Methods for Handling Missing Data in Research Studies

A project submitted in partial fulfilment of the requirements for the degree of

Bachelor of Science Honours

in

Mathematical Statistics

Department of Statistics

Rhodes University

by

Chotsani Siyabonga Isaac

September 2023

Supervisor: Dr Chinomona Amos

Abstract

Missing data are more prevalent and almost unavoidable phenomena in research. Ignoring missing data can results in misleading or incorrect results and conclusions. This paper investigate available methods of handling missing data in research studies. Over the pasts years a variety of methods of handling missing data have been developed, mainly to mitigate problems that arises when analysing data with missing data. The available methods do not perform the same in terms of handling missing data, some of this methods give best estimated values for the missing values some do not. According to Rubin (1976) missing values can be classified in to three main categories that is Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not At Random (MNAR) These are called missing data mechanisms, to correctly choose the right method to handle missing values the correct classification for the missing values needs to made. The deletion method, single imputation method and the multiple imputation method will be investigated in this study. In this paper a synthetic data set will be used for analysis. The results obtained in this study shows missing data in research studies can lead to poor results and conclusions if proper steps are not taken to handle the missing data points. From the investigated methods of handling missing data the results show that multiple imputation method handles missing data better than other methods.

Keywords: **Missing data, Imputation, Deletion, Respondents, Nonrespone.**

Contents

List of Figures

List of Tables

List of Abbreviations

Missing Completely At Random - **MCAR**

Missing At Random – **MAR**

Missing Not At Random – **MNAR**

Expectation Maximisation – **EM**

Raw Maximum Likelihoods – **RML**

Statistical Analysis System – **SAS**

Statistical Package – **SP**

Statistical Package for the Social Sciences – **SPSS**

Analysis of a Moment Structure - **AMOS**

Single Imputation –**SI**

Multiple Imputation –**MI**

Analysis of variance -**ANOVA**

Aikaike Information Criterion -**AIC**

Bayesian Information Criterion -**BIC**

Acknowledgments

This project would have not been possible if it was not the support I got from many people around me. I would to take this opportunity to thank my supervisor Dr Amos Chinomona for the perfect guidance, the advice he gave me, and the patience he had with me, without his assistance this research would not have been possible. I would like to also thank the entire statistics department staff members for the holistic support we got. To the Tomorrow Trust Foundation I like to acknowledge and thank you for the financial support and the psychological support you gave me. Lastly I like to thank my family especial my mom, my friends and my classmates for the encouragement and support I am very grateful.

Chapter 1

Introduction

1.1 Background

Research studies often involve collection of data. Most of the collected data is used to make informed decisions about various phenomena. However, the opposite is true if the collected data is characterized by or contains missing observations; inferences from this data may lead to incorrect conclusions. In this research methods of handling missing data in research will be studied. During the process of collecting data a lot could go wrong or lead to the data collected having missing observations. Sometimes data can be missing as a result of many contributing factors. For instance when a survey is conducted with the use of a questionnaire, where respondents can choose not to participate or not to respond to some of the survey questions. The validity of the collected data can be influenced by a lot of factors some of which are explained by Hansen and Hurwitz (1946), who discussed how the cost of conducting a survey can influence the outcome leading to biased conclusions.

In this project we will investigate the different methods and techniques that were developed to mitigate the challenges of missing observations in data analysis. This is because the effects of missing data impact negatively on the quality of the results from the data analysis. With the available methods we will discuss how each of them ensures that the data containing missing observations leads to better conclusions. There are a lot of available methods developed to mitigate the effects of missing data. The imputation and the likelihood methods together with deletion methods (list-wise deletion method and the pairwise deletion method) as explained in Cheema (2014) will be discussed. In particular we will investigate and assess which one is the best out of all the other available methods.

1.2 Aim

To investigate methods of handling missing data in research and how this methods improves the validity and the significance of the results.

1.3 Objective

- To investigate different methods of handling missing data in research studies.
- To critically assess how useful are the methods in improving the significance of the results.

1.4 Significance of the study

This study is significant because if we know which methods leads to the best results in a data set characterised by missing values, it will be easy to decide which method to use when we have a certain type of missingness in the data. Results may be improve.

Chapter 2

Method

2.1 Introduction

One of the most common data collection methods in research studies are surveys. As explained by Scheuren (2004) a survey is a method of collecting information from a sample of individuals. The sample is just a fraction or a subset of the population being studied. There are a variety of data collection methods available, these methods have their advantages and disadvantages hence it is important to choose wisely the appropriate method. Data can be collected using personal interviews, telephone interviews, mail surveys and web surveys as explained by Owens (2002).

2.1.1 Personal interviews

Face to face interviews can be used when participants are reachable easily. As explained by Owens (2002) this method generally brings high quality response, since it allows for the interviewer and the interviewee to engage in longer and complex interviews. Respondents are most likely to participate and its easy for them to fully concentrate on the interview. The disadvantage of face to face interviews is that respondents can respond feeling under pressure and sometimes feeling intimidated.

2.1.2 Telephone interviews

Collecting data using telephonic interviews is one way that is used in different situations, for instance covid-19 era with all the restrictions. This data collection method is advantageous in the sense that it is cheap to conduct, one can reach participants that are far easily and quickly. Data can be collected effectively and quickly, as respondents can respond immediately. The disadvantage with this method is that biased information can be collected, because this method is only effective in people with access to phones as explained by Colombotos (1969).

Sometimes the questionnaires can be too complex leading to participants unable to answer some questions.

2.1.3 Mail surveys

Sending mail to people to answer surveys is another useful method for collecting data. Generally it is less expensive to conduct surveys this way. One can send one survey questionnaire to a wide range of people to participate. Respondents can always consult with others to give the best response. With these advantages there are disadvantages as well. Oftentimes data collection can be slow resulting in a delay on the study. Most of the people will ignore survey received through emails resulting in the survey having a low response rate and a small sample size as stated by Owens (2002).

2.1.4 Web surveys

The advantages of collecting data using web surveys is almost the same as the mailing survey method. The down side of this method is that not all people have computers and other people might have connectivity challenges. These disadvantages can results in poor participation in the surveys.

It is noted that data collection or surveys can be done in may different ways. All the available methods come with their advantages and disadvantages. From these methods according to De Leeuw (2001) the collected data can have missing observations due to many contributing factors.

2.2 Non-response in Surveys

The biggest contributor in surveys to having missing data is the effect of non-response. Ignoring non-response in sample surveys or census can lead to poor quality results as explained by Gary (2007). As stated by Lohr (2021) the best way to handle non-response is to prevent it. However if non-response has occurred methods to handle non-response have to be employed. This is done in attempt to increase the accuracy of the results compared to when the non-response were unattended.

There are two types of non-response in surveys the unit non-response and item non-response according to Lohr (2021). Unit non-response is when the entire observation unit is missing. This means that for a particular survey a certain participant would not have responded to any of the survey questions. Item non-response is when some measurements are present but

not all of them have been provided. For example a survey participant can chose to respond to some of the survey questions but not all of them.

The two types of non-response can occur for many different reasons as explained by Lohr (2021). Unit non-response can occur when the interviewer is unable to reach certain participants because of issues such as connectivity or the person can choose not to participate in the survey entirely. A research that was done by Yan and Curtin (2010) discovered that unit non-response is most prevalent in household related surveys. This type of non-response can occur as a result of some survey questions being sensitive to the participants and hence they opt to skip those questions.

2.3 Classification of Missing Data

Methods of handling missing data are often applied at the data analysis stage. Before applying these methods the missing data needs to be classified according to the reasons for missingess. There are three categories used to classify this data, Missing At Random (MAR), Missing Completely At Random (MCAR) and Missing Not At Random (MNAR) as explained by Little and Rubin (1987).

2.3.1 Missing Completely At Random (MCAR)

According to Little and Rubin (1987) MCAR occurs when the missing observation is not related to any observation in the collected data. Conceptually data that are classified under MCAR most of the time are not linked to questions in the survey. For an example when a question is asked related to income denoted by x_1, another question related to profession denoted by x_2. Under MCAR the reason for x_1 to have a missing response is not because of x_1 nor x_2. As explained by Little (1988) when missing data are classified under MCAR, Little's Test of Missingness can be used to check if the missing data truly falls in the MCAR category.

