Missing Data Analysis with the Mahalanobis Distance

by

Elaine M. Berkery, B.Sc.

Department of Mathematics and Statistics,
University of Limerick

A thesis submitted for the award of M.Sc.

Supervisor: Dr. Kevin Hayes

Submitted to the University of Limerick, Ireland, 2016
Declaration

This thesis is presented in fulfilment of the requirements for the degree of Master of Science. It is entirely my own work, completed without collaboration with others except my supervisor, Dr. Kevin Hayes. Where use has been made of the work of other people it has been fully acknowledged and referenced accordingly.

Signature:  

-------------------------------
Elaine Berkery

April 2016
Acknowledgements

I would firstly like to thank my supervisor, Dr. Kevin Hayes, for allowing me to carry out this research project with him and for his guidance and help in submitting this thesis.

To my colleagues and many friends in the department of Mathematics and Statistics, thank you for your help and support throughout my time here. A special word of thanks to Peg, Jo and Sinead for going out of their way to help me whenever I needed it.

To my sisters, brother and friends for your eternal support and kindness when it was needed. You truly have ridden this roller coaster with me and for that I thank you.

Lastly and most importantly, I wish to thank my parents, Willie and Eileen, without your love and support I would not have made it to today. I cannot thank you enough and couldn’t wish for better.
Abstract

Missing data occur regularly when data are collected for a variety of reasons such as participants refusing to answer question in surveys or machines failing to record measurements in a manufacturing process. The fact that data are missing cannot be ignored. Removing observations and analysing only a complete dataset can affect the results of any subsequent analysis. Many methods have been developed to deal with the problems that arise as a result of having missing values including the widely used method of multiple imputation.

This thesis examines one such method of generating imputed datasets using multiple imputation and a distance measure known as the Mahalanobis distance. Using the Mahalanobis distance identifies similar observations, which are fully observed, to those with missing values from which to draw estimates of those missing values. Amendments to a currently used method are proposed, the results compared to simulated data and applied to a real dataset. It also outlines the importance and usefulness of visualisation in missing data analysis.

Additional to this missing data work, a study was carried out on Growing Up in Ireland data and the ability of both children and their primary care givers at rating their BMI whilst simultaneously accounting for the missing data that exists in this dataset.
Contents

1 Introduction ........................................... 3
  1.1 Outline of Chapters ..................................... 5
  1.2 Datasets Used .......................................... 6
    1.2.1 Biomechanical Jump Data ............................... 6
    1.2.2 Growing Up in Ireland ............................... 7

2 Missing Data Analysis ................................... 8
  2.1 Introduction ........................................... 8
  2.2 Historical Review ....................................... 9
  2.3 Patterns of Missing Data ............................... 11
  2.4 Mechanisms of Missing Data ............................ 13
  2.5 Overcoming Missing Data Problems ....................... 15
    2.5.1 Ad-hoc methods ..................................... 16
    2.5.2 EM algorithm ........................................ 18
    2.5.3 Multiple Imputation ................................ 19
    2.5.4 Available Software .................................. 23
  2.6 Visualising Missing Data ............................... 24
  2.7 Summary ................................................ 30

3 Mahalanobis Distance ..................................... 31
  3.1 Uses of the Mahalanobis Distance ....................... 32
  3.2 SOLAS for Missing Data ............................... 37
    3.2.1 Modifications to Mahalanobis Distance Matching ....... 40
    3.2.2 Shortcomings and Benefits .......................... 41
Chapter 1

Introduction

The occurrence of missing data is ubiquitous in applied statistics and can occur in all types of datasets for a variety of different reasons. However, standard statistical methods of analysing data have been developed with the intention to be used on fully observed datasets and from a pedagogical point of view, much of the teaching of statistics concentrates on the study of fully observed data. As a result, missing data analysis has evolved to become a specialised area within the field of statistics.

Reasons for missing data are manifold. For example, in chemical experiments, certain results may be unrecorded due to the breakdown of a machine that does not depend on the experimental process itself. In a health survey, respondents may refuse to report what their weight is. In a household survey, respondents with an income level above a certain cut-off point may choose not to report their income because of the level it is at. Regardless of the reason for missing data, it is imperative that the missing data is acknowledged and dealt with accordingly by using one of a variety of methods available to the analyst. This thesis aims to propose another viable option for dealing with the missing data problem.

Missing data cause a loss of vital information and can introduce bias into inferences if not treated correctly when carrying out analysis (Little and Rubin, 2002). Simply deleting the observations with missing cases is not a viable method of dealing with the missing data and this thesis aims to introduce and develop some of the
current methods of dealing with those missing values. Because of its prevalence in 
statistics, many methods have been developed to deal with the problem of missing 
data. Such methods include multiple imputation (Rubin, 1978) and the expectation 
maximisation (EM) algorithm (Dempster et al., 1977). The focus of this thesis will 
be on multiple imputation but both methods, among others, will be outlined.

Rubin (1976) was the first to introduce the concept of the mechanism of missing 
data and showed how these assumptions affect the subsequent analysis. Since that 
paper the study of missing data has grown but the assumptions outlined by Rubin 
(1976) are of huge importance still. The definitions of the mechanisms of missing 
data outlined in that paper form the backbone of much research today.

Following on from his definition of mechanisms of missing data, Rubin (1978) 
first introduced the idea of multiple imputation, a widely used method in dealing 
with missing data up to this day. Multiple imputation allows the data analyst to 
create several fully observed datasets, thus capturing some of the uncertainty that 
is associated with imputation. The analyst can then analyse each complete dataset 
using standard statistical techniques before combining the results of any analysis 

Further to these mechanisms of missing data, the pattern of missing data is also 
important when analysing missing data. Various patterns can occur for a variety of 
reasons. For example, a monotone missing data pattern, explained in more detail in 
chapter 2, occurs often in the case of longitudinal surveys where participant dropout 
happens before the end of the study and do not return at any point after dropout.

The EM algorithm developed by Dempster et al. (1977) doesn’t create multiple 
datasets like those in multiple imputation but instead replaces missing values with 
estimates for these values, the relevant parameters needed are then estimated and 
the missing values re-estimated using these parameters. This process is repeated 
until convergence is reached. The EM algorithm, while initially developed in the
missing data sphere has gone on to be used across many areas of statistics including factor analysis (Becker et al., 1997) and principal component analysis (Tipping and Bishop, 1999).

**SOLAS**, a statistical software package developed by Statistical-Solutions (2014), implements a method of performing multiple imputation using a distance metric called the Mahalanobis distance (Mahalanobis, 1936). This thesis aims to investigate this method of multiple imputation further and propose changes and modifications to it accordingly. These changes are implemented and compared to the original method on simulated data and on a real dataset.

Throughout the duration of this research, some of the work has been presented at various conferences including the Conference of Applied Statistics in Ireland 2014, the Research Students Conference 2014 and the inaugural MissData Conference 2015.

### 1.1 Outline of Chapters

Chapter 2 introduces and discusses missing data analysis in some detail. Patterns of missing data and mechanisms of missing data (Rubin, 1976) are described along with details on historical and current methods of dealing with missing data. These methods include multiple imputation (Rubin, 1978) and the expectation maximisation (EM) algorithm (Dempster et al., 1977). This chapter also outlines the importance and usefulness of using visual aids and techniques (Templ et al., 2012) in the analysis of missing data and how these aids can particularly help in identifying the patterns and mechanisms of missing data that will be outlined.

Chapter 3 introduces the Mahalanobis distance (Mahalanobis, 1936) and its use throughout statistics. The chapter shows where and how it is currently used in the context of missing data within the software package **SOLAS** (Statistical-Solutions, 2016) for the purpose of multiple imputation. This chapter goes on to propose
changes and alterations to how this distance measure can be used in the context of missing data and multiple imputation. It includes detailed algorithms of how the process currently undertaken within the SOLAS software package is understood and algorithms on how modifications of this method are undertaken and implemented.

In chapter 4, the method proposed in section 3.2 along with the current method are implemented on simulated multivariate normal datasets and the results analysed. The results are analysed using criteria regularly used to compare various methods of dealing with missing data as can be seen in Collins et al. (2001) and Schafer and Graham (2002) among others.

Chapter 5 introduces a separate body of work on the Growing Up in Ireland (GUI) dataset in relation to how well primary care givers recognise their childrens levels of obesity and Body Mass Index (BMI). Some missing data exist in this dataset and these missing values are accounted for using current methods of multiple imputation for categorical variables.

1.2 Datasets Used

1.2.1 Biomechanical Jump Data

A dataset containing information on an experiment that examined the effects of cold water immersion, also known as cryotherapy, on jumping movement and force measures was provided by the Biomechanics Research Unit in the University of Limerick. Subjects were recorded performing a simple vertical jump using force plates and high speed cameras. This dataset contains many missing values for a number of reasons. Participants may have missed the measurement plates on landing after a jump was carried out, or the machine may not have recorded the values correctly.
1.2.2 Growing Up in Ireland

Growing Up in Ireland (GUI)(ESRI, 2009) is a national study of children who grow up in the Republic of Ireland and is the most significant study of its kind to take place in the country. This study is following the progress of almost 20,000 children across Ireland to collect a host of information to help improve our understanding of all aspects of children and their development. A huge number of variables were collected on the subjects who took part in the study across a wide range of areas from social and economic to cultural environments. As there was a large number of participants and many variables, missing data was a common feature throughout the data. For the purposes of any analysis carried out to date there is a small subset of those variables that were of particular interest. Moving forward with analysis, any work done can be expanded to include the rest of the available data. More detail will be provided on the variables in later chapters when more analysis is carried out.
Chapter 2

Missing Data Analysis

2.1 Introduction

Datasets containing missing data are an ubiquitous problem in practice (Rubin, 1976) and occur across a wide range of research areas from psychology and education to clinical trials. The most widely used methods of statistical analysis are used mainly in the setting of fully observed datasets and these methods have been widely taught to statistics students. However, outside the classroom, research into the problem of missing data is an area in statistics that has grown hugely since the 1970’s and 1980’s (van Buuren, 2012), much of it enabled because of advances in computing power.

In reality, missing data occur regularly and for a variety of reasons. When faced with the problem of missing data, the analyst needs to understand and answer a number of questions before proceeding. Why are missing data present? Are there patterns that exist within these missing data? How should analysis be carried out on a dataset with missing values?

There are many reasons why missing data arise and may be specific to a particular problem, experiment or survey carried out. These missing values could occur randomly or may be due to a certain set of circumstances being in play at the time of observation. For example, in filling out a household survey, the respondent may
choose not to disclose the what their income is or in recording the temperature of a bioreactor in the manufacturing process of a drug, a thermometer may malfunction. If the reason why there is missing data can be identified, preventative action can be taken to avoid missing data in future surveys or experiments.

The underlying reasons for missing data and the relationship between variables with missing data and fully observed variables can give some answers to the specific question that is of interest to the analyst. These underlying interactions require the analyst to look at the patterns of missing data and what are known as mechanisms of missing data (Rubin, 1976) which will be introduced later in this chapter following this brief introduction.

When patterns and mechanisms of missing data have been correctly identified, it is then possible for the analyst to decide on the most appropriate way to proceed in the statistical analysis of the dataset taking into account the missing data. There are two “state of the art” methods (Schafer and Graham, 2002) that are widely used, multiple imputation and maximum likelihood, among many others.

2.2 Historical Review

While the growth of missing data research began in earnest from the 70’s onwards, the research in the area began a long time before this. Allan and Wishart (1930) proposed a formula for the least squares estimate of one missing value in a randomised block design and also in a Latin square design. Another of the earliest papers on dealing with missing data was by Wilks (1932) in which maximum likelihood estimates of the population parameters for a bivariate Normal population are developed. Yates (1933) developed a method of dealing with missing data in ANOVA where missing values are replaced by their least squares estimates \( \hat{y}_i = x_i \hat{\beta}_s \) where \( \hat{\beta}_s = (X^T X)^{-1} (X^T Y) \) applied to the \( r \) rows of \((X, Y)\) that have \( y_i \) observed. Matthai (1951), Lord (1955) and Anderson (1957) extended Wilks (1932) approach of maximum likelihood estimates to the trivariate Normal case.
Iteration methods to find missing values have also been developed historically. Hartley (1956) proposed a method for dealing with one missing value that substituted three trial values for the missing value and calculating the error sum of squares for each trial value. He suggested that this method could be used iteratively when more than one missing value exists. Healy and Westmacott (1956) proposed another iterative method that can be used in any analysis where least squares estimates are derived. Missing values are substituted with trial values, a complete case analysis is carried out, predicted values are obtained for the missing values and are then used as new substitutes for the missing values. A complete case analysis is then performed and the process continues until the error sum of squares no longer decreases. To increase the speed of convergence in this method, Pearce (1965) and Preece (1971) suggested improvements.

The basic theory upon which much of modern missing data analysis is based, particularly the two “state of the art” methods outlined by Schafer and Graham (2002), began in the 1970’s. First with Rubin (1976), where the ideas of patterns of missing data and mechanisms of missing data were introduced. These ideas are discussed further in Sections 2.3 and 2.4 respectively. The breakthrough came in Rubin (1978) when multiple imputation was proposed for the first time. The idea of multiple imputation was that a number of complete datasets were created rather than just one which allowed the analyst to take into account the level of uncertainty associated with imputation or filling in values. Even though the concepts were developed in these two seminal papers, a full introduction to the ideas only came later in Rubin (1987). Two of the more prominent methods of multiple imputation are joint modelling (Rubin and Schafer, 1990) and fully conditional specification (van Buuren et al., 2006).

The second “state of the art” method is the expectation maximisation algorithm (EM). Similar algorithms to the EM algorithm have been proposed before, such as that in Hartley (1958) before the seminal paper was published by Dempster et al.
(1977) introducing a general algorithm which has since been shown to have many uses outside of missing data analysis (Becker et al., 1997; Tipping and Bishop, 1999).

With the amount of research in the area of missing data growing, so too has the quantity of software to deal with the issue. Packages such as MICE (van Buuren and Groothuis-Oudshoorn, 2011), VIM (Templ et al., 2011) and Amelia (Honaker et al., 2011) implement various methods of dealing with missing data in R (R Development Core Team, 2008), including both multiple imputation and maximum likelihood.

### 2.3 Patterns of Missing Data

Before beginning an analysis on data containing missing values and in order to decide how to proceed with any analysis, it is important to understand why there are missing values. Looking at the missing data pattern in a dataset can help with these decisions. Little and Rubin (2002) outline how certain methods of dealing with missing data are suitable for use with particular missing data patterns. Define $Y = (y_{ij})$ as an $(n \times K)$ rectangular dataset with the $i$th row $y_i = (y_{i1}, \ldots, y_{iK})$ and $y_{ij}$ is the value of subject $i$ for variable $j$. To denote missing data, a missing data indicator matrix $M = m_{ij}$ is defined, such that $m_{ij} = 0$ if $y_{ij}$ is observed and $m_{ij} = 1$ if $y_{ij}$ is missing. $M$ can then be used to identify the pattern of missing data. Examples of such missing data patterns are illustrated in Figure 2.3.

In Figure 2.3(a), univariate nonresponse refers to missing values occurring in only one variable. This is one of the most basic patterns of missing data and could occur in reality as a result of low response rate to a particular question on a survey. An extension of the univariate nonresponse is the multivariate nonresponse in Figure 2.3(b). This is also known as unit nonresponse and occurs when no data exist for a particular respondent in a survey other than the data that was known before distribution of the survey. Item nonresponse occurs when certain questions are answered by a respond but others are not. This can occur in a haphazard fashion and no real obvious pattern is evident. An example of an item nonresponse
Figure 2.1: Patterns of “missingness”. Rows correspond to observations, columns to variables (Little and Rubin, 2002).

pattern can be seen in Figure 2.3(d). From the point of view of the subsequent analysis, unit nonresponse is more crippling than item nonresponse. In the case of item nonresponse, some analysis can still be carried out on the data collected on these respondents however no information is available on the former case.

The monotone pattern in Figure 2.3(e) is a feature common to longitudinal studies. Respondents are tracked over a period of time and the same questions are asked at different time points. Some subjects in the study may choose to drop out of it before the end and do not return, resulting in the pattern seen above. This can be particularly prevalent in health studies if the health of a subject deteriorates to an extent that they can no longer participate or in some cases even, to death. The pattern in Figure 2.3(e) is known as file matching. This pattern often occurs as a result of combining a number of datasets from two sources. $Y_1$ is common to both sources but $Y_2$ is only present in the first of the two sources and $Y_3$ is only present in the second of the sources. It is important to be aware of this pattern when carrying out analysis as some parameters representing the association between the variables cannot be estimated from the data. The final pattern in Figure 2.3(f), called factor analysis occurs when one variable is completely observed and the other is completely missing. This approach can be used in the case of “latent” variables, when the latent
variable is never observed and treated as a missing data problem.

2.4 Mechanisms of Missing Data

While identifying the pattern of missing data that occurs is of help to the statistician, it is even more important to identify the mechanism that has lead to the missing data. There are three mechanisms of missing data that were first introduced by Rubin (1976) and now form the backbone of all missing data analysis. The properties of missing data methods depend on the missing data mechanism. Again, $M$ is a missing data indicator matrix, such that $m_{ij} = 1$ if $y_{ij}$, as defined in the Section 2.3, is missing and $m_{ij} = 0$ if $y_{ij}$ is observed.

