UNIVERSITY OF GHANA
COLLEGE OF BASIC AND APPLIED SCIENCE

COMPARISON OF IMPUTATION METHODS FOR MISSING VALUES IN LONGITUDINAL DATA

BY

JOHNSON KATSEKPOR

(ID.NO.10556546)

THIS THESIS IS SUBMITTED TO THE UNIVERSITY OF GHANA, LEGON IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE AWARD OF MPHIL STATISTICS DEGREE

JUNE, 2017

DECLARATION

I hereby declare that this submission is my own work towards the award of the M. Phil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

JOHNSON KATSEKPOR	.....	.....
Student	Signature	Date
(10556546)		

Certified by:

DR. ANANI LOTSI	.....	.....
Supervisor	Signature	Date

Certified by:

DR. LOUIS ASIEDU	.....	.....
Supervisor	Signature	Date

DEDICATION

To God Almighty, in whom lies all the treasures of wisdom and knowledge.

ABSTRACT

Longitudinal data are common in various sectors where repeated measurements on a dependent variable are collected for all subjects. Missing data pattern are caused when most planned measurements are unavailable for some subjects. The dropout process may cause three missing values mechanism, namely: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). The missing values have influence on quantitative study that can be serious, leading to biased estimates of parameters, information loss, reduced statistical power, increased standard errors, and weakened generalization of findings. This thesis compared the performance of seven (7) techniques of imputing missing values under the assumptions of MCAR and MAR mechanisms. The study adopted the little's test to check whether a dataset with missing values is MCAR or MAR. The techniques for solving missing values problems were compared using the Generalized Estimating Equation (GEE) model for the complete dataset, the coefficient of determination and root mean squared error (RMSE). The study discovered that when large (above 10%) or small (below 10%) values are missing at random (MAR), it is important to use multiple imputation or expectation maximization to replace missing values in the dataset. The pairwise deletion is the best under MCAR mechanism. Listwise deletion and the hot deck imputation methods performed poorly under the MCAR mechanism. It is recommended that researchers should understand the patterns of missing values in dataset and clearly recognize missing data problems and the situations under which they occurred. However, further research is needed to find a better method for imputing missing not at random (MNAR) with multiple imputation. This thesis focused on missing values in a longitudinal dataset. However, future research using categorical data is a step in right the direction.

ACKNOWLEDGMENT

I thank the Almighty God who has given me the care, knowledge and the opportunity to pursue education up to this level.

I am highly indebted to my project supervisors Dr. Anani Lotsi, and Dr. Louis Asiedu for their countless guidance, advice and constructive criticisms throughout this work. I would also like to express my profound gratitude to Mr. Mawutor Fleku, who helped me to settle on the thesis topic and his immense and continuous encouragement and support. Additional, I am grateful to all the lecturers of Department of Statistics.

Finally, I thank my family for their prayers and support throughout the period of my study. May the Almighty God Bless Us All.

CONTENTS

LIST OF ABBREVIATION

AMP ..... Arbitrary Missing Patterns

CD4 ..... Cluster of Differentiation 4

EM ..... Expectation Maximization

GEE ..... Generalized Estimation Equation

GLM ..... Generalized Linear Models

HMC ..... Homogeneity of Mean and Covariance

IM ..... Imputation Methods

IRLS ..... Iteratively Reweighted Least Squares

LOCF ..... Last Observation Carried Forward

LVCF ..... Last Value Carried Forward

MAR ..... Missing at Random

MCAR ..... Missing Completely at Random

MMRM ..... Mixed Effects Model for Repeated Measurements

MI ..... Multiple Imputation

ML ..... Maximum Likelihood

MMP ..... Monotone Missing Patterns

MNAR ..... Missing Not at Random

MSE ..... Mean Square Error

NIDS ..... National Income Dynamic Study

NA ..... Not Available

OLS ..... Ordinary Least Squares

RMSE ..... Root Mean Square Error

SALDRU ..... Southern Africa Labour and Development Research Unit

SSE ..... Sum of Squares Error

SST ..... Sum of Squares Total

UMP ..... Univariate Missing Data Pattern

WGEE ..... Weighted Generalized Estimating Equation

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

According to Hedeker and Gibbons (2006), longitudinal study is an observational research technique in which data is collected for the same subjects repeatedly over a time period. Longitudinal research projects can spread over years or even decades. For instance, people with HIV may be followed over time and monthly measures such as CD4 (cluster of differentiation 4) counts, or viral load are collected to describe immune status and disease burden respectively. Such repeated measures data are correlated within subjects and thus involve special statistical methods for valid analysis and inference. Longitudinal studies are now regularly used in biology, psychology, social, public health and clinical research (Singer & Willett, 2002).

There are many types of longitudinal research designs, including prospective (cohort or follow up) designs, retrospective (case-control) designs, observational designs, and experimental designs. The prospective longitudinal research design is used to collect data on subjects going forward in time. However, subjects are sampled with and without risk factors, they are monitored over a period of time to repeatedly measure a defined outcome variable. The retrospective longitudinal study design does just the contrary. The retrospective longitudinal study is used to collect data on subjects going backwards in time where the outcome variable for both cases (those already known to have disease based on their outcome) and controls (those already known to not have the disease) is repeatedly collected backwards in time. For example, a researcher may look for a trend when he finds the medical records of previous years.

Cohort studies sample a cohort, and defined as a group undergoing some event in a certain period of time and performing a cross-section through some time intervals. In Panel studies, cross-sectional data are also used and similar group of individuals are compared at intervals through time, but the sample is not always classified as a cohort, as it can be a group of subjects that fail to share a similar event. Hence, a cohort study can be a panel study, but a panel study is not always classified as a cohort study.

Longitudinal research has many benefits over cross-sectional studies (Hedeker & Gibbons, 2006). First, in order to achieve the same statistical power, smaller subjects are needed in longitudinal studies. This is because the repeated measurements from a single subject deliver more information than a single measurement of a single subject. Second, in a longitudinal research, each subject can assist as his or her own control. Generally, intra-subject variability is much less than inter-subject variability. Third, investigators are able to separate timing effects from cohort effects. Finally, longitudinal studies can give information on individual change, which could not be provided by cross-sectional studies. However, longitudinal studies are also having their own challenges. There are several reasons, that is both practical and theoretical, which make the longitudinal analysis very difficult. Such reasons include, but are not limited to, correlation between repeated outcome measurements, missing data, irregularly timed data, mixture of static and time varying covariates, and availability of software for model fitting. This thesis focuses on missing data in longitudinal study.

1.1 Background to the Study

Missing data are a collective problem in several research areas. In survey-based study, respondents may deliberately refuse to answer certain questions due to privacy reasons, or even reluctance of participants to attend the meeting for

evaluation. For example, in clinical research, missing data are also a main issue, especially in longitudinal clinical trials. In survival analysis, the data lack usually occurs in the covariates.

Incomplete data and Missing values are common in social, clinical and institutional research. Missing values problems occur regardless of how cautiously the researcher designs the experiment. Applying wrong methodology to solve missing values problem can reduce the sample size. As a result, this affects the precision of confidence intervals, causes biased estimates and reduces statistical power of the research. It is very important for investigators to use correct methods of analysis to solve missing values problems in order to sustain internal (power of research) and external (generalization of sample results to larger population) validity of research. Properly dealing with missing values can be problematic as it involves cautious scrutiny of the data to detect the type and pattern of missing values, and also a perfect understanding of how the different imputation methods work.

Longitudinal studies are extensively used in clinical trial and public health to measure results over time period within individuals as well as differences in variations across individuals. Unfortunately, missing data are predominant in longitudinal studies. The inference built merely upon the observed outcomes may be seriously flawed. How to handle missing values is a very important problem in longitudinal research. According to most papers, it is clear that no common method can be considered definitive for any missingness situations. Instead, different methods need to rely on the different patterns and types of missing data. It is very significant for the researcher to find and report any patterns of missingness (Schafer & Graham, 2002). Doing so not only helps the consumer of the research to understand the data more completely, but it also defends the choice of the data imputation method used by the researcher.

1.2 Problem Statement

Missing values usually occur in longitudinal studies. They occur when in the current observation there is no data values that are kept for the variable. The problem of attrition or drop-out are mainly involve in missing data, that is, before intended completion of longitudinal study or research, some individuals drop out or withdraw from the study. Hence, the record for the subject dismisses prematurely.

Little (1995) proposed that missing data have three significant implications for longitudinal data analysis. First, when data are missing in longitudinal study, the data set is certainly not balanced over the time period since not all individuals have similar number of repeated measurements at a common set of occasions. This imbalance data makes the methods of analysis change from the one of balanced data. Secondly, there must be some loss of information and also reduction in the sample size when there are missing values. The missing values spread sporadically over several subjects and how highly correlated the missing data are with the observed data will affect loss of precision. Finally, under certain circumstances, missing value can contributes to bias and thereby lead to misleading inferences about changes in the mean response. The higher attrition is likely to have bias and the potential for serious bias makes the longitudinal analysis more complicated.

Selecting the most suitable technique to solve the problem of missing data during analyses is one of the most difficult decisions researchers go through. Most often, missing values are overlooked rather to use suitable imputation technique to replace them.

In view of the above mentioned problems, the following data imputation methods such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, Last

observation carried forward (LOCF) are based on the assumption that data are MCAR and also Multiple imputation, and Expectation maximization are also based on the assumption that data are MAR will be compared to know the best imputation method to solve the problem of missing values. It is against this background that this study is being undertaken to research and compare the best imputation technique for missingness in longitudinal data.

1.3 Objectives

The main objective of the study is to identify the best methods for solving missing data problems in the context of Generalised Estimating Equation and to compare various imputation techniques for handling missing data. The specific objectives of the study are:

- To investigate potential reasons why values are missing in longitudinal data (NIDS data).
- To identify the best imputation method for handling missing values in longitudinal data.
- Develop a practical and useful procedure for estimating missing values in incomplete data.

1.4 Significance of the Study

Missing values problems occur in many areas of statistical applications. For example, missing value exits when an individuals withdraw from survey research prematurely before the study intended completion. Data imputation methods are significant because inference based on overlooking missing mechanisms undermines efficiency and often results in biases and misleading conclusions. Furthermore, the total effect of missingness in many variables often results in exclusion of a substantial proportion of the actual sample, which in turn leads to

a substantial loss of precision and statistical power.

Proper methods to solve missing values problems in longitudinal data may lessen the loss of precision and power resulting from exclusion of individuals with incomplete predictor variables and also reduce bias. Good Imputation method uses all the information collected and helps the researcher to perform accurate analysis of a complete data set using standard complete data techniques. Using proper imputation method to replace missing values also build good associations between variables and give valid estimates of the parameters.

1.5 Research Questions

Questions that would be answered in this research are:

- Which imputation method is the best for handling missing values in longitudinal studies?
- What are the main reasons why values are missing in longitudinal data?
- What are the imputation techniques used to handle missing values in longitudinal data?

1.6 Methodology

In order to know the best imputation method, data imputation methods, namely; Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, Last observation carried forward (LOCF), Multiple imputation, and Expectation maximization will be used to replace missing values which would be created artificially and assessed.

The Generalized Estimating Equations (GEEs) represent an extension to the generalized linear model to accommodate correlated data. The GEEs are a

widely used method for estimating the covariate coefficients of marginal models for repeated binary responses. We explored GEE to analysed the full data and also each of the imputation methods to know the best imputation method. The root mean square error (RMSE) and coefficient of determination (R^2) will also help to assess the performance of each imputation method.

1.7 Scope of the Study

This thesis used national income dynamics study (NIDS) data from South Africa (i.e. wave 1 and wave 2). It made reference to only one stage of repeated study. Specifically, populations included in this study only consisted of Two Waves, that is a study on the same population at only two time points. Missing values were, throughout the thesis, assumed ignorable. This assumption is based on the fact that no one knows the reason why data are missing in the database that are considered in this work. There is also no known method that can test the validity of the assumption. The best that can be done is to assume it is ignorable and this assumption is suggested until more methods of testing this are available (Schafer & Olsen, 1998). Relaxing this assumption will essentially imply a replacement with a similar assumption, which cannot be tested. When data are missing for reason beyond the control of researchers, no one can tell if the assumption is still valid. Additionally, data may be missing by more than one mechanisms. It will therefore be assumed in this thesis that departures from MAR are minimal and will not cause a big degradation in the accuracies of the prediction.

1.8 Sources of Data

Data from the National Income Dynamics Study (NIDS) from South Africa was used to investigate the best imputation method for handling missing values in longitudinal data. NIDS is a programme to compile comprehensive longitudinal information on persons selected for the study and to find out who is moving ahead

and who is falling behind. These data are also key for research and policy makers.

The first study was conducted in 2008, and the data were compiled into the Wave 1 dataset. The second, Wave 2 dataset was compiled after the second visit was made to the same group of people between 2010 -2011.

1.9 Limitations of the Study

- Creating percentage of missing values in a random way was very challenging. This is because most of the statistical software resulted in creating the missing values in a non-random way.
- Financial and other resources are also not available to personally collect primary data (longitudinal data) to compare the result to the real life data (secondary data).

1.10 Organization of the study

This study will be organized into five chapters. Chapter one introduces the entire research work. It expounds the background of the study, problem statement, objectives of the study, the significance of the study, research questions, a brief methodology of the study, scope of the study, sources of the data and limitations of the study. Chapter two presents the literature review, which looks at work done by other researchers. Chapter three is concerned with the detailed methodology used in the study. Chapter four is results of data analysis. It consists of graphical and tabulation of results for discussion. Chapter five, looks at the summary of findings, conclusions, limitations, recommendations, and future research proposal of the study.

CHAPTER 2

LITERATURE REVIEW

2.1 Missing Data

Missing data are observations which exist but were not recorded or recorded and then lost. In clinical studies missing data often result from withdrawal, attrition and inability to follow up. In other settings the missing data could be generated through a coarsening design. An example of a coarsening design is when continuous data are deliberately collected in intervals to maximize the chance of response. A classic example is collecting income in ranges, see Heitjan & Rubin, 1991.

Practically "all methods of statistical analysis are affected by problems with missing values. It is well known that the use of wrong methods for handling missing values can lead to bias in parameter estimates (Jones, 1996), bias in standard errors and test statistics (Glasser, 1964), and unproductive use of the data (Afifi & Elashoff, 1966). There are many reasons why data may be missing from a complete dataset, for instance, unable to find certain characteristics (Van Hulse & Khoshgoftaar, 2008). The most common and simple technique of handling missingness in a dataset is to overlook either the projects or the features with missing observations". This method causes the loss of important information and leads to imprecise cost estimation models.

Missing values may have an influence on a research external validity and limit its generalization of particular population. Hence, it is significant to search and find out how to deal and solve problems with missing values. According

to Cohen et al. (2003), when researchers use conventionally proper approaches for dealing with missingness in datasets, different methodologies may result to different conclusions.

Gad & Ahmed (2006), "claimed that disregarding the missing values in this case leads to biased conclusions. Furthermore, when an attribute has a missing value in a test case, it may or may not be meaningful to take the extra effort in order to achieve a value for that attribute(s). There are many approaches of solving the problem of missing values.

Missing values are common across the social sciences (Juster & Smith, 1998), and family studies is no exclusion. King, Hopnaker, Joseph, & Scheve (2001) found that about 50% of the members in political research data have missing values, and family research frequently approximates this level of missing values. However, several data sets employed in articles or journals have major issues with missing data. This is not false even for large datasets use by the public such as the National Survey of Families and Households, the National Longitudinal Survey of Youth, the General Social Survey, the Panel Study of Income Dynamics, and the Survey of Income and Program Participants". There are many classes of research by which missing data is a problem. This thesis focuses on survey analysis, but missing data are also an issue for experimental designs and administrative data as well.

To solve missing values properly, it is important to recognize the types and how data is missing from the dataset. The most reason for missing data is non-response, which according to Umbach (2005), can stem from a lot of reasons. Example, errors can occur when coding or entering data, respondents unable to provide answer to research questions, and the constraint of research design to elicit responses (Umbach, 2005).

Kim & Yates (2003) simulated seven common missing value techniques but did not find any superior technique after his research. Feelders (1999) concluded that, imputation methods work better than surrogate split when he researched on the performance of surrogate split and imputation. Batista & Monard (2003) after comparing the four (4) separate missing values techniques, concluded that, ten (10) close neighbor imputation performed well than other techniques "in most situations.

In the background of cost sensitive classification trees, Zhang, Qin, Ling, & Sheng (2005) researched four different missing data methods based on their performances on five data sets with artificially generated random missing values. They also summarized that the internal node technique performed well than the other three techniques scrutinized. Fujikawa & Ho (2002) compared numerous imputation methods based on preliminary clustering algorithms to probabilistic split on simulations based on some real data sets and found comparable performance. A weakness of all of the above studies is that they focused only on the restrictive MCAR situation.

Twala & Cartwright (2005) explore the ensembles of imputation methods in order to increase effort prediction accuracy and classifier learning effectiveness. The issue addressed in the study is the influence of missing value to the prediction accuracy. In 2010, the same study under the same topic available addressing the concern of good quality data is required to improve prediction accuracy (Cartwright & Twala, 2010).

Twala et al (2006) examine the randomization of decision tree structure algorithms to enhance forecasting accuracy. The major purposes are examining the influence of lost of data on forecast accuracy and how to ensemble missing

values methods can be employed to enhance the effort of prediction accuracy. Moreover, the outcomes fail to work when the data set is not large with various characteristics and produces different performance as the proportion of missingness increased.

Twala (2009) used computer simulations based on real data sets to compare the properties of different missing value methods, comprising using complete cases, single imputation of missing values, likelihood-based multiple imputation (where missing values are imputed several times, and the results of fitting trees to the different generated data sets are combined), probabilistic split, and surrogate split. He researched MAR, MCAR, and NMAR missing values and also generating procedures, although dependence of missing values on the dependent variable was not scrutinized. Multiple imputation and probabilistic split performed well, the were no much difference found between the techniques when the proportion of missingness was low. As expected, MCAR missing value mechanism did not caused more problems for methods, while NMAR missing values caused more problems. Also concluded by Kalousis & Hilario (2000), missing values extent over many predictors is very serious if it is focused on only one.

Twala, Jones, & Hand (2008) recommended technique similar in creating a different method for missing data, and came out that its performance was competitive with that of likelihood-based multiple imputation.

Green et al.(2001) (among others) recognized two alternatives ways to handle missing values: data imputation, where values are assessed to fill in missing values, and marginalization, where missing values are overlooked. Moreover, values imputed cannot be taken as consistent as the main observed data. Troyanskaya et al.(2001) suggested this when researching on various imputation

techniques for biological data: "However, it is important to exercise caution when drawing critical biological conclusions from data that is partially imputed. Estimated data should be flagged where possible to avoid drawing unwarranted conclusions.

Schrapler's (2001) impact focuses on the longitudinal development of item nonresponse for gross earnings and dealings of individual concerns. Comparable to (Lillard et al., 1986) and tom (Biewen, 2001) he discovers evidence that those in low social positions (also females and the young) tend to withhold income information. Again, just as in the data of Sousa-Poza & Henneberger (2000), respondents appear to be much more uncooperative in front of females than males. Schrapler summarizes that the friendship between respondent and interviewer is very importance, as with improving trust the item-nonresponse rate falls off over a time period".

2.2 Missing Data Mechanisms

The missing data mechanism defines the association between the missing values of the data and the values of the variables in the data matrix, i.e. whether the missing values depend on the underlying values of the variables in the data set. Gelman & Hill (2007) posit several reasons data may be missing. There are various assumptions concerning missing data mechanisms:

2.2.1 Missing Completely at Random (MCAR)

The probability of dropout is independent of the observed data and the missing data. That is $f(R_i/Y_i, X_i)=f(R_i)$. A typical example is that a subject moved to a different location where the treatment can't be continued. MCAR happens when any data of a variable have the same likelihood of being missing. It also be said when there is randomly missing of data values in the dataset and no reason

why a particular value is missing. The existences of respondents not responding to sampling surveys are common, and in Gelman & Hill's research assumed MCAR as the mechanism of non-response (Rubin, 1987). MCAR missing value shows that, no exact evidence could be gotten from the other responses of how the missing value should be. MCAR is simplest but most restrictive dropout mechanism. Under MCAR mechanism, the observed data value can be treated as a random sample of all data. If a data set is MCAR, there is no impact on bias and most standard approaches of analysis are valid (e.g. complete case analysis, generalized estimating equation, etc). Little's test (Little, 1988) can be used to test whether missingness is MCAR or not. This test compares the distribution of observed variables between dropouts and completers. MCAR probably caused by skipping a question or missing a trial. If the null hypothesis is not rejected, then we can clearly say that the assumptions of MCAR are fulfilled. Unfortunately, MCAR is often not plausible in longitudinal studies, especially in clinical trials.

2.2.2 Missing at Random (MAR)

The likelihood of dropout is only dependent on the observed data but not dependent on missing data. That is $f(R_i/Y_i, X_i)$=$f(R_i/Y_{i(obs)}, X_i)$. Where the observed dependent response vector is $Y_{i(obs)}$ and the observed covariate vector is X_i. DeSouza (2009) proposed that MAR can also be known as outcome-dependent MAR and/or covariate-dependent MAR. Some investigators think that dropout due to lack of efficacy will cause MAR. For instance, the assumption MAR can be achieved if the likelihood of missing data on income depended on the age of persons, but within age interval the likelihood of missing income was not related to income. MAR can be assumed to be semi-MCAR. The different between MAR and MCAR is that variable can be predicted from other available data in the datasets with MAR. When data are MAR, omitting cases with missing data is accepted because doing so will reduce the bias of the conclusions. MAR assumption is more accurate than MCAR in most study. The primary method

for clinical trial is usually based on this assumption. Under MAR, observed cases are no longer a random sample of full data, thus some standard approaches of analysis (e.g. complete case analysis, GEE) can not provide unbiased and efficient results for this situation any more. Mixed effects model for repeated measurements (MMRM), weighted generalized estimating equation (WGEE) and multiple imputation (MI) are three available and popular methods for MAR.

