class in a disability analysis, which indicates that the missing rate is around 2%. From my point of view, given this missing rate it is appropriate to apply imputation methods to impute the missing values.

2.3 Implementation in R

This section explains how the original complete data set is generated using R. The original complete data set is composed of 10 independent variables and one response variable. The 10 independent variables consist of 7 continuous variables and 3 categorical variables. In order to produce multivariate normally distributed continuous random variables with the help of R, the function *mvrnorm* from the *MASS* package is applied. This function has three necessary arguments, namely the sample size n, the mean vector (μ), and a square covariance matrix (Σ), which should all be specified in advance. As mentioned in section 2.2.1, these continuous variables should be correlated, thus a random correlation matrix is generated by specifying Σ. The *corrplot* package is applied to graphically display a correlation matrix, which indicates correlation coefficients among the continuous variables. The generated matrix is shown in Figure 2.5.

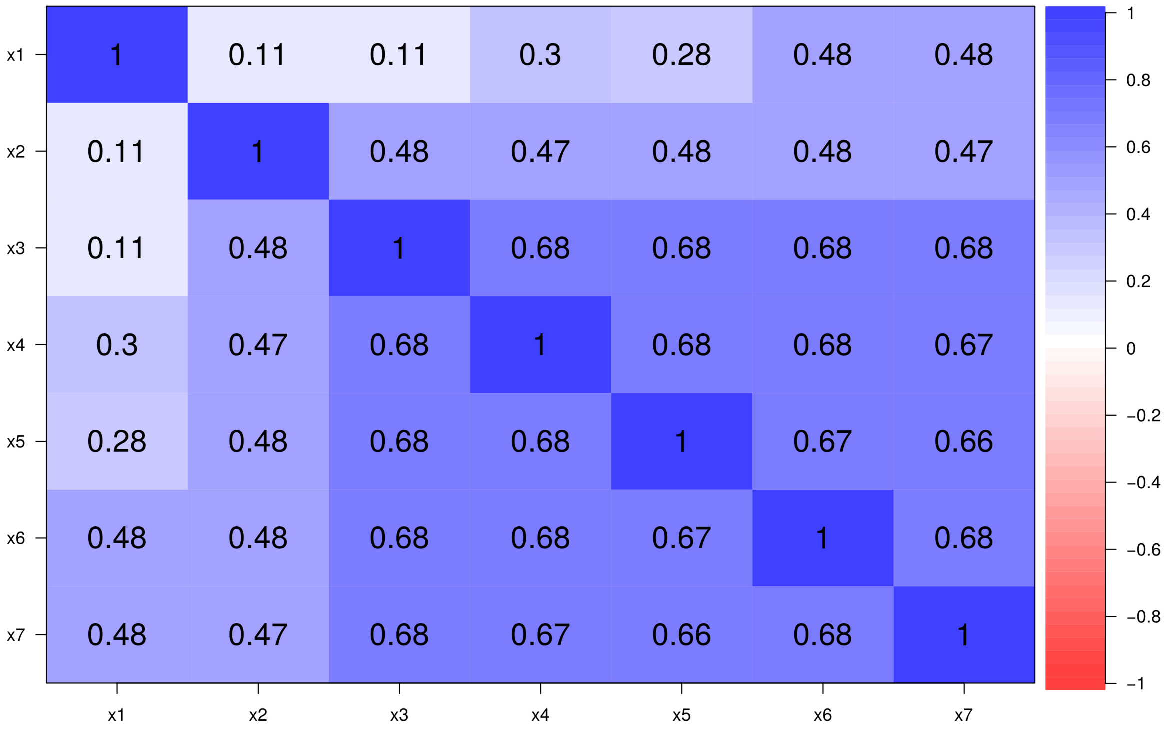


Figure 2.5: Correlations among independent continuous variables in one data set

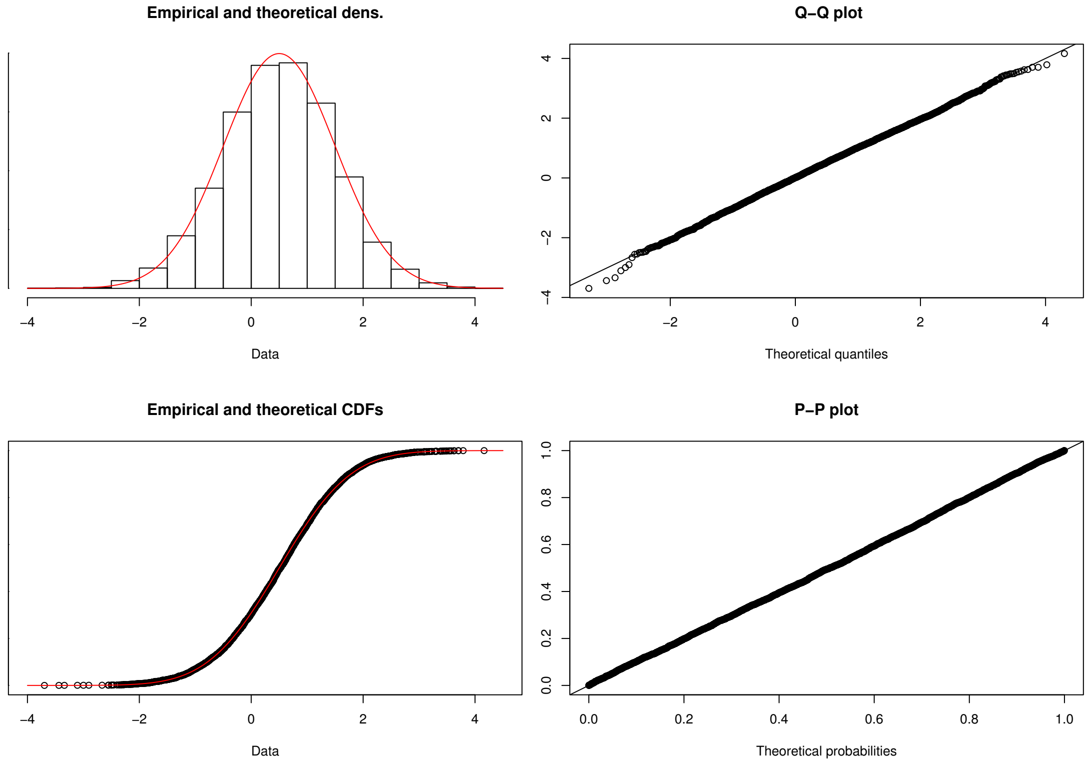


Figure 2.6: Comparison between empirical and theoretical distributions

As mentioned above, seven continuous variables with multivariate normal distributions are simulated. The *fitdist* function from the *fitdistrplus* package is used, which enables the fit of a parametric univariate distribution to non-censored or censored data by the maximum likelihood method. A quantil-quantil-Plot (Q-Q plot) is shown in Figure 2.6 and compares two probability distributions, namely the theoretical distribution and the empirical distribution. The points in the Q-Q plot represent the distribution of the simulated data. The linearity of the points suggests that the data fit a normal distribution. In conclusion, each continuous variable is univariate normally distributed, implying that the seven continuous variables comprise a multivariate normal distribution.

Table 2.1: Construction of all independent variables

Var	Distribution	detail
X= $[x_1,x_2,...,x_7]^T$	Multivariate normal distribution	$\mathbf{X} \sim N_7(\mu,\Sigma)$ $\mu = \mathrm{E}[\mathbf{X}] = [\mathrm{E}[X_1],\mathrm{E}[X_2],\ldots,\mathrm{E}[X_7]]^{\mathrm{T}}$ $\Sigma =: \mathrm{E}[(\mathbf{X}-\mu)(\mathbf{X}-\mu)^{\mathrm{T}}]$
Gender	Binomial distribution	$f(Gender) = \begin{cases} 0.3, & \text{if } Gender = \text{Male}, \\ 0.7, & \text{if } Gender = \text{Female}. \end{cases}$
Occu	Multinomial distribution	$f(Occu) = \begin{cases} 0.1, & \text{if } Occu = \text{A}, \\ 0.2, & \text{if } Occu = \text{B}, \\ 0.65, & \text{if } Occu = \text{C}, \\ 0.05, & \text{if } Occu = \text{D}. \end{cases}$
Smoker	Binomial distribution	$f(Smoker) = \begin{cases} 0.4, & \text{if } Smoker = \text{Yes}, \\ 0.6, & \text{if } Smoker = \text{No}. \end{cases}$

In addition to the continuous variables mentioned above, three categorical variables are also simulated, namely gender, occupation class, and smoker status. Gender and smoker status fit binomial distributions, while occupation class fits the multinomial distribution. Table 2.1 presents the constructions of all independent variables.

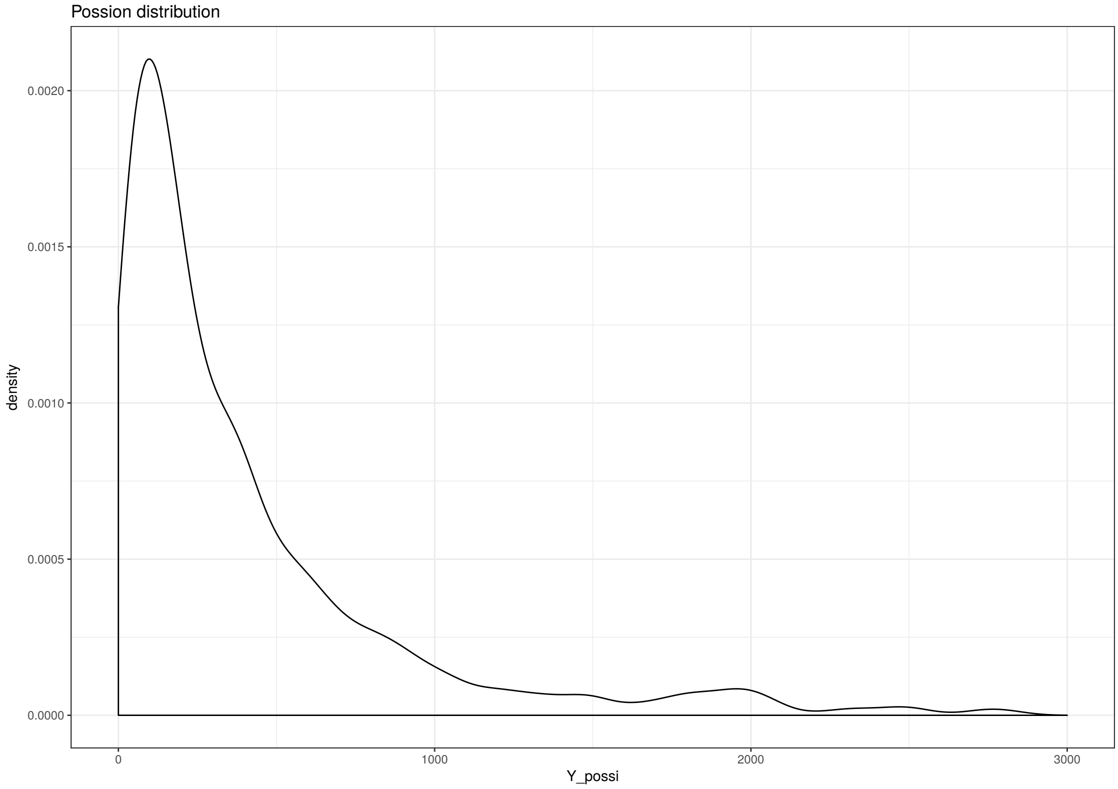


Figure 2.7: Graphical representation of the distribution of the response variable

This simulation study is based on the Poisson regression model, which implies that the response variable Y fits the Poisson distribution. Given all simulated independent variables, the regression coefficients β should be determined in order to generate a Poisson distributed response variable. The coefficients β are defined as follows.

$$\begin{aligned}\beta &= (\beta_0,\beta_1,\beta_2,\beta_3,\beta_4,\beta_5,\beta_6,\beta_7,\beta_8,\beta_9,\beta_{10})^T \\ &= (0.1,2,2,2,0,0,0,0,0,0,1)^T\end{aligned}$$

As mentioned in section 2.2.1, given the independent variables and the regression coefficients, the response variable Y is generated by the use of the log link function. The corresponding log link function used in this study is expressed as follows.