2.3.2 Missing At Random (MAR)

Data that are classified as MAR are missing based on another observable case, such as underlying factors contributing to respondents not answering questions. MAR observations in simple terms are observations that can be predicted. Other survey participants can choose to skip some of the questions due to many underlying factors. For instance high earning people can choose not to disclose their earnings for security reasons. According to a research that was conducted by Turrell (2000) high income earners are more likely not to respond in income related questions. Consider the example below in Table 2.1 where 5 people answered

a survey relating to income based on age and gender. Responded 2 who is above 30 years old did not disclose his monthly income. Looking at the available information, one could attempt to predict that his monthly income is high, because looking at the available information all males over 30 years are high earners. Therefore in the case of MAR the reason for one variable to be missing can be as a result of another variable as explained in Aquilino (1991).

Table 2.1: Income based on age and gender

Responded	Gender	Age	Monthly Income
1	Male	over 30	High
2	Male	over 30	xxx
3	Male	over 30	High
4	Female	under 30	low
5	Female	under 30	low

2.3.3 Missing Not At Random (MNAR)

Data that are MNAR is identified when it does not meet the criteria of MCAR or MAR. Subjective data analysis is needed to classify data under the MNAR category, no statistical test are needed like with MCAR. Under MAR there may be correlation between an observable data and the reason of missing data, while under MNAR data can be attributed to an unobservable factor that directly contributes to the reason of having missing data. This can be the survey question that cause the missing response in the survey as explained Von Elm et al. (2007).

2.4 Methods of Handling Missing Data

Methods of handling missing have drastically increased in recent years as explained in Enders (2013). Researchers have a wide variety of methods to choose from in order to get the best out of surveys with missing data, that is by performing the proper statistical analysis to get a good representation of the data set. In the following subsections the traditional methods of dealing with missing data and the modern methods will be discussed.

2.4.1 Deletion Methods

Deletion methods are the most common methods of handling missing data in many research areas. This is because they are easy to implement and are standard choices in some statistical software packages.

List-wise deletion method (LD), focuses on removing data cases that have one or more missing observations. According to Enders (2010), the only advantage of LD is that it is convenient

and eliminates the need to use more sophisticated methods. The LD method assumes that the data are MCAR and can sometimes lead to misleading parameter estimates when this assumption does not hold. This method is argued to be wasteful because it just discards cases with missing values, affecting the study sample size, hence there is loss of valuable statistical information.

Pair-wise deletion (PD), attempts to reduce data loss by deleting cases on a case-by-case basis. When the data set has a low to moderate correlation, the pair-wise deletion performs better than the list-wise deletion method. This method also assumes that the data are MCAR. However if the data are not MCAR using the PD method can lead to poor results from the analysis of the data. The PD method has many unique problems as well. Example adapted from Enders (2010) using a variate of subsets of cases comes with problems of measures of association. To show this consider the sample co-variance formula

$$\sigma_{\hat{XY}} = \frac{\Sigma \left(x_i - \hat{\mu_X}\right)\left(y_i - \hat{\mu_Y}\right)}{N-1}. \tag{2.1}$$

PD utilises the subset of cases with complete data on X and Y to compute the co-variance. Most statistical packages use the same sub-sample to calculate the variate means. It is also possible to calculate $\hat{\sigma_X}$ from the cases that have data on X and to calculate $\hat{\sigma_Y}$ from the cases that have data on Y. The same problem arises when computing the denominator of the correlation coefficient,

$$r = \frac{\sigma_{\hat{XY}}}{\sqrt{\sigma_{\hat{X}^2}\sigma_{\hat{Y}^2}}}, \tag{2.2}$$

Statistical packages use the subset of cases with complete data on both X and Y to calculate the variance. As explained by Enders (2010) the latter approach is problematic since it can produce correlation values that are more than 1 or less than -1.

2.4.2 Single Imputation Methods

Single imputation method (SI) is considered as the predecessor of Multiple imputation method (MI) that will be discussed in the next section. This data handling method attempts to fill in missing data values by estimating the missing data points. These estimates are then filled into the data set with missing values for analysis as explained by Bennett (2001). The most common imputation methods are explained below.

Last value carried over method is commonly used in longitudinal data. Here the last value in the data set will be carried forward and then imputed for the missing value. This method is only applicable when the data are classified as MCAR.

Mean substitution is used when data are available about certain participants with a partic-

ular covariant for instance age. If the same covariant is missing for another participant, the mean will be imputed from the available data to fill in for the missing participant. This approach assumes that the data are MCAR.

Regression methods, this method involves creating a regression equation based on a complete variable from the same data set. This variable is treated as an outcome and all the other variables are used as predictor variables. For participants with missing observation, the predicted values from the regression equation created are then used as replacement. This method assumes data are MAR. Other single imputation methods are the Hot-deck imputation and Cold-deck imputation.

2.4.3 Multiple Imputation

This method was first proposed by Little and Rubin (1987). MI is a more robust technique for dealing with missingness in data sets compared to SI. MI method is a process that replaces missing data points by a vector $\boldsymbol{D} \geq 2$ of imputed values. The D values are arranged such that D complete data sets can be used to form a vector of imputations. As explained by Enders (2010) replacing each missing value by the second component units vector creates the second complete data set. When D sets of imputations are repeated random draws from the predictive distribution of missing values under a particular model for non-response, the D complete data inference can be merged to create one inference that correctly reflects uncertainty as a results of non-response under the model. When the imputations are from more than one model for the non-response, the merged inferences under the model can be contrasted across models to show the sensitivity of the inference to model the non-response.

The analysis of multiple imputed data set is direct. As explained by Enders (2010) each complete imputed data set is analysed using the same complete data technique that would be used in the absence of non-response. Let $\hat{\theta_d}, d = 1, ..., D$ be D complete data estimates and their associated variance for the estimated parameter θ, calculated from D repeated imputations on one model. The combined estimate is

$$\bar{\theta_D} = \frac{1}{D} \sum_{d=1}^{D} \hat{\theta}_d. \tag{2.3}$$

Since imputations in MI are the conditional draws rather than condition means, from good imputation models they provide good estimates for a wide range of estimates. The averaging over D imputed data sets in Equation (2.3) improves the estimates compared to those obtained from a single data set with conditional draws. The variability associated with these estimates comprises of two components, the within imputation variance

$$\bar{W_D} = \frac{1}{D}\sum_{d=1}^{D} W_d, \tag{2.4}$$

and the between imputation components,

$$B_D = \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\theta}_d - \bar{\theta_D}\right)^2. \tag{2.5}$$

The total variability associated with $\bar{\theta_D}$ is

$$T_D = \bar{W_D} + \frac{D+1}{D}B_D, \tag{2.6}$$

where $\left(1+\frac{1}{D}\right)$ is an adjustment for finite D. Therefore

$$\hat{\gamma} = (1+\frac{1}{D})\frac{B_D}{T_D} \tag{2.7}$$

is an estimate of the fraction of information about θ missing because of non-response. For big sample sizes and scalar θ, the relevant distribution for the interval estimates and significance test is a t distribution,

$$\left(\theta - \hat{\theta_D}\right) T^{\frac{1}{2}} \tilde{} \; t_v \tag{2.8}$$

where v are the degrees of freedom given by

$$(D-1)\left(1+\frac{1}{D+1}\frac{\bar{W_D}}{B_D}\right)^2, \tag{2.9}$$

based on the s Satterwaite approximation according to Rubin and Schenker (1986). For small data set the degrees of freedom to be used are given by

$$v* = \left(v^{-1} + \hat{v_{obs}^{-1}}\right)^{-1}, \tag{2.10}$$

where

$$\hat{v_{obs}} = (1-\hat{\gamma_D})\left(1+\frac{v_{com}+1}{v_{com}+3}\right)v_{com}, \tag{2.11}$$

and v_{com} represents the degrees of freedom for approximate or exact t inferences about θ when there are no missing values as explained by Barnard and Rubin (1999).The data set are then analysed using standard statistical techniques.

The use of MI has become more prevalent in surveys and non-survey studies, according to Little and Rubin (1987) and Schafer (1997). One of the advantages of MI compared to the likelihood method as given in Section 2.4.4 is that it is easy to implement. Little and Rubin (1987) further showed that the efficiency of an estimate based on n imputations is given as

$$E_{ff} = (1 + \frac{\lambda}{n})^{-1} \tag{2.12}$$

where λ is the fraction of information for the quantity being tested that is missing and E_{ff} is the relative efficiency of using this method. Table 2.2 below shows the relative efficiency of the MI method with different λ values and different number of imputations. The values in the table are produced from applying Equation 2.12 The table shows that for small fraction of missing values and more imputations the MI method is more effective.