2.2.3 Missing not at Random (MNAR)

The probability of dropouts is dependent on the unobserved data and also the observed data or missing values do depend on unobserved values. That is $f(R_i/Y_i, X_i)$=$f(R_i/Y_{i(mis)}, X_{i(obs)})$. Under MNAR, the dropout procedure is also dependent on the missing values given observed measures. Example: to achieve the assumption of MNAR individual in high income class are less likely to report their income. Another instance of this is from medical research when any particular treatment of disease causes discomfort to the patient, the probability of that patient walking out or dropping out will be high (Rubin, 1987).

2.3 Tests of the Missing Data Mechanism

To test for MCAR is to compare the means of the observed values between a pair of missing and non-missing groups using pairwise t-tests, however this can result in multiple comparison problems. To avoid issues with Type I error, little (1988) proposed a global chi-square statistic that uses all of the data available. Park & Davis (1993) extended Little's test for longitudinal missing data using a Wald test, an improvement over prior studies that used a weighted least squares estimation. They explored a wide variety of methods such as weighted GEE's, non-parametric estimation of conditional scores, and modeling conditional distributions. Chen & Little (1999) extended Little's test to generalized estimating equations (GEEs) which forgoes distributional

assumptions. This method reduced bias when data was MAR but increased variance when data was MCAR.

Jamshidian & Schott (2007) also tested the equality of means and covariance in structural equation models but allowed for the partition of cases that belong to more than one pattern of missingness. Not dividing groups by missing patterns means that the researcher needs to manually define the groupings a priori. Kim & Bentler (2002) extended little's test to structural equation models by assessing homogeneity of means and covariances (HMC) using generalized least squares estimation. The rejection of HMC means rejection of MCAR but not vice-versa. In structural equation modeling, a single mean and covariance structure is modeled, but if HMC is not true, then a single set of mean and covariance parameters will not be able to represent the population mean and covariance. Jamshidian & Jalal (2010) imputed the missing data for each group and then applied F-tests based on the Hawkins test as well as on a non-parametric test of homoscedasticity. The authors' method involves imputing the missing data for each group and then applying a complete data method to the imputed data.

2.4 Ignorable Mechanism

Rubin (1987) classified missing values in a dataset as ignorable or not. Ignorable missing involves the combined term of missing data mechanisms MCAR and MAR, and also the parameters of the data model and for the missing data mechanism must be different. This means without respect to the missing data distribution, data can be analyzed. The likelihood of dropout and the actual missing response are related (which is unobserved because it is missing) or to other unobserved quantities, the missing data due to dropout is known as MNAR or non-ignorable (Schafer, 1997). The opposite situation of non-ignorable missing is considered to be missing not at random (MNAR), and must be handled with care.

The ignorable mechanisms can be satisfied with two conditions. These include, first the data are MAR. Secondly, the parameters that direct the process of missingness are unrelated to the parameters to be estimated. According to Rubin (1987), ignorable means that the variable can be overlooked or deleted in the model building. In the situation of MAR, the ignorable missing data mechanism occurs when variables are less significant or partially depended to the model than other variables. This assumption has similar mechanism as the causal framework, in which overlooking something can be done if a lot of clear evidence is established and also gathering more information. So, in these situations, less variables which are correlated can be ignored. For instance, when we forecast someone's athletic performance. The variable of most suitable color would most likely be unrelated to the prediction model; hence, ignoring this variable will likely have less negative outputs on the prediction of model's accuracy.

2.5 Patterns of Missingness

The "missing data pattern describes which values in the data matrix that are actually missing, and can help in the choice of method for handling the missing data. Missing data patterns are usually divided into monotone missing pattern (MMP) and arbitrary missing patterns (AMP). Monotone missing data often happen because of attrition in longitudinal studies, where dropping out is define as a means that all the observations will be missing in a study. A particular situation of MMP is the univariate missing data pattern (UMP). This is where only a variable in the data set suffer from missing observations, (see Table 2.1). An AMP also occurs when the data matrix cannot be ordered as in MMP, (see Table 2.1). One example of AMP is item non response in surveys where respondents for some reason have failed to answer one or more questions, but missing values in one variable does not necessarily implies that all following variables are missing. (Little & Rubin, 2002)

Sijtsma & Van der Ark (2003) discussed a test statistics for researching to know whether or not the pattern of missing values in a respondent by item data matrix is random. Since this is an asymptotic test, they researched whether it was important in small but realistic sample sizes. They also discussed two known imputation techniques, person mean and two-way imputation, and they suggested two new imputation techniques, response-function and mean response-function imputation. These techniques are based on few assumptions about the structure of the data. An empirical data example with simulated missing values revealed that the new technique "Response-Function" was superior to the techniques "Person Mean", "Two-Way", and "Mean Response-Function" in recovering from incomplete data several statistical properties of the original complete data. The Two-Way and Response-Function methods are useful both when item score missingness is ignorable and nonignorable".

Table 2.1: Patterns of Missingness

Missing Monotone				Missing Arbitrarily			
K1	K2	K3	K4	K1	K2	K3	K4
√	√	√	√	√	√	NA	√
√	√	√	√	NA	√	√	NA
√	√	√	√	√	NA	√	NA
√	√	√	√	√	√	NA	NA
√	√	√	NA	√	√	NA	√
√	√	NA	NA	√	√	√	NA
√	NA	NA	NA	NA	√	√	√

NA	Missing Values
√	Not missing

Assumptions and patterns of missing values helped to decide the methods that can be used to deal with missing data.

2.6 Methods for Handling Missing Data

There are so many methods in handling missing values. Especially, various techniques have been suggested and developed to handle missing values in longitudinal clinical trials. For example, Musil et al (2002) "investigated with simulations to compare the complete case method, mean imputation method, regression method, and EM algorithm method, and concluded that regression method produced good estimates while mean imputation method was the least efficient method. In contrast, Engel & Diehr (2003) concluded that the last observation carried forward (LOCF) method was the most effective method out of 14 imputation methods. Also, Tufis (2008) conducted the imputation comparison among mean imputation method, EM algorithm method, and multiple imputation (MI) method and concluded that MI method was the most efficient method to estimate missing values. Zhou et al (2001) compared three imputation methods, that is multiple imputation method, complete case method and mean imputation method. They concluded better standard deviation estimates for MI method than mean imputation method. Cheung (2007) came out that complete case technique outperformed in most of his experimental settings compared with EM algorithm method and MI method. There is no consensus on which method is uniformly effective than the other methods and research is still going on to study and search for the best imputation methods in different settings for developing a guideline to determine suitable methods for handling missing values. This thesis will also serves as addition to the research.

However, in real trials with missing data there are few methods that are actually used. This thesis will find the best imputation method for handling missingness in longitudinal data. The most frequently used techniques for dealing with missing values problems are described as follows:

2.6.1 Listwise Deletion

The most widely used technique of dealing with missingness in a dataset is listwise deletion, it is also known as complete-case analysis (Schafer & Graham, 2002). The computer program automatically deletes any case that has missing data for any bivariate or multivariate analysis when listwise deletion method is used. Although each variable may be missing only a small percentage of responses, collectively as cases are deleted, large portion of the data may not be used. This reduction in sample size interprets into reduced statistical power and brings into question how representative the remaining sample is of the population being studied.

This systematic loss of data with listwise deletion, increased the risk of bias if there is any pattern to the missing data, a risk that is lessened only when the data are MCAR (Pigott, 2001; Schafer & Graham, 2002). Some investigators have characterized listwise deletion as the least desirable data imputation method because of these biases and have warned against its use (Graham et al.,2001).

2.6.2 Pairwise Deletion

Pairwise deletion called available case analysis is a popular alternative to listwise deletion in linear models. In Pairwise deletion, all data available are used to estimate the parameters of the model. When a researcher looks at univariate descriptive statistics of a dataset with missingness, he or she is using available case analysis, he examines the means and variances of the variables observed throughout the data set. The Pairwise deletion is a process that focuses on the variance-covariance matrix. Each element of that matrix is calculated from all data available for that element. To estimate the parameters of interest in the data, different sets of cases are used. Kim & Curry (1977) concluded that

estimates can be enhanced by using all the available cases not the complete cases, others (Anderson, Basilevsky, & Hum, 1983; Haitovsky, 1968; Little, 1992; Little & Rubin, 1987) have summarized various issues with the process.

Data are MCAR, when the remaining observations are representative of the actual identified data set, Little (1992) pointed out that case available analysis gives consistent estimates (the correct point estimates, see e.g., Cox & Hinckley, 1974) when there is correlation in variables in regression models. For highly correlated variables, available case analysis gives estimates that are inferior to complete case results as illustrated in simulations by Haitovsky (1968).

2.6.3 Mean Substitution

The method of mean substitution imputes the missing values using the mean of the available observed values. This method has the potential of introducing biases as well as underestimating variability (Carpenter & Kenward, 2004).

Graham & Hofer (2000) described mean substitution as archaic. To use this technique, we replace the total sample mean of a variable for all of the missingness in that variable. For instance, if the mean age of the individuals in the study is 67.3 years, then 67.3 is used to substitute all missing values for age for any case in the dataset. Mean substitution is a fast and simple way to recover cases. The estimate of the mean for the variable is not affected by using the sample mean. However, this imputation technique relied on an MCAR assumption, which was discussed earlier in this chapter. Furthermore, the standard deviation and variance of the estimate is lowered, contributing in biased and deflated standard errors (McDonald, Thurston, & Nelson, 2000; Pigott, 2001; Streiner, 2002). The various changes of this technique share the similar advantages and disadvantages of substitution of the sample mean.

2.6.4 Hotdecking

This method identifies a subject with complete information in the dataset or data with similar on identified correlated attribute to a person with incomplete information and uses that person's information to replace the missing value. This method works well when the sample is large so that a similar characteristic is easily found and also when the variable used to sort the data is highly predictive of the variable with the missingness (Streiner, 2002).

In hotdeck imputation method, the standard deviation of the variable with the inserted values better estimates the standard deviation value for the variable without the substituted values. However, standard deviations are still likely to be reduced overall (Streiner, 2002). When using hotdecking to replace missing values, bias is likely to occur in regression equations than when calculating measures of central tendency. Another drawback of hotdecking is its difficulty to implement; programming requires a great deal of time and labor.

Durrant (2005) reviewed and discussed the advantages and disadvantages of several imputation techniques for handling item-nonresponse in the social sciences and found that, certain forms of fractional and multiple hot deck methods perform well with regards to bias and efficiency of a point estimator and robustness against model misspecifications but standard parametric imputation methods were not found adequate.

2.6.5 Last Observation Carried Forward (LOCF)

The LOCF technique substitutes every missing data with its corresponding last observed value and is the simplest imputation method. The technique usually used the assumptions of MCAR in longitudinal research of continuous outcomes. There is a notion that the outcome fails to change after the last observed value.

Hence, there is no time effect since the last observed data. LOCF is a common technique that is mostly used in solving missing values issues. This is because, it is simple to use and can be done easily as well. Moreover, the sample size does not change compare to the complete case method.

2.6.6 Multiple Imputation

Multiple Imputation (MI) is generally an appropriate method for dealing with incomplete data. When it was first suggested, it was mainly thought of as a way to handle nonresponse in a complex survey context (Rubin, 1987). Multiple imputation did not take much long to become accepted and useful in other areas as well (Rubin, 1996). The use of MI has spread and includes in various statistical areas like clinical trials (Wood et al., 2004), epidemiology (Klebanoff & Cole, 2008) and longitudinal research (Spratt et al., 2010). According to Rubin (1987), there is no much benefit in producing and analyzing only a few imputed datasets if the rate of missing values is not very high. Multiple imputation inference includes three distinct phases: imputation, analysis and pooling. These steps are illustrated as:

Imputation: Impute (fill in) the missing entries of the incomplete data sets, not once, but m times ($m = 3$). Imputed values are drawn for a distribution (that can be different for each missing entry). This step results is m complete data sets.

Analysis: Analyze each of the m completed data sets. This step results in m analyses.

Pooling: Integrate the m analysis results into a final result. Simple rules exist for combining the m analyses. The results from the m complete data sets are combined for the inference. The pooling step consists of computing the mean over the m repeated analysis, its variance, and its confidence interval or P value. In general, these computation are relatively simple".

2.6.7 Maximum Likelihood

The method of maximum likelihood is a routine procedure for obtaining estimators for unknown parameters from a set of data x_1,x_2,...x_n. The maximum likelihood method is a common method for estimating missing value (Little & Rubin, 1987; Schafer & Olsen, 1998; Schafer,1997) and is centered on an accurate statistical model of the data. When the maximum likelihood technique is used for the task of imputing the missing data, the commonly used model is the multivariate, Gaussian mixture model. Likelihood techniques may be categorised into single imputations and multiple imputations (Schafer, 1997; Little and Rubin, 1987). Maximum likelihood imputation method of missing value can be viewed as a method to maximize the likelihood: $L(\theta/Y_{obs})= \int f(Y_{obs}, Y_{mis}/\theta)Y_{mis}$, where Y_{obs} and Y_{mis} represent the observed data and the missing data respectively and θ here is some control parameter of interest (Little & Rubin, 1987). Most of the ML methods, involve calculating the matrix of second derivatives of the log likelihood, which become very complex in the presence of missing data (Little & Rubin, 1987). One method that does not need these second derivatives to be estimated when a dataset is characterised by incomplete data, is the Expectation Maximisation (EM) algorithm.

2.6.8 Expectation Maximisation

The expectation maximisation algorithm was introduced by Dempster et al.(1977) and was intended at solving issues of complexity, related with maximum likelihood techniques. "Expectation maximisation combines statistical methodology with algorithmic implementation and has gained much attention recently in many missing data problems. Some researches have proved that expectation maximization works better than methods such as listwise, pairwise data deletion, and mean substitution because it assumes that incomplete cases have data missing at random rather than missing completely at random (Allison,

2002; Rubin, 1978).

This technique is one of several maximum likelihood (ML) methods. In all ML approaches, observed data are used to estimate parameters, which are then used to estimate the missing scores. These ML strategies have proven superiority to deletion, nonstochastic imputation, and stochastic regression imputation procedures (Roth, 1994) for multivariate normal distributions. Ghahramani & Jordan (1994) offered an improved EM algorithm that can process missing values in datasets. The maximum likelihood model parameters are estimated concurrently by this method, data cluster assignments, and missing structures for the values. Each of these techniques suffers from an inability to discount imputed value because of their lack of full reliability.

Dempster, Laird, & Rubin (1977) suggested the use of an iterative solution, known as the EM algorithm, to find a parameter of the estimate (such as the means and covariance matrix) when closed form solutions to the maximization of a likelihood are not possible. Little & Rubin (1987) and Schafer (1997) offer the theory of the EM algorithm for missing data analysis assuming multivariate normal data.

2.7 Measures of Performance for Imputation Methods

The Mean squared error (MSE), the root mean squared error (RMSE) and the coefficient of determination will be used as criteria to assess the performance of the best imputation methods.

2.7.1 Mean squared error (MSE)

Mean square error (MSE) is the mean of the squared differences. It is the average squared difference between the estimated parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the corresponding true parameters (β_0 and β_1) derived from the original data set. It shows how the estimator is close to the true value. MSE is also equal to the sum of variance and the squared bias of the estimated parameters. The MSE provides a useful measure of the overall accuracy, as it incorporates both measures of bias and variability. The more effective imputation technique will have lower MSE.

2.7.2 The Root Mean squared error (RMSE)

The Root Mean squared error (RMSE) is defined as the square root of the MSE. The RMSE is a valuable measure of total precision or accuracy and can help to know how each imputation technique is performing. In general, the more efficient method would have a lower RMSE (Huang & Carriere, 2006).

2.7.3 Coefficient of Determination (R^2)

The coefficient of determination is used to measures how well a model explains and predicts future outcomes. The coefficient of determination in statistical analysis, also known as R-squared, is used as a guideline to assess the accuracy of the model. It is the degree of variability in factor. It also measures how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain. It is relied on mainly in trend analysis and is represented as a value between zero and one. The closer the value is to one, the better the fit, or relationship, between the two factors.

2.8 Generalized Estimating Equations

We explored the Generalized Estimating Equations (GEE) to analysed the full dataset and also each imputation methods in order to know the best imputation methods.

The Generalized Estimating Equations (GEEs) first introduced by Liang & Zeger (1986), represent an extension to the generalized linear model to accommodate correlated variables. The GEEs are widely used for estimating the covariate coefficients of marginal models for repeated binary responses. The GEE is an estimation method that estimates a common scale parameter and a working correlation matrix of the result variables, treating them as nuisance parameters. GEE can also be used for both continuous and discrete results, but is mostly used for discrete outcomes and even then, most real-life applications are correlated binary outcomes (Diggle, Liang, & Zeger 1994).

Liang & Zeger (1986) and Prentice (1988) have introduced moment based GEEs, which only required a specification of the form of the first two moments (the probabilities of success and their correlations) of the vector of binary responses Fitzmaurice, & Lipsitz, (1995). The GEE models are built on probability of an event of interest and correlations as espoused by Liang and Zeger (1986) & Prentice (1988). In specifying only the marginal distribution of the outcome variable, GEE will produce estimates of population parameters only (modeling of population mean only). Hence, GEE cannot be used in settings where subject specific estimation and hypothesis testing are required".

Specifically, when regression is the main focus, Liang & Zeger (1986) suggested an extension of generalized linear models to the study of longitudinal data. Their method employed a generalized linear model for the marginal distribution of A_{it},

where A_{it}(t=1,2,...,n_i). They did not specify any form of joint distribution for the repeated measures. However, they presented estimating equations that resulted in consistent estimates of the regression parameters and as well as their variances. They modeled the marginal distribution, but did not consider the conditional distribution which takes into account previous observations. The methods they proposed reduced to maximum likelihood when the A_{it} are multivariate Gaussian.

Some have argued that binary response models that condition on all or some of binary responses variables are useful for studying only particular types of dependencies. To estimate marginal response probabilities, the conditioning is not needed. In cases where the analysis involves paired data, parametric methods to these data appear not to be complicated, otherwise they are complicated. Hence, Prentice (1988) advocated a generalized estimating equation approach for inference on response probabilities and correlations.

If the observed variation in the data is not reliable with the theoretical difference determined by the model assumed, one main technique is to present additional dispersion parameters to give explanation for the additional changes in the dataset. In this situation, we can still come out with parameter estimates by solving the resulting "score" equations, but the corresponding "likelihood" is no longer a true likelihood but rather a quasi-likelihood, and these "score" equations are called generalized estimating equations (GEEs). For clustered or longitudinal dataset, Liang & Zeger (1986) "suggested using a working correlation matrix in the GEEs. That is, the same assumptions is kept by GEEs about the mean and covariance structures as in quasi-likelihood methods but introduce the working covariance matrix which may depend on few nuisance parameters to simplify the correlation structure.

Zeger & Liang (1986) offered a methodology for discrete and continuous

longitudinal data that employs the quasi-likelihood technique due to the scarcity of multivariate distributions for non-normal distributed data. They again stated that there should be a known function of the marginal expectation of the response variable that displays a linear function of the explanatory variables. They also assumed that the variance is a known function of the mean. They specify a working correlation matrix for observations for each subject. The above procedure leads to generalized estimating equations which produces consistent estimators of the regression coefficients and of their variances".

For example, with binary outcomes it is assummed that the logit of the probability of response k depends linearly on the covariates, the variance is just $k(k-1)$ outcome distribution results in simple methods for regression analyses of Gaussian, Gamma, Poisson, binomial and ordinal data (McCullagh and Nelder, 1983, as cited in Zeger & Liang, 1986).

2.8.1 Marginal Odd Ratios

Instead of using marginal correlations in modeling the association between a pair of binary responses, Lipsitz et al. (1991), Liang et al. (1992) and Carey et al. (1993) suggested the use of a marginal odds ratio. To them, the odds ratio is a natural metric for measuring association and may also be easier to interpret, when dealing with binary responses. Cessie & Houwelingen, (1994) recommended the use of different measures of dependence in modeling for logistic regression for correlated binary data.

Fitzmaurice & Lipsitz (1995) also proposed a serial pattern models for the odd ratio. An attractive feature of these models is that they allow the repeated responses for each individual to be unevenly spaced in time and the times of measurement to vary between individuals. Also, the number of observations per individual may vary.

Molenberghs & Lesaffre (1999) suggested a marginal model of multivariate categorical data. However, these marginal models may not provide proper estimation of parameters because of lack of efficient specification of the dependence of binary outcomes in the model. Again, Azzalini (1994) proposed a marginal model based on binary Markov Chain for a single stationary process $(Y_1 - Y_T)$ where Y's take values of 0 or 1 at subsequent times.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter deals with a description of the methods employed for this study. At this stage, the study describes the source of data and the research design. The researcher explores the methodological framework in the Generalized Estimating Equations (GEEs). It further describes how to test missing data mechanism (MCAR and MAR). The chapter continues by discussing the assumptions of MCAR and MAR. It will also discuss the classifications of missing data under the assumptions of various missing mechanism. Aside these, the chapter touches on the mean square error (MSE), root mean squared error (RMSE) and the coefficient of determination as performance assessment procedures to compare each imputation method and know the best. It also entails an extensive chronology of the data analysis procedure.

3.2 Data Description

This study illustrated the application to real life data by using data from the National Income Dynamics Study (NIDS) from south Africa. In South Africa, the first national panel research conducted was the NIDS. The Southern Africa Labour and Development Research Unit (SALDRU) in the School of Economics at the University of Cape Town is responsible in executing this survey. The research took a national sample of 28,000 respondents from closely 7,300 households across the entire country when it started in 2008. In every two years, the survey is repeated with these same household members and observes

the livelihoods of individuals and households over time period. The NIDS shows South Africa dynamic household structure, changes in people living situations and the well-being of members in the household in a way that no other study in South Africa has been able to do.