Data is said to be missing completely at random (MCAR) if the missing observation is independent of all other values of the data $Y$. The missing data mechanism is characterised by the conditional distribution of $M$ given $Y$, say $f(M|Y, \phi)$, where $\phi$ denotes unknown parameters

$$f(M|Y, \phi) = f(M|\phi) \text{ for all } Y, \phi.$$  \hfill (2.1)

Raghunathan (2004) notes that MCAR is a very strong assumption to make and is rarely satisfied in reality. To demonstrate MCAR, consider, for example, a health survey where a patient’s age, blood pressure and cholesterol levels are recorded. The variable cholesterol contains missing values and both age and blood pressure are fully observed. If the mechanism is MCAR, then the missing values do not depend on the values of either age or blood pressure.

Data is said to be missing at random (MAR) if the missing values depend on the values observed in $Y$. Define $Y_{\text{obs}}$ as the observed entries of $Y$ and $Y_{\text{mis}}$ as the missing entries. We can express MAR using the probability statement

$$f(M|Y, \phi) = f(M|Y_{\text{obs}}, \phi) \text{ for all } Y_{\text{mis}}, \phi.$$  \hfill (2.2)
Consider the same example as before. If the mechanism is MAR, missing values of cholesterol depend on values of another variable, say blood pressure. It may be the case that if a participant’s blood pressure was lower than a certain value, cholesterol wasn’t measured for that participant.

The final mechanism is called missing not at random (MNAR) and occurs when the distribution of $M$ depends on the value of the missing observation in the data $Y$. Consider a simple univariate random sample where some of the units are missing. Let $Y = (y_1, ..., y_n)'$, where $y_i$ is the value of a random variable for unit $i$, and $M = (M_1, ..., M_n)$, where $M_i = 0$ for units that are observed and $M_i = 1$ for units that are missing. If the joint distribution $(y_i, M_i)$ is independent across all units then the probability that a unit is observed does not depend on the values of $Y$ or $M$ for other units. Then,

$$f(Y, M|\theta, \phi) = f(Y|\theta)f(M|Y, \phi) = \prod_{i=1}^{n} f(y_i|\theta) \prod_{i=1}^{n} f(M_i|y_i, \phi),$$

(2.3)

where $f(y_i|\theta)$ is the density of $y_i$, indexed by unknown parameters $\theta$ and $f(M_i|y_i, \phi)$ is the density of a Bernoulli distribution for the binary indicator $m_i$ with probability $\Pr(M_i = 1|y_i, \phi)$ that $y_i$ is missing. If missing data is independent of $Y$, that is $\Pr(M_i = 1|y_i, \phi) = \phi$, a constant that does not depend on $y_i$, then the mechanism of missing data is MCAR (or in this case equivalently MAR). If the mechanism depends on $y_i$ the mechanism of missing data is MNAR as it depends on $y_i$ that is missing.

Again consider the previous example, if the missing data mechanism is MNAR, the missing cholesterol values depend on the values of cholesterol themselves. In this situation, it could happen that the device that measures cholesterol cannot read values above a certain level and so whether cholesterol is observed or missing depends on what that value for cholesterol was itself.

The missing data mechanism is not an assumption or characteristic about an
entire dataset, rather it relates to a specific analysis that is carried out and the variables used in this analysis (Baraldi and Enders, 2010). As a result, a dataset could contain all three mechanisms of missing data depending on what analysis was carried out and what variables are used. Mechanisms of missing data are outlined further in Allison (2002) and Enders (2010).

In order to identify a missing data mechanism, it is possible to carry out a test on data to determine if it is MCAR. Little (1988b) proposed a method of identifying if data are MCAR which uses all available data and reduces to a t-test when the data are bivariate with missing values confined to one variable. It is not, however, possible to identify or test to check if data are MAR (Baraldi and Enders, 2010; Potthoff et al., 2006; Harel and Zhou, 2007). Instead most methods of dealing with missing data assume that the mechanism at play is MAR. As knowing if the mechanism is MNAR requires knowledge of the underlying variables, it is not possible to test for this mechanism either.

2.5 Overcoming Missing Data Problems

In order to deal with missing data, the analyst must undertake some additional work to obtain the parameters of interest that relate to the data in question. The most common ways of dealing with missing data are, firstly, to ‘fill in’ the missing values, know as imputation and introduced initially by Rubin (1978) or secondly, to use a maximum likelihood approach using the Expecation-Maximisation (EM) algorithm, introduced by Dempster et al. (1977), to estimate the parameters of the model. The main focus of this thesis will be on multiple imputation although the general idea behind the EM algorithm will be introduced in section 2.5.2 as well as several ad-hoc methods which, although not the best way of dealing with missing data, are commonly used throughout research and described in Little and Rubin (2002) and Kalton and Kaspryzk (1982) among many others.
2.5.1 Ad-hoc methods

Before the introduction of multiple imputation methods, other various methods of “filling in” the missing values were, and in some fields continue to be used. Karahalios et al. (2012) reviewed published papers that had results derived from datasets with missing values and noted that some of the methods used to deal with missing data were known to produce biased results such as last observation carried forward and mean imputation. These methods, among others, are documented below. When the data is filled in, the complete case dataset is then analysed using standard statistical methods.

**Complete case analysis:** Under complete case analysis, any units in a study that contain missing values in any variable are removed and only those participants who have responded to all variables are included in the analysis. This is also known as listwise deletion. An advantage of complete case analysis is its simplicity, there is no computation required to obtain a fully observed dataset. However, the loss of information by deleting partially observed units can result in a loss in precision and if the mechanism of missing data is not MCAR, bias may be introduced in results. Another version of complete case analysis is called pairwise deletion. In this case, subjects are removed on an analysis by analysis basis, depending on the variables used in a particular analysis.

**Mean imputation:** The mean of the observed values in each variable is calculated and is subsequently substituted for any missing values within that variable. An advantage of this method is that it is not computationally expensive, is simple to implement and unbiased for the mean if the data is MCAR. However, this method results in the variance being underestimated and it is inappropriate for categorical variables. The distribution of the data is also changed and mean imputation results in biased results if the data are MAR.

**Regression imputation:** Regression is carried out on the complete case dataset
with the resulting regression equation then used to estimate missing values. In this case variance is again underestimated and as a result associations are biased. The correlations between variables used in the regression equations are artificially increased and again the variance is underestimated.

**Indicator method:** This method does not exclude any variables from an analysis but instead an indicator or dummy variable is added to the statistical model being used to indicate if an observation is observed or missing. This method of imputation however, results in biased estimates, even in the MCAR case.

**Last observation carried forward:** This practice is used mainly for longitudinal data, particularly in medical studies where participant dropout occurs after a period of time and the participant does not return. The last value observed for a participant is carried into all future time points that are missing for that unit. This practice results in incorrect means and variances.

Although deletion methods are considered some of the worst ways of handling missing data (Wilkinson, 1999), the practice occurs regularly in reality. Peugh and Enders (2004) investigated how missing data were dealt with in educational research. In 160 papers that reported data containing missing values, 96% used some form of listwise or pairwise deletion, 5 papers used mean imputation and 1 used regression imputation. In no instances were maximum likelihood or multiple imputation used. Wood et al. (2004) analysed how missing data are handled in clinical trials. The trials considered in their analysis were all published in the *British Medical Journal, Journal of the American Medical Association, The Lancet* and *the New England Journal of Medicine*. Of the 71 trials reviewed, 63 reported missing data yet only one trial carried out multiple imputation. The vast majority again carrying out a complete case analysis. Eekhout et al. (2012) reviewed the handling of missing data in epidemiology research papers. In 81% of the studies considered a complete case analysis was carried out while only 8% used multiple imputation. In cancer research, Burton and Altman (2004) analysed 100 papers, 81 of which had missing
data but only one of these papers reported using multiple imputation in dealing with the missing values.

2.5.2 EM algorithm

The Expectation Maximisation (EM) algorithm was first introduced by Dempster et al. (1977) and made it possible to compute maximum likelihood estimates in missing data problems. Little and Rubin (1987) outline many examples of the EM algorithm in their book and Enders (2001) summarises the use of maximum likelihood approaches including the EM algorithm. The basic steps to the procedure are outlined as follows:

1. Replace the missing values with estimated values.
2. Estimate the parameters of the data.
3. Re-estimate the missing values using the parameters just found, assuming that these are correct.
4. Repeat this process until convergence is reached.

Each iteration of the EM algorithm has two steps, the expectation step and the maximisation step. The M step performs a maximum likelihood estimation of $\theta$, the parameter of interest, as if there was no missing data and uses the standard computational method as maximum likelihood estimation from $\ell(\theta|Y)$.

The E step finds the conditional expectation of the missing data given the observed data and the current estimated parameters and then substitutes these expectations for the missing values. Differentiating the EM algorithm from ad hoc methods of filling in missing values is the idea that the missing data are not $Y_{mis}$ but functions of $Y_{mis}$ in the complete data loglikelihood $\ell(\theta|Y)$.

If $\theta^{(t)}$ is the current estimate for a parameter $\theta$, the E step finds the expected complete data loglikelihood if $\theta$ were $\theta^{(t)}$. 
\[ Q(\theta|\theta^{(t)}) = \int \ell(\theta|Y)f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)})dY_{mis}. \]

The M step finds \( \theta^{(t+1)} \) by maximising the expected complete data loglikelihood found above:

\[ Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}), \text{ for all } \theta. \]

The EM algorithm while introduced first to deal with missing data has since been used in a wide range of statistical areas including factor analysis (Becker et al., 1997) and principal component analysis (Tipping and Bishop, 1999).

### 2.5.3 Multiple Imputation

Each of the ad-hoc methods outlined in section 2.5.1 results in one complete dataset which can then analysed using regular methods of statistical analysis. Multiple imputation, introduced first in Rubin (1978) results in a number of complete datasets which can be analysed separately and then the results combined using rules set out by Rubin (1987). The standard procedure when carrying out multiple imputation is to follow the following steps:

1. Choose an imputation model that will generate the multiple complete datasets.

2. Choose the number of imputations to be generated, say \( m \). More information on how the number of imputations should be chosen are outlined further in this section.

3. When each of the \( m \) fully observed datasets have been generated, the relevant analysis should be carried out on each dataset.

4. Once the analysis has been carried out on each of the \( m \) datasets, the results of each of the \( m \) datasets should be combined using the rules outlined in Rubin (1987).
Types of Multiple Imputation models

There are a large number of models that are implemented within the multiple imputation framework. Some of these models include fully conditional specification (van Buuren et al., 2006), also known as chained equations (Azur et al., 2011) and joint modelling (Rubin and Schafer, 1990).

**Fully conditional specification (FCS):** FCS require the user to specify \( P(Y_{mis}|Y_{obs}, X, M) \) before imputations are drawn using Markov Chain Monte Carlo (MCMC) techniques. This chained equations method (van Buuren and Oudshoorn, 1999) of multiple imputation takes a variable by variable approach. The imputation model is specified for each variable using the other variables in the dataset as predictors and the imputed values of that variable are then used in the imputation of the next variable. This process is repeated using a Gibbs sampling procedure until convergence is reached.

**Joint Modelling:** Joint modelling was proposed by Rubin and Schafer (1990) and require the user to specify the joint model, \( P(Y, X, M) \), derive \( P(Y_{mis}|Y_{obs}, X, M) \) before using MCMC techniques to draw imputations for \( Y_{mis} \). The joint modelling approach yields correct statistical inference under the assumed joint model but there is a lack of flexibility and can impute impossible values.

**Conditional Gaussian:** When dealing with both continuous and categorical data with missing values, a conditional Gaussian approach (Schafer, 2010) allows imputation on both types of data. A log-linear model is specified for the categorical data and conditional on this a multivariate Gaussian distribution is assumed for the continuous predictors.

**Hot deck imputation:** Recorded units in the sample are then substituted for missing values based on certain matching criteria. This is a very common method of dealing with missing values in survey practice. The use of the Mahalanobis
distance as a method of choosing similar respondents that will be discussed further in chapter 3. Hot deck imputation can be used as a method of either simple (where one complete dataset is generated) or multiple imputation.

**Choosing the number of Imputations**

When deciding on how many fully observed datasets to impute, it is surprising to find that a small number of imputations can be sufficient. Rubin (1987) showed that the efficiency of an estimate for the parameter, $\theta$, based on $m$ imputations is

$$\left(1 + \frac{\gamma}{m}\right)^{-1},$$

where $\gamma$ is the rate of missing data for the quantity being estimated. The rate of missing data, $\gamma$, is described by

$$\gamma = \frac{r + 2/(df + 3)}{r + 1},$$

where $df$, $\bar{U}$ and $B$ are defined in equations 2.8, 2.5 and 2.6 below and

$$r = \frac{(1 + m^{-1})B}{\bar{U}}.$$

Some examples of the rates of efficiency for various values of $\gamma$ and $m$ can be seen in table 2.1.

<table>
<thead>
<tr>
<th>$m$</th>
<th>$\gamma$</th>
<th>$0.1$</th>
<th>$0.3$</th>
<th>$0.5$</th>
<th>$0.7$</th>
<th>$0.9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97</td>
<td>91</td>
<td>86</td>
<td>81</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>94</td>
<td>91</td>
<td>88</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>97</td>
<td>95</td>
<td>93</td>
<td>92</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.1: Rate of efficiency for various values of $\gamma$ and $m$. 

21
Combining the Results of Each Imputation

To combine the results of each imputation the rules set out in Rubin (1987) must be implemented. The parameter estimate and their standard errors from each of the \( m \) imputed datasets must be saved. Suppose that \( \hat{Q}_j \) is the parameter of interest e.g. a regression coefficient, from each of \( j \) datasets with \( j = 1, 2, ..m \) and \( U_j \) is the variance for the \( Q_j \)th parameter. The estimate for the parameter \( Q \) from the \( m \) multiple imputations is

\[
\bar{Q} = \frac{1}{m} \sum_{j=1}^{m} \hat{Q}_j. \tag{2.4}
\]

The within imputation variance

\[
\bar{U} = \frac{1}{m} \sum_{j=1}^{m} \hat{U}_j, \tag{2.5}
\]

and \( B \) is the between imputation variance

\[
B = \frac{1}{m-1} \sum_{j=1}^{m} (\hat{Q}_j - \bar{Q})^2 \tag{2.6}
\]

Consequently, the total variance associated with \( \bar{Q} \) is denoted by

\[
T = \bar{U} + (1 + \frac{1}{m})B \tag{2.7}
\]

and the degrees of freedom, \( df \), are given by

\[
df = (m - 1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2. \tag{2.8}
\]

Formulas 2.4 - 2.8 will be utilised further in Chapter 4 to combine the results of analysis carried out on simulated datasets using the proposed method suggested in Chapter 3.

Advantages of Multiple Imputation

- Once the multiple imputations have been obtained, complete data methods can be used to analyse the data.
• The data collector’s knowledge can be incorporated to reflect uncertainty around what values should be imputed.

• The uncertainty in values that are imputed are captured.

• Increases the efficiency of the estimation compared to single imputation.

Disadvantages of Multiple Imputation

• More work is required to produce multiple fully complete datasets compared to single imputation and it is computationally more exhaustive.

• More computer space is needed to store the multiple imputed datasets.

• More work is required to analyse each of the imputed datasets before combining the results using the rules set out above.

2.5.4 Available Software

While there are undoubtedly many software packages available to deal with datasets containing missing data, a short summary of some of the more popular ones are outlined below.

*MICE:* MICE (Multivariate Imputation by Chained Equations) (van Buuren and Groothuis-Oudshoorn, 2011) is a software package available for download and use within the R (R Development Core Team, 2008) programming language. MICE carries out multiple imputation using fully conditional specification (FCS). Under FCS, the $P(Y_{mis}|Y_{obs},\phi,M)$ is specified and imputations are then drawn using Markov Chain Monte Carle (MCMC) techniques.

*Amelia II:* Amelia II (Honaker et al., 2011) is a software package also freely available for download and use within the R farmework. It allows users to impute missing data using a combination of the EM algorithm and a bootstrapping technique that has been proposed by the authors.
SOLAS: SOLAS (Statistical-Solutions, 2016) is a stand alone software package provided by Statistical Solutions Ltd. They provide a number of multiple and single imputation methods for implementation as well as a range of graphical features. One such multiple imputation technique will be explained and explored further in chapter 3.2.

SAS: SAS (SAS Institute Inc., 2003) software provides a number of procedures for dealing with missing data. Multiple imputation, MCMC simulations, generalised estimating equations and structural equation models are all available for use on datasets containing missing data.

STATA: STATA (StataCorp, 2015), data analysis and statistics software, also has features that allows users to handle missing data. These features include chained equations, MCMC methods and a range of regression methods.

2.6 Visualising Missing Data

Visualising data is an extremely important way of obtaining information about any dataset and this is no different when working in the context of missing data. Here, the use of visual aids can be of huge importance. It allows the analyst to simultaneously carry out preliminary data analysis and gain an insight into the structure of the missing and non-missing elements of the dataset as well as the relationship between missing and non missing parts (Templ et al., 2012). Visual tools are especially used in the identification of missing data patterns and the mechanisms of missing data that were introduced in chapter 2.