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_{1000} \end{pmatrix} = \begin{pmatrix} x_{(1,1)} & x_{(1,2)} & \cdots & x_{(1,10)} \\ x_{(2,1)} & x_{(2,2)} & \cdots & x_{(2,10)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{(1000,1)} & x_{(1000,2)} & \cdots & x_{(1000,10)} \end{pmatrix} \times (\beta_0, \beta_1, ..., \beta_{10})^T \tag{16}$$

In the next step, based on particular properties of the Poisson distribution $\lambda = \mathrm{E}(X) = \mathrm{Var}(X)$, the *rpois* function from the *stats* package is used, which generates multi-Poisson random variables based on an Aitchison composition. Figure 2.7 visualizes the distribution of the response variable Y, implying that Y fits the Poisson distribution.

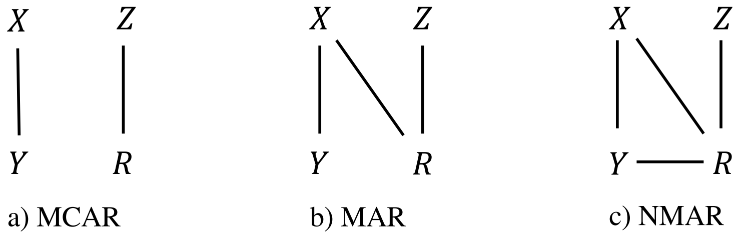
3 Missing data

3.1 Three Types of Missing Data Mechanism

Three main factors determine the risk of bias due to missing data, namely the proportion of missing data, the reasons why data are missing, and the type of missing data mechanism, which is of greatest importance. The extent to which the missing data bias statistical results is dependent on the type of missing data mechanism. For example, if the missing data are MCAR, the data sample can still be considered as representative of the population because the joint distribution is the same for the complete data and the observed subset. Alternatively, if values are missing in a systematic way then the observed data cannot represent the population. For instance, consider an example where researchers are carrying out a study to analyze the relationship between education level and income level, with the assumption that individuals whose education level is relatively lower are likely not to answer the question "What is your salary?" In this case, if those data that are MAR are not taken into consideration then the analysis is prone to a wrong conclusion regarding the relationship between education level and income level. Accordingly, it is vital to understand missing data mechanisms when comparing different imputation methods.
The relationship between missing variables and the underlying values of variables in the data set is based on the corresponding missing data mechanism. In 1976, Little and Rubin proposed a theoretical framework which led to the generally accepted classification method used today (Roderick J. A. Little and Donald B. Rubin(2002)). Figure 3.1 further explains the differences between the three missing data mechanisms, namely MCAR, MAR, and NMAR (Schafer & Graham (2002)), where X represents variables that are completely observed, Y represents a variable that is partly missing, Z represents the element of the causes of missingness unrelated to both X and Y, and R represents the missingness. Figure 3.1 a) explains the MCAR mechanism, which implies that there is no relationship between the missing data mechanism and the values of any variable in the data set, whether missing or observed. The second mechanism, shown in Figure 3.1 b), is MAR, and indicates that there is a systematic relationship between the tendency of missing values and the observed data instead of the missing data. Figure 3.1 c) presents NMAR data (nonignorable nonresponse), which are neither MAR nor MCAR data (Polit, D.F. and Beck,

C.T. (2012)). In other words, if the missing data are non-random and are dependent on the missing variables, then they are classified as NMAR.

Schafer & Graham (2002)



a) MCAR b) MAR c) NMAR

Figure 3.1: Graphical representations of a) missing completely at random (MCAR), b) missing at random (MAR), and c) not missing at random (NMAR).

In the following sections, descriptions of different types of missing data mechanisms and their corresponding consequences are explained in detail using mathematical notations. In addition, the simulation processes of these mechanisms are also introduced.

3.1.1 Missing Completely at Random

If the events that lead to any specific data item being missing are independent not only of observable variables but also of unobservable parameters of interest, and if they occur completely at random, then the corresponding missing values in a data set are **MCAR** (Polit, D.F. and Beck, C.T.). An example of this type of missing data is an accident whereby researchers carelessly lose a few questionnaires when studying risk factors for high blood pressure. In this case, it is not possible to assume that the missing questionnaires (i.e. missing values) are related to the value of blood pressure or to other variables, thus the missing data can be considered as a random subset of the data.

Of the three missing data mechanisms, MCAR is the only type that can be tested for. As mentioned above, the joint distribution is the same for the complete data set and the observed subset, which is why MCAR is not a problematic missing data mechanism. Therefore, in this case there is no need to make adjustments for missing data because by using the observed data or the whole data set, the approximate results would be reached. This kind of handling method, of simply using the observed data, is called complete case analysis. When MCAR data occurs, these missing data can be ignored and it is not necessary to include the modeling of the missing data mechanism in the estimation process. However, MCAR data is often an ideal situation which is unlikely to occur in reality.

Several notations and terms are used to further explain the differences between MCAR and other missing data mechanisms. If missingness is unrelated to the values of the data, whether missing or observed, this means that the data are MCAR and can be denoted mathematically as:

$$f(M|Y,\phi) = f(M|\phi)\ for\ all\ Y,\ \phi$$

where $Y = (y_{ij})$ is defined as the complete data as mentioned in the previous section, $M = (M_{ij})$ stands for the missing-data indicator matrix, and ϕ is the unknown parameter.

3.1.2 Missing at Random

In contrast to MCAR data, the **MAR** mechanism occurs when the missingness is not completely random, and can be explained by at least one other variable with complete information. In this case, the missingness probability is related to some of the observed data instead of the missing data itself. This type of missing data mechanism occurs more often in reality, but unlike MCAR it cannot be tested. For the MAR mechanism the distribution of the observed data and the complete data are generally not the same. Hence the observed data cannot be applied for analysis, or biased estimates would occur.

For example, if questionnaire respondents with a higher education level are more likely to report their income than those who have a relatively lower education level, then it is reasonable to consider that a missing income level value can be attributed to the MAR mechanism. In this circumstance the education level is completely observed, which implies that this variable has complete information.

As in the previous section, the observed components Y_{obs} and the missing components Y_{mis} are defined and the missing data mechanism MAR can be expressed as follows:

$$f(M|Y,\phi) = f(Y_{mis}|\phi)\ for\ all\ Y_{mis},\ \phi$$

While MAR is less restrictive than MCAR, it still depends on the values of other variables. Both of the two mechanisms described above are random missing data mechanisms.

3.1.3 Not Missing at Random

The last type of missing data mechanism is **NMAR**. When the missing data are NMAR, the missingness has an exclusive relationship with the missing data. In other words, the missingness probability is allowed to be dependent on the missing values themselves. To further explain this, the example mentioned in section 3.1.2 can be applied again. As described, the missing income level values, which can be considered as MAR data, are related to education level. However, regarding the assumption that those respondents with a higher income level are more likely to report their income level than those with a relatively lower income level, this can be classified as NMAR instead of MAR, because the missing income level values are not related to other variables which have complete information, but rather depend on the missing values themselves.
In addition, several terms and notations are applied to distinguish NMAR from other types of missing data mechanisms:

$$f(M|Y,\phi) = f(Y_{obs}|\phi)\ for\ all\ Y_{mis},\ \phi$$

If data are NMAR then this missing data mechanism cannot be ignored, as this mechanism must be modeled as part of the estimation process. However, it is not easy to determine the optimal modeling method because the observed data do not contain information on this mechanism. Unlike MCAR and MAR, NMAR is not a random missing data mechanism.

3.2 Implementation in R

The three missing data mechanisms discussed above can be simulated by applying R. This section illustrates the R code used in this simulation study. The R code is written based on the theoretical differences identified between the three missing data mechanisms that are explained in the previous section.

As mentioned in section 2.2.2, there are 10 different variables, and 7 of them are continuous variables, which are $x_1, x_2, ..., x_7$. The rest of the variables are categorical variables, including gender (*Gender*), occupation class (*Occu*), and smoker status (*Smoker*). The following R-code simulates three different missing data mechanisms, regarding two types of variables.

MCAR

```
# for variable x1
set.seed(111)
x1.miss.tag <- rbinom(1000,1,0.5)
Data.MCAR$x1[x1.miss.tag == 1] <- NA

# for variable Gender
set.seed(888)
Gender.miss.tag <- rbinom(1000,1,0.5)
Data.MCAR$Gender[Gender.miss.tag == 1] <- NA
```

The variable x_1 has a normal distribution. Since the missingness is independent both of observed variables and of unobserved variables, MCAR is completely random. Therefore, an object (*miss.tag*) should be defined by using the function *rbinom* from the package *stats*. This function generates the required number of random values of given probability from a specified sample. The simulation process of the variable *Gender* is identical to that of x_1.

MAR

```
# for variable x1
set.seed(1111)
x1.miss.tag.MAR <- rbinom(1000,1,0.7)
Data.MAR$x1[Data.MAR$Y_possi <= 530
            & x1.miss.tag.MAR==1] <- NA

# for variable Gender
set.seed(108)
Gender.miss.tag.MAR <- rbinom(1000,1,0.55)
Data.MAR$Gender[Data.MAR$Y_possi <= 1000
                & Gender.miss.tag.MAR ==1 ] <- NA
```

MAR occurs when the missingness can be accounted for by one or more other variables with complete information. In the simulated data set, the response variable Y (*Y_possi*) does not have missing values, and missing values frequently occur in the variable x_1 when the value of the response variable equals to or is less than 530 in this case. The simulation process of the variable *Gender* is also same as that of x_1.

NMAR

```
# for variable x1
set.seed(101)
x1.miss.tag.NMAR <- rbinom(1000,1,0.8)
Data.NMAR$x1[Data.NMAR$x1 <= 0.8
             & x1.miss.tag.NMAR ==1 ] <- NA

# for variable Gender
set.seed(108)
Gender.miss.tag.NMAR <- rbinom(1000,1,0.7)
Data.NMAR$Gender[Data.NMAR$Gender == "Female"
                 & Gender.miss.tag.NMAR ==1 ] <- NA
```

When data are NMAR, this missing data mechanism is neither MCAR nor MAR because the tendency of a value to be missing is related to its values. Regarding the continuous variable x_1, missing values occur more frequently in the case that the value itself is equal to or less than 0.8. The R code shown above simulates a data set with a missing rate of around 50%. The missing rate can be adjusted by altering the argument (*prob*) in the function *rbinom*, thus the purpose of simulating different missing rates can be achieved.

4 Variable Selection

In statistics stepwise regression is applied to fit regression models, whereby predictive variables are chosen by an automatic procedure (Efroymson,M. A. (1960)). Before further describing this method as it is used in this thesis, the purpose and necessity of adopting variable selection is explained in detail.

Through variable selection, the "best" subset of variables or predictors are selected. Variables should be selected in studies for three reasons, which are listed and explained as follows.

1. If there are a large number of predictor variables in the multiple regression model and there are certain correlations between these variables, then they cannot independently predict the dependent variable. In other words, too many predictor variables would predict the dependent variable at the same time. However, this simply cannot be accomplished due to the linear relationships existing between these variables. Under this circumstance, these redundant variables can lead to multicollinearity, which in multiple regression models is a phenomenon in which one predictor variable can be linearly predicted from the others with a high degree of accuracy.

2. It is known that the more predictor variables there are in a regression model, the more information they can represent. Nevertheless, unnecessary and thus redundant predictors add noise to the estimation of other important quantities. Moreover, degrees of freedom are wasted. According to (Julian J. Faraway(2009)) a smaller model may generate more precise estimates and predictors.

3. Variable selection should be considered during algorithm design, especially for larger or relatively more complex algorithms, which usually require more computing time. Therefore, in order to reduce the required computing time it is necessary to conduct variable selection. Firstly, it can find the most important variables and keep them in the regression model. Secondly, it can also identify the comparatively less important variables and remove them from the model in order to achieve the goal of reducing the calculation time as much as possible.