The main characteristic of the research is its ability to follow respondents as they relocate to different households. NIDS is a programme to compile comprehensive longitudinal information on respondents selected for the study and to find out who is moving ahead and who is falling behind. This data is also key for research and policy makers. The NIDS data constitute areas such as health, education, labour market and birth history. The 2008 data was compiled into the Wave 1 dataset. The second, Wave 2 dataset was compiled after the second visit was made to the same group of people between 2010 -2011. This study focused on the work status of persons selected for the research. The binary response variable measured therefore was whether an individual was employed or not at the time of visit. Specifically, (Employed=0, Not employed=1)
Covariates chosen for this exercise include:

- Gender of respondent (Male= 0, Female = 1)
- Education (Educated= 0, Not educated = 1)
- Age (18-30 years = 0, 31-57 years = 1) and
- Marital Status (Married=0, Not married=1)

The covariates chosen were tested to find out whether there is significant impact for conditional, marginal models and joint models.

3.3 Research Design

The research design adopted in this study include the following listed below:

- Per the above data in a long format, GEE model is fitted;

- Missingness are artificially created in the general data (ie 5%, 10%, 15%, 20%, and 30% missingness created);
- The little's test was employed to check whether a dataset with missing values is MCAR or MAR;
- Various imputation methods are adopted under MCAR and MAR mechanisms to replace the missing values created above;
- After each imputation methods has been used, a GEE model was fitted to re-estimate the coefficients and their standard errors;
- The models were compared based on their standard errors.

Below is a schematic representation of step by step procedure of the research design;

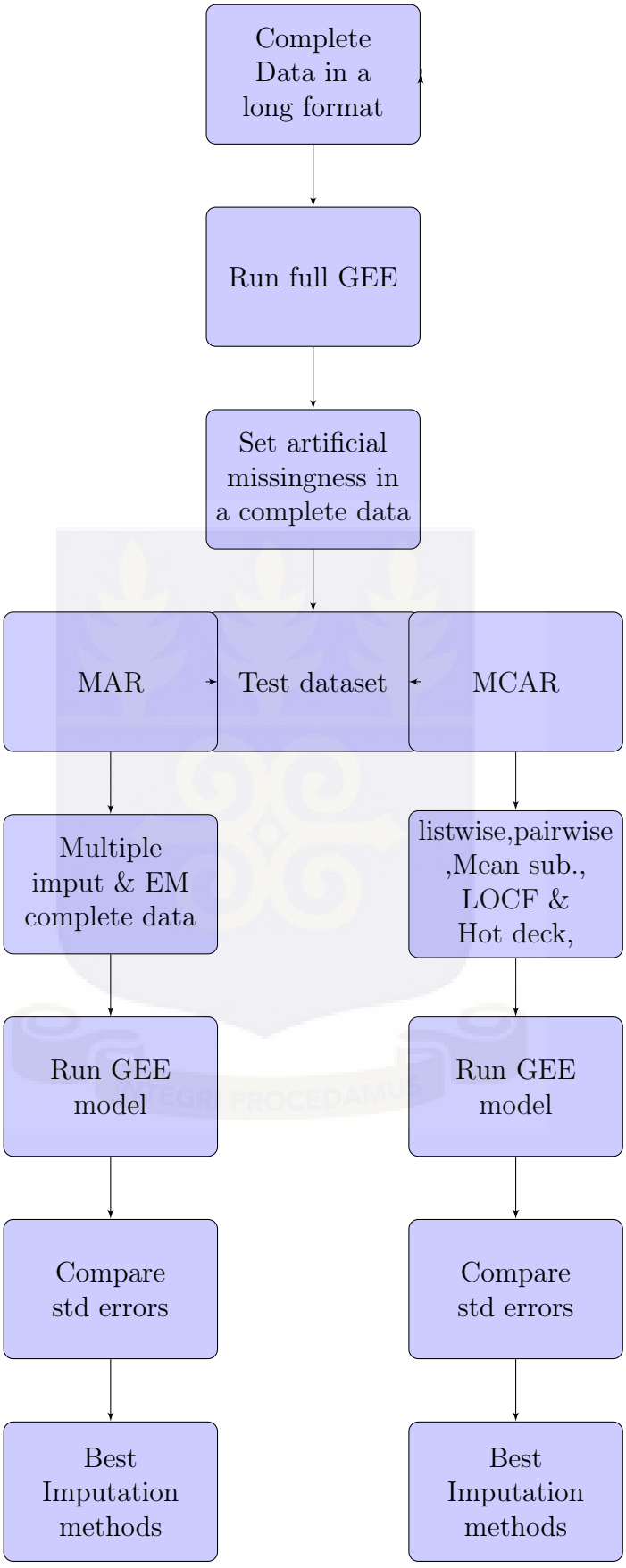

Figure 3.1: Step by step procedure of the research design

3.4 Models Used For Analysis

The "ordinary least squares (OLS) is a technique for estimating parameters which are unknown in a linear regression model, the major aim is to minimize the sum of squares of the differences between the observed responses in the given dataset and those predicted by a linear function of a set of explanatory variables. Ordinary least squares (OLS) regression is also a generalized linear modeling method that may be used to model a single response variable which has been recorded on at least an interval scale. The technique may be important to single or multiple explanatory variables and categorical explanatory variables that are well coded. Regressors are exogenous when OLS estimator is consistent, and optimal in the class of linear unbiased estimators when the residuals are homoscedastic and serially uncorrelated. The method of OLS gives minimum-variance mean-unbiased estimation when the errors have variances that are finite under these conditions. Under the assumption that the residuals are normally distributed, OLS is the maximum likelihood estimator.

The individual national income from South Africa (i.e.wave 1 and wave 2) was correlated, hence the ordinary least squares (OLS) regression assumptions was violated. This is because these correlations needed to be taken into consideration in modeling; otherwise the standard errors of the estimates would be underestimated for the between-subject and overestimated for the within-subject effects. To analyze discrete and correlated data, the generalized estimating equations (GEE) were recommended by Liang and Zeger (1986) as an extension of generalized linear models (GLM). Its strength is that it models a known function of the marginal expectation of the dependent variable as a linear function of explanatory variables. The advantage of GEE is obvious when the number of observations is large in relation to the number of waves within subjects as in the case of the dataset to be used in this study". Missing data

will, throughout the thesis, be assumed ignorable under GEE. This assumption is based on the fact that no one knows the reason why data are missing in the database that are considered in this work. Ignorable missing is a combined term that involves the missing data mechanisms MCAR and MAR.

3.5 Generalized Estimating Equations (GEE) Models

The GEE is a semi-parametric regression approach which uses moment-based inference, it was first introduced by Liang and Zeger (1986). It is an extension of generalized linear models that account for correlated responses. Instead of attempting to specify a model for the whole multivariate distribution of a data vector, GEE only models the first moment, specifically the mean response $E(Y_{it})$ at each visit t for the i^{th} subject. The part of the model that specifies the correlation is treated as a nuisance and not of scientific interest.

Let Y_{ij} ; $j = 1, ..., n_i$, $i = 1, ..., k$, k represents the j^{th} measurement on the i^{th} subject. There are n_i measurements on subject i and $\sum_{i=1}^{k} n_i$ total measurements. The link function is use to modeled correlated data and "linear predictor setup (systematic component) as the independence case. The random component is also known by the variance functions as in the independence case, but also the covariance structure of the correlated measurements must be modeled. We let the vector of measurements on the i^{th} subject be $Y = \mu_i = [Y_{i1}...Y_{ini}]'$ with corresponding vector of means $\mu_i = [\mu_{i1}...\mu_{ini}]'$ and also let V_i be an estimate of the covariance matrix of Y_{ij}. The Generalized Estimating Equation for estimating β is an extension of the independence estimating equation to correlated data and is given by:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (Y_i - \mu_i(\beta)) = 0 \tag{3.1}$$

A simple characteristic of GEE models is that the joint distribution of a subject's response vector y_i, does not need to be specified. Instead, it is only the marginal distribution of y_{ij} at each time point that needs to be specified. To further explain this, assuming that there are two time points and that we are dealing with a continuous normal outcome. GEE informs us to assume that the distribution of y_{i1} and y_{i2} are two univariate normal, rather than assuming that y_{i1} and y_{i2} form a joint bivariate normal distribution. Thus, the GEE avoids the need for multivariate distributions by only assuming a functional form for the marginal distribution at each time point.

The advantage of GEE models is that the structure of covariance (correlation matrix) is assumed as a nuisance. Under GEE, multivariate normality suffices univariate normality but the reverse is not always the same. The emphasis is seen on the regression of Y on X. However, the generalized estimating equation models give consistent and asymptotically normal solutions for the regression coefficients $\beta(s)$, even with misspecification of the structure of covariance in the longitudinal data. GEE models is an extension of GLMs for correlated data, the GEE specifications entail those of GLM with one addition. So, first, the linear predictor is given as:

$$\eta_{ij} = x_{ij}^{'}\beta \tag{3.2}$$

Where x_{ij} is the covariate vector for subject i at time j. We then consider the link function as:

$$g(\mu_{ij}) = \eta_{ij} \tag{3.3}$$

Mean response:

$$E(y_{ij}) = \mu_{ij} \tag{3.4}$$

As in GLMs, the general choices here are the identity, logit, and log link for continuous, binary, and count data, respectively". The variance is then described as a function of the mean, namely,

$$V(\mu_{ij}) = \varphi v(\mu_{ij}) \tag{3.5}$$

Where $v(\mu_{ij})$ is a known variance function and φ is a scale parameter that may be known or estimated.

3.5.1 The GEE Estimation (Working Correlations)

If A_i is an $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ as the j^{th} diagonal element, as specified above, we define $n_i \times n_i$ working correlation matrix (of the n_i repeated measures) for the i^{th} subject (i.e. Y_i) as $R(\alpha)$. Hence, the working variance-covariance matrix for Y_i will be:

$$V(\alpha) = \varphi A_i^{1/2} R_i(\alpha) A_i^{1/2} \tag{3.6}$$

For the case of outcomes that are normally distributed with homogeneous variance across time, is given as:

$$V(\alpha) = \varphi R_i(\alpha) \tag{3.7}$$

In the case of normal outcomes, Park (1993) improves this to heterogeneous variance across time by making the scale parameter φ_j to change across time period $(j = 1, ..., n)$. The GEE estimator of β is the solution of :

$$\sum_{i=1}^{N} D_i'[V(\hat{\alpha})]^{-1}(y_i - \mu_i) = 0 \tag{3.8}$$

Where $\hat{\alpha}$ is a consistent estimate of α and $D_i = (\frac{\partial \mu_i}{\partial \beta})$ and therefore equation (3.8) becomes:

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta} \right) (V(\hat{\alpha}))^{-1} [y_i - \mu_i] = 0 \tag{3.9}$$

This is an improvement on estimating equation for β in any GLM, which is given in (3.9). Therefore, the GEE solution can be seen as a natural generalization of the GLM solution for correlated data. As an example, in the normal case, for equation (3.9), that is:

$$\begin{gathered} U(\beta) = \textstyle\sum_{i=1}^{N} (\frac{\partial \mu_i}{\partial \beta})' (V(y_i))^{-1} [y_i - \mu_i] = 0 \\ \mu_i = X_i \beta \\ D_i = X_i \\ V(\alpha) = R_i(\hat{\alpha}) \end{gathered} \tag{3.10}$$

The solution for the parameter β (by making β a subject) gives;

$$\beta = [\sum_{i=1}^{N} X_i' [R_i(\hat{\alpha})]^{-1} X_i]^{-1} \sum_{i=1}^{N} X_i' [R_i(\hat{\alpha})]^{-1} y_i]^{-1}] \tag{3.11}$$

Equation (3.11) depends on the mean and variance of y and is called quasi-likelihood estimates. Working the GEE includes iterating between the quasi-likelihood solution for estimating β and a robust technique of finding α as a function of β. Basically, it includes:

- Given estimates of $R_i(\alpha)$ and φ, calculate estimates of β using iteratively reweighted least squares (IRLS).
- Given estimates of β, obtain estimates of $R_i(\alpha)$ and φ. For this, calculate Pearson (or Standardized) residuals

$$r_{ij} = \frac{(y_{ij} - \mu_{ij})}{\sqrt{[V(\hat{\alpha})_{ij}]}} \tag{3.12}$$

and use these residuals to consistently find α and φ. Liang and Zeger (1986) suggest the estimators for various different working correlation structures. Upon convergence, it is of interest to obtain standard errors associated with the estimated regression coefficients in order to perform hypothesis tests and construct confidence intervals. These standard errors are obtained as the square root of the diagonal elements of the matrix $V(\hat{\beta})$.

3.5.2 GEE for Binary Longitudinal Data

Longitudinal binary data often arise when repeated measurements (ie positive or negative to certain tests) are made on the same subject over time. Assuming the "random variable Y_{ij} signifies a sequence of binary measurements at time j, where $j = 1, ..., N$ for subject i, where $i = 1, ..., n$. The observed value y_{ij} is a realization of the binary response variable Y_{ij}, and we adopt independence across subjects. The focus of this study is on the marginal models that describe the binary outcome vector, given a set of predictor variables. Let π_{ij} represent the marginal probability of observing a 'success' for subject i at time j, that is $\pi_{ij} = E(Y_{ij}) = P(Y_{ij} = 1)$,

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - Y_{ij})}} \tag{3.13}$$

be Standardized deviation between the data and the model predictor for subject i at time j, and $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}) = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3})$,... be relations among responses. According to Bahadur (1961), the model can be represented by;

$$f(y_i) = c(y_i) \prod_{j=1}^{N} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \tag{3.14}$$

Where $y_i = (y_{11}, ..., y_{1N})$ is a vector of measurements for subject i, and

$$c(y_i) = 1 + \sum_{j_1<j_2} \rho_{ij_1j_2} e_{ij_1} e_{ij_2} + \sum_{j_1<j_2<j_3} \rho_{ij_1j_2j_3} e_{ij_1} e_{ij_2} e_{ij_3} + ... + \rho_{i1...N} e_{i1...e_N} \tag{3.15}$$

With $e_{ij} = \frac{Y_{ij}-\pi_{ij}}{\sqrt{\pi_{ij}(1-Y_{ij})}}$, the joint likelihood mass function is thus the product of individual mass functions and the correlation factor $c(y_i)$. GEE is important in marginal models, this is because by accepting working assumptions about the association structure, one only needs correctly specifying the univariate marginal distributions.

Again, for a binary response Y_{ij}, assume we have X_{ij} is a p - dimensional vector of complete covariates. Supposing the logit link function, the mean structure of the binary model can be expressed as;

$$logit = \{P(Y_{ij} = 1|X_{ij}, \beta)\} = X_{ij}^{'}\beta, \tag{3.16}$$

Where β is the vector of model parameters". The GEE model in estimate its optimum coefficient assumes the quasi likelihood estimator.

3.5.3 Quasi-likelihood Estimator

Quasi-likelihood is a methodology for regression that involves the specification of relationships between mean response and covariates and between mean response and variance. Thus it does not assume a probability distribution as in the case of full likelihood.

Assuming Y_i to be the response variable for each subject $i = 1, ..., N$ and X_i be $p \times 1$ vector of covariates. Let β be $(p \times 1)$ vector of regression parameters to be estimated. Define $\mu_i = E(Y_i|X_i)$ to be the conditional expectation of Y_i and a function of covariates and regression parameters, so that $\mu_i = h(X_i^{'}\beta)$. The inverse of h is the link function which relates the mean response to the linear predictor $X_i^{'}\beta$. For quasi-likelihood, variance of each Y_i, denoted as v_i is a known function of the expectation μ_i, so that $v_i = f(\mu_i).\Phi$. The scale parameter Φ is treated as a nuisance parameter. The quasi-likelihood estimator is the solution

to the equations:

$$S_k\beta = \sum_{i=1}^{N} \frac{\partial \mu_i}{\partial \beta} v_i^{-1}(Y_i - \mu_i) = 0 \tag{3.17}$$

where $k = 1, ..., p$.

Estimators of regression parameters, $\hat{\beta}$ are obtained by iteratively reweighted least squares method.

3.5.4 Marginal Models

A standard GEE is known as a marginal model. Marginal models extend generalized linear models to longitudinal data and are typically used when the inference is population-based, rather than individual-based. The term "marginal" means that in the model specification the expected value of the response variable Y, depends only on covariates (fixed effects) and does not depend on subject specific random effects nor directly on previous responses of the subject. Since the purpose is to describe the changes in population mean rather than changes within subjects, within-subject correlation is regarded as a nuisance characteristic. Regression parameters and within-subject correlation is modeled separately (Fitzmaurice et al.,2004).

Let assume we have N subjects who are measured repeatedly. Y_{ij} denotes the response variable for the i^{th} subject on the j^{th} measurement occasion. A realisation of each Y_{ij} is observed at time t_{ij}. The response variable can be continuous, binary, multinomial or a count. We assume the data are unbalanced (the number of repeated measurements can be different for subjects and /or they can be measured at different occasions) and that there are n_i repeated measurements for the i^{th} subject.

The response variable is a $n_i \times 1$ vector;

$$\mathbf{Y_i} = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{ini} \end{pmatrix}, i = 1, ..., N. \tag{3.18}$$

Y_i are assumed to be independent, but observations within the subject are not assumed to be independent. Associated with each response at a given time point j, there is a $p \times 1$ vector of covariates;

$$\mathbf{X_{ij}} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, i = 1, ..., N; j = 1, ..., n_i \tag{3.19}$$

Which can be collected into a $n_i \times$ p matrix of covariates

$$\mathbf{X_i} = \begin{pmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{ini} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \dots & X_{i1p} \\ X_{i21} & X_{i22} & \dots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{ini1} & X_{ini2} & \dots & X_{inip} \end{pmatrix}, i = 1, ..., N \tag{3.20}$$

They can be either time-invariant or time-dependent. Time-invariant variable is fixed within a subject at the same value irrespective of time point j, whereas time-dependent variable is varying with time for each subject.

3.5.5 Fitting Generalized Estimating Equations

We estimate generalized estimating equations using the Fisher scoring technique. Define $x_{it} = (1, x_i, t)$ as the covariate matrix where we find the probability of our variable of interest using the logit link function;

$$P(y_{it}) = \mu_{it} = \frac{exp(x_{it}^{'}\beta)}{1 + exp(x_{it}^{'}\beta)} \tag{3.21}$$

We employ generalized estimating equations to find $\hat{\beta}$ that best fits the above. $u_{it} = E(y_{it}) = p_{it}$ where $(p_{it}) = \beta_0 + \beta_1 x_1 + \beta_2(t-1) = x_{it}^{'}\beta$. We define the expected information matrix as ;

$$I(\beta) = \sum_{i=1}^{k} D_i^{'} V_i^{-1} D_i \tag{3.22}$$

3.6 Testing the Missing Data Mechanism

In practice, investigators are often confronted with analyzing data sets that have missing values. To appropriately analyze such data sets involves a proper idea of the missing data mechanism. If data are missing completely at random (MCAR), then many missing value analysis methods lead to valid conclusion. Thus, tests of MCAR are needed. A brief overview of the literature showed that much of the work on missing data tests derive from Little's (1988) concept of testing equality of means or covariance across subgroups.

3.6.1 Describing Little's Test of MCAR

Kim & Bentler (2002) extended little's test to structural equation models by assessing homogeneity of means and covariances (HMC) using generalized least squares estimation. Since Little's test is the foundation for all homogeneity of means and covariance (HMC) MCAR tests, a brief exposition to Little's method is warranted. We begin with an index of notation to be used throughout this

thesis and in subsequent methods:

- i is the particular subject number
- j is an index of the variable
- y_{ij} is a particular subject's response for a particular variable
- m_{ij} is the missing response for a particular subject
- k is an index of the $K = 2^p$ missing patterns
- r_k is a vector of the particular missing response pattern
- r_{ij} is the single response within the k^{th} pattern for the j^{th} variable
- p is the number of variables
- p_k is the number of observed variables for missing pattern k
- N is the total number of observations

The procedure for Little's test of MCAR can be summarized using the following steps:

- The Expectation-Maximization (EM) algorithm takes into account missing data to generate a maximum likelihood estimate. Given a dataset $Y = [y_{ij}]$ that contains the set of observed and missing variables, enter the $Y : N \times p$ matrix into the EM algorithm. The EM maximum likelihood "estimate for the population mean vector μ and variance-covariance matrix Σ is defined as $\tilde{\mu}$ and $\tilde{\Sigma}$ respectively.

  To calculate the chi-square statistic, the expected mean vectors and variance covariance matrices have to be partitioned for every response pattern to contain only the observed values. Define $\tilde{\mu}_{obs.k} = \tilde{\mu}^T D_k$ where D_k is $p \times pk$ selection matrix with only one per column that selects observations with no missing values. This results in a vector of pk means for every k^{th} response pattern". Similarly, define $\tilde{\Sigma}_{obs.k} = D_k^T \tilde{\Sigma} D_k$ as the $(pk \times pk)$ variance-covariance matrix for every k^{th} response pattern.

- Group cases according to the missing pattern to get observed means for each group. Create a set of matrices S_k for $k = 1, ..., k$ where for each matrix is a subset of Y having all the cases that are found with particular pattern of missingness. Define N^{rk}, which is the number of cases that belong to a k^{th} missing response pattern.

- Take the difference of each vector of observed means estimated in Step 2 from the overall EM-estimated means estimated in Step 1 weighted by the EM-estimated variance-covariance matrix to obtain the goodness of fit statistic Little's test of MCAR:

  Little's test of MCAR =

$$\sum_{k=1}^{N} N^{rk}(\bar{y}_{obs.k} - \tilde{\mu}_{obs.k})^T \tilde{\Sigma}^{-1}_{obs.k}(\bar{y}_{obs.k} - \tilde{\mu}_{obs.k}) \tag{3.23}$$

  Where N^{rk} is the number of observed samples for the k^{th} missing response pattern, and the chi-square statistic has degrees of freedom $\sum_{k=1}^{N} pk - p$ where pk is the number of observed variables for all K patterns.

When the Little's test of MCAR has p value that is larger than the value of alpha given, then neither the hypothesis of normality nor the hypothesis of MCAR is rejected. If little's test of MCAR is less than alpha value of 0.05 we reject the null hypothesis and conclude that the data are not missing completely at random, hence missing at random (MAR). Data are MCAR when the pattern of missing values does not depend on the data values.