The R (R Development Core Team, 2008) package VIM (Templ et al., 2011) is extremely useful when it comes to visualising missing data with the capability to focus on analysing just one variable with another as much imputation is done one variable at a time, or alternatively to focus on the multivariate data structure. For illustration purposes, I have used the GUI dataset to show how the visual aids can
be used in missing data analysis. An outline of the variables used can be found from table 5.1.

The pattern of missing data is much easier to identify compared to the mechanism of missing data as each different pattern type can be identified directly by observation rather than having to consider underlying reasons as is the case with mechanisms of missing data. When using visual aids to identify mechanisms of missing data, there are a number of limitations that should be taken into account. As can be expected, identifying missing value mechanisms requires some knowledge of the missing data values themselves and this is not always possible. When missing values occur across multiple variables, the missing data mechanism is even more difficult to identify. Additionally, for non-correlated data, it is difficult to distinguish between MCAR and MNAR and when considering correlated data, it is not possible to differentiate between MNAR and MAR.

Aggregation plots can be used to indicate how many missing values are present within each variable and also to identify the most and least frequent combinations of variables with missing values. The plots that show the combinations of missing values can also be used to identify the patterns of missing data at play in the dataset. This knowledge can be useful for the analyst if generating missing values in artificial data for simulation studies. In the case of Figure 2.2, 81% of cases are fully observed across all 6 variables included.

Histograms, barplots and spinograms, some of which can be seen in Figure 2.3 can be used to plot the values of one variable against the number of missing and observed values in another variable. Spinograms can be used to see the relative frequencies within the bins of the histograms on the horizontal axis while on the vertical axis, the proportion of missing data on the other variable is represented. Here, as the BMI of the primary care giver increases, the proportion of missing values in the self reported BMI also increases.
Figure 2.2: Aggregation plot. The plot on the left shows the proportion of missing values on each variable used for this example while the plot on the right shows the variables where different combinations of missing data occur.

Parallell box plots can be especially useful for identifying instances of MAR situations by plotting the distributions of a variable with whether they are observed or missing on another. In Figure 2.4 it look as though there may be a difference between the ages of those people who are observed and missing on variable $MMB1$. This may be an indicator of the mechanism MAR. Across all other variables, there doesn’t seem to be a difference dependent on that variable being observed or missing.

Standard scatterplots can be used and enhanced in the case of missing data to include additional information in the margins of the scatterplot. The values that are observed on one variable can be included in the margin along with a box plot of the distribution of the data for all observed values and the distribution of the data that are only observed on one variable. These can also indicate situations of MAR. In Figure 2.5 an example of this type of margin scatter plot is displayed.

While the histograms, spinograms and scatterplots are useful in the case where only two variables are considered, there are ways of visualising multivariate data.
Figure 2.3: On the left is the histogram of the variable \textit{intPCGBMI} with colour coding for that variable as missing (red) or observed (blue) on the variable \textit{srPCGBMI}. On the right are the same two variables, this time shown on a spinogram.

The scatterplot matrix above in Figure 2.5 can be enhanced in some ways to analyse multiple variables in what are called scatterplot matrices. Figure 2.6 plots a combination of four variables together, missing data is indicated by rug marks on the horizontal and vertical axes. One variable can be highlighted to investigate its relationship with all others. In this case, variable \textit{MMB1} is chosen as the significant variable. In each of the scatterplots, the red crosses indicate observations that are missing for the variable \textit{MMB1}. On the diagonal, the density of each of the variables is displayed in blue, while in red is the density of that variable for observations that are missing on the variable \textit{MMB1}.

The final plot under consideration, an example of which can be seen in Figure 2.7 for the analysis of multivariate missing data analysis makes use of a mosaic plot and is most useful when considering categorical or ordinal variables. The mosaic is first divided up into the relative frequencies of the combinations of variables before the tile of each combination if then divided into the the proportion of missing and ob-
Figure 2.4: Here, values of the variable age are compared using parallel box plots, depending on whether they are observed or missing in other variables.

While these plots help to visually detect some of the mechanisms of missing data that may be evident in a dataset, it is my hope to enhance some of them to carry out statistical tests that will automatically test for differences between populations when a variable is missing or observed on another and immediately notify the user of the mechanism they are dealing with.
Figure 2.5: Here, values of the variable \textit{intPCGBMI} are plotted against the variable \textit{srPCGBMI}. In the margins are the values of the variables that are observed in one but not in the other, along with box plots of the distributions of the data that are observed (in blue) and missing (in red), for each variable.

Figure 2.6: Here, values of the variable \textit{intPCGBMI} are plotted against the variable \textit{srPCGBMI}. In the margins are the values of the variables that are observed in one but not in the other, along with box plots of the distributions of the data that are observed (in blue) and missing (in red).
Figure 2.7: Mosaic plot of the variables $CQ22$ and $mma5ap2$ with proportions of missing and observed values in variable $intChildBMI$.

2.7 Summary

In summation of this chapter, it is important to consider a number of things when dealing with missing data problems:

1. The pattern of missing data should be considered first as some methods of dealing with missing data are suited to particular patterns.

2. The mechanism of missing data should then be identified where possible. Tests can be carried out to determine if the mechanism is MCAR. Many methods of dealing with missing data make the assumption that data is MAR as MNAR requires some knowledge of the missing values themselves.

3. Visual aids can be used to help determine patterns and mechanisms of missing data. While visual tools cannot provide conclusive proof, they can be an excellent indicator.

4. The most suitable method to deal with missing data should then be chosen, be that a maximum likelihood method or multiple imputation method.
Chapter 3

Mahalanobis Distance

The Mahalanobis distance (Mahalanobis, 1936) is a measure used to calculate how similar one set of points are, to another set, taking into account the covariance matrix between the variables used. It allows for a different scale and variance in each variable in a multivariate dataset. The Mahalanobis distance follows a Chi-squared distribution (Brereton, 2015) with \( d \) degrees of freedom where the data is multivariate normal with \( d \) dimensions.

In multivariate statistics, it is often assumed that the row vectors \( x'_i, i = 1, \ldots, n \) of an \( n \times p \) matrix \( X \) are independent realisations from the multivariate normal density \( \text{MVN}_p(\mu, \Sigma) \). The maximum likelihood estimates of the \( p \) dimensional mean vector, \( \mu \), and \( p \times p \) variance matrix \( \Sigma \), are \( \hat{\mu} = \bar{x} = \left( \frac{1}{n} \right) X'u \), and \( \hat{\Sigma} = X'CX/n \), respectively, where \( I \) is the \( n \times n \) identity matrix, \( u \) is the vector of ones in \( \mathbb{R}^n \) and the \( n \times n \) matrix \( C = I_n - \frac{1}{n} uu' \). (The matrix \( CX \) has rows \( x'_i - \bar{x}' \) for \( i = 1, \ldots, n \).) Typically the unbiased estimate \( S = X'CX/(n - 1) \) of \( \Sigma \) is used in place of \( \hat{\Sigma} \).

If \( S \) is non-singular, the Mahalanobis squared distance of the \( p \) dimensional vector \( x_i \) is defined as \( D_i^2 = (x_i - \bar{x})'S^{-1}(x_i - \bar{x}) \). The 5% and 1% critical values for testing for the discordancy of a single outlier in a sample of multivariate normally distributed observations using the test statistic \( M = \max_i \{D_i^2\} \), are given in Table XXXII in Barnett and Lewis (1984).
The Mahalanobis squared distance can also be defined as a dissimilarity measure between two random vectors $x$ and $\tilde{x}$ from the same distribution with the covariance matrix $S$ as

$$D_{x\tilde{x}}^2 = (x - \tilde{x})' S^{-1} (x - \tilde{x}).$$  \hspace{1cm} (3.1)

The functional form of the MVN$_p(\mu, \Sigma)$ density is

$$f_X(x) = (2\pi)^{-n/2} \det(\Sigma)^{-1/2} \exp\left\{ -\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu) \right\}, \text{ for } x \in \mathbb{R}^p. \hspace{1cm} (3.2)$$

The exponent $-\frac{1}{2} (x - \mu)' \Sigma^{-1} (x - \mu)$ specifies an ellipsoid in $\mathbb{R}^p$, which, in the case of $p = 2$ are a series of continuous ellipses. Here the Mahalanobis distance is measuring distance of observation vectors $x$ to the centre of the ellipsoid. In other words, Mahalanobis distance is capturing the remoteness of $x$ from $\mu$, taking into account the geometry of $\Sigma$.

The Mahalanobis distance in equation (3.1) reduces to the Euclidean distance if the correlation $\rho_{ij} = 0$ between each pair of variables, and each variable has unit variance i.e. the inverse variance-covariance matrix, $S^{-1}$ is in fact the identity matrix.

### 3.1 Uses of the Mahalanobis Distance

The Mahalanobis distance is used in a wide rage of statistical areas including the detection of outliers and robust alternatives, principal component analysis, discriminant analysis and cluster analysis.

In the detection of outliers (Rousseeuw and Leroy, 2005) within datasets, the Mahalanobis distance is particularly useful when the dimension of a dataset or matrix is greater than 2 as a visual perception of the data is no longer of use. Outliers are then defined as observations with a Mahalanobis distance of greater than the cut off of $\sqrt{\chi^2_{d,\alpha}}$. Consider the example below in Figure 3.1, multivariate normal data of $d = 2$ dimensions is simulated to illustrate the concept of using the Mahalanobis
distance as an outlier detector. The ellipses of constant Mahalanobis distances of 1, 2 and 3 are drawn in black while the cut off point \( \sqrt{\chi^2_{2,0.95}} = 2.4477 \) is shown in red. Any points that lie outside this cut off are then considered as outliers. In Figure 3.1 there are two outliers.

![Ellipses Having Equal Mahalanobis Distance From Center](image.png)

Figure 3.1: Plot of simulated data for two variables together with ellipses (in black) of equal Mahalanobis distances from the centre point. In red is the cut off ellipse, outside of which points are classified as outliers.

However, extreme values or clusters of outliers can have an effect on the Mahalanobis distance calculated, as both the sample mean and the covariance matrix use these outliers in their calculations. As a result, what is known as a masking effect occurs, and points which are actually outliers are not detected as such.

Consequently, a number of robust methods of calculating the mean and variance-covariance matrices measures have been developed in order to better detect these outliers. Stahel (1981) and Donoho (1982) proposed a robust estimator for the mean and variance with a high breakdown point i.e. the fraction of outliers that a method can handle. They used the minimum volume ellipsoid (MVE) estimator introduced
by Rousseeuw (1985). The centre of the MVE covering half of the observations is
used for $\hat{x}$ and $S$ is calculated using the same ellipsoid.

Another robust method of calculating the Mahalanobis distance is to use the mini-
mum covariance determinant method developed by Rousseeuw (1984) which also has
a high breakdown point.

The Mahalanobis distance is also used within the principal component analysis
(PCA) framework. PCA (Jolliffe, 2002) is a method used to identify a small number
of uncorrelated variables in a dataset from a larger set of data. It is used commonly
to reduce the number of variables in a dataset, particularly if the existing variables
are highly correlated. It is also used to reduce the number of predictors that may
be needed to be used in a statistical model. The new variables created using PCA,
$z_1, z_2, ..., z_p$ are linear combinations of the original variables:

\[
\begin{align*}
    z_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \\
    z_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \\
    & \vdots \\
    z_p &= a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p
\end{align*}
\]

and $a_{ij}$ is the weight of the $j$th variable for the $i$th principal component.
The Mahalanobis distance is used in principal components in that the squared Ma-
hanobis distance is equal to the sum of the squares of the scores of all non-zero
standardised principal components.

A limitation of the Mahalanobis distance is that the use of the inverse variance-
covariance matrix in its calculation can cause a number of problems. Multicollinear-
ity can exist in datasets that contain a large number of variables which may result in
a variance-covariance matrix that cannot be inverted. Also, the number of variables
in the dataset must be less than the sample size. As a result it may be useful to
reduce the number of variables in the dataset using PCA first before calculating the
Mahalanobis distances.

Mahalanobis distance also occurs in the area of discriminant analysis (DA) (De Maesschalck et al., 2000). DA is a classification technique used to classify individuals into one of two or more populations using the variables that best discriminate between the populations. A training sample is used whereby the population an individual belongs to is known in the training sample and the population membership of future individuals is unknown. Only continuous predictors can be used and it is assumed that the variables are multivariate normally distributed. Variables used for classification are $x_1, x_2, ..., x_p$. For $p$ predictor variables, a value $Z$, called Fisher’s discriminant function is given by

$$Z = a_1x_1 + a_2x_2 + ... + a_px_p.$$  

The mean values for $Z$ are calculated for each population, $\bar{Z}_1$ and $\bar{Z}_2$, in the training sample. A cut-off value $C$ must be chosen to divide individuals into each of the two populations where the value for

$$C = \frac{\bar{Z}_1 + \bar{Z}_2}{2}.$$  

To measure how far apart the two populations are the Mahalanobis distance between the two can then be calculated. The coefficients $a_1, a_2, ..., a_p$ are selected such that the Mahalanobis distance between the two populations are maximised. The larger the Mahalanobis distance between the two populations is, the easier it is to discriminate between the two.

Cluster analysis (Johnson et al., 1992) is the process of grouping observations together that are more similar to others in the same group as themselves than those in other groups. A number of methods of measuring similarity between variables can be used. The most commonly used is the Euclidean distance outlined in equation 3 however, the Mahalanobis distance is often used here too. Hierarchical methods are one such technique that is used to cluster data points into categories. Hierarchical
methods can be either “bottom up” or “top down”. “Bottomup”, whereby we start with $n$ clusters and $n$ is the number of observations on a dataset. The two closest clusters are then combined and this is continued until there is only one cluster or the desired number of clusters left, the desired number is usually known beforehand. “Top down” is a decisive method where there is 1 cluster to begin with and the process then removes the most dissimilar observations.

The Mahalanobis deviates also have a relationship in terms of the general linear model. Consider the general linear model as defined in Christensen(1996),

$$ Y = X\beta + \varepsilon \quad (3.3) $$

where $Y$ is an $n \times 1$ vector of observations, $X$ is an $n \times p$ matrix of $p$ variables, $\beta$ is a $p \times 1$ vector of parameters and $\varepsilon$ is an $n \times 1$ vector of errors. This vector of errors, $\varepsilon$, is assumed to be distributed $N(0, \sigma^2 I)$ i.e. that the expected value of the errors is 0 and the covariance of $e$ is $\sigma^2 I$ where $\sigma^2$ is an unknown parameter. This results in the covariance matrix being a diagonal matrix as it is multiplied by the identity matrix

$$ I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} $$

For example, analysis of variance (ANOVA), in its simplest form can be expressed as a general linear model. In the ANOVA case, the $n \times p$ matrix $X$, is a matrix consisting entirely of 0’s and 1’s as described below. The ANOVA model can be described as

$$ y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (3.4) $$

where $i$ is the number of experimental units or groups being investigated and $j$ is the number of observations within each of the experimental units. $j$ may not
necessarily be the same for each group \( i \). Consider the example where \( i = 1, 2, 3 \) and \( j = 1, N_1, N_2, N_3 \) and for the purposes of this example \((N_1, N_2, N_3) = (2, 2, 2)\)

\[
\begin{bmatrix}
y_{11} \\
y_{12} \\
y_{21} \\
y_{22} \\
y_{31} \\
y_{32}
\end{bmatrix}
= 
\begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\alpha_3 \\
\beta
\end{bmatrix}
+ 
\begin{bmatrix}
\varepsilon_{11} \\
\varepsilon_{12} \\
\varepsilon_{21} \\
\varepsilon_{22} \\
\varepsilon_{31} \\
\varepsilon_{32}
\end{bmatrix}
\]

The general linear model can be manipulated to obtain

\[ Y - X\beta = \varepsilon \quad (3.5) \]

and when considering the Mahalanobis distance in terms of the general linear model in equation 3.5, the squared Mahalanobis distance is defined as

\[
D^2 = (Y - X\beta)'(\sigma^2 I)^{-1}(Y - X\beta).
\]

Haslett and Hayes (1998) show how the residuals, \( \hat{e}_A \), the conditional or leave-\( A \) out prediction residuals, and \( \tilde{e}_A \), the marginal or classical residuals of set \( A \) contribute to the Mahalanobis distances of a subset \( Y_A \) from the rest of the model which has been fitted without reference to \( Y_A \).

### 3.2 SOLAS for Missing Data

While each of the areas of statistics mentioned above make use of the Mahalanobis distance, it is also used in the area of missing data. Vacek and Ashikaga (1980) proposed the Mahalanobis distance for distance-function matching and this is further referred to in Little (1988a).

One software package that specialises in missing data imputation is called SOLAS.
The SOLAS software boasts a wide range of methods to carry out multiple imputation for missing data, one of which is a Mahalanobis distance matching method. This method proceeds using the following steps.

