Bayesian data augmentation using MCMC: Application to missing values imputation on cancer medication data

By

Thamsanqa Innocent Ndlela

Supervisor : Dr Siaka Lougue

A thesis submitted in fulfillment of the requirement for the Masters Degree in Statistics

School of Mathematics, Statistics and Computer Science

University of KwaZulu-Natal

Westville Campus

South Africa

Declaration

I, Thamsanqa Innocent Ndlela, declare that this thesis titled, 'Bayesian data augmentation using MCMC: Application to missing values imputation on cancer medication data' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work for others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Mr. T.I Ndlela Signed_____ Date_____
Dr. S. Lougue Signed_____ Date_____

Abstract

Missing data is a very serious issue that negatively affect inferences and findings of researchers in data science and statistics. The ignorance of missing data or deletion of cases that contain missing observations may lead to reducing statistical power, loss of information, increasing standard errors of estimates and increases estimation bias in data analysis. One of the advantages of using imputation methods is to keep the full sample size, which makes the results to be more precise. Amongst all the missing data imputation techniques, data augmentation is not so popular in the literature and very few articles mentioned the use of the technique to account for missing data problems. Data Augmentation technique can be used for imputation of missing data in both Bayesian and classical statistics. In the classical approach, data augmentation is implemented through EM algorithm that uses maximum likelihood function to impute and estimate unknown parameters of a model. EM algorithm is a useful tool for a likelihood-based decision when dealing with missing data problems. The Bayesian data augmentation approach is used when it is not possible to directly estimate a posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{x}_{ov})$, of the parameters, $\boldsymbol{\theta}$ given the observed data $\mathbf{x}_{ov}$ due to the missing data in $\mathbf{x}$. This study aims to contribute to a better understanding of Bayesian data augmentation and improve the quality of estimates and precision of the analysis of data with missing values. The General Household Survey [GHS 2015] is the main source of data in this study. All the analyses are made using the software **R** and more precisely the package ***mix***. In this study, we have find that Bayesian data augmentation can solve the problem of missing data in cancer drug intake data. The Bayesian data augmentation performs very well in improving modelling of cancer drug affected by missing data.

Keywords: *Data augmentation, EM algorithm, Missing data, Cancer, Bayesian approach*

Acknowledgements

My sincere gratitude goes to my supervisor, Dr. Siaka Louge, for his wisdom and guidance. His supervision opened up the fascinating world of statistics and supported my enthusiasm to learn. I wish to thank the University of KwaZulu-Natal postgraduate bursary for funding my research, without your support this thesis would not be possible. I would also like to thank my friends, family and my girlfriend, for their continuing support and encouragement throughout the last few years and my life.

Contents

List of Figures

List of Tables

Glossary of Acronyms

AIC	Akaike Information Criterion
CI	Confidence Interval
DA	Data Augmentation
EM	Expectation Maximization
GHS	General Household Surveys
IPF	Iterative Proportional Fitting
LOCF	Last Observatio Carried Forward
MAR	Missing At Random
MCAR	Missing Complete At Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
ML	Maximum Likelihood
MLE	Maximum Likelihood Estimates
MNAR	Missing Not At Random
OLS	Ordinary Least Squares
OR	Odds Ratio
SE	Standard Error
SEM	Structural Equation Models

Chapter 1

Introduction

1.1 Background

Missing data is a very serious issue that negatively affects the findings and inferences of researchers in data science and statistics, such as in operation management (Tsikriktsis, 2005), psychology (Graham, 2009) and epidemiology (Cattle et al., 2011) amongst others. Most researchers and analysts are interested in making inferences based on the whole target population rather than a sample that is not representative of the population. The consequence of ignoring of missing data or deletion of cases that contain missing values is that, it may reduce the sample size, statistical power, loss of information, thereby increasing standard errors of estimates and increases estimation bias, especially when the proportion rate of that missing data is high (De Leeuw et al., 2003; Enders, 2010; McKnight et al., 2007). To handle missing data problems in the given data set, mostly investigators use imputation methods to impute missing values in the data set. The benefit of using imputation methods keeps the full sample size n, which makes the results less biased and precise. Missing data is common in many data sets and cannot be avoided in data based research, even if great effort is put into planning and data collection plan (Allison, 2002; Regoeczi & Riedel, 2003; Rudas, 2005; Stumpf, 1978).

The analysis of missing data leads to a basic theory of missing data problem on how does missing data affect the inference and prediction on the data sets. A traditional way of handling missing data is to discard them from the analysis; most of the statistical software packages provide this method by default such as SPSS, SAS and STATA. The use of traditional method may substantially influence and affect data analysis especially when dealing with a sample with a high proportion of

missing data. On reviewing the literature on missing data, different methods have been developed to handle missing data (Little & Rubin, 2002; Schafer & Graham, 2002; Tsikriktsis, 2005). It means that the results, without using imputation in this case, are less appropriate to make a conclusion about a data. Once all missing data have been imputed, it allows the Researchers to analyse the data set, using standard methods for complete data. There are many standard methods used to deal with missing data problem, and these use the different forms of imputation technique or algorithm such as multiple imputations, expectation-maximization (EM) algorithm and other imputation methods. The extent of damages or misleading results, caused by missing data depends on the percentage of missing data in the data sample, although there is no fixed rule as to how much of the amount of missing data allowed. Cohen (1983) state that 5% to 10% of missing data is considered to be low or small and high if at least 40% of data is missing (Raymond & Roberts, 1987).

Before tackling missing data, Researchers need to know the reason why data is missing in the dataset (Carpenter & Kenward, 2012; Little, 1988a). In general, there are several reasons that categorise data to be missing in the data set. These reasons include one of these listed reasons:

- data entry errors.
- failure to complete the whole questionnaire.
- refuse to respond to certain questions in the survey such as income level, etc.

It is important to consider the mechanism and item nonresponse in terms of how it should be treated in the analysis (de Leeuw & Huisman, 2003). Missing data mechanisms need to be one of these mechanisms : missing completely at random (MCAR), missing not at random (MNAR) or missing at random (MAR). These are all methods that are used to handle missing data. The pattern of data loss is assumed to be random (either MCAR or MAR). Identifying the underlying missing mechanism is one of the important things in dealing with missing data problems since it gives information on how missing data should be tackled. MCAR data have been found to be less likely to produce serious bias in the parameter estimates, regardless of the methods used to handle missing data (Graham, 2009; Musil et al., 2002), while MNAR data is difficult to identify and handle because true values of missing values are unknown (Little & Rubin, 2002).

In the past years researchers found the influential paper based on the development or construction of MCMC data augmentation. The Markov Chain Monte Carlo

method is defined as stochastic version of expectation-maximization (EM) algorithm. The expectation-maximization (EM) algorithm is the most general method for tracing missing observations by finding maximum likelihood estimates (MLEs) of the parameter that we estimate (Dempster et al., 1977; Meng & Van Dyk, 1997; Krishnan & McLachlan, 1997). The EM algorithm is a useful tool for a likelihood-based decision where we are dealing with missing data problems. However, there are challenges using maximum likelihood to make inferences, such as finding their standard errors in situation with many parameters (Meng & Rubin, 1991), i.e if the data contains nuisance parameters and having a small size where the asymptotic theory of MLE may not hold. The EM algorithm was more desirable before statisticians found its limitations specifically when the sample size is small. They then introduced the more efficient Bayesian approach method usually called a data augmentation algorithm for incomplete data analysis. This type of algorithm can overcome the limitations of EM and it give more precise estimates, especially when the given sample size is very small (Tanner & Wong, 1987). The Bayesian data augmentation algorithm may use EM-algorithm as a starting point in the algorithm first step.

