EVALUATION OF SINGLE AND MULTIPLE MISSING DATA IMPUTATION TECHNIQUES: A COMPARATIVE APPLICATION ON BMI DATA

By
Lethani Mbongeni Ndwandwe
210507780

A dissertation submitted in fulfilment of the academic requirements of

Master of Science in Statistics

School of Mathematics, Computer Science and Statistics

College of Agriculture, Engineering and Science

Supervisor: Dr Siaka Lougue

Co-supervisor: Ms Annapurna Hazra

November 2017

PREFACE

The research contained in this thesis was completed by the candidate while based in the Discipline of Statistics, School of Mathematics, Computer Sciences and Statistics of the College of Agriculture, Engineering and Science, University of KwaZulu-Natal, Westville Campus, South Africa. The research was financially supported by College of Agriculture, Engineering and Science.

The contents of this work have not been submitted in any form to another university and, except where the work of others is acknowledged in the text, the results reported are due to investigations by the candidate.

....

Signature (Dr Siaka Lougue) Date

DECLARATION: PLAGIARISM

I, Lethani Mbongeni Ndwandwe, declare that:

(i) the research reported in this dissertation, except where otherwise indicated or acknowledged, is my original work;

(ii) this dissertation has not been submitted in full or in part for any degree or examination to any other university;

(iii) this dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;

(iv) this dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced;

(v) where I have used material for which publications followed, I have indicated in detail my role in the work;

(vi) this dissertation is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included;

(vii) this dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the References sections.

…… ……

Signed: Lethani Mbongeni Ndwandwe Date:

ABSTRACT

Missing data are a common occurrence in various fields of data science and statistics. The research into missing data is one of the most important topics in applied statistics, especially in academic, government and industry-run clinical trials. However, this data loss can result in an inadequate basis for study inferences. Dealing with missing data involves neglecting or imputing unobserved values. However, the methods used to deal with the missingness in a data set may bias the results and lead to results which do not reflect a true picture of the reality under investigation in a study.

This thesis discusses the various missing data mechanisms and how missing values can be inferred. The main objective of this thesis is to evaluate the performance of several single and multiple imputation methods for a continuous dataset to find the best imputation techniques. Based on a complete survey data (2014 Lesotho Demographic Household Survey), missingness was created in the response variable (BMI) using three missing data mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Missing values were then imputed using three single imputation methods and two multiple imputation methods, namely: mean substitution, hot-deck and regression, multiple linear regression and predictive mean matching (PMM), respectively. The analysis indicated that the PMM imputation method is more precise and can also produce lower estimated standard error compared to other methods.

DEDICATION

I would like to dedicate this thesis to my late parents, Nelisiwe Dlamini and Mr Ndwandwe.

I also dedicate this work to my late brother, Mzinto Ndwandwe who took care of me in difficult times and to my little sister, Ncamsile Ndwandwe, who used her last cent so that I can further my education.

ACKNOWLEDGMENTS

First and foremost I would like to thank my Lord and Saviour for making all that I do possible and for constantly reminding me that anything could happen.

A special thanks goes to my supervisor, Dr Siaka Lougue, who literally carried me with his support and guidance.

To my co-supervisor, Ms Annapurna Hazra, thank you for all the effort, guidance, time and support you have given me over these years, from my undergraduate studies till today.

I would also like to thank my family, friends and colleagues that God blessed me with for always being there for me.

Finally, I would like to thank the Sol Plaatje University for believing in me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY OF ACRONYMS

BPCA	Bayesian principal component analysis
BMI	Body mass index
CATI	Computer-assisted telephone interviewing
CAPT	Computer-assisted personal interviewing
EM	Expectation-maximisation
FKM	Fuzzy K-means
IOA	Index of agreement
KNN	K-nearest neighbour
MAE	Mean absolute error
MAR	Missing at random
MCAR	Missing completely at random
MICE	Multiple imputations by chained equations
MNAR	Missing not at random
NNHD	Nearest neighbour hot-deck
PMM	Predictive mean matching
R	Correlation coefficient
RHD	Random hot-deck
RMSE	Root mean square error
SCE	Supervised classification error
SHD	Sequential hot-deck
SKNN	Sequential K-nearest neighbour
SRS	Simple random sampling
SVD	Singular value decomposition
UCE	Unsupervised classification error

CHAPTER 1: INTRODUCTION

1.1 Background

Missing data are a common occurrence in various fields of data science and statistics. Research in the field of missing data is one of the most important topics in applied statistics, especially in academic, government and clinical trials. This thesis focuses on the performance of the three single and two multiple imputation techniques for continuous data. Having missing values in a dataset can have serious consequences in the analysis of the data. For example, Katz (2015) argues that data loss can result in an inadequate basis for study inferences. Furthermore, data loss can lead to a small sample size, damaging the precision of confidence intervals, which could bias estimates of parameters, decrease statistical power and produce high standard errors (Dong and Peng, 2013). Dong and Peng (2013) further suggest that the best way to deal with missing data is to avoid missing information in the first instance. This means that the data collector must try by all means possible to ensure that the survey questions are all answered.

No matter the precautions taken, it is nevertheless difficult to collect data with no missing values. This situation gives rise to an important question regarding how missing data must be dealt with. Most researchers suggest that imputing missing data using statistical techniques is an option in dealing with missing data. These imputation techniques can largely improve data quality and various imputation techniques have thus been developed to tackle this problem. However, it is very difficult for a data scientist to identify or choose the best technique for each situation. In this study, missingness was created on a complete survey dataset from the 2014 Lesotho Demographic Household Survey in the response variable body mass index (BMI) based on three missing data mechanisms namely: missing completely at random (MCAR); missing at random (MAR); and missing not at random (MNAR). Five imputation techniques, namely mean substitution, hot-deck, regression, multiple linear regression and predictive mean matching (PMM), were evaluated to determine the most effective technique.

During the analysis phase, researchers usually apply a technique that omits all the cases containing missing information. However, this technique is not recommended since it can lead to results which are not truly representative of the data collected (Nakagawa and Freckleton, 2008). Many studies reveal that the imputation of missing information depends on the three missing data

mechanisms: MCAR, where the probability that a value is missing does not depend on missing neither observed; MAR, where the probability that a value is missing does depend only on observed; and MNAR, where the probability of an instance having a missing value for a variable could depend on the value of that variable.

1.2 Aim of the study

This study was initiated to investigate the topic of various missing data mechanisms and how missing values can be inferred. The main aim and objective of this thesis is to evaluate the performance of single and multiple imputation methods for a continuous dataset on a single variable which contains missing values, in this case, BMI. There are several imputation techniques that can be used to improve the quality of dataset. In this study, we consider five of the most popular imputation techniques to find one that can better reduce the nonresponse biased, namely: mean substitution, hot-deck, regression, multiple imputation linear regression and PMM methods. These imputation techniques were selected since they are the most widespread and commonly used in handling the missing data that is continuous.

1.3 Missing values on BMI

Body mass index (BMI) is a common metric used to assess an individual's body fat, it is calculated as weight (kg)/height (m^2). BMI cannot be calculated when information on either height or weight is missing. Absence of complete data for the whole of an individual's study period is a potential bias risk and standard complete-case approaches may lead to biased estimates. Considering the dataset collected in recently surveys in Lesotho, BMI found to be the most variable that possesses a lot of missingness. This study is initiated to find the best imputation technique to improve the quality of data (by reducing nonresponse).

1.4 Structure of the thesis

This thesis is organised as follows:

Chapter 1: This chapter presents the introduction to the study.

Chapter 2: This chapter provides a review of the relevant literature, and focuses on the geographic aspects of Lesotho as well as its population, the prevalence of obesity and underweight, the factors that affect BMI, and the concerns of several researchers regarding methods to be used in dealing with missing data. The chapter also presents an overview of the weighting methods for nonresponse.

Chapter 3: This chapter consists of an overview of sources and patterns of missing data.

Chapter 4: In this chapter, the data set analysed in this study is outlined, together with the design of the experiments, the experimental procedure and the criteria for evaluation.

Chapter 5: This chapter consists of the results of the current study.

Chapter 6: This concluding chapter provides a discussion of the major findings of this thesis and presents the conclusions that can be drawn from the study.

CHAPTER 2: LITERATURE REVIEW

It was reported that above 180,000 vulnerable children in Lesotho are in need of essential services such as health care, access to education, and psychosocial support. Nutritional deficiencies and conditions are a challenge because of widespread poverty, food insecurity, and inadequate access to services. Because of these reasons, people feel sensitive and end up not answering questions regarding their BMI in the survey. Therefore, BMI variable results in having a lot of missing information that can be used for analysis. The next section gives the biography of Lesotho.

2.1 Lesotho – geography and population

Lesotho is a very small country which is situated inside of the borders of South Africa. The population of the country in 2013 was estimated at 1.9 million where Basotho people were dominating every area of the country with the high percentage of 99.7% (Van den Berg et al. 2014). The total area of the country spans approximately 30, 355 square kilometres, of which approximately 10% is arable land (Lesotho Demographic and Health Survey 2004). The country is well known by its nickname 'kingdom in the sky' due to the mountainous terrain that predominates. Lesotho is a beautiful country with a clear environment and no forest: only 1% of the land area is forested (Lesotho Demographic and Health Survey 2004;). The map of Lesotho is provided in figure 2.1 (Bureau of Statistics, Population Statistics Unit Ministry of Development Planning).

Figure 2.1: Map of Lesotho (Bureau of Statistics, Population Statistics Unit Ministry of Development Planning)

Figure 2.2 below illustrates the trend in population growth in Lesotho from 2006 to 2016, and shows that the population has been increasing year after year.

Figure 2.2: Population of Lesotho from 2006 to 2016 (www.tradingeconomics.com, 2017)

For the last decade (2006 to 2016) the population of the country has been increasing dramatically. In fact, the statistics indicates that there are now more births compared to the number of deaths in the country.

2.2 Obesity and underweight in Lesotho

One of the measure of obesity is body mass index (BMI) which is calculated as weight in kilograms divided by height in meters squared. Usually a person with a BMI of 30 and above is considered as obese. According to the 2014 Lesotho Demographic and Health Survey, it was found that 45% of women in Lesotho are obese, 4% are thin, and 51% of women have a normal BMI. The world health organisation (WHO) have noticed that most of the Sub-Saharan African countries are facing the problem of obesity. Recently studies have revealed that obesity from young age can have a significant factor on your adulthood (Van den Berg et al. 2014). Van den Berg (2014) further discussed obesity in Lesotho specifically amongst the youth, these researchers argued that there is no research that has been initiated to focus on obesity amongst the youth of Lesotho. Figure 2.3 shows the trend in women's obesity and underweight in the period of ten years (2004 to 2014).

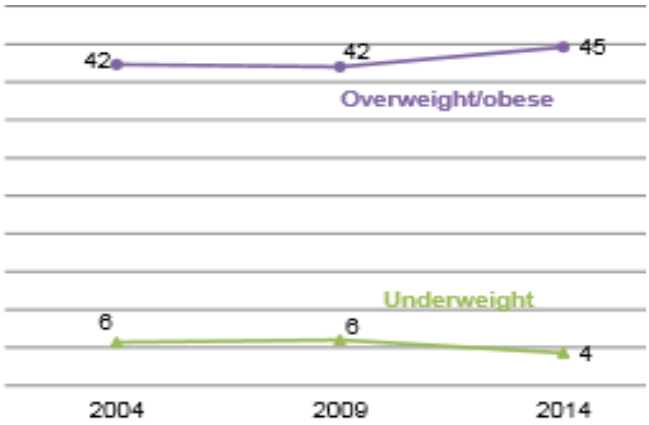

Figure 2.3: Trends in women's obesity (Ministry of Health [Lesotho] and ICF International)

In the figure it is indicated that the percentage of obese women from 2004 to 2009 was 42% and subsequently during the period 2009 to 2014 this percentage increased to 45%. The percentage of thin people from 2004 to 2009 was 6% which then decreased during 2009 to 2014 to 4%.

An examination of the malnutrition status of Lesotho indicates that food shortages seem to be the main cause of underweight. The country does not have enough food to provide for all communities, and is currently in a position where it requires efficient support from food security agencies so that it can combat this shortage of food. The shortage of food in the country has a significant effect on malnutrition of people of Lesotho. According to the 2014 Lesotho Demographic and Health Survey, 33% of children under the age of 5 suffer from stunted development, which can be attributed to malnutrition. Figure 2.4 below shows the depth of the food deficit from the year 2004 to 2014.

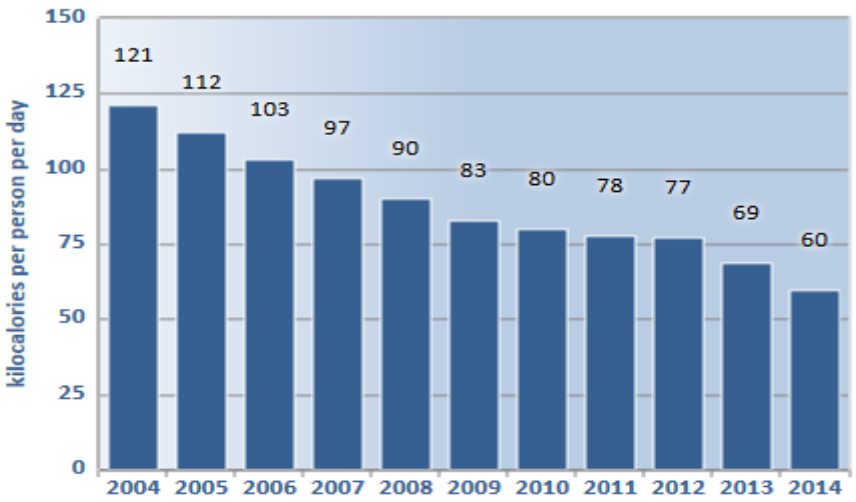

Figure 2.4: Depth of the food deficit (http://lesotho.opendataforafrica.org/ursdlbc/food-deficit)

It is evident that from 2004 the kilocalories consumed per person per day have been decreasing dramatically. Attaining food security means guaranteeing the quality and continuity of food access, in addition to quantity, for all household members.

2.3 Factors affecting BMI level

Gender, age, education, physical activity, residence type, marital status and environmental factors are some of the factors that can have a significant effect on BMI. Most studies that have been conducted have indicated that median men have higher BMI values than median women while other studies have seen it in a different way (Sattar et al. 2013). Another noteworthy factor is that age and gender are both associated with weight gain: the older a person is, the more likely they are to be overweight, and women are usually more overweight than men. This further indicates that both age and gender can play a major role on BMI (Sattar et al. 2013). Sattar (2013) further argued that BMI increases with age and usually remains steady or decreases at the age of 60 and above.

Furthermore, in their study marital status was one of the major factor associated with BMI (comparing married and unmarried people), where 22.9% of married people had a BMI which is greater than 30 compared to 6.6% of unmarried people. This therefore indicates that marital status can be a significant factor for BMI. Sattar et al. (2013) also discussed the influence of income and residence type as other factors that can influence BMI, finding that people who earn more are at risk of higher BMI (BMI which is greater than 30) as compared to those with low income. In addition, their results showed that 12.9% of those who live in rural areas have a BMI less than

18.5% compared to 16.6% of those who live in urban areas with BMI greater than 30. In the United States of America, obesity is most likely to occur among adults from rural areas than those from urban areas (Sattar et al. 2013).