4.1 Best Subsets Regression

In this section, the best subsets regression and stepwise regression are discussed in detail. The reasons for applying stepwise regression instead of the best subsets regression are then explained.
The best subsets regression, which is also known as the "all possible model," is an automatic process that can be applied to assist in choosing from a large number of independent variables. The best subsets regression procedure considers all possible combinations of independent variables and fits all possible models based on these remaining independent variables after conducting variable selection. For example, if there are 10 independent variables in the regression model, then it fits 1,024 models. In other words, if there are p independent variables in the model, the best subsets regression takes each variable into consideration and determines whether or not these variables can remain in the model. Accordingly, in total there are 2^p possible models (Patrick Royston, Willi Sauerbrei(2008)).
The results of comparisons between all possible models indicate that the best subsets regression is the optimal fitting model with one independent variable, two independent variables, three independent variables, and so on. Therefore, the best subsets regression is considered advantageous as it can present different sizes of fitted models with one variable up to the full model. The subset of predictors that performs best can be determined after a certain criterion is met, which is either the adjusted R^2 or Mallows' C_p. The Mallows' C_p for selecting P regressors from a set of $K > P$ is defined as:

$$C_p = \frac{SSE_p}{S^2} - N + 2P$$

where:

- $SSE_p = \sum_{i=1}^{N} (Y_i - Y_{pi})^2$ is the error sum of squares for the model,
- Y_{pi} is the predicted value of Y,
- S^2 is the residual mean square,
- N is the sample size.

The adjusted R^2 is defined as:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- R^2 is the coefficient of determination,
- p is the total number of explanatory variables in the model ,
- n is the sample size,

4.2 Stepwise Regression

Besides the best subsets regression, stepwise regression is another option for variable selection. Compared with the best subsets regression, the stepwise regression procedure automatically selects a model by adding or removing predictor variables step by step. Whether to add or remove variables depends on their corresponding statistical significance, which implies that the most statistically significant variables would be added and the least significant variable in the model would be removed. In this case, a single regression model is eventually produced instead of many possible combinations of independent variables.

Stepwise Regression

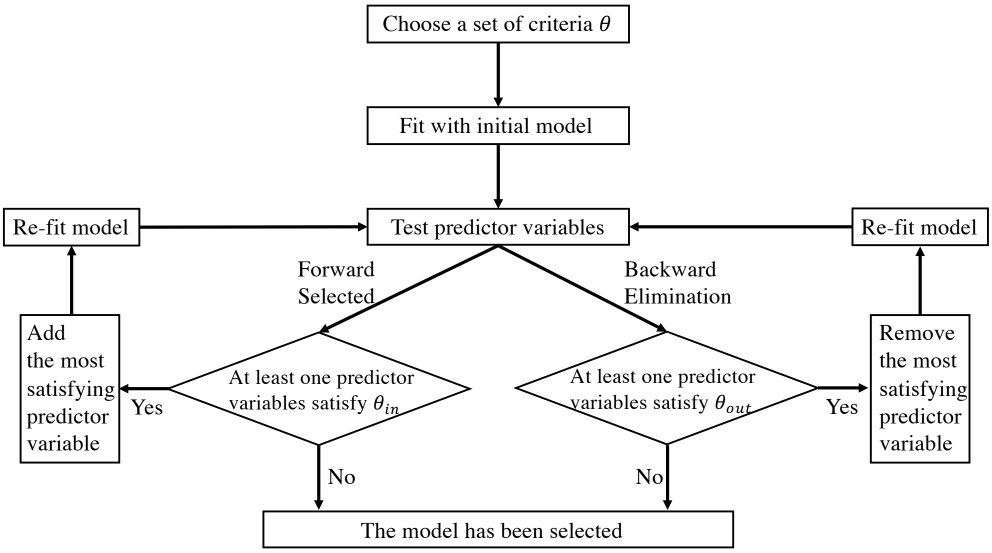

Choose a set of criteria θ
Fit with initial model
Re-fit model
Test predictor variables
Re-fit model
Forward Selected
Backward Elimination
Add the most satisfying predictor variable
Yes
At least one predictor variables satisfy θ_{in}
At least one predictor variables satisfy θ_{out}
Yes
Remove the most satisfying predictor variable
No
No
The model has been selected

Figure 4.1: A schematic diagram of stepwise regression.

The flowchart shown in Figure 4.1 explains the process of stepwise regression and shows the two main approaches that it applies, namely forward selection and backward elimination. In short, stepwise regression generally consists of two steps:

Step 1:
From a set of criteria, a specific criterion should be chosen to determine whether predictor variables should be added or removed. With this set of criteria, all of the possible models can be fitted and the best regression model can be chosen. Possible criteria are the Bayes information criterion (BIC), the Akaike information criterion (AIC), cross-validation (CV), and Mallows' C_p. In practice, AIC and BIC are the most frequently used methods. In general (Akaike, H. (1974)) (Wit, Ernst(2012)):

$$AIC = 2k - 2ln(\hat{L})$$

while

$$BIC = ln(n)k - 2ln(\hat{L})$$

where:

- $\hat{L}$ is the maximized value of the likelihood function of the model,
- n is the number of observations or the sample size,
- k is the number of parameters estimated by the model.

By comparing the formulae of the two criteria, it can be seen that the formula of the BIC is similar to that of the AIC, only with a different penalty for the number of parameters. To be more specific, in the AIC the penalty is $2k$, while in the BIC the penalty is $ln(n)k$. Hence, it is important to choose a fixed criterion as the principle for model selection. A comparison between AIC and BIC is conducted by Burnham and Anderson (Burnham & Anderson (2004)), according to which the AIC can be derived in the same Bayesian framework as the BIC simply by using different prior probabilities. In the Bayesian derivation of the BIC each candidate model has a prior probability of $1/R$ (where R is the number of candidate models), which, however, should be a decreasing function of k. Therefore, such a derivation is "not sensible." In addition, the abovementioned authors also demonstrate a number of simulation studies indicating that in practice the AIC tends to be more advantageous than the BIC. For this reason the AIC is used in this paper instead of the BIC.

Step 2:

Many available methods can be chosen to fit the most appropriate regression model, such as forward selection, backward elimination, block-wise selection, and so on. From these options this thesis focuses on forward selection and backward elimination, which are both considered as statistical regression methods.
Notably, forward selection begins with no predictor variables; these are added step by step following the order of correlation with the response variable, from the highest to the lowest. When none of the remaining predictor variables are significant, the procedure stops to add a new predictor variable into the regression model, which means that the selected model is determined. In contrast, backward elimination is the reverse process of forward selection, as it begins with all predictor variables in the regression model. These are removed step by step according to their significance level. The predictor variable with the lowest significance level is supposed to be removed first. If no insignificant predictor variables remain in the regression model then backward elimination stops; this is the difference between the forward selection and backward elimination procedures. The number of predictor variables should be considered as one of the main determining factors when choosing between forward selection and backward elimination.
For instance, when a large number of variables are present in the model the forward selection method is recommended rather than the backward elimination method, because in the latter case the model would initially include all predictor variables, which could lead to the problem that unnecessary variables may also be included. However, the number of predictor variables in this simulation study is not large, and after a series of tests it is determined that the same selected model can be acquired using both forward selection and backward elimination. In conclusion, both methods are suitable for use in this simulation study.

As mentioned above, the best subsets regression and stepwise regression are both possible alternatives for variable selection. However, only stepwise regression is used in this simulation study due to the following:

- Stepwise regression returns a single best selected model constructed using the p-values of the predictor variables. In contrast, the best subsets regression assesses all possible models and presents different sizes of fitted models along with some criteria. In this bachelor's thesis the PRESS statistic is calculated based on the selected model. Consequently, the use of stepwise regression is more suitable.

- Furthermore, stepwise regression is faster than other automatic model selection methods, which is advantageous as it reduces the required computing time.

5 Methodology

In general, there are two approaches to handling problems related to missing data. The first option is to simply omit units with missing data, and is known as complete case analysis. A second option is to infill missing values, and is called imputation. These methods maintain the complete sample size, which is considered beneficial for reducing biases and increasing precision when appropriate methods are applied. However, imputation also has drawbacks. According to Dempster and Rubin (1983):

> *"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases."*
>
> – Dempster and Rubin (1983)

Imputation is the procedure of using substituted values to replace missing data, for which a predictive distribution is created based on the observed data. Generally, two types of methods are used to generate this distribution: single imputation methods and multiple imputation methods.

5.1 Single Imputation Methods

A single imputation method implies the use of a single estimate to impute a missing value, for which a variety of approaches can be applied. These include mean substitution, hot deck imputation, and cold deck imputation. Considering its conceptual simplicity and simple operation, single imputation is widely applied. Compared to listwise deletion, single imputation methods can maintain the same number of observations as the original complete data set.

However, this type of imputation method also has its disadvantages. If the missing data are not classified as MCAR then biased parameter estimates are likely to be produced by a single imputation method, for example means, correlations, and regression coefficients. It is possible that the imputed values produced using a single imputation method would probably be more biased than values produced by listwise deletion.

In this thesis mean substitution is the applied single imputation method, which is not recommended in practice. Thus, the emphasis of applying this approach is not placed on its imputation effectiveness. Rather, mean substitution acts as a measurement criterion used to study the upper limit of the missing rate, with which any imputation method is not recommended for application because the imputed complete data set is almost entirely unrepresentative of the characteristics of the original complete data set.

Mean substitution

Mean substitution or mean imputation is the most straightforward method to impute, whereby each missing value is replaced with the mean of the observed values for this variable. This method is widely used in questionnaire manuals. The greatest benefit of this method is that it does not reduce the complete sample size but does lead to the reduction of variability in the data, which implies that the standard deviations and variance estimates are likely to be underestimated. However, restricting the variability also decreases the significance of the covariances and correlation. Biased estimates are often produced using this method, regardless of the underlying missing data mechanism (Enders, 2010; Eekhout et al, (2013)).

In general, there are two types of mean imputation, namely item-mean imputation and person-mean imputation. By applying person-mean imputation, the mean of an individual's total completed items is substituted for those items with missing values, to a certain degree. Meanwhile, item-mean imputation substitutes the mean response of the entire sample that responded to the item. In this simulation study, item-mean imputation is applied to impute missing values.

5.2 Multiple Imputation Methods

According to Royston (2004) an appropriate imputation method should be able "to inject the correct degree of randomness into the imputations and to incorporate that uncertainty when computing standard errors and confidence intervals for parameters of interest." Traditional single imputation methods such as mean substitution cannot yet fulfill these criteria, for two main reasons. First, they do not take into consideration the randomness of values, which is based on specific distributions. Second, standard errors are not considered. Although single imputation methods can technically be applied to impute all missing data, this would distort the true distribution of variables.
Unlike single imputation methods, multiple imputation (MI) methods can fulfill the criteria mentioned above and have a relatively wider application in practice. With the use of MI methods, instead of replacing each missing value in a data set with only one randomly imputed value, which does not reflect the uncertainty relating to the imputation model, each missing value is replaced with several imputed values. When model-based imputation is applied it can reflect both to what extent the imputed values vary from the observed values, which is also called the sampling variability, and the uncertainty relating to the regression coefficients existing in the model. To do so, MI creates more than one imputed value for each missing value. The created values are predicted from a regression model that is different to a small degree, which can reflect sampling variability.
Created by Rubin in 1987, the procedure for conducting multiple imputation for missing data is introduced as follows. This method generally consists of six steps. First, an appropriate regression model that incorporates random variation should be built. Second, the first step should be repeated several times. Third, a standard and complete MI method should be applied to conduct the analysis on each data set. In order to acquire a single point estimate, the next step is to average the values of the parameter estimates across the missing value samples. Subsequently, the standard errors should be obtained by averaging the squared standard errors of the missing value estimates, and the variance of the missing value parameter across the samples should be calculated. Finally, the two quantities in MI for missing data should be combined to calculate the standard errors. Before performing MI for missing data, certain conditions should be satisfied. The first condition is that the data should be MAR or MCAR, which im-

plies that the missingness probability is related to some of the observed data with complete information instead of the missing data themselves. The second condition is that the model should be appropriate and that it should match other models. However, in reality these two conditions tend not to be entirely satisfied. For example, the missing data mechanism is unlikely to be classified as a certain type of mechanism in practical data sets. Therefore, this simulation study performs MI methods under the circumstance of different missing data mechanisms. Three powerful R-packages that can help to realize MI methods are explained in the following sections.

5.2.1 With MICE Package

Introduction

In contrast to single imputation, MI considers statistical uncertainty when imputing missing values. One of the three powerful R-packages that handle missing data is **multivariate imputation by chained equations (MICE)**, also called "sequential regression multiple imputation" or "fully conditional specification." This is one of the most important methods used to address and impute missing data. In consideration of the flexibility of chained equations, MICE can handle various types of variables in the data set, such as continuous variables, categorical variables, and mixed-type variables. If the distribution of each variable in the data set is already established, this method is more applicable. For example, if a variable fits the normal distribution then specific approaches can be defined in advance to impute the missing values of this variable by using the MICE function. Even if no appropriate multivariate distribution can be found, MICE remains an applicable option; this implies that MICE is suitable for data sets composed of mixed-type data. In conclusion, for the application of MICE the specific distribution of each variable in the data set should be defined in advance, which is based on a univariate distribution. The R-package MICE uses the FCS algorithm, which imputes each variable with missing values in the data set by conducting several repetitions.