The data augmentation (DA) is the term that was introduced by Tanner & Wong, 1987 as the estimation procedure for dealing with the unobserved values by completing the observed data with missing data. The data augmentation is an MCMC procedure that improves the quality of data by adding the missing data in the model. Very few studies implement a Bayesian data augmentation techniques framework to simulate missing observations or to solve the problem of missing data. This approach allows for sampling both the missing data and model parameter from their posterior full conditional distributions (Tanner & Wong, 1987). The data augmentation algorithm provides the way of improving the quality of data and inferences, especially powerful when the sample size n is small. The basic idea of data augmentation algorithm is to extend the observed data $\mathbf{x}_{ov}$ by adding the missing data $\mathbf{x}_{mv}$. If both $\mathbf{x}_{ov}$ and $\mathbf{x}_{mv}$ are known, we can now compute or sample from the extended posterior distribution $P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv})$. The Bayesian data augmentation approach can be implemented where we are not able to directly simulate or determine a posterior distribution $P(\theta \mid \mathbf{x}_{ov})$, of the parameters, θ given the observed data $\mathbf{x}_{ov}$ due to the missing data in $\mathbf{x}$. The mutual dependency between $P(\theta \mid \mathbf{x}_{ov})$ and $P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov})$ leads to determining an iterative algorithm to calculate the posterior distribution $P(\theta \mid \mathbf{x}_{ov})$, which is the method of successive substitution of the functional analysis for solving an operator fixed point equation and it can be used with Monte-Carlo approximation in each substitution step. In the past years, Researchers introduce this approach to augement observed and unobserved data so that it makes it more

easy to analyse data with latent values using Bayesian approach (Tanner & Wong, 1987).

Despite the popularity of Bayesian data augmentation, but there is still no clear guidance on this method based on the performance of missing values imputation when the imputation is done on explanatory variables for both continuous and categorical type containing missing data. This study attempted to evaluate the performance of this method when data was; missing at random, missing completely at random or not missing at random and based on the proportion rate of missing data. Also in this study we considered the case where missing values are observed on the explanatory variables (binary variables). The ignorability assumption of missing data was applied throughout this thesis. This assumption is based on the observed data, so that we can be able obtain the estimates of missing values. Therefore, the assumptions of MCAR, MAR were only considered (only random). The assumptions above are applied to investigate whether or not the missing data mechanisms have an impact on the performance of Bayesian data augmentation technique under different proportions of missing data. The performance of Bayesian data augmentation is assessed when missing values are missing at random, completely missing at random or not missing at random under different proportion rate of missing data on the variables of interest.

1.2 Statement of the problem

The General Household Survey (GHS) data of South Africa is used to make decisions about a population under Cancer-Medication study. However, the use of survey data is limited by the following problems:

- The small sample size and high percentage of missing data approaches produce inadequate results according to Schafer & Olsen (1998).
- If we obtain invalid parameter estimates and it may lead to inappropriate inferences.

There are several approaches that have been proposed for dealing and analysing survey data that is affected by missing data. Faced with this form of problem, how then, can a researcher handle cases with missing data without resorting to include only cases with complete information and deleting those with missing data deletion (default in statistical packages). The most common way of tackling missing data problem is to ignore those cases from the analysis. This method is referred to as case deletion method or complete case analysis and can lead to lower statistical power

and biased parameter estimates when the proportion rate of missing data is too high (Graham, 2009). How can missing data be handled to obtain best results so as not go to the risk of getting biased, skewed and sometimes misleading inferences?. The data augmentation approach is the most useful method if sample size of the data is small and missing percentage is high according to (McKnight et al., 2007). There are several methods proposed to handle or impute missing data instead of deleting missing values to keep the sample size constant such as single imputation methods, multiple imputations and model-based methods. The data augmentation algorithm was introduced in the form of multiple imputation concept using Markov Chain Monte Carlo framework.

This study used a 2015 General Household Survey (GHS) data of South Africa data set, this is a complicated survey with a complex sampling design and weighting procedure that needs to be used during the analysis. There are many studies that show that when the survey data sets contain weight variables, weighted results are chosen because they give less bias in the estimates than unweighted results because usually sampling is not done in proportion to the size of the particular region, but rather equally across all regions, this can lead to certain survey characteristics being over or under represented in the sample (Korn & Graubard, 1995). Thus, the use of the weight correct this. The performance of the data augmentation is evaluated by comparing the baseline data set and imputed data sets under different proportion rate of missing data and missing mechanisms.

1.3 Aim and objective

In real life, data is highly affected by the problem of missing data. The collection of household data about cancer is a very difficult task, which may lead to missing data problem. In this study we focus on data augmentation as one of the imputation technique that can be used to improve data quality of cancer medication intake in South Africa that is affected by missing data. This method of imputation is not so popular in both classical and Bayesian statistics framework to handle missing data problem. The classical methods like EM algorithm and others fail to provide precise estimates since the asymptotic theory of MLEs are sometimes not satisfied. The main aim of this Research is to investigate how Bayesian data augmentation contributes to handling the incomplete data analysis problem under different missing mechanisms such as MCAR, MAR, and MNAR. Moreover, the focus of this study is to evaluate how using the Markov Chain Monte Carlo (MCMC) Bayesian data aug-

mentation technique improves the quality of data, with missing data depending on the nature of the missing (MCAR, MAR, and MNAR). According to Tanner & Wong (1987), this method tends to improve the statistical power of the data analysis especially if the sample size is small such as with cancer medication data. The primary objective of this study was to determine the performance of data augmentation (DA) method when the data are MCAR, MNAR and MAR that are separated into various rates of missing proportions. This study explores performance of data augmentation technique by fitting a Survey logistic regression model and assessing the analysis for each model (MCAR, MAR, MNAR).

The other specific objectives are:

1. To compare the theoretical performance of data augmentation in handling missing covariates.
2. To investigate whether the proportion rates of missing data in the cancer medication data can impact on the performance of MCMC data augmentation algorithm.
3. To explore the performance of data augmentation technique under different models (MCAR, MNAR and MAR) with different rates of missingness.

Throughout the thesis, statistical analyses are done using both R and SAS software. The package **MIX** in R to impute missing data (Data augmentation) and proc survey logistic procedure in SAS is used to fit the model, check adequacy, determine parameter estimates and so on.

1.4 Significance of study

Cancer is a serious pandemic that affects a large portion of the population in the world and especially in the developing countries such as South Africa. An important goal is that Medical Health institutions around South Africa distribute cancer medications efficiently to all patients who are affected by cancer. Therefore, it is critical to study how people in South Africa respond in order to control or monitor this disease. This study is important in terms of investigating the factors that affect people during the process of controlling cancer. The collection of Household Survey [GHS] data about cancer is a very difficult task, since respondents are not at liberty to disclose the information about cancer matters. Missing data is difficult to avoid in the household surveys [GHS data] and particularly affected by sensitive questions in general. A data set with missing data, may lead to inadequacy of results and false conclusions in any statistical analysis. Hence, the use of data augmentation to handling missing data that is supported by statistical theory can enhance the accuracy of the estimation of cancer-medication intake in South Africa. Empirical applications of Bayesian data augmentation started to grow rapidly after data augmentation was introduced by Tanner & Wong (1987). The data augmentation method is a Bayesian approach to replace the EM algorithm for incomplete data analysis. Data augmentation is the Bayesian method and is used as one of the multiple imputation methods. The findings from this study will help Research and Medical Health institutions in South Africa to assess the use of cancer-medication. That however depends on the quality of dataset. Therefore, fitting a survey logistic regression model without taking missing data into account may produce biased estimates of parameters and that leads to incorrect decisions.

1.5 Limitations of study

In this study, there are many possible factors that can affect whether or not cancer patients medication (e.g. the stage of illness, the type of cancer etc.) but the information of these factors was not collected and therefore is unavailable to be used in this study.

1.6 Research questions

This study is trying to answer the following questions about the MCMC data algorithm and imputation of missing data in cancer-medication data set:

1. What is the performance of MCMC data augmentation in terms of regression estimates and standard errors when data hold under different missing mechanisms which is MCAR, MNAR or MAR different rates of missingness in the Survey Logistic Regression.
2. What is performance of MCMC data augmentation in terms of Log likelihood ratio and AIC when data are MCAR, MNAR and MAR with different rates of missingness compared to original data.
3. Does MCMC data augmentation improve the quality of analysis when data are MCAR, MNAR and MAR with different rates of missingness in the survey logistic regression.