3.7 Classifications of Missing Data Under the Assumptions of Various Missing Mechanism

This study adopts the little's test MCAR to check whether a dataset with missing values is MCAR or MAR. Little's test of MCAR provides tests for the MCAR and MAR assumption. If we failed to reject the null hypothesis under

the little test of MCAR, then we can conclude that imputation methods such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) rely "on the assumption that the pattern of missingness does not depend on the data values in the dataset. This is also called missing completely at random, or MCAR. Therefore, all the techniques for estimation give consistent and unbiased estimates of the correlations and covariances when the data are MCAR. Violation of the MCAR assumption can lead to biased estimates produced by the methods of handling missing data.

When the p value is less than the alpha value, we reject the null hypothesis under the little test of MCAR and say that imputation methods such as Multiple imputation and Expectation maximization rely on the assumption that the pattern of missingness is associated to the observed data only in the dataset and this condition is known as missing at random, or MAR. The assumption permits estimates to be well adjusted using available information. For instance, in a research of income and education, the subjects that have low education may have more missing income values. In this situation, the data are MAR, not MCAR. On the other hand, for MAR, the likelihood that income is recorded depends on the subject's educational level. The probability may vary by education but not by income within that level of education. If the probability that income is recorded also varies by the value of income within each level of education (for example, people with high incomes don't report them), then the data are neither MCAR nor MAR. This is not a common situation, and if it applies, none of the methods is proper.

3.8 Methods of Handling Missing Data Under the MCAR Assumptions

3.8.1 Listwise Deletion

Listwise deletion (also called case deletion) is a common and simple solution to missing values. It is in the default in most standard statistical packages. If one fails to meet the assumption of MCAR, then listwise deletion is conservative because the smaller sample size will increase the standard errors and reduce the level of significance. Therefore, in situation where the assumptions of MCAR are met, the conservatism simple means increasing the risk of a Type II error. Such reduced power may not be as serious with a large sample. If the data fails to meet the assumption of MCAR, listwise deletion may caused biased estimates. Table 3.1 below shows the approaches of listwise deletion when values are missing in longitudinal data:

Table 3.1: Dataset with missing values

Rows	JK	J1	J2	J3	J4
A	54	NA	5	67	9
B	76	26	94	65	77
C	73	67	45	NA	53
D	88	8	88	33	47

NA: missing value (Not Available).

Since row A and row C have missing values, we delete row A and row C. We maintain Row B and raw D for the final analysis because they have no missing values. Table 3.2 below shows the data set after listwise deletion.

Table 3.2: Dataset after listwise deletion

Rows	JK	J1	J2	J3	J4
B	76	26	94	65	77
D	88	8	88	33	47

3.8.2 Pairwise Deletion

Pairwise deletion tries to use all existing data by discarding cases on an analysis-by-analysis basis. This technique is frequently described in the context of a covariance (or correlation) matrix, whereby each variance and covariance term is computed by using all cases with complete data on a given variable or variable pair. For instance, the covariance between income and depression might be relied on 50% of the individuals who gave answer for both of the items, but the covariance between number of children and age of oldest child might be based on 99% of the participants who gave answered for both of those items.

The Matrices below illustrate a simple two-variable data matrix with only one variable subject to nonresponse. In the pairwise deletion, all the available cases would be used to find the mean of y_1, but only the cases that are complete would contribute to an estimate of y_2, and the correlation between y_1 and y_2, Different sets of cases are used to estimate parameters of interest in the data. Matrices Illustration of missing data restricted to one variable.

$$\begin{pmatrix} y_{11} & y_{21} \\ y_{12} & y_{22} \\ \vdots & \vdots \\ y_{1m} & y_{2m} \end{pmatrix} \quad m \textit{ Complete cases}$$

$$\begin{pmatrix} y_{1(m+1)} & - \\ \vdots & \vdots \\ \vdots & \vdots \\ y_{1n} & - \end{pmatrix}$$

$n - m$ Cases with observations on y_1

Pairwise Deletion Estimates:

$$\bar{y}_1 = \sum_{i=1}^{n} y_{1i}$$

$$\bar{y}_2 = \sum_{i=1}^{m} y_{2i}$$

$$S_1^2 = \frac{\sum_{i=1}^{n}(y_{1i}-\bar{y}_1)^2}{n-1}$$

$$S_2^2 = \frac{\sum_{i=1}^{m}(y_{2i}-\bar{y}_2)^2}{m-1}$$

$$r_{xy}^2 = \frac{1}{m-1}\frac{\sum_{i=1}^{m}(y_{1i} - \bar{y}_{i(m)})(y_{2i} - \bar{y}_{2i} - \bar{y}_2)}{S_{1(m)}S_2} \tag{3.24}$$

Where $\bar{y}_{1(m)}$ and $S_{1(m)}$ are the mean and standard deviation of y_1 calculated from m complete cases.

3.8.3 Mean Substitution

In this method, missing values are imputed with the mean value of that variable on the basis of the non-missing values for that variable. This method assumes that data are MCAR and results in biased means when this assumption is false. Again, imputing the mean value into cases may lower the variance of the variable, which also attenuates covariance that the variable has with other variables. This technique may contributes to biased of the means with data that are MAR" and underestimates variance and covariance in all cases. Table 3.3 and table 3.4 show examples of the mean substitution.

Table 3.3: Example of the mean substitution

Rows	JK	J1	J2	J3	J4
A	54	{**NA**}	5	67	9
B	76	26	94	65	{**NA**}
C	73	67	45	{**NA**}	53
D	88	8	88	33	47
Means	**73**	**34**	**58**	**55**	**37**

NA: missing value (Not Available).

Table 3.4: Complete dataset after using mean substitution

Rows	JK	J1	J2	J3	J4
A	54	{**34**}	5	67	9
B	76	26	94	65	{**37**}
C	73	67	45	{**55**}	53
D	88	8	88	33	47

3.8.4 Hot Deck Imputation

Hot deck imputation is a technique where missing data are replaced with data that comes from other records of the same data. In concept, the missing data are substituted by data obtained from the data with similar cases (Schafer, 1997). The hot deck imputation has two major steps, as outlined below:

- Records are first subdivided into classes. This can be done using a number of different skills such as clustering and nearest neighbour techniques.
- Incomplete records are filled with values that fall within the same class.

Example of Hot Deck Imputation:

Table 3.5 shows data set with missing values and how Hot deck imputation method was used to replace missing values that comes from similar cases.

Table 3.5: Handling of missing values using Hot Deck Imputation

Dataset with missing values		Dataset after Hot Deck Imputation	
Weight(JK)	**Height(J2)**	**Weight(JK)**	**Height(J2)**
260	(**66**)	260	66
190	68	190	68
NA	(**66**)	(**260**)	66
215	72	215	72
145	(**62**)	145	62
NA	(**62**)	(**145**)	62

NA: missing value (Not Available).

3.8.5 Last Observation Carried Forward (LOCF)

The last observation carried forward (LOCF) also known as last value carried forward (LVCF) the last observation a participant gave is entered into the empty cells that follow (and hopefully the degrees of freedom are adjusted accordingly). Similar strategies involve "replacing missing observations with the participant's mean over the trials on which data are present, or basing imputed values on trends from past trials. All these approaches carry with them assumptions about what the data would have looked like if the participant had not dropped out, and none of them is to be recommended.

3.9 Methods of Handling Missing Data Under the MAR Assumptions

3.9.1 Multiple Imputation

Multiple imputation (MI) is generally a proper method for dealing with incomplete data. According to Molenberghs et al. (2007), multiple imputation replaces each missing value with two or more possible values. Furthermore, it introduces random variation which enhance the possibility to have unbiased estimates of all parameters. Multiple imputation is a step ahead of estimates of maximum likelihood (ML). Multiple imputation offers multiple data sets for analysis. Multiple imputation is that an average of the completed-data likelihood over unknown missing values can be used to estimate observed-data likelihood. To explain further, both analyses are approximately equal and variation across the different datasets signifies the imputation uncertainty (He, 2010).
Multiple imputation (MI) involves four different stages, these include:

- Check the data and determine pattern: The first step is to look at the data and based on the information existing determine the pattern of missingness.

The pattern of missing data can be monotonic or non-monotonic. Knowing the pattern of missing data, the second step can be determined.

- Imputation: The second step is imputing multiple values for each missing value to create M complete datasets. This step selects the value randomly to fill the missing value from the predictive distribution of missing data given observed data.

- Analysis: The third step is to do analysis for each completed dataset separately to generate M sets of estimates. Every set of estimates may slightly differ from each other.

- Pooling: The fourth and the last step is to combine all sets of estimates to generate a total estimate and calculate the variation among parameter estimates.

In summary, we conclude the procedure for analysis as:

Impute (Creates multiple possible datasets) $\Longrightarrow$ Analyze (Run analysis on each dataset) $\Longrightarrow$ Pool (Find average of estimates)

The overall knowledge behind MI is really simple and straightforward. Instead of imputing one value for each missing observation, $m \geq 2$ plausible values are imputed creating m complete data sets. The distinction between the imputed data sets reflects the uncertainty caused by imputations. Let K, W, B and T be the quantity of interest, within imputation variance, between imputation variance and total variance associated with K, respectively. Again, $\hat{K}_d$, and $\hat{W}_d$, $d = 1, 2, 3...m$ are the estimated quantity and the variance that quantity in the dth imputed data set. The combined point estimate of the parameter of interest, $\bar{K}$, is the average of the m estimates from the different completed data sets:

$$\bar{K} = \frac{1}{m} \sum_{d=1}^{m} \hat{K}_d \tag{3.25}$$

The variability associated with K consists of two parts. The first part is the within imputation variance, that is the average of the calculated variances associated with K in each imputed data set:

$$\bar{W} = \frac{1}{m}\sum_{d=1}^{m}\hat{W}_d \tag{3.26}$$

The second part is the between imputation variance, which is the variance of the m estimates of K ,

$$\bar{B} = \frac{1}{m-1}\sum_{d=1}^{m}(\hat{K}_d - \bar{K})^2 \tag{3.27}$$

The sum of variance is then a modified sum of the within and between imputation variances,

$$\bar{T} = \bar{W} + \left(1 + \frac{1}{m}\right) B, \tag{3.28}$$

Where $(1 + \frac{1}{m})$ is a correction for the finite number of imputed data sets. The test statistics and confidence intervals are based on a t- distribution with

$$v = (m-1)\left(1 + \frac{1}{m+1}\frac{\bar{W}}{B}\right)^2 \tag{3.29}$$

degrees of freedom. Confidence intervals associated with K are calculated as:

$$\hat{K}_m \pm t_v(\alpha/2)\sqrt{T} \tag{3.30}$$

And the test statistic for testing hypothesis concerning K is

$$\frac{\hat{K} - K}{\sqrt{T}} \sim t_v \tag{3.31}$$

Although the knowledge behind MI relies on a bayesian foundation, Rubin (1987) showed that under some general conditions, MI will also yield valid conclusions

from a frequentist perspective. The conditions which will result in proper imputation are the following:

- As m becomes large, $(\bar{K} - K)/\sqrt{T}$ should approximately follow a standard normal distribution, $N(0, 1)$, with Y and R regarded as fixed. Where $Y = (Y^{obs}, Y^{mis})$ be the complete data matrix consisting of the observed and unobserved values and R is a missing data indicator matrix.

- As m becomes large, $\bar{W}$ should be a consistent estimate of W with random Y and R fixed.

- The between-imputation variance, $\bar{B}$, be stable over the repeated samples and lower than that of K.

3.9.2 Expectation Maximization (EM)

Expectation maximization (EM) provides the means and covariance matrix estimates, which are to be used to find consistent estimates of the parameters of interest. It is based on an expectation step and a maximization step, which are repeated many times until maximum likelihood estimates are obtained. In EM process, large sample size are needed and that the data are missing at random (MAR). It is a principled method for handling missing data. Dempster, Laird, & Rubin (1977) in their seminal paper on this method coined the term EM.
According to Little & Rubin (1987), the EM algorithm formalizes a relatively old ad hoc idea for handling missing data:

- replace missing values by estimated values,

- estimate parameters,

- re-estimate the missing values assuming the new parameter estimates are correct,

- re-estimate parameters, and so forth, until iterating convergence.

Expectation maximisation has also been proven to work better than methods such as listwise, pairwise data deletion, and mean substitution because it assumes that incomplete cases have data missing at random rather than missing completely at random (Allison, 2002; Rubin, 1978). The distribution of the complete data, Y, can be represented as follows:

$$f(Y/\theta) = f(Y_{obs}, Y_{mis}/\theta) = f(Y_{obs}/\theta)f(Y_{mis}Y_{obs}/\theta) \tag{3.32}$$

Where $f(Y_{obs}, Y_{mis}/\theta)$ is the density of the observed data and $f(Y_{mis}Y_{obs})$ is the density of the missing data, given the observed data (Little & Rubin, 1987). The log likelihood of equation (3.36) below is written as:

$$l(\theta/Y) = l(\theta/Y_{obs}, Y_{mis}) = l(\theta/Y_{obs}) + ln(f(Y_{mis}Y_{obs}/\theta) \tag{3.33}$$

So that the objective is to optimize $l(\theta/Y_{obs}, Y_{mis})$, using the control parameter, θ. Now, we will let the current estimate of the parameter θ be denoted by $\theta^{(t)}$ Optimisation of equation (3.33) is an iterative process, of two steps, namely, the E-step and the M-step.

- The E-step (Expectation step) determines the expected log likelihood of the data, as if the parameter θ was truly the current estimate, $\theta^{(t)}$, as:

$$Q(\theta/\theta^{(t)}) = \int l(\theta/Y)f(Y_{mis}/Y_{obs}, \theta) = \theta^{(t)}dY_{mis} \tag{3.34}$$

- The M-step (Maximisation step) finds $\theta^{(t+1)}$ by maximising equation (3.33) as:

$$Q(\theta^{(t+1)}/\theta^{(t)}) \geqslant Q(\theta/\theta^{(t)}), \qquad \forall\theta \tag{3.35}$$

The EM algorithm is a general method capable of fitting models to incomplete data and capitalizes on the relationship between missing data and the known

parameters of a data model. If the parameters of the data model were known, then it would be possible to get unbiased predictions for the missing values. This interdependence between model parameters and missing values suggests an iterative method where, firstly, the missing values are predicted based on assumed values for the parameters and these predictions are used to update the parameter estimates, and the process is repeated until convergence. The sequence of parameters converges to maximum-likelihood estimates that implicitly average over the distribution of the missing values.

- Maximum Likelihood

The maximum likelihood technique handles missing data as random variables to be integrated out of a likelihood function instead of deleting or filling in incomplete cases. In other hand, the maximum likelihood method is to maximize the observed likelihood, that is the marginal density for the observed data. However, when Newton's method is used to obtain the MLE of the observed likelihood, two problems arise:

- It can be difficult in computing the second order of partial derivatives of the likelihood;
- The likelihood does not necessarily increase for every iteration.

Maximum likelihood imputation of missing data can be viewed as a method to maximise the likelihood which is;

$$L(\theta/Y_{obs}) = \int f(Y_{obs}, Y_{mis}/\theta) dY_{mis} \tag{3.36}$$

Where Y_{obs} and Y_{mis} represent the observed data and the missing data respectively and θ here is some control parameter of interest (Little & Rubin, 1987). One method that does not need these second derivatives to be calculated when a dataset is characterised by incomplete data, is the Expectation Maximisation (EM) algorithm.

3.10 Performance Assessment Procedures to Compare Various Imputation Methods

3.10.1 Mean Square Error (MSE)

The mean squared error (MSE) of an estimator measures the average of the squares of the errors. The MSE is a risk function, which correspond to the expected value of the squared error loss or quadratic loss.

Mean square error (MSE) is also the mean of the squared differences. It the average squared difference between the estimated parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the corresponding true parameters (β_0 and β_1) obtained from the original data set. It also shows how close the estimator is to the true value. The MSE provides a useful measure of the overall accuracy, as it incorporates both measures of bias and variability. If $\hat{Y}$ is a vector of n predictions, and Y is the vector of observed values corresponding to the inputs to the function which generated the predictions, then the MSE of the predictor can be estimated by: $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$. This is an easily computable quantity for a particular sample. The MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$.

The MSE can be written as the sum of the variance of the estimator and the squared bias of the estimator, providing a useful way to calculate the MSE and implying that in the case of unbiased estimators, the MSE and variance are equivalent: $MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$.

3.10.2 The Root Mean Squared Error (RMSE)

RMSE is the square root of the MSE. The RMSE is the important measure of the total precision or accuracy and can be used to evaluate the performance

of each imputation technique. The RMSE measures the differences between values predicted by a model or an estimator and the values actually observed. The RMSE represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. In general, the more effective method would have a lower RMSE (Huang & Carriere, 2006).

The RMSE of an estimator $\hat{\theta}$ with respect to an estimated parameter θ is the square root of the mean square error:

$$RMSE(\hat{\theta}) = \sqrt{MSE(\hat{\theta})} = \sqrt{E[(\hat{\theta} - \theta)^2]}. \tag{3.37}$$

For an unbiased estimator, the RMSE is the square root of the variance, known as the standard deviation. The RMSE of predicted values $\hat{y}_t$ for times t of a regression's dependent variable y_t is computed for n different predictions as the square root of the mean of the squares of the deviations:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{3.38}$$

3.10.3 Coefficient of Determination (R^2)

The coefficient of determination also known as the R-square statistic (R^2) in a regression model, measures the proportion of variability in the response that is explained by the regressor variables. In a linear regression model with intercept, R^2 is defined as:

$$R^2 = 1 - \frac{SSE}{SST} \tag{3.39}$$

where SSE is the residual sum of squares and SST is the total sum of squares corrected for the mean. R^2 statistics also play a significant indirect role in regression calculations. For example, the proportion of variability explained by regressing all other variables in a model on a particular regressor can give insights into the interrelationship among the regressors". The higher the value of R^2, the greater the proportion of the variability explained by fitting the data to the model. This implies that the regression model is a good fit.

3.11 Data Analysis Procedure

In longitudinal data, we employ several (repeated) observations of many subjects, over different periods of time. The repeated observations tend to be correlated from the same subject. One way to show this correlation is through dynamic patterns. Data from the National Income Dynamics Study (NIDS) is a programme to compile comprehensive longitudinal information on persons selected for the study and to find out who is moving ahead and who is falling behind in terms of their income level. This study only consisted of "Two Waves". Throughout the thesis, Missing data will be assumed ignorable under Generalized Estimating Equations (GEE). By applying the GEE methods and with the aid of R and IBM SPSS 20 statistical software, we will be able to compute and display meaningful information which will give the grounds to interpret the empirical results. Pactitti (1998) cited in Amaratunga et al. (2002), outlines a quantitative data analysis plan that follows the following pattern:

- Raw data assessment;
- Data entry;
- Data processing;
- Communicating findings;
- Data interpretation and;

- Completing data analysis

In this study the steps outlined by Pactitti (1998) served as the guideline for our data analysis. First, the raw data was assessed for any discrepancies with the help of Microsoft Excel. After this process, we transformed the data from wide to long format using IBM SPSS 20 statistical package. However, for us to process the data and use the various methods of handling missing data into meaningful statistical results, the study employs the R and IBM SPSS 20 statistical package. At the fourth stage of the data analysis procedure, we will present and simplify the results in the form of tables and through graphical representations for easy exposition. Finally, the empirical results will be interpreted with reference to the tables and the figures in congruence with literature.

CHAPTER 4

RESULTS OF DATA ANALYSIS

4.1 Introduction

This chapter of the study covers the presentation of empirical results and its statistical interpretation. The chapter start by giving the descriptive statistics of the National Income Dynamics Study (NIDS) data from South Africa (Wave one and two), marginal model-GEE, missing data mechanism test, comparison of imputation methods under MCAR and MAR mechanism, comparison of methods for handling missing values under GEE model, performance assessment of imputation methods using the coefficient of determination, and finally comparison of imputation methods using the root mean square error (RMSE). All analysis was done using the R and IBM SPSS 20.

4.2 Descriptive Statistics

Table 4.1 shows the descriptive statistics of data from National Income Dynamics Study (NIDS) from South Africa (WAVE 1 and 2) and its selected variables.

Table 4.1: Summary of Descriptive Statistics

			Freq	Percent	Variance	Mean	Std.Dev
WAVE	Valid	1	3382	50.0	0.250	1.50	0.500
		2	3382	50.0			
		Total	6764	100.0			
Variables	**Codes**	**Description**					
Work status	0	working	2462	36.4	0.232	0.64	0.481
	1	not working	4302	63.6			
Gender	0	male	2299	34.0	0.224	0.66	0.474
	1	female	4465	66.0			
Marital status	0	married	3264	48.3	0.250	0.52	0.500
	1	not married	3500	51.7			
Age	0	18 to 30	3788	56.0	0.246	0.44	0.496
	1	31 to 57	2976	44.0			
Education	0	educated	464	6.9	0.064	0.93	0.253
	1	not educated	6300	93.1			

The descriptive statistics as evidenced in table 4.1 reveals 6764 samples selected for this study. It is evident that 36.4% of persons between the ages of 18 and 57 interviewed in wave one and two are employed whereas 63.6% are unemployed. It also shows that out of 6764 respondents interviewed, 2299 are male and 4465 are female representing 34.0% and 66.0% respectively. The descriptive analysis shows that 48.3% interviewed are married while 51.7% of the respondents are not

married. In terms of education, 464 respondents representing 6.9% are educated while 6300 respondents representing 93.1% are uneducated.