Assumption

Regarding the application of MICE, two assumptions should be taken into consideration. The first assumption is that the missing data mechanism is MAR, which implies that the missingness probability is not related to the missing data but is related to some of the observed data (Schafer & Graham (2002)). If data are not MAR, biased results are likely to be obtained when applying MICE. However, in order to compare the performance of MICE under the circumstances of different missing data mechanisms, MICE is also implemented in the cases of MCAR and NMAR data.

The second assumption concerns the size of the data set. In practice, data sets tend to be large in size, which implies that they include thousands of observations and hundreds of variables (He et al.(2009); Stuart et al.(2009)). Furthermore, in these large data sets a high variety of variables often exists. Based on the large size of data sets, a large joint model for all of the various types of variables should be fitted. With the help of the flexibility of MICE a series of regression models is run for each variable with missing data, which are based on the distribution of each variable. For the purposes of operability and objective comparison between different imputation methods, the data set in this simulation study is not large.

Algorithm

Generally, implementing MICE involves five basic steps. First, each variable with missing values in the data set is substituted using single imputation methods such as mean substitution, after which the imputations can be considered as temporarily occupying the missing place.

Second, the substituted values are set back to missing while the observed values of other variables remain the same. These substituted values should be imputed using a new estimated regression model.

In the third step the observed values of the target variable, the missing values of which should be imputed, are considered as the response variable in a new estimated linear regression model, in which all of the other variables are independent variables. Since several variables in a data set may have missing values, a series of linear regression models is generated; these models are conducted under the same assumption.

Fourth, the predictions are obtained from the regression model mentioned in the last step and are applied to replace the missing values. Following this replacement, the predictions and the observed values of this variable are considered as independent variables in the estimation of the subsequent regression model, in which the next variable with missing values to be imputed becomes the dependent variable.

In the last step the abovementioned stages are repeated for each variable with missing values in the data set, like a cycle. Once each missing value is replaced with predictions from regression models, the data set is complete and without missing values.

One cycle can generate only one imputed complete data set. The number of cycles determines the number of imputed data sets that can be produced, which can be defined by the user. In this simulation study the cycle is repeated five times, generating five imputed complete data sets with the same size. The following paragraph further explains how this algorithm is applied in the simulation study.

Implementation in R

Two types of data sets are used in this simulation study: one consists of seven continuous variables, and the other consists of seven continuous variables and three categorical variables. With the help of the $aggr()$ function from the VIM package, the missingness pattern is visualized as shown in Figure 5.1. The two figures therein present the missing rate of each variable and the frequencies of the combination of missing variables in the data set with mixed-type variables.

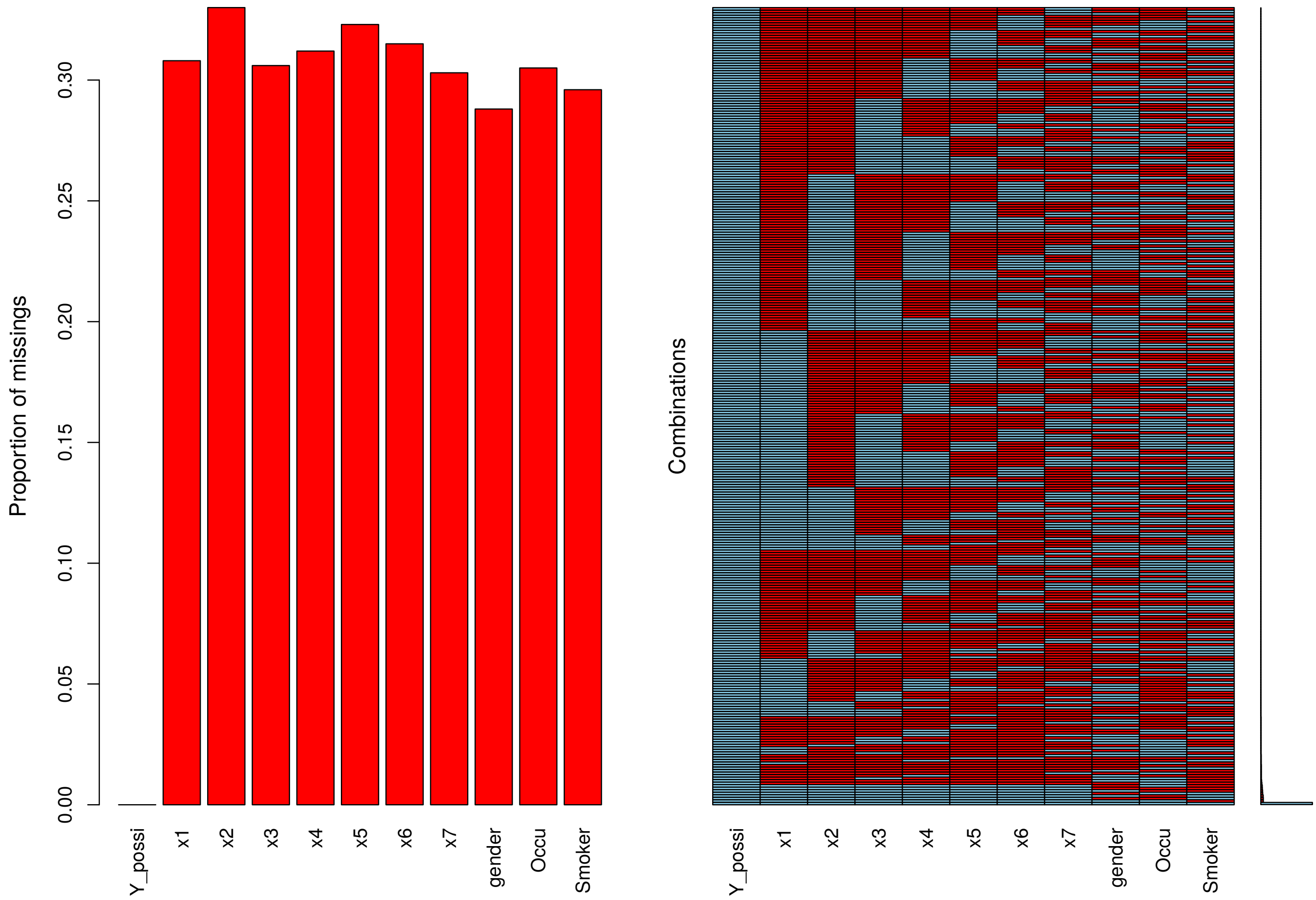


Figure 5.1: Missing pattern when the data are MAR and the missing rate is 30% in the data set with mixed-type variables

The left-hand figure shows that the variable (Y_{possi}) does not have missing values, whereas the other variables all have missing values with missing rates of around 30%. This is one of the simulated cases. Due to the number of variables and different missing situations of these variables the combination of the variables is complex, as can be seen in the right-hand figure.

Figure 5.2, proposed by (Van Buuren and Groothuis-Oudshoorn,(2011)), illustrates the application of MICE in R.

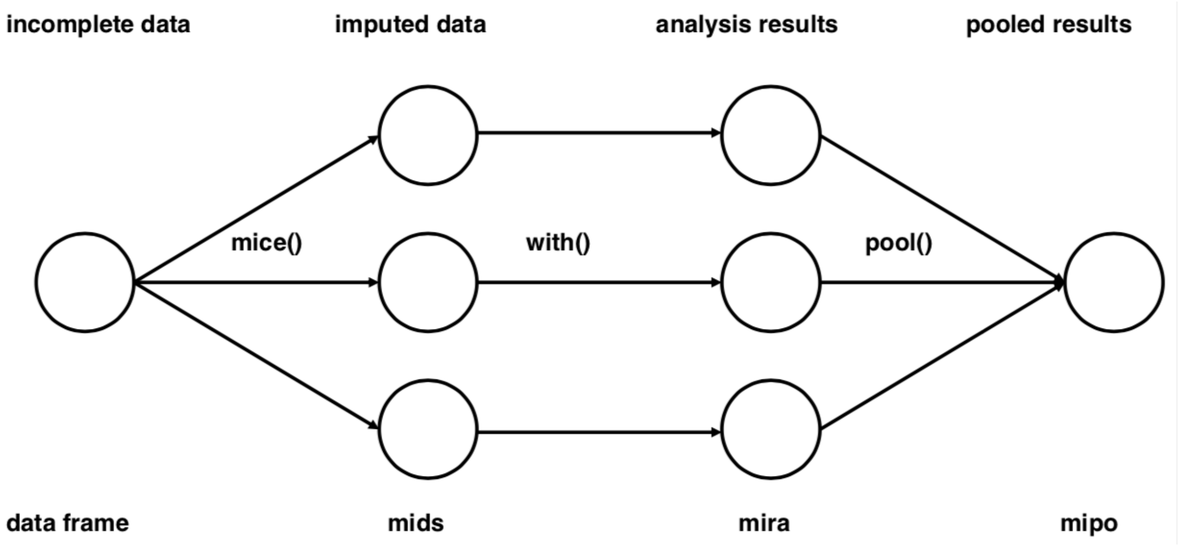


Figure 5.2: Graphic demonstration of the main steps of MICE

The *mice*() function imputes each missing value by using the algorithm described above. There is an argument *m*, the number of multiple imputations, with a used default value of $m = 5$, which indicates that five imputed complete data sets with the same size are generated. The with() function produces an analysis report regarding each individual data set. All of the results are combined with the pool() function based on Rubin's rule (Rubin, 1987).

Before applying the *mice*() function, appropriate methods should be defined for each variable based on the types of variables and the corresponding distributions. For this purpose, several possible methods can be applied as shown in Table 5.1.

Table 5.1: List of univariate imputation methods

Method	Name of variable	Type of variable	Type of regression
pmm	x1,...,x7	any	Predictive mean matching
norm	x1,...,x7	numeric	Bayesian linear regression
logreg	gender,Smoker	binary	Logistic regression
polyreg	Occu	unordered	Polytomous logistic regression

In the data set the two variables gender (*gender*) and smoker status (*Smoker*) have the binomial distribution, based on which the logistic regression model is used and the *logreg* method is applied to the *mice*() function. Another categorical variable, occupation class (*Occu*), has the multinomial distribution. On this basis the multinomial logistic regression model is used and the *polyreg* method is the most appropriate.

The seven continuous variables have the multivariate normal distribution, for which two possible methods can be used, namely *pmm* and *norm*. In order to choose the better method, the following figures compare the imputation effects between the two methods by using the *densityplot*() function.