1.7 Hypotheses

The following hypotheses were tested:

1. The Bayesian augmentation techniques, yield similar parameter estimates for both categorical and continuous variables containing missing data.
2. The performance of Bayesian data augmentation is not affected by the different missing mechanisms (MCAR, MAR, MNAR).
3. Missing data patterns have no impact on the performance of Bayesian augmentation technique under different proportion rate of missing data.

1.8 Thesis outline

The thesis is organized in 6 chapters as follows

Chapter 1: This chapter is the introduction to this thesis.

Chapter 2: represents the literature review of data augmentation and missing data problems. It defines missing data and the key assumptions such missing data mechanisms (MCAR, MAR, MNAR), data patterns and observed-data likelihood. The methods of imputation are reviewed. The Single/Traditional imputations that were reviewed are: Listwise, Pairwise deletion, Mean method, Hot-Deck method, Cold-Deck method and Regression method. The imputation-based methods that were considered are: Maximum Likelihood, Expectation Maximization (EM) and Markov Chain Monte Carlo (MCMC) technique. Also, it discusses previous work to missing data and work related to data augmentation.

Chapter 3: This chapter gives a detailed explanation of the data; the source, the variables and the sampling design. In addition, some basic exploratory analyses to explore relationships between the variables are presented. Then, discusses the data augmentation background, how it works and advantages and disadvantages of this algorithm and also reviews convergence properties and rate of convergence of data augmentation. It also, discusses model-based methods that are classified into two, which is: Non-Bayesian data augmentation and Bayesian data augmentation. It gives the basic theory of data augmentation algorithms models under the methods for multivariate normal, cross-classified categorical data and mixed data.

Chapter 4: This chapter contains the design of the application study based on Cancer Medication data and offers a description of the dataset based on the complete data and incomplete data under different percentages used in the analysis. We investigate the performance of the data augmentation algorithm in handling incomplete Cancer-Medication data under different missing data mechanisms (MAR, MCAR and MNAR) and different missing percentages of data. The simulation study including the design, data generation and evaluation criteria used in the analysis are presented. In particular, we fit the Binary Survey Logistic Regression model on the imputed data sets for different percentages and compare the results with the complete data. Discussion of results obtained from a complete case analysis and a Bayesian data augmentation is presented.

Chapter 5: gives the general conclusion of the thesis. The findings are summarized and remarks of the thesis. Finally, we present references list at the end of the thesis.

Chapter 2

Literature Review on missing data

2.1 Review of the Fundamental Research

2.1.1 Theoretical Foundation for the study of missing data

Terminology

The Missing data problem is not easily avoided in the data set and usually occurs at two levels: either at the unit level or the item level. A unit-level nonresponse occurs when the respondent refuses to reveal or give information that is collected by interviewer. For example, a respondent may refuse to answer the survey questions, or does not show up to do a survey. This type of situation, termed unit nonresponse, is the subject of many studies in the literature, in various fields, documented by work on selection bias in econometrics (Heckman, 1976; Berk, 1983) and much previous work has been done in statistics (Rja & Rubin, 1987; Little & Rubin, 1989; Little, 1988b). While the unit non-response is a common problem to solve, it is not the focus of this study (Little & Rubin, 1989). This study only focuses on the problem of the item non-response. An item nonresponse refers to the incomplete data collected from a respondent. For example, when a respondent may have not answered some of the questions on a survey, but answered the rest of the questions. The problem of missing data at the item level can be solved by looking at three different aspects: the proportion of missing data, the missing data mechanisms, and patterns of missing data. Each aspect is being discussed below.

Define missing data

The relationship between missingness of the data and the observed data of the variables in the matrix data is described by missing data mechanisms, i.e. to determine whether the missingness depends on or is independent of the underlying values of the variables in the given data set. Little & Rubin (1989) name this process as "mechamism of missingness". This process is interpreted as the probability distribution of the missing data (Little & Rubin, 2002, 2014; Allison, 2002). There are three different types of missing mechanisms, which is missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Let $\mathbf{x}$ denote the matrix data set with n rows and k columns. Since the matrix $\mathbf{x}$ has a missingness, the matrix must be partitioned into two parts, which is the observed data $\mathbf{x}_{ov}$ and missing data $\mathbf{x}_{mv}$ i.e. $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$. Now let $\mathbf{x}_{ov} = (x_{ij,ov}, i = 1, 2, 3, ..., n)$ and $\mathbf{x}_{mv} = (x_{ij,mv}, j = 1, 2, 3, ..., k)$. Now according Rubin (1976), we can define the *missing-data indicator* $\mathbf{R} = r_{ij}, i = 1, 2, ..., n, j = 1, 2, ..., k$ with $n \times k$ dimensions and binary indicators:

$$r_{ij} = \begin{cases} 1 & \text{if } x_{ij} \text{ is missing} \\ 0 & \text{if } x_{ij} \text{ is is observed} \end{cases}$$

We treat the indicators r_{ij} as the observation of a random variable with an underlying distribution. We can model the missing data indicators distribution conditional on $\mathbf{x}$, $P(\mathbf{R} \mid \mathbf{x}, \varphi)$ where φ are vector of the unknown parameters.

Proportion of missing data

The proportion rate of missing data or percentage of missing data is directly connected to the quality of statistical inferences under statistical framework. There is currently no literature providing a guidline of an acceptable percentage of missing information in a data set in order to make valid statistical inferences in the analysis. Some examples of this is reviewed in the following literature: (J. L. Schafer & Olsen, 1998; J. L. Schafer, 1999) establishes that a proportion of missing rate of 5% or less is insignificant, thus it has no big influence in the statistical inferences. Bennett (2001) argues that the analysis is more likely to produce biased estimates when the data set contain more than 10% missing information. However, the rate of missing data, is not the main reason that researchers use to assess the missing data problem.

Whereas, Tabachnick et al. (2001) assume that the missing data mechanisms and the missing data patterns have more influence on research outcomes than the proportion rate of missing data.

Missing data mechanisms

According to D. B. Rubin (1976), there are three mechanisms under which missing data can occur: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR).

Missing Completely at Random (MCAR)

The data is said to follow the MCAR missing mechanism if the missingness is independent or unrelated to any values observed or missing in the data set. The conditional distribution can be expressed as follows:

$$P(\mathbf{R} \mid \mathbf{x}, \varphi) = P(\mathbf{R} \mid \varphi) \quad \forall \mathbf{x}, \varphi \tag{2.1}$$

Therefore, the above missing mechanism is called Missing Completely At Random or MCAR. If missing data holds under MCAR assumption, it is defined as a random sample of the complete data. Ignoring missing data under MCAR will not introduce bias, but will increase the Standard Error (SE) of the sample parameter estimates due to decrease in sample size.

Missing at Random (MAR)

A type of missing data mechanism that is considered having weaker assumption (J. L. Schafer & Graham, 2002) is that of missing at random (MAR). A data set $\mathbf{x}$, then data on $\mathbf{x}$ said to hold under the assumption of missing at random(MAR) if the missingness only depends on the observed, $\mathbf{x}_{ov}$, but distinct to the missing, $\mathbf{x}_{mv}$ (Allison, 2000). The missing data at random (MAR) can be expressed as follows

$$P(\mathbf{R} \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}, \varphi) = P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi) \quad \forall\, \mathbf{x}_{mv}, \varphi \tag{2.2}$$

This expression shows that the missing data on $\mathbf{x}_{mv}$ does not depend on $\mathbf{x}_{mv}$ itself, but only on $\mathbf{x}_{ov}$ and the parameter φ. If the joint prior distribution $P(\theta \mid \varphi) = P(\theta)P(\varphi)$, the inference of θ under Bayesian can be obtained by using observed data-

likelihood $P(\mathbf{x}_{ov} \mid \theta)$. The missing is considered as ignorable under this assumption. Many Advanced or Modern missing data techniques (e.g., EM, MI, MCMC) assume MAR always. The assumption violation is always expected in most cases (J. L. Schafer & Graham, 2002). Fortunately, previous work found that the violation of the MAR assumption does not influence parameter estimates that much (Collins et al., 2001). However, MAR assumption is more possible when data are unobserved by design. Under Bayesian data augmentation (MCMC) it only works effectively under assumption of ignorability (M. A. Tanner & Wong, 1987a; J. L. Schafer, 1997).