In another study that was conducted to understand the drivers of overweight and obesity in developing countries, specifically in South Africa, Puoane et al. (2002) argued that South African people are not concerned about their body weight, especially those aged 15 and above. The greater the actual body weight of participants in the study, the more the self-perception deviated from the true value. They further highlighted that these incorrect perceptions can also be differentiated between population groups where: women can have a higher BMI compared to men; those who live in urban areas have higher BMI compared to those in rural areas; adults people high BMI compared to children; women with a better education have less BMI compared to those with lower grades, on average (Butzlaff and Minos, 2016; Puoane et al. 2002).

2.4 Comparison of missing data imputation techniques

Since there are so many imputation techniques that can be used to tackle the missingness, Rubin (1976) argued that there is no better technique for all types of datasets. This section of the theses provides the findings about what other researchers found in their comparisons using different imputation methods in different kind of datasets. However, most of the imputation techniques they considered were not used in our study but few methods such as mean substitution, KNN which is the extension of the hot-deck are also used in our study. Tavakoli et al. (2011) also evaluated and compared the imputation techniques based on single and multiple imputation methods.

Zainuri et al. (2015) evaluated and compared some of the imputation methods to deal with missing values occurring in an air quality dataset. The imputation methods evaluated include: the expectation-maximisation (EM) method, singular value decomposition (SVD), K-nearest neighbour (KNN) method, mean substitution and median substitution. The purpose of the study was to determine the best imputation methods of the five procedures used for air quality data in Malaysia, and to subsequently establish whether there was any significant difference between the techniques that were applied at eight stations in Peninsular Malaysia. R software (free software) was used to obtain the analysis for the study. Three imputation indicators: correlation coefficient (R), index of agreement (IOA) and mean absolute error (MAE) were used to compare the

performance of each imputation method. All the methods were tested using different types of datasets from eight monitoring stations located at the central, southern and northern regions of Peninsular Malaysia with missing data comprising 5%, 10%, 15%, 20%, 25% and 30% of the various datasets used. After the analysis was obtained, the results indicated that three of the methods, EM, KNN and sequential K-nearest neighbour (SKNN) were the most powerful imputation methods, irrespective of the station and percentage of missing data. All three performance indicators arrived at the same conclusion that these three imputation methods are the best methods for imputing the missing data (Zainuri et al. 2015).

Tavakoli et al. (2011) evaluated and compared the imputation techniques based on single and multiple imputation to study the role of stress in relation to social support and mood, and also investigated whether mediator effects biased the bond. The methods that were involved in the study were: no imputation, single imputation and MI. The results of the study by Tavakoli et al. (2011) indicated that there were no significant differences between the data with no imputation and the data after imputation was applied for missing values. However, the study did show that there were significant results found in terms of different effect sizes after the imputation was applied (Tavakoli et al. 2011)

Schmitt et al. (2015) evaluated and compared six imputation techniques, namely: mean substitution, KNN, fuzzy K-means (FKM), SVD, Bayesian principal component analysis (BPCA), and multiple imputations by chained equations (MICE). Mean substitution consists of substituting the missing data for a given variable by the average of all observed values of that variable; KNN is an extension of hot-deck imputation that defines for each sample or individual a set of K-nearest neighbours and then substitutes the missing data for a given variable by averaging observed values of its neighbours; FKM is an extension of KNN based on fuzzy K-means clustering. SVD and bPCA are based on eigenvalues. Last but not least, MICE are an iterative algorithm based on chained equations that practises an imputation model specified separately for each variable and involving the other variables as predictors (Schmitt et al. 2015). For their study, four real datasets in different sizes were used for the analysis, small data was considered to be data with less than ten variables, large data was the data with more than ten variables and another data form was MCAR and based on four evaluation criteria: root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. The results

suggested that the four imputation methods, namely mean, KNN, SVD and MICE, are not that well organized. MICE are based on a much more complex algorithm and performance of these imputations appears to be related to the size of the dataset. MICE can perform quickly and effectively when applied in a small dataset, but it can be time consuming when it is applied in large datasets. Finally, the conclusion was that BPCA and FKM perform more effectively compared to other methods that were used in the study (Schmitt et al. 2015).

2.5 Weighting methods for nonresponse

2.5.1 Introduction

The basic idea behind weighting methods is to make the observed values as similar as possible to the original sample in terms of the distribution of some variables. Weighting adjustments are commonly applied in surveys to compensate for nonresponse and to make weighted sample estimates adapt to external values (Lohr, 2009). Lohr (2009) argued that the best way to reduce nonresponse bias in sample surveys is to apply a method of nonresponse weighting adjustment. This procedure can only be applied by the multiplication of the sampling weight of the respondent and the inverse of the estimated response probability, or weights can be defined as reciprocals of the inclusion probabilities so that an estimator of the population total becomes

$$\sum_{i=1}^{n} w_i y_i,$$

where $w_i = (\pi_i)^{-1} = \frac{N}{n}$ is the weight for an individual *i*, y_i is the response of an individual *i* to the variable of interest *y* (in this case BMI).

Here $\pi_i = \left(\frac{n}{N}\right)$ is the probability of being selected in the sample, *N* and *n* are the total population and sample size respectively.

For stratification purposes, weights can be given by

$$w_i = \frac{N_h}{n_h},$$

where N_h and n_h are the total population, and sample size in the stratum *h*.

Weights can also be used to adjust for nonresponse bias, R_i is defined as the indicator variable that an individual *i* is selected in the sample with $P(R_i = 1) = \pi_i$ (Lohr, 2009). Let Z_i be a random variable that the i^{th} selected individual responds. If R_i is independent of Z_i, then

$$P(individual\ i\ selected\ in\ sample\ and\ responds) = P(R_i = 1, Z_i = 1) = \pi_i \emptyset_i,$$

where $\emptyset_i$ denotes the probability that individual *i* responds.

The above expression can also be written in the mathematical expression:

$$P(R_i = 1, Z_i = 1) = P(Z_i = 1 | R_i = 1) P(R_i = 1) = \pi_i \emptyset_i.$$

Since weighting methods for nonresponse does not depend on Y, then it is assumed to be MAR. Estimating $\emptyset_i$ for each individual in the sample gives the final weight as:

$$w_f = \frac{1}{\pi_i \hat{\emptyset}_i}.$$

2.5.2 Weighting class adjustments

Lohr (2009) also noticed that some researchers use a common method to adjust weights for nonresponse bias and that this procedure creates homogeneous weighting groups or cells of sample members, for both respondents and nonrespondents. The known variables in the sample are taken up to create the weighting adjustment groups and it is assumed that all the respondents and nonrespondents in the same group are also similar. The basic procedure of forming the groups is to cross-tabulate the set of the present variables. Within each group, the respondents' weights are increased to take on the weights of the nonrespondents. Consider the example below where variable age is grouped, suppose that the age of each member is known from a selected sample and it has a sampling weight:

$$w_i = {}^{1}/_{\pi_i}$$

and the estimated response probability for each group is measured by:

$$\hat{\emptyset}_{group} = \frac{sum\ of\ weights\ for\ respondents\ in\ the\ group}{sum\ of\ weights\ for\ selected\ sample\ in\ the\ group}$$

Table 2.1: Breakdown of sample by age and weight

Age					
	15-24	25-34	35-44	45-64	65+
Sample size	62	78	51	46	63
Respondents	33	52	42	34	63
Sum of weights for sample	2340	2780	1902	1530	2430
Sum of weights for respondents	988	1033	871	780	2430
$\widehat{\emptyset}_{group}$	0.422	0.372	0.458	0.509	1.000
Weight factor	2.369	2.688	2.183	1.964	1.000

Table 2.1 Illustrate the weighting class adjustment factors, where the probability of response is assumed to be the same within each weighting class, with the implication that within a weighting class, the probability of response does not depend on dependent variable *y*.

To determine the estimation of total population through weight class adjustment:

$$x_{gi}=\begin{cases} 1 \text{ if individual } i \text{ is in } group\ g \\ 0 \qquad\qquad \text{otherwise} \end{cases}$$

Then the new weights (called predicted weights) for respondent individual *i* is given by:

$$\widetilde{w}_i = \frac{1}{\pi_i}\sum_{g=1}^{G}\frac{x_{gi}}{\widehat{\emptyset}_g},$$

where w_i represents the sampling weight for individual *i*, $\widehat{\emptyset}_g$ is the response probability for each group and $\widetilde{w}_i = \frac{1}{\pi_i \widehat{\emptyset}_g}$ if individual *i* is in a group g and $\widetilde{w}_i = 0$ if individual *i* is a nonrespondent.

From the above the estimated total population is:

$$\hat{t}_{wg} = \sum_{i=1}^{n}\widetilde{w}_i y_i.$$

Also, the estimated population mean is given by:

$$\hat{\bar{y}}_{wg} = \frac{\hat{t}_{wg}}{\sum_{i=1}^{n} \tilde{w}_i},$$

which is the division of the estimated population total by the summation of the predicted weights.

In simple random sampling (SRS), if n_g denotes the number of sample individuals in group g, n_{gRes} denotes the number of respondents in group g, and $\bar{y}_{gRes}$ is the average for the respondents in group g, then:

$$\hat{\emptyset}_g = \frac{n_{gRes}}{n_g}$$

and

$$\hat{t}_{wg} = \sum_{i=1}^{n} \sum_{g=1}^{G} w_i \, \hat{\emptyset}_g x_{gi} y_i$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{G} \frac{N}{n} \frac{n_g}{n_{gRes}} x_{gi} y_i$$

$$= N \sum_{g=1}^{G} \frac{n_g}{n} \bar{y}_{gRes}.$$

2.6 Poststratification

Poststratification includes adjusting the sampling weights so that they sum to the population sizes within each post-stratum. The method assists in decreasing the biasness of nonresponse and underrepresented groups in the population. The main advantage of this method is that it can lessen variance estimates. Poststratification was designed to balance a sample's covariate distribution in the situation of complete response, but it is often used in practice to diminish nonresponse or coverage layers between the sample frame and target population (Holt and Smith, 1979).

The most important aspect of poststratification is that it is similar to weighting class adjustment, however the difference is that population counts are used to adjust weights. After the sample has been taken from the population using SRS, all individuals are grouped into *H* post-strata. Suppose

there is population N_h individual in poststratum h, then sample n_h needs to be selected using SRS and n_{hRes}. Now the poststratification estimator for $\bar{y}_U$ is found to be:

$$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{hRes}$$

and the weighting class adjustment for $\bar{y}_U$ is given by:

$$\bar{y}_{wg} = \sum_{h=1}^{n} \frac{n_h}{n} \bar{y}_{hRes},$$

where $\bar{y}_{hRes}$ is the average of the respondents in the poststrata, n_h is the selected sample in poststrata. Poststratification and weighting class adjustment have similarities, however the difference is that N_h for poststratification is known but for weighting class adjustment it is not known (Lohr, 2009).

CHAPTER 3: SOURCES AND PATTERN OF MISSING DATA

3.1 Unit and item nonresponse

Nonresponse actually refers to failure to acquire a measurement on one or more study variables for one or more observation(s) in the survey. There are two main sources of missing data and these can be distinguished by unit and item nonresponse. Unit nonresponse refers to a situation where the entire unit is missing in the survey (a person was interviewed but chose not to answer any of the questions asked). Item nonresponse means that person was interviewed and gave some answer where he/she could but failed to answer some of the questions in the survey. Nonresponse has been a persistent problem in surveys, data analysts and researchers usually have to deal with nonresponse before they even proceed for analysis. However, it is important for researchers to understand the potential impact of nonresponse on the ability of surveys to describe large populations. More importantly is to ensure that nonresponse in the survey is reduced and that an adequate response rate is encouraged. A recent study by Groves (2006) revealed that a response rate of at least 50 % is considered adequate for analysis and reporting. A response rate of 60 % can be considered as good, a response rate of 70 % is very good. Therefore, for any researcher it is very important to pay attention to response rates (Groves, 2006). It is important to be able to distinguish between the two types of nonresponses (unit and item nonresponse). In most survey research, researchers have noticed that unit nonresponse is the main issue when compared to item nonresponse, however, they have developed statistical techniques that will help to address the problem of unit nonresponse (Yan and Curtin, 2010).

3.2 Factors that influence response rate and data accuracy

There are various reasons as to why data might be missing. In this section of the thesis these possible reasons will be discussed. Some of these reasons could be: a respondent might refuse to respond to a particular question due to the sensitivity of the question if it is related to a topic such as income or drugs. Data might be lost due to the carelessness of a data collector failing to enter data correctly on the system; the data may be missing because equipment collapsed during data collection, for example, a dropped call if the interview was telephonic. Weather conditions also play a crucial role in missingness of data, for instance, it may be difficult to collect data in certain areas in summer because people may be busy with activities such as garden work or ploughing.

The most difficult situation is when a respondent refuses to participate in the survey but does not provide a reason (Lohr, 2009).

The next section discusses some factors that play an important role in reducing nonresponse before and during survey.

Some of factors can be grouped as follows:

• **Survey content**- Surveys on personal information such as drug use and financial problems may have a higher number of people refusing to participate, because they may be unwilling to reveal their personal information (Lohr, 2009).

• **Time of survey**- The seasons of the year can also play a crucial role in causing nonresponse. For an example, in summer, especially in December you find that people have decided to go for vacation so it might not be a good timing for data collection (Lohr, 2009).

• **Interviewers**- The behaviour of the interviewer to an interviewee can play a magnificent role in gaining cooperation from a sample unit. The way he/she interacts with the respondent can help in reducing nonresponse in the survey. It is possible for an interviewer to fail to simplify the questions to a respondent, which can have a significant influence in receiving a poor response rate (Lohr, 2009).

• **Data-collection method**- Lohr argued that some of the instruments that are used in surveys may also be reasons that leads to have poor responses. He continued mentioned the use of telephone and mail systems, of which these two are often have low response rates. But two instruments were then introduced to help the improvement of data accuracy in the data that was collected using telephone and in-person survey. These two instruments are Computer-Assisted Telephone Interviewing (CATI) and Computer-Assisted Personal Interviewing (CAPI) (Lohr, 2009).

• **Questionnaire design**- Creating good measures involves both writing good questions and organising them to form the questionnaire. Wording may have a serious impact on receiving accurate responses. Questionnaire design is not an easy process because it requires serious attention to many details at once, so accuracy is vital. A creatively designed form can have a great impact in increasing data accuracy and also reducing item nonresponse (Dillman, 2008; Lohr, 2009).

• **Respondent burden**- Some people may be unwilling to respond to surveys and it is therefore better to begin with closed questions and to end with opened questions when designing a questionnaire. It is advisable to have short questions that are clear and easy to understand. Shorter questions require fewer details, which plays an important role in reducing respondent burden (Lohr, 2009).

• **Survey introduction**- Some people might not be familiar with the topic of the survey, and it is therefore advisable that the main aim and objective of the study is easy to understand. The researcher's name, the name of any organisations represented, the nature of the information that the researcher is attempting to find out must be explained, and respondents should be assured that all responses will be completely anonymous. Respondents should also be made aware of how survey results may benefit the community (Lohr, 2009).

• **Follow-up**- Follow up calls mostly occur when the data collection was achieved through mail and telephone surveys. People think in different ways, and there may be people who refuse to answer no matter how often they may be contacted. It is nevertheless important to make follow up calls to collect further data from a survey (Lohr, 2009).

After the availability of the dataset and all the above mentioned factors were followed but only to find that missingness still does exist, then it where the imputation techniques comes. It should be noted that no matter how the precaution is taken, it always going to be difficult to collect the dataset with no missing values on it. That is why this study focuses on the imputation of missingness.