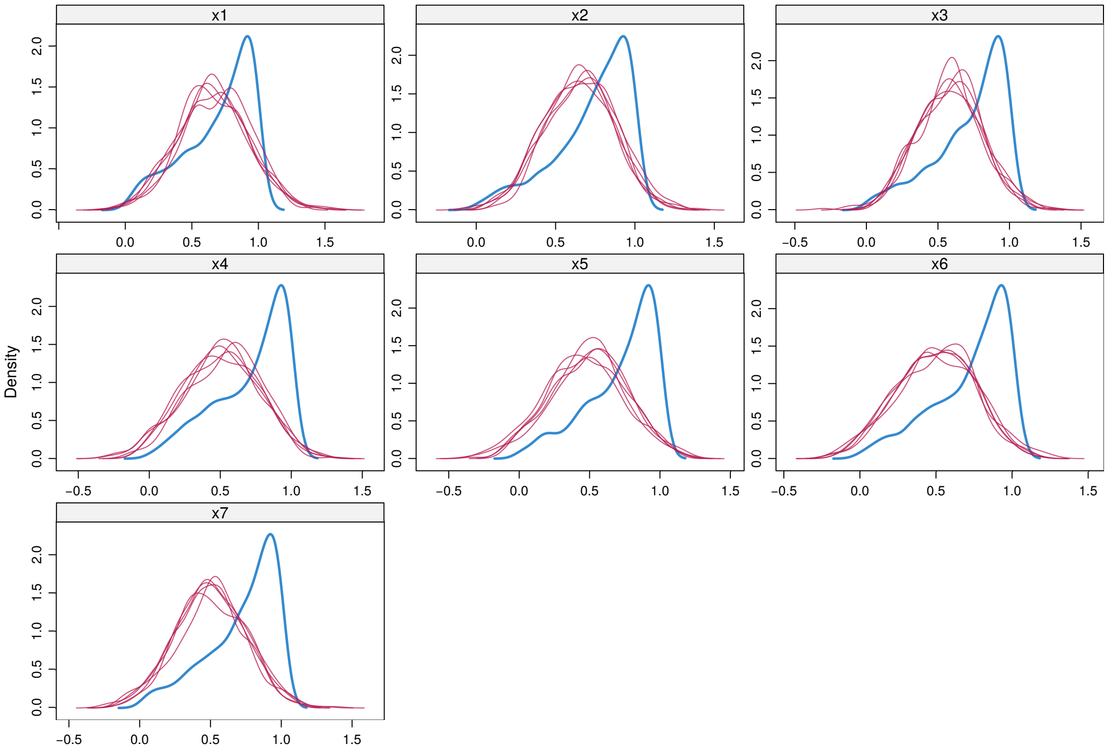


Figure 5.3: Density of observed data and imputed data by applying *norm* with MAR

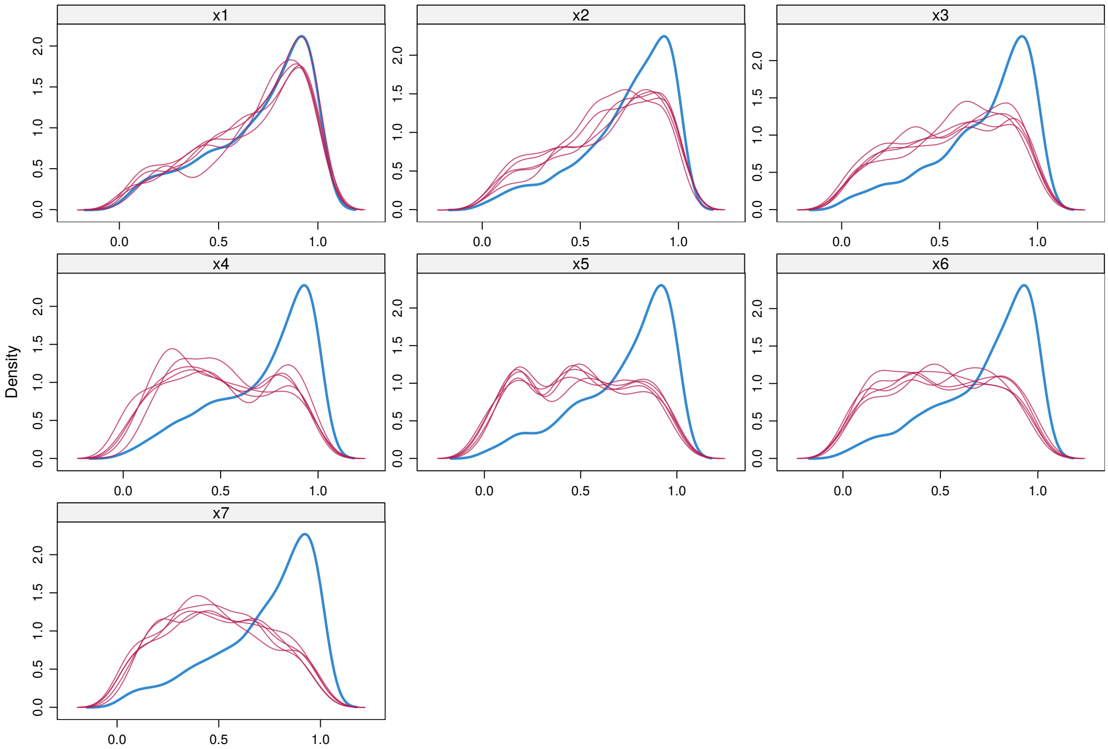


Figure 5.4: Density of observed data and imputed data by applying *pmm* with MAR

In these figures the density of the observed data is shown in blue, while the density of the imputed data is shown in red for each data set. In conclusion, the imputation effects of the two methods exhibit no distinguishing difference, but the *pmm* method is generally a better option.

5.2.2 With AMELIA II Package

Introduction

Amelia II is a complete R-package developed by James Honaker, Gary King, and Matthew Blackwell (J. Honaker et al. (2011)), and is used for the MI of missing data. This package is similar to MICE; indeed, the two packages are both MI methods. In comparison to listwise deletion and single imputation methods, Amelia II is able to significantly reduce the bias in variances and covariances.

Although MICE and Amelia II are both MI methods, certain differences exist between them. For instance, the algorithm applied by Amelia II is the expectation-maximization with bootstrapping (EMB) algorithm, which is a unique bootstrapping approach whereby the expectation-maximization

(EM) algorithm works on multiple bootstrapped samples of the original incomplete data. The missing values are then imputed based on the estimated bootstrapped parameters. Furthermore, if potentially useful information is available that can be used as Bayesian priors, Amelia II can use this information as an additional boost to improve imputation models. Because Amelia II has a relatively high efficiency and the ability to handle a large number of variables, it is one of the MI methods used in this simulation study.

Assumption

There are two basic assumptions in the application of Amelia II. Similar to MICE, Amelia II also assumes that data are MAR. According to (J. Honaker et al. (2011)) a special case of MAR is MCAR, whereby the missing values are created completely randomly and the missingness is not at all dependent on all of the variables. As a result, Amelia II is also suitable for the MCAR missing data mechanism. Another notable point is that for MAR the missingness of one variable depends on other variables, so additional information about these variables helps to predict the missing values of the variable being considered.
A second assumption of Amelia II is that the complete data set, which includes both the observed values and missing values, fits the multivariate normal distribution. In this simulation study the joint distribution of all continuous variables is the multivariate normal distribution with a given mean and covariance, which satisfies this assumption. However, there are also categorical variables in the simulated data set which do not fit the multivariate normal distribution. Nevertheless, Amelia II works just as effectively on these variables (Schafer and Olsen 1998)).

Algorithm

Amelia II combines the bootstrapping approach and the EM (**E**xpectation-**M**aximization) algorithm. As mentioned above, Amelia II is an MI method within which bootstrapping can generate multiple data sets. Bootstrapping is an efficient statistical method that can produce multiple bootstrapped samples, which fulfills the purpose of MI. A great advantage of applying bootstrapping is that it does not consider the distribution of the data. When the size of the original sample is small this approach is especially useful to

estimate bootstrapped parameters. Unlike other statistical methods, which determine confidence intervals with the known mean or standard deviation of the population, bootstrapping utilizes only the sample itself.
Based on the original incomplete data set, resampling is conducted. To be more specific, each element of a bootstrapped sample is extracted from the original incomplete data set, after which the element is returned to the data set and extracted again. Then, each generated bootstrapped data set has the same dimension as the original incomplete data set. Through entirely random resampling, bootstrapped data sets are generated that are mutually exclusive and independent from other bootstrapped data sets, and that are each different from the original incomplete data set.

Figure 5.5 illustrates the procedure of ***Amelia** II*.

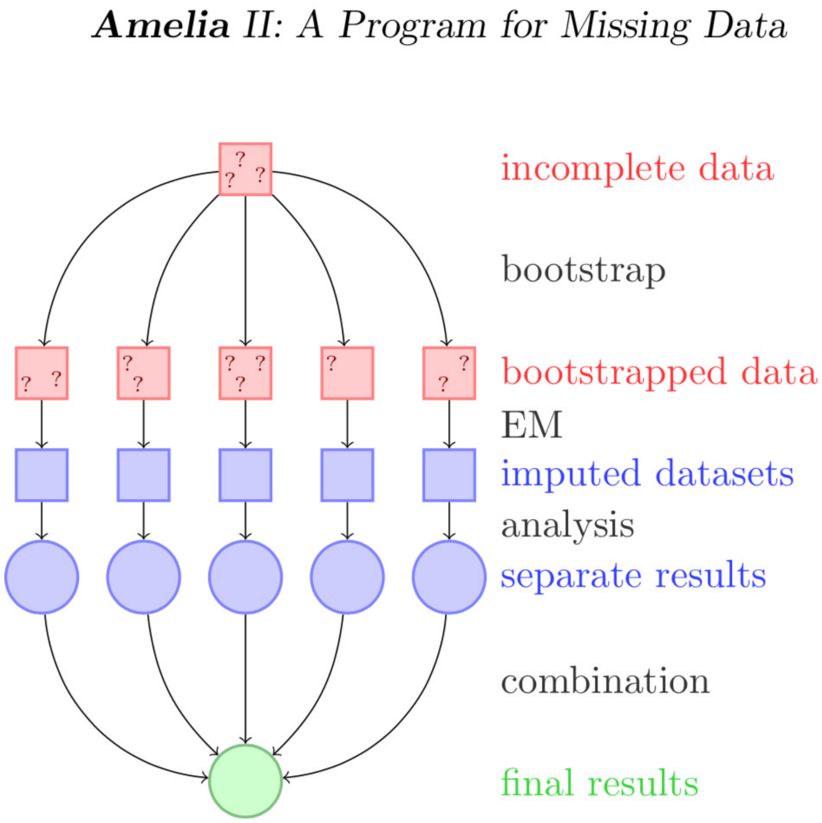


Figure 5.5: Graphic demonstration of multiple imputation with the EMB algorithm from (Honaker et al.(2011)).

Each bootstrapped data set should have missing values. These are imputed using the EM algorithm, an iterative process consisted of two steps. The first is the expectation step (E) and the second is the maximization step (M). In the first step E, a starting value for the estimation of parameter $\theta = (\mu, \sigma)$ is assumed, based on which the values that would replace the missing values are predicted. After this, an initial imputed complete data set is generated, based on which the new parameter $\hat{\theta}_{ML}$ is computed by using the maximum likelihood estimate (MLE) to maximize the log-likelihood θ found on the first E step. If the distribution of the latent variables in the next E step can be determined by the new parameter $\hat{\theta}_{ML}$ in the M step, the next E step can be conducted. The process above would be repeated, which constructs an iteration. Generally, the iteration process would end when the values of the parameter estimates on the successive E step and M step are the closest, indicating that they are convergent. In short, the iteration process stops when convergence occurs. The rate of convergence depends on the missing rate in the data set, which implies that the number of iterations should increase with the increase of missing values in the data set. For instance, if there are no missing values in the data set, then convergence would occur instantly. During each iteration, only the missing values should be replaced while the values of the observed data should remain constant.

Implementation in R

For the purpose of conducting Amelia II as an MI method certain information should be supplied, such as the function of the original incomplete data set, the desired number of multiple imputed data sets *m*, and the types of variables.

Regarding the types of variables, nominal variables should be handled in a rather different way to continuous variables. In the simulated data set there are several nominal variables, which should additionally be specified by the argument *noms*. By setting this argument, Amelia II is able to determine the number of categories p of a multinomial variable, and thus replace $p-1$ binary variables to specify each possible category. In the multivariate normal imputation method these $p-1$ variables, whose missing values are imputed, are treated as other continuous variables.

Table 5.2: List of possible statement for the regression model in Amelia II

Method	**Regression**	**Type of variable**
logit	binomial (logit)	binary
ls	linear (least squares)	continuous
normal	linear (MLE)	continuous
poisson	poisson	count data
probit	binomial (probit)	binary

The possible statements for the regression models can be specified as the table above

In this simulation study, four different missing rates are simulated. Under the circumstance of missing rates 30% and 50% with certain missing data mechanisms, perfect collinearity is likely to occur. Thus, there would be an error report and the process would be unable to carry on. In order to solve this problem, a necessary logical argument should be specified, namely *incheck*, which determines whether or not the inputs to the function should be checked before performing imputation. The default setting of *incheck* is ***TRUE***, but it should be set to ***FALSE*** if perfect collinearity occurs. Another numerical argument *empri* is also a possible solution to this problem, which decreases the covariance of the simulated data while keeps the means and variances constant for high missing rates. For different types of data sets a number of other arguments can be used. These alternative arguments are not relevant to this simulation study so they are not applied here, but they are covered in (Honaker et al. (2011)).

5.2.3 With missForest Package

Introduction

According to (D. J. Stekhoven and P. Bühlmann (2011)) most current imputation methods have certain pitfalls. These methods are suitable for either continuous variables or categorical variables, leading to a lack of consideration of the interactions between different types of variables within mixed-type data sets. Therefore, another imputation method, **missForest**, was developed by (D. J. Stekhoven et al. (2011)). MissForest is an iterative method for imputing missing values based on the random forest algorithm, which is nonparametric. This characteristic is adopted by missForest, which makes it possible for this method to handle different types of variables simultaneously.