Observe-data likelihood

According to D. B. Rubin (1976); Little (1988a) and Little & Rubin (1987), if we use MAR and distinctness assumption it allows us to re-write the observe data likelihood as follows

$$P(\mathbf{x}_{ov}, \mathbf{R} \mid \theta, \varphi) = \int P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi) P(\mathbf{x}_{ov} \mid \theta) d\mathbf{x}_{mv} \tag{2.3}$$

$$P(\mathbf{x}_{ov}, \mathbf{R} \mid \theta, \varphi) = P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi) \int P(\mathbf{x}_{ov} \mid \theta) d\mathbf{x}_{mv} \tag{2.4}$$

$$P(\mathbf{x}_{ov}, \mathbf{R} \mid \theta, \varphi) = P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi) P(\mathbf{x}_{ov} \mid \theta) \tag{2.5}$$

Under MAR, the likelihood of the observed data can be separated into two partition and one of partition contains unknown parameters of θ and the other partition contain nuisance parameters φ. When the parameters φ and θ are distinct, then an inference about parameter θ based on likelihood estimation will not be affected by nuisance parameter φ or $P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi)$. According to Little & Rubin (1987) the likelihood function can be obtained by ignoring the missing data mechanism

$$L(\theta \mid \mathbf{x}_{ov}) \propto P(\mathbf{x}_{ov} \mid \theta) \tag{2.6}$$

This likelihood function is referred as observed data likelihood, where we assume Ignorability.

Missing Not at Random (NMAR)

If the missingness in the data cannot be assumed to be neither MCAR or MAR and the conditional distribution of $\mathbf{R}$ depends on the missing data $\mathbf{x}_{mv}$ even after including observed data. Thus, mathematically can be expressed as follows :

$$P(\mathbf{R} \mid \mathbf{x}, \varphi) = P(\mathbf{R} \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}, \varphi) \quad \forall \mathbf{x}, \varphi \tag{2.7}$$

Therefore, the data said to be NMAR. Under this assumption, the missing data mechanism cannot be ignored and it must be included in the imputation model (Little & Rubin, 2002).

Ignorability

The missing data mechanism is considered as ignorable if the given data satisfies the MAR assumption and the parameters that control data missing method are the parameters in the model to be predicted or estimated (McKnight et al., 2007). Also the level of percentage of missingness should not be too high.

Distinctness of parameters

The parameter, θ of the data model and the nuisance parameter φ of the missing data mechanism are distinct, if it can be separated into two products of the joint parameter space(θ, φ). It is the product of the parameter space of θ and the product of the parameter space of φ (Schafer, 1997). The missing data mechanism is said to be ignorable, if they are both MAR and distinctness (Little $\&$ Rubin, 1987).

The observed-data posterior

The inferences in the Bayesian point of view are based on the posterior distribution of the model. The posterior distributions are the distribution of the unknown parameters conditional to observed quantities i.e $P(\theta, \varphi \mid \mathbf{R}, \mathbf{x}_{ov})$. The parameters of the model are (θ, φ) and $(R, \mathbf{x}_{ov})$ are the observed quantities.

Using the Bayes Theorem, now we can obtain the posterior distribution:

$$P(\theta, \varphi \mid \mathbf{R}, \mathbf{x}_{ov}) = \frac{P(\mathbf{R}, \mathbf{x}_{ov} \mid \theta, \varphi)P(\theta, \varphi)}{\int\int P(\mathbf{R}, \mathbf{x}_{ov} \mid \theta, \varphi)P(\theta, \varphi)d\theta d\varphi} \tag{2.8}$$

where $P(\theta, \varphi)$ is the joint prior distribution and the normalizing constant is given by the denominator of the quantity (2.9) , which is

$$\int\int P(\mathbf{R}, \mathbf{x}_{ov} \mid \theta, \varphi)P(\theta, \varphi)d\theta d\varphi \tag{2.9}$$

Assuming the MAR missing mechanism, then this assumption allows to substitute (2.5) into (2.8) to obtain

$$P(\theta, \varphi \mid \mathbf{R}, \mathbf{x}_{ov}) \propto P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi)P(\mathbf{x}_{ov} \mid \theta)P(\theta, \varphi) \tag{2.10}$$

Under the Bayesian framework we always make decisions based on the parameter θ alone, so to compute the marginal posterior based on θ, we integrate the posterior distribution (2.10) with respect to nuisance parameter φ. When the parameter θ and nuisance φ are independent to each other, the prior distribution can be factored as follows

$$P(\theta, \varphi) = P_\theta(\theta)P_\varphi(\varphi) \tag{2.11}$$

Under ignorability assumption, the marginal posterior for θ alone can be obtained by

$$P(\theta \mid \mathbf{x}_{ov}, \mathbf{R}) = \int P(\theta, \varphi \mid \mathbf{R}, \mathbf{x}_{ov})d\varphi \tag{2.12}$$

$$P(\theta \mid \mathbf{x}_{ov}, \mathbf{R}) \propto P(\mathbf{x}_{ov} \mid \theta)P_\theta(\theta) \int P(\mathbf{R} \mid \mathbf{x}_{ov}, \varphi)P_\varphi(\varphi)d\varphi \tag{2.13}$$

$$P(\boldsymbol{\theta} \mid \mathbf{x}_{ov}, \mathbf{R}) \propto L(\theta \mid \mathbf{x}_{ov})P_\theta(\theta) \tag{2.14}$$

Thus, under the ignorability assumption, the posterior distribution of the parameter θ is given by

$$P(\boldsymbol{\theta} \mid \mathbf{x}_{ov}) \propto L(\boldsymbol{\theta} \mid \mathbf{x}_{ov})P_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \tag{2.15}$$

where $P(\boldsymbol{\theta} \mid \mathbf{x}_{ov}, \mathbf{R}) = P(\boldsymbol{\theta} \mid \mathbf{x}_{ov})$ since $\mathbf{R}$ does not appear on the quantity (2.15) right

side. Now, equation (2.15) must referred as the observed-data posterior.

Missing data patterns

The statisticians frequently encounter missing-data problem in real data, due to different data situations and inefficient data collection procedures that are being implemented during the collection of data. The missing data is a serious problem in most surveys across all field of study. In Survey Methodology, this would mainly be caused by item non-response, in Biology this can be caused by impure samples used. There are some different methods for analysis which are intended for missing data pattern. A missing data patterns are very useful in missing-data problem because it gives the actual location of the missing data observations in the incomplete data matrix **x**. It is also useful in a choice making based on which imputation method to use in the given data and variables types that is imputed also depends on missing data patterns. The missing data patterns are divided into monotone and arbitrary missing patterns.

Monotone missing pattern:

A data set matrix **x** with variables $x_1, x_2, .., x_k$ is said to have a monotone missing pattern when the variable x_j is missing for a specific unit jth cell, which implies that all the following variables $x_{j+1}, x_{j+2}, ..., x_l$, where $l > j$ are missing for that specified unit cell in the model (Institute, 2005).

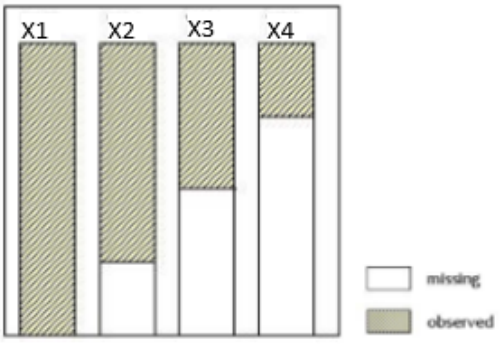

Figure 2.1 – Monotone missing-data pattern

The monotone missing pattern usually occurs more under the longitudinal studies,where, if the interviewee fails to answer at one point, then data is missing on subsequent measures.