3.3 Patterns of missing data

The aim of data patterns is to provide the amount of dataset and also the structured of the data. According to Little and Rubin (2002), there are three patterns of missing data: univariate, monotone and arbitrary. Little and Rubin (2002) discuss the theory behind each pattern of missing data and the focus of each type of dataset. A challenge with these missing data patterns is that it may not be clear what the causes of the missingness are.

Univariate missingness pattern: The dataset is said to have univariate missingness pattern if the data possess *k* variables such as X_1, X_2 X_3 …,X_k and every variable is fully observed except for one variable which has some missing values. Consider the example of a univariate pattern shown in Table 3.1 below:

Table 3.1: Illustration of univariate missingness pattern

X_1	X_2	X_3	X_4	X_5
obs	obs	obs	obs	obs
obs	obs	obs	obs	obs
obs	obs	obs	obs	NA
obs	obs	obs	obs	NA
obs	obs	obs	obs	NA

Monotone missingness pattern: A dataset is said to have monotone missingness pattern if variables with missing values are imputed sequentially with covariates obtained from their corresponding sets of preceding variables (Kombo et al. 2017). If an individual is missing variable X_j then it is assumed that the very same individual is also missing all subsequent variables $X_k, k > j$ (Enders, 2010). Consider the example below in Table 3.2 of a monotone pattern:

Table 3.2: Illustration of monotone pattern

X_1	X_2	X_3	X_4	X_5
obs	obs	obs	obs	obs
obs	obs	obs	obs	NA
obs	obs	obs	NA	NA
obs	obs	NA	NA	NA
obs	NA	NA	NA	NA

Arbitrary: The data set is said to have an arbitrary missing pattern if the missing data occurs in any variable for any participant simply at random. Researchers find it very difficult to work with arbitrary missing pattern or to analyse the data of this structure (Enders, 2010; Lohr, 2009). Consider the example of an arbitrary pattern shown in Table 3.3 below:

Table 3.3: Illustration of arbitrary pattern

$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$
obs	obs	NA	obs	NA
obs	obs	NA	obs	NA
obs	NA	obs	NA	NA
obs	NA	obs	NA	NA
obs	NA	obs	NA	NA

3.4 Missing data mechanism

Rubin (1976) introduced three types of missing data: missing complete at random (MCAR), missing at random (MAR) and missing not at random (MNAR), this can be classified as missing data mechanism (Huisman, 2009). These three missing data mechanisms describe relationships between measured variables and the probability of missing data.

Missing data mechanisms have precise probabilistic and mathematical meanings but what is significant is that every mechanism has a different explanation as to why data are missing. Let V denote the complete data matrix with element V_{ik} in the i^{th} row and k^{th} column, where $i = 1, \dots, n$ and $k = 1, \dots, M$. In the presence of missing data, V_{obs} denotes observed values of the matrix V and V_{mis} denotes missing values. Let R denote a matrix with elements (Lohr, 2009):

$$r_{ik} = \begin{cases} 1 \text{ if } V_{ik} \text{ is observed} \\ 0 \text{ if } V_{ik} \text{ is missing} \end{cases}.$$

Now, each of the three missing data mechanisms, namely MCAR, MAR and MNAR, are discussed in depth below.

3.4.1 Missing completely at random (MCAR)

Missingness in a dataset is said to be MCAR if the probability that a value is missing depends neither on missing nor observed values. In the case of an MCAR situation, it is assumed that each individual shares the same probability of missing value (Houchens, 2015). As an example, a dataset can be lost in the laboratory when a tube containing blood samples breaks, which will result

in the researcher not being able to obtain measurements for blood parameters. Other examples of MCAR occur when a participant misses the administration of a survey due to scheduling difficulties. Other unrelated reasons may also contribute to a participant missing the administration of a survey, such as a doctor's appointment, or questionnaires being lost in the participants' emails (Donders et al. 2006).

The MCAR mathematical expression can be given by:

$$P(R|V_{obs}, V_{mis}) = P(R).$$

The advantage of MCAR data is that the complete case analysis remains unbiased. The researcher may lose power for the design, but the estimated parameters are not biased by the absence of data. To deal with data that is MCAR is not problematic since listwise/pairwise deletions are also options in handling missing values. This mechanism suggests that the distribution of the outcome variable has no difference between two groups ($R_i = 0$ and $R_i = 1$). This mechanism is treated as the strongest assumption compared to others (Little and Rubin, 2002).

3.4.2 Missing at random (MAR)

A missing dataset is said to be missing at random (MAR) if the probability that a value is missing depends only on observed values. For an example, old people who have reached the age 50 and above might feel inferior to respond to questions about their sexual activity. However, if the data contains age of all respondents, these data may still be regarded as MAR (if it is reasonable to assume that the response of the sexual activity question itself does not affect to probability of missingness.(Eekhout et al. 2012; Vatanen, 2012). The MAR mathematical expression can be given by:

$$P(R|V_{obs}, V_{mis}) = P(R|V_{obs}).$$

Recent studies reveal that multiple imputation techniques and maximum likelihood are the best ways of dealing with data that are MAR. To determine whether MAR is a good assumption for data or not, the first option is to check the correlation between independent variable and dependent

variables. Once the correlation between independent and dependent variables becomes stronger, then it implies a weak MAR assumption.

3.4.3 Missing not at random

A dataset is said to be MNAR if the missing data are neither MCAR nor MAR (Van den Berg et al. 2014). In other words, a dataset is said to be MNAR if the probability of missing a value is dependent on the missing value itself (Little and Rubin, 2002). It becomes a serious issue when the probability of the missingness depends on the response variable itself, which is what MNAR is related to. As an example, consider a self-report drug assessment administered to mine workers. MNAR data would result if heavy drug users are more likely to skip questions out of fear of being reprimanded. This mechanism can play an important role in the level of bias in statistical analyses. The MNAR mathematical expression can be given by:

$$P(R|V_{obs}, V_{mis}) = P(R|V_{mis}).$$

MNAR data is the most difficult to estimate and model compared to the other two missing data mechanisms, MCAR and MAR (Little and Rubin, 2002; Scheffer, 2002).

3.4.4 Ignorable and nonignorable nonresponse mechanisms

A missing data mechanism can also be categorised as ignorable or nonignorable. There are certain conditions that need to be taken into consideration for a missing data mechanism to be ignorable. The first condition is that the dataset must be missing complete at random (MCAR) or missing at random (MAR). The next condition, is to let Θ and ϑ be the two independent parameters, where Θ represents parameter of interest and ϑ represents parameter of incomplete data set processes. In fact, these two parameters are independent if the following condition takes place:

$$P(\Theta, \vartheta) = P(\Theta)P(\vartheta).$$

The above expression means that the joint distribution of the two parameters is equivalent to the product of the two independent parameters. If one of the two conditions does not apply, then the missing data mechanism becomes nonignorable (Boyko, 2013).

CHAPTER 4: DATA AND METHODS

4.1 Data

The data used in this study were obtained from the Lesotho Demographic Household Survey 2014. The data were used to calculate several measures of nutritional status, specifically maternal height and weight questions and subsequently BMI information, which is calculated as weight in kilograms divided by height in meters squared. For this study, the focus is on the BMI of women aged 15-49. From the dataset, variable BMI (variable of interest) was categorised into three groups: those who have a BMI of less than 25 were grouped as underweight; those with a BMI of between 25 and 35 were grouped as overweight; and the ones who have a BMI of greater than 35 were grouped as obese.

4.2 Methodology

The methodology used in this study involved three steps. The first step consisted of the creation of missing values from a complete dataset using three missing data mechanisms, namely MCAR, MAR and MNAR. The second step involved the imputation of missing data using five different techniques: mean substitution, hot-deck, regression, multiple linear regression and PMM methods. The last step consisted of evaluating and comparing the performance of the five imputation methods.

4.2.1 Simulation of the missing data mechanisms

Using a real dataset, cells are systematically deleted following MCAR, MAR and MNAR mechanisms following nine (9) scenarios of missingness with each scenario simulated three (3) times. The missing values under MCAR, MAR and MNAR were generated across three degrees of missingness (5%, 10%, 25%). The BMI variable was chosen to be variable of interest to experience the missingness (Hendry et al. 2017).

In the MCAR mechanism scenarios, values were remove at random across all the categories in the BMI variable. Provided the probability of missing $P(V_{mis}) = P$, for $P = 0.05, 0.1$ and 0.25.

To simulate the MAR mechanism, the missingness was created according to its association with 'educational attained, marital status and residence type' respectively. Data was randomly deleted from BMI variable such that P(R|no education) = 0.1, P(Primary education) = 0.1, P(R|secondary education) = 0.1, P(R|tertiary education) = 0.1, P(R|married) = 0.15, P(R|not-marrried) = 0.15, P(R|rural) = 0.15 and P(R|urban) = 0.15. These deletions were carried out for all three amount of missingness (Hendry et al. 2017).

MNAR was obtained by creating missing values where the probability of missingness is a function of the dependent variable (BMI). Deletion from the BMI variable was carried out such that P(R|bmi underweight) = 0.2, P(R|bmi overweight) = 0.2 and P(R|bmi obese) = 0.6. These deletions were repeated for the repetition for the three amount of missingness like in MAR (Hendry et al. 2017).

4.2.2 Summary and comparisons of the missing mechanisms and imputations

Parameter estimates resulting from the imputation techniques from all three missing data mechanisms are compared to the complete dataset results to assess bias in estimation. For a descriptive statistics perspective, allowing sampling variability, results (mean, median and standard deviation) for each of these nine scenarios were generated three times and results were averaged and the average was compared to the one of the original dataset. To find the differences between the means, Z – test hypothesis was conducted to draw statistical inferences.

For the multivariate analysis, mean square error (MSE), R-squared and adjusted R-squared were used as the criteria comparison to see how close when compared to the results of original data. Smaller values for MSE indicate closer agreement between predicted and observed results, and an MSE of 0.0 indicates perfect agreement.

R-package was used for the creation of missingness and analysis of the three single imputation methods which includes the mean substitution, hot-deck, and regression imputation methods. Stata software was used for the analyses of the two multiple imputation (MI) methods which includes the linear regression method and PMM methods.

4.3 Imputation

Imputation methods substitute the missing values by plausible values so that it can produce the complete dataset which will grant the data analyst the opportunity to continue with the analysis using standard analysis methods and software (Huisman, 2014). Little and Rubin (2002) developed some of these imputation methods and have been useful for so many years. However, the question arises as to why the missing values must be imputed. The reason for imputing the missing values is not to eradicate the missing but to reduce nonresponse bias. In this study the main focus is on the two general popular approaches: single and multiple imputation techniques. Single and multiple imputations require certain criteria to be satisfied before adoption. Imputation actually plays an important role in ensuring that missingness is reduced from the dataset and also to prepare for analysis. These methods are also responsible for the recreation of a balanced design such that techniques that were applied for analysing a full observed dataset can be applied in many situations. In fact, there are several methods to deal with or to handle missing data, but currently

imputation is the most recommended method (Rubin, 1976). Single imputation methods can work extremely well in a dataset that has few missing values as compared to a dataset with many missing values. Allowing single imputation for a dataset that has many missing values will cause systematic errors since the reflection of uncertainty is not covered. Imputation has several advantages and disadvantages. The advantage of imputation is that it assists the analyst to proceed with the analysis using standard analyses and some of the statistical software. One disadvantage of imputation is that it can lead to the biasness of parameters, such as variances and it results in the researcher treating the imputed values as if they were real values from the survey. Therefore, this study was initiated to find the best imputation technique

4.4 Data handling (listwise and pairwise deletion methods)

Listwise and pairwise are considered as the most common methods of dealing with missing data, however both have their advantages and disadvantages. Before these two techniques (listwise and pairwise) are used for imputation, it is assumed that the data must be MCAR, meaning that the probability of missing in the dependent variable is related neither independent nor dependent (Peugh and Enders, 2004). This proves that such methods can only be unbiased when missing data are MCAR (Liu and Gopalakrishnan, 2017). Peugh and Enders (2004) argued that these two simple data handling methods are among the most disadvantageous methods available for practical

applications. The biggest problem is that researchers or data analysts once they come across with a dataset that has some missing values, they waste no time but to use these methods for analysis (Kim and Curry, 1977).

4.4.1 Listwise deletion method

Listwise deletion method only focuses on the data that are fully observed (completed data set) and it involves deletion of the entire observation which possesses the missing value in any of the variables. Consider the example of a listwise method shown in Table 4.1 below, the shaded part indicates the data that has been deleted:

Table 4.1: Illustration of a listwise deletion method

X_1	X_2	X_3	X_4
obs	obs	NA	obs
obs	obs	obs	obs
obs	obs	obs	obs
NA	obs	obs	NA
obs	obs	obs	obs
NA	obs	obs	NA

Table 4.1 shows that wherever the missing value has occurred then the entire individual is deleted and not only the missing value. This method is still used in many fields, especially in medical fields. This technique has several disadvantages: firstly, it produces biased estimates and parameters if the data is not MCAR; and secondly, it may cause serious damage in loss of a statistical power (Humphries, 2013).

4.4.2 Pairwise deletion method

In the pairwise deletion method, only the missing values are eliminated and not the entire observation, which means that present values in the same observation are used for analysis. Statistical software programs such as SPSS do include cases that possess missing values on the variable(s) under analysis, but specifically remove the missing values and not the entire observation. In table 4.2, again the shaded part shows the data that has been deleted.

Table 4.2: Illustration of a pairwise deletion method

X_1	X_2	X_3	X_4
obs	obs	NA	obs
obs	obs	Obs	obs
obs	obs	Obs	obs
NA	obs	Obs	NA
obs	obs	Obs	obs
NA	obs	Obs	NA

Table 4.2 shows that this technique only removes the missing values (highlighted in blue) from each variable but not the entire individual. The advantage of the pairwise technique is that it increases the statistical power and seems like it would be good compared to listwise method, because it does make use of all available data (Graham, 2012). This method also assumes that the data is MCAR. Even though the pairwise method is recommended over the listwise method, but the method also has its own disadvantages, like producing standard errors that are underestimated or overestimated (Peugh and Enders, 2004).

4.5 Single Imputation

Single imputation refers to the gathering of common traditional missing data methods where the researcher imputes the missing values with superficially suitable replacement values. Researchers have developed many different kinds of single imputation techniques, but in this section five of mostly used single imputation techniques are discussed. The following are five different types of single imputation methods.

- Deductive imputation method
- Mean substitution imputation method
- Hot-deck imputation method
    - ✓ Sequential hot-deck imputation method
    - ✓ Random hot-deck imputation method
    - ✓ Nearest neighbour hot-deck imputation method

- Cold-deck method
- Regression method

4.5.1 Deductive imputation method

The deductive imputation method may only be considered when a value is deduced with certainty and can be completed during the data collection, office data capturing or during data processing. This method can only be applied if there is only one possible solution for the particular question, for example, if a respondent did not answer the gender question but did fill in the Mr/ Ms/Mrs question, then it can be deducted that the respondent is male or female. Another example is, if the question 'do you smoke cigarettes?' was left blank, but the following question 'How many cigarettes?' contained the response '6 cigarettes a day', then the answer to the first question is deducted from the second question of which it is 'YES' (Croft, 2008). Most researchers do not encourage people to apply this technique because it could lead to biases in the data. However, in some situations it is possible, and indeed desirable to deduce the response from other information in the questionnaire.