In MICE, parametric regression models are needed and assumptions about the distribution of data are considered as prior knowledge. In the application of MICE it is necessary to specify the appropriate approach for each imputed variable in advance. If the assumptions are not correct then biased imputation results are likely to be produced. For example, it is assumed that a continuous variable fits the normal distribution; in fact this variable cannot perfectly fit normal distribution, which would lead to the estimation of problematic parameters. In addition, if there are complicated interactions, nonlinear relation structures, or high correlation between the regression model variables in the data sets, then predictions made using MICE tend to be less accurate. Consequently, the quality of the imputation would be decreased. As mentioned above missForest is a nonparametric method, which implies that it does not need to make assumptions about structural aspects of the data. Therefore, biased imputation results would not be caused by improper assumptions when applying missForest.

Meanwhile, one of the assumptions of Amelia II is that the complete data set, which includes both the observed values and missing values, fits the multivariate normal distribution. However, in reality not all variables that need to be imputed are continuous variables. Although Amelia II can also be applied to impute categorical variables, its performance is less satisfying than that of missForest, which is based on a random forest and thus is more able to deal with mixed-type data. In conclusion, missForest is highly practical in the circumstances of a large number of variables, complex interactions, and nonlinear relation data structures. Additionally, another great advantage of missForest is that it can provide an out-of-bag (OOB) error, a method that can numerically display the prediction error of random forests. With the help of OOB, the number of variables being randomly sampled at each split and the number of decision trees in the random forest can be adjusted.

Algorithm

In order to deal with mixed-type data, missForest uses a nonparametric approach called random forest. Nonparametric statistical approaches do not need to fit an appropriate distribution or the corresponding distribution parameters. Random forest is a method that constructs a large number of classification and regression trees (CART). Depending on the purpose, the CART algorithm generates two types of decision tree: classification trees and regression trees. A classification tree performs binary splits of the data, with each split generated based on only one variable. In consideration of randomness, each classification tree has a different sub-set. Hence, each tree generates different results and votes for a particular class. Meanwhile, the regression tree minimizes the sum of mean squared errors of the response variable and outputs a mean prediction for continuous variables.
The CART algorithm has two main disadvantages, namely instability and overfitting. To handle these problems random forest generates boot-strapped data sets from the original data set. According to the types of variables, each bootstrapped data set produces either a classification tree or a regression tree. Subsequently, m variables are selected in every twig or knot, and the most appropriate split can be determined. Each decision tree is generated based on its corresponding bootstrapped data set, which is combined by random forest. Eventually, random forest combines the results of all decision trees and generates the final outcome by selecting the class that has the greatest number of votes.
In this case, the random forest algorithm is applied to estimate the missing values based on the remaining observed data. For the imputation of missing values, the R-package *missForest* developed by (D. J. Stekhoven et al. (2011)) is applied. According to the developers, suppose there is a covariable matrix $\mathbf{X}(n \times q)$, where $X = (X_1, X_2, ..., X_q)$, then for the variable X_s, $s \in \{1, ..., q\}$, its missing values are denoted as $i_{miss}^{(s)} \subseteq \{1, ..., n\}$, whereas its given observed values are denoted as $i_{miss}^{(s)} \subseteq \{1, ..., n\}$.

Algorithm Impute missing values with random forest.

```
Require: $\mathbf{X}$ an $n \times p$ matrix, stopping criterion $\gamma$
Make initial guess for missing values;
$\mathbf{k} \leftarrow$ vector of sorted indices of columns in $\mathbf{X}$
   w.r.t. increasing amount of missing values;
while not $\gamma$ do
    $\mathbf{X}_{old}^{imp} \leftarrow$ store previously imputed matrix;
    for $s$ in $k$ do
        Fit a random forest: $\mathbf{y}_{obs}^{(s)} \sim \mathbf{x}_{obs}^{(s)}$;
        Predict $\mathbf{y}_{mis}^{(s)}$ using $\mathbf{x}_{mis}^{(s)}$;
        $\mathbf{X}_{new}^{imp} \leftarrow$ update imputed matrix,using predicted $\mathbf{y}_{mis}^{(s)}$;
    end for
    update $\gamma$.
end while
return the imputed matrix $\mathbf{X}^{imp}$
```

The algorithm of *missForest* is described as above. The data set can be divided into four parts as follows.

1. y_{obs}^{s} The observed values of the variable X_s
2. y_{miss}^{s}: The missing values of the variable X_s
3. x_{obs}^{s}: The values of all other variables except X_s in the place of $i_{obs}^{(s)}$
4. x_{miss}^{s}: The values of all other variables except X_s in the place of $i_{miss}^{(s)}$

In the application of *missForest*, there is a stopping criterion γ, which is met as soon as the difference between the newly imputed data matrix and the former one increases for the first time. This is the case for both continuous and categorical variables. The difference for the set of continuous variables N is defined as

$$\Delta_N = \frac{\sum_{j \in N}(\mathbf{X}_{new}^{imp} - \mathbf{X}_{old}^{imp})^2}{\sum_{j \in N}(\mathbf{X}_{new}^{imp})^2} \tag{17}$$

and as for the set of categorical variables F, the difference is defined as

$$\Delta_F = \frac{\sum_{j\in F}\sum_{i=1}^{n} I_{\mathbf{X}^{imp}_{new}\neq\mathbf{X}^{imp}_{old}}}{\#\mathbf{NA}} \tag{18}$$

where #*NA* is the number of missing values in the categorical variables.

Implementation in R

In this part, the application of *missForest* in R to impute missing values will be explained in detail. As mentioned above, *missForest* is an implementation of random forest algorithm.

```
MissForest<- missForest(missForest,maxiter = 10, ntree = 100,
mtry = floor(sqrt(ncol(missForest))), replace = TRUE)
```

There are a number of arguments that can be specified by applying *missForest* function. The following table illustrates some possible arguments used in this simulation study.

Table 5.3: List of possible arguments applied in *missForest* function

Argument	Setting	Description
xmis	missForest	The imported data set
maxiter	10	Maximun number of iterations equals to 10
ntree	100	100 trees to grow in each forest
mtry	floor(sqrt(ncol(missForest)))	The number of variables in the simulated data set
replace	TRUE	Bootstrap sampling is performed with replacements

The *missForest* function returns a list object with three components: "ximp", "OOBerror", and "error". Here "ximp" represents the imputed complete data set, "OOBerror" represents the out-of-bag estimated imputation error, and "error" stands for the true imputation error. In this simulation study the imported data set with missing values is a data frame. The "$" operator is used to return the imputed complete data set from which the OOBerror rates are returned, consisting of two statistical measurements. The first is

the proportion falsely classified (PFC), which is used for categorical variables. The second measurement is the normalized root mean squared error (NRMSE), which is used for continuous variables.
The NRMSE is a normalization of the mean squared error (MSE), which measures the average of the squares of the errors and is also known as the average squared difference between the estimated values and the observed values in the data set. The MSE can be computed as

$$\mathrm{MSE} = \mathrm{MSE}(\hat{\theta}) = \mathrm{E}((\hat{\theta} - \theta)^2) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{19}$$

The root-mean-square error (RMSE), which is the square root of MSE, is also a significant measurement. It is denoted as

$$\mathrm{RMSE} = \mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)} \tag{20}$$

With the help of normalizing, comparison between data sets with different scales can be improved(S. Oba et al. (2003)). NRMSE is denoted as

$$\mathrm{NRMSE} = \mathrm{NRMSE}(\hat{\theta}) = \sqrt{\frac{\mathrm{MSE}(\hat{\theta})}{var(Y_i)}} = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{var(Y_i)}} \tag{21}$$

In this case, the effect of imputation improves with the decrease of NRMSE.

```
MissForest$OOBerror
NRMSE          PFC
0.0004937029 0.3061538373
```

The R-code above is an example of the result of OOBerror, which shows the effect of imputation with MAR and 50% missing rate. The value of NRMSE is 0.000494, which implies the predictive power for random forest and the model can explain the average deviation ±0.05% of the range.

5.3 Predicted Residual Error Sum of Squares (PRESS)

In this thesis the PRESS statistic is the chosen criterion to measure the effect of different imputation methods. The coefficient of determination is an alternative used to measure the quality of fit in regression, but it does not have predictive power. Hence it is not suitable in this case.
The coefficient of determination, denoted as R^2, represents the proportion of the variance in the response variable, which can be predicted from the independent variable. In other words, this index measures how well the model fits with given observations without the ability to make predictions about future values. The coefficient of determination is expressed as follows.

$$R^2 \equiv \frac{SS_{reg}}{SS_{tot}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{22}$$

where

$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares,
$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2$ is the regression sum of squares,
$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$ is the sum of squares of residuals.

In regression analysis, as one of the CV methods, the PRESS statistic is a measure of how well a model fits a sample of observations which themselves were not used to estimate the model. To be more specific, in CV the data set is divided into two separate parts, namely the training data and the test data. Based on the training data a predicted model is fitted, from which the predicted values are calculated and compared to observed values in the test data. As the following formula shows, the PRESS statistic is calculated as the sums of squares of the prediction residuals for the observations.

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{23}$$

The fitted model based on the training data affects the values of the PRESS statistic to a large degree, thus the more appropriate the model, the lower

the value of the PRESS statistic. There are missing values in both the training data and test data. If these missing values are imputed by any imputation method, the imputed complete data set inevitably deviates from the original complete data set on some level, which negatively influences the quality of the training data. Based on the biased training data, the quality of the fitted predictive model may be unsatisfying, which in turn affects the value of the PRESS statistic. Therefore, this thesis uses the PRESS statistic as the criterion to measure different imputation methods.
The PRESS statistic is calculated as the sums of squares of the prediction residuals, which have a considerable value in the case of large samples. This is a disadvantage for graphical organization. Accordingly, this thesis uses a logarithmic transformation of the PRESS statistic to present the results of the comparison.

6 Analyses and Results

As mentioned in section 5, the MI methods used in this simulation study are realized by applying three powerful R-packages, namely **MICE**, **Amelia II**, and **missForest**, while the single imputation method used is **mean substitution (mean imputation)**. All of these imputation methods are compared based on two different types of data sets under two dimensions, which are missing rates and missing data mechanisms. In this section the detailed results of comparison between these methods are presented in the form of boxplots, which illustrate the value of the log-transformed PRESS statistic for a more intuitive demonstration. The label "replace" in the boxplots denotes "replace with column mean", which indicates mean substitution.

6.1 Data Set with Only Continuous Variables

6.1.1 In the MCAR Data Set

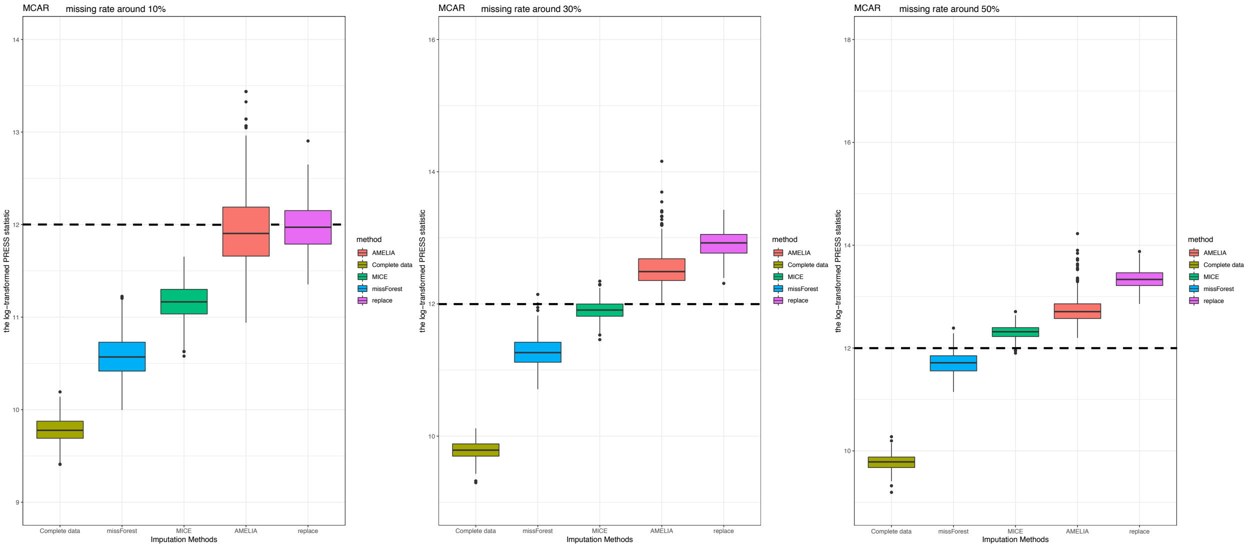


Figure 6.1: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism MCAR

1. With missing rates of around 10%, 30%, and 50%, the order of performance of the imputation methods ranked from best to worst is fixed: missForest, MICE, Amelia II, mean substitution.