Non-monotone missing pattern:

The dataset is said to be non-monotone missing data pattern if we can not re-order the incomplete variables in the matrix **x**, such that the variable x_j is missing for a specific unit j^{th} cell, which implies that not all the following variables $x_{j+1}, x_{j+2}, ..., x_l$, where $l > j$ are missing for that specified unit cell in the model.

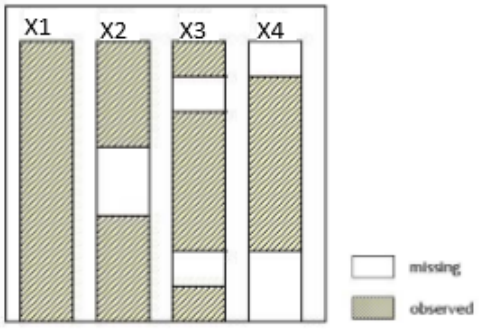

Figure 2.2 – Non-monotone missing-data pattern

An arbitrary missing pattern data set is a data set that is neither a monotone missing pattern nor a non-monotone missing pattern.

2.1.2 General Review for existing Simple missing data methods

In analysing the problem of missing data in the given dataset, a number of methods have existed in Statistics framework to tackle this problem in past decades. The methods of imputations can be classified into two types i.e Simple or Traditional and Advanced techniques. The Simple imputation methods includes the deletion methods, and Single imputation techniques (D. B. Rubin, 1997). This technique is called Simple because it can easily be used in standard statistical software (for example SAS, R, Stata and others). The problem with Simple approaches is that they are acceptable only if dealing with small percentage of missingness but is not advisable to be used when the missing data percentage is very high. The advanced techniques are the likelihood based methods and including Multiple imputation. In general, the simple methods have disadvantage over advanced methods in terms of imputation accuracy.

The Simple/Traditional methods

The Deletion methods

- **Listwise deletion**
  The Listwise or case deletion is the most common used approach by statisticians to dealing with missing data problem by excluding all rows with missing data and remaining with observed data only. Under this approach, analysis is only based on the observed data only. This approach is well known as the complete case analysis or case deletion. The advantage of using Listwise deletion approach is that the remaining dataset is complete. Thus, this complete data has a reduced sample size and power, caused by excluding rows with missing data. The method of Listwise deletion is the most commonly used to handle missing data problem (Peugh & Enders, 2004), and thus most of statistical software packages set this approach as a default option for analysis. The Listwise deletion sacrifices a large amount of data (Malhotra, 1987; Stumpf, 1978). The loss of large data will reduce the statistical power (Gilley & Leone, 1991) and may reduce the precision of the estimated parameters (Donner, 1982). Thus, when the data hold under missing at random (MAR), the type II error rates may be inflated (Raymond, 1986). Researchers find that using this method may introduce bias in the estimation of the estimates of the model parameters. Under the assumption of MCAR, a Listwise deletion approach produces an unbiased estimate and precise results of the model. When the given data do not hold under the assumption of MCAR, the method may produce bias in the estimates of parameters. The Listwise deletion is reasonable to be used if the sample is large enough, no problem with power and hold under the assumption of MCAR. The Listwise is not an optimal strategy to be implemented if the above assumptions listed above are violated. However, this method in statistics is still applied frequently in several fields of research in both medical and epidemiological studies.

- **Pairwise deletion**
  The pairwise deletion is a more enhanced version of Complete-Case Analysis: Use all available data in a more efficient way, to estimate the parameters of interest in the model. This method is well known by Available-Case analysis (AC). This method helps researcher to be able to examine the covariance matrix of p number of variables containing variance estimates and means based on the observed data of each variable. According to Allison (2001) regression model estimates under MAR based on Available-Case analysis can be seriously biased (unlike model estimates based on Complete-Case analysis). Kim & Curry

(1977) argue that by using Available-Case estimates can be improved instead of using complete case, others A. B. Anderson, Basilevsky, Hum, et al. (1983); A. B. Anderson, Basilevsky, & Hum (1983); Haitovsky (1968); Little (1992); Little (1988a)) give us more information about the problems of Available-Case Analysis method. The problems associated with pairwise deletion implementation are that the estimated correlations may lie outside the acceptable range $(-1, 1)$ and that the R^2 may be less than 0 or larger than 1 (Raymond, 1986; Cohen et al., 1983) and also include Little (1988a). When the data hold under MCAR missing mechanism, the remaining observations are the originally identified data set. The available case analysis provides consistent estimates (the correct point estimates, see e.g., Cox & Hinkley (1979)) shown by Little (1992), especially if variables are moderately correlated in regression models. Thus, if variables are highly correlated with each other, pairwise deletion method provides estimates that are inferior to listwise deletion results, shown in Haitovsky (1968) literature.

Single Imputations

The single imputation method is another method of imputing missing values in the data set with a missing data problem. The advantage of using this approach is that it keeps the sample size constant unlike the complete case analysis mentioned previously. The single imputation approach does not discard missing values. The inference will be based on the imputed data set using the statistical tools as in the full data set case. These methods are fully explained by Little & Rubin (2000) and J. L. Schafer & Graham (2002) to review them .Imputation methods are normally used to fill in the missing observations of the variable. Imputation of missing values may have advantages and disadvantages, the problem with imputation is that it may lead to affect the biasness of the parameters such as standard deviations by taking the imputed observations as if they were real observations from the original data set.

Single imputation methods:

Mean method: Mean imputation (unconditional) method takes the average of all the present values in that variable and uses it to fill in the missing data. Mean method is regarded as a simple and straight forward imputation technique, for instance let say the three items are missing from the ten item-scale, then the mean of the seven item would be calculated and used to impute for the three missing items. This technique is normally not acceptable to deal with missing data and another very strong disadvantage is that it underestimate the population variance (Fayers et al., 1998; Song & Shepperd, 2007)

Hot-Deck method: Hot-Deck imputation method is the continuation of cell mean imputation. This method actually takes the observed value given by other respondents to impute the missing value for another respondent. For example, if sex, race and level of education have been completed but age is missing, a random respondent with the same sex, race and level of education is chosen from the respondents who match, and that respondent's age is entered for the missing data (Andridge & Little, 2010).

Cold-Deck method: Cold deck imputation just uses the information from other external sources to impute the missing values. These values can be constructed with the use of historical data, subject-matter expertise etc. For example, from the previous data a particular participant provided information about his education level but for the current data same participant did not provide the information about his education level, so Cold deck imputation method uses participant's previous information to impute for the current one. This technique is not trusted in reducing biasness form the dataset (Lohr, 2009).

Regression method: Regression imputation is also known as conditional mean imputation, which substitutes the missing values with predicted scores from a regression equation. It has the similar disadvantage to mean method, it also underestimate the population variance but produces better results compared to mean imputation. Linear regression is used if the variable is continuous, a logistic regression if the variable is binary and a Poisson regression if the variable is a count variable, etc (Song & Shepperd, 2007).

Last observation carried forward (LOCF):
The Last observation carried forward (LOCF) is a well-known imputation method

that is commonly used by most of the researchers to tackle the problem of missing data in the dataset. This approach works as follows: Whenever a data is missing, all missing values are replaced by the last observed value within the same subject. This approach assumes that the results would not have changed from the last observed value. This assumption is material and an inappropriate assumption can influence the validity of this method negatively. LOCF method is used the most because of its simplicity to use, but other researchers reveal there is strong evidence not to use it all the time. The method may introduce bias in the results according to (Molnar et al., 2008). Molenberghs & Kenward (2007) stated that LOCF method can be applied to both patterns of missing data (monotone and non-monotone). The literature of Craig et al. (2003) and Siddiqui & Ali (1998) has more information about the method, and also Molenberghs & Kenward (2007) provide the review about the issues of this method.