4.5.2 Mean substitution imputation method

The mean substitution imputation (unconditional) method takes the average of all the observed values in that variable and uses it to fill in the missing data. The mean substitution method is regarded as a simple and straightforward method among imputation techniques. This method is not recommended, and researchers have seen that it is the worst of all possible strategies. One of the disadvantages is that injecting the mean substitution as imputed value reduces variance on the variable and can dramatically cause a very serious problem with covariance and correlations. This technique is usually not acceptable to handle missing data and it also underestimates population variance (Fayers et al. 1998; Song and Shepperd, 2007). Consider the example shown in Table 4.3 below:

Table 4.3: Data with missing values

Participants	Age	Gender	Education	Residence
1	22	male	2	rural
2	31	female	1	urban
3	28	male	1	rural
4	19	male	3	rural
5	NA	male	3	rural
6	34	female	2	urban
7	20	male	2	urban
8	NA	female	2	rural
9	39	female	2	urban
10	32	female	1	urban
11	41	female	2	rural
12	15	male	1	rural
13	40	female	2	rural
14	21	female	2	rural
15	18	male	3	urban
16	NA	female	3	rural
17	NA	female	2	urban
18	35	male	1	urban
19	25	male	1	urban
20	30	female	2	rural

The dataset in table 4.3 consists of twenty individuals and four variables (age, gender, education level attained, residence type) of which three variables have a completed dataset and age has some missing values in it.

Example: individuals 5, 8, 16 and 17 did not provide their age, so for the mean substitution imputation technique, the average of all other participants with completed dataset in variable age is taken and used to impute for individual 5, 8, 16 and 17. The mathematical expression for mean method can be given by:

$$Mean\ Impute = \frac{\sum_{i=1}^{n_{obs}} Age_{obs,i}}{n_{obs}},$$

where Age_{obs} are observed values (of Age) and n_{obs} is total number of participants with complete values of Age.

4.5.3 Hot-deck imputation method

The hot-deck imputation method actually takes the observed value given by other respondents to impute the missing value for another respondent. For example, if sex, race and level of education have been completed but age is missing, a random respondent with the same characteristics like sex, race and level of education is selected from the respondents, and that respondent's age is used for the missing value (Allen and Seaman, 2010). Notice that hot-deck continues from the mean substitution imputation method but this does not mean that these two methods share any features, they are totally different. Consider the example of a hot-deck in tables below:

Table 4.4: Illustration of dataset with some missing values

STUDENT NO:	AGE	GENDER	MODULE	MARK
211503699	25	MALE	MATHS	NA
211513099	21	NA	STATS	71

Table 4.4 above identifies cases with missing values (NA's) from the dataset.

Table 4.5: Illustration of dataset with completed values

STUDENT NO:	AGE	GENDER	MODULE	MARK
211513619	25	MALE	MATHS	52
211523430	25	MALE	MATHS	49
211543088	25	MALE	MATHS	85
211514000	25	MALE	MATHS	66
211511111	25	MALE	MATHS	63
211515019	21	FEMALE	STATS	69
211513022	21	MALE	STATS	69
211513005	21	MALE	STATS	69
211511077	21	FEMALE	STATS	69
211514093	21	MALE	STATS	69
211523049	21	FEMALE	STATS	69

Table 4.5 shows a complete dataset where all student numbers with similar characteristics are grouped together.

Now, one of these records is chosen at random, and the MARK value is 'borrowed' for student #211503699, specifically those with the same characteristics, and again any of the entries on GENDER is randomly selected and substituted with student #21153099.

Hot-deck imputation procedures have several methods to determine how to choose a donor unit. In this case three of them are going to be considered, which are sequential hot-deck (SHD), random hot-deck (RHD) and nearest neighbour hot-deck imputation (NNHD).

4.5.3.1 Random hot-deck (RHD) imputation method

This technique involves simply choosing a donor (observed) of the same variable and using it to impute the missing value. The same donor may be used several times if there are too many participants with missing information in that variable, alternatively other imputation methods may be applied (Allen and Seaman, 2010). For example, if one of the participants did not give information about his or her educational level, this technique allows the researcher to choose randomly from the data any of those who provided their educational level and use it for imputation of the missing value.

4.5.3.2 Sequential hot-deck (SHD) imputation method

The sequential hot-deck method sanctuaries the arranging methodology of the unweighted procedure, but the advantage of this method is that its gives all the respondents an opportunity to become donors. In fact, sampling weights can also be used to assist in the limitation of a single donor to be used several times for imputation (Lohr, 2009). Respondents and non-respondents are first separated into two files and sorted (randomly, or by auxiliary variables). Sample weights of the non-respondents are rescaled to sum to the total of the respondent weights. For example, if an individual did not provide the information about his/her education level, the closest donor from the top till the bottom can be substituted for that participant.

4.5.3.3 Nearest Neighbour hot-deck (NNHD) method

This technique simply selects a donor that is very close to the one of the missing and substitutes it there. The NNHD method is considered as the most well organised method when compared to the random and sequential hot-deck imputation methods. Since the NNHD uses the supplementary knowledge given by the values of the independent variables (*x*-values), this procedure does not pick up a donor simply at random but requires y-respondents and x-values (Chen and Shao, 2000). For example, consider an individual who did not provide the information about his/her education level attained, the NNHD method allows the researcher to take the closest donor (respondents) with the same characteristics and to impute the information to that particular individual.

4.5.4 Cold-deck imputation

Cold-deck imputation uses the information from other external sources to impute the missing values (Lohr, 2009). These values can be constructed with the use of historical data, subject-matter expertise, etc. For example, from the previous data one of the respondents may have provided information about his/her education level attained, but it may have happened that for the current data he/she then failed to provide the information about his/her education level attained. The cold-deck imputation method uses the respondent's previous information to impute for the current one. It has been found that it is very difficult to use the cold-deck imputation method alone, but as an alternative it can be used as a starting point if the researcher wants to use the hot-deck imputation method (Wang, 2003). The tables below show how cold-deck imputation can be applied.

Table 4.6: Data collected in previous years

Income	Age	Gender	Education
R10k	19	female	NA
R13k	21	male	1
R15k	23	male	2

Table 4.7: Data collected recently

Income	Age	Gender	Education
R10k	NA	female	1
R13k	28	NA	NA
R15k	31	male	2

Table 4.8: Cold-deck imputation method

Income	Age	Gender	Education
R10k	19	female	1
R13k	28	male	1
R15k	31	male	2

Table 4.8 shows the completed dataset obtained after cold-deck method was used.

4.5.5 Regression imputation method

Regression imputation is also known as conditional mean substitution imputation, which substitutes the missing values with predicted scores from a regression equation. The basic idea of this procedure is intuitively appealing; it uses knowledge from complete variables to fill the incomplete variables. This technique uses a regression equation to produce all the predictions, then adds a random error to each of the predictions to complete the imputation of the missing values. Let Y be a continuous variable which satisfies the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e_i.$$

The above model is fitted using completed data with no missing values for the variable Y and its covariates $X_1, X_2, \ldots, X_p$ where p represents the remaining variables.

To obtain a random error the normal distribution is applied with the mean of zero and a variance equal to the residual variance from the prior regression analysis. The process is all dependent on the variable that needs to be imputed (dependent variable).

4.6 Multiple imputations

Multiple imputation procedure is assumed to be more superior compared to single imputation procedures. Researchers noted that single imputation procedures failed to produce unbiased standard error estimates, and then introduced the multiple imputation procedures (Schafer and Graham, 2002; Tonini et al.). This technique was first introduced by Rubin (1977) and expanded on in his book '*Multiple imputation for nonresponse in surveys'* in 1987. Multiple imputation means that the imputations are done more than once $(M > 1)$ taking different random errors to add on each imputation. Using this technique, the process of imputation is repeated many times and all the repetitions are combined to provide a single estimate. Multiple imputation was found to be the most popular technique for handling missing data. It should be noted that imputed values should not be treated as the original values since they were created from a completed data set. According to Rubin (1977), multiple imputation consists of three steps:

1. **Imputation:** Nonrespondents are imputed *M* times to generate *M* completed data set.
2. **Analysis:** The completed data set from step 1 are analysed to produce parameter estimates for each imputed value.
3. **Pooling:** All the parameter estimates from different imputed values in step 2 are grouped together to obtain a final estimate (Yuan, 2010).

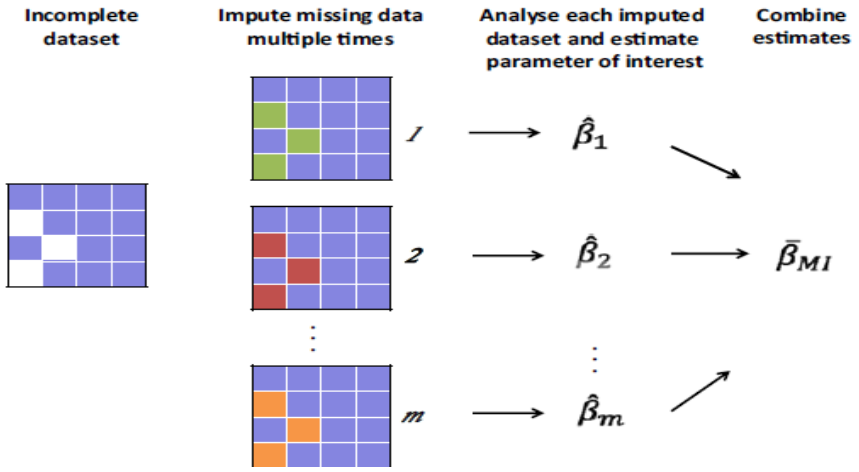

Figure 4.1: Multiple imputation (Lee and Simpson, 2014)

Multiple imputation has been considered to be one of the most applicable tools for general purpose handling of missing data by many different data analysts and researchers (Little and Rubin, 2002). Little and Rubin (2002) provide the key steps of how multiple imputation is generated:

Step 1: Impute missing values using an appropriate model
Step 2: Repeat the first step *M* times (usually 3-5 times), producing *M* complete data sets.
Step 3: Perform the desired analysis on each data set using standard complete data methods.
Step 4: Average the values of the parameter estimates across the *M* samples to produce a single point estimate.
Step 5: Calculate the standard errors by firstly averaging the squared standard errors of the *M* estimates and calculating the variance of the *M* parameter estimates across samples, and last combine the two quantities using a simple formula.

This section of the thesis employs a theory based on three different multiple imputation methods:

- Multiple imputation linear regression for a continuous variable
- Predictive mean matching (PMM) for continuous variable
- Logistic regression for a binary variable

4.6.1 Multiple imputation linear regression method

For the multiple imputation linear regression method , a regression model is fitted for each variable with missing values, with the previous variables as covariates. Based on the resulting model, a new regression model is then fitted and is used to impute the missing values of the dependent variable BMI (Yuan, 2010). Let *Y* be a continuous variable which satisfies the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + e_i.$$

Consider the partition of $X = (X_0', X_m')$ into $n_0 \times 1$ and $n_1 \times 1$ vectors containing the observed and missing values. Now consider the similar partition $Z = (Z_0, Z_m)$ into $n_0 \times q$ and $n_1 \times q$ submatrices.

The following steps show how the linear regression imputation method works:

Step 1: Fit the regression model on the complete dataset (X_0, Z_0) to obtain parameter estimates $\hat{\beta}$, and $\hat{\sigma}^2$.

Step 2: Simulate new parameters $\beta^* \; and \; \sigma^{2*}$ from their joint posterior distribution under ventional nonformative improper prior $P(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$. This can be obtained from using two following steps

$$\sigma^{2*} \sim \hat{\sigma}^2 (n_0 \times q \,) / \chi^2 n_0 \times q$$

$$\beta^* | \sigma^{2*} \sim N\{\hat{\beta}, \hat{\sigma}^2 (Z_0', Z_0)^{-1}\}$$

Step 3: Obtain one set of imputed values, X_m^1 by simulating from $N\left(X_m \beta^*, \sigma^{2*} I_{n_1 \times n_1}\right)$.

Step 4: Repeat step 2 and step 3 above correspond to simulating from the posterior predictive distribution of the missing data $P(X_m | X_0, Z_0)$.
(Stata, 2009; Schenker *et al*., 2010).

4.6.2 Predictive mean matching (PMM)

Predictive mean matching (PMM) is one of the imputation techniques that is used to impute the missing values of a continuous variable. The PMM method is a partly parametric approach that first predicts the values for the missing data *Y* using a linear prediction model. The method uses the predictions (linear regression on observed data) to choose the nearest observed value to impute for the missing value. Randomisation can be introduced by defining a set of values that are closest to the predicted value and choosing one value out of that set, at random, for imputation (Durrant, 2005). Vink (2014) further demonstrated the five steps regarding how PMM multiple imputation can be obtained. Let Y_{resp} and X_{resp} denote the observed values from a dependent variable *Y* and independent variables *X*, respectively .

Step 1: Use linear regression of Y_{resp} given X_{resp} to estimate $\hat{\beta}, \hat{\sigma}^2 \; and \; \hat{e}$ by means of ordinary least squares.

Step 2: Draw σ^{2*} and β^* from their posterior distributions, as for $\sigma^{2*} = \frac{(\hat{e}'\hat{e})}{V}$ where V denotes the chi-square distribution χ^2 with $n_{resp} - r$ degrees of freedom. Where β^* can be obtained from multivariate normal distribution with mean $\hat{\beta}$ and estimated covariance matrix $\sigma^{2*}(X'_{resp}X_{resp})^{-1}$

Step 3: Compute $\hat{Y}_{resp} = X_{resp}\hat{\beta}$ and $\hat{Y}_{nonresp} = X_{nonresp}\beta^*$. where $Y_{nonresp}$ denotes the missing values of variable *Y*

Step 4: For each missing value $\hat{Y}_{nonresp,i}$ where $i = 1, \dots, n_{nonresp}$.

(a) Find $\Delta = |\hat{Y}_{resp,k} - \hat{Y}_{nonresp,i}|$ for all *k*, where $k = 1, \dots, n_{resp}$.

(b) Now for imputation, one of the component $\Delta^1, \dots, \Delta^5$ can be chosen at random and the corresponding Y_{resp} can be taken as the imputation.

Step 5: Step 1 to 4 should be repeated *M* times (Vink et al. 2014).

4.6.3 Logistic regression for a binary variable

The logistic regression imputation method is used to fill in the missing values of a binary variable. This technique is found to be a parametric method that assumes a basic logistic model for those variables that are imputed but when other predictors are known. In the case of the logistic regression imputation method, this technique simply fits the logistic model using the present values (observed values) together with covariates. After the logistic model has been fitted and all the parameter estimates have been obtained, then the posterior predictive distribution of the parameters can be built. Posterior predictive distribution then produces a new logistic model that can be used to impute the missing values (Dolovich et al. 2011). Maximum likelihood can also be used to estimate the logistic regression using the completed dataset. Again, posterior distribution of the parameters can be used in the completed dataset to create a random draw. Once the logistic model has been fitted then all the probabilities are obtained for every missing and a Bernoulli draw is created for that probability, producing imputed values of 0 and 1 (Allison, 2005). The following section outlines the steps and formulas regarding how the logistic regression imputation technique can be obtained.

Firstly a univariate variable $Y = (y_1, y_2, \dots y_n)'$ that follows a logistics regression model must be examined.

$$P(Y_i \neq 0|X_i) = exp(X'_i\beta)/1 + exp(X'_i\beta),$$

where $X = (x_{i1}, x_{i2}, \dots x_{in})'$ records values of estimates of the univariate variable (Y) for an individual i and β is the $q \times 1$ vector of the unknown regression coefficient (Raghunathan et al. 2001).

Taking into account that the variable of interest Y (BMI) contains some missing values that need to be filled in. Now the partition $Y = (Y_0', Y_m')$ into $n_0 \times 1$ and $n_1 \times 1$ vectors must be considered, containing a full data set with completed values and missing. Again, the same partition of $X = (X_0, X_m)$ into $q_0 \times 1$ and $q_1 \times 1$ submatrices must be considered.