Consider the multivariate data in the rectangular layout:

\[
\begin{bmatrix}
  z_{11}^{\text{obs}} & \cdots & z_{1q}^{\text{obs}} & x_{11} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  z_{n1}^{\text{obs}} & \cdots & z_{nq}^{\text{obs}} & x_{n1} & \cdots & x_{np} \\
  z_{11}^{\text{mis}} & \cdots & z_{1q}^{\text{mis}} & \tilde{x}_{11} & \cdots & \tilde{x}_{1p} \\
  \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
  z_{m1}^{\text{mis}} & \cdots & z_{mq}^{\text{mis}} & \tilde{x}_{m1} & \cdots & \tilde{x}_{mp}
\end{bmatrix}
\]

where the superscripts “obs” and “mis” on the matrices \(Z^{\text{obs}}\) and \(Z^{\text{mis}}\) indicate observed values and missing values for the criterion variables \(Z_1, \ldots, Z_q\), and the corresponding matrices \(X\) and \(\tilde{X}\) represent complete records on \(p\) associated covariates \(X_1, \ldots, X_p\). It should be noted that \(Z^{\text{mis}}\) may not be fully missing and that various patterns of missing data will be evident within this matrix.

Consider the following example to illustrate the matrix above. There exists a matrix with data measured across four variables for six units:

\[
Y = \begin{bmatrix}
  3 & 2 & 7 & 5 \\
  1 & 2 & \text{NA} & \text{NA} \\
  2 & 4 & 8 & 6 \\
  4 & 4 & 6 & \text{NA} \\
  3 & 4 & \text{NA} & 5 \\
  2 & 3 & 6 & 7
\end{bmatrix},
\]

Converting this example matrix \(Y\) into the format as outlined above we obtain
As mentioned previously, the $Z^{mis}$ matrix is not completely missing but rather has some observed and some missing values.

The Mahalanobis distance matching imputation algorithm in SOLAS proceeds by first calculating

$$D_{ij}^2 = D^2_{x_i,\tilde{x}_j} = (x_i - \tilde{x}_j)'S^{-1}(x_i - \tilde{x}_j)$$

for every $i = 1,\ldots,n$ and $j = 1,\ldots,m$. There are several ways in SOLAS to use the values of $D_{ij}^2$ to select vectors to form what are known as donor pools for each missing case $j$.

1. Fix the value of $K$ as a pre-set integer. SOLAS uses the default value $K = 10$.
   For each value of $j$ select the $K$ vectors $x_i$ with the smallest values of $D_{ij}^2$.

2. Choose $K$ as a pre-specified percentage of the donor pool. SOLAS uses a default value of ten percent, $K = 0.1 \times n$. For each value of $j$ select the $K$ vectors $x_i$ with the smallest values of $D_{ij}^2$.

Once the donor pool has been chosen, the imputed values for the missing cases are then randomly selected from the corresponding variables in the donor pool using the approximate Bayesian bootstrap method (Rubin and Schenker, 1986).

While the multiple imputation implemented by SOLAS is a simple, fast and very implementable approach compared to other multiple imputation methods that could be used, there are a number of drawbacks to the process. The documentation on the website and in the accompanying white paper (Grannell, 2011) are very sparse.
and do not have all of the relevant information to deduce exactly what process is being implemented. The process has not been properly critiqued and this is what this thesis aims to do.

When investigating the process carried out by SOLAS, and outlined further in Grannell (2011), it is not obvious whether the sample variance-covariance matrix $S$ used in $D^2_{ij}$ is based on: the $n \times p$ data matrix $X$; or the $(n - 1) \times p$ data matrix $X_{-i}$, that is, $X$ with row $i$ deleted; or the augmented $(n + m - 2) \times p$ matrix
\[
\begin{bmatrix}
  X_{-i} \\
  \tilde{X}_{-j}
\end{bmatrix},
\]
where $\tilde{X}_{-j}$ is matrix $\tilde{X}$ with row $j$ deleted. The possible combinations that $S$ may be calculated using, as outlined above, may have an effect on the robustness of the imputations calculated. If $S$ is calculated using the $n \times p$ matrix $X$, it is computationally efficient, however may result in some bias in the Mahalanobis distance calculated. This bias could be eliminated by calculating $S$ using $(n - 1) \times p$ data matrix $X_{-i}$ however it would require more computing power to do this.

Operations on how $Z^{mis}$ are imputed are not detailed at any point on either the SOLAS webpage or in any published documents. It is unclear if all variables in $Z^{mis}$ are imputed jointly or on a variable by variable case with an imputed variable being used in the imputation of the next variable.

### 3.2.1 Modifications to Mahalanobis Distance Matching

An alternative method of defining the donor pool may be to assign the vectors $z_{i,obs}^j$, for each value of $j$ in the donor pool the sampling weights
\[
w_{ij} = \frac{p_{ij}}{\sum_{i=1}^n p_{ij}}
\]
where
\[
p_{ij} = 1 - Pr(D^2 > \chi^2_d)
\]
where $d$ is the number of columns in the matrix $X$ as obtained from using the partition function in Algorithm 1, presented in section 3.3 and $D^2$ is the calculated
Mahalanobis distance in question. The probability of getting a random number greater than the Mahalanobis distance calculated between two units, given that the Mahalanobis distances in the general case have a $\chi^2_d$ distribution, is defined as $p_{ij}$. Intuitively the weights used are then a measure of dissimilarity rather than a measure of similarity.

These weights are then used to draw a proportional random sample of size $K$ from the donor pool. This procedure is not implemented in SOLAS and guidelines about selecting the value of $K$ are not available but mentioned in demonstrative videos that values are selected at random. Intuitively, this proposed method should capture a greater level of inherent variability in the tails as well as the centre of the population being sampled. Extending this idea, rather than limiting the value of $K$, to obtain a better level of variability, all cases should be included in the donor pool.

3.2.2 Shortcomings and Benefits

When considering both the SOLAS procedure and the proposed modifications, there are a number of shortcomings that the methods have while there are also a number of benefits to them.

Shortcomings

1. If a variable is not predictive of another variable containing a missing value is included in the calculation of the Mahalanobis distance it may lead to the Mahalanobis distance metric being dominated by that variable.

2. As only previously observed values are chosen for the imputed values, when data are MNAR in particular, the true value of parameters may be grossly under/over estimated.

3. As the modified method uses all fully observed data as its donor pool and selects from this donor pool using a probability weight, there is likely to be more variability in the parameters estimated.
Benefits

1. Impossible values will never be chosen as replacements for missing values as the imputed values are all observed in the dataset previously.

2. As the modified method uses all fully observed data as its donor pool and selects from this donor pool using a probability weight, the imputed values capture more of the uncertainty around what the missing values may be.

3. Does not require huge computing power compared to other methods of dealing with missing data.

3.3 Algorithms

The following algorithms are a detailed explanation of how both the SOLAS procedure and the proposed procedure outlined in Section 3.2.1 are implemented in R (R Development Core Team, 2008). The R code corresponding with each of Algorithms 1-5 can be found in Appendix B.

Algorithm 1 separates a matrix with missing data into the format shown in the multivariate data layout described in section 3.2. The partitioning of the data containing missing values is the first step in allowing the calculation of Mahalanobis distances between units of data in a matrix.

The second step in the process is described in Algorithm 2. Here the Mahalanobis distances are calculated from the observed variables of units with missing values to the corresponding variables in units that are fully observed. Once the Mahalanobis distances are calculated, they are then turned into the dissimilarity measure described in Equation 3.7.

Upon obtaining the weights, these are then used for the selection of values used in generating the multiply imputed datasets. The probability of selecting a particular value is determined by Equation 3.6 and the implementation of this is described
in Algorithm 3.

To implement the process carried out in SOLAS, Algorithm 1 is still used as above. The Mahalanobis distances are calculated as below in Algorithm 4 and the imputed values are selected with equal probability as described in Algorithm 5.

**Algorithm 1** Partition of matrix containing missing values

```plaintext
function Partition(Y)
    rows ← complete.cases(Y)
    columns ← complete.cases(Y')
    X ← Y[rows, columns]
    X̂ ← Y[!rows, columns]
    Zobs ← Y[rows, !columns]
    Zmis ← Y[!rows,!columns]
return (X, X̂, Zobs, Zmis)
end function
```

**Algorithm 2** Calculate the Mahalanobis distances from observations with missing values to those fully observed

```plaintext
function Calc.MDs(Y)
    Q ← partition(Y)
    S^{-1} ← (var(X))^{-1}
    m ← number of rows in \( \hat{X} \)
    if m=1
        for i ← 1 to number of rows of X
            MDs ← \sqrt{(x_i - \hat{x}_1)'S^{-1}(x_i - \hat{x}_1)}
    else if m>1
        for i ← 1 to number of rows of X
            MDs ← \sqrt{(x_i - \hat{x}_1)'S^{-1}(x_i - \hat{x}_1)}
        for j ← 2 to m
            for i ← 1 to number of rows of X
                MDs = cbind(MDs,MDs ← \sqrt{(x_i - \hat{x}_j)'S^{-1}(x_i - \hat{x}_j)})
    MDs = MDs'
    MDsProb ← 1 - pchisq(MDs, df = number of columns of X)
    comment : MDsProb is the probability of getting a random number greater than the MDs value inputted given that the Mahalanobis Distances have a Chi-squared distribution with degrees of freedom equal to the number of columns of X.
return (MDsProb)
end function
```
Algorithm 3 Create imputed values based on values obtained using CALC.MDs

function IMPMDSIM1B(Y, m = number of fully imputed datasets)
    M ← is.na(Y) × 1
    N ← unique combinations of M
    Q ← PARTITION(Y)
    ZobsX ← cbind(X, Zobs)
    ZmisXtilde ← cbind(X, Zmis)
    MXZ ← is.na(ZmisXtilde) × 1
    NXZ ← unique combinations of MXZ
    n ← number of rows in NXZ
    a ← number of rows in MXZ
    ZXnew ← XobsX
    ans = numeric(0)
    for i ← 1 to n
        ZXnew ← ZobsX
        for j ← 1 to a
            if MXZ[j,] == NXZ[i,]
                ZXnew = rbind(ZXnew, ZmisXtilde[j,])
        ans = cbind(ans, calc.MDs(ZXnew))
    f = positions of fully observed units in Y
    f ← cbind(f, c(1: number of rows in f))
    s = numeric(0)
    for i ← 1 to (number of columns of ans)
        s = cbind(s, sample(1 : (number of columns of ans),
                       m, replace = TRUE, prob = ans[, i]))
    h = numeric(0)
    for i ← 1 to (number of columns of ans)
        h = cbind(h, f[c[s[, i]], 1])
    miscols = positions of columns with missing data in Y
    y.new = Y
    o = an empty array with c columns, r rows and m dimensions
    k = positions of units which missing values in Y
    for j ← 1 to m
        for i ← 1 to number of columns in ans
            miss = what columns have missing values in Y[[k[i]],]
            y.new[k[i], c(miss)] = Y[h[, i]][j,][c(miss)]
    o[, , j] = y.new
    return (o)
end function
Algorithm 4 Calculate the Mahalanobis distances from observations with missing values to those fully observed as in the SOLAS software package

```r
function CALC.MDSOLAS(Y)
    Q ← \text{PARTITION}(Y)
    S^{-1} ← (\text{var}(X))^{-1}
    m ← \text{number of rows in } \hat{X}
    \text{if } m=1
        \text{for } i ← 1 \text{ to number of rows of } X
            MDs ← \sqrt{(x_i - \hat{x}_1)'S^{-1}(x_i - \hat{x}_1)}
    \text{else if } m>1
        \text{for } i ← 1 \text{ to number of rows of } X
            MDs ← \sqrt{(x_i - \hat{x}_1)'S^{-1}(x_i - \hat{x}_1)}
        \text{for } j ← 2 \text{ to } m
            \text{for } i ← 1 \text{ to number of rows of } X
                MDs = cbind(MDs, MDs ← \sqrt{(x_i - \hat{x}_j)'S^{-1}(x_i - \hat{x}_j)})
        MDs = MDs'
end function
```

45
Algorithm 5 Create imputed values based on values obtained using CALC.MDs

```r
function impMDsim1a(Y, m = number of fully imputed datasets)
  M ← is.na(Y) × 1
  N ← unique combinations of M
  Q ← partition(Y)
  ZobsX ← cbind(X, Zobs)
  ZmisXtilde ← cbind(X, Zmis)
  MXZ ← is.na(ZmisXtilde) × 1
  NXZ ← unique combinations of MXZ
  n ← number of rows in NXZ
  a ← number of rows in MXZ
  ZXnew ← ZobsX
  ans = numeric(0)
  for i ← 1 to n
    ZXnew ← XobsX
    for j ← 1 to a
      if MXZ[j,] == NXZ[i,]
        ZXnew = rbind(ZXnew, ZmisXtilde[j,])
    end for
    ans = cbind(ans, calc.MDsolas(ZXnew))
  end for
  ord = numeric(0)
  for i ← 1 to (number of columns of ans)
    ord = cbind(ord, order(ans[i,]))
  end for
  comment : order each column of ans from lowest to highest
  g = first m × 2 rows of each column of ans
  f = positions of fully observed units in Y
  h = numeric(0)
  for i ← 1 to (number of columns of ord)
    h = cbind(h, f[order(g[i,])])
  end for
  s = numeric(0)
  for i ← 1 to (number of columns of ord)
    s = cbind(s, sample(1 : (m × 2, m, replace = TRUE)))
  end for
  miscols = positions of columns with missing data in Y
  y.new = Y
  o = an empty array with c columns, r rows and m dimensions
  k = positions of units which missing values in Y
  for j ← 1 to m
    for i ← 1 to number of columns in ord
      miss = what columns have missing values in Y[[k[i]],]
      y.new[k[i], c(miss)] = Y[h[i], ][s[j, i], ][c(miss)]
    end for
    o[, j] = y.new
  end for
  return (o)
end function
```


Chapter 4

Application of proposed algorithm

To test the proposed adjustment in Section 3.2 to the SOLAS algorithm, it is necessary to compare results achieved from imputation with the standard method used. To do this, data will be simulated and imputation carried out on each dataset using each of the two methods and the results compared. The results are combined using Rubin’s rules outlined in Equations 2.4 to 2.7. To see the impact of each method on various types of missing data mechanisms, the simulated data had missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) mechanisms imposed on it with rates of missingness varying from 10% to 50%. Both methods are also implemented on a real biomechanics dataset and the results of each method compared.

4.1 Simulation Study

Similar to the set up of the simulation study in Collins et al. (2001), independent random samples of size $n = 200$, $n = 500$ and $n = 1000$ are drawn from a multivariate normal distribution with the properties outlined in Appendix A.1. Each dataset has three variables, $X$ which is always observed, $Y$ which can be either observed or missing and $Z$ which is also observed and may be a potential cause for missingness in $Y$ when referring to the MAR missing data mechanism.

To simulate the data, the following parameters were arbitrarily chosen and spec-
ified in the model, $\mu_X = 5$, $\sigma_X = 1$, $\mu_Y = 5.2$ and $\sigma_Y = 1.44$. $Z$ which may be the cause of missingness is set to have $\mu_Z = 0$ and $\sigma_Z = 1$. The correlation between the combinations of variables are set to $\rho_{XY} = 0.6$, $\rho_{XZ} = 0.54$ and $\rho_{YZ} = 0.9$.

To impose the mechanisms of missing data, MCAR meant that missing data was imposed on variable $Y$ independently of $X$, $Y$ or $Z$ at the various rates of missing data. To implement the MAR mechanism, the probability of $Y$ being missing depended on the value of $Z$. This has been implemented by dividing the distribution of the $Z$ variable into quartiles and imposing different rates of missing data onto each of the corresponding $Y$ values. To impose a rate of missing data on $Y$ overall of 10%, and assuming that $Y$ is more likely to be missing the larger the value of $Z$ is, the probability that a $Y$ value is missing in each of the four quartiles is (0.04,0.08,0.12,0.16). To impose a rate of missing data of 25% on $Y$, the probability of a $Y$ value being missing was given a probability in each of the four quartiles of (0.1,0.2,0.3,0.4) and for a missing data rate of 50%, the probabilities were (0.2,0.4,0.6,0.8). To implement the MNAR mechanism, the probability of a value being missing depends on the underlying value itself. This time the $Y$ variable is sectioned into quartiles and the same proportions of missing values are applied as those for the MAR mechanism.

### 4.1.1 Evaluation Criteria

To compare the results of both methods, a number of evaluation criteria are used. These evaluation methods are commonly used in comparing methods used for imputing missing data in numerous papers (Collins et al., 2001; Schafer and Graham, 2002; Zhu, 2014; Schmitt et al., 2015). The first evaluation criteria used is the standardised bias (SB). Bias is the difference between the parameter estimate and the true parameter. The standardised bias is described as

$$ SB = 100 \times \frac{(\text{average estimate} - \text{parameter})}{SE} $$

where $SE$ is the standard deviation of the estimates. The SB is reported as a per-
centage for interpretability because all parameters are not on the same scale. For example, a standardised bias value of 50% means that the estimate falls, on average, half a standard deviation above the true parameter value. Negative values indicate that the estimate falls on average below the true parameter value.

The second evaluation criteria that will be used is the root mean square error (RMSE). This is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\text{average estimate} - \text{parameter})^2}{n}}$$

where \( n \) is the number of simulations carried out. Root mean square error is on the same scale as the parameter in question.

The third measurement that the two methods will be compared on is coverage. Coverage is defined as the proportion of simulated datasets that return a 95% confidence interval containing the true parameter value. Coverage values lower than 90% are considered poor. On top of the 95% confidence interval, the average length of these confidence intervals will also be reported. When comparing two methods, if the rate of coverage is similar for both, then the method with the shorter confidence interval is preferable. Shorter intervals indicate better accuracy and higher power.