2.1.3 Review for existing advanced missing data techniques

Model-based approach

The data analyst and researchers most of the time uses model-based approaches when trying to determine the unknown parameters $\boldsymbol{\theta}$. The model-based approach is when the likelihood functions is used to find inferences for unknown parameters that are given as $\boldsymbol{\theta}$. Let the vector $\mathbf{x}$ to be a complete data (with no missing data), then the likelihood function for unknown parameters $\boldsymbol{\theta}$ given $\mathbf{x}$ is specified by assuming the model to be

$$L(\boldsymbol{\theta} \mid \mathbf{x}) \propto P(\mathbf{x} \mid \boldsymbol{\theta}) \tag{2.16}$$

However, if there is a missingness in original data $\mathbf{x}$, let's try to consider those missing data in the model so that one can obtain precise and unbaised estimates. To model such data, let's assume that the missing data mechanism is MAR (denoted by $P(R \mid \varphi, \mathbf{x})$. The MAR allow to form a model for full likelihood function with the joint parameters $\boldsymbol{\theta}$ and φ given observed data $\mathbf{x}_{ov}$ and missing data indicator $\mathbf{R}$ (D. B. Rubin, 1996),

$$L(\boldsymbol{\theta}, \varphi \mid \mathbf{x}_{ov}, R) \propto \int P(\mathbf{x}_{ov}, \mathbf{x}_{mv} \mid \boldsymbol{\theta}) P(\mathbf{R} \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}, \varphi) d\mathbf{x}_{mv} \tag{2.17}$$

by assuming that the parameters $\boldsymbol{\theta}$ and φ are not identical in nature. To specify for Bayesian framework, let's start by including prior knowledge distributions on the distinct parameters $\boldsymbol{\theta}$ and φ in order to compute the posterior distributions

$$L(\boldsymbol{\theta}, \varphi \mid \mathbf{x}_{ov}, \mathbf{R}) \propto P(\theta, \varphi) \int P(\mathbf{x}_{ov}, \mathbf{x}_{mv} \mid \theta) P(\mathbf{R} \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}, \varphi) d\mathbf{x}_{mv} \tag{2.18}$$

When the prior distributions on $\boldsymbol{\theta}$ and φ are independent and the missing data mechanism is MAR and ignorable for inferences about unknown parameters θ (D. B. Rubin, 1996). According to D. B. Rubin (1996), this means the following equation can expressed as

$$P(\boldsymbol{\theta} \mid \mathbf{x}_{ov}) \propto P(\boldsymbol{\theta}) P(\mathbf{x}_{ov} \mid \boldsymbol{\theta} a) \tag{2.19}$$

The posterior distribution, $P(\boldsymbol{\theta} \mid \mathbf{x}_{ov})$, may not be easy to be computed. However, now we can simulate parameter estimates of unknown θ from posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{x}_{ov})$, for us make posterior inferences for unknown parameters given by $\boldsymbol{\theta}$. However, it is difficult to directly simulate sample from poste-

rior distribution $P(\boldsymbol{\theta} \mid \mathbf{x}_{ov})$, due to missing data in $\mathbf{x}$. This motivates Bayesian data augmentation method to make it possible to determine posterior inferences.

Data augmentation techniques

Data augmentation methods avoid several shortcoming associated with deletion methods. Data augmentation methods estimates model parameters from the available data observations as well as from either on the underlying distribution or probability model. In comparisons, some of the single imputation and data augmentation methods do not replace unobserved observations. Under estimation of parameters, data augmentation algorithm augment by taking into account the latent data and the observed data. In the missing data problems framework, Maximum Likelihood (ML), Expectation Maximization (EM), Markov Chain Monte Carlo (MCMC) are considered to be augmentation methods. According to McKnight et al. (2007), the Markov Chain Monte Carlo (MCMC), Maximum Likelihood (ML) and Expectation Maximization (EM) methods classication as augmentation methods is not clear-cut. The Markov Chain Monte Carlo (MCMC) method has been referred to as an augmentation method within the context of multiple imputation (Allison, 2002). The ML and EM methods are defined as model-based methods according to Little & Rubin (1987). These procedures mentioned above have also been referred to as data augmentation by J. L. Schafer (1997). We now focus on some of these methods under augmentation methods, namely ML, EM and Markov Chain Monte Carlo (MCMC) version of EM.

Maximum Likelihood (ML)

Maximum Likelihood (ML) was not originally designed to deal with missing data issues in a way such as LOCF and multiple imputation technique. The ML is usually used for estimating parameters under structural equation models (SEM) and ordinary least squares in regression model. The ML is a method that can be used for handling missing data. Little & Rubin (2002) , give the application of ML to missing data problems. Furthermore, in different situations, ML has proven to be an appropriate technique for dealing with missing data issue. Under (MAR or MCAR) missing mechanism , the missing data are ignorable, then ML is adequate to be used, and it gives unbiased estimates (Arbuckle, 1996; Allison, 2002). Therefore, the Maximum Likelihood is fairly easy to describe under assumption of MAR or MCAR. If the assumptions hold,

ML estimators for missing data produce unbiased estimates in large samples, asymptotically efficient estimates (small standard errors) and satisfy asymptotic normality which is to say that estimates approximate a normal distribution which can then be used to exploit a normal approximation for statistical inference, such as finding confidence intervals and p-values (McKnight et al., 2007). The ML can be used in most statistical software including R, SAS, SPSS, S-Plus and more.

Expectation maximization (EM)

The Expectation Maximization algorithm was manly introduced by Dempster et al. (1977) . Expectation Maximization algorithm can be defined as the process of computing and imputing missing values of each observation under the variable based on the selected probability distribution. According to Little & Rubin (2002), the Expectation Maximization algorithm is a common iterative method for Maximum Likelihood estimation under missing data problems. This algorithm hold under the MAR assumption. The basic idea of the EM is to tackle the problem of missing data and solve the complications of estimates that are related to the Maximum Likelihood estimation.

The EM algorithm used the following steps to handle missing data problems:

1. impute missing values for missing data by using predicted values simulated by Maximum Likelihood setting.
2. predict parameter estimates based on data simulated in step 1 above.
3. estimate again parameters based on the parameter estimates obtained from previous step 2.
4. estimate again parameters based on the estimate obtained from the data from step 3, and continue multiple times the same process, iterating the process until the converges stage is reached.

The EM algorithm iteration consists of two steps: expectation step and the maximization step (Little & Rubin, 2002). To finish each step the algorithm iterates multiple times repeatedly until a convergence criterion is met. The theoritical review of the information about the steps of EM algorithm are found in Dempster et al. (1977) and Little & Rubin(2002). On the convergence state, the fitted parameters are equivalent to a local maximum of a likelihood function according to Dempster et al. (1977) . This algorithm has two disadvantages: firstly, performance of the algorithm is very slow to converge. Sec-

ondly, it fails to directly measure the precision for the maximum likelihood estimates. Many proposed techniques introduced to overcome listed drawbacks, and these techniques are documented by Louis (1982); McLachlan & Krishnan (1997); D. B. Rubin (1991); Baker (1992).

2.2 Related work

In this section, we briefly review the literature mostly related to this study. In particular, we summarize missing data literature review and data augmentation technique reviews. For past years, the data augmentation algorithm has been used to tackle the problem of missing data that improves the accuracy and precision of data (data quality) according to J. L. Schafer (1997). Early literature on missing data that used to handle missing values includes a review by Afifi & Elashoff (1966) and Hartley & Hocking (1971) review. The applications literature that was introdroduced afterwards is Expectation Maximization (EM) by Dempster et al. (1977). Data imputation and augmentation procedures (see, Rubin,1987; Tanner & Wong, 1987), reviews the most powerful software of solving the computational difficulties in practical data sets to improve quality of data analysis. These techniques are implemented to tackle the problem of missing data in statistical analysis. In the 1980s, many resources for tackling missing problems were developed. More efficienct computers and new techniques were introduced for Bayesian simulation (J. L. Schafer, 1997). Only after a decade, Little & Rubin (1987) and D. B. Rubin (1987a) reviewed the shortcomings of case deletion methods and single imputations and introduced new technique called Multiple Imputations (MI). Multiple Imputations (MI) was introduced with the aim of improving computational performance and effectiveness of handling missing data (J. L. Schafer & Olsen, 1998). Since then, there have been no new studies or existing literature regarding an acceptable proportion rate of missing data for valid and quality of statistical inferences. According to J. L. Schafer (1999), a missing rate of 5% or less would be acceptable because it doesnt affect the analysis that much. Bennett (2001) argues that a missing data of 10% or more would lead to biased results. The review of Tabachnick and Fidel (2012) claimed that missing data mechanism and missing data patterns have a greater impact on analysis in comparison with the varying proportion of missing data.