The following steps for logistic regression imputation must be considered:
Step 1: The logistic regression model using completed data sets must be fitted and all parameter estimates $\hat{\beta}$, and their asymptotic sampling variance, $\widehat{U}$ must be yielded.
Step 2: β^* must be drawn from the large-sample normal approximation, $N(\hat{\beta}, \widehat{U})$ to its posterior distributions but assuming non-informative prior $P(\beta) \propto costant$.
Step 3: One set of filled in values, Y_m^1, must be obtained by simulating from logistic distribution:

$$P(y_i = 1) = exp(X_{im}'\beta_*)/\{1 + exp(X_{im}'\beta_*)\}.$$

To obtain a greater number of imputations (M), steps 2 and 3 must be repeated. It must be kept in mind that β_* is draw from the asymptotic approximation to its posterior distribution (Rubin, 1987).

4.7 Combining inferences from imputed data sets

The multiple imputation involves three steps: imputing, analysing and pooling. Therefore, the inferences must be combined, as well as the multiple sets of parameter estimates, standard errors and test statistics to obtain final parameter estimate. Then using the combining rule which was introduced by Rubin (1987) the average of the estimates must be calculated across multiple imputations, as well as the variances of the estimates. Suppose inferences about scalar ϕ are to be made, W denotes within imputation variance, B denotes between imputation variance and T is the total variance associated with ϕ. The uncertainty about the results from the single imputed dataset is reflected by W, but B reflects the uncertainty due to the missing data. Let $\hat{\phi}_i$ and $\widehat{W}_i$ be the

completed data estimates where $i = 1, 2, \ldots, M$. To group the final parameter estimates for ϕ, the average is calculated from the completed data set and is given by:

$$\bar{\phi} = \sum_{i=1}^{M} \frac{\hat{\phi}_i}{M}$$

and

$$W = \sum_{i=1}^{M} \frac{\widehat{W}_i}{M},$$

where W is the within-imputation variance component that is achieved as the average of the complete dataset variance estimates

$$B = \sum_{i=1}^{M} \frac{\left(\hat{\phi}_i - \bar{\phi}\right)^2}{M - 1}$$

and B is the between-imputation variance.

$$\mathrm{T} = W + (1 + 1/\mathrm{M})\mathrm{B}.$$

Total variance is the variance estimates which is related to average of ϕ (meaning $\bar{\phi}$).

From the total variance, $(1 + 1/M)B$ estimates the increase in variance due to nonresponse occurred.

If B becomes higher than W then there is a greater efficiency, this indicates that the greater the number of imputations increases the more accurate estimates are achieved (Marshall et al. 2009).

The RMSE is a useful measure tool of overall precision or accuracy and can be used to evaluate the performance of each imputation method. It represents the sample standard deviation of difference between original values and imputed difference (Ferrari and Ozaki, 2014). In general, the technique that would be more effective would be the one with a lower RMSE. The RMSE is obtained by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}.$$

CHAPTER 5: RESULTS

5.1. Descriptive analysis

5.1.1. Descriptive analysis of the original data

The first step in analysing the data consisted of the descriptive analysis of the variable of interest, meaning BMI using the original data without any missing values. The results of this preliminary analysis are enclosed in Table 5.1.

The p-value related to the Shapiro-Wilk test for normality is greater than 0.05, which indicates that the variable BMI from original complete dataset were normally distributed. This result implies the use of mean and standard deviation tests for the summary statistics and descriptive analysis.

Table 5.1: Summary statistics for BMI in the original complete data

Summary statistics	**Original data**
Mean	24.9200
Median	24.0500
Std-dev	6.7110
IQR	7.248698
Shapiro-Wilk test	W=0.907 P-value = 0.0521

Results of the summary statistics of BMI from the original dataset show that the mean BMI is 24.92 with a standard deviation of 6.711. This results show that in general people have a healthy weight (scientifically called normal weight) because the BMI is in range 15 - 25.

5.1.2. Descriptive analysis of data with missing values imputed

In this section, it is a question of using descriptive statistics methods for examining which of the mean imputation, hot-deck, regression and multiple imputation techniques provide the best results and under which condition (type of missingness and percentages of missing).

Table 5.2: Summary statistics for BMI imputed with mean imputation, hot-deck, regression and Multiple linear regression and Multiple (PMM) imputation methods

Mean substitution technique									
Measure of central tendency	**MCAR (BMI)**			**MAR (BMI)**			**MNAR (BMI)**		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Mean	24.11	24.99	25.66	24.67	25.43	25.22	24.51	25.62	25.58
Median	24.29	24.23	25.26	24.64	24.24	23.97	24.17	24.75	24.32
Std-dev	7.088	7.063	7.096	7.211	7.033	7.062	7.183	7.083	7.038
Hot-deck imputation technique									
Mean	24.51	25.68	25.45	24.77	25.68	25.46	24.91	25.41	25.32
Median	24.12	24.75	24.32	24.28	24.22	24.54	24.18	24.87	24.35
Std-dev	7.215	7.166	6.996	7.153	7.306	7.168	7.021	7.684	7.112
Regression imputation technique									
Mean	24.61	25.77	25.27	25.28	25.19	25.38	24.85	25.59	25.12
Median	24.34	24.55	24.35	24.74	24.38	24.57	24.56	24.64	24.74
Std-dev	7.083	7.356	7.352	7.407	7.719	6.749	7.418	7.187	7.240
Multiple imputation (Linear regression) technique									
Mean	25.14	25.86	25.26	24.68	25.77	24.88	24.19	25.77	24.88
Median	24.34	25.11	25.42	24.34	24.38	24.38	24.28	24.78	24.81
Std-dev	7.155	7.422	7.158	6.665	6.973	7.075	6.868	7.038	7.230
Multiple (PMM) imputation technique									
Mean	23.91	25.82	24.85	24.70	25.23	24.64	24.41	24.94	24.77
Median	24.09	24.45	23.86	24.13	24.24	24.34	24.046	24.65	24.21
Std-dev	6.614	7.153	7.320	7.060	6.679	6.808	6.392	6.833	6.131

The table 5.2 includes results of descriptive analysis of BMI with missing values imputed using different techniques from different missing mechanisms. Results summarized in this table are massive and the interpretation will be on based on the mean differences between imputed data and original data.

Hypothesis Test for a Difference between two Means (μ_1 and μ_2).

Let μ_1 denote the mean of the original dataset and μ_2 denote the mean of the imputation method from different missing mechanism across three degrees of missingness (5%, 10% and 25%).

Given the summary statistics above and the sample size of 3631 for both original dataset and imputed dataset, we begin by finding the difference between original data and Mean substitution imputed data under MCAR at 5% missing.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Test statistics: $$Z_o = \frac{\bar{X}_1 - \bar{X}_1 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

To find the critical values:

$\propto = 0.05$, $\frac{\propto}{2} = 0.025$ and 1- $\frac{\propto}{2}$ = 0.975. H_0 is rejected if the p-value is less than $\propto$ (level of significance) otherwise we don't.

Table 5.3: Results of tests for comparison of mean of the completed data and imputed data based on 5 imputation methods applied to 5%, 10% and 25% missingness created by MCAR, MAR and MNAR.

Imputation methods	**MCAR**			**MAR**			**MNAR**		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Mean substitution	2×10^{-11}	0.6672	2×10^{-8}	0.126	1.6×10^{-3}	0.0628	0.012	2×10^{-12}	2×10^{-9}
Hot-deck	0.012	2×10^{-9}	1×10^{-3}	0.3576	2×10^{-9}	1×10^{-3}	0.9522	3.8×10^{-3}	0.0138
Regression	0.0562	2×10^{-10}	0.034	0.03	0.1118	3.6×10^{-3}	06744	2×10^{-10}	0.2224
Multiple imputation linear regression	0.1770	2×10^{-14}	0.0366	0.126	2×10^{-12}	0.8026	2×10^{-12}	2×10^{-12}	0.8026
PMM	2×10^{-14}	2×10^{-12}	0.6672	0.1738	0.0488	0.0768	$\times 10^{-4}$	0.8966	0.3222

Comparing results of mean substitution method under MCAR, MAR and MNAR across all three degrees of missingness (5%, 10% and 25%), results show evidence significant difference between the values of original results (meanBMI = 24.92) and this technique under MNAR data at 5%, 10% and 25% (24.51, 25.62 and 25.58) respectively. Our null hypothesis is actually rejected at this point. Apart from that the results show that there are no significant differences between the values of original data and mean substitution method under MAR at 5% and 25%, the values were close to the one of the original data. Very interestingly, the findings also indicated that there is no difference between the values of this method under MCAR at 10% (meanBMI = 24.99) when compared to the results of the original, p-value of 0.6672 which is greater 0.05.

Regarding the hot-deck technique, under MCAR, MAR and MNAR at 5%, 10% and 25% missing, the results show that hot-deck method performed better under MNAR at 5% (meanBMI = 24.91). This shows that BMI imputed values are very close to the values of the BMI original data. BMI imputed data under MCAR and MNAR across two degrees of missingness (10% and 25%) are far higher compared to the values of the true data. Indeed, at 10% and 25% the values are 25.68 and 25.45 for MCAR and 25.41 and 25.32. This indicate a significant difference between these two comparisons. These results reveal that Hot-deck method provide good result compared to the mean substitution method. Indeed, the mean BMI (24.91) of hot-deck method under MNAR at 5% have the closest value to the original data compared to mean BMI of mean substitution method with 24.99 (MCAR 10%).

In addition to findings presented above, the results in table 5.2 indicate that both regression and Multiple imputation (linear regression) techniques perform better with 5% and 25% missing under MCAR (24.61 and 25.14) and MNAR (25.12 and 24.88) respectively. These findings show that no significant differences are found between these two techniques and original BMI data, values were not far from the values of the original BMI variable. The results also highlighted that the null hypothesis is rejected under MCAR at 10% and 25% in both techniques, this means that there is a significant difference between the values of this technique compared to the values of original data, in fact, the values under these mechanisms are too high compared to the values of the original data.

The last technique to be investigated in this analysis is the predictive mean matching method (PMM). Results reveal that PMM produces the closest value to the original data in situation of 10% missing under MNAR (meanBMI = 24.94), results indicates that there is no significant

difference between the imputed BMI values and original BMI data. Actually this technique provides good results under all three mechanisms at 25% missingness. Indeed, the hypothesis test confirmed that imputed BMI values of this technique 25% under all mechanisms has no differences between the values of original BMI data.

In nutshell, findings from table 5.2, reveal that in situation of 5% missing in a data, Hot-deck method is the best method (meanBMI of 24.91) under MNAR compared to all the above mentioned techniques. In situation of 10% missing, PMM method (meanBMI = 24.94) is the best of all the methods under MNAR mechanism. In the case of 25% missing, results in table 5.2 show that Multiple imputation linear regression method (meanBMI = 24.88) under MNAR provides better results than all the other imputation techniques.

5.2 Multivariate analysis

The comparison of imputation technique at descriptive analysis level is limited. Indeed, it only focuses on comparing the values of mean BMI produced by each imputation technique and the mean BMI of the original data. It is important to analyse the impact of each imputation technique using a multivariate analysis. In this section, linear regression of BMI (dependent variable) and age, education attained, marital status and type of residence as covariates are undertaken in different situations to examine the performance of different imputation methods.

To compare results of the model with imputed data and the original data at multivariate level, the parameters of the linear regression are examined. Three criteria are utilized in this section to reach our objective. Firstly, the best model is the one with the adjusted R-squared R^2 that is very close to the original data and MSE which closer to zero. Secondly, it is the model with significant variables identical to significant variables in the original data. Thirdly and finally, it is the model with regression coefficients very close to the original model. It is based on these three criteria that the best models are selected in this section.

5.2.1. Multivariate analysis of original data

Table 5.4: Linear regression results for BMI using the original dataset

Complete original dataset			
R-square = 0.1365	Adjusted R-square = 0.1355		MSE = 0.505
Independent variables	Coefficients	SE	p-value
Constant	14.084	1.037	< 2e-16
Age	0.181	0.014	< 2e-16
Education (ref = no education)			
Primary	3.981	0.942	2.46e-05
Secondary	5.414	0.944	1.07e-08
Tertiary	6.068	1.016	2.60e-09
Marital status (ref= unmarried)			
Married	2.211	0.275	1.33e-15
Residence type (ref=rural)			
Urban	–0.541	0.249	0.0305

The table 5.4 summarizes the linear regression output on original data with no missing values. These results indicate that 13.55% (adjusted R^2) of the variability in BMI is explained by this model where the explanatory variables are age, education attained, marital status and residence type. Output illustrate that the overall model was statistically significant, and all predictors included in the model are statistically significant at 0.05 level of significance (P-value < 0.05).

The above results shows that the MSE is 0.505, which iindicate that the BMI values are scattered wildly around the regression line, so MSE of 0.505 is as good as it gets (and is in fact, the line of best fit).

The regression coefficient associated with age suggests that a one-year increase in age is associated with a 0.181-unit increase in BMI. Regarding the influence of education level attained on BMI, results show that people with primary, secondary and tertiary education level have 3.981, 5.414 and 6.068 higher BMI compared to those with no education respectively. This result reveals that when the level of education increases, BMI increases. Results show that BMI's are significantly different from marital status. Indeed, married people are likely to have 2.211 higher BMI than not married people. It also appeared that people who live in urban areas are likely to have a 0.541 lower BMI as compared to those in rural areas. This result highlights the fact that living environment can have significant effect on BMI level.

5.2.2. Multivariate analysis of imputed data

Table 5.5: Linear regression results for BMI using mean substituted data

Mean substitution method (MCAR)									
	R-squared = 0.1287 Adjusted R-squared=0.1273(5%) MSE = 0.752			R-squared = 0.1271 Adjusted R-squared = 0.1251 (10%) MSE = 0.825			R-squared = 0.1225 Adjusted R-squared=0.1215(25%) MSE = 0.818		
covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.870	15.362	17.025	0.918	0.895	0.877	< 2e-16	< 2e-16	< 2e-16
Age	0.170	0.140	0.196	0.013	0.011	0.0015	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	3.676	4.656	5.021	0.828	0.812	0.781	9.19e-06	1.19e-16	2.11e-06
Secondary	4.947	5.365	5.624	0.831	0.822	0.752	2.82e-09	2.82e-09	2.32e-10
Tertiary	5.843	6.784	6.871	0.902	0.900	0.891	1.05e-10	2.56e-11	7.05e-15
Marital status (ref=not married)									
Married	2.152	3.125	2.333	0.265	0222	0.210	6.17e-16	10.9e-06	6.19e-16
Residence type (ref=urban)									
urban	-0.626	-0.325	-0.414	0.229	0.215	0.199	0.0063	0.00021	0.00003
Mean substitution method (MAR)									
	R-squared = 0.1134 Adjusted R-squared=0.1121(5%) MSE = 0.711			R-squared = 0.1132 Adjusted R-squared = 0.1111 (10%) MSE = 0.814			R-squared = 0.1215 Adjusted R-squared=0.1195(25%) MSE = 0.723		
covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	18.116	18.215	19.222	0.795	0.699	0.678	< 2e-16	< 2e-16	< 2e-16
Age	0.101	0.132	0.131	0.011	0.010	0.009	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	2.051	3.112	3.254	0.717	0.701	0.091	0.00423	0.00003	0.000001
Secondary	2.853	3.542	3.666	0.719	0.699	0.671	7.43e-05	3.43e-10	3.43e-05
Tertiary	3.346	4.051	4.001	0.719	0.699	0.671	1.88e-05	9.28e-03	1.58e-05
Marital status (ref=not married)									
Married	1.971	2.015	2.515	0.229	0.221	0.196	< 2e-16	< 2e-16	< 2e-16
Residence type (ref=urban)									
urban	-0.311	-0.352	-0.333	0.198			0.11667	0.092	0.071
Mean substitution method (MNAR)									
	R-squared = 0.1243 Adjusted R-squared=0.1223(5%) MSE = 0.705			R-squared = 0.1238 Adjusted R-squared = 0.1213 (10%) MSE = 0.733			R-squared = 0.1275 Adjusted R-squared=0.1245(25%) MSE = 0.759		
covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.603	15.025	15.321	1.018	1.004	0.981	< 2e-16	< 2e-16	< 2e-16
Age	0.165	0.175	0.179	0.014	0.019	0.011	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	3.707	3.881	4.325	0.926	0.855	0.814	6.41e-05	5.41e-07	6.11e-01
Secondary	5.088	5.310	7.198	0.928	0.881	0.844	4.47e-08	2.47e-05	1.47e-08
Tertiary	5.793	6.164	6.525	0.998	0.899	0.811	7.20e-09	4.20e-09	5.20e-08
Marital status (ref=not married)									
Married	2.172	2.443	3.547	0.270	0.211	0.917	1.34e-15	1.20e-09	5.20e-10
Residence type (ref=urban)									
urban	-0.509	-0.625	-0.666	0.245	0.215	0.916	0.0381	0.0012	0.0002

Output of multivariate analysis presented in table 5.5 show the results of the mean imputation technique for different missing data mechanisms at three different percentages (5%, 10% and 25%).