Table 2.2: Relative Efficiency

	λ				
n	10%	20%	30%	50%	70%
3	0.9677	0.9375	0.9091	0.8571	0.8108
5	0.9804	0.9615	0.9434	0.9091	0.8772
10	0.9901	0.9804	0.9709	0.9524	0.9346
20	0.9950	0.9901	0.9852	0.9756	0.9662

Figure 2.1 below summarises how the multiple imputation method works the diagram was adopted from Enders (2010).

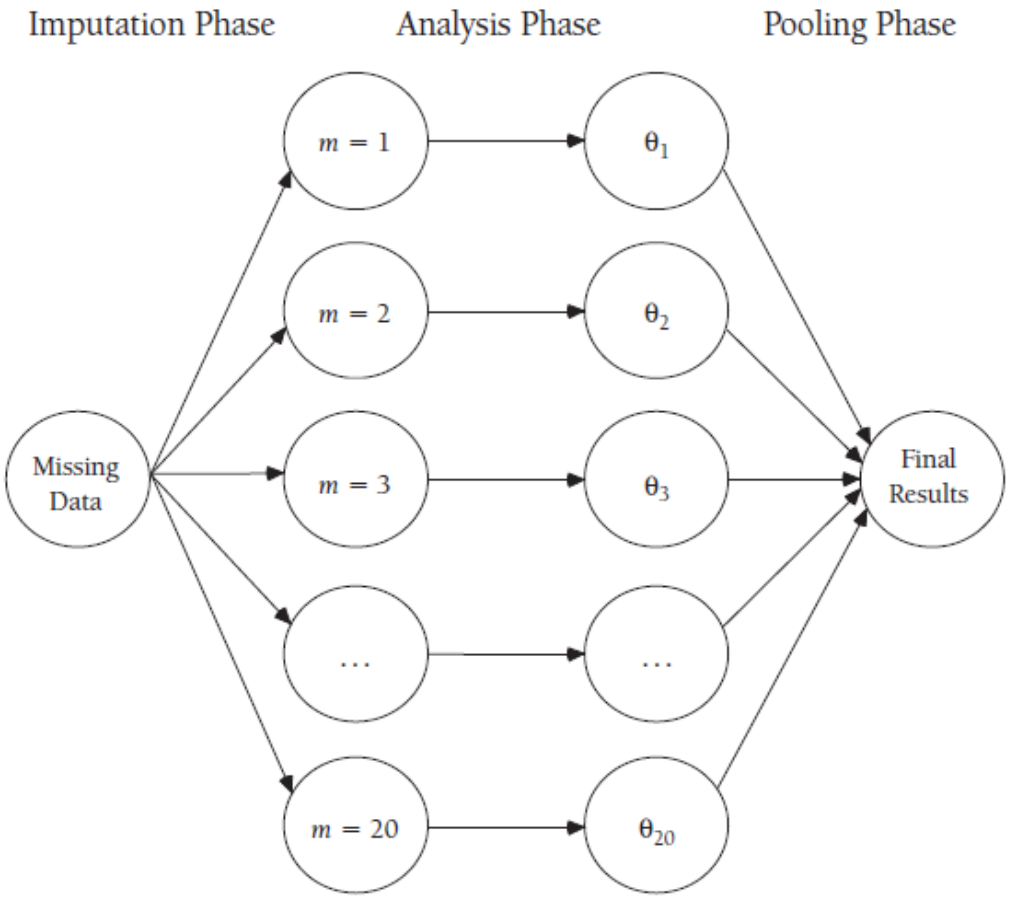

Figure 2.1: Multiple imputation steps

As seen in the figure above the multiple imputation method consist of three main steps the imputation phase, the analysis phase and the pooling phase. In the imputation phase the data set containing missing values is duplicated into multiple copies and the missing values are imputed with different missing values in a predictive framework. During the analysis phase the model parameters are estimated using the complete data sets generated from the imputation phase. The last phase which is the pooling phase the parameter estimates and standard errors are combined into one complete results with no missing values.

2.4.4 Likelihood Based Methods

The likelihood based methods are more robust compared to the imputation methods discussed in section 2.4.3 above and they are argued to have good statistical properties. Pigott (2001) explains that when a researcher has a complete data set, it is easy to use statistical methods to calculate quantities like the mean and linear regression coefficients. Pigott (2001) further explains that these quantities are maximum likelihood estimates, based on maximising the likelihood of the observed data. The same approach is applicable when the observed data contains missing observations. Dempster et al. (1977) proposed methods to find estimates when the data set has missing observations. These methods are discussed below.

Expectation Maximisation (EM) approach is an iterative procedure consisting of two steps in each iteration, the expectation step and the maximisation step. The EM method aims to estimate the missing values in the data set. In the expectation step, the distribution of the missing values are determined using the available data. In the maximisation step the expected values for the missing data from the expectation step are used. The maximum likelihood function is then computed as if there is no data missing to estimate the new parameter. The new estimated parameters are then substituted back to the expectation step and a further maximisation step is performed. The process iterates in the these two steps until convergence is obtained. This method assumes that the data are MAR and according to Schafer (1997) most statistical packages such as SAS and SPSS use this method.

Raw Maximun Likelihood (RML) method uses the available information, such as the mean and the variance of the data to generate estimates of the missing data using the maximum likelihood estimation. This method only calculates means and variances for the available covariates. Then a statistical package uses this as imputed for further analysis. The RML method is similar to the EM method discussed above, except that it does not have the expectation step and it converges faster. If the data are MAR, the RML method generates means and variances that are less biased compared to other methods. The RML method is available in software packages such as SAS, SPSS, and structural modeling programs such as AMOS and LISREL as explained in Arbuckle (1995) and Jöreskog and Sörbom (1989).

There are complications that arise from dealing with the process that creates missing data. Rubin (1976) provides a mathematical precise treatment that are not based on the likelihood. According to Rubin (1976) if $\boldsymbol{Y}$ denotes a complete data set with no missing values, then $\boldsymbol{Y} = (Y_{obs}, Y_{mis})$, where Y_{obs} indicates the observed values and Y_{mis} indicates the missing values. Let $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$ be the density function of the joint distribution of Y_{obs} and Y_{mis}. The marginal probability density of Y_{obs} can be obtained by integrating out the missing data Y_{mis},

$$f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}. \tag{2.13}$$

The likelihood of θ based on Y_{obs} ignoring the missing data mechanism is defined to be the function of θ proportional to $f(Y_{obs}|\theta)$,

$$L_{ign}(\theta|Y_{obs}) \propto f(Y_{obs}|\theta), \quad \theta \in \Omega_\theta. \tag{2.14}$$

Inference about θ can be based on the $L_{ign}(\theta|Y_{obs})$, giving the mechanism that leads to incomplete data to be ignored as discussed below. Ignorable RML estimates are obtained by maximising Equation 2.14 with respect to θ. Ignorable Bayes inference for θ based on Y_{obs} is obtained by including a prior distribution $p(\theta)$ for θ and the inference will be on the posterior distribution

$$L_{ign}(\theta|Y_{obs}) \propto p(\theta) \times L_{ign}(\theta|Y_{obs}). \tag{2.15}$$

Generally an indicator function can be included in the distribution indicating whether each element of $\boldsymbol{Y}$ is missing or observed. Define the indicator function of $\boldsymbol{Y}$ taking values 1 and 0, 1 representing missing values and 0 indicating observed values. As indicated by an example from Enders (2010) if $\boldsymbol{Y} = y_{ij}$ for a matrix with n observations measured for k variables, one can get

$$M_{ij} = f(x) = \begin{cases} 1, & y_{ij} \text{ missing}, \\ 0, & y_{ij} \text{ observed}. \end{cases} \tag{2.16}$$

The complete model treats M as a random variable and specifies the joint distribution of M and Y. The distribution mechanisms for missing data which is indexed by an unknown parameter ψ and $\Omega_{\theta,\psi}$ is the space parameter of θ, ψ

$$f(Y, M|\theta, \psi) = f(Y|\theta) \times (M|Y, \psi), \quad (\theta, \psi) \in \Omega_{\theta,\psi}. \tag{2.17}$$