There are several methods that were introduced and implemented to solve the problem of missing data in statistics. i.e listwise deletion, pairwise deletion, single im-

putations (Mean, Hot-Deck, Cold-Deck and Regression Imputation) and advanced techniques including ML, EM, MCMC. The shortcomings of both deletion and single imputation methods have been reviewed by Little & Rubin (1989). Bayes reproducing techniques such as Markov Chain Monte Carlo (MCMC) and data augmentation were developed in the late 80s period. Recently, most researchers are more focused on advanced techniques that avoid the specification of a full parametric model for the population (Robins et al., 1994). The advantage of MCMC is that it is more flexible than the classical methods since it falls under Multiple Imputation that produce minimal standard errors as compared to earlier methods according to McKnight et al. (2007). The Markov Chain Monte Carlo method has been applied in many statistical problems, overview on Gilks (1996). To apply MCMC method we start by choosing a probability model and parameter estimates, which are estimated using the Bayesian posterior distribution based on the likelihood function of the probability model of interest of the observed data incorporated with a prior distribution based on our beliefs about the distribution. The Markov Chain Monte Carlo (MCMC) method of data augmentation algorithm is implemented in order to reproduce the posterior distribution observations, from which the imputed values can be drawn. The whole iterative process of imputation is repeated a multiple K times (e.g., 50) to simulate independent data sets points (J. L. Schafer & Olsen, 1998). In the context of multiple imputation techniques, J. L. Schafer (1997) applied the MCMC method by using the data augmentation algorithm introduced by Tanner & Wong (1987). The term data augmentation was introduced by Tanner & Wong (1987) they were implying to iterative methods for sampling by adding the missing data to complete data. This Bayesian data augmentation method is the unpopular multiple imputation method for tackling missing data amongst all techniques from deletion method to likelihood based methods of dealing with missing data problem. This method can be implemented under SAS procedure called PROC MI. The paper of Raghunathan (2004) also discussed the ML method and its shortcoming for application purposes due to technical problems. Peng & Zhu (2007) continue later to do the analysis by comparing the MI technique and EM technique in the missing data analysis. These two techniques were compared for dealing with missing data in categorical covariates in logistic regression. The results were then compared to those obtained when Complete Case Method was used.

An important review to our study was that developed independently by both Tanner & Wong (1987) and Lui et al. (1994a). In generally MCMC data augmentation method was derived from the Gibbs Sampler method to obtain two steps, viz Imputation (I) and Posterior (P). The theory behind the derivation of Gibbs Sampler

is shown by J. L. Schafer (1997). The MCMC method uses available data excluding missing data from the observed data to compute a corresponding posterior distribution of interest by using the idea of data augmentation algorithm (Tanner & Wong, 1987). Thus, these techniques use a likelihood-based sampling procedure to perform the imputation of missing data. The current literature shows that most statisticians have built up techniques that are based on distributional models for handling missing data such as ML and MCMC method. These missing data techniques are documented in the statistical literature on missing data (Little, 1992; J. L. Schafer, 1997) and Little & Rubin (1987). The main aim to implement a data augmentation method is to simulate unobserved data in order to make estimation or calculations manageable in the given data set. The Bayesian data augmentation has grabbed more attention of reseachers in different fields of study to handle missing data problems of real-life applications under data analysis, including, Albert & Chib (1993), Meng & Van Dyk (1999), Holmes et al. (2006) and Frühwirth-Schnatter et al. (2009). The EM is taken as a starting point of data augmentation. The Expected Maximizing (EM) algorithm defined as a deterministic algorithm, which is the most common method that uses maximum likelihood to solve the problem of missing data. The EM algorithm was introduced by Dempster et al. (1977), which is a seminal article in the statistical framework for maximizing a likelihood function. According to Dempster et al. (1977), these algorithms are also called the two-step algorithm because of two important steps called expectation (E-step) and Maximization (M-step). The data augmentation schemes is built-up to be a stochastic version EM algothm and it's started to grow rapidly in terms of application in the statistical literature since it can solve more complicated missing data problems (J. L. Schafer & Graham, 2002; McKnight et al., 2007).

For example, the first application of data augmentation was introduced by Tanner & Wongs (1987) in the statistical analysis literature for sampling both missing data and model parameter from their posterior full conditional distributions. In a Physics literature review, data augmentation was applied by Swendsen & Wang (1987) for sampling from the Ising and Potts models (Derrida et al., 1994) and their generalization. The data augmentation in Physics literature is called the Auxiliary variables method. Also in Physics, data augmentation is called Auxiliary variables and Swendsen & Wang (1987) use this method to improve the speed of the iterative simulation of variables. In Statistics, Tanner & Wong (1987) used data augmentation scheme to make random simulations feasible and it was implemented in missing data problems. The other examples that apply data augmentation are: The data augmentation can be applied in the context of multiple imputation techniques,

J. L. Schafer (1997) applied the MCMC method by using the data augmentation algorithm introduced by Tanner & Wong (1987). Multiple Imputation has been defined in full and reviewed in detail in several papers (J. L. Schafer, 1997; J. L. Schafer & Graham, 2002; Sinharay et al., 2001; J. L. Schafer & Olsen, 1998). The MCMC data augmentation algorithm has recently been applied to missing data imputation and this has become a fact of life in different disciplines of science including medical studies (Piantadosi, 1997; Green et al., 1997; Friedman et al., 1998) and epidemiological studies (Kahn & Sempos, 1989; Clayton & Hills, 2013; Lilienfeld & Stolley, 1994; Selvin, 1996). Missing data in surveys, psychometry and econometrics are discussed in Fowler (1988); K. Schafer et al. (1993); D. B. Rubin (1987a) and (J. S. Rubin et al., 1995), to name but a few literature. The St. Louis Risk Research Project was an observational study to assess the affects of parental psychological disorders on child development. In the preliminary study, 69 families with 2 children were studied (J. L. Schafer, 1997). J. L. Schafer (1997) data augmentation algorithm was used to implement the mixed variables to solve the issue of missing data for 69 families. The psychometricians, data users and other statisticians; professionals are concerned about the presence of missing data, in large or small data sets since it affects the parameter estimation and inferences made. It brings biasness in parameter estimates of the model. The interested Researcher is encouraged to review J. L. Schafer & Graham (2002) for a comprehensive review of methods for dealing with missing data. In addition to the J. L. Schafer & Graham (2002) paper, there are a number of other comprehensive discussions regarding specific types of missing data, (J. L. Schafer, 1997; J. L. Schafer & Olsen, 1998; Bernaards & Sijtsma, 1999; Sinharay et al., 2001; Peng & Zhu, 2007) . Data augmentation (MCMC) is one of the methods of solving the problem of missing data (J. L. Schafer, 1997). More attention was based on the independent work of K. H. Li (1985a,b), who explains that multiple imputation of missing values , is very similar in its formal structure of data augmentation method. The main goal of the article of Li's to explain the accuracy of the data augmentation in making the Bayesian decisions based on the parameter estimates. Li's work focuses in the sources of examples, which are based on the imputation of missing values in given data. This research study will check whether there is any improvement in the quality of cancer medication data by comparing the result of data augmentation (MCMC) technique under different missing data mechanisms and different missing proportions to the complete data.