Let start examining the results of the table 5.5 by focusing only on situation of 5% missing. Using the adjusted R-squared as the evaluation criteria, findings show that mean substitution method provides better results under MCAR 5% with an adjusted R^2 of 12.73% (the original

data adjusted R^2 is (13.55%), followed by MNAR 5% with adjusted R^2 of 12.23% and MAR 5% with adjusted R^2 of 11.21%.

The MSE values of this technique under three mechanisms across all three missingness (5%, 10% and 25%) are significantly higher compared to the results of the original.

The regression coefficients used as an evaluation criterion lead to similar conclusion that this technique provides the best results under MCAR 5%. For example, the regression coefficient related to age under MCAR 5% is 0.170 which is closer to the value of the original data (0.181) compared to MNAR and MAR.

At 10% missing, this technique provides better results under MCAR 10% (Adjusted R-squared = 12.51%) followed by MNAR 10% (Adjusted R-squared = 12.13%) mechanisms.

Considering only 25% in the comparison of the 3 missing mechanisms, mean substitution method appears to produce better findings under MNAR at 25% missing with the adjusted R-squared value of 12.45% (original data adjusted R^2 is 13.55%). The regression coefficients associated to the model also confirmed that mean substitution technique produces better results under MNAR at 25%. For example, the coefficient related to age is 0.179 which is closer to the original data (0.181).

In general, the output shows that in situation of unknown percentage of missing mean substitution method should be recommended for the imputation of missing values if they are MCAR. Indeed, this technique perform better at 5% and 10% under MCAR and also relatively perform better under MCAR at 25%.

Table 5.6: Linear regression results for BMI using hot-deck imputed data

Hot-deck imputation method (MCAR)									
	R-squared = 0.1261 Adjusted R-squared= 0.1241(5%) MSE = 0.725			R-squared = 0.1272 Adjusted R-squared = 0.1261 (10%) MSE = 0.720			R-squared = 0.1295 Adjusted R-squared=0.1275(25%) MSE = 0.713		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.792	15.365	17.252	1.038	0.992	0.896	< 2e-16	< 2e-16	< 2e-16
Age	0.162	0.188	0.191	0.014	0.0112	0.0092	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	3.972	4.115	4.512	0.943	0.894	0.862	2.63e-05	1.63e-15	8.63e-11
Secondary	5.151	5.548	6.542	0.945	0.868	0.784	5.46e-08	3.16e-08	7.46e-08
Tertiary	5.843	5.914	6.725	1.017	0.987	0.963	1.01e-08	3.65e-07	1.11e-09
Marital status (ref=not married)									
Married	2.286	2.325	2.551	0.276	0.211	0.196	6.17e-16	2.17e-13	6.11e-12
Residence type (ref=urban)									
urban	-0.605	-0.699	-0.625	0.250	0.214	0.198	0.0157	0.0012	0.00005
Hot-deck imputation method (MAR)									
	R-squared = 0.1070 Adjusted R-squared = 0.1055(5%) MSE = 0.865			R-squared = 0.1045 Adjusted R-squared = 0.1025 (10%) MSE = 0.896			R-squared = 0.1122 Adjusted R-squared=0.1108(25%) MSE = 0.827		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	17.974	18.652	18.785	1.046	1.025	1.009	< 2e-16	< 2e-16	< 2e-16
Age	0.111	0.123	0.153	0.014	0.0093	0.0071	3.71e-14	3.01e-16	8.17e-10
Education (ref=no education)									
Primary	2.244	3.254	3.661	0.951	0.892	0.784	0.018380	0.00154	0.000075
Secondary	3.203	3.655	4.125	0.953	0.911	0.844	0.000785	0.00125	0.02655
Tertiary	4.220	4.254	4.545	1.025	1.022	1.008	3.97e-05	3.25e-03	4.7e-20
Marital status (ref=not married)									
Married	1.887	2.155	2.482	0.278	0.314	0.260	1.32e-11	2.07e-19	7.17e-11
Residence type (ref=urban)									
urban	-0.640	-0.663	-0.695	0.252	0.222	0.211	0.0111	0.0054	0.000091
Hot-deck imputation method (MNAR)									
	R-squared = 0.1335 Adjusted R-squared = 0.1315(5%) MSE = 0.621			R-squared = 0.1321 Adjusted R-squared = 0.1307 (10%) MSE = 0.658			R-squared = 0.1311 Adjusted R-squared=0.1301(25%) MSE = 0.687		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	13.885	14.052	15.546	1.077	1.021	1.005	< 2e-16	< 2e-16	< 2e-16
Age	0.177	0.178	0.199	0.012	0.010	0.0083	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	4.125	5.250	5.641	0.979	0.842	0.784	2.61e-05	1.1e-16	6.1e-18
Secondary	5.561	5.698	6.451	0.981	0.863	0.810	1.58e-08	5.5e-11	2.8e-18
Tertiary	6.256	6.458	7.254	1.054	1.099	0.987	3.23e-09	4.52e-09	6.3e-16
Marital status (ref=not married)									
Married	2.243	2.125	1.987	0.283	0.215	0.203	3.43e-15	2.53e-17	3.25e-12
Residence type (ref=urban)									
urban	-0.564	-0.652	-0.698	0.257	0.211	0.198	0.0287	0.0065	0.0035

The table 5.6 reveals findings of hot-deck imputation method and performance under different missing mechanisms and different percentages.

According to table 5.6, hot-deck imputation method appears to deliver much better results compared to mean substitution method. This finding is established by the values of the adjusted R-squared and also the significance of all the predictors included in the model (p-values are less than 0.05). In fact, hot-deck imputation method produces better results under MNAR at 5%, 10% and 25% with the adjusted R-squared of 13.15%, 13.07% and 13.01% respectively. This result is confirmed by the values of the regression coefficients associated with marital status for example. The regression coefficients associated with married people are 2.243, 2.125

and 1.987, these coefficients present little difference with value of the original (2.211). These results also highlighted that the MSE values are higher compared to the one of the true value but still good since there not far from zero.

Table 5.7: Linear regression results for BMI using data imputed by regression method

Regression imputation method (MCAR)									
	R-squared = 0.1421 Adjusted R-squared = 0.1392(5%) MSE = 0.635			R-squared = 0.1423 Adjusted R-squared = 0.1411 (10%) MSE = 0.626			R-squared = 0.1485 Adjusted R-squared=0.1465(25%) MSE = 0.629		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.246	15.052	15.624	14.246	2.054	3.054	< 2e-16	< 2e-16	< 2e-16
Age	0.174	0.172	0.176	0.174	0.056	0.062	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	4.058	4.325	5.120	4.058	2.625	3.255	1.04e-05	2.07e-06	1.04e-01
Secondary	5.363	5.663	5.692	5.363	1.864	1.965	6.28e-09	6.28e-09	6.11e-09
Tertiary	6.039	6.325	6.871	6.039	1.255	1.352	1.22e-09	3.42e-09	1.23e-04
Marital status (ref=not married)									
Married	2.253	2.650	3.262	2.253	3.254	3.251	6.17e-16	7.18e-13	8.0e-16
Residence type (ref=urban)									
urban	-0.541	-0.624	-0.644	-0.541	0.235	0.421	0.0265	0.00027	0.00052
Regression imputation method (MAR)									
	R-squared = 0.1611 Adjusted R-squared = 0.1609(5%) MSE = 0.752			R-squared = 0.1635 Adjusted R-squared = 0.1621 (10%) MSE = 0.688			R-squared = 0.1695 Adjusted R-squared=0.1675(25%) MSE = 0.701		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	18.116	19.325	19.625	1.011	2.152	2.325	< 2e-16	< 2e-16	< 2e-16
Age	0.101	0.125	0.162	0.014	0.002	0.0003	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	2.051	3.521	3.762	0.918	2.635	2.862	1.04e-05	2.04e-01	2.14e-07
Secondary	2.853	3.658	4.125	0.921	3.125	3.552	6.28e-09	4.25e-19	2.78e-02
Tertiary	3.346	4.842	4.325	0.991	4.018	4.521	1.22e-09	5.23e-03	1.52e-07
Marital status (ref=not married)									
Married	1.971	2.251	2.351	0.268	0.421	0.625	< 2e-16	< 2e-16	< 2e-16
Residence type (ref=urban)									
urban	-0.311	-0.356	-0.421	0.244	0.462	0.662	0.0265	0.0041	0.0495
Regression imputation method (MNAR)									
	R-squared = 0.1421 Adjusted R-squared = 0.1411(5%) MSE = 0.725			R-squared = 0.1575 Adjusted R-squared = 0.1551 (10%) MSE = 0.765			R-squared = 0.1595 Adjusted R-squared=0.1585(25%) MSE = 0.728		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	13.881	14.025	15.021	1.015	1.115	1.254	< 2e-16	< 2e-16	< 2e-16
Age	0.178	0.186	0.195	0.014	0.035	0.049	< 2e-16	< 2e-16	< 2e-16
Education (ref=no education)									
Primary	4.119	4.215	4.632	0.922	0.846	0.812	8.24e-06	2.54e-16	2.31e-06
Secondary	5.543	5.621	5.982	0.924	0.851	0.835	2.22e-09	2.02e-07	4.12e-12
Tertiary	6.223	6.841	6.941	0.994	0.886	0.899	4.45e-10	1.58e-09	1.25e-11
Marital status (ref=not married)									
Married	2.227	2.321	2.625	0.269	0.352	0.451	< 2e-16	< 2e-16	< 2e-16
Residence type (ref=urban)									
urban	-0.561	-0.652	-0.682	0.244	0.325	0.562	0.0218	0.0035	0.00088

Regression imputation methods are applied to data with different missing mechanisms. Results are included in the table 5.7 and compared with results of the previous imputation techniques summarized in table 5.5 and 5.6.

Findings show that regression imputation method produced higher values of adjusted R^2 at 5%, 10% and 25% for all three missing mechanism (minimum adjusted $R^2$of 13.92%) compared to the original data where the adjusted R-squared was 13.55%. For data with 5% missing, the output indicate that regression method perform better under MCAR 5% (adjusted $R^2 = 13.95\%$) followed by MNAR 5% (adjusted $R^2 = 14.11\%$).

Focusing in 10% missing, it appears that the adjusted R-squared produced by regression imputation method are quite different (higher) compared to the one of the original data. However, this method perform better at 10% under MCAR (adjusted $R^2 = 14.11\%$) followed by MNAR (adjusted $R^2 = 15.51\%$). The regression coefficient related to tertiary education among others, is in line with the findings. Indeed, coefficient related to tertiary education level at 10% under MCAR is 6.325 followed by MNAR (6.841) while the regression coefficient of the original data is 6.068.

When the percentage of missing reached 25%, this method performed very poorly with big differences in adjusted R-squared values and regression coefficients compared to the original data.

The above data is scattered wildly around the regression line, so MSE value are as good as it gets.

Multiple imputation techniques as described in the methodology section are theoretically stronger than other methods. Because of the nature of our dependent variable, the multiple imputation linear regression is the first method of this category to be looked at. Results are consined in the table 5.8

Table 5.8: Linear regression results for BMI using data imputed by multiple imputation linear regression method

Multiple imputation (Linear regression) technique (MCAR)									
	R-squared = 0.1385 Adjusted R-squared = 0.1365(5%) MSE = 0.658			R-squared = 0.1401 Adjusted R-squared = 0.1389 (10%) MSE = 0.47			R-squared = 0.1431 Adjusted R-squared=0.1420(25%) MSE = 0.625		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.25	15.021	15.632	0.842	1.025	1.092	0.000	0.000	0.000
Age	0.186	0.187	0.191	0.012	0.041	0.049	0.000	0.000	0.000
Education (ref=no education)									
Primary	4.153	4.652	5.352	0.883	0.971	1.050	0.000	0.000	0.000
Secondary	5.166	5.628	5.981	0.911	1.052	1.159	0.000	0.000	0.000
Tertiary	5.762	6.021	6.821	0.963	1.070	1.260	0.000	0.000	0.000
Marital status (ref=not married)									
Married	1.728	1.789	2.015	0.243	0.325	0.421	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.544	-0.624	-0.640	0.240	0.312	0.515	0.024	0.039	0.0549
Multiple imputation (Linear regression) technique (MAR)									
	R-squared = 0.1175 Adjusted R-squared = 0.1162(5%) MSE = 0.802			R-squared = 0.1225 Adjusted R-squared =0.1201 (10%) MSE = 0.782			R-squared = 0.1245 Adjusted R-squared=0.1225(25%) MSE = 0.755		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	14.37	16.021	16.524	1.126	1.205	1.025	0.000	0.000	0.000
Age	0.184	0.187	0.189	0.017	0.021	0.034	0.000	0.000	0.000
Education (ref=no education)									
Primary	3.952	4.012	4.652	0.752	0.811	0.820	0.000	0.000	0.000
Secondary	4.891	5.321	5.563	0.863	0.897	0.910	0.000	0.000	0.000
Tertiary	5.196	6.522	6.845	0.958	1.025	1.360	0.000	0.000	0.000
Marital status (ref=not married)									
Married	2.054	2.321	3.025	0.322	0.396	0.401	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.536	-0.633	-0.699	0.235	0.325	0.352	0.088	0.0901	0.095
Multiple imputation (Linear regression) technique (MNAR)									
	R-squared = 0.1315 Adjusted R-squared = 0.1295(5%) MSE = 0.586			R-squared = 0.1322 Adjusted R-squared = 0.1301 (10%) MSE = 0.577			R-squared = 0.1331 Adjusted R-squared=0.1310(25%) MSE = 0.569		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	13.87	17.325	17.965	0.843	0.995	1.031	0.000	0.000	0.000
Age	0.183	0.185	0.191	0.012	0.032	0.036	0.000	0.000	0.000
Education (ref=no education)									
Primary	3.922	4.015	4.625	0.699	0.755	0.795	0.000	0.000	0.000
Secondary	4.623	5.236	5.552	0.705	0.861	0.8862	0.000	0.000	0.000
Tertiary	4.989	5.632	5.924	0.821	0.965	0.986	0.000	0.000	0.000
Marital status (ref=not married)									
Married	1.785	1.925	2.069	0.235	0.302	0.332	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.579	-0.606	-0.625	0.240	0.365	0.415	0.016	0.032	0.0421

Under multiple imputation linear regression method in table 5.7, MCAR at 5%, 10% and 25% missing have the adjusted R^2 of 13.65%, 13.89% and 14.20% respectively. As for MAR, the adjusted R^2 are 11.62%, 12.01% and 12.25% at 5%, 10% and 25% respectively. Finally, the adjusted R^2 for MNAR at 5% is 13.01%, at 10% is 12.95% and at 25% it is 13.10%. This

output reveals a pattern that, as the percentage of imputed missing increases in the data, the adjusted R^2 together with regression coefficients increases for all the missing data mechanisms. The results show that multiple imputation linear regression provides better results under MCAR at 5% missing compared to MAR at 5% and MNAR at 5%. This was confirmed by the adjusted R^2 (13.65%) which is the closest to the original data (13.55%) and regression coefficients associated with education (primary = 4.153, secondary = 5.166 and tertiary =5.762).