The parameters of interest that will be evaluated are \( \mu_Y, \sigma_Y^2, \beta_{X,Y}, \beta_{Y,X}, \rho_{XY} \) and \( \rho_{YZ} \). \( \beta_{X,Y} \) is the regression parameter for the regression of \( X \) on \( Y \) which represents a completely observed dependent variable and an incomplete predictor. \( \beta_{Y,X} \) is the regression parameter for the regression of \( Y \) on \( X \) which represents a completely observed predictor variable and an incomplete dependent variable. The various parameters are estimated and combined across the multiple imputed datasets using Rubin’s rules and as recommended in Marshall et al. (2009). In Tables 4.1 - 4.9, “A” refers to the results obtained from the process currently carried out within SOLAS and described in Algorithms 4 and 5 and “B” refers to the results obtained using the proposed method, described in Algorithms 2 and 3.
4.1.2 Missing Completely at Random

Table 4.1 below contains the evaluation criteria for MCAR data, with \( n = 200 \), simulated 1000 times. In estimating \( \mu_Y \), method B (the proposed method) outperforms or performs as well as method A (the SOLAS method) regardless of the rate of missing data. However, when looking at the other parameters of interest, method A generally gives better results than method B particularly as the rate of missing data increases. It is interesting to note that for \( \rho_{XY} \) and \( \rho_{YX} \), regardless of the rate of missing values, neither method performs well with coverage values falling below the recommended 90% in all cases. For the regression parameter values, method A performs well for \( \beta_{XY} \), which has a fully observed dependent variable. Method B performs reasonably well, albeit with higher SB values than method A, until the rate of missing values increases to 50%. For \( \beta_{YX} \), which has a completely observed predictor variable, method A performs well in terms of coverage rates until the rate of missing data reaches 50%. The SB values for this parameter are all negative regardless of the method indicating that both methods are underestimating the true parameter value.

When the size of the dataset increases to \( n = 500 \) in Table 4.2, similar results can be observed. For parameter \( \mu_Y \), method B outperforms or performs as well as method A. For \( \beta_{YX} \), method A outperforms method B for 10% and 25% missing values but along with method B performs poorly once the rate of missing data reaches 50%. For \( \beta_{XY} \), method A outperforms method B across the three missing data rates. For \( \rho_{XY} \), method A performs better than B for 25% and 50% missing data. In terms of coverage, method A and B perform equally well when the rate of missing data is at 10% but SB is larger for method B than A.

For \( n = 1000 \), the results can be seen in Table 4.3. Method B again performs better than method A when estimating the parameter \( \mu_Y \) but method A is better for other parameters especially when the rate of missing data increases. Method A performs well for \( \beta_{YX} \) when the rate of missing data is 25% or less, for \( \beta_{XY} \) and \( \rho_{XY} \) for all rates of missing data, and for \( \rho_{YX} \) when the rate of missing data is
10%. Method A performs poorly for $\rho_{YX}$ with rates of missing values of 25% or 50%. Method B performs badly for both correlation parameters and when rates of missing data exceed 25% for both regression parameters.
<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_Y$ Standardised bias</td>
<td>-3.088</td>
<td>-3.016</td>
<td>2.28</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.101</td>
<td>0.102</td>
<td>0.109</td>
</tr>
<tr>
<td>Coverage</td>
<td>95</td>
<td>95.3</td>
<td>92.8</td>
</tr>
<tr>
<td>Width</td>
<td>0.40</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>$\sigma^2_Y$ Standardised bias</td>
<td>-17.65</td>
<td>-21.55</td>
<td>-52.38</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.216</td>
<td>0.217</td>
<td>0.237</td>
</tr>
<tr>
<td>$\beta_{YX}$ Standardised bias</td>
<td>-20.71</td>
<td>-56.79</td>
<td>-51.76</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.084</td>
<td>0.092</td>
<td>0.094</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.5</td>
<td>93.7</td>
<td>92.9</td>
</tr>
<tr>
<td>Width</td>
<td>0.329</td>
<td>0.342</td>
<td>0.340</td>
</tr>
<tr>
<td>$\beta_{XY}$ Standardised bias</td>
<td>2.72</td>
<td>-27.66</td>
<td>-3.55</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.6</td>
<td>95.3</td>
<td>95.5</td>
</tr>
<tr>
<td>Width</td>
<td>0.161</td>
<td>0.169</td>
<td>0.169</td>
</tr>
<tr>
<td>$\rho_{XY}$ Standardised bias</td>
<td>-13.07</td>
<td>-54.34</td>
<td>-36.41</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Coverage</td>
<td>62.1</td>
<td>59.9</td>
<td>57.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td>$\rho_{YX}$ Standardised bias</td>
<td>-50.48</td>
<td>-232.22</td>
<td>-129.81</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.017</td>
<td>0.040</td>
<td>0.027</td>
</tr>
<tr>
<td>Coverage</td>
<td>55.7</td>
<td>5.5</td>
<td>28.7</td>
</tr>
<tr>
<td>Width</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 4.1: Evaluation criteria for MCAR data, $n = 200$
<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>( \mu_Y )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>0.596</td>
<td>0.378</td>
<td>-0.377</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.066</td>
<td>0.067</td>
<td>0.069</td>
</tr>
<tr>
<td>Coverage</td>
<td>99.7</td>
<td>99.7</td>
<td>99.5</td>
</tr>
<tr>
<td>Width</td>
<td>0.400</td>
<td>0.402</td>
<td>0.398</td>
</tr>
<tr>
<td>( \sigma_{XY}^2 )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-10.075</td>
<td>-26.33</td>
<td>-49.475</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.133</td>
<td>0.135</td>
<td>0.153</td>
</tr>
<tr>
<td>( \beta_{YX} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-13.024</td>
<td>-82.38</td>
<td>-44.245</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.05</td>
<td>0.065</td>
<td>0.059</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.4</td>
<td>90.3</td>
<td>92.3</td>
</tr>
<tr>
<td>Width</td>
<td>0.206</td>
<td>0.216</td>
<td>0.212</td>
</tr>
<tr>
<td>( \beta_{XY} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>4.184</td>
<td>-46.079</td>
<td>3.003</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.026</td>
<td>0.028</td>
<td>0.026</td>
</tr>
<tr>
<td>Coverage</td>
<td>94</td>
<td>94</td>
<td>96.1</td>
</tr>
<tr>
<td>Width</td>
<td>0.100</td>
<td>0.106</td>
<td>0.104</td>
</tr>
<tr>
<td>( \rho_{XY} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-6.783</td>
<td>-81.024</td>
<td>-27.784</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.029</td>
<td>0.037</td>
<td>0.031</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.7</td>
<td>90</td>
<td>93.9</td>
</tr>
<tr>
<td>Width</td>
<td>0.113</td>
<td>0.117</td>
<td>0.114</td>
</tr>
<tr>
<td>( \rho_{YX} )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-32.499</td>
<td>-358.46</td>
<td>-105.73</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.099</td>
<td>0.037</td>
<td>0.015</td>
</tr>
<tr>
<td>Coverage</td>
<td>92.9</td>
<td>3.2</td>
<td>79.1</td>
</tr>
<tr>
<td>Width</td>
<td>0.034</td>
<td>0.044</td>
<td>0.037</td>
</tr>
</tbody>
</table>

Table 4.2: Evaluation criteria for MCAR data, \( n = 500 \)
## Rate of Missingness

<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>3.046</td>
<td>2.935</td>
<td>-0.258</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.045</td>
<td>0.046</td>
<td>0.047</td>
</tr>
<tr>
<td>Coverage</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Width</td>
<td>0.399</td>
<td>0.398</td>
<td>0.398</td>
</tr>
<tr>
<td>$\sigma_Y^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-13.043</td>
<td>-44.408</td>
<td>-43.924</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.094</td>
<td>0.103</td>
<td>0.104</td>
</tr>
<tr>
<td>$\beta_{YX}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.036</td>
<td>0.055</td>
<td>0.041</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.7</td>
<td>83.4</td>
<td>92.8</td>
</tr>
<tr>
<td>Width</td>
<td>0.145</td>
<td>0.152</td>
<td>0.149</td>
</tr>
<tr>
<td>$\beta_{XY}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>5.711</td>
<td>-67.995</td>
<td>5.844</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.018</td>
<td>0.021</td>
<td>0.018</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.7</td>
<td>91.8</td>
<td>95.1</td>
</tr>
<tr>
<td>Width</td>
<td>0.07</td>
<td>0.075</td>
<td>0.073</td>
</tr>
<tr>
<td>$\rho_{YX}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-2.661</td>
<td>-117.86</td>
<td>-22.496</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.02</td>
<td>0.03</td>
<td>0.022</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.9</td>
<td>81.6</td>
<td>93.8</td>
</tr>
<tr>
<td>Width</td>
<td>0.079</td>
<td>0.083</td>
<td>0.08</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-39.375</td>
<td>-534.54</td>
<td>-89.033</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.007</td>
<td>0.037</td>
<td>0.009</td>
</tr>
<tr>
<td>Coverage</td>
<td>93.1</td>
<td>0</td>
<td>82.9</td>
</tr>
<tr>
<td>Width</td>
<td>0.024</td>
<td>0.03</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Table 4.3: Evaluation criteria for MCAR data, $n = 1000$
4.1.3 Missing at Random

Moving on to MAR data, when \( n = 200 \) in Table 4.4, for \( \mu_Y \), method A performs better than method B when the rate of missing data is 10% or 25%. Both methods perform poorly when 50% of the data is missing. The same trend can be seen for \( \beta_{YX} \). For \( \beta_{XY} \), both methods perform well in terms of coverage but the SB values for method A would suggest that this is a better method at estimating this parameter. For \( \rho_{XY} \), method A performs better than method B when the rate of missing data is 10% or 25%. Both methods perform poorly when 50% of the data is missing. Considering \( \rho_{YX} \), method A for 10% missing data is the only case where good results are seen.

Upon increasing the sample size to \( n = 500 \) in Table 4.5, good rates of coverage are obtained for estimating \( \mu_Y \) under both methods when the rate of missing data is 10% or 25% however method A has better SB values than method B. Method B performs poorly when the rate of missing values increases to 50%. For \( \beta_{YX} \), method A performs well for 10% and 25% missing data but poorly when the rate of missing data is 50%. Method B only performs well with 10% of data missing and in that case the SB values are better for method A. For \( \beta_{XY} \), method A has coverage rates greater than 90% for all three rates of missing values while method B again only reaches this threshold for a missing data rate of 10%. For the correlation parameters, method B is unreliable to use in any circumstance while method A performs well for both parameters when the rate of missing data is 10% and for \( \rho_{XY} \) when the rate of missing data is 25%.

When \( n = 1000 \) in Table 4.6, \( \mu_Y \) is well estimated by both methods for rates of missing data of 10% and 25%. However method B performs poorly when there is 50% missing data. For the regression parameters, method B does not perform well in any case. Method A does well for \( \beta_{YX} \) with 10% and 25% missing values and in all cases for \( \beta_{XY} \). For the correlation parameters, method B is unreliable to use in any circumstance while method A performs well for both parameters when the rate
of missing data is 10% and for $\rho_{XY}$ when the rate of missing data is 25%.

<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>Rate of Missingness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10% missing</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td>A</td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-8.233</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.107</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.1</td>
</tr>
<tr>
<td>Width</td>
<td>0.402</td>
</tr>
<tr>
<td>$\sigma_Y^2$</td>
<td>-17.894</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.218</td>
</tr>
<tr>
<td>$\beta_{XY}$</td>
<td>-16.858</td>
</tr>
<tr>
<td>Standardised bias</td>
<td>0.084</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.04</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.329</td>
</tr>
<tr>
<td>$\beta_{X Y}$</td>
<td>3.62</td>
</tr>
<tr>
<td>Standardised bias</td>
<td>0.04</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.044</td>
</tr>
<tr>
<td>Coverage</td>
<td>96.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.16</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td>-10.597</td>
</tr>
<tr>
<td>Standardised bias</td>
<td>0.046</td>
</tr>
<tr>
<td>RMSE</td>
<td>95</td>
</tr>
<tr>
<td>Coverage</td>
<td>0.18</td>
</tr>
<tr>
<td>Width</td>
<td>0.058</td>
</tr>
</tbody>
</table>

Table 4.4: Evaluation criteria for MAR data, $n = 200$
<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_Y$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-3.428</td>
<td>-45.06</td>
<td>-19.91</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.064</td>
<td>0.07</td>
<td>0.068</td>
</tr>
<tr>
<td>Coverage</td>
<td>99.8</td>
<td>99.8</td>
<td>99.8</td>
</tr>
<tr>
<td>Width</td>
<td>0.4</td>
<td>0.402</td>
<td>0.4</td>
</tr>
<tr>
<td>$\sigma_Y^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-13.1</td>
<td>-29.79</td>
<td>-43.06</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.137</td>
<td>0.142</td>
<td>0.153</td>
</tr>
<tr>
<td>$\beta_{YX}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-12.304</td>
<td>-82.273</td>
<td>-42.99</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.053</td>
<td>0.067</td>
<td>0.058</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.3</td>
<td>90</td>
<td>92.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.206</td>
<td>0.216</td>
<td>0.211</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.026</td>
<td>0.028</td>
<td>0.027</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.4</td>
<td>94.5</td>
<td>94.7</td>
</tr>
<tr>
<td>Width</td>
<td>0.100</td>
<td>0.106</td>
<td>0.104</td>
</tr>
<tr>
<td>$\rho_{YX}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-7.199</td>
<td>-81.76</td>
<td>-22.216</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.03</td>
<td>0.038</td>
<td>0.031</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.9</td>
<td>88.1</td>
<td>93.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.113</td>
<td>0.117</td>
<td>0.114</td>
</tr>
</tbody>
</table>

Table 4.5: Evaluation criteria for MAR data, $n = 500$
<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>Rate of Missingness</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-8.17</td>
<td>-66.80</td>
<td>-14.52</td>
<td>-180.44</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.047</td>
<td>0.056</td>
<td>0.05</td>
<td>0.102</td>
</tr>
<tr>
<td>Coverage</td>
<td>100</td>
<td>99.9</td>
<td>99.9</td>
<td>98.3</td>
</tr>
<tr>
<td>Width</td>
<td>0.399</td>
<td>0.398</td>
<td>0.397</td>
<td>0.39</td>
</tr>
<tr>
<td>$\sigma^2_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-14.86</td>
<td>-48.76</td>
<td>-49.37</td>
<td>-130.75</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.094</td>
<td>0.102</td>
<td>0.110</td>
<td>0.161</td>
</tr>
<tr>
<td>$\beta_{YX}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-11.38</td>
<td>-118.84</td>
<td>-47.64</td>
<td>-316.67</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.038</td>
<td>0.057</td>
<td>0.043</td>
<td>0.121</td>
</tr>
<tr>
<td>Coverage</td>
<td>94</td>
<td>81.9</td>
<td>92.5</td>
<td>20</td>
</tr>
<tr>
<td>Width</td>
<td>0.145</td>
<td>0.152</td>
<td>0.149</td>
<td>0.164</td>
</tr>
<tr>
<td>$\beta_{XY}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-0.188</td>
<td>-70.17</td>
<td>7.214</td>
<td>-174.84</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.019</td>
<td>0.023</td>
<td>0.019</td>
<td>0.036</td>
</tr>
<tr>
<td>Coverage</td>
<td>93.6</td>
<td>88.5</td>
<td>95.1</td>
<td>71</td>
</tr>
<tr>
<td>Width</td>
<td>0.07</td>
<td>0.075</td>
<td>0.073</td>
<td>0.084</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-7.94</td>
<td>-116.68</td>
<td>-25.86</td>
<td>-299.31</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.022</td>
<td>0.032</td>
<td>0.022</td>
<td>0.067</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.3</td>
<td>77.7</td>
<td>92.3</td>
<td>14.8</td>
</tr>
<tr>
<td>Width</td>
<td>0.08</td>
<td>0.083</td>
<td>0.08</td>
<td>0.088</td>
</tr>
<tr>
<td>$\rho_{YX}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-31.16</td>
<td>-535.54</td>
<td>-91.89</td>
<td>-1188.18</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.007</td>
<td>0.037</td>
<td>0.01</td>
<td>0.094</td>
</tr>
<tr>
<td>Coverage</td>
<td>93.1</td>
<td>0</td>
<td>80.2</td>
<td>0</td>
</tr>
<tr>
<td>Width</td>
<td>0.024</td>
<td>0.032</td>
<td>0.025</td>
<td>0.044</td>
</tr>
</tbody>
</table>

Table 4.6: Evaluation criteria for MAR data, $n = 1000$
4.1.4 Missing Not at Random

The final case to consider was MNAR. Table 4.7 contains simulated MNAR data when the sample size $n = 200$. In estimating $\mu_Y$, both methods perform well when the rate of missing data is 10%. Only, method A performs well when the rate of missing data is 25% and neither methods perform well when the rate of missing data is 50%. For $\beta_{XY}$, both methods perform well with coverage rates greater than 90% for all rates of missing values but method A has better SB values. For $\beta_{YX}$, method A performs reasonably well for rates of missing data of 10% and 25% and method B performs well for 10% missing values but poorly for the other rates. For the correlation parameters, method B is unreliable to use in any circumstance for $\rho_{YX}$ while method A performs well when the rate of missing data is 10%. Method A performs well when the rate of missing data is 10% and 25% for $\rho_{XY}$ while method B only performs well for this parameter when the rate of missing data is 10%.