Sometimes the distribution of the missing data mechanism is know and ψ is redundant. More specification on the joint distribution are discussed in Enders (2010). The actual observed data have values of the variables (Y_{obs}, M). To get the distribution of the observed data the

joint density of Y and M has to be integrated with respect to Y_{mis}. That is,

$$f\left(Y_{obs}, M|\theta, \psi\right) = \int f\left(Y_{obs}, Y_{mis}|\theta\right) f\left(Y_{obs}, Y_{mis}, \psi\right) dY_{mis}. \tag{2.18}$$

The likelihood function of θ and ψ is any function of the two parameters θ and ψ proportional to Equation 2.18.

$$L\left(\theta, \psi|Y_{obs}, M\right) \propto f\left(Y_{obs}, M|\theta, \psi\right). \left(\theta, \psi\right) \in \Omega_{\theta,\psi} \tag{2.19}$$

The question now is when inference on θ will be based on Equation 2.19 or on Equation 2.14 Note that if the distribution of the missing data mechanism is independent on the missing values, that is

$$L\left(M|Y_{obs}, Y_{mis}, \psi\right) = f\left(M|Y_{obs}, \psi\right) \text{for all } Y_{mis}, \tag{2.20}$$

then Equation 2.18 becomes

$$f\left(Y_{obs}, M|\theta, \psi\right) = \int f\left(Y_{obs}, Y_{mis}|\theta\right) f\left(Y_{obs}, Y_{mis}, \psi\right) dY_{mis}. \tag{2.21}$$

$$= f\left(M|Y_{obs}, \psi\right) f\left(Y_{obs}|\theta\right). \tag{2.22}$$

In practical application θ and ψ are different such that the joint space parameter of (θ, ψ) is the product of the space parameter of θ and ψ, $\Omega_{\theta,\psi} = \Omega_\theta \times \Omega_\psi$. When the data are MAR and the two space parameter are different, then the likelihood based inference for θ from Equation 2.19 and the likelihood based inference for θ from Equation 2.15 will be the same since the likelihoods are proportional to each other.

Figure 2.2 shows how to classify missing data and which method to use in dealing with missing data in research. For missing data that are MCAR possible methods that can be used are the Single imputation method, Multiple imputation and Deletion methods. If the missing data are MAR likelihood method is the appropriate method to use. Markov Chain imputation and the pattern mixture models can be used when the missing data does not hold for the assumptions of MCAR and MAR an explained by Bennett (2001).

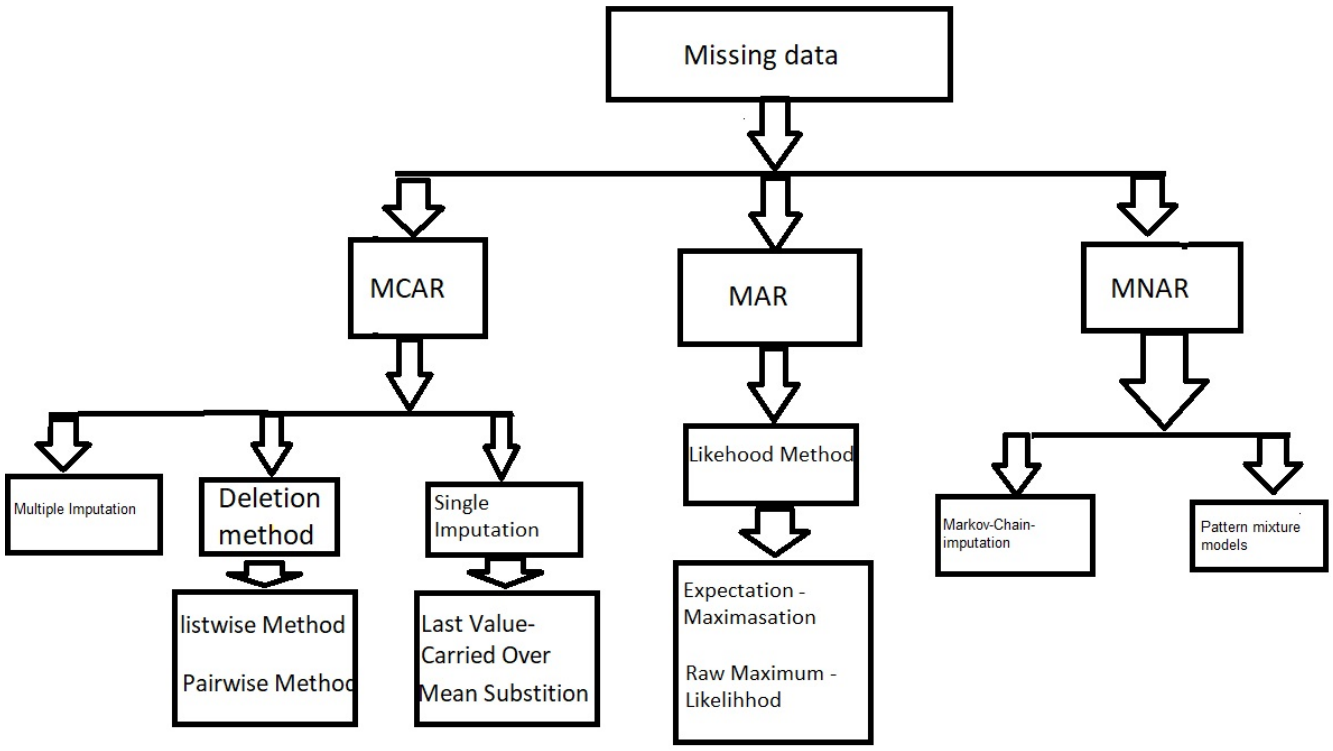

Figure 2.2: Decision tree for classifying missing data

2.5 Data Analysis Method

This section presents the data analysis methods that will be used in Chapter 3. The analysis performed will then be used to compare the performance of the methods of handling missing data discussed in Section 2.4 above. There are a variety of statistical methods available to choose from to perform different analysis, depending on the data at hand and research objective. In the following section analysis of variance (ANOVA) will be performed to analyse a synthetic data set, to assess if there is a statistically significant difference in the mean values of the variable among the groups. The analysis will be performed using RStudio® version 4.2.2.

ANOVA was developed and introduced in 1925 by Ronald A. Fisher as explained by Kao and Green (2008). The aim of ANOVA is to identify whether differences exist between means for more than two groups.There are three assumptions that needs to be satisfied before performing ANOVA. The first assumptions is that the data must be normally distributed in all the groups being tested. Multiple methods can be used to check for normality as explained by Kao and Green (2008) such as formal tests like the Kolmogorov-Smirnov test or Shapiro-Wilks test. The second assumption is that the variance in all the groups are equal. The Levene test can be used to assess homogeneity of variance. The last assumptions is that the observations are independent, that is they are not correlated.

After ANOVA is performed and if the results suggest that a significant difference between means exists, then post-hoc analysis can be performed to determine which groups are different. As explained by Homack (2001) post-hoc tests are conducted to determine which groups differ if the null hypothesis of equal means is rejected. From an example given by Homack (2001), if there are three groups it can happen that two groups have equal means and the third group mean is different from the two groups , post-hoc tests can be used to determine which groups differ.

Chapter 3

Data Analysis, Results and Discussions

3.1 Introduction

This chapter presents the data analysis of the simulated data set with nested designs. This data set was simulated using RStudio® version 4.2.2. As explained in Section 2.5 the statistical method that will be used to analyse the data set is analysis of variance (ANOVA) in a mixed effects modeling framework. In order to perform the ANOVA certain assumptions need to be met, as discussed in Section 2.5. In practice generally it is not easy to get a data set that meets all the assumptions for the ANOVA, because of this reason a synthetic data set will be used. Kellner et al. (1999) explains the reason of simulating data. In particular simulation is done as an aid for decision making and to mitigate risks of encountering problem when analysing a data set that may not be well structured. The data set was simulated in the context of of a hypothetical educational scenario, where the effect of teaching methods and schools on student test results is being studied.

The data set consists of three teaching methods method 1, method 2 and method 3. Two schools school A and school B, are included with each school having three classes (nested within the schools), whereas each class has 60 students. The test results (in percentage) are considered the response variable. The complete data set will be analysed, then missing values will be systematically introduced in the data set to satisfy the missing data mechanisms. The introduced missing values will be handled using the appropriate methods guided by Figure 2.2, then the results will be discussed.