At 10% missing, this technique produced better results in terms of coefficients under all three missing mechanisms but with the use of adjusted R^2, it appears that multiple linear regression method provided good results under MCAR (13.86%) compared to MAR and MNAR mechanisms. These results indicate that multiple linear regression imputation method is a recommended method to tackle the data with MCAR no matter the percentage of missing.

PMM method is also an important multiple imputation technique which was considered in this study. Results obtained by analysing PMM imputed data in several conditions are included in the table 5.9.

Table 5.9: Linear regression results for BMI using PMM imputed data

Predictive mean matching method (MCAR)									
	R-squared = 0.1345 Adjusted R-squared = 0.1345(5%) MSE = 0.555			R-squared = 0.1375 Adjusted R-squared = 0.1352 (10%) MSE = 0.425			R-squared = 0.1395 Adjusted R-squared=0.1375(25%) MSE = 0.365		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	10%	25%	**5%**	**10%**	**25%**
Constant	14.26	15.254	15.405	0.826	0.862	0.901	0.000	0.000	0.000
Age	0.188	0.183	0.192	0.012	0.035	0.044	0.000	0.000	0.000
Education (ref=no education)									
Primary	4.562	3.998	4.125	0.895	0.925	1.025	0.000	0.000	0.000
Secondary	5.587	5.012	5.233	0.961	1.002	1.145	0.000	0.000	0.000
Tertiary	6.321	6.325	6.524	1.063	1.302	1.232	0.000	0.000	0.000
Marital status (ref=not married)									
Married	1.708	1.985	2.256	0.234	0.444	0.471	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.519	-0.625	-0.640	0.235	0.352	0.410	0.027	0.029	0.04
Predictive mean matching method (MAR)									
	R-squared = 0.1235 Adjusted R-squared = 0.1225(5%) MSE = 0.704			R-squared = 0.1261 Adjusted R-squared = 0.1251 (10%) MSE = 0.685			R-squared = 0.1295 Adjusted R-squared=0.1275(25%) MSE = 0.698		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	13.93	14.025	14.302	0.905	1.025	1.063	0.000	0.000	0.000
Age	0.172	0.179	0.191	0.014	0.016	0.023	0.000	0.000	0.000
Education (ref=no education)									
Primary	4.215	4.623	4.815	0.883	0.887	0.952	0.000	0.000	0.000
Secondary	5.352	5.961	6.359	0.920	0.962	0.971	0.000	0.000	0.000
Tertiary	5.766	6.532	6.995	0.987	1.032	1.120	0.000	0.000	0.000
Marital status (ref=not married)									
Married	2.202	2.403	3.662	0.263	0.277	0.325	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.555	-0.571	-0.590	0.269	0.286	0.333	0.040	0.052	0.071
Predictive mean matching method (MNAR)									
	R-squared = 0.1345 Adjusted R-squared = 0.1355(5%) MSE = 0.432			R-squared = 0.1375 Adjusted R-squared = 0.1365 (10%) MSE = 0.399			R-squared = 0.1385 Adjusted R-squared=0.1375(25%) MSE = 0.365		
Covariates	Coefficients			SE			p-value		
	5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Constant	13.35	13.75	13.95	0.856	0.952	1.050	0.000	0.000	0.000
Age	0.181	0.185	0.187	0.012	0.032	0.033	0.000	0.000	0.000
Education (ref=no education)									
Primary	3.986	4.106	4.220	0.926	0.983	1.056	0.000	0.000	0.000
Secondary	5.416	5.222	5.698	0.963	0.999	1.009	0.000	0.000	0.000
Tertiary	6.069	6.150	6.325	1.097	1.120	1.325	0.000	0.000	0.000
Marital status (ref=not married)									
Married	2.212	2.359	2.663	0.238	0.369	0.398	0.000	0.000	0.000
Residence type (ref=urban)									
urban	-0.540	-0.661	-0.768	0.246	0.310	0.457	0.014	0.026	0.041

The output in table 5.9 indicate some interesting results especially under MNAR 5%. Results under MNAR has produces results that matches with the one of the original data.

For data with 5% missing, findings highlighted that PMM produced best results under MNAR (adjusted $R^2 = 13.55\%$) compared to all other imputation techniques included in this study.

The regression coefficient also agreed with the results. For example, the coefficient associated to the variable age (0.181) matches the results of the original data (0.181).

In the case of 10% missing, PMM perform better under MCAR (adjusted R^2 =13.52%) followed by MNAR (13.65%). The regression coefficient of the variable age confirms that PMM operates well under 10% MCAR (coef = 0.183) followed by 10% MNAR (coef = 0.185).

In situations where a data is suffering 25% missing, PMM technique appears to provide the best performance under two different missing mechanisms (MCAR and MNAR) by considering the adjusted R-squared values (13.75% for each). After taking another evaluation criterion into action (regression coefficients related to age), the findings reveal that this technique can perform much better under MNAR mechanism. Indeed, the regression coefficient associated with age is 0.187 which closer to the original data (0.181).

After multivariate analysis of different imputation methods under all missing mechanisms with 5%, 10% and 25% missingness, the results are compressed in the following paragraphs.

Findings reveal that for 5% MCAR data, multiple imputation linear regression method is the best method to use followed by PMM technique. In situation of 10% and 25% MCAR data PMM appears to be the right choice.

When a data is affected by 5%, 10% and 25% MAR, this study recommends the use of PMM technique followed by multiple imputation linear regression method. These results reveal that any data with MAR problem will be best treated using multiple imputation techniques.

In the situation of MNAR 5% missing data, findings reveal that PMM is the best technique to tackle such data followed by hot-deck method. At 10% and 25% missing, PMM method is still providing better performance than other imputation techniques included in the study.

This technique provided MSE values that are closer to zero but slightly less than actually MSE value from the original data.

In overall, in the light of the above mentioned results, for any data affected by missing values, if the missing mechanism is unknown and or the percentage of missing is also unknown this study suggests the use of PMM method.

CHAPTER 6: DISCUSSION AND CONCLUSIONS

This research addressed two aspects, the first was the creation of missing data mechanism (MCAR, MAR and MNAR) across three degrees of missingness (5%, 10% and 25%) on dependent variable (BMI) from the original complete dataset with no missing values. A second aspect of the study was to impute the created missing values on BMI variable (variable of interest) using different types of imputation techniques. Two sets of imputation techniques, single and multiple imputation, were involved. The analysis in this study was carried out using the free software of R and STATA. For the evaluation of results, the Shapiro-Wilk test was used to check the normality from both the original complete dataset and three missing data mechanisms at three different percentages. The results of the study show that all datasets (original data, MCAR, MAR and MNAR) were normally distributed.

The main aim of this study was to find the best imputation technique by comparing the results of each imputation technique to the results of a complete dataset (original data) using explanatory and multivariate analyses. At the descriptive level, comparisons were made regarding the BMI means of five imputation methods based on three missing data mechanisms. The results show that the best imputation technique for the data that have values MNAR at 5% is the hot-deck method (meanBMI of 24.91) followed by PMM method (meanBMI = 24.85) under MAR, mean substitution, regression and multiple imputation linear regression methods being the least advantageous. At 10% and 25% missing, PMM (meanBMI = 24.94) under MNAR and Multiple imputation linear regression (meanBMI = 24.88) under both MNAR and MAR are the best imputation methods respectively. In summary, descriptive analysis results reveal that if the researcher is not aware of the missing data mechanism, then the PMM technique could be the best imputation technique to consider.

Multivariate analyses results indicated that all the variables included in the model for the original data were statistically significant determinants of BMI (P-value < 0.05). In this case, the adjusted R-squared, MSE and regression coefficients are used as the evaluation criteria. For data that had values MCAR at 5% missing, the results show that multiple imputation linear regression is the best technique followed by PMM method. At 10% and 25% missing, PMM continued proving better performance than other methods. For data that have value MAR at 5%, 10% and25%, this study recommends the use of PMM technique followed by multiple imputation linear regression method. Finally, for the data that had values MNAR, the results show that the PMM method is the best imputation technique compared to all other methods,

followed by the hot-deck. Since multivariate analysis is statistically more powerful than descriptive analysis, this study recommends the use of the PMM technique in the event that the researcher is not aware of the missing mechanism.

From the results, it is clear that there is no better technique for all types of datasets but in the case of studies related to BMI, this study proposes the use of the PMM technique for the imputation of missing values. These results also indicated that the smaller the percentage of missingness the better results can be obtained. It is suggested that the results of this study can be extended to any continuous dataset for any health-related issue, however this extension of the results needs to be undertaken with caution.

REFERENCE LIST

1. Amine E, Baba N, Belhadj M, Deurenbery-Yap M, Djazayery A, Forrester T, Galuska D, Herman S, James W, MBuyamba J, Katan M. Diet, nutrition and the prevention of chronic diseases: report of a Joint WHO/FAO Expert Consultation. World Health Organization; 2002 Jan 1.

2. Allen, I.E. & Seaman, J.E. 2010. Imputation explanation: Find the best way to handle missing data in surveys. *Quality Progress*, *43*(7), 58-60.

3. Allison, P.D. 2005. Imputation of categorical variables with PROC multiple imputation . *SUGI 30 Proceedings*, 113(30), 1-14.

4. Boyko, J. 2013. *Handling data with three types of missing values* (PhD thesis). Department of Statistics, University of Connecticut, Storrs, CT.

5. Butzlaff, I. & Minos, D. 2016. Understanding the drivers of overweight and obesity in developing countries: The case of South Africa. GlobalFood Discussion Papers.

6. Chen, J. & Shao, J. 2000. Nearest neighbor imputation for survey data. *Journal of Official Statistics,* 16(2), 113-131.

7. Croft, T. 2008. DHS data editing and imputation. [Unpublished]. Paper presented at the Demographic and Health Surveys World Conference Washington DC, August 5-7 1991.

8. Donders, A.R.T., van der Heijden, G.J., Stijnen, T. & Moons, K.G. 2006. A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology, 59*, 1087-1091.

9. Dong Y. & Peng, C.Y. 2013. Principled missing data methods for researchers. *SpringerPlus, 2*(1), 222.

10. Durrant, G.B. 2005. Imputation methods for handling item-nonresponse in the social sciences: a methodological review. *ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute. NCRM Methods Review Papers NCRM/002*.

11. Eekhout, I., de Boer, R.M., Twisk, J.W., De Vet, H.C. & Heymans, M.W. 2012. Missing data: A systematic review of how they are reported and handled. *Epidemiology, 23*, 729-732.

12. Enders, C.K., 2010. *Applied missing data analysis*. New York: Guilford Press.

13. Fayers, P.M., Curran, D. & Machin, D. 1998. Incomplete quality of life data in randomized trials: Missing items. *Statistics in Medicine,* 17, 679-696.

14. Ferrari, G.T. & Ozaki, V. 2014. Missing data imputation of climate datasets: Implications to modeling extreme drought events. *Revista Brasileira de Meteorologia, 29*, 21-28.

15. Graham, J.W., 2012. Analysis of missing data. In *Missing data* (pp. 47-69). Springer, New York, NY.

16. Groves, R.M. 2006. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 7(5), 646-675.

17. Hendry, G.M., Zewotir, T., Naidoo, R.N. and North, D., 2017. The effect of the mechanism and amount of missingness on subset correspondence analysis. *Communications in Statistics-Simulation and Computation*, *46*(9), pp.7100-7115.

18. Huisman, M., 2009. Imputation of missing network data: some simple procedures. *Journal of Social Structure*, *10*(1), pp.1-29.

19. Holt, D. & Smith, T.F. 1979. Post stratification. *Journal of the Royal Statistical Society. Series A (General)*, 142, 33-46.

20. Houchens, R. 2015. Missing data methods for the NIS and the SID. HCUP Methods Series Report# 2015-01. Agency for Healthcare Research and Quality [accessed on June 22, 2015]. Available at http://www.hcup-us. ahrq. gov/reports/methods/methods. jsp.

21. https://tradingeconomics.com/lesotho/population-imf-data.html.

22. Huisman, M., 2014. Imputation of Missing Network Data: Some Simple Procedures. *Encyclopedia of Social Network Analysis and Mining*, pp.707-715.

23. Humphries, M. (2013). Missing data and how to deal: An overview of missing data. Population Research Center, University of Texas. Available online at: https://liberalarts.utexas.edu/prc/_files/cs/Missing-Data.pdf.

24. Katz, T.L. 2015. Missing data in clinical trials forum. *Clin. Invest. (Lond.), 5*(8), 681-685.

25. Kim, J-O. & Curry, J. 1977. The treatment of missing data in multivariate analysis. *Sociological Methods and Research,* 6(2), 215-240.

26. Kombo, A.Y., Mwambi, H. and Molenberghs, G., 2017. Multiple imputation for ordinal longitudinal data with monotone missing data patterns. *Journal of Applied Statistics*, *44*(2), pp.270-287.

27. Lee, K.J. & Simpson, J.A. 2014. Introduction to multiple imputation for dealing with missing data. *Respirology*, *19*(2), 162-167.

28. Little, R.J.A. & Rubin, D.B. 2002. Bayes and multiple imputation. In R.J.A Little & D.B. Rubin (Eds.). *Statistical analysis with missing data* (200-220). New Jersey: John Wiley & Sons, Inc.

29. Liu, Y. & Gopalakrishnan, V. 2017. An overview and evaluation of recent machine learning imputation methods using cardiac imaging data. *Data, 2*(1), 8.

30. Lohr, S.L. 2009. *Sampling: Design and analysis*. USA: Cengage Learning.

31. Marshall, A., Altman, D.G., Holder, R.L. & Royston, P. 2009. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Medical Research Methodology,* 9, 57.

32. Ministry of Health [Lesotho] & ICF International. 2016. *Lesotho demographic and health survey, 2014*. Maseru: Lesotho Ministry of Health & ICF International.

33. Ministry of Health and Social Welfare (MOHSW) [Lesotho], Bureau of Statistics (BOS) [Lesotho], and ORC Macro. 2005. Lesotho Demographic and Health Survey 2004. Calverton, Maryland: MOH, BOS, and ORC Macro.

34. Nakagawa, S. & Freckleton, R.P. 2008. Missing inaction: the dangers of ignoring missing data. *Trends in Ecology & Evolution, 23*(11), 592-596.