4.3 Marginal Model-GEE

Using the marginal model, from table 4.2 below, four covariates such as Gender, Marital status, Age and Educational status were significant, meaning that they contribute significantly to the state of employment status. The fitted marginal model is

$$logit\left(\hat{p}_1\right) = log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) = -0.041 + 0.124X_1 - 0.146X_2 + 0.081X_3 + 0.602X_4 \tag{4.1}$$

where X_1= Gender, X_2= Marital status, X_3 = Age, X_4 = Educational status, $\hat{p}_1$ represents the estimated probability of recording an "employed" response at the second measurement and $1-\hat{p}_1$ represents the estimated probability of not recording an "employed" response at the second measurement. The estimated intercept is -0.041 representing the estimated logit when gen = 0, mari= 0, age= 0, and edu=0. This means that the respondent had no age group, no gender, no marital status and no educational status which is impossible in this particular study. The estimated coefficient for the variable age is 0.081 meaning that for respondents who are in age range 18-30 versus those in the age bracket 31-57, the expected change in the log odds is 0.081, given that education, marital status and gender stays constant.

Based on the odds ratio scale, the odds for an individual in the 18-30 year group to be employed is $exp(-0.041 + 0.124 - 0.146 + 0.602) = 1.185$ and the odds for an individual in the 31-57 year age group is $exp(-0.041 + 0.124 - 0.146 + 0.081 + 0.602) = 1.859$. Therefore the odds of 1.19 means that individuals in the 18-30 year group are 1.19 times more likely

to be employed at the second visit than those in age group 31-57. In terms of probabilities, the probability of an individual in the 18-30 year group to be employed at the time of the second visit will be $\frac{1.185}{(1+1.185)} = 0.54$ and that of an individual in the age group 31-57 will be 0.46.

All four covariates were significant at a 5% level of significance indicating that if the same sample were run for 100 times 95 of them will have all four estimated coefficients being significant. The interpretation for education status, marital status and gender of the respondents were left undone, since that is not the main objective for this study.

Table 4.2: Fitted model using data from NIDS: GEE Model

Estimate	β values	Odds Ratio	Std.err	Wald	Pr(>|W|)	
(Intercept)	-0.04090	0.95993	0.10032	0.166	0.68353	
Gen	0.12411	1.13214	0.05402	5.278	0.0216	*
Mari	-0.14627	0.86392	0.05056	8.369	0.00382	**
Age	0.08058	1.08392	0.05123	2.474	0.03574	*
Edu	0.60229	1.82630	0.09171	43.132	5.12E-11	***

Note: * Significant at 5%, ** Significant at 1%, *** Significant at 0.1%

4.4 Missing Data Mechanism Test

To properly analyze missing data sets requires the knowledge of how the data is missing (i.e in a random way or non-random way). This will help us to classify missing data under the assumptions of various missing mechanism. In this study, percentage of missing values such as 5%, 10%, 15%, 20% and 30% were artificially created in a random and non-random way using complete life data (NIDS data) from south Africa. In order to use the required imputation method to handle each percentage of missingness, little's test of MCAR was employed. Table 4.3 shows the output of little's MCAR test on the percentages of missing values artificially created.

Table 4.3 shows that the significant values for all the percentages of missingness are greater than the alpha values of 0.05, hence we fail to reject the null hypothesis.

Table 4.3: Output of little's MCAR test: under MCAR

Percentages of missing values created					
MCAR test	5%	10%	15%	20%	30%
Chi-sq	53.16	51.31	53.43	62.03	54.32
D.f	48	58	61	65	69
Sig	0.282	0.721	0.744	0.582	0.902

We conclude that the data values in the dataset was randomly missing and there is no reason why a specific value in the dataset is missing. This condition is known as missing completely at random (MCAR). This means that various imputation methods to handle these missingness depend on the assumption that the pattern of missing values does not depend on the data values. In addition, all methods for estimation give consistent and unbiased estimates of the correlations and covariances when the data are MCAR.

From Table 4.4, the significant values for all the percentages of missingness are less than alpha value of 0.05, we then reject the null hypothesis and conclude that the data are not missing completely at random, but missing at random (MAR).

Table 4.4: Output of little's MCAR test: under MAR

Percentages of missing values created					
MCAR test	5%	10%	15%	20%	30%
Chi-sq	72.24	84.08	120.11	99.34	99.56
D.f	53	58	64	66	71
Sig	0.041	0.031	0.000	0.005	0.015

This shows that the probability of dropout is only dependent on the observed

data but not dependent on missing data. In MAR, imputation methods to handle the percentage of missing values depend on the assumption that the pattern of missingness is related to only the observed data. Through this assumption, estimates can be adjusted using available information.

The little test MCAR has the following decision rule:

- P value ≥ 0.05, we fail to reject the H_0. Conclude that missing data mechanism is MCAR.
- P value < 0.05, we reject the H_0. Conclude that missing data mechanism is MAR.

After the little MCAR test, the various imputation methods stated in table 4.5 below will be used to replace missing values artificially created in the complete dataset. This will help us to compare and know the best imputation techniques under a specific missing data mechanism.

Table 4.5: Imputation Methods for Handling Missing Values

MCAR	MAR
Listwise deletion	Multiple imputation
Pairwise deletion	Expectation maximization
Mean substitution	
Hotdecking	
Last observation carried forward (LOCF)	

4.5 Comparison of Imputation Methods for Handling Missing Values Under GEE Model

In order to compare various imputation methods and know the best, we compare each imputation method used to handle percentage of missingness to the general

GEE model for the complete dataset, which is;

$$logit\left(\hat{p}_1\right) = log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) = -0.041 + 0.124X_1 - 0.146X_2 + 0.081X_3 + 0.602X_4. \tag{4.2}$$

The procedures for the comparison are stated below:

- We compare the model for each method used to handle the percentage of missing values to the general GEE model. This is done by finding the coefficient difference for each imputation methods used to handle percentage of missing value.
- Coefficient difference is calculated by subtracting each coefficient from the coefficients of the general model.
- We compute the average of the coefficients difference for each imputation method.
- We sum the average coefficients difference for all the percentage of missingness for each imputation methods.
- To compare, we select the best imputation methods by picking the smallest average number.

Table 4.6 shows how we compared 5% missingness under listwise deletion to the general GEE model and finally computing the average coefficients difference.

Table 4.6: 5% missingness under listwise deletion to the general GEE model

Estimate	Complete Data	Std.err	Estimate for listwise	Std.err	Difference
(Intercept)	-0.04090	0.10032	-0.008168	0.101281	
Gen	0.12411	0.05402	0.114559	0.055021	0.009551
Mari	-0.14627	0.05056	-0.153269	0.051736	0.007000
Age	0.08058	0.05123	0.070577	0.052134	0.010003
Edu	0.60229	0.09171	0.582699	0.092548	0.019591
				Average	**0.011540**

4.5.1 Comparison of Methods for Handling Missing Values Under MCAR Mechanism

Missing data can frequently occur in a longitudinal data analysis. In a real-world data analysis, the missing data can be MCAR, MAR, or MNAR depending on the reasons that lead to data missing. In this thesis, method for handling missing data such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) under MCAR mechanism were compared. To evaluate the performance of these five imputation methods, we first use the total average coefficient difference for each imputation method and adjudge the smallest values as the best method. Table 4.7 below shows how each methods for handling missing values under MCAR mechanism performed.

From Table 4.7, the best imputation method under MCAR is Last observation carried forward (LOCF), which recorded the minimum average coefficient of 0.265092.

Table 4.7: Performance of methods for handling missing values under MCAR mechanism

Percentages	**Listwise**	**Pairwise**	**MeanSubs**	**HotDeck**	**LOCF**
	Averages	Averages	Averages	Averages	Averages
5	0.01154	0.009105	0.00726	0.01526	0.01182
10	0.03046	0.009101	0.009105	0.01741	0.01036
15	0.0298	0.008105	0.01001	0.02009	0.01461
20	0.11236	0.007705	0.01316	0.02497	0.01697
30	0.238305	0.23831	0.238192	0.23692	0.211332
Overall	**0.422465**	**0.272328**	**0.277727**	**0.31465**	**0.265092**

Among the five imputation methods compared under MCAR mechanism using the average coefficient, listwise deletion is the poorest method. Mean substitution and pairwise deletion performed well when small percentage of missing values occurred in a dataset. In concluding, when small percentages such as 5% and 10% of missingness occurred in the dataset under MCAR mechanism, it is advisable

to use the mean substitution or pairwise imputation methods to replace missing values in the dataset. Again, when large percentage of missingness occurred in a dataset under MCAR mechanism, the Last observation carried forward (LOCF) method will give consistent and unbiased estimates, hence it will be prudent to go for LOCF imputation method when the percentage of missingness is large under MCAR mechanism.

4.5.2 Comparison of Methods for Handling Missing Values Under MAR Mechanism

Missing at random (MAR) occurred when missingness is conditional on observed factors and is independent of the unobserved data. It also occurs when the probability of any variable instance to be missing is the same for all units. In this study, we replace missing values under MAR mechanism using the multiple imputation and expectation maximization. We again use the total average coefficient difference for each imputation method and adjudge the smallest values as the best method. Table 4.8 below shows how each methods for handling missing values under MAR mechanism performed.

Table 4.8: Performance of methods for handling missing values under MAR mechanism

Percentages	**EM**	**Multiple Imp**
	Averages	Averages
5	0.00998	0.0056275
10	0.02697	0.0084925
15	0.02681	0.0158875
20	0.038761	0.046273
30	0.057474	0.0173875
Overall	**0.159995**	**0.093668**

Table 4.8 indicates the performance of multiple imputation and expectation maximization methods for handling missing values under the assumptions of MAR mechanism using the average coefficient difference. From the table, multiple

imputation outperformed expectation maximization method by 5%, 10%, 15% and 20% missingness in the dataset. This means that, when you have a large or small values of data missing at random (MAR), it is important to use multiple imputation to replace missing values in the dataset. Furthermore, multiple imputation gives unbiased inference and accurate conclusion when used to replace missing values under MAR mechanism. We conclude by saying that, multiple imputation do better than expectation maximization under MAR mechanism.

4.5.3 Comparison of Methods Under MCAR and MAR Mechanism

In order to know the best imputation method under both MCAR and MAR mechanism, the study compared the average coefficient difference for both MCAR and MAR mechanism. We observed that, multiple imputation is the best when the seven imputation methods were compared. Multiple imputation and expectation maximization performed well at common amounts of missing data. They are useful to reduce attrition bias. This observation is also supported by the bar graph in figure 4.1 below:

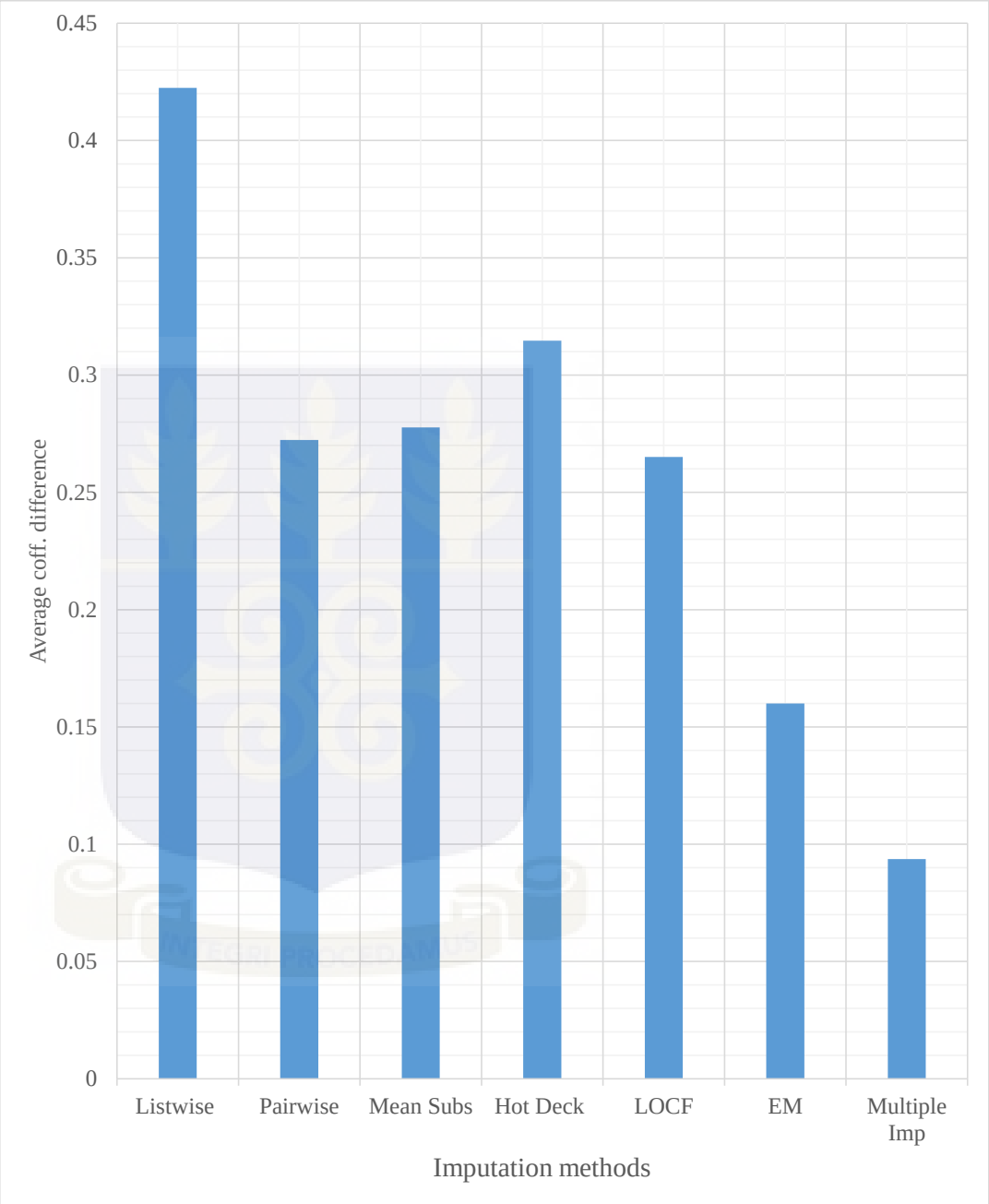

Figure 4.1: Comparison of methods under MCAR and MAR

4.6 Comparison of Imputation Methods Using the Coefficient of Determination (R^2)

The coefficient of determination measures the proportion of variability in the response that is explained by the regressor variables. It is also how well a model explains and predicts future outcomes. Coefficient of determination depends on mostly in trend analysis and is represented as a value between zero and one. The closer the value is to one, the better the fit, or relationship between the two factors.

The R-Squared for the complete dataset is 0.00835, meaning 0.84% of the total variation in employment status was explained by the regression model. This study seeks to compare imputation methods, hence individual percentage of missingness of the various imputation methods may be doing well if their coefficient of determination values are closer to the R Squared value of the complete dataset which is 0.84%. To also know the best imputation methods, average coefficients of determination of the various imputation methods will be compared and we select the best. The higher the average coefficient of determination, the better the method for handling missingness in the data. Table 4.9 shows how each imputation methods for handling missing values under MCAR mechanism performed using the coefficient of determination (R^2).

Table 4.9: Performance of imputation methods using the R-Squared under MCAR

Percentages	Listwise	Pairwise	MeanSubs	HotDeck	LOCF
5	0.007943	0.98354	0.008136	0.008092	0.007684
10	0.006582	0.95453	0.007798	0.007766	0.008189
15	0.006534	0.87438	0.007681	0.007791	0.007586
20	0.006732	0.94231	0.007713	0.007373	0.008218
30	0.500048	0.87398	0.534737	0.531458	0.581234
Overall	**0.527839**	**4.62874**	**0.566065**	**0.56248**	**0.612911**
Averages	**0.105568**	**0.925748**	**0.113213**	**0.112496**	**0.122582**

From table 4.9, when small percentage of values (5% or 10%) are missing complete at random (MCAR) from a dataset, it will be prudent to use the mean imputation or the LOCF methods. This is because under 5% missing values using the mean imputation to replace missingness in the data, the total variation in employment status explained by the regression model is 0.81% which is closer to the coefficient of determination for the complete dataset value of 0.84%. On the other hand, when 10% of missing values were replaced by the LOCF imputation method, the coefficient of determination recorded was 0.82% also closer to the complete dataset value of 0.84%. Replacing large sets of missing values in a dataset, it is important to use the pairwise imputation method to replace missing values in the dataset under MCAR mechanism. From the above table, when 30% values were missing, the pairwise imputation method recorded R squared of 87%. Thus 87% of the total variation of employment status was explained by the regression model. In all, pairwise method performed well in replacing missing values under MCAR mechanism. To achieve proper statistical inference, it is advisable to use either the mean imputation or the LOCF to replace missing values when the lost in the dataset is small (5% or 10%) and the pairwise imputation when the missing values in the dataset is large. Listwise and hot deck imputation methods performed poor under MCAR mechanism. For this reason, they may lead to bias in parameter estimates and improper statistical interpretation of the analysis.

Table 4.10 shows how each imputation methods for handling missing values under MAR mechanism performed using the coefficient of determination.

Table 4.10: Performance of imputation methods using the R-Squared under MAR

Percentages	EM	Multiple Imp
5	0.008934	0.008771
10	0.009776	0.008192
15	0.009536	0.007559
20	0.01002	0.008074
30	0.01091	0.008102
Overall	**0.049176**	**0.040698**
Aerages	**0.0098352**	**0.0081396**

Table 4.10 shows that multiple imputation method outperformed expectation maximization when small percentage of values are missing in the dataset under MAR mechanism. This is because the R squared of 5% data missing at random (MAR) is 0.87% which is closer to the R squared of the complete dataset which is 0.84%. Expectation maximization performed well when values of 15% are missing in the dataset under MAR mechanism. Comparably, all the two imputation methods (ie EM and multiple imputation) performed well in all the percentage of missingness when data is missing at random (MAR). Hence when large or small values are missing in a dataset under MAR mechanism, it will be proper to use either expectation maximization or multiple imputation method to replace missing values in the dataset.

4.7 Comparison of Imputation Methods Using the Root Mean Square Error (RMSE)

RMSE measures the differences between values predicted by a model. The RMSE shows sample standard deviation of the differences between "predicted values and observed values. It is also defined as the square root of the MSE. The MSE and RMSE are integral components in regression models. As such, they are natural measures to use in many forecast error evaluations that use regression-based and statistical methods. One advantage that RMSE has over

MSE is that its scale is similar as the forecast data. Instead of reporting in terms of the average of squared errors, as is the case for MSE, errors reported by the RMSE are representative of the size of an average error". The best imputation method has lower RMSE.

Table 4.11 below shows how each imputation methods for handling missing values under MCAR mechanism performed using the RMSE.

Table 4.11: Performance of imputation methods using the RMSE under MCAR

Percentages	Listwise	Pairwise	MeanSubs	HotDeck	LOCF
5	0.4792586	0.102646	0.4767173	0.4791523	0.4790837
10	0.4801008	0.1032809	0.4755424	0.4795628	0.4795842
15	0.4806464	0.1029053	0.4730848	0.4796394	0.4796476
20	0.4806464	0.1026656	0.4708293	0.4801907	0.4797823
30	1.42E-14	1.49E-08	1.14E-13	1.14E-13	1.14E-13
Overall	**1.9206522**	**0.411497815**	**1.8961738**	**1.9185452**	**1.9180978**

Table 4.11 indicates the performance of imputation methods using the RMSE. Listwise and hotdeck imputation have the RMSE value of 1.9206522 and 1.9185452 respectively, which is the worst performance compared to the various imputation methods under MCAR mechanism. Pairwise deletion performed well in all the percentages of missingness artificially created. This means that, pairwise deletion will do well in both small and large amount of missing values in a dataset under MCAR mechanism. The table also reveals that, the mean substitution and the last observation carried forward (LOCF) did well under the MCAR mechanism. It is important to use either pairwise deletion, the mean substitution or the last observation carried forward (LOCF) imputation methods to replace missing values in a dataset under MCAR mechanism. There is no extreme different using either the MSE or the RMSE to assess the performance of various imputation methods under MCAR mechanism.

Table 4.12 shows how each imputation methods for handling missing values under MAR mechanism performed using the RMSE.

Table 4.12: Performance of imputation methods using the RMSE under MAR

Percentages	EM	MultipleImp
5	0.4784878	0.4793353
10	0.476997	0.4794315
15	0.4760377	0.4794619
20	0.4755165	0.4797079
30	0.4741574	0.4795612
Overall	**2.3811964**	**2.3974978**

Table 4.12 shows that expectation maximization recorded the smallest RMSE compared to the multiple imputation. The table evidenced that, using the RMSE to compare the performance between expectation maximization and multiple imputation method, the expectation maximization performed well in all percentages of missing values compared to the multiple imputation method under MAR mechanism. Multiple imputation did not performed bad compared to expectation maximization. It is important to know that, the two imputation methods did creditably well under MAR mechanism. In situation when large or small amount of data are missing at random, it will be appropriate to use either the expectation maximization or multiple imputation method to replace missing values in the dataset.

CHAPTER 5

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter summarizes the results and discussion and makes conclusions that reflect the research objectives. Recommendations and suggestions are made accordingly for further research.

5.2 Summary

Missing data, often overlooked during data analysis, give biased outcomes, making it hard to give effective and efficient interpretations about a population to guide both practitioners and researchers. Researchers neglect missing data when they do not recognize the importance of the problem or lack the knowledge of possible solutions to solve the problem. Data may be missing for three reasons as stated below:

- Non-coverage: the sample do not represent the population to which the researcher wishes to generalize. Some portions of the intended population were not covered.
- Subject non-response: This is also known as unit non-response, some subjects are included in the sample but fail to provide any information, even to demographic items.
- Item non-response: Some subjects in the sample fail to give all the information for the items.

In this thesis, the study classified and compared methods of handling missing values under the assumptions of various missing mechanism. Listwise deletion, pair-wise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) imputation need the missing data to be MCAR for unbiased estimate. Multiple imputation and expectation maximization method require missing data to be MAR to achieve proper statistical inference. Multiple imputation and expectation maximization comprise of bayesian technique. EM algorithm is an iterative algorithm that estimates the parameters which maximize the log likelihood when there are missingness in the dataset. When there is large fractions of missing data in the dataset, it can be slow to converge. MI also has similar disadvantage as EM algorithm.