3.2 Analysis

In this section the original complete data set with no missing observations will be analysed. This will be followed by analysis on the same data set with missing values introduced in Section 3.3. As explained in section 2.5 the analysis will be done using RStudio® version 4.2.2.

Table 3.1: Summary statistics for the test scores

Min	1st Qu	Median	Mean	3rd Qu	Max
55.00	67.75	75.50	75.50	82.00	97.00

Table 3.1 shows the summary statistics of the test scores from the two schools with three different teaching methods being used. The table shows that the data set is normally distributed around the mean.

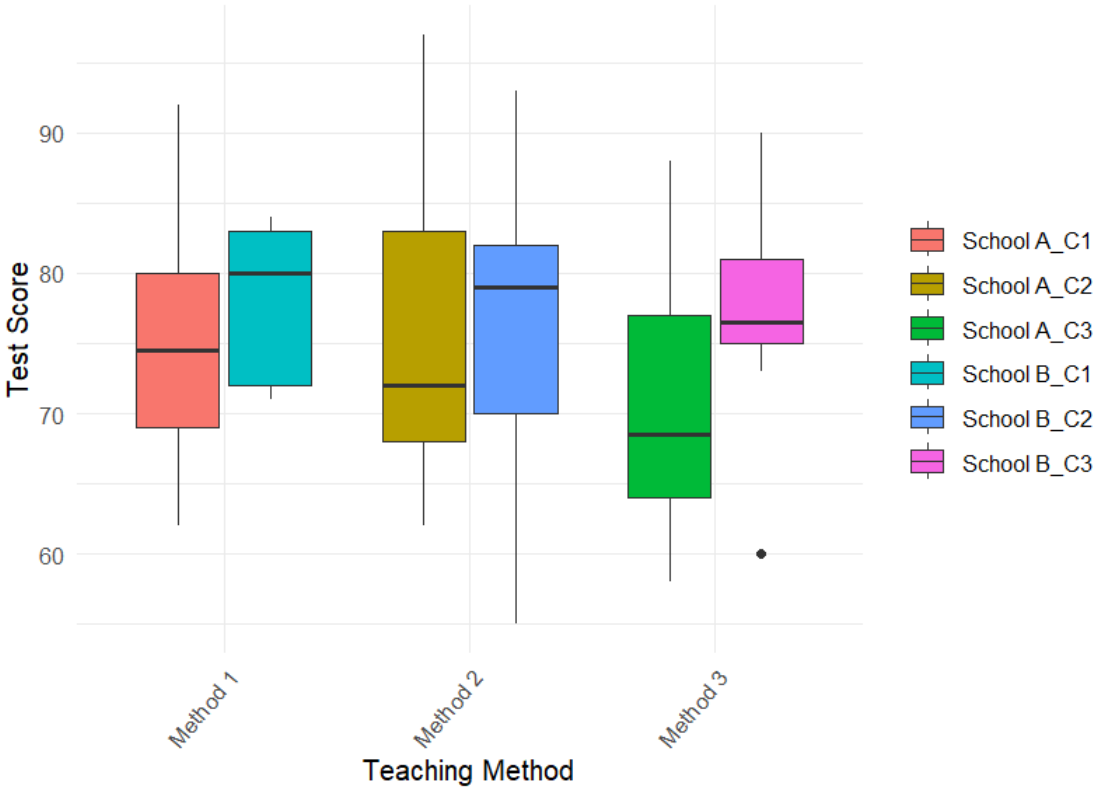

Figure 3.1: Box plot of test scores by teaching method and class

Figure 3.1 shows the box plots of the test scores from the different schools with different teaching methods. It is clear from the box plots that the test scores do not have outliers in teaching method 1 and teaching method 2 and teaching method 3 only has one outlier from school B.

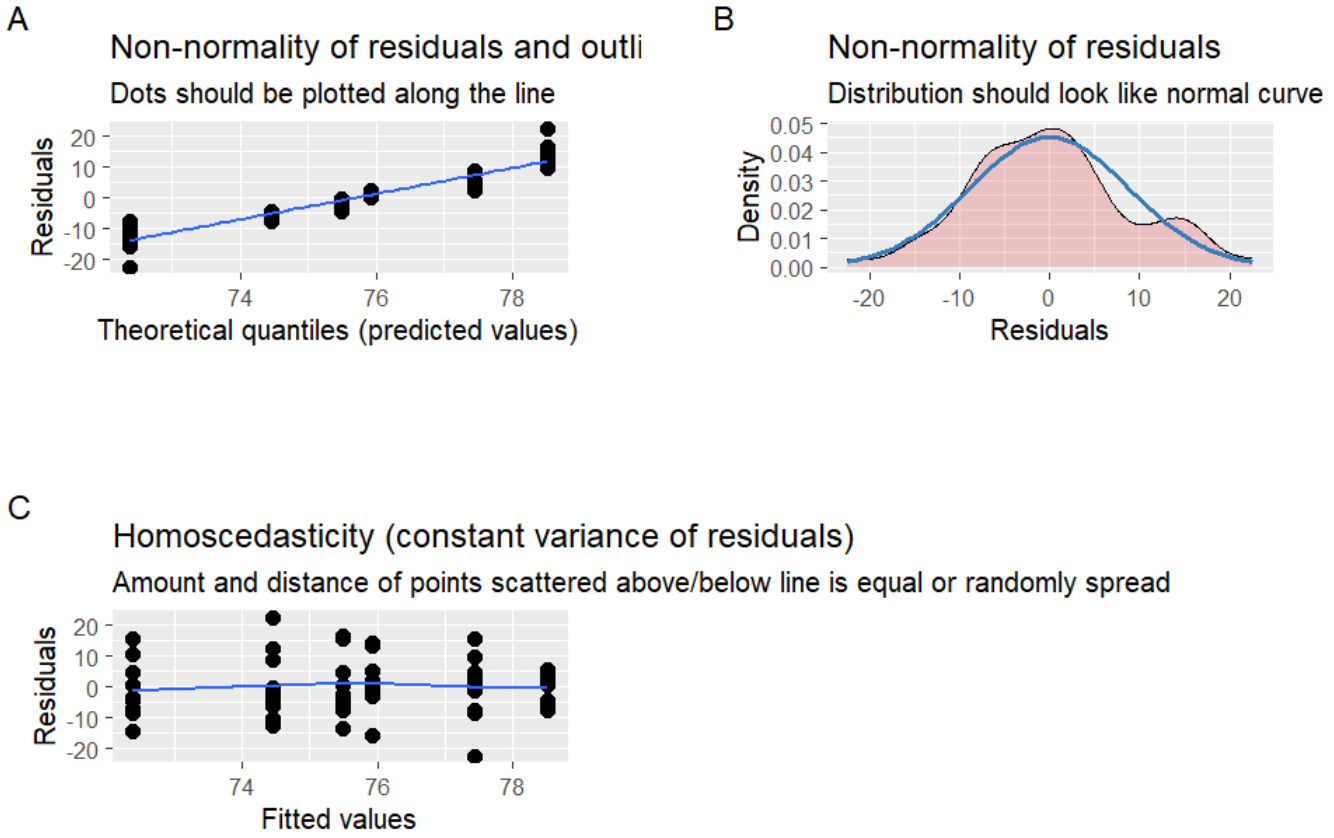

Figure 3.2: Normality plots check

Performing the shapiro wilk's test, on the test scores, $p - value = 0.2585$, at the 5% level significant conclusion can be made that the data is normally. The shapiro wilks's test results is consistent with the conclusion made from 3.1, that the data set is normally distributed. Performing the leveneTest on the data set to test for equality of variances $p - value = 0.0.2931$, at the 5% level significant this means that the data set meet the assumption of homoscedasticity. This is further supported by the Figure 3.2 above which shows no violation of the assumptions of ANOVA. Figure 3.2 A showing the residuals vs theoretical quantiles the dots are plotted along the straight line, Figure 3.2 B shows the distribution looks like a normal curve. Further more Figure 3.2 C shows that the data set is homoscedastic. Assuming the observation are independent ANOVA will be performed.

The data set will be analysed by fitting a linear mixed effect model by restricted maximum likelihood (REML) .

From the fitted model these results are obtained for the random effect (school), the estimated standard deviation is approximately 2.3906, this indicates that there is random variability in test scores between different schools. For classes within the schools the standard deviation is 0.6132, this indicates that there is random variability in test scores between different classes

within the same school. The standard deviation of the residual variability is 8.9089.