2. When the missing rate is around 10%, the values of the log-transformed PRESS statistic of missForest and MICE are relatively further from

12, while these values of Amelia II and mean substitution are around 12. With an increased missing rate the values of the log-transformed PRESS statistic of all imputation methods gradually become closer to 12 and even exceed 12. This implies that the imputation effect of all imputation methods exhibits a decreasing trend with an increased missing rate.

3. When the missing rate reaches around 10%, the performances of Amelia II and mean substitution display almost no difference.

6.1.2 In the MAR Data Set

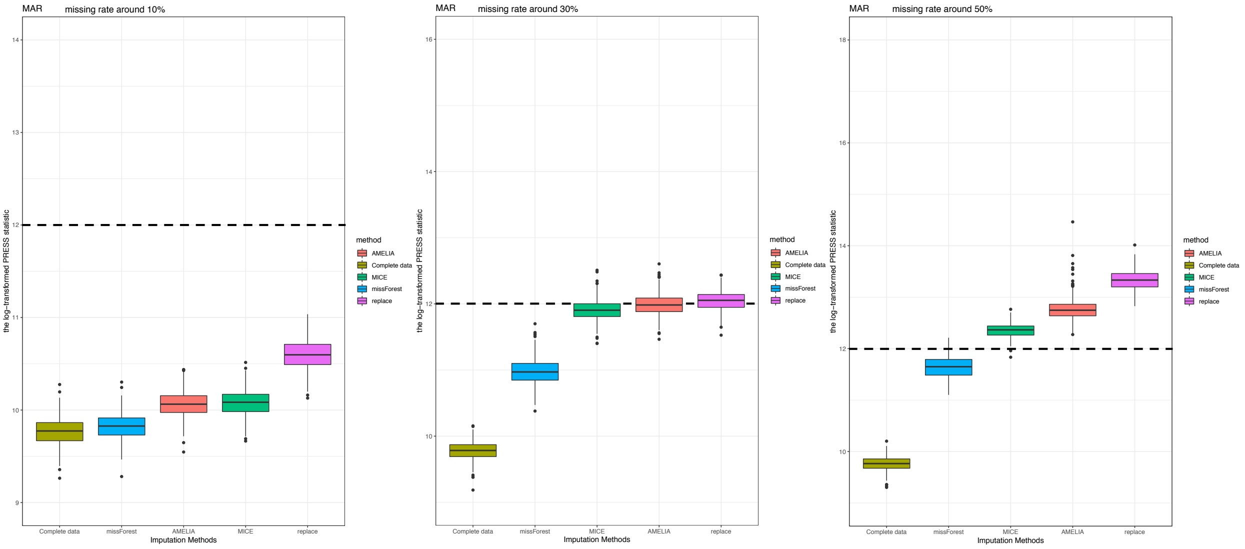


Figure 6.2: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism MAR

1. With a missing rate of around 10%, the order of performance of the imputation methods ranked from best to worst is: missForest, Amelia II, MICE, mean substitution. With missing rates of around 30% and 50% the corresponding order is: missForest, MICE, Amelia II, mean substitution. From the boxplots it can be observed that missForest always performs best, while mean substitution always produces the worst effects of imputation.

2. When the missing rate is around 10% the values of the log-transformed PRESS statistic of all imputation methods are furthest from 12, implying that the aggregate effect of imputation reaches its optimum. As

with the MCAR situations, the value of the log-transformed PRESS statistic gradually increases, implying that the performances of all imputation methods worsen with an increased missing rate.

3. Of the considered methods, when the missing rate is around 10% the value of the log-transformed PRESS statistic of missForest differs least from that of the original complete data set. This indicates that missForest has the best performance. Hence, applying missForest in this situation is able to restore the original complete data set to the greatest extent.

4. Disregarding missForest, there is no fixed order of the imputation effect of the other three imputation methods. Indeed, when the missing rate reaches around 30% the performances of these three imputation methods exhibit almost no difference.

5. The difference between the imputation effect of missForest and those of the other three imputation methods does not present a fixed pattern.

6.1.3 In the NMAR Data Set

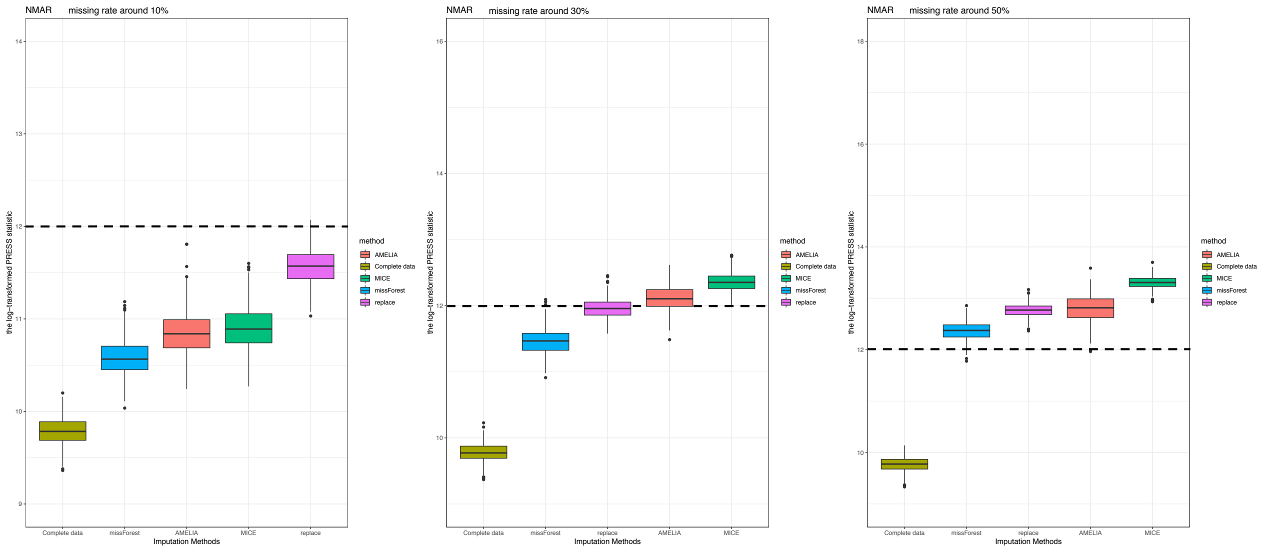


Figure 6.3: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism NMAR

1. With a missing rate of around 10% the order of performance is: missForest, Amelia II, MICE, mean substitution. Noticeably, at missing rates of around 30% and 50%, as a single imputation method mean

substitution performs better than two other imputation methods. The possible reasons for this will be explained in the conclusion section. To compare the performance of MI methods with missing rates of around 10%, 30%, and 50%, the order is: missForest, Amelia II, MICE.

2. When the missing rate is around 10%, the values of the log-transformed PRESS statistic of all imputation methods are furthest from 12. When the missing rate reaches around 30%, the log-transformed PRESS statistic values of Amelia II and MICE exceed 12, and as the missing rate hits around 50% these values all exceed 12, thus proving that the imputation effect declines.

3. When the missing rate is around 10% the imputation effects of Amelia II and MICE are almost equal. Excepting missForest, the imputation effects of the other three imputation methods demonstrate no significant difference when the missing rate is around 30% and 50%.

6.2 Data Set with Continuous and Categorical Variables

6.2.1 In the MCAR Data Set

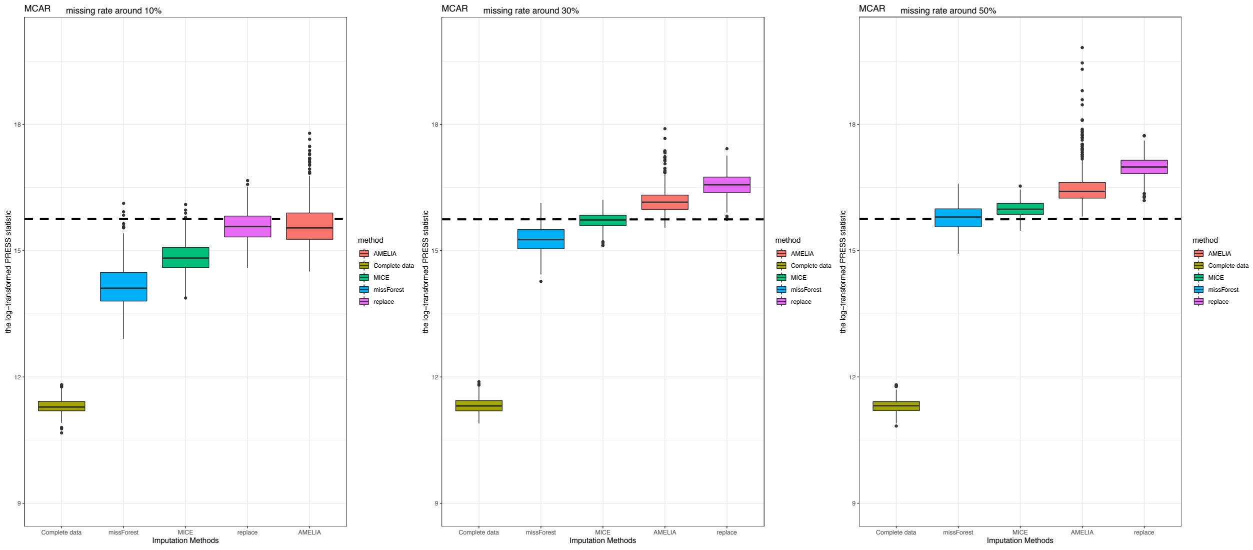


Figure 6.4: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism MCAR

1. With a missing rate of around 10%, the order of performance is: missForest, MICE, mean substitution, Amelia II. At missing rates of around 30% and 50% the corresponding order is: missForest, MICE, Amelia II, mean substitution. The results suggest that the performance of MICE is always better than that of Amelia II. If only the MI methods are compared then the order is fixed: missForest, MICE, Amelia II.

2. When the missing rate is around 10% the log-transformed PRESS statistic values of all imputation methods barely exceed 16. As the missing rate increases, these values become closer to 16 and even surpass 16, which implies that the imputation effect of all imputation methods worsens.

3. When the missing rate is around 10% the log-transformed PRESS statistic value of missForest differs considerably from that of the other three imputation methods. As the missing rate increases this difference gradually reduces, which indicates that the imputation effect of missForest approaches that of the other methods.