The Generalized estimating equation model, the coefficient of determination, and root mean squared error (RMSE) were used as criteria to assess the performance of the seven imputation methods under missing data mechanism. The GEE model for the complete dataset is;

$$logit\left(\hat{p}_1\right) = log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) = -0.041 + 0.124X_1 - 0.146X_2 + 0.081X_3 + 0.602X_4 \tag{5.1}$$

where X_1= Gender, X_2= Marital status, X_3 = Age, X_4 = Educational status, $\hat{p}_1$ represents the estimated probability of recording an "employed" response at the second measurement and $1-\hat{p}_1$ represents the estimated probability of not recording an "employed" response at the second measurement. Based on the GEE model performance criteria for MCAR missing data, listwise deletion is the poorest method. Mean substitution and pairwise deletion performed well when small percentage of missing values occurred in a dataset. The Last observation carried forward (LOCF) method is the best method under MCAR mechanism. From the analysis, it is clear that LOCF do well when large percentage of missing values occur in the dataset. Again, using the GEE model as performance

assessment criteria for MAR missing data, multiple imputation outperformed expectation maximization method by 5%, 10%, 15% and 20% missingness in the dataset.

Using the coefficient of determination to compare various imputation methods under MCAR mechanism, when small percentage of values (5% or 10%) are missing complete at random (MCAR) from a dataset, it will be prudent to use the mean imputation or the LOCF methods. Also, replacing large sets of missing values in a dataset, it is important to use the pairwise imputation method under MCAR mechanism. In all, pairwise method performed well in replacing missing values under MCAR mechanism using the coefficient of determination. Under MAR mechanism using the coefficient of determination, all the two imputation method (ie EM and multiple imputation) performed well in all the percentage of missingness.

Listwise and hotdeck imputation have the RMSE value of 1.9206522 and 1.9185452 respectively, which is the worst performance compared to the various imputation methods under MCAR mechanism. From the analysis, it reveals that Pairwise deletion performed well in all the percentages of missingness artificially created under MCAR mechanism. The analysis also shows that, the mean substitution and the last observation carried forward (LOCF) did well under the MCAR mechanism. It is advisable to use either pairwise deletion, the mean substitution or the last observation carried forward (LOCF) imputation methods to replace missing values in a dataset under MCAR mechanism. It is also evident in the analysis that, using the RMSE to compare the performance between expectation maximization and multiple imputation method, the expectation maximization performed well in all percentage of missing values compared to the multiple imputation method under MAR mechanism.

5.3 Conclusion

The findings from the study discovered that when we have large (above 10%) or small (below 10%) values of data missing at random (MAR), it is important to use multiple imputation or expectation maximization to replace missing values in the dataset. This is because multiple imputation and expectation maximization give unbiased inference and accurate conclusion when used to replace missing values under MAR mechanism. The findings also revealed that, the differences in performance between multiple imputation and expectation maximization are not much, hence the two imputation methods (EM and MI) doing well under MAR mechanism.

Comparing imputation methods under MCAR mechanism, the analysis shows that pairwise deletion is the best. The mean substitution and last observation carried forward (LOCF) also did creditably well. This shows that, when large percentage of missing values occurred in a dataset under MCAR mechanism, the pairwise deletion, last observation carried forward (LOCF) or mean substitution method will give consistent and unbiased estimates, hence it will be prudent to go for either one of these imputation methods when the percentage of missingness is large or small under MCAR mechanism. Listwise deletion and the hot deck imputation methods performed poor under the MCAR mechanism, for this reason, it should not be encouraged in replacing missing values in a dataset. This may lead to bias in parameter estimates of the analysis.

5.4 Recommendations

The following recommendations are made both in the area of policy formulation and future studies based on the findings and conclusions made from the study.

- The study recommend that when data is missing at random (MAR),

multiple imputation or expectation maximization method should be used to replace missing values in the dataset. Under missing complete at random (MCAR), the pairwise deletion, or mean substitution or last observation carried forward (LOCF) is recommended to replace either small or large amount of missing values in the dataset. This will help to achieve proper statistical inference in data analysis.

- Researchers must determine whether the cause and pattern of the missing data will seriously weaken the quality of the inferences derived and which technique is suitable for handling missingness in the dataset. Examining carefully the factors causing missing data and the missing data pattern, helps researchers to decide if and how to best deal with missing values in a study.

- Researcher must think of the research designs and data collection strategies that minimize missing data when planning data collection since data collection plays a major role in the problem of missing data for a particular study. Careful planning in data collection process can decrease the amount of missing data. How to handle the missing data and how to reduce the amount of missing data at the analysis stage are the major factors that must be measured when planning and designing a study for data collection.

- All report on research must report the reasons for and the amount of missingness as well as what data imputation mechanism was employed during the analysis.

5.5 Further Studies

- The study recommended that future researchers find a better technique for imputing missing not at random (MNAR) with multiple imputation. This is vital because of the hope many researchers have in this method due to

the advantages that multiple imputations have among the other imputation methods compared in this research.

- More research should be done to examine sensitivity analysis since it is a significant issue in modeling incomplete longitudinal data when MNAR holds and this should be conducted regularly. In this context, a comparison between different sensitivity analysis models need further analysis.

- This thesis focused on missing values in a longitudinal dataset. However, future research using categorical data is a step in right the direction.

- In clinical sector especially, the study proposed missing values caused by death is treated as MNAR, assuming that the missing values after death can be imputed. This assumption may be implausible. How to handle death as a reason of dropout is a confusing problem that needs to be solved in the future.

REFERENCES

Afifi, A. A., & Elashoff, R. M. (1966). Missing observations in multivariate statistics: Review of the literature. *Journal of the American Statistical Association*, 61, 595–604.

Allison, P. D. (2002). Missing Data: Quantitative Applications in the Social Sciences, Thousand Oaks, CA: Sage.

Amaratunga E., Avery, R. Eisenbeis and Sinkey, J. (2002). *Applications of Discriminant Analysis in Business*, Banking & Finance, 3rd Edition.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, 81, 767-775.

Bahadur, R. R. (1961). Studies in item analysis and prediction, Stanford mathematical studies in the social sciences VI. Chapter *A representation of the joint distribution of responses to n dichotomous items*. Stanford University Press, Stanford, USA.

Batista, G.E.A.P.A., & Monard, M.C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519–533.

Biewen, M. (2001). Item non-response and inequality measurement: Evidence from the German earnings distribution, *AllgemeinesStatistischesArchiv85*, 409-425.

Carpenter, J., Kenward, M.G., Evans, S. & White, I. (2004) Last Observation Carry-Forward and Last Observation Analysis. *Statistics in Medicine*, 23, 3241-3242.

Cartwright, M., & Twala, B. (2010). "Ensemble missing data techniques for software effort prediction". *Intelligent Data Analysis*, 14: 299-331.

Cartwright, M., Twala, B., & Menzies. (2005). *Ensemble Imputation Methods for Missing Software Engineering Data. METRICS*

Cessie, S., & Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society* C, 43, 95-108.

Chen, H., & Little, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika*, 86 (1), 1.

Cheung, M. W.-L. (2007). "Comparison of Methods of Handling Missing Time-Invariant Covariates in Latent Growth Models under the Assumption of Missing Completely at Random", *Organizational Research Methods*, Vol. 10, No. 4, pp. 609-634.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cox, D.R., & Hinckley, D.V. (1974). Theoretical Statistics. London: Chapman & Hall.

Dempster, A.P., Laird, N.M., & Rubin, D. B. (1977). Maximum likelihood

estimation from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Association*, B39, 1-38.

DeSouza, C.M., & Legedza A.T.R. (2009). An overview of practical approaches for handling missing data in clinical trials, *Journal of Biopharmaceutical Statistics*, 19: 1055-1073.

Diggle, P.J., Liang, K.L., & Zeger, S.L. (1994). *Analysis of longitudinal data. Oxford*: Oxford University Press; 247 p.

Durrant, G.B. (2005). Imputation Methods for Handling Item-Nonresponse in the social sciences: A Methodological Review. Southampton: University of Southampton.

Engel, J.M., & Diehr, P. (2003). "Imputation of Missing Longitudinal Data: A Comparison of Methods", *Journal of Clinical Epidemiology*, Vol. 56, No. 10, pp. 968-976.

Feelders, A. (1999). Handling missing data in trees: Surrogate splits or statistical imputation? *Principles of Data Mining and Knowledge Discovery*, 1704:329–334.

Fitzmaurice, G. M., & Lipsitz, S. R. (1995). A model for binary time series data with serial odds ratio patterns. *Applied Statistics*, 51-61.

Fitzmaurice, G.M., Laird, N.M. & Ware, J.H. (2004). *Applied Longitudinal Analysis*, Hoboken, New Jersey: John Wiley & Sons.

Fujikawa, Y., & Ho, T.B. (2002). Cluster-based algorithms for filling missing values. *Lecture Notes in Artificial Intelligence*, 2336:549–554.

Gad, A. M., & Ahmed, A. S. (2006). Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, *Computational Statistics and Data Analysis*, 50(10), 2702-2714.

Gardenier, J.S., & Resnik, D.B. (2002). *The misuse of statistics:* concepts, tools, and a research agenda. Account Res, 9:65-74.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel / hierarchical models.* Columbia University, NY: Cambridge University Press.

Ghahramani, Z. & Jordan, M. I. (1994). Learning from incomplete data. Tech. Rep., Massachusetts Inst. of Technology Artificial Intelligence Lab.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, 59, 834–844

Green, P. D., Barker, J., Cooke, M. P. & Josifovski, L. (2001). *Handling missing and unreliable information in speech recognition.* In Proc. of AISTATS 2001.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society*, B30, 67 - 82.

He, Y. (2010). Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter. *Circulation Cardiovascular Quality and Outcomes*, 3: 98-105.

Hedeker, D., & Gibbons, R.D. (2006). *Longitudinal Data Analysis*, Wiley

Publications.

Heitjan, D. F., & Rubin, D. B. (1991). Ignorability and coarse data. *Institute of Mathematical Statistics*, (19), 2244 - 2253.

Huang, R., & Carriere, K.C. (2006). Comparison of Methods for Incomplete Repeated Measures Data Analysis in Small Samples. *Journal of Statistical Planning and Inference*, 136, 235-247.

Hulse Van, J., & Khoshgoftaar, T.M. (2008). A comprehensive empirical evaluation of missing value imputation in noisy software measurement data, *The Journal of Systems & Software*, 81(5), 691-708.

Jamshidian, M., & Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika*, 1-26.

Jamshidian, M., & Schott, J. (2007). Testing equality of covariance matrices when data are incomplete. *Computational statistics & data analysis*, 51 (9), 4227 - 4239.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91,222–230.

Juster, F. T., & Smith, J. P. (1998). Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92, 27.

Kalousis, A., & Hilario, M. (2000). Supervised knowledge discovery from incomplete data. Cambridge, UK. Proceedings of the 2nd International Conference on Data Mining 2000, WIT Press.

Kim, H., & Yates, S. (2003). Missing value algorithms in decision trees. In H. Bozdogan, editor, *Statistical Data Mining and Knowledge Discovery*, Chapman & Hall/CRC, Boca Raton, Fla, pages 155–172.

Kim, K., & Bentler, P. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67 (4), 609-623.

King, G., Hopnaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95, 49–69.

Klebanoff, M. A., & Cole, S. R. (2008). Use of Multiple Imputation in Epidemiological Literature. *American Journal of Epidemiology*, 168, 355- 357.

Kotz, S., & Johnson, N. L. (1982 – 1988). *Encyclopedia of statistical sciences*. Vol. 1 - 9. -Wiley.

Liang, K., & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

Lillard, L., James P.S., & Finis, W. (1986). What do we really know about wages? The Importance of Non reporting and Census Imputation, *Journal of Political Economy*, 94(3), 489-506.

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1991). Generalized estimating

equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, 78(1), 153-160.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 1198 - 1202.

Little, R. J., & Rubin, D. B. (2002). *Analysis with Missing Data*. Hoboken, New Jersey: Wiley.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227 - 1237.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies, *Journal of the American Statistical Association*, 90, 1112-1121.

Little, R.J.A., & Rubin, D.B (1987). *Statistical Analysis With Missing Data*. New York: John Willey.

McCullagh, P., & Nelder, J. A. (1989). Generalized Linear Models, no. 37 in *Monograph on Statistics and Applied Probability*.

McDonald, R. A., Thurston, P.W., & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. *Organizational Research Methods*, 3, 71—92.

Molenberghs, G., & Leasffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, 18, 2237-2255.

Molenberghs, G., Kenward, M.G. (2007). *Missing Data in Clinical Studies*, John Wiley & Sons, Ltd, Chichester, UK

Musil, C.M., Warner, C.B., Yobas, P.K., & Jones, S.L. (2002). "A Comparison of Imputation Techniques for Handling Missing Data", *Western Journal of Nursing Research*, Vol. 24, No. 7, pp. 815-829.

Pactitti, B. J. (1998) *"Organisational Learning in Research and Development Organisations*: A Study of New Product Development Projects", Unpublished Phd Thesis, University of Manchester, Manchester.

Park, T., & Davis, C. (1993). A test of the missing data mechanism for repeated categorical data. Biometrics, 49 (2), 631- 638.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, 7, 353-383

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 1033-1048.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537–570.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*, New York: J. Wiley & Sons.

Rubin, D. B. (1996). Multiple Imputation after 18 + years. *Journal of American Statistical Society*, 91, 473-489

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* New York: Chapman & Hall.Wiley & Sons.

Schafer, J. L., & Olsen, M.K. (1998). Multiple imputation for multivariate missing-data problems: A data analysts perspective, *Multivariate Behavioural Research*, 33(4), 545–571.

Schrapler, J.P. (2001). Respondent Behavior in Panel Studies. A Case Study of the German Socio-Economic Panel (GSOEP), DIW Discussion Papers No. 244, DIW - Berlin.

Sijtsma, K., & Van der Ark, L.A. (2003). *Investigation and Treatment of Missing Item Scores in Test and Questionnaire Data.* Tilburg: Tilburg University.

Singer, J.D., & J. B. Willett, J.B. (2002). *Applied Longitudinal Data Analysis*: Modeling Change and Event Occurrence, Oxford University.

Sousa-Poza, A., & Henneberger, F. (2000). Wage data collected by telephone interviews: an empirical analysis of the item nonresponse problem and its implications for the estimation of wage functions, *SchweizerischeZeitschriftfürVolkswirtschaft und Statistik* 136(1), 79-98.

Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for mulitple imputation in longitudinal studies. *American Journal of Epidemiology*, 172, 478-487.

Streiner, D. L. (2002).The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*, 47, 68-75.

Stuart, A., & Ord, J. K. (1991). Kendall's advanced theory of statistics, volume 2: *classical inference and relationships* - Edward Arnold.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). *Missing value estimation methods for DNA microarrays. Bioinformatics*, 17, 520–525.

Tufis, C.D. (2008). "Multiple Imputation as a Solution to the Missing Data Problem in Social Sciences", Calitatea Vietii, Vol. 1-2, pp. 199-212.

Twala, B. (2009). An empirical comparison of techniques for handling incomplete data using decision trees. *Applied Artificial Intelligence*, 23:373–405.

Twala, B., Cartwright, M., & Shepperd, M. (2006). *Ensemble of Missing Data Techniques to Improve Software Prediction Accuracy. ICSE.*

Twala, B., Jones, M.C., & Hand, D.J. (2008). Good methods for coping with missing data in decision trees. *Pattern Recognition Letters*, 29:950–956.

Umbach, P. D. (Ed.) (2005). *Survey research: Emerging issues.* New Directions for Institutional Research (Vol. 127). San Francisco: Jossey-Bass.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? *A review of published randomized controlled trials in major medical journals.* Clinical Trials, 1, 368-376.

Zeger, S.L., & Liang, K. Y. (1986). *The analysis of discrete and continuous longitudinal data. Biometrics*, 42, 121-130.

Zhang, S., Qin, Z., Ling, C.X., & Sheng, S. (2005). Missing is useful: Missing values in cost-sensitive decision trees. *IEEE Transactions on Knowledge and Data Engineering*, 17:1689–1693.

Zhou, X.H., Eckert, G.J., & Tierney, W.H. (2001). "Multiple Imputation in Public Health Research", *Statistics in Medicine*, Vol. 20, No. 9-10, pp. 1541-1549.

APPENDIX A

GEE Model for Methods of Handling Missing Data

Table 5.1: GEE Model for Methods of Handling Missing Data Under MCAR Mechanism

		Pairwise		Mean Sub		LOCF		Listwise		Hot Deck	
Percentage missing	Coefficients:	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err	Estimate	Std.err
	(Intercept)	-0.03079	0.09934	-0.0245	0.1	-0.01899	0.09985	-0.00817	0.101281	-0.0353	0.09983
	Gen	0.12046	0.05401	0.10938	0.054	0.1269	0.05421	0.114559	0.055021	0.0972	0.05394
5%	Mari	-0.13579	0.0509	-0.1489	0.051	-0.13452	0.05054	-0.15327	0.051736	-0.1323	0.05051
	Age	0.07165	0.05135	0.06954	0.051	0.07207	0.05142	0.070577	0.052134	0.0688	0.05086
	Edu	0.58893	0.09083	0.60164	0.091	0.57806	0.09116	0.582699	0.092548	0.6107	0.09116
	(Intercept)	-0.03079	0.09934	-0.0308	0.099	-0.04789	0.10166	0.02557	0.10362	-0.0235	0.0988
	Gen	0.12046	0.05401	0.12046	0.054	0.11172	0.05449	0.08757	0.05643	0.1197	0.05391
10%	Mari	-0.13579	0.0509	-0.1358	0.051	-0.13605	0.05083	-0.134	0.05281	-0.1144	0.05051
	Age	0.07165	0.05135	0.07165	0.051	0.07083	0.05163	0.05919	0.05359	0.0743	0.05125
	Edu	0.58893	0.09083	0.58893	0.091	0.61136	0.09272	0.55065	0.095	0.5752	0.09045
	(Intercept)	-0.03079	0.09934	-0.0322	0.099	-0.03945	0.10113	0.04084	0.10498	-0.0332	0.09856
	Gen	0.12046	0.05401	0.12094	0.054	0.10975	0.05484	0.0849	0.05772	0.1132	0.05393
15%	Mari	-0.13579	0.0509	-0.131	0.051	-0.11889	0.05083	-0.14585	0.05405	-0.104	0.05032
	Age	0.07165	0.05135	0.07874	0.051	0.07093	0.05181	0.0764	0.05491	0.0825	0.05123
	Edu	0.58893	0.09083	0.58253	0.091	0.59524	0.09202	0.52691	0.09597	0.5771	0.09013
	(Intercept)	-0.03079	0.09934	-0.0433	0.099	-0.07915	0.10321	0.000603	0.110657	-0.0516	0.09948
	Gen	0.12046	0.05401	0.13171	0.054	0.11798	0.05516	0.077233	0.059298	0.1243	0.05379
20%	Mari	-0.13579	0.0509	-0.1275	0.051	-0.11434	0.05073	-0.12638	0.055172	-0.0864	0.05025
	Age	0.07165	0.05135	0.07477	0.051	0.07278	0.05232	0.086675	0.056426	0.0914	0.05106
	Edu	0.58893	0.09083	0.5818	0.09	0.62431	0.09394	0.562793	0.100921	0.5733	0.09061
	(Intercept)	-4.96E+01	4.16E-01	51.5807	0.073	5.16E+01	7.14E-02	5.16E+01	7.92E-02	51.582	0.070551
	Gen	1.86E-12	4.05E+00	0.00024	0.037	3.50E-04	3.58E-02	-1.31E-05	3.94E-02	0.0036	0.035231
30%	Mari	5.12E-12	3.91E+00	-0.0001	0.025	1.48E-04	2.55E-02	3.37E-06	2.81E-02	-0.0007	0.024327
	Age	-3.79E-12	3.30E+00	-0.0001	0.034	1.41E-05	3.28E-02	6.46E-06	3.63E-02	-0.0006	0.032324
	Edu	-1.30E-12	1.27E+01	-3E-06	0.069	-6.19E-04	6.70E-02	-1.18E-06	7.43E-02	0.0006	0.066302

Table 5.2: GEE Model for Methods of Handling Missing Data Under MAR Mechanism

		Multiple Imputation		EM	
Percentage missing	**Coefficients:**	**Estimate**	**Std.err**	**Estimate**	**Std.err**
	(Intercept)	0.063	0.1005	-0.0637	0.10083
5%	Gen	-0.123	0.0543	0.13353	0.05418
	Mari	0.144	0.0509	-0.14401	0.05069
	Age	-0.082	0.0528	0.07847	0.05138
	Edu	-0.62	0.0918	0.62843	0.09177
	(Intercept)	0.032	0.1058	-0.03577	0.10205
	Gen	-0.106	0.0554	0.11162	0.05464
10%	Mari	0.156	0.0528	-0.18345	0.0508
	Age	-0.083	0.0517	0.09182	0.05154
	Edu	-0.606	0.0972	0.64926	0.09354
	(Intercept)	0.02	0.1043	-0.006646	0.10305
	Gen	-0.107	0.0568	0.112152	0.055033
15%	Mari	0.15	0.0516	-0.190983	0.050959
	Age	-0.101	0.0519	0.100474	0.051759
	Edu	-0.58	0.0983	0.632961	0.094545
	(Intercept)	-0.002	0.1026	0.003285	0.103364
	Gen	-0.097	0.0561	0.093397	0.05542
20%	Mari	-0.086	0.0521	-0.20592	0.051078
	Age	0.149	0.0524	0.0943	0.051868
	Edu	-0.573	0.0944	0.653251	0.094854
	(Intercept)	0.033	0.1071	0.007142	0.104618
	Gen	-0.09	0.0583	0.080263	0.056086
30%	Mari	0.147	0.0524	-0.224968	0.051479
	Age	-0.11	0.054	0.111297	0.052162
	Edu	-0.597	0.0957	0.678922	0.096161

GEE Coefficients for Various Percentages of Missing Data Values

Table 5.3: Coefficients of Various Percentages of Missing Data Values Under MCAR and MAR