Table 3.2: Fixed effect results

	Value	Std.Error	p-value
Intercept	77.00	2.090056	0.0000
Teaching_Method 2	-1.05	1.738294	0.6072
Teaching_Method 3	-2.85	1.738294	0.2428

Table 3.2 shows the fixed effect results from the fitted linear mixed effect model. The table presents the coefficients and statistical information for the predictor variable Teaching Method in the model. The Intercept is significant with $p - value = 0.000$. There are two levels of the Teaching Method variables, Method 2 and Method 3. The coefficients for these levels are -1.05 and -2.85, respectively relative to Method 1.The $p - values$ indicate that they are not significant compared to Method 1 at 5 % level of significance.

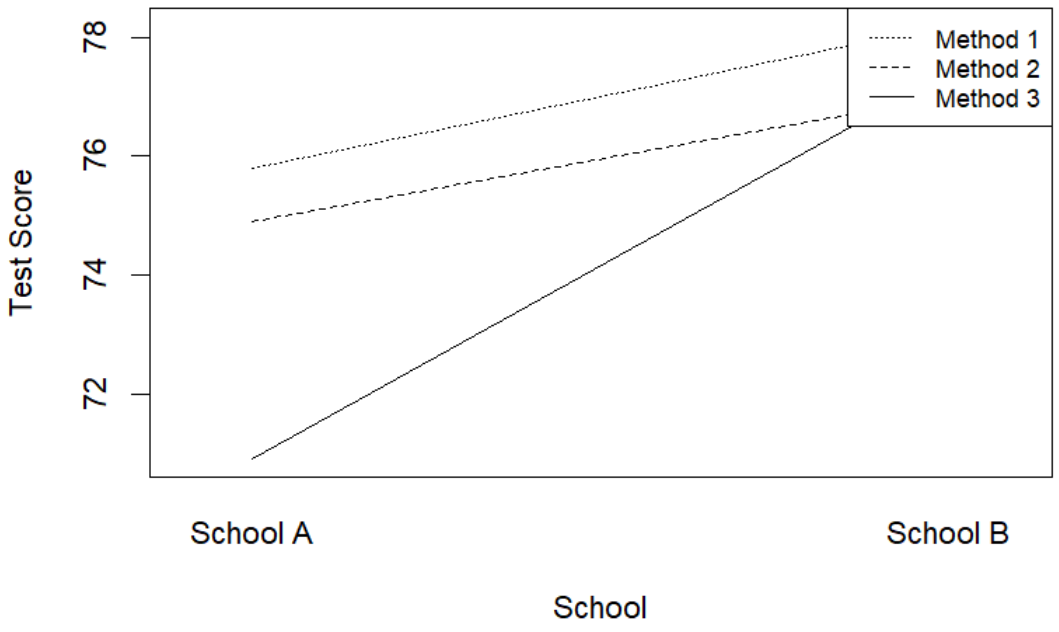

Figure 3.3: Interaction plot for original data set

To further investigate the interaction effect of the teaching methods per school on test scores. Figure 3.3 shows an interaction plot between the three different teaching methods on test results. The results show no significant interaction between Method 1 and Method 2 whereas there is significant interaction between Method 2 and Method 3. Teaching Method 3 has a steeper slopes than Method 1 and Method 2 suggesting that Method 3 has a stronger effect

on the test scores.

Table 3.3: Model fit table

AIC	BIC	LogLik
1303.101	1322.158	-645.5506

Table 3.3 will be discussed in Section 3.4.

3.3 Missing Data Introduced

This section focuses on analysing the synthetic data with missing values introduced. In this section the methods discussed in Chapter 2 will be implemented.

Table 3.4: Summary statistics with missing values

Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
55.00	69.00	76.00	75.54	82.00	97.00	45

Table 3.4 show the summary statistics of the data with missing values. The original synthetic data set has 180 observations and 25% of missing values will be systematically introduced in all the teaching methods. This will further mean that the total observation of test results will have 25% of missing values, the new number of observations will be 135. Faber and Fonseca (2014) explain that doing analysis on an incomplete data set can at times give false and inaccurate results, such as accepting that the research hypothesis is true when it is actually false and vice versa. The developed methods of dealing with the missing data will be used to reduce risks of getting inaccurate results from analysing the data set with missing values.

3.3.1 Deletion Method

Assuming the introduced missing values on the data set are MCAR, the deletion method is performed. After performing the deletion method the data set will be analysed by fitting a linear mixed effect model by REML.

From the fitted model results, the estimated standard deviation is approximately 2.0840, indicating random variability in the test scores between different schools. For classes within the schools, the standard deviation is 0.00051, indicating a small random variability in test scores between other classes within the same school. The standard deviation of the residual variability is 8.8923.

Table 3.5: Fixed effect results for deletion method

	Value	Std.Error	p-value
Intercept	76.83262	1.983550	0.0000
Teaching_Method 2	-2.25305	1.874907	0.3525
Teaching_Method 3	-2.00000	1.874662	0.3978

Table 3.5 shows the fixed effect results from the fitted linear mixed effect model by REML. The table shows the coefficients and statistical information for the predictor variable Teaching Method. The intercept is significant with $p - value = 0.000$. There are two levels of the Teaching Method variables, the coefficients for these levels are -2.25305 and -2.00000, respectively relative to Method 1. The $p - values$ indicate that they are not significant compared to Method 1 at 5 % level of significance.

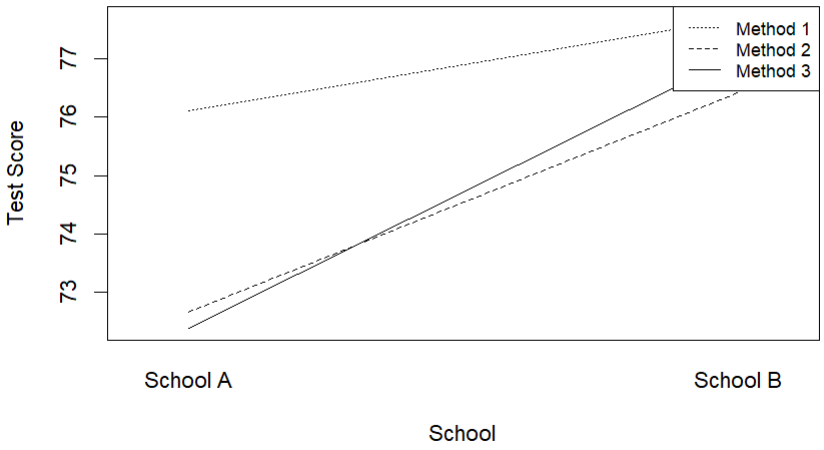

Figure 3.4: Interaction plot for deletion method

Figure 3.4 shows an interaction plot between the three different teaching methods on test results. The results show no significant interaction between method 1 and method 2 whereas there is a significant interaction between method 2 and method 3 since these two lines do cross at a particular point.

Table 3.6: Model fit table with deletion method

AIC	BIC	LogLik
976.4504	993.7472	-482.2252

Table 3.6 will be discussed in Section 3.4.

3.3.2 Single Imputation Method

Since the data is normally distributed, the mean imputation will be the more appropriate method to use. A mixed effects model will be fitted from data with missing data mean imputed.

The results obtained show an estimated standard deviation of approximately 1.560306, indicating that random variability exists in test scores between different schools. For classes within the schools the standard deviation is 0.000476, indicating a small random variability in test scores between different classes within the same school. Standard deviation of the residual variability is 7.7182.

Table 3.7: Fixed effect results for single imputation method

	Value	Std.Error	p-value
Intercept	76.61852	1.486655	0.0000
Teaching_Method 2	-1.73333	1.409160	0.3437
Teaching_Method 3	-1.50000	1.409160	0.3986

Table 3.7 shows the fixed effect results after performing single imputation. The table shows the coefficients and statistical information for the predictor variable. The intercept for the model is 76.61852, with a standard error of 1.486655. $P - value = 0.0000$ indicate that it is statistically significant at 5% level of significance. There are two levels of the Teaching Method variables, Method 2 and Method 3. The coefficients for these levels are -1.73333 and -1.50000, respectively relative to Method 1.The $p - values > 0.05$ this indicates that the other two teaching methods are not significant compared to Method 1 at 5 % level of significance.