35. Peugh, J.L. & Enders, C.K. 2004. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research,* 74(4), 525-556.

36. Puoane, T., Steyn, K., Bradshaw, D., Laubscher, R., Fourie, J., Lambert, V. & Mbananga, N. 2002. Obesity in South Africa: the South African demographic and health survey. *Obesity Research, 10*(10), 1038-1048.

37. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J. & Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85-95.

38. Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. New York: John Wiley & Sons.

39. Sattar, A., Baig, S. & Bashir, B. 2013. Factors affecting BMI: Assessment of the effect of sociodemographic factors on BMI in the population of Ghulam Mohammad Abad Faisalabad. *Professional Medical Journal, 20*(6), 956-964.

40. Schafer, J.L. & Graham, J.W. 2002. Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

41. Schenker, N., Raghunathan, T.E., Chiu, P-L., Makuc, D.M., Zhang, G. & Cohen, A.J. 2010. Multiple imputation of family income and personal earnings in the National Health Interview Survey: Methods and examples. *National Center for Health Statistics*.

42. Schmitt, P., Mandel, J., & Guedj, M. 2015. A comparison of six methods for missing data imputation. *Journal of Biometrics and Biostatistics, 6*(1), 1-6.

43. Song, Q. & Shepperd, M. 2007. A new imputation method for small software project data sets. *Journal of Systems and Software, 80*, 51-62.

44. Stata, A., 2009. STATA MULTIPLE-IMPUTATION REFERENCE MANUAL RELEASE 13.

45. Sturge, M.D. & Horsley, V. 1911. On some of the biological and statistical errors in the work on parental alcoholism by Miss Elderton and Professor Karl Pearson, F.R.S. *British. Medical. Journal*, January 14, 72-82.

46. Tavakoli, A.S., Scharer, K. & Hussey, K. (2011) Compare imputation and no imputation to examine mediator effect for social support of mothers of mentally ill children. Paper 215-2011. SAS Global Forum.

47. Van den Berg, V.L., Seheri, L. and Raubenheimer, J., 2014. Body mass index of 16-year olds in urban Maseru, Lesotho. *African journal of primary health care & family medicine*, *6*(1), pp.1-14.

48. Vatanen, T., 2012. Missing value imputation using subspace methods with applications on survey data.

49. Vink, G., Frank, L.E., Pannekoek, J. & Buuren, S. 2014. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica, 68,* 61-90.

50. Wang, J. 2003. *Data mining: opportunities and challenges,* IGI Global.

51. White, I.R., Royston, P., & Wood, A.M. 2011. Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine, 30*, 377-399.

52. Yan, T., & 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.

53. Yuan, Y.C. 2010. Multiple imputation for missing data: Concepts and new development (Version 9.0). *SAS Institute Inc, Rockville, MD*, 49, 1-11.

54. Zainuri, N.A., Jemain, A.A., & Muda, N. 2015. A comparison of various imputation methods for missing values in air quality data. *Sains Malaysiana, 44*, 449-456.

Appendix

Summary statistics for BMI in the original complete data and data with missingness

Summ STATS	**Original data**	**MCAR (BMI)**			**MAR (BMI)**			**MNAR (BMI)**		
		5%	**10%**	**25%**	**5%**	**10%**	**25%**	**5%**	**10%**	**25%**
Mean	24.9200	24.5400	24.923	24.771	24.2100	24.53	24.51	24.2152	24.915	25.521
Median	24.0500	24.0600	24.136	24.051	23.2900	23.71	23.34	24.0496	23.532	23.75
Std-dev	6.7110	6.4000	6.7	6.772	6.390	6.412	6.07	6.3935	0.487	0.817
IQR	7.248698	7.07527	7.207	7.237	6.404877	6.77	6.631	7.04869	7.248	6.921
Shapiro-Wilk test	W=0.907 P-value =2.2e - 16	W=0.9043 3 P-value =2.2e - 16	W=0.9 P-value =2.2e16	W=0.87 P-value =2.2e16	W=0.874 P-value =2.2e - 16	W=0.78 P-value =2.2e-16	W=0.81 P-value =2.2e16	W=0.9 P-value =2.2e-16	W=0.87 P-value =2.2e-16	W=0.907 P-value =2.2e-16

Summary statistics for BMI imputed with mean substitution imputation method

Mean substitution technique									
Missing completely at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.280	23.51	24.55	24.211	23.52	27.25	23.6300	24.72	28.63
Median	24.31	24.16	24.41	24.21	24.18	24.31	25.11	24.35	26.33
Std-dev	7.809	6.423	7.032	7.165	7.211	6.913	6.881	6.888	7.552

Summary statistics for BMI imputed with mean substitution imputation method

Mean substitution technique									
Missing at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.12	24.81	25.09	23.850	25.12	27.31	24.33	25.02	26.31
Median	24.310	24.24	24.16	23.8500	24.88	25.21	24.1100	23.65	24.15
Std-dev	7.141	6.457	7.666	6.115	7.121	7.963	7.125	6.525	7.135

Summary statistics for BMI imputed with mean substitution imputation method.

Mean substitution technique									
Missing not at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.340	25.01	24.19	24.2300	25.11	27.51	24.51	25.31	26.12
Median	24.140	24.444	23.911	23.741	24.68	25.81	24.10	23.80	25.11
Std-dev	6.425	7.113	7.012	7.125	7.821	9.003	6.925	7.105	7.885

Summary statistics for BMI imputed with hot-deck imputation method

Hot-deck imputation technique									
Missing completely at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.230	25.11	24.19	24.4100	25.33	27.32	24.7100	25.51	26.53
Median	24.190	24.151	24.26	23.80	24.74	25.721	24.1400	23.65	25.16
Std-dev	7.011	7.403	7.332	7.865	7.221	7.603	7.025	7.165	8.200

Summary statistics for BMI imputed with hot-deck imputation method

Hot-deck imputation technique									
Missing at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.430	25.11	24.77	24.2300	25.31	27.51	24.330	25.42	26.62
Median	24.020	24.24	24.10	23.95	24.31	24.41	24.130	24.44	25.06
Std-dev	7.101	7.123	7.135	6.805	8.011	8.103	7.225	7.065	7.215

Summary statistics for BMI imputed with hot-deck imputation method

Hot-deck imputation technique									
Missing not at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.430	25.11	25.18	24.50	24.61	27.11	24.51	25.12	26.32
Median	24.16	24.12	24.26	24.26	25.25	25.11	23.8500	24.13	25.06
Std-dev	6.511	7.053	7.100	7.520	8.011	8.821	7.015	6.865	7.565

Summary statistics for BMI imputed with regression imputation method

Regression imputation technique									
Missing completely at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.240	25.00	24.59	24.5100	25.41	27.41	24.1900	25.11	26.52
Median	24.410	24.06	24.54	24.200	24.25	25.21	23.920	24.851	24.76
Std-dev	6.124	7.003	7.122	6.805	6.121	8.143	6.385	8.465	7.205

Summary statistics for BMI imputed with regression imputation method

Regression imputation technique									
Missing at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.240	26.51	25.09	23.9100	25.41	26.25	24.5200	25.22	26.42
Median	23.910	25.23	25.10	24.510	24.58	26.01	24.510	24.20	24.35
Std-dev	6.213	6.553	7.565	7.525	8.001	7.633	5.999	6.215	7.035

Summary statistics for BMI imputed with regression imputation method

Regression imputation technique									
Missing not at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	25.140	25.41	23.99	24.5500	25.01	27.21	23.69	25.13	26.55
Median	24.970	24.26	24.46	23.8800	24.18	25.08	23.980	25.35	24.88
Std-dev	6.322	7.113	8.582	7.105	6.125	7.333	6.352	7.816	7.552

Summary statistics for BMI imputed with Multiple linear regression imputation method

Multiple linear regression technique									
Missing completely at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.22	26.01	25.19	24.5100	25.55	27.51	23.56	26.21	26.02
Median	24.410	24.06	24.54	24.500	25.11	25.71	24.280	24.75	24.22
Std-dev	6.882	7.153	8.032	7.775	7.220	6.673	7.444	6.805	7.225

Summary statistics for BMI imputed with Multiple linear regression imputation method

Multiple linear regression technique									
Missing at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	25.540	24.61	23.89	24.1100	26.21	27.01	25.00	24.63	25.021
Median	24.210	25.36	24.16	24.580	24.51	26.01	24.1400	23.75	25.26
IQR	7.411	6.453	7.532	6.865	8.121	7.633	7.325	6.865	8.235

Summary statistics for BMI imputed with Multiple linear regression imputation method

Multiple linear regression technique									
Missing not at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.40	23.51	24.66	25.2100	25.31	27.00	25.3500	24.67	27.23
Median	23.990	24.81	24.06	23.850	24.28	26.21	24.2400	23.95	26.26
Std-dev	6.887	7.143	6.582	7.100	9.001	8.013	6.355	7.335	9.001

Summary statistics for BMI imputed with PMM imputation method

Predictive mean matching (PMM) technique									
Missing completely at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	23.650	23.89	24.19	25.5100	24.63	27.31	23.810	24.63	26.13
Median	24.110	24.13	24.02	23.7500	24.88	24.71	23.4400	23.85	24.30
Std-dev	7.010	7.311	6.522	7.765	7.021	6.673	7.250	6.445	7.266

Summary statistics for BMI imputed with PMM imputation method

Predictive mean matching (PMM) technique									
Missing at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	23.540	26.01	24.55	24.530	25.03	26.13	24.500	24.33	25.10
Median	23.910	24.06	24.41	23.950	24.65	24.11	23.9500	24.80	24.26
Std-dev	6.515	6.663	7.002	6.995	7.111	6.633	7.525	6.865	8.035

Summary statistics for BMI imputed with PMM imputation method

Predictive mean matching (PMM) technique									
Missing not at random (BMI)									
Measure of central tendency	**3 simulations at 5% missing**			**3 simulations at 10% missing**			**3 simulations at 25% missing**		
	1st	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Mean	24.440	24.51	24.29	24.210	24.61	26.00	24.600	24.61	25.12
Median	24.08	24.11	23.95	23.860	25.08	25.01	24.4100	23.95	24.26
Std-dev	7.231	7.413	7.532	7.865	7.021	7.613	7.325	7.515	6.555

Results of MCAR after fitting linear regression model.

Missing complete at random (5%)			
Multiple R-square = 0.1337	Adjusted R-square = 0.1319	F-statistics: 76.88	
Covariates	Coefficients	SE	p-value
Constant	14.246	1.054	< 2e-16
Age	0.174	0.015	< 2e-16
Education (ref=no education)			
Primary	4.058	0.954	2.18e-05
Secondary	5.362	0.956	2.24e-08
Tertiary	6.039	1.032	5.41e-09
Marital status (ref=not married)			
Married	2.253	0.283	2.61e-15
Residence type (ref=rural)			
Urban	-0.540	0.256	0.0351
Missing complete at random (10%)			
Multiple R-square = 0.1317	Adjusted R-square = 0.1311	F-statistics: 67.79	
Covariates	Coefficients	SE	p-value
Constant	15.254	1.025	< 2e-16
Age	0.193	0.012	< 2e-16
Education (ref=no education)			
Primary	4.652	0.985	2.5e-17
Secondary	5.965	0.988	4.1e-11
Tertiary	6.784	0.994	5.3e-06
Marital status (ref=not married)			
Married	2.514	0.277	5.1e-04
Residence type (ref=rural)			
Urban	-0.623	0.196	0.0084
Missing complete at random (25%)			
Multiple R-square = 0.1299	Adjusted R-square = 0.1278	F-statistics: 72.58	
Covariates	Coefficients	SE	p-value
Constant	17.241	1.059	< 2e-16
Age	0.197	0.035	< 2e-16
Education (ref=no education)			
Primary	5.214	1.002	3.62e-15
Secondary	6.325	0.985	7.02e-08
Tertiary	6.985	0.969	8.21e-12
Marital status (ref=not married)			
Married	1.987	1.874	2.28e-13
Residence type (ref=rural)			
Urban	-0.632	0.163	0.00965

Results of MAR after fitting linear regression model.

Missing at random (5%)			
Multiple R-square = 0.1217	Adjusted R-square = 0.1196	F-statistics: 57.99	
Covariates	Coefficients	SE	p-value
Constant	14.063	1.264	< 2e-16
Age	0.190	0.017	< 2e-16
Education (ref=no education)			
Primary	3.991	1.158	0.000576
Secondary	5.214	1.159	7.19e-06
Tertiary	5.662	1.237	4.93e-06
Marital status (ref=not married)			
Married	1.972	0.303	9.50e-11
Residence type (ref=urban)			
Urban	-0.477	0.283	0.091883
Missing at random (10%)			
Multiple R-square = 0.1156	Adjusted R-square = 0.1138	F-statistics: 77.91	
Covariates	Coefficients	SE	p-value
Constant	15.066	1.058	< 2e-16
Age	0.195	0.041	< 2e-16
Education (ref=no education)			
Primary	5.236	1.123	0.000365
Secondary	6.562	0.996	3.65e-03
Tertiary	7.001	0.976	5.23e-16
Marital status (ref=not married)			
Married	2.015	0.253	0.00000535
Residence type (ref=urban)			
Urban	-0.726	0.222	0.07812
Missing at random (25%)			
Multiple R-square = 0.1317	Adjusted R-square = 0.1299	F-statistics: 47.99	
Covariates	Coefficients	SE	p-value
Constant	17.059	1.421	< 2e-16
Age	0.175	0.0214	< 2e-16
Education (ref=no education)			
Primary	4.051	1.025	4.3e-11
Secondary	5.632	1.325	5.2e-02
Tertiary	5.941	1.362	4.33e-09
Marital status (ref=not married)			
Married	2.035	0.363	9.50e-11
Residence type (ref=urban)			
Urban	-0.576	0.253	0.061883

Results of MAR after fitting linear regression model.

Missing not at random (5%)			
Multiple R-square = 0.1342	Adjusted R-square = 0.1325	F-statistics: 77.40	
Covariates	Coefficients	SE	p-value
Constant	13.885	1.077	< 2e-16
Age	0.177	0.012	< 2e-16
Education (ref=no education)			
Primary	4.125	0.979	2.61e-05
Secondary	5.561	0.981	1.58e-08
Tertiary	6.256	1.054	3.23e-09
Marital status (ref=not married)			
Married	2.243	0.283	3.43e-15
Residence type (ref=urban)			
urban	-0.564	0.257	0.0287
Missing not at random (10%)			
Multiple R-square = 0.1269	Adjusted R-square = 0.1245	F-statistics: 68.00	
Covariates	Coefficients	SE	p-value
Constant	15.663	1.098	< 2e-16
Age	0.182	0.0452	< 2e-16
Education (ref=no education)			
Primary	3.999	1.023	0.0000213
Secondary	4.652	0.996	0.0000238
Tertiary	5.625	0.895	3.13e-17
Marital status (ref=not married)			
Married	2.231	0.325	< 2e-16
Residence type (ref=urban)			
urban	-0.645	0.189	0.00358
Missing not at random (25%)			
Multiple R-square = 0.1347	Adjusted R-square = 0.1315	F-statistics: 84.12	
Covariates	Coefficients	SE	p-value
Constant	15.352	1.111	< 2e-16
Age	0.198	0.042	< 2e-16
Education (ref=no education)			
Primary	4.598	1.160	4.23e-07
Secondary	5.669	0.988	4.1e-16
Tertiary	6.254	0.987	1.13e-23
Marital status (ref=not married)			
Married	3.251	0.257	0.0000365
Residence type (ref=urban)			
urban	-0.674	0.222	0.000232