When $n = 500$ for MNAR data in Table 4.8, both methods perform well when the rate of missing values is 10%. However, only method A performs well when the rate of missing data is 25% and neither method performs well when the rate of missing data is 50%. For $\beta_{XY}$, method A performs well for all rates of missing data and for $\beta_{YX}$, performs well for missing data rates of 10% and 25%. Method B is unreliable under all three rates for parameter $\beta_{YX}$ and only performs well for a rate of missing data of 10% for $\beta_{XY}$. For the correlation parameters, method B is unreliable to use in any circumstance while method A performs well when the rate of missing data is 10% or 25% for $\rho_{XY}$, and 10% for $\rho_{YX}$.

When $n = 1000$, both methods estimate $\mu_Y$ well when the rate of missing data is 10% or 25% but poorly for 50% missing data. The SB values suggest that method A is more accurate in these cases. For $\beta_{YX}$ and $\beta_{XY}$, method B performs poorly in all circumstances while method A performs well for both parameters when the rate of missing data is either 10% or 25% but poorly for a rate of 50%. For $\rho_{XY}$ and $\rho_{YX}$, method B is unreliable to use in any circumstance while method A performs.
well when the rate of missing data is 10% and 25% for $\rho_{XY}$ and 10% for $\rho_{YX}$.

<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>Rate of Missingness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10% missing</td>
<td>25% missing</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-18.59</td>
<td>-36.20</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.112</td>
<td>0.117</td>
</tr>
<tr>
<td>Coverage</td>
<td>92</td>
<td>91.6</td>
</tr>
<tr>
<td>Width</td>
<td>0.402</td>
<td>0.411</td>
</tr>
<tr>
<td>$\sigma^2_Y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-19.04</td>
<td>-22.69</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.211</td>
<td>0.212</td>
</tr>
<tr>
<td>$\beta_{YX}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMSE</td>
<td>0.083</td>
<td>0.091</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.1</td>
<td>93.7</td>
</tr>
<tr>
<td>Width</td>
<td>0.329</td>
<td>0.34</td>
</tr>
<tr>
<td>$\beta_{XY}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>1.55</td>
<td>-31.28</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.039</td>
<td>0.04</td>
</tr>
<tr>
<td>Coverage</td>
<td>96.2</td>
<td>95.5</td>
</tr>
<tr>
<td>Width</td>
<td>0.16</td>
<td>0.168</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-10.02</td>
<td>-52.99</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.045</td>
<td>0.051</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.4</td>
<td>94.6</td>
</tr>
<tr>
<td>Width</td>
<td>0.179</td>
<td>0.186</td>
</tr>
<tr>
<td>$\rho_{YX}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-53.69</td>
<td>-238.6</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.017</td>
<td>0.04</td>
</tr>
<tr>
<td>Coverage</td>
<td>92.1</td>
<td>34.3</td>
</tr>
<tr>
<td>Width</td>
<td>0.057</td>
<td>0.072</td>
</tr>
</tbody>
</table>

Table 4.7: Evaluation criteria for MNAR data, $n = 200$
<table>
<thead>
<tr>
<th>Parameter Estimated</th>
<th>Rate of Missingness</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-23.89</td>
<td>-60.4</td>
<td>-78.06</td>
<td>-182.31</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.068</td>
<td>0.077</td>
<td>0.088</td>
<td>0.143</td>
</tr>
<tr>
<td>Coverage</td>
<td>99.9</td>
<td>99.3</td>
<td>97.7</td>
<td>86.5</td>
</tr>
<tr>
<td>Width</td>
<td>0.399</td>
<td>0.400</td>
<td>0.398</td>
<td>0.402</td>
</tr>
<tr>
<td>$\sigma^2_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-21.47</td>
<td>-40.30</td>
<td>-50.77</td>
<td>-95.65</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.132</td>
<td>0.138</td>
<td>0.148</td>
<td>0.179</td>
</tr>
<tr>
<td>$\beta_{YX}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-19.84</td>
<td>-93</td>
<td>-46.61</td>
<td>-277.48</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.052</td>
<td>0.068</td>
<td>0.059</td>
<td>0.125</td>
</tr>
<tr>
<td>Coverage</td>
<td>96</td>
<td>89.6</td>
<td>93</td>
<td>51.9</td>
</tr>
<tr>
<td>Width</td>
<td>0.206</td>
<td>0.216</td>
<td>0.211</td>
<td>0.233</td>
</tr>
<tr>
<td>$\beta_{XY}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>0.821</td>
<td>-51.91</td>
<td>3.54</td>
<td>-130.41</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.025</td>
<td>0.027</td>
<td>0.027</td>
<td>0.041</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.6</td>
<td>94.8</td>
<td>93.8</td>
<td>84</td>
</tr>
<tr>
<td>Width</td>
<td>0.100</td>
<td>0.106</td>
<td>0.104</td>
<td>0.119</td>
</tr>
<tr>
<td>$\rho_{XY}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-12.48</td>
<td>-88.24</td>
<td>-27.32</td>
<td>-218.23</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.03</td>
<td>0.038</td>
<td>0.032</td>
<td>0.07</td>
</tr>
<tr>
<td>Coverage</td>
<td>95.4</td>
<td>87.7</td>
<td>92.2</td>
<td>44</td>
</tr>
<tr>
<td>Width</td>
<td>0.113</td>
<td>0.118</td>
<td>0.114</td>
<td>0.125</td>
</tr>
<tr>
<td>$\rho_{YX}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standardised bias</td>
<td>-41.98</td>
<td>-317.46</td>
<td>-109.89</td>
<td>-837.25</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.01</td>
<td>0.038</td>
<td>0.015</td>
<td>0.094</td>
</tr>
<tr>
<td>Coverage</td>
<td>92.3</td>
<td>1.8</td>
<td>77.8</td>
<td>0</td>
</tr>
<tr>
<td>Width</td>
<td>0.035</td>
<td>0.045</td>
<td>0.037</td>
<td>0.062</td>
</tr>
</tbody>
</table>

Table 4.8: Evaluation criteria for MNAR data, n = 500
<table>
<thead>
<tr>
<th>Rate of Missingness</th>
<th>10% missing</th>
<th>25% missing</th>
<th>50% missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter Estimated</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$\mu_Y$ Standardised bias</td>
<td>-30.65</td>
<td>-83.37</td>
<td>-105.21</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.049</td>
<td>0.062</td>
<td>0.066</td>
</tr>
<tr>
<td>Coverage</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Width</td>
<td>0.399</td>
<td>0.398</td>
<td>0.396</td>
</tr>
<tr>
<td>$\sigma_Y^2$ Standardised bias</td>
<td>-10.33</td>
<td>-43.16</td>
<td>-54.52</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.094</td>
<td>0.101</td>
<td>0.111</td>
</tr>
<tr>
<td>$\beta_{YX}$ Standardised bias</td>
<td>-11.34</td>
<td>-119.07</td>
<td>-46.44</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.038</td>
<td>0.057</td>
<td>0.04</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.5</td>
<td>81.3</td>
<td>93.6</td>
</tr>
<tr>
<td>Width</td>
<td>0.146</td>
<td>0.152</td>
<td>0.149</td>
</tr>
<tr>
<td>$\beta_{XY}$ Standardised bias</td>
<td>-1.72</td>
<td>-73.39</td>
<td>14.15</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.018</td>
<td>0.022</td>
<td>0.019</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.3</td>
<td>89.4</td>
<td>94.1</td>
</tr>
<tr>
<td>Width</td>
<td>0.07</td>
<td>0.075</td>
<td>0.073</td>
</tr>
<tr>
<td>$\rho_{XY}$ Standardised bias</td>
<td>-9.07</td>
<td>-119.96</td>
<td>-19.70</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.021</td>
<td>0.032</td>
<td>0.021</td>
</tr>
<tr>
<td>Coverage</td>
<td>94.4</td>
<td>78.4</td>
<td>94.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.08</td>
<td>0.083</td>
<td>0.08</td>
</tr>
<tr>
<td>$\rho_{YX}$ Standardised bias</td>
<td>-28.88</td>
<td>-531.85</td>
<td>-97.57</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.007</td>
<td>0.037</td>
<td>0.009</td>
</tr>
<tr>
<td>Coverage</td>
<td>93.2</td>
<td>0</td>
<td>82.2</td>
</tr>
<tr>
<td>Width</td>
<td>0.024</td>
<td>0.03</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Table 4.9: Evaluation criteria for MNAR data, $n = 1000$
4.1.5 Conclusions on Simulated Data

1. The proposed method, B, performs well for estimating the values of parameter $\mu_Y$, especially when the data is MCAR, outperforming method A.

2. When the rate of missing data increases to 50%, for parameters other than $\mu_Y$, both methods struggle to produce satisfactory results regardless of the mechanism of missing data.

3. Both methods are better able to predict $\beta_{XY}$ than $\beta_{YX}$. $\beta_{XY}$ is the regression of $X$ on $Y$ which was a completely observed dependent variable and an incomplete predictor while $\beta_{YX}$ is a regression with a completely observed predictor variable and an incomplete dependent variable.

4. Both methods perform particularly poorly when the data is MNAR.

5. While method A performs better than method B in many of the cases analysed, it does not automatically imply that it is a suitable method for dealing with missing data. It struggles to deal with high rates of missing data. This is of particular importance for users of SOLAS (Statistical-Solutions, 2014) as it is already implemented in this software.

4.2 Application to Biomechanical Jump Data

The biomechanics jump data was introduced briefly in Section 1.2.1. This dataset contains data collected to examine the effects of cold water immersion, also known as cryotherapy (Bleakley et al., 2012) on jumping movement and force measurement. Missing data occur in this dataset for a number of reasons. As a participant carries out the required jump, they may not land correctly on the plates that record force measurement. Similarly, they may jump in such a way that the infrared cameras that measure various joint angles is unable to record a particular jump. The task involved dropping from a 30 cm bench, and performing an immediate rebound-jump to tap a target which is suspended at a previously recorded maximum rebound-jump height. The suspended target acts as a trigger for a directional cueing system which
indicates to the subject on landing which direction (left or right) they must run to. The run/cutting manoeuvre is completed at a 45 angle. A number of kinetic and kinematic data were collected on this task using force platforms and infrared cameras. Twenty trials were carried out by each of 12 participants in each direction before cold water immersion and again 20 were carried out after immersion. Table 4.10 explains each of the variables that were collected.

![Figure 4.1](image)

Figure 4.1: The plot on the left shows the proportion of missing values on each variable used for this example while the plot on the right shows the variables where different combinations of missing data occur i.e. the pattern of missing data.

Of the 480 trials that were carried out (12 participants each doing 20 trials pre immersion and 20 trials post immersion), 440 were fully observed. This left 40 trials where there was some amount of missing data, 8.33%. Figure 4.1 outlines the variables on which these missing data occur and the pattern of missing data in the entire dataset. There are 32 of the 40 cases where the data is missing for all kinetic and kinematic data, information is just known on whether the trial was pre or post immersion and if the participant had to cut left or right.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>The participant in question.</td>
</tr>
<tr>
<td>Condition</td>
<td>Whether the task carried out was pre-immersion or post immersion.</td>
</tr>
<tr>
<td>Direction</td>
<td>The direction a subject had to run in upon landing after rebound-jump.</td>
</tr>
<tr>
<td>Trial</td>
<td>The number of the trial undertaken. 1 indicates the first trial carried out by a participant.</td>
</tr>
<tr>
<td>PGRF</td>
<td>Peak ground reaction force.</td>
</tr>
<tr>
<td>TimePGRF</td>
<td>Time to peak ground reaction force.</td>
</tr>
<tr>
<td>FPTime</td>
<td>Contact time to force plate.</td>
</tr>
<tr>
<td>TimeComplete</td>
<td>Time to complete the task.</td>
</tr>
<tr>
<td>MaxKnee</td>
<td>Maximum knee angle.</td>
</tr>
<tr>
<td>MaxAnkle</td>
<td>Maximum ankle angle.</td>
</tr>
<tr>
<td>MaxHip</td>
<td>Maximum hip angle.</td>
</tr>
<tr>
<td>TotKnee</td>
<td>Total knee range of motion.</td>
</tr>
<tr>
<td>ToAnkle</td>
<td>Total ankle range of motion.</td>
</tr>
<tr>
<td>TotHip</td>
<td>Total hip range of motion.</td>
</tr>
<tr>
<td>FirstPtKnee</td>
<td>Knee angle at initial ground contact.</td>
</tr>
<tr>
<td>TotAnkle</td>
<td>Ankle angle at initial ground contact.</td>
</tr>
<tr>
<td>TotHip</td>
<td>Hip angle at initial ground contact.</td>
</tr>
</tbody>
</table>

Table 4.10: Variable Descriptions

In order to determine if the cold water immersion has had an effect on the kinetic and kinematic data collected, the SOLAS algorithm and the proposed modified algorithm were implemented on the dataset. Average differences were calculated be-
between pre and post results to investigate differences between results before and after immersion before these results were combined using Rubin’s rules. The confidence intervals for the average difference of values pre and post cryotherapy are displayed in Tables 4.11 and 4.12.

For the SOLAS method, the following table of confidence intervals for the difference between pre-cryotherapy results and post-cryotherapy results was obtained.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Confidence Interval</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGRF</td>
<td>(-41.357, 105.074)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TimePGRF</td>
<td>(-0.013, 0.061)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>FPTtime</td>
<td>(-0.058, -0.010)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>TimeComplete</td>
<td>(-0.063, -0.002)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>MaxKnee</td>
<td>(-2.148, -0.419)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>MaxAnkle</td>
<td>(-0.858, 1.734)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>MaxHip</td>
<td>(-2.067, 0.874)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotKnee</td>
<td>(-0.548, 1.963)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotAnkle</td>
<td>(-0.549, 2.090)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotHip</td>
<td>(0.668, 3.583)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>FirstPtKnee</td>
<td>(-2.442, -0.522)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>FirstPtAnkle</td>
<td>(-1.514, 1.628)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>FirstPtHip</td>
<td>(-3.438, -1.184)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
</tbody>
</table>

Table 4.11: Difference as a result of cryotherapy accounting for missing data using SOLAS method

A number of variables showed a difference in results before and after cold water immersion. FPTime increased from before immersion to after immersion. Time to complete the task also increase from before immersion to after immersion. The maximum knee angle reached also increased. Total hip range of motion decreased from before cryotherapy to after cryotherapy. The knee angle at initial ground contact increased, as did hip angle at initial ground contact. For all other variables there was no difference observed for pre and post cryotherapy results.

The results on the same biomechanics dataset, this time accounting for missing values using the proposed modifications to the SOLAS method can be seen in Table 4.12. Similar to the results in Table 4.11, a number of variables showed a difference in results before and after cold water immersion. FPTime increased from before immersion to after immersion. Time to complete the task also increased.
Table 4.12: Difference as a result of cryotherapy accounting for missing data using modified SOLAS method

<table>
<thead>
<tr>
<th>Variable</th>
<th>Confidence Interval</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGRF</td>
<td>(-26.829, 132.600)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TimePGRF</td>
<td>(-0.021, 0.057)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>FPTime</td>
<td>(-0.059, -0.017)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>TimeComplete</td>
<td>(-0.063, -0.004)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>MaxKnee</td>
<td>(-2.122, -0.396)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>MaxAnkle</td>
<td>(-0.991, 1.790)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>MaxHip</td>
<td>(-2.102, 0.820)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotKnee</td>
<td>(-0.414, 2.105)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotAnkle</td>
<td>(-0.577, 2.187)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>TotHip</td>
<td>(0.527, 3.574)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>FirstPtKnee</td>
<td>(-2.394, -0.507)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
<tr>
<td>FirstPtAnkle</td>
<td>(-1.609, 1.703)</td>
<td>No difference as a result of cryotherapy</td>
</tr>
<tr>
<td>FirstPtHip</td>
<td>(-3.429, -1.123)</td>
<td>Difference between pre and post cryotherapy</td>
</tr>
</tbody>
</table>

Regardless of the method used to handle the missing data, the same conclusions were reached. This is largely due to the fact that the rate of missing data is quite low, the highest rate of missing data on any one variable is 8.33%. It is also very likely that the mechanism of missing data is MCAR as the reason that a missing value may occur is likely to have been down to how the jump was carried out by the participant as opposed to values of the missing data themselves or any other variables in the dataset.

Therefore, it is possible to conclude that on a dataset such as this biomechanics data, both the SOLAS method and the modified method both perform well and are suitable for use in this situation.
Chapter 5

Growing Up in Ireland

The Growing Up in Ireland (GUI) dataset was introduced in section 1.2.2. It is a national study of children who grow up in the Republic of Ireland and is the most significant study of its kind to take place in the country. The study is following the progress and development of almost 20,000 children throughout Ireland, collecting data on a huge number of variables. For the purposes of this analysis there is a small subset of those that are of particular interest and they are outlined in further detail in table 5.1.