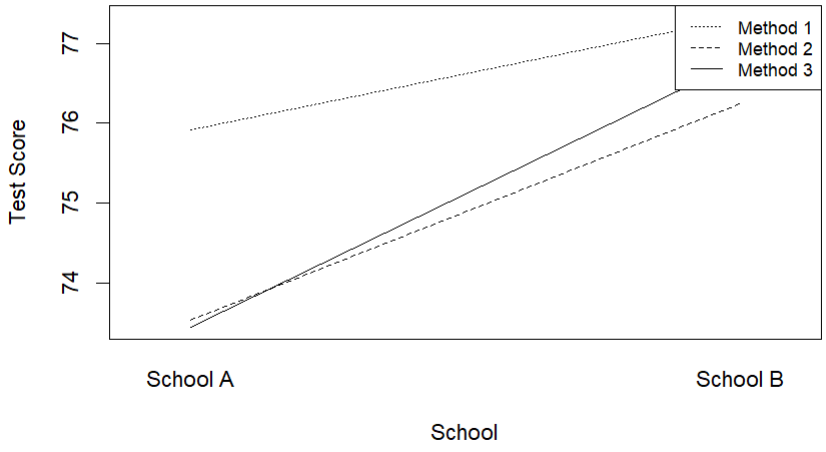

Figure 3.5: Interaction plot for single imputation method

Figure 3.5 looks the similar to Figure 3.4, therefore the same observation as in Figure 3.4

above can be made.

Table 3.8: Model fit table with single imputaion method

AIC	BIC	LogLik
1251.562	1270.619	-619.781

Table 3.8 will be discussed in Section 3.4.

3.3.3 Multiple Imputation Method

Multiple imputation (MI) is performed to impute and fill in the missing values, then the results will be analysed.

From the fitted model the estimated standard deviation is approximately 1.151354 indicating that there is random variability in test scores between different schools. For classes within the schools the standard deviation is 0.000467 indicating that there is random variability in test scores between different classes. The standard deviation of the residual variability is 8.2935.

Table 3.9: Fixed effect results for multiple imputation method

	Value	Std.Error	DF	t-value	p-value
Intercept	75.98333	1.345064	174	56.49048	0.0000
Teaching_Method 2	1.01667	1.514193	2	0.67142	0.5711
Teaching_Method 3	0.50000	1.514193	2	0.33021	0.7726

Table 3.9 shows the fixed effect results model. The table presents the coefficients and statistical information for the predictor variable Teaching Method. The Intercept for the model is 75.98333, with a standard error of 1.345064.The $p - value = 0.0000$ indicating is statistically significant intercept at 5% level of significance. There are two levels of the Teaching Method variables, Method 2 and Method 3. The coefficients for these levels are 1.01667 and 0.50000, respectively relative to Method 1. The $p - values > 0.05$ indicating that the other two teaching methods are less significance compared to Method 1 at 5 % level of significance.

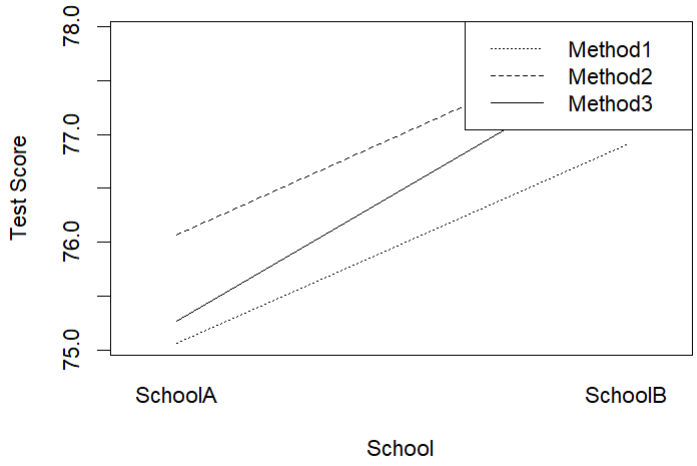

Figure 3.6: Interaction plot for multiple imputation method

Figure 3.6 shows the interaction plot after performing the multiple imputation. The plot shows that there is no significant interaction between Method 1 and method 2. Method 3 is more steeper than the other two methods showing possible interaction of method 3 with the other methods. Overall from this figure a conclusion can be made that Method 3 do interact with Method 1 and Method 2.

Table 3.10: Model fit table with deletion method

AIC	BIC	LogLik
1276.474	1295.53	-632.2368

Table 3.10 will be discussed in Section 3.4.

3.4 Results and Discussion

This section presents and discusses the results obtained in Section 3.3. The discussion will based on comparing the results from the different methods of handling missing data, together with results obtained from the complete data set.

Table 3.11: p-values for analysed data

	Complete data	Deletion method	SI method	MI method
Intercept	0.00	0.00	0.00	0.00
Teaching method 2	0.6072	0.3525	0.3437	0.5711
Teaching method 3	0.2428	0.3978	0.3986	0.7726

Table 3.11 shows the $p-values$ of the results obtained from the fitted mixed effect model. From the complete data set and with the data handling methods discussed above after introducing missing values. Comparing the $p-values$ obtained from different methods to assess

how these methods impact the significance of the results compared to the complete data results. It is noted that in all cases, the $p-value$ for the intercept is 0.00 which suggests that the intercept is highly statistically significant across all methods. For Method 2, all three methods applied have lower $p-values$ compared to the complete data results. This indicate that the effect of Method 2 is more statistically significant when using imputed data compared to complete data. For Method 3, the complete data method has the lowest $p-value$, indicating a higher level of significance compared to the imputed data methods. This suggests that the effect of Method 3 may be less statistically significant when data handling methods are used.

In comparing the different methods of handling missing data looking specifically at Table 3.11 it is not clear to see which missing data handling method performs better. To see which method on the fitted model fits the data better than the other methods the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) will be used.

Table 3.12: AIC and BIC values

	Complete Analysis	Deletion Method	SI Method	MI Method
AIC	1303.101	976.4504	1251.562	1276.474
BIC	1322.158	993.7472	1270.619	1295.53

In statistical analysis the results in Table 3.12 are used to check which model fits the data best. In particular a fitted model with the smallest AIC or BIC value is considered better. In this paper the AIC and BIC values will be used to compare model fit for the models obtained from the different methods of handling missing data against the complete case analysis. The complete data set without missing values has AIC = 1303.101 and BIC = 1322.158, because this results are obtained from complete data set, the results will be considered good. The method of handling missing data that have results that are closest to the complete data is considered best. With reference to Table 3.12, it can be seen that the MI method dealt with the missing values far much better that the deletion and SI method. That is when comparing the Deletion method, SI method, and MI method to the complete analysis results, the MI method results are almost the same as the complete analysis results compared to the other 2 methods.

Chapter 4

Conclusion and Future Studies

4.1 Concluding remarks

When faced with missingness in a data set it is important to carefully consider what method to use to handle the missing values. Missing data can be classified into three main categories as explained by Rubin (1976), MCAR, MAR and MNAR. It is essential to correctly categorise missing data. This is because correctly identifying the type of missingness will lead to the correct method of handling missing data. If missing values are incorrectly classified, a wrong choice of a method for handling the missing data can be used resulting in misrepresentation of the data. From this misleading results can be obtained for the study.

In this paper deletion, single imputation, multiple imputation and maximum likelihood method were discussed. A simulated complete data set was obtained, missing values were introduced systematically to the data and methods of handling missing values were applied. The results obtained from using the different methods all differ compared to the results obtained when the data did not have missing vallues. The analysis shows that from the applied methods of handling missing data the multiple imputation fits the data better compared to the other methods used. The analyses were done in RStudio® version 4.2.2.

4.2 Future studies

This paper was based on a synthetic data set and the data analysis method that was used was ANOVA and interaction plots were considered. From the synthetic data set missing values were introduced and the deletion, single imputation and multiple imputation methods were used to handle the missing values. Real-world data set can be used in future studies and other methods of handling missing data can be considered, such as the likelihood based methods and other methods. Since in real-world data set generally it is not easy to get a data set that is normally distributed, other statistical methods can be used including non-parametric methods.

References

Aquilino, W. (1991). Telephone Versus Face-to-Face Interviewing for Household Drug use Surveys. *International Journal of the Addictions*, 27(1):71–91.

Arbuckle, J. L. (1995). Amos for Windows: Analysis of Moment Structures (Version 3.5)[Computer Software]. Chicago: SmallWaters.

Barnard, J. and Rubin, D. B. (1999). Miscellanea. Small-Sample Degrees of Freedom with Multiple Imputation. *Biometrika*, 86(4):948–955.