General (full data)				
intercept	Gen	Mari	Age	Edu
-0.0409	0.12411	-0.14627	0.08058	0.6023

Missing Completely at Random

	Listwise					Pairwise				
	intercept	Gen	Mari	Age	Edu	intercept	Gen	Mari	Age	Edu
5%	-0.00817	0.114559	-0.15327	0.07058	0.582699	-0.0308	0.1205	-0.13579	0.07165	0.5889
10%	0.02557	0.08757	-0.134	0.05919	0.55065	-0.0308	0.1205	-0.13579	0.07165	0.5889
15%	0.04084	0.0849	-0.14585	0.0764	0.52691	-0.0218	0.1311	-0.1442	0.07265	0.4788
20%	0.000603	0.077233	-0.12638	0.08668	0.562793	-0.0311	0.2311	-0.1356	0.03421	0.4611
30%	5.16E+01	-1.31E-05	3.37E-06	6.46E-06	-1.18E-06	0.0011	0.1102	0.2356	0.02341	0.1266

	Mean sub					Hotdecking				
	intercept	Gen	Mari	Age	Edu	intercept	Gen	Mari	Age	Edu
5%	-0.02448	0.109	-0.1489	0.06954	0.60164	-0.0353	0.0972	-0.1323	0.06882	0.6107
10%	-0.03079	0.12	-0.1358	0.07165	0.58893	-0.0235	0.1197	-0.1144	0.07429	0.5752
15%	-0.0322	0.121	-0.131	0.07874	0.58253	-0.0332	0.1132	-0.104	0.0825	0.5771
20%	-0.04327	0.132	-0.1275	0.07477	0.5818	-0.0516	0.1243	-0.0864	0.09139	0.5733
30%	51.58074	2.00E-04	-0.0001	-0.0001	-2.80E-06	51.5819	0.0036	-0.0007	-0.0006	0.0006

	LOCF				
	intercept	Gen	Mari	Age	Edu
5%	-0.01899	0.1269	-0.13452	0.07207	0.57806
10%	-0.04789	0.11172	-0.13605	0.07083	0.61136
15%	-0.03945	0.10975	-0.11889	0.07093	0.59524
20%	-0.07915	0.11798	-0.11434	0.07278	0.62431
30%	5.16E+01	3.50E-04	1.48E-04	1.41E-05	-6.19E-04

Missing at Random

	Multiple Imputation					EM				
	intercept	Gen	Mari	Age	Edu	intercept	Gen	Mari	Age	Edu
5%	.063	-.123	.144	-.082	-.620	-0.0637	0.1335	-0.14401	0.07847	0.6284
10%	.032	-.106	.156	-.083	-.606	-0.0358	0.1116	-0.18345	0.09182	0.6493
15%	.020	-.107	.150	-.101	-.580	-0.0066	0.1122	-0.19098	0.10047	0.633
20%	-.002	-.097	-.086	.149	-.573	0.00329	0.0934	-0.20592	0.0943	0.6533
30%	.033	-.090	.147	-.110	-.597	0.00714	0.0803	-0.22497	0.1113	0.6789

Table 5.4: 5% Multiple Imputation Under MAR Mechanism

Parameter Estimates

Imputation Number		B	Std. Error	95% Wald Lower	95% Wald Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Fraction Missing Info.	Rel Efficiency
Original data	(Intercept)	.036	.1028	-.165	.238	.123	1	.726		
	Gen	-.106	.0553	-.215	.002	3.693	1	.055		
	Mari	.133	.0515	.032	.234	6.709	1	.010		
	Age	-.067	.0524	-.169	.036	1.628	1	.202		
	Edu	-.599	.0940	-.783	-.414	40.557	1	.000		
	(Scale)	1								
1	(Intercept)	.056	.1005	-.141	.253	.307	1	.580		
	Gen	-.126	.0540	-.232	-.020	5.464	1	.019		
	Mari	.154	.0504	.055	.252	9.298	1	.002		
	Age	-.092	.0512	-.192	.008	3.235	1	.072		
	Edu	-.614	.0916	-.793	-.434	44.889	1	.000		
	(Scale)	1								
2	(Intercept)	.065	.1000	-.131	.261	.427	1	.514		
	Gen	-.123	.0541	-.229	-.017	5.171	1	.023		
	Mari	.143	.0505	.044	.242	7.985	1	.005		
	Age	-.063	.0513	-.164	.037	1.518	1	.218		
	Edu	-.628	.0912	-.807	-.449	47.343	1	.000		
	(Scale)	1								
3	(Intercept)	.063	.1002	-.134	.259	.389	1	.533		
	Gen	-.128	.0541	-.234	-.022	5.588	1	.018		
	Mari	.142	.0504	.043	.240	7.889	1	.005		
	Age	-.090	.0513	-.190	.011	3.046	1	.081		
	Edu	-.613	.0914	-.792	-.434	44.961	1	.000		
	(Scale)	1								
4	(Intercept)	.066	.1007	-.131	.263	.428	1	.513		
	Gen	-.118	.0541	-.224	-.012	4.751	1	.029		
	Mari	.137	.0505	.038	.236	7.374	1	.007		
	Age	-.079	.0513	-.180	.021	2.390	1	.122		
	Edu	-.624	.0918	-.804	-.444	46.189	1	.000		
	(Scale)	1								
5	(Intercept)	.067	.1006	-.131	.264	.439	1	.508		
	Gen	-.120	.0540	-.226	-.015	4.975	1	.026		
	Mari	.144	.0504	.045	.243	8.184	1	.004		
	Age	-.086	.0511	-.187	.014	2.866	1	.090		
	Edu	-.622	.0917	-.801	-.442	45.924	1	.000		
Pooled	(Intercept)	.063	.1005	-.134	.260			.530	.002	1.000
	Gen	-.123	.0543	-.229	-.017			.023	.007	.999
	Mari	.144	.0509	.044	.244			.005	.017	.997
	Age	-.082	.0528	-.186	.021			.120	.060	.988
	Edu	-.620	.0918	-.800	-.440			.000	.006	.999
	(Scale)	1.000	0.0000					.[a]		

Dependent Variable: Work

a. No variance among imputations.

Table 5.5: 10% Multiple Imputation Under MAR Mechanism

Parameter Estimates											
				95% Wald		Hypothesis Test			Fraction	Relative	Relative
Imputation Number		B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Missing Info.	Increase Variance	Efficiency
Original data	(Intercept)	.058	.1064	-.151	.266	.295	1	.587			
	Gen	-.120	.0564	-.231	-.010	4.538	1	.033			
	Mari	.149	.0527	.046	.253	8.025	1	.005			
	Age	-.071	.0534	-.176	.033	1.790	1	.181			
	Edu	-.632	.0973	-.822	-.441	42.166	1	.000			
	(Scale)	1									
1	(Intercept)	.054	.1012	-.144	.253	.289	1	.591			
	Gen	-.113	.0540	-.219	-.008	4.421	1	.035			
	Mari	.169	.0507	.070	.269	11.181	1	.001			
	Age	-.076	.0510	-.176	.024	2.208	1	.137			
	Edu	-.635	.0921	-.815	-.454	47.435	1	.000			
	(Scale)	1									
2	(Intercept)	.060	.1012	-.138	.258	.351	1	.554			
	Gen	-.111	.0540	-.217	-.006	4.255	1	.039			
	Mari	.170	.0507	.071	.269	11.253	1	.001			
	Age	-.085	.0512	-.186	.015	2.769	1	.096			
	Edu	-.636	.0923	-.817	-.456	47.559	1	.000			
	(Scale)	1									
3	(Intercept)	.026	.1013	-.172	.225	.068	1	.794			
	Gen	-.109	.0539	-.215	-.004	4.110	1	.043			
	Mari	.139	.0505	.040	.238	7.577	1	.006			
	Age	-.080	.0511	-.180	.021	2.428	1	.119			
	Edu	-.590	.0923	-.771	-.409	40.878	1	.000			
	(Scale)	1									
4	(Intercept)	-.011	.1013	-.210	.187	.012	1	.912			
	Gen	-.087	.0542	-.193	.020	2.556	1	.110			
	Mari	.148	.0507	.049	.248	8.540	1	.003			
	Age	-.094	.0512	-.194	.007	3.345	1	.067			
	Edu	-.574	.0926	-.756	-.393	38.471	1	.000			
	(Scale)	1									
5	(Intercept)	.029	.1010	-.169	.227	.082	1	.774			
	Gen	-.110	.0541	-.216	-.004	4.160	1	.041			
	Mari	.154	.0506	.055	.253	9.280	1	.002			
	Age	-.080	.0511	-.180	.020	2.455	1	.117			
	Edu	-.597	.0925	-.778	-.416	41.648	1	.000			
	(Scale)	1									
Pooled	(Intercept)	.032	.1058	-.176	.240			.765	.089	.093	.983
	Gen	-.106	.0554	-.215	.002			.055	.049	.050	.990
	Mari	.156	.0528	.053	.260			.003	.081	.085	.984
	Age	-.083	.0517	-.184	.018			.109	.022	.022	.996
	Edu	-.606	.0972	-.798	-.415			.000	.102	.108	.980
	(Scale)	1.000	0.0000					.[a]			

Dependent Variable: Work

Model: (Intercept), Gen, Mari, Age, Edu

a. No variance among imputations.

Table 5.6: 15% Multiple Imputation Under MAR Mechanism

Parameter Estimates

Imputation Number		B	Std. Error	95% Wald Lower	95% Wald Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
Original data	(Intercept)	.028	.1097	-.187	.243	.067	1	.796			
	Gen	-.120	.0579	-.233	-.007	4.305	1	.038			
	Mari	.148	.0540	.043	.254	7.566	1	.006			
	Age	-.075	.0547	-.182	.033	1.865	1	.172			
	Edu	-.607	.1001	-.804	-.411	36.784	1	.000			
	(Scale)	1									
1	(Intercept)	.009	.1001	-.188	.205	.007	1	.932			
	Gen	-.103	.0541	-.209	.003	3.607	1	.058			
	Mari	.153	.0506	.054	.252	9.149	1	.002			
	Age	-.104	.0512	-.205	-.004	4.170	1	.041			
	Edu	-.577	.0914	-.756	-.398	39.829	1	.000			
	(Scale)	1									
2	(Intercept)	-.007	.0997	-.203	.188	.006	1	.941			
	Gen	-.083	.0538	-.189	.022	2.383	1	.123			
	Mari	.136	.0504	.038	.235	7.326	1	.007			
	Age	-.102	.0511	-.202	-.002	3.982	1	.046			
	Edu	-.559	.0912	-.738	-.381	37.648	1	.000			
	(Scale)	1									
3	(Intercept)	.059	.1010	-.139	.257	.344	1	.558			
	Gen	-.110	.0539	-.215	-.004	4.152	1	.042			
	Mari	.161	.0504	.062	.260	10.190	1	.001			
	Age	-.090	.0511	-.190	.010	3.113	1	.078			
	Edu	-.630	.0919	-.810	-.450	47.060	1	.000			
	(Scale)	1									
4	(Intercept)	.026	.1004	-.171	.223	.066	1	.797			
	Gen	-.112	.0543	-.219	-.006	4.261	1	.039			
	Mari	.154	.0508	.054	.253	9.182	1	.002			
	Age	-.097	.0510	-.197	.003	3.643	1	.056			
	Edu	-.588	.0914	-.767	-.409	41.397	1	.000			
	(Scale)	1									
5	(Intercept)	.012	.1020	-.188	.212	.014	1	.905			
	Gen	-.127	.0540	-.233	-.021	5.519	1	.019			
	Mari	.144	.0505	.045	.243	8.118	1	.004			
	Age	-.112	.0510	-.212	-.012	4.820	1	.028			
	Edu	-.547	.0934	-.730	-.364	34.328	1	.000			
	(Scale)	1									
Pooled	(Intercept)	.020	.1043	-.185	.224			.851	.072	.074	.986
	Gen	-.107	.0568	-.219	.005			.060	.099	.105	.981
	Mari	.150	.0516	.048	.251			.004	.042	.042	.992
	Age	-.101	.0519	-.203	.000			.051	.030	.030	.994
	Edu	-.580	.0983	-.774	-.387			.000	.134	.146	.974
	(Scale)	1.000	0.0000					.[a]			

Dependent Variable: Work

Model: (Intercept), Gen, Mari, Age, Edu

a. No variance among imputations.

Table 5.7: 20% Multiple Imputation Under MAR Mechanism

Parameter Estimates											
				95% Wald		Hypothesis Test			Fraction	Relative	Relative
Imputation Number		B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Missing Info.	Increase Variance	Efficiency
Original data	(Intercept)	.051	.1120	-.169	.270	.203	1	.652			
	Gen	-.094	.0592	-.210	.022	2.518	1	.113			
	Age	-.072	.0561	-.182	.038	1.637	1	.201			
	Mari	.146	.0552	.038	.254	6.987	1	.008			
	Edu	-.638	.1027	-.839	-.437	38.570	1	.000			
	(Scale)	1									
1	(Intercept)	-.001	.1008	-.199	.196	.000	1	.989			
	Gen	-.117	.0540	-.223	-.011	4.663	1	.031			
	Age	-.074	.0509	-.174	.026	2.114	1	.146			
	Mari	.138	.0505	.039	.237	7.494	1	.006			
	Edu	-.563	.0928	-.744	-.381	36.738	1	.000			
	(Scale)	1									
2	(Intercept)	.011	.0996	-.185	.206	.011	1	.915			
	Gen	-.080	.0542	-.186	.027	2.163	1	.141			
	Age	-.100	.0510	-.200	.000	3.865	1	.049			
	Mari	.133	.0504	.034	.232	6.972	1	.008			
	Edu	-.577	.0905	-.754	-.399	40.601	1	.000			
	(Scale)	1									
3	(Intercept)	-.033	.1007	-.230	.165	.105	1	.746			
	Gen	-.096	.0543	-.203	.010	3.158	1	.076			
	Age	-.087	.0513	-.188	.013	2.897	1	.089			
	Mari	.158	.0505	.059	.257	9.772	1	.002			
	Edu	-.547	.0917	-.727	-.367	35.540	1	.000			
	(Scale)	1									
4	(Intercept)	.008	.1012	-.191	.206	.006	1	.940			
	Gen	-.101	.0540	-.207	.005	3.478	1	.062			
	Age	-.086	.0512	-.187	.014	2.832	1	.092			
	Mari	.159	.0506	.060	.259	9.915	1	.002			
	Edu	-.591	.0925	-.772	-.410	40.823	1	.000			
	(Scale)	1									
5	(Intercept)	.005	.1016	-.194	.205	.003	1	.957			
	Gen	-.093	.0542	-.199	.013	2.936	1	.087			
	Age	-.085	.0512	-.185	.016	2.731	1	.098			
	Mari	.159	.0508	.059	.258	9.767	1	.002			
	Edu	-.590	.0931	-.772	-.408	40.165	1	.000			
	(Scale)	1									
Pooled	(Intercept)	-.002	.1026	-.203	.199			.984	.036	.037	.993
	Gen	-.097	.0561	-.207	.013			.083	.071	.074	.986
	Age	-.086	.0521	-.189	.016			.097	.039	.040	.992
	Mari	.149	.0524	.047	.252			.004	.073	.076	.986
	Edu	-.573	.0944	-.759	-.388			.000	.049	.050	.990
	(Scale)	1.000	0.0000					.[a]			

Dependent Variable: Work

Model: (Intercept), Gen, Age, Mari, Edu

a. No variance among imputations.

Table 5.8: 30% Multiple Imputation Under MAR Mechanism

Parameter Estimates

Imputation Number		B	Std. Error	95% Wald Lower	95% Wald Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Fraction Missing Info.	Relative Increase Variance	Relative Efficiency
Original data	(Intercept)	.040	.1157	-.187	.267	.118	1	.731			
	Gen	-.093	.0622	-.214	.029	2.212	1	.137			
	Mari	.140	.0579	.026	.253	5.830	1	.016			
	Age	-.091	.0587	-.206	.025	2.379	1	.123			
	Edu	-.631	.1060	-.839	-.423	35.409	1	.000			
	(Scale)	1									
1	(Intercept)	.007	.1009	-.191	.205	.005	1	.946			
	Gen	-.062	.0544	-.169	.044	1.317	1	.251			
	Mari	.156	.0503	.057	.254	9.588	1	.002			
	Age	-.097	.0508	-.197	.002	3.673	1	.055			
	Edu	-.597	.0922	-.778	-.417	41.985	1	.000			
	(Scale)	1									
2	(Intercept)	.065	.1008	-.132	.263	.421	1	.516			
	Gen	-.108	.0537	-.213	-.002	4.006	1	.045			
	Mari	.151	.0507	.052	.250	8.860	1	.003			
	Age	-.128	.0512	-.228	-.027	6.226	1	.013			
	Edu	-.619	.0927	-.801	-.437	44.561	1	.000			
	(Scale)	1									
3	(Intercept)	.072	.1012	-.126	.270	.505	1	.477			
	Gen	-.109	.0546	-.216	-.002	3.960	1	.047			
	Mari	.125	.0507	.026	.224	6.083	1	.014			
	Age	-.095	.0512	-.195	.005	3.435	1	.064			
	Edu	-.624	.0923	-.805	-.444	45.793	1	.000			
	(Scale)	1									
4	(Intercept)	-.002	.1002	-.198	.195	.000	1	.986			
	Gen	-.081	.0544	-.188	.026	2.213	1	.137			
	Mari	.149	.0505	.050	.248	8.707	1	.003			
	Age	-.102	.0510	-.202	-.002	3.974	1	.046			
	Edu	-.573	.0917	-.753	-.393	39.049	1	.000			
	(Scale)	1									
5	(Intercept)	.021	.0994	-.173	.216	.046	1	.829			
	Gen	-.090	.0543	-.197	.016	2.761	1	.097			
	Mari	.155	.0507	.055	.254	9.295	1	.002			
	Age	-.126	.0511	-.226	-.026	6.111	1	.013			
	Edu	-.572	.0904	-.749	-.395	40.109	1	.000			
	(Scale)	1									
Pooled	(Intercept)	.033	.1071	-.178	.244			.760	.126	.136	.975
	Gen	-.090	.0583	-.205	.025			.124	.140	.152	.973
	Mari	.147	.0524	.044	.250			.005	.072	.074	.986
	Age	-.110	.0540	-.216	-.003			.043	.112	.119	.978
	Edu	-.597	.0957	-.785	-.409			.000	.083	.087	.984
	(Scale)	1.000	0.0000					.[a]			

Dependent Variable: Work

Model: (Intercept), Gen, Mari, Age, Edu

a. No variance among imputations.

Patterns of Missing values

Variable Summary

	Missing		Valid
	N	Percent	N
age	77	1.10%	6687
education	72	1.10%	6692
Marital status	71	1.00%	6693
work status	61	0.90%	6703
gender	59	0.90%	6705

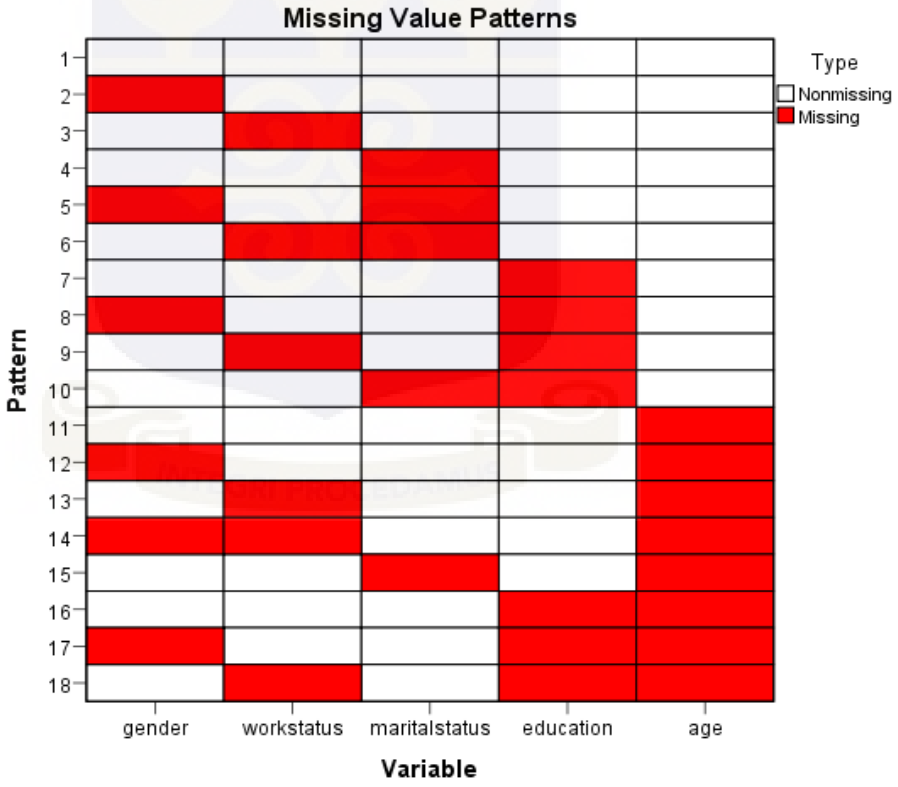

Figure 5.1: 5% Pattern of missing values at random (MAR)

Variable Summary

	Missing		Valid
	N	Percent	N
Work	172	2.50%	6592
Edu	166	2.50%	6598
Mari	129	1.90%	6635
Gen	112	1.70%	6652
Age	99	1.50%	6665