Of the 8,568 children in the nine-year-old cohort, 8,465 of those children had female primary care givers broken into the following categories:

- 8,358 biological mothers
- 54 adoptive mothers
- 29 step-mothers
- 20 foster mothers
- 3 other relatives
- 1 unrelated

For this analysis only children with female primary care givers were considered in order that like subjects were being compared at all times. However, the analysis
<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mma5ap2</td>
<td>Sex of the child</td>
</tr>
<tr>
<td>IOTFgrade</td>
<td>Child’s International Obesity Taskforce (IOTF) grade</td>
</tr>
<tr>
<td>CQ22</td>
<td>Child’s opinion of their own body weight.</td>
</tr>
<tr>
<td>MMD9</td>
<td>Primary care giver’s image of the child’s body weight.</td>
</tr>
<tr>
<td>MMF7</td>
<td>Primary care giver’s opinion of their own body weight.</td>
</tr>
<tr>
<td>intPCGBMI</td>
<td>Primary care giver’s Body Mass Index (BMI) - derived from measured data.</td>
</tr>
<tr>
<td>MMagep1</td>
<td>Age of primary care giver.</td>
</tr>
<tr>
<td>MMB1</td>
<td>Child’s birth weight in kgs.</td>
</tr>
<tr>
<td>srPCGBMI</td>
<td>BMI of primary care giver derived from self reported data.</td>
</tr>
<tr>
<td>intChildBMI</td>
<td>Child’s BMI from measured data.</td>
</tr>
</tbody>
</table>

Table 5.1: Variable Descriptions

was re-run with male primary care givers included and there were minimal changes in the results.

In the Growing Up in Ireland dataset, the variable “Child’s IOTF grade” was broken into 6 categories originally. There were three “thinness” categories, for children who had BMI of less than 16, 17 and 18.5 respectively. These categories have been grouped together into one, called “underweight”. There were also two categories for children who had BMI of greater than 25 and greater than 30. These have also been grouped together for the purposes of analysis into one category called “overweight”. There was one “normal weight” category for those with a BMI of between 18.5 and 25 which remained a single category.

The variable “Child’s image of self” category was originally recorded with five different levels ranging from “Very skinny” to “Very overweight”. The levels “very
skinny” and “a bit skinny” were grouped together in one category called “underweight”. The group “Just the right size” was relabelled as “normal” and “a bit overweight” and “very overweight” were joined in a category called “overweight”.

The variables “Primary care givers image of self” and “Primary care givers image of child” both had seven levels on the obesity scale. The levels “very underweight”, “moderately underweight” and “slightly underweight” were grouped into one variable called “underweight”. “About the right weight” was renamed “normal” and the remaining three categories “slightly overweight”, “moderately overweight” and “very overweight” were combined in a category called “overweight”.

5.1 Kappa Statistic

Kappa (Cohen, 1960) is a single number statistic that measures the strength of agreement between two raters on the same subject. It compares the observed agreement to what would be expected if the ratings are independent and can be computed for nominal or ordinal variables. If variables are nominal, an unweighted value for kappa is calculated. However, if categories within a variable can be ordered, the variable is ordinal and a weight (Cohen, 1968) can be applied to kappa to allow for deviation from agreement between raters.

As kappa aims to reduce a table of agreement between two raters to a single statistic it can best be described as an exploratory variable to aid with further investigation and analysis. Reducing a table to one number can result in the loss of a certain amount of information and therefore cannot always be recommended as the best way of analysing the inter-rater agreement. Other such methods of analysing inter-rater agreement include odds ratios and modelling the relationship between the two raters using logistic regression models.

Kappa can range from values between 0 and 1 with higher values of kappa indicating stronger agreement between two raters. While deciding how strong the
relationship is between two raters is entirely arbitrary, Landis and Koch (1977) suggested values of kappa and their corresponding strength of agreement. Their suggested categories define values between 0 and 0.2 as slight agreement, 0.21-0.4 as fair, 0.41-0.6 as moderate, 0.61-0.8 values have substantial agreement and between 0.81 and 1 there is almost perfect agreement.

The formula to obtain weighted kappa is as follows:

$$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

(5.1)

with $p_{o(w)}$ being the observed weighted proportion of agreement calculated by

$$p_{o(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}p_{ij}$$

(5.2)

where $p_{ij}$ is the sample cell probability i.e. the probability of being in row $i$ and column $j$, and $p_{e(w)}$, the expected weighted proportion of agreement, is got by finding

$$p_{e(w)} = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}p_i.p_j$$

(5.3)

Here, $p_i$ is the probability of being observed in row $i$ and $p_j$ is the probability of being observed in column $j$. The weights applied in the calculation for kappa also vary between 0 and 1. Two of the most popular choices for weights (Agresti, 2010) can be seen below. The choice of weights penalises values the further they are from the diagonal, the diagonal being where agreement between two raters occurs. Fleiss and Cohen (1973) suggested possible weights to be taken as

$$w_{ij} = 1 - \frac{(i - j)^2}{(k-1)^2}$$

(5.4)

while Cicchetti and Allison (1971) suggested the weights of

$$w_{ij} = 1 - \frac{|i - j|}{|k - 1|}$$

(5.5)
where \( k \) is the number of categories into which a subject can be assigned. When comparing the different raters (child and primary care giver) with how the child is actually classified by their IOTF grade, the table of agreement will be a \( 3 \times 3 \) table. The resulting matrices of weights for the \( 3 \times 3 \) case can be seen below. For equation 5.4, the matrix of weights are on the left and for equation 5.5 the weights are seen here on the right:

\[
\begin{pmatrix}
1 & 0.75 & 0 \\
0.75 & 1 & 0.75 \\
0 & 0.75 & 1
\end{pmatrix}
\]

\[
\begin{pmatrix}
1 & 0.5 & 0 \\
0.5 & 1 & 0.5 \\
0 & 0.5 & 1
\end{pmatrix}
\]

### 5.1.1 Kappa Analysis on GUI dataset

Applying the kappa statistic to the Growing Up in Ireland data, the strength or ability of the primary care giver and and the child to correctly classify the child as “underweight”, “normal” or “overweight” can be tested. In the following tables kappa “1” is calculated using the weight in equation 5.5 while kappa “2” refers to the weight specified in equation 5.4.

<table>
<thead>
<tr>
<th>Child’s Image of Self</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
<th>Kappa Value</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>228</td>
<td>261</td>
<td>12</td>
<td>1</td>
<td>0.16</td>
<td>0.14</td>
<td>0.18</td>
</tr>
<tr>
<td>Normal</td>
<td>1150</td>
<td>3923</td>
<td>218</td>
<td>2</td>
<td>0.25</td>
<td>0.23</td>
<td>0.26</td>
</tr>
<tr>
<td>Over</td>
<td>149</td>
<td>1665</td>
<td>380</td>
<td>2</td>
<td>0.5</td>
<td>0.48</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 5.2: Childs classification of themselves compared to child’s IOTF grade.

<table>
<thead>
<tr>
<th>PCG’s Image of Child</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
<th>Kappa Value</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>204</td>
<td>294</td>
<td>5</td>
<td>1</td>
<td>0.44</td>
<td>0.38</td>
<td>0.42</td>
</tr>
<tr>
<td>Normal</td>
<td>620</td>
<td>4586</td>
<td>110</td>
<td>2</td>
<td>0.5</td>
<td>0.48</td>
<td>0.51</td>
</tr>
<tr>
<td>Over</td>
<td>37</td>
<td>1133</td>
<td>1050</td>
<td>2</td>
<td>0.5</td>
<td>0.48</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Table 5.3: Primary care givers classification of child compared to child’s IOTF grade.

When comparing the childs IOTF grade with the childs image of themselves the kappa values obtained are low indicating slight and at most fair agreements between the two raters, depending on which weights are used (Table 5.2). The primary care
giver is a much better rater of a child’s weight as the kappa values for this rater when compared with the child’s IOTF grade range from between 0.42 to 0.46 and 0.48 to 0.51 in Table 5.3.

A further dimension of analysis that may be of interest to investigate is where primary care givers are grouped by their BMI, and subsequently their opinion of the child then analysed with the child’s IOTF grade.

<table>
<thead>
<tr>
<th>PCG’s Image of Child</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
<th>Kappa Value</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>5</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0.32</td>
<td>0.13</td>
<td>0.51</td>
</tr>
<tr>
<td>Normal</td>
<td>9</td>
<td>57</td>
<td>2</td>
<td>2</td>
<td>0.40</td>
<td>0.22</td>
<td>0.58</td>
</tr>
<tr>
<td>Over</td>
<td>0</td>
<td>9</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.4: Primary care givers (whose BMI indicates they are underweight) classification of child compared to child’s IOTF grade.

<table>
<thead>
<tr>
<th>PCG’s Image of Child</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
<th>Kappa Value</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>112</td>
<td>186</td>
<td>4</td>
<td>1</td>
<td>0.38</td>
<td>0.35</td>
<td>0.41</td>
</tr>
<tr>
<td>Normal</td>
<td>303</td>
<td>2347</td>
<td>58</td>
<td>2</td>
<td>0.44</td>
<td>0.41</td>
<td>0.47</td>
</tr>
<tr>
<td>Over</td>
<td>12</td>
<td>404</td>
<td>299</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5: Primary care givers (whose BMI indicates they are normal weight) classification of child compared to child’s IOTF grade.

<table>
<thead>
<tr>
<th>PCG’s Image of Child</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
<th>Kappa Value</th>
<th>Lower CI</th>
<th>Upper CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>74</td>
<td>89</td>
<td>1</td>
<td>1</td>
<td>0.46</td>
<td>0.44</td>
<td>0.49</td>
</tr>
<tr>
<td>Normal</td>
<td>283</td>
<td>1971</td>
<td>45</td>
<td>2</td>
<td>0.51</td>
<td>0.49</td>
<td>0.54</td>
</tr>
<tr>
<td>Over</td>
<td>24</td>
<td>649</td>
<td>675</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6: Primary care givers (whose BMI indicates they are overweight) classification of child compared to child’s IOTF grade.

The three tables, 5.4, 5.5 and 5.6, would suggest that primary care givers who are overweight according to their BMI are better raters of children than primary care givers who are normal weight according to their BMI. It cannot be said if primary care givers who are overweight are better raters than those who are underweight due to the small number of primary care givers who fall into the category of underweight.
5.1.2 Advantages and Disadvantages of Kappa

Advantages

• Agreement between two raters can be summarised in a single number compared with other agreement measures.

• Kappa can be computed for both ordinal and nominal variables by introducing weights to the calculations.

Disadvantages

• The values of both kappa and weighted kappa depend strongly on the marginal distributions.

• The values for weighted kappa can only be compared if the weights used are the same for all calculations.

As kappa is just a one number statistic, some of the information that can be obtained from the contingency tables of agreement and disagreement may be lost and subsequently it may be of use to carry out a more detailed model based investigation into what is happening within the data. Logistic regression has become the standard way of analysing ordinal variables and will be outlined more in Section 5.2.

5.2 Logistic Regression

Logistic regression has become a standardised way of analysing ordinal variables (Agresti, 2010). For ordinal variables there are a number of types of logistic regression that can be modelled on the data at hand with one of the most popular being cumulative logistic regression. It is cumulative logistic regression models that have been fitted to the Growing Up in Ireland data in question and the models for various pairs of variables together can be seen below in equation 5.6. The structure of cumulative logistic regression is that there are various outcome categories for the response variable and the model is built on the observed values for the explanatory variables so
that the model used has the form:

$$\text{logit}[P(Y_i \leq j)] = \alpha_j + \beta'x_i = \alpha_j + \beta_1 x_{i1} + \beta_2 x_{i2}...$$  \hspace{1cm} (5.6)

$\beta$ is a vector of parameters that describe the explanatory variables and $Y_i$ is the outcome category for the response variable. In the case of our data, the outcome categories are whether a variable is under (denoted as group 0), normal (group 1) or over (group 2) weight. $j$ goes from 1 to $c - 1$ where $c$ is the number of categories in the response variable. The vector of parameters, $\beta$, are calculated by maximum likelihood estimation using an iterative approach such as the Newton-Raphson method (Albert and Anderson, 1984).

If $p$ is probability of an occurrence, then the odds are $o = p/(1-p)$. Subsequently the log of the odds can be modelled in the form $\alpha + \beta X$. If $x = 0$ then the following equation is formed:

$$\log\left(\frac{p}{1-p}\right) = \alpha$$ \hspace{1cm} (5.7)

and solving for $p$ the following is obtained:

$$p = \frac{e^\alpha}{1 + e^\alpha}$$ \hspace{1cm} (5.8)

Using equation 5.8 the fitted values for combinations of variables can be computed depending in the coefficients in each particular model. Applying this cumulative logistic regression model to the data we have extracted from the GUI dataset, the following matrices of probability estimates are obtained and shown in Tables 5.7 and 5.8:

<table>
<thead>
<tr>
<th>Child’s Image of Self</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0.5</td>
<td>0.48</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.21</td>
<td>0.74</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Over</td>
<td>0.07</td>
<td>0.77</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.7: Cumulative logistic regression of IOTF grade with child’s image of themselves
For example, given that a child graded as having normal weight by the IOTF grade, they will consider themselves to be underweight with probability 0.21, have normal weight with probability 0.74 and be overweight with probability 0.05. Given that a child graded as having normal weight by the IOTF grade, the primary care giver will consider the child to be underweight with probability 0.1, have normal weight with probability 0.86 and be overweight with probability 0.04.

### 5.3 Missing Data Analysis on GUI

To date, all the analysis done on the GUI dataset has been on the observed values and as outlined in chapter 2 this is not the best way to deal with datasets containing missing values. Within the Growing Up in Ireland dataset there are some missing values that should be taken into account. Table 5.9 shows the missing data pattern that occurs in the variables used throughout this chapter.

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>PCG’s Image</th>
<th>Child’s Image</th>
<th>IOTF grade</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Observed</td>
<td>Observed</td>
<td>Observed</td>
<td>8080</td>
</tr>
<tr>
<td>Observed</td>
<td>Observed</td>
<td>Observed</td>
<td>Missing</td>
<td>Observed</td>
<td>403</td>
</tr>
<tr>
<td>Observed</td>
<td>Observed</td>
<td>Missing</td>
<td>Observed</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td>Observed</td>
<td>Missing</td>
<td>Missing</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0</td>
<td>0</td>
<td>84</td>
<td>432</td>
<td>516</td>
</tr>
</tbody>
</table>

Table 5.9: Missing data pattern for the four variables listed above

Before choosing the number of imputations that will be carried out, Table 2.1 should be considered. In the case of the GUI dataset, the rate of missing data for the IOTF grade variable is 0.05 and less again for the child’s image of themselves, resulting in 99% efficiency with 5 imputations. The imputation of the ordinal cate-
gorical variables was carried out by specifying a proportional odds logistic regression model within the chained equations framework. This can be conducted in R (R Development Core Team, 2008) using the mice.impute.polr function from the mice package (van Buuren and Groothuis-Oudshoorn, 2011). Re-running the cumulative logistic regression and combining the results of the five imputed dataset according to Rubin’s rules outlined in section 2.5.3, tables 5.10 and 5.11 were obtained.

<table>
<thead>
<tr>
<th>Child’s Image of Self</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0.48</td>
<td>0.50</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.21</td>
<td>0.74</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Over</td>
<td>0.07</td>
<td>0.77</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.10: Cumulative logistic regression on the imputed data for IOTF grade and child’s image of self.

<table>
<thead>
<tr>
<th>PCG’s Image of Child</th>
<th>IOTF</th>
<th>Under</th>
<th>Normal</th>
<th>Over</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under</td>
<td>0.59</td>
<td>0.41</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>0.1</td>
<td>0.84</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>Over</td>
<td>0.01</td>
<td>0.58</td>
<td>0.41</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.11: Cumulative logistic regression on the imputed data for IOTF grade and primary care giver’s image of the child.

Upon imputation of the data and combining the results, the logistic regression estimates do not differ very much compared to the complete analysis. This is due to a number of factors, largely because the proportion of data that is missing is relatively small to have had a large impact that would have dramatically changed the results.
5.4 Conclusions

1. Children as raters of their own weight do not classify themselves well with many overweight children likely to see themselves as having normal weight.

2. Primary care givers in general are much better at rating their children correctly into the classes of under, normal and over weight than the children themselves.

3. On analysing the ability of primary care givers to correctly classify their children depending on the BMI of the parent, it was found that primary care givers who are themselves overweight are better raters of their child’s weight than primary care givers who have a normal BMI. As the sample size is very small for primary care givers who are underweight, it is not possible to say conclusively how they compare to primary care givers with an overweight BMI.

4. The results of analysis carried out, taking into account the missing data, did not yield any different results. This was due to the very small proportion of missing data in the variables used in the analysis. The method used for imputation in this case was a proportional odds logistic regression model which allows for the prediction of categorical variables.
Chapter 6

Conclusions

This thesis introduces the concept of missing data and its development through to the present day. Along side missing data, the Mahalanobis distance measure is introduced and its application in the area of missing data is outlined. One such area is in multiple imputation with the Mahalanobis distance used as a metric to identify similar subject to those containing missing data. This idea is used in software called SOLAS. An investigation into the performance of the SOLAS method was carried out. Recommendations on modifications that could be made to the SOLAS method were proposed and a simulation analysis was carried out to investigate this proposed method. Upon completion of these investigations, it was concluded that both methods perform adequately in certain situations and poorly in others. Broadly, both methods do best when the data are MCAR and the rate of missing data are 10%. The SOLAS method performs better in many cases than the proposed modified method when the rate of missing data increases to 25%. When the rate of missing data increases to 50%, for many of the parameters analysed, both methods fail to evaluate them well.