Bennett, D. A. (2001). How Can I Deal with Missing Data in my Study? *Australian and New Zealand Journal of Public Health*, 25(5):464–469.

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, 84(4):487–508.

Colombotos, J. (1969). Personal Versus Telephone Interviews: Effect on Responses. *Public Health Reports*, 84(9):773.

De Leeuw, E. (2001). Reducing Missing Data in Surveys: An Overview of Methods. *Quality and Quantity*, 35:147–160.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Publications.

Enders, C. (2013). Dealing with Missing Data in Developmental Research. *Child Development Perspectives*, 7(1):27–31.

Faber, J. and Fonseca, L. M. (2014). How Sample Size Influences Research Outcomes. *Dental Press Journal of Orthodontics*, 19:27–29.

Gary, P. R. (2007). *Adjusting for Nonresponse in Surveys*, pages 411–449. Springer Netherlands, Dordrecht.

Hansen, M. H. and Hurwitz, W. N. (1946). The Problem of Non-Response in Sample Surveys. *Journal of The American Statistical Association*, 41(236):517–529.

Homack, S. R. (2001). Understanding what ANOVA Post Hoc Tests Are, Really.

Jöreskog, K. and Sörbom, D. (1989). *LISREL 7: A Guide to The Program and Applications*. Spss.

Kao, L. S. and Green, C. E. (2008). Analysis of Variance: Is there a Difference in Means and What Does it Mean? *Journal of Surgical Research*, 144(1):158–170.

Kellner, M., Madachy, R., and Raffo, D. (1999). Software Process Simulation Modeling: Why? What? How? *Journal of Systems and Software*, 46(2-3):91–105.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data With Missing Values. *Journal of the American statistical Association*, 83(404):1198–1202.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. John Wiley and Sons.

Lohr, S. L. (2021). *Sampling: Design and Analysis*. CRC press.

Owens, L. (2002). Introduction to Survey Research Design. In *SRL Fall 2002 Seminar Series*, volume 1.

Pigott, T. D. (2001). A Review of Methods for Missing Data. *Educational Research and Evaluation*, 7(4):353–383.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63(3):581–592.

Rubin, D. B. and Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples With Ignorable Nonresponse. *Journal of the American Statistical Association*, 81(394):366–374.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. CRC Press.

Scheuren, F. (2004). What is a Survey? American Statistical Association Alexandria.

Turrell, G. (2000). Income Non-Reporting: Implications for Health Inequalities Research. *Journal of Epidemiology & Community Health*, 54(3):207–214.

Von Elm, E., Douglas, G., Egger, M., Stuart, J., Peter, C., and Jan, P. (2007). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies. *The Lancet*, 370(9596):1453–1457.

Yan, T. and Curtin, R. (2010). The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research*, 22(4):535–551.

Appendix : R Script, Chapter

```r
install.packages("nlme")

install.packages("tidyverse")

install.packages("sjPlot")

install.packages("gridExtra")

install.packages("ggplot2")

library(nlme)

library(tidyverse)

library(sjPlot)

library(ggplot2)

# Set random seed

set.seed(123)

# Generate the data

num_students <- 60

num_classes_per_school <- 3

num_schools <- 2

teaching_methods <- c("Method 1", "Method 2", "Method 3")


nestdata <- data.frame(

  Student = paste("S", 1:num_students, sep=""),

  Class = rep(paste("C", 1:num_classes_per_school, sep=""),
```

```
                each=num_students/(num_classes_per_school*num_
                    schools)),

  School = rep(rep(paste("School", c(" A", " B"), sep=""),

                   each=num_students/(num_classes_per_school*num_
                       schools))),

  Teaching_Method = rep(rep(teaching_methods,

                            each=num_students/(num_classes_per_
                                school*num_schools)),

                        times=num_classes_per_school*num_schools)
                            ,

  Test_Score = round(rnorm(num_students, mean=75, sd=10))
)

nestdata$Class <- with(nestdata, paste(School, Class, sep="_"))

# Saving the data set

write.csv(nestdata, "Modified_dataset.csv", row.names = FALSE)

cat("Dataset saved as 'modified_dataset.csv'\n")


data.nest <- read.csv("C:/Users/siyab/OneDrive/Documents/modified
    _dataset.csv")

# Create a box plot

ggplot(data.nest, aes(x = Teaching_Method, y = Test_Score, fill =
     Class)) +

  geom_boxplot() +

  labs(x = "Teaching Method", y = "Test Score", fill = "") +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 50, hjust = 1))

# Fit a mixed-effects model
```

```
data.nest.lme <- lme(Test_Score ~ Teaching_Method, random = ~ 1 |
    School/Class,
                     data = data.nest)

summary(data.nest.lme)


interaction.plot(data.nest$School, data.nest$Teaching_Method,
   data.nest$Test_Score,

                 xlab = "School", ylab = "Test Score", legend =
                    FALSE)

unique_methods <- unique(data.nest$Teaching_Method)

# Define line types and colors for the legend

line_types <- c(3, 2, 1)  # 3 = solid, 2 = dashed, 1 = dotted

line_colors <- c("black", "black", "black")

# Create a legend

legend("topright", legend = unique_methods, col = line_colors,

       lty = line_types, cex = 0.8)

plot_grid(plot_model(data.nest.lme, type = "diag"))


# Missing values

missing.data <-
  read.csv("C:/Users/siyab/OneDrive/Documents/FIXmodified_dataset
     _with_missing.csv")

summary(missing.data)

nes <-missing.data

nes1 <- nes$Test_Score # accessing the test results

summary(nes1)

p <- data.frame(na.omit(nes1)) #removing the missing observation

p1 <- summary(p) #summary of the dataset without missingness
```

```
data <- nes

# Remove missing values

data_cleaned <- data[complete.cases(data), ]

summary(data_cleaned$Test_Score)

data.cleaned.lme <- lme(Test_Score ~ Teaching_Method,
                        random = ~ 1 | School/Class,
                        data = data_cleaned)

summary(data.cleaned.lme )

interaction.plot(data_cleaned$School, data_cleaned$Teaching_
   Method,
                 data_cleaned$Test_Score,

                 xlab = "School", ylab = "Test Score", legend =
                    FALSE)

unique_methods <- unique(data_cleaned$Teaching_Method)

line_types <- c(3, 2, 1)  # 3 = solid, 2 = dashed, 1 = dotted

line_colors <- c("black", "black", "black")

# Create a custom legend with line types and colors

legend("topright", legend = unique_methods, col = line_colors,

       lty = line_types, cex = 0.8)


# Single imputation method

imputed <- mean(nes$Test_Score,na.rm=TRUE) # calculating the mean

nes[is.na(nes$Test_Score),"Test_Score"] <-
  imputed #replacing missing values with the mean

imputednest <- nes

imputed.data.lme <- lme(Test_Score ~ Teaching_Method, random = ~
   1 | School/Class,
                        data = imputednest)

summary(imputed.data.lme)
```

```
interaction.plot(imputednest$School, imputednest$Teaching_Method,
    imputednest$Test_Score,

                 xlab = "School", ylab = "Test Score", legend =
                    FALSE)

unique_methods <- unique(imputednest$Teaching_Method)

line_types <- c(3, 2, 1)  # 3 = solid, 2 = dashed, 1 = dotted

line_colors <- c("black", "black", "black")

# Create a custom legend with line types and colors
legend("topright", legend = unique_methods, col = line_colors,
       lty = line_types, cex = 0.8)

# Multiple imputation method with mice package

micedata <-
  read.csv("C:/Users/siyab/OneDrive/Documents/complete_dataset_
     with_imputed.csv")


micedata.lme <- lme(Test_Score ~ Teaching_Method, random = ~ 1 |
   School/Class,
                       data = micedata)

summary(micedata.lme)

interaction.plot(micedata$School, micedata$Teaching_Method,
   micedata$Test_Score,

                 xlab = "School", ylab = "Test Score", legend =
                    FALSE)

unique_methods <- unique(micedata$Teaching_Method)


line_types <- c(3, 2, 1)  # 3 = solid, 2 = dashed, 1 = dotted

line_colors <- c("black", "black", "black")

legend("topright", legend = unique_methods, col = line_colors,

       lty = line_types, cex = 1)
```

9C__Users_siyab_OneDrive_Desktop_Research_R_fullcode5.R