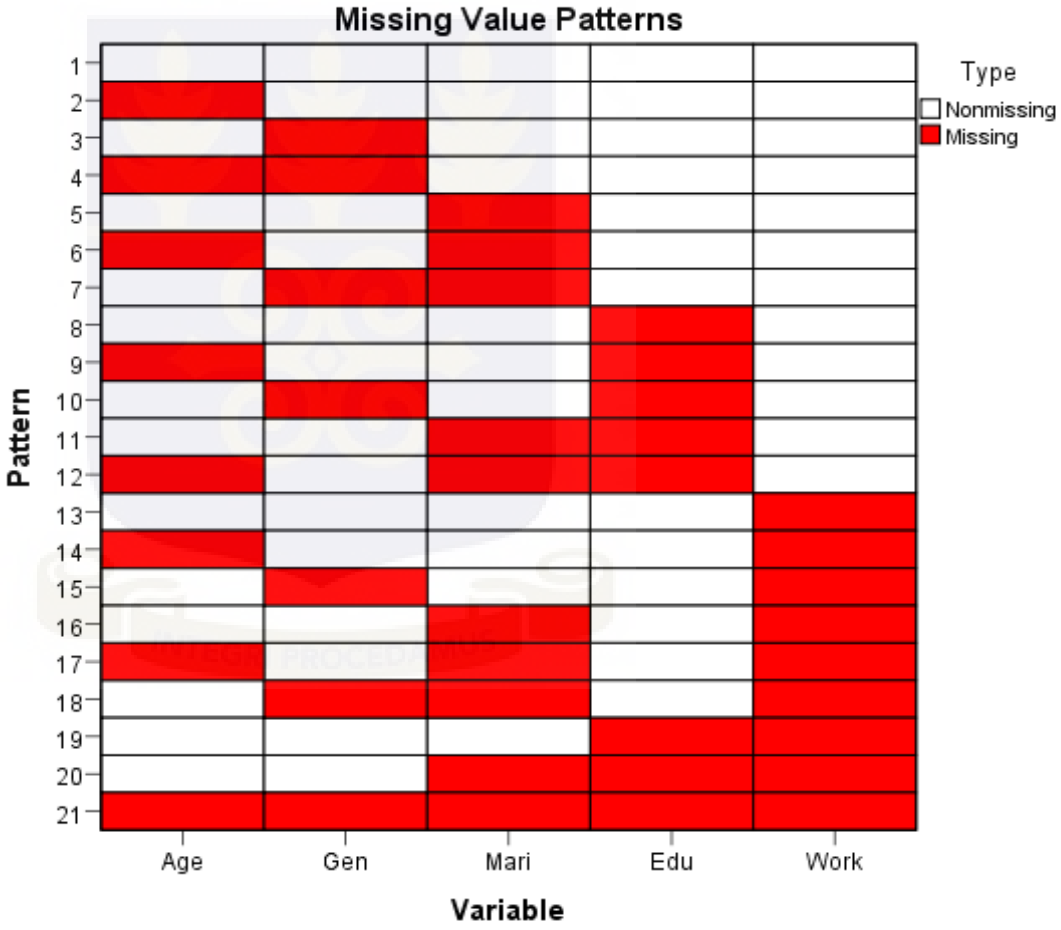

Figure 5.2: 10% Pattern of missing values at random (MAR)

	Missing		Valid
	N	Percent	N
Edu	516	7.60%	6248
Age	391	5.80%	6373
Work	382	5.60%	6382
Mari	382	5.60%	6382
Gen	360	5.30%	6404

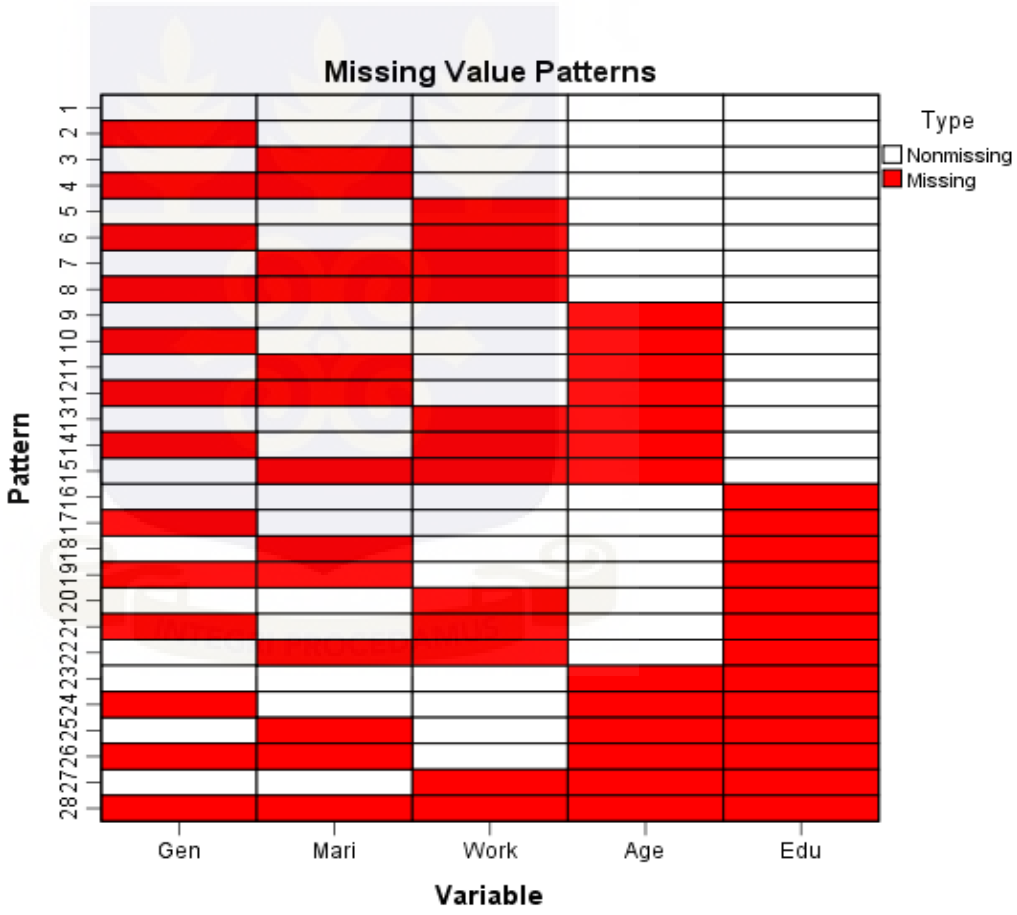

Figure 5.3: 30% Pattern of missing values at random (MAR)

Variable Summary

	Missing		Valid
	N	Percent	N
Mari	78	1.20%	6686
Work	70	1.00%	6694
Gen	69	1.00%	6695
Age	66	1.00%	6698
Edu	57	0.80%	6707

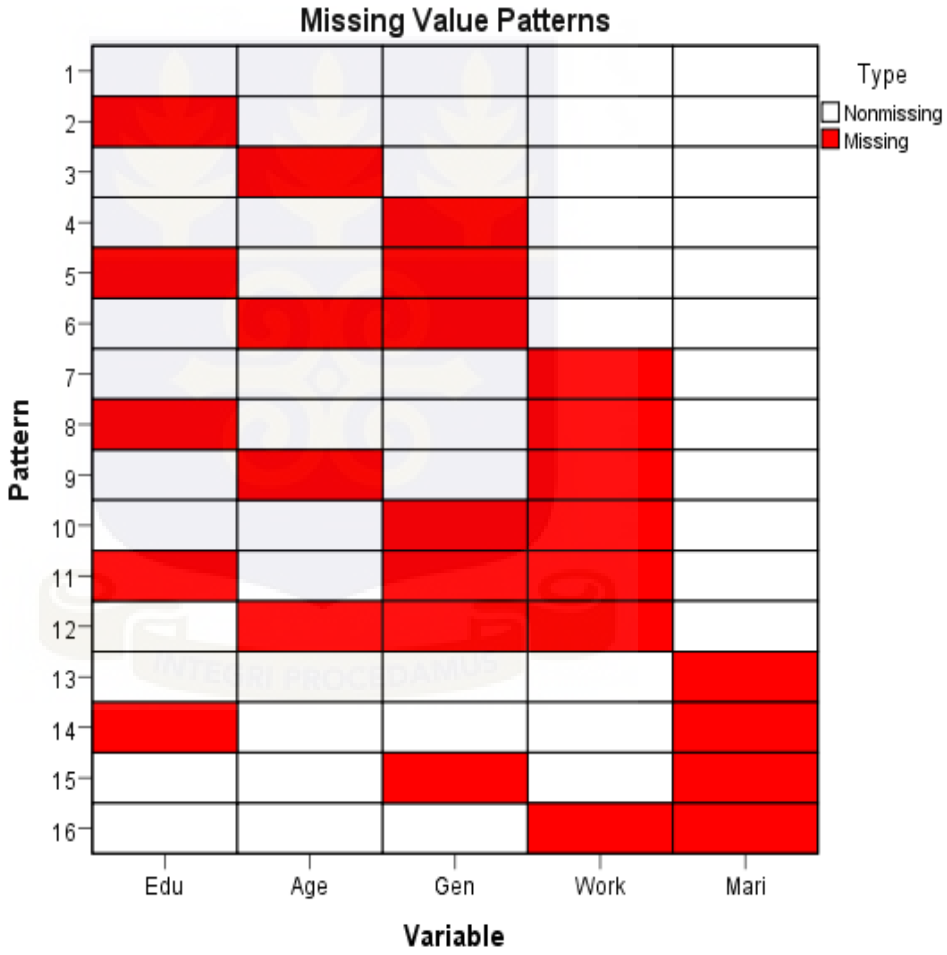

Figure 5.4: 5% Pattern of missing values completely at random (MCAR)

Variable Summary

	Missing		Valid
	N	Percent	N
Mari	184	2.70%	6580
Edu	156	2.30%	6608
Age	130	1.90%	6634
Work	111	1.60%	6653
Gen	97	1.40%	6667

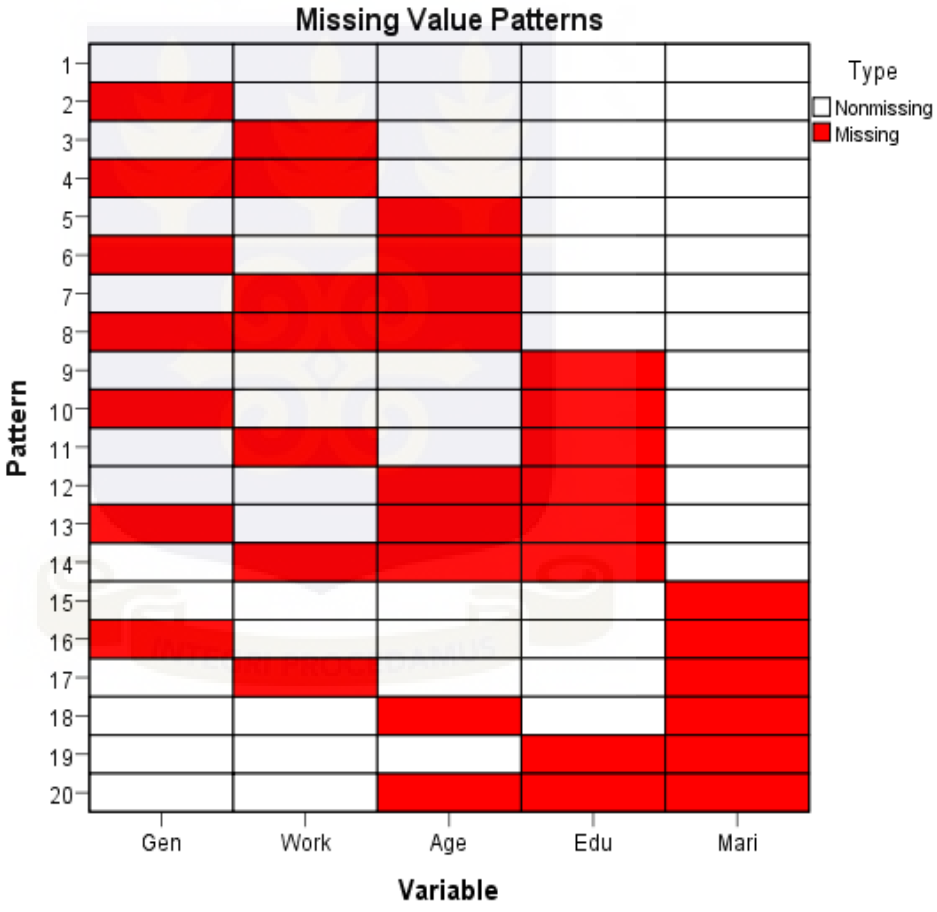

Figure 5.5: 10% Pattern of missing values completely at random (MCAR)

Variable Summary

	Missing		Valid
	N	Percent	N
Mari	503	7.40%	6261
Edu	409	6.00%	6355
Work	394	5.80%	6370
Age	381	5.60%	6383
Gen	344	5.10%	6420

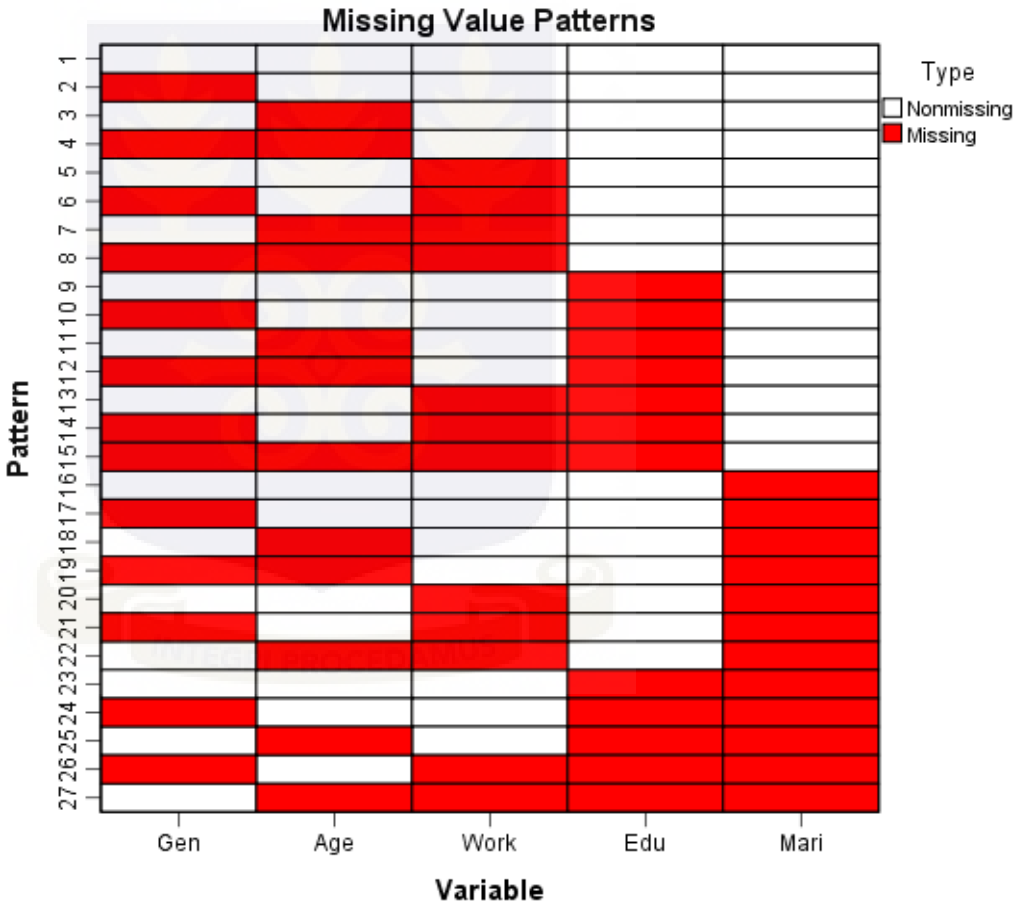

Figure 5.6: 30% Pattern of missing values completely at random (MCAR)

APPENDIX B

Codes Used in Data Analysis

```
##################################
Little's Test of MCAR
##################################

nn=read.csv("C:/Users/USER/Desktop/gg.csv",na.strings=",")
head(nn)
attach(nn)
str(nn)
library(BaylorEdPsych)
library(mvnmle)
LittleMCAR(nn)


##################################
Patterns of Missingness
##################################

library(mice)
library(VIM)
p<-aggr(nn,col=c('navyblue','red'),numbers=TRUE,SortVars=TRUE,labels=names
(nn),Cex.axis=0.7,gap=3,ylab=c("Histogram of missing data", "Pattern"))


##################################
Number of Missingness in a Data
##################################

Call for p
```

```
##################################
Pattern and distribution of complete and incomplete observations
##################################
```

```
nn=read.csv("C:/Users/USER/Desktop/cre30.csv",na.strings=",")
names(nn)
head(nn)
attach(nn)
str(nn)

library(mice)
library(VIM)
library(Hmisc)
library(Amelia)
library(MASS)
library(geepack)
library(lattice)
library(ggplot2)

md.pattern(nn)
p=md.pairs(nn)
p

marginplot(nn[,c("Age","Work")], col=c("blue","red","orange"), cex=1.5,
cex.lab=1.5, cex.numbers=1.3, pch=19) pbox(nn,pos=5)
```

```
##########################################
```

Generalised Estimating Equation

```
##########################################
```

```
dat=read.table("C:/Users/USER/Desktop/Re.txt")
names(dat)=c("pid","wav","Gen","Mari","Age","Edu","Work")
names(dat)
dat
head(dat)
attach(dat)
str(dat)

library(MASS)
library(geepack)
geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,data=dat,family=binomial,
id=pid)
summary(geeglm.out)
geeglm.out
logLik(geeglm.out)
```

```
##########################################
```

Hotdeck Imputation

```
##########################################
```

```
Ra=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
Ra
names(Ra)
attach(Ra)
head(Ra)
str(Ra)
```

```
library(VIM)
x=hotdeck(Ra,variable=c("Gen","Mari","Age","Edu","Work"))
x

##########################################
GEE for Hotdeck Imputation
##########################################

library(MASS)
library(geepack)

geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,data=x,family=binomial,id=pid)
summary(geeglm.out)
geeglm.out

##########################################
Last Observation Carried Forward (LOCF)
##########################################

pro=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
names(pro)
attach(pro)
head(pro)
str(pro)

library(zoo)
la=na.locf(pro)
la

##########################################
GEE for Last Observation Carried Forward (LOCF)
##########################################
```

```
library(MASS)
library(geepack)
geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,data=la,family=binomial,id=pid)
summary(geeglm.out)
geeglm.out
```

##

Listwise Deletion

##

```
good=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
names(good)
attach(good)
head(good)
str(good)
Fm=good[complete.cases(good),]
Fm
```

##

GEE for Listwise Deletion

##

```
library(MASS) library(geepack) geeglm.out<-geeglm(formula=Work Gen+Mari+Age+
Edu,data=Fm,family=binomial,id=pid)
summary(geeglm.out)
geeglm.out
```

##

Mean Substitution

##

```
ka=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
names(ka)
```

```
attach(ka)
head(ka)
str(ka)

mean.subst <- function(a) {
a[is.na(a)] <- mean(a, na.rm=TRUE)
a
}
pro <- apply(ka,2,mean.subst)
pro
lam=data.frame(pro)
lam
```

##################################

GEE for Mean Substitution

##################################

```
library(MASS)
library(geepack)
geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,family=binomial,
data=lam,id=pid)
summary(geeglm.out)
geeglm.out
```

##################################

Pairwise Deletion

##################################

```
Joh=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
names(Joh)
attach(Joh)
head(Joh)
```

```
str(Joh)
pan=cov(Joh,use="pairwise.complete.obs")
pan
Remm=data.frame(pan)
Remm

##################################
GEE for Pairwise Deletion
##################################

library(MASS)
library(geepack)

geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,family=binomial,
data=Remm,id=pid)
summary(geeglm.out)
geeglm.out
logLik(geeglm.out)

##################################
Multiple Imputation
##################################

Dan=read.csv("C:/Users/USER/Desktop/cre5.csv",na.strings=",")
names(Dan)
head(Dan)
attach(Dan)
str(Dan)

library(mice)
library(VIM)
library(Hmisc)
```

```
library(Amelia)
library(MASS)
library(geepack)
library(lattice)
library(ggplot2)
van=mice(Dan, m=5)
van
van$imp$Gen$
mu=complete(van,'long',inc=TRUE)
mu
geeglm.out=with(van,geeglm(formula=Work Gen+Mari+Age+Edu,
data=van,family=binomial,id=pid))
summary(geeglm.out)
geeglm.out
pool(geeglm.out)
summary(pool(geeglm.out))
```

```
##################################
Expectation maximization (EM)
##################################
```

```
dat=read.csv("C:/Users/USER/Desktop/ass30.csv",na.strings=",")
names(dat)
head(dat)
attach(dat)
str(dat)

library(norm)
library(VIM)
library(MASS)
library(geepack)
```

```
library(cat)

em.norm=function(dat,mit,sit)
{
dat2=dat[!is.na(dat)]
dat3=dat[is.na(dat)]
n=length(c(dat2,dat3))
r=length(dat2)

###Initial values###
mut=mit
sit=sit

###To define the log-likelihood function###
kan=function(y,mu,sigma2,n)
{
−0.5 ∗ n ∗ log(2 ∗ pi ∗ sigma2) − 0.5 ∗ sum((y − mu)^2)/sigma2
}
kan2=kan(dat2,mut,sit,n)
repeat{

###E-step###
edat=sum(dat2)+(n-r)*mut
edat2=sum(dat2^2) + (n − r) ∗ (mut^2 + sit)

###M-step###
mut1=edat/n
sit1=edat2 / n - mut1^2
```

```
###Update parameter values###
mut=mut1
sit=sit1
abn=kan(dat2,mut,sit,n)

###print current parameter values and likelihood###
cat(mut,sit,abn,"/n")

###Stop if converged###
if(abs(kan2-abn)<0.001)
kan2=abn
}
return(c(mut,sit))
}

###function en.norm ends here###
Sam = em.norm(dat,0,0.2)

geeglm.out<-geeglm(formula=Work Gen+Mari+Age+Edu,data=sam,
family=binomial,id=pid)
summary(geeglm.out)
geeglm.out
logLik(geeglm.out)
```

##

Mean Square Error (MSE)

##

```
Library (nlme)
mse <-(mean(geeglm.out$residuals^2))
mse
```

##

Root Mean Square Error (RMSE)

##

Library (nlme)

r.rmse <- sqrt(mean($residuals$(geeglm.out)^2$)

r.rmse