6.2.2 In the MAR Data Set

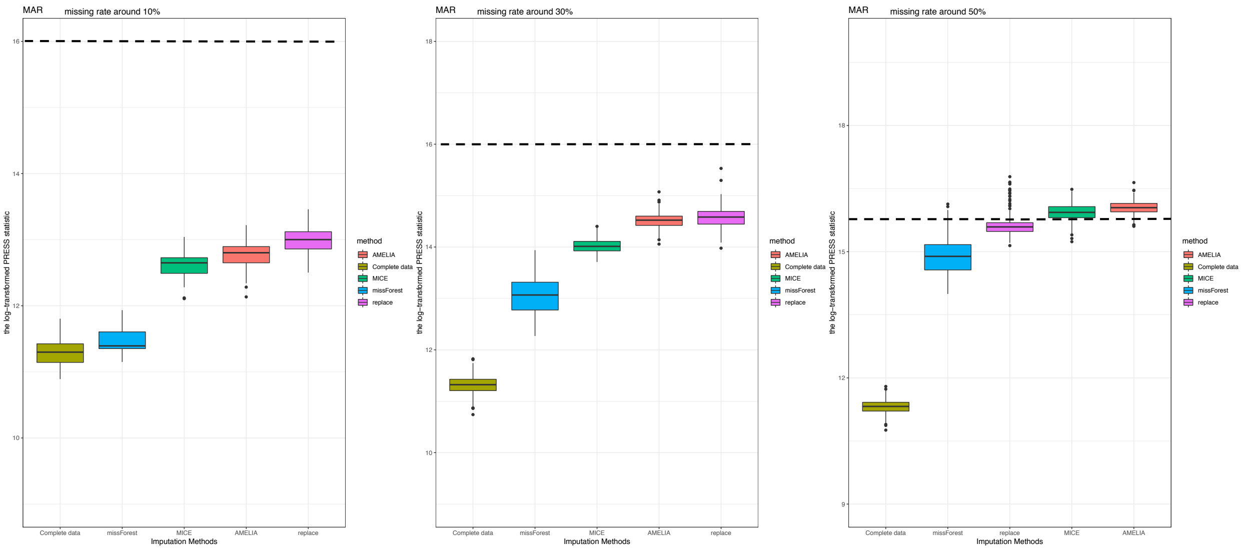


Figure 6.5: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism MAR

1. At missing rates of around 10% and 30% the performance order is: missForest, MICE, Amelia II, mean substitution. At around a 50% missing rate the corresponding order is: missForest, mean substitution, MICE, Amelia II. Notably, the imputation effect of MICE is always better than that of Amelia II, even though both approaches are based on the assumption that the missing data mechanism is MAR.

2. Similar to the situations with MCAR data, the log-transformed PRESS statistic values of all imputation methods gradually increase as the missing rate increases. When the missing rate reaches around 50%, the log-transformed PRESS statistic values of almost all imputation methods becomes closest to or even exceeds 16.

3. When the missing rate is around 10%, the log-transformed PRESS statistic value of missForest exhibits the least difference from that of the original complete data set. Hence, applying missForest under this circumstance is recommended.

4. The difference between the imputation effects of missForest and the other methods gradually decreases, which implies that the imputation effect of missForest approaches that of the other three imputation methods.

6.2.3 In the NMAR Data Set

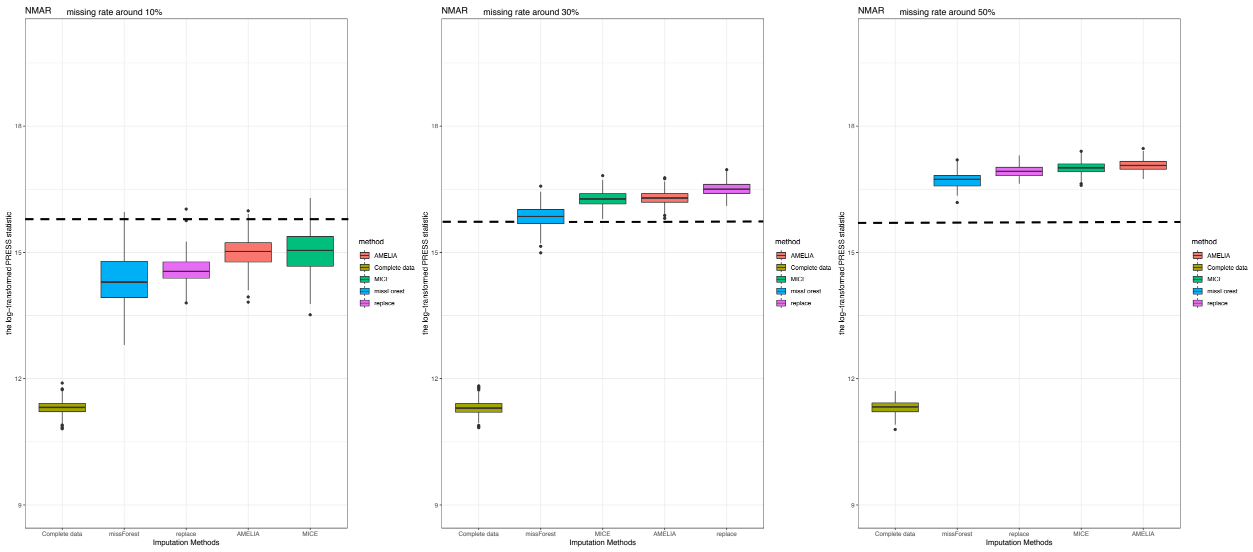


Figure 6.6: Performances of different imputation methods with missing rate around 10%, 30%, and 50% under the missing data mechanism NMAR

1. At a missing rate of around 10% the performance order is: missForest, mean substitution, Amelia II, MICE. At a missing rate of around 30% the corresponding order is: missForest, MICE, Amelia II, mean substitution. Finally, at the missing rate of around 50% the corresponding order is: missForest, mean substitution, MICE, Amelia II.

2. When the missing rate is around 10%, the log-transformed PRESS statistic values of all imputation methods are slightly lower than 16. With the increase of the missing rate, these values become closer to and even exceed 16, implying that the aggregate performance of all imputation methods diminishes.

3. Excluding missForest, there is no fixed order of the imputation effect of the remaining three imputation methods.

4. At missing rates of 10%, 30%, and 50%, the four imputation methods have the closest imputation effect to that demonstrated for MAR and MCAR.

6.3 Conclusions

Notably, the following comparison results are all drawn based on the log-transformed PRESS statistic, which is used as the criterion in this study. If other criteria are applied, comparison results may differ.

The common aspects identified by the comparisons based on two data sets are listed as follows.

1. Compared to the other three imputation methods, missForest exhibits the best performance under all circumstances.
2. The log-transformed PRESS statistic values of all imputation methods are closest to those of the original complete data set when the missing rate is around 10%. With an increased missing rate, the difference between the log-transformed PRESS statistic values of all imputation methods and those of the original complete data set increases. This implies that the imputation effect of all imputation methods displays a decreasing trend with the increase of the missing rate.
3. The values of the log-transformed PRESS statistic of the original complete data set are almost the same under all circumstances. This value is not affected by changing the missing rate because no missing value is created in the original complete data set, thus no imputation method is conducted. Therefore, these values can be considered as a standard to enable comparison of the imputation method performances.

The differences observed from the comparisons based on two data sets are listed as follows.

1. For the continuous data set the interval of the log-transformed PRESS statistic value of the original complete data set is $(9,10)$, and that of the mixed-type data set is $(11,12)$. All log-transformed PRESS statistic values for continuous variables are lower than the same values for mixed-type data. A possible reason is that the number of predictor variables differs between these two data sets, which can affect the log-transformed PRESS statistic values considerably. To be more specific, there are seven variables in the continuous data set, while the mixed-type data set contains 10 values.

2. For the continuous data set, the dotted line represents a log-transformed PRESS statistic value equal to 12 and acts as a fixed reference. Meanwhile, the dotted line representing the value 16 is a fixed reference for the mixed-type data set.

According to the results, single imputation methods appear to perform better than MI methods in certain situations. Possible reasons are explained in the following.

A certain amount of information is stored in the observed values of a data set. For instance, if a data set has missing values then it contains a fixed amount of information. When complete case analysis is applied to deal with the missing values, some information is lost. In contrast, if single imputation methods such as mean substitution are applied to replace missing values by using the existing information in the data set, then the standard deviation is reduced.

A key problem of mean substitution is that it does not take imputation uncertainty into account. As a result, standard errors computed from the imputed data are systematically underestimated. Thus, MI was proposed to deal with this problem.

By applying MI methods, multiple data sets with the same size are generated by stochastic resampling, such as bootstrapping, after which these data sets are imputed. The pooled standard deviation and pooled coefficient are derived from the pooled multiple imputed data sets. With the application of bootstrapping, random noise is added to the prediction; single imputation methods do not achieve this. Compared to single imputation methods, which are intended to impute the missing values as precisely as possible, MI methods aim to impute without underestimating the standard deviation by adding variance to the prediction.

In short, although single imputation methods occasionally seem to perform better, in general MI methods exhibit a better imputation effect.

7 References

References

[1] M. Y. Ivory, M. Hearts: An Empirical Foundation for Automated Web Interface Evaluation. Ph.D. thesis, University of California at Berkeley, 2001

[2] Derrick A. Bennett: How can I deal with missing data in my study.
Australian and New Zealand Journal of Public Health, 2009

[3] Ting Yan, Richard Curtin: The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective.
International Journal of Public Opinion Research, Volume 22, Issue 4, 1 December 2010, Pages 535:551, 2010

[4] S. Fielding, P. M. Fayers and C. R. Ramsay: Investigating the missing data mechanism in quality of life outcomes: a comparison of approaches.
Health and Quality of Life Outcomes, 2009

[5] Roderick J. A. Little, Donald B. Rubin: Statistical Analysis with Missing Data, Second Edition.
ISBN: 978-0-471-18386-0, 2002

[6] Polit, D.F. and Beck, C.T. : Nursing Research: Generating and Assessing Evidence for Nursing Practice. 9th Edition.
Philadelphia :Wolters Kluwer Health/Lippincott Williams & Wilkins, 2012

[7] J. Banks; J. Carson; B. Nelson; D. Nicol: Discrete-Event System Simulation. Prentice Hall.
ISBN: 0-13-088702-1, 2001

[8] Efroymson, M. A.: "Multiple regression analysis," In: A. Ralston and H. S. Wilf, Eds., Mathematical Methods for Digital Computers.
John Wiley, New York, 1960.

[9] Julian J. Faraway : Linear Models with R (Texts in Statistical Science) Second Edition.
ISBN: 978-1439887332, 2009

[10] Patrick Royston , Willi Sauerbrei : Multivariable Model - Building: A Pragmatic Approach to Regression Anaylsis based on Fractional Polynomials for Modelling Continuous Variables.
ISBN: 978-0-470-02842-1, 2008

[11] Akaike, H. : A new look at the statistical model identification.
IEEE Transactions on Automatic Control,19 (6), 1974

[12] Wit, Ernst; Edwin van den Heuvel; Jan-Willem Romeyn : 'All models are wrong...': an introduction to model uncertainty.
Statistica Neerlandica. 66 (3): 217-236, 2012

[13] Burnham, K. P.; Anderson, D. R. : Multimodel inference: understanding AIC and BIC in Model Selection.
Sociological Methods & Research, 33: 261-304, 2004

[14] Daren S. Starnes, Daniel S. Yates, David S. Moore : The Practice of Statistics.
New York : W.H. Freeman, c2012.
ISBN: 9781429262583, 2012

[15] Thomas P. Minka : Estimating a Dirichlet distribution. 2003

[16] Frank A. Haight : Handbook of the Poisson distribution.
New York: John Wiley & Sons.
ISBN 10: 0471339326, 1967

[17] Schafer JL: Multiple imputation: a primer.
Statistical Methods in Medical Research.
PMID: 10347857 DOI: 10.1177/096228029900800102, 1999

[18] Derrick A. Bennett: How can I deal with missing data in my study?.
Australian and New Zealand Journal of Public Health.25(5):464-469. 2009

[19] Yiran Dong: Principled missing data methods for researchers.
SpringerPlus DOI:10.1186/2193-1801-2-222, 2013

[20] Enders, C. K.: Applied missing data analysis.
The guilford press ISBN 978-1-60623-639-0, 2010

[21] Eekhout et al,: Missing data in a multi-item instrument were best handled by multiple imputation at the item score level.
Journal of Clinical Epidemiology 67(3) DOI: 10.1016/j.jclinepi, 2013

[22] Patrick Royston: Multiple imputation of missing values.
The Stata Journal
4, Number 3, pp. 227-241, 2004

[23] He et al. : Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide.
Statistical Methods in Medical Research
2009:1-18. Epub ahead of print, 2009

[24] Stuart et al. :
Multiple Imputation With Large Data Sets: A Case Study of the Children's Mental Health Initiative
American Journal of Epidemiology.169:1133-1139, 2009

[25] Stef van Buuren, Karin Groothuis-Oudshoorn :
mice: Multivariate Imputation by Chained Equations in R
Journal of Statistical Software 45(3): 1-67, 2009

[26] Schafer JL, Olsen MK:
Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective
Multivariate Behavioral Research, 33(4), 545-571, 1998

[27] S. Oba et al. :
A Bayesian missing value estimation method for gene expression profile data
Bioinformatics, vol. 19, no. 16, pp. 2088-2096, 2003

Declaration

Herewith I declare that I have written this thesis completely by myself and did not use neither other sources nor resources except the listed ones. Moreover, I have not handed in this thesis elsewhere and also have not published it yet.

Munich, February 04, 2019

Rui Yang