As the biomechanics dataset has a rate of missing data less than 10%, it is appropriate to analyse it using the methods introduced and proposed in Chapter 3. Combining the results of the analysis using the rules set out in Rubin (1987), changes were noticed in a number of the variables collected as a consequence of cold water immersion. Both methods indicated that there were differences for the following vari-
ables: $FP\text{Time}$, $Time\text{Complete}$, $Max\text{Knee}$, $Tot\text{Hip}$, $First\text{PtKnee}$, and $First\text{PtHip}$.

In summary of the analysis carried out on the GUI data, using metrics that assess the level of agreement between two raters, it is possible to conclude that primary care givers who are classified as overweight according to their BMI are better raters of their children’s weight than those primary care givers who are normal weight. Additionally, it was also shown that children themselves are not good raters of their weight, with many overweight children identifying themselves as normal weight.

In future work, it may be of interest to expand the use of the Mahalanobis distance matching metric in the capacity of multiple imputation so that it is appropriate for use with categorical variables. The conditional multivariate distribution in Appendix A.2. could be made use of for this research. It could also be further investigated if varying the calculation of the covariance matrix used in the calculation of the Mahalanobis distance, as outlined in Section 3.2, would have an effect on the results obtained in the simulations of Chapter 4.
Bibliography


Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture 1*(2), 129–142.

Appendix A

Probability

A.1 The Multivariate Normal Distribution

Definition A.1 (Multivariate Gaussian). The $p$–dimensional random vector $Y$ has a multivariate normal distribution if and only if its density is of the form

$$f_Y(y) = (2\pi)^{-p/2} \det(V)^{-1/2} \exp \left\{ -\frac{1}{2} (y - \mu)' V^{-1} (y - \mu) \right\}, \quad \text{for } y \in \mathbb{R}^p$$

where $\mu$ is a fixed $p$–dimensional vector and the $p \times p$ matrix $V$ is symmetric and positive definite.

The vector $\mu$ and matrix $V$ are parameters of this distribution. We shall see that $\mu$ is the expected value of $Y$, and $V$ is the variance matrix of $Y$. The following direct construction of a multivariate normal random vector from a set of independent normal random variables shows that such a distribution exists.

Step 1: Let $Z_1, Z_2, \ldots, Z_p$ be $p$ independent random scalars each with standard $N(0,1)$ density function, and put

$$Z = \begin{bmatrix}
Z_1 \\
Z_2 \\
\vdots \\
Z_p
\end{bmatrix}.$$ 

Note that $E(Z) = 0$ and $\text{var}(Z) = I$. 
Step 2: The density function of $Z$ is identifiable as the joint density function of $Z_1, Z_2, \ldots, Z_p$ which by independence, is

$$f_Z(z) = f_{Z_1, Z_2, \ldots, Z_p}(z_1, z_2, \ldots, z_p) = (2\pi)^{-p/2} \exp\left\{ -\frac{1}{2} z' z \right\},$$

where we use the fact that $z' z = \sum_i z_i^2$.

Step 3: Given a symmetric positive definite matrix we can always find a positive definite matrix $A$ to express it as $AA'$. As $V$ is symmetric and positive definite we can therefore find $A$ such that $V = AA'$.

Step 4: Set $X = AZ$. The Jacobian of this transformation is $\det(A)^{-1} = \det(V)^{-1/2}$ and so the density function of $X$ is

$$f_X(x) = (2\pi)^{-p/2} \det(V)^{-1/2} \exp\left\{ -\frac{1}{2} x' V^{-1} x \right\}.$$

Step 5: Finally put $Y = X + \mu$ to give $f_Y$ as in the definition.

Proposition A.1. Any multivariate normal random vector $Y$ can be represented as a linear transformation of a standard multivariate normal random vector $Z$, that is, $Z \sim N(0, 1)$, together with a shift: $Y = AZ + \mu$.

Proof. The result is established from the five steps above. \hfill \Box

Corollary A.2. The multivariate normal random vector $Y$ has expected value and $E(Y) = \mu$ and variance $\text{var}(Y) = V$.

Proof. Set $Y = AZ + \mu$ and so by the linearity of the expectation, $E(Y) = AE(Z) + \mu$. But $E(Z) = 0$. As $\text{var}(Z) = I$, it follows that $\text{var}(Y) = \text{var}(AZ) = A\text{var}(Z)A'$ and, by construction, the last term is just $V$. \hfill \Box

Corollary A.3. If $Y \sim N$ then the linear combination $v'Y \sim N$ for all fixed $p$–dimensional vectors $v$. 

89
Proof. As $Y = A\mathbf{Z} + \mu$ then

$$v'Y = v'(A\mathbf{Z} + \mu) = (A'v)'\mathbf{Z} + v'\mu = \sum \alpha_j Z_j + v'\mu.$$ 

But this is just a linear combination of normal random scalars which is also normally distributed. \qed

Corollary A.4. If $E(Y) = 0$ then $Y'\text{var}(Y)^{-1}Y \sim \chi^2_p$.

Proof. We have $Y = A\mathbf{Z}$ and $\text{var}(Y) = AA'$ so

$$Y'\text{var}(Y)^{-1}Y = (AZ)'(AA')^{-1}(AZ) = Z'Z = \sum_{j=1}^{p} Z_j^2,$$

is a sum of $p$ independent chi-squared random variables each with 1 degree of freedom; hence the quadratic form has a chi-squared distribution with $p$ degrees of freedom. \qed

A.2 The Conditional Multivariate Normal

Consider the $(p+q)$-dimensional multivariate normal distribution whose variates are written as the partitioned vector $(\mathbf{Y} \mathbf{X})$ where $\mathbf{Y}$ contains $p$ elements and $\mathbf{X}$ contains $q$ elements. The mean vector and variance-covariance matrix are correspondingly partitioned as

$$E\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} E(\mathbf{Y}) \\ E(\mathbf{X}) \end{pmatrix} = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}$$

and

$$\text{var}\begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} \text{var}(\mathbf{Y}) & \text{cov}(\mathbf{Y}, \mathbf{X}) \\ \text{cov}(\mathbf{X}, \mathbf{Y}) & \text{var}(\mathbf{X}) \end{pmatrix} = V$$

In general, the conditional density function of $\mathbf{Y}$ given $\mathbf{X}$ is fixed, say $x$, is

$$f_{Y|X=x} = \frac{f_{Y,X}(y,x)}{f_X(x)}.$$
In the present case

\[ f_X(x) = (2\pi)^{-q/2} \det[\text{var}(X)]^{-1/2} \exp \left\{ -\frac{1}{2} (x - \mu_x)'\text{var}(X)^{-1}(x - \mu_x) \right\} . \]

For the computation of the conditional density \( f_{Y|X} \) we shall need the joint density \( f_{Y,X} \) in terms of the submatrices of \( V \). From section (??) we have that

\[ \det(V) = \det[\text{var}(X)] \times \det[\text{var}(Y) - \text{cov}(Y,X)\text{var}(X)^{-1}\text{cov}(X,Y)] , \]

and that

\[
V^{-1} = \begin{pmatrix}
Q^{-1} & -Q^{-1}\text{cov}(Y,X)\text{var}(X)^{-1} \\
-\text{var}(X)^{-1}\text{cov}(X,Y)Q^{-1} & \text{var}(X)^{-1} + \text{var}(X)^{-1}\text{cov}(X,Y) \\
\times Q^{-1}\text{cov}(Y,X)\text{var}(X)^{-1}
\end{pmatrix}
\]

where \( Q = \text{var}(Y) - \text{cov}(Y,X)\text{var}(X)^{-1}\text{cov}(X,Y) \).

The joint density can be written

\[
f_{Y,X} = (2\pi)^{-\left(\frac{p+q}{2}\right)} \det[\text{var}(X)]^{-\frac{1}{2}} \\
\times \det[\text{var}(Y) - \text{cov}(Y,X)\text{var}(X)^{-1}\text{cov}(X,Y)]^{-\frac{1}{2}} \\
\times \exp\left\{ -\frac{1}{2} \left[ (y - \mu_y)'Q^{-1}(y - \mu_y) \\
-(x - \mu_x)'\text{var}(X)^{-1}\text{cov}(X,Y)Q^{-1}(y - \mu_y) \\
-(y - \mu_y)'Q^{-1}\text{cov}(Y,X)\text{var}(X)^{-1}(x - \mu_x) \\
+(x - \mu_x)'\text{var}(X)^{-1}\text{cov}(X,Y)Q^{-1}\text{cov}(Y,X)\text{var}(X)^{-1}(x - \mu_x) \\
+(x - \mu_x)'\text{var}(X)^{-1}(x - \mu_x) \right]\right\}
\]

Upon division by \( f_X(x) \) the term \((2\pi)^{-q/2} \det[\text{var}(X)]^{-1/2}\) in the normalizing constant and the final quadratic term of the exponent vanish. Inspection of the

91
remaining terms show that the conditional density can be written

\[
f_{Y|X} = (2\pi)^{-\frac{d}{2}} \det[\text{var}(Y) - \text{cov}(Y, X)\text{var}(X)^{-1}\text{cov}(X, Y)]^{-\frac{1}{2}}
\times \exp\left\{ -\frac{1}{2} \left[ (y - \mu_Y - \text{cov}(Y, X)\text{var}(X)^{-1}(x - \mu_X))^\prime \times (\text{var}(Y) - \text{cov}(Y, X)\text{var}(X)^{-1}\text{cov}(X, Y))^{-1} \times (y - \mu_Y - \text{cov}(Y, X)\text{var}(X)^{-1}(x - \mu_X)) \right] \right\}
\]

This is the density of a multivariate normal distribution with mean vector

\[
E_{Y|X}(Y|X) = E(Y) + \text{cov}(Y, X)\text{var}(X)^{-1}(x - E(X)) \quad (A.1)
\]

and (conditional) variance-covariance matrix

\[
\text{var}_{Y|X}(Y|X) = \text{var}(Y) - \text{cov}(Y, X)\text{var}(X)^{-1}\text{cov}(X, Y). \quad (A.2)
\]
Appendix B

R Code

B.1 Algorithm 1

```r
partition = function(Y) {
  rows = complete.cases(Y)
  columns = complete.cases(t(Y))
  X = Y[rows,columns]
  if (length(which(!rows==1))==1) {
    Xtilde = matrix(Y[!rows,columns], nrow = 1)
  } else {
    Xtilde = (Y[!rows,columns])
  }
  Zobs = Y[rows, !columns]
  Zmis = Y[!rows, !columns]
  return(list(X = X, Xtilde=Xtilde, Zobs=Zobs, Zmis=Zmis, rows=rows, columns=columns ))
}
```

B.2 Algorithm 2

```r
calc.MDs = function(M, vc = "one"){
  Q = partition(M)
  Q$X = as.matrix(Q$X)
  Q$Xtilde = as.matrix(Q$Xtilde)
  if (vc == "one"){
```
\[ S \text{inv} = \text{solve}(\text{var}(QX)) \]

\[ m = \text{nrow}(QXtilde) \]

\[
\text{if } (m == 1) \{
    \text{MDs} = \text{mahalanobis}(QX, \text{center} = c(QXtilde), \text{cov} = S\text{inv}, \text{inverted} = \text{TRUE})
    \text{MDs} = \text{as.matrix}(\text{MDs})
    \text{MDs} = 1 - \text{pchisq}(\text{MDs}, \text{df}=\text{ncol}(QXtilde))
\}
\text{else} \{
    \text{MDs} = \text{mahalanobis}(QX, \text{center} = c(QXtilde[1,]), \text{cov} = S\text{inv}, \text{inverted} = \text{TRUE})
    \text{for} (j \text{ in } 2:m) \{
        \text{MDs} = \text{rbind}(\text{MDs}, \text{mahalanobis}(QX, \text{center} = c(QXtilde[j,]), \text{cov} = S\text{inv}, \text{inverted} = \text{TRUE}))
    \}
    \text{MDs} = t(\text{MDs})
    \text{MDsProb} = 1 - \text{pchisq}(\text{MDs}, \text{df}=\text{ncol}(QXtilde))
\}
\]

\section*{B.3 Algorithm 3}

\begin{verbatim}
impMDsim1b = function(x, varc = "one", m = 5){
M = is.na(x) * 1
N = unique(M)

if (is.data.frame(x)) x = as.matrix(x)

Q = partition(x)
ZobsX = cbind(QX, Q$Zobs)
ZmisXtilde = cbind(QXtilde, Q$Zmis)
a = dim(ZmisXtilde)[1]

MXZ = is.na(ZmisXtilde) * 1
NXZ = unique(MXZ)
n = dim(NXZ)[1]
a = dim(NXZ)[1]
\end{verbatim}
if (is.data.frame(ZobsX)) ZobsX = as.matrix(ZobsX)
if (is.data.frame(ZmisXtilde)) ZmisXtilde = as.matrix(ZmisXtilde)

ZXnew = ZobsX
ans = numeric(0)

for (i in 1:n){
  for (k in 1:a){
    if (identical(MXZ[k,],NXZ[i,]) == TRUE){
      ZXnew = rbind(ZXnew, ZmisXtilde[k,])
    }
  }
  ans = cbind(ans, calc.MDs(ZXnew, vc = varc))
}

f = matrix(c(which(rowSums(M)==0), c(1:nrow(ans))), ncol = 2)

s = numeric(0)
for (i in 1:ncol(ans)){
  s = cbind(s, sample(1:nrow(ans), m, replace = TRUE, prob = ans[,i]))
}

h = numeric(0)
for (i in 1:ncol(ans)){
  h = cbind(h, f[c(s[,i],1)])
}

miscols = which(Q$columns == FALSE)
x.new = x

o = array(NA, dim = c(nrow(x), ncol(x),m))

k = which(rowSums(M) != 0)
for (j in 1:m){
  for (i in 1:ncol(ans)){

95
miss = which(is.na(x[(k[i]) ,]))==TRUE)
x.new[k[i] ,c(miss)] = x[h ,i] [j ,][c(miss)]
}
o[ ,j] = x.new
}
return(o)

B.4 Algorithm 4

calc.MDsA = function(M, vc = "one"){
  Q = partition(M)
  Q$X = as.matrix(Q$X)
  Q$Xtilde = as.matrix(Q$Xtilde)
  if (vc == "one"){
    Sinv = solve(var(Q$X))
    m = nrow(Q$Xtilde)
    if (m == 1) {
      MDs = mahalanobis(Q$X, center = c(Q$Xtilde), cov = Sinv, inverted = TRUE)
      MDs = as.matrix(MDs)
    } else {
      MDs = mahalanobis(Q$X, center = c(Q$Xtilde[1 ,]) , cov = Sinv, inverted = TRUE)
      for (j in 2:m) {
        MDs = rbind(MDs, mahalanobis(Q$X, center = c(Q$Xtilde[j ,]) , cov = Sinv, inverted = TRUE))
      }
      MDs = t(MDs)
    }
  }
  return(MDs)
}
Algorithm 5

```r
impMDsim1a = function(x, varc = "one", m = 5) {
  M = is.na(x) * 1
  N = unique(M)

  if (is.data.frame(x)) x = as.matrix(x)

  Q = partition(x)
  ZobsX = cbind(Q$X, Q$Zobs)
  ZmisXtilde = cbind(Q$Xtilde, Q$Zmis)
  a = dim(ZmisXtilde)[1]

  MXZ = is.na(ZmisXtilde) * 1
  NXZ = unique(MXZ)
  n = dim(NXZ)[1]
  a = dim(MXZ)[1]

  if (is.data.frame(ZobsX)) ZobsX = as.matrix(ZobsX)
  if (is.data.frame(ZmisXtilde)) ZmisXtilde = as.matrix(ZmisXtilde)

  ZXnew = ZobsX
  ans = numeric(0)

  for (i in 1:n) {
    ZXnew = ZobsX
    for (k in 1:a) {
      if (identical(MXZ[k,], NXZ[i,]) == TRUE) {
        ZXnew = rbind(ZXnew, ZmisXtilde[k,])
      }
    }
    ans = cbind(ans, calc.MDsA(ZXnew, vc = varc))
  }

  ord = numeric(0)
}
```
for (i in 1:ncol(ans)) {
    ord = cbind(ord, order(ans[, i]))
}
g = ord[c(1:(m*2)),]
f = matrix(c(which(rowSums(M) == 0), c(1:nrow(ans))), ncol = 2)
h = numeric(0)
for (i in 1:ncol(ord)) {
    h = cbind(h, f[c(g[, i]), 1])
}
s = numeric(0)
for (i in 1:ncol(ord)) {
    s = cbind(s, sample(1:(m*2), m, replace = TRUE))
}

miscols = which(Q$columns == FALSE)

x.new = x

o = array(NA, dim = c(nrow(x), ncol(x), m))
k = which(rowSums(M) != 0)
for (j in 1:m) {
    for (i in 1:ncol(ord)) {
        miss = which(is.na(x[(k[i],)]) == TRUE)
        x.new[k[i], c(miss)] = x[h[, i], ][s[j, i], ][c(miss)]
    }
    o[, , j] = x.new
}
return(o)