Data augmentation application to missing data problem includes the following reviews: Statistical literature review data augmentation methods as independent techniques in the work of Meng & Van Dyk (1999). The Markov Chain Monte Carlo method is a Monte Carlo integration method using Markov Chains (W.-x. Zhang, 2003). The method is valid only under the assumption of multivariate normality which implies that valid imputations may be generated by linear regression equations (J. M. Robins et al., 1994). W.-x. Zhang (2003) suggested the method of formulating the data with the MCMC approach and illustrating how MCMC can be conducted to impute the missing data. The Markov Chain Monte Carlo (MCMC) method is based on draws of random samples from a target probability distribution given in the problem. There are several articles that overview this method such as Besag & Green (1993) and methodological and theoretical papers include Damlen et al. (1999); Higdon (1998); Mira & Tierney (1997) and Roberts et al. (1997). The Auxiliary variables method and data augmentation method are identical in their general forms. Albert & Chib (1993) developed a data augmentation application to estimate the probit model where the observed data is viewed as censored realizations of latent utility. The data augmentation methods are also implemented to simplify analysis in Hierarchical Bayesian models, where we let augmented variables be equal to unobserved data (Tan et al., 2009). J. L. Schafer (1997) described an imputation approach of categorical data (MCMC) based on the multinomial distribution. J. L. Schafer (1997) applied Multiple Imputation Categorical data technique (MCMC) to victimization status for households in the National Crime Survey to solve missing data problem. In theory Multiple Imputation Categorical data is only appropriate when the given data contain categorical variables, then J. L. Schafer (1997) find that as the number of categorical variables increases in the model, the posterior distribution become improper, that making it impractical for use of real world problems. Fuchs (1982) analyzed data from the Protective Services Project for Older Persons, a longitudinal study designed to measure the impact of enriched social casework services on the well-being of elderly clients (Blenkner et al., 1971). This study caused considerable controversy in the social work literature. Fischer (1973) argued that the enriched services seemed to be detrimental to the clients, because the mortality rate for the experimental group was actually higher than for the control group.

Chapter 3

Methodology

The methodology chapter outlines the methods used in this study. Section 3.1 describes the data set used in this study. Section 3.2 explains the sampling method and data collection procedure used. Section 3.3 highlights the Bayesian method used for unknown parameter estimation. Section 3.4 explains the data augmentation theory, and how this method simulates missing values. Section 3.5 gives the theory behind the non-Bayesian data augmentation methods such as ML, EM etc. Section 3.6 give the explanation of MCMC and how do we incorporate it with data augmentation. Section 3.7 explains the simulation process of missing data and data preparation steps under different missing mechanisms (MCAR, MAR, MNAR) scenarios. Section 3.8 gives all statistical tools used in the study. Section 3.9 gives the theory behind the statistical computation package in R which is mix, how it works. Section 3.10 explain data augmentation for mixed data, the mathematical aspect and theory. Section 3.11 explain the missing data imputation process and performance measurement of different models.

3.1 Source of Data

The South Africa General Household Survey (2015) is the source of data in this study and it is used to do all the statistical analysis. These are datasets available online and are accessible as an open source on the home page of Statistics South Africa. One may download it or write a request to access the data for research purposes. The accessibility from this source is approved by government officials (StatSA). South Africa General Household Survey (2015) is responsible for collecting and disseminating accurate, nationally representative data on population in South Africa to measure the level of development and monitor the performance of various government

programs and projects such as building houses, creating job opportunities, assessing healthy facilities etc (StatSA, 2015).

3.1.1 Description of the General Households Survey 2015 Data

The General Household Survey (GHS) was conducted by Statistics of South Africa from January to December 2015. The survey subjects are based on gathering information on households with respect to six aspects: housing, agriculture, food security, health and social development, household access to services, facilities and education. The main purpose of GHS2015 is to measure the level of development and monitor the performance of various government programs and projects such as building houses, creating job opportunities etc. It also provides national and provincial indicators in various living conditions and draws a comparison between the General Household Survey of the previous years and the current survey results. The data used was collected on everyone on the household, where the household was used as the unit of observation.

The data contain the total **N=74450** number of observations for people who responded to the survey in South Africa. The data used in this study was collected from all private households and residents in workers hostels in all nine provinces of South Africa. The survey does not include collective quarters such as student hostels, old age homes and others. This survey was conducted by interviewing people within households using questionnaires by StatSA. This study only focuses on data (patients) who have cancer and are using cancer medication. The sample is composed of a total of 194 patients, who responded and have cancer and are also users of chronic cancer-medication. Among this selected sample,175, (90.21%) use cancer medication and 19 (9.79%) are not users of cancer medication.

3.2 Study Population and Sampling Procedure

The data used was collected on everyone on the household, where the household was used as the unit of observation. A multi-stage design was used in this survey, which is based on a stratified design with probability proportional to size selection of primary sampling units (PSUs) at the first stage and sampling of dwelling units (DUs) with systematic sampling at the second stage. After allocating the sample to the provinces, the sample was further stratified by geography (primary stratification), and by population attributes using Census 2001 data (secondary stratification). During the first phase of the survey, sampled dwelling units were visited and informed about the coming survey as part of the publicity campaign.

PSUs are enumeration areas (EAs) from the census list, that had a household count of more than 25 or less (excluding workers hostels, convents and monasteries). In addition, EAs in the census that have less than 60 dwelling units were combined to form PSUs. The GHS is a survey conducted in South Africa with three major data sets: household, persons and workers. There are nine provinces in South Africa: Western Cape (WC), Eastern Cape (EC), Northern Cape (NC), and Free State (FS), KwaZulu-Natal (KN), North West (NW), Gauteng (Gau), Mpumalanga (Mpu) and Limpopo (Lim). Two hundred and thirty-three enumerators (233) and 62 provincial and district coordinators participated in the survey across all nine provinces. An additional 27 quality assurors were responsible for monitoring and ensuring questionnaire quality. The actual interviews took place four weeks later. A total of 74450 households (including multiple households) were successfully interviewed during face-to-face interviews.

3.2.1 The cleaning process

The data South Africa General Household Survey (2015) didn't come in the form that is ready to be used (clean form), for analysis. In this section, I am going to explain how I cleaned the data to be in an acceptable format to be used for the imputation method, called the data augmentation algorithm.

Figure 3.1 below demonstrate how we arrived at a sample size of 204 after the data was refined.

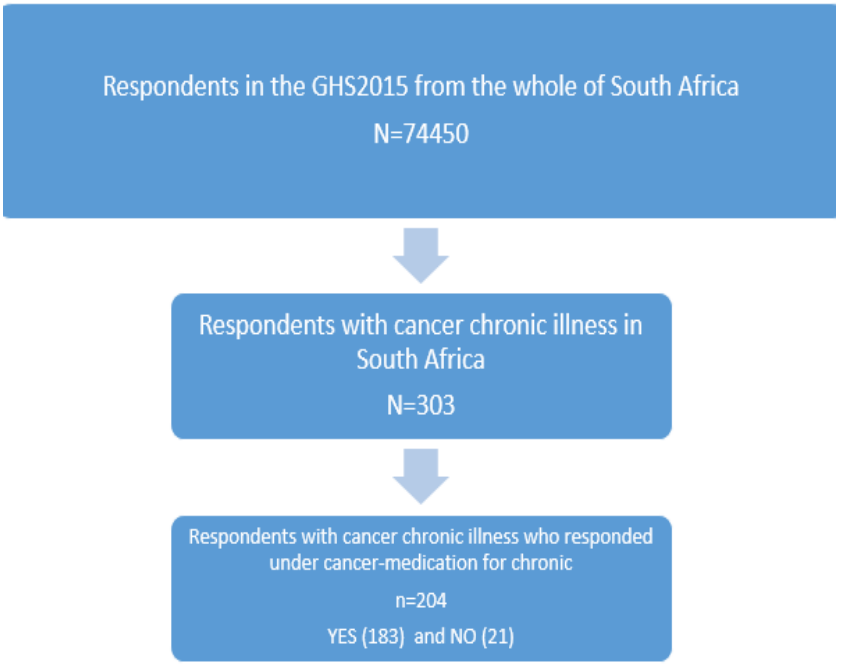

Figure 3.1 – Data refining stages

The data contains the **N=74450** which equals the number of observations for people who responded to the survey in South Africa. The sample size of $n = 303$ is the sample size of people who have cancer chronic illness in South Africa. In order to make the dataset suitable for the analysis, we further reduced the sample size from $n = 303$ to $n = 204$ for a sample of people who use cancer medication for chronic cancer. In summary then as noted above, we focused on individuals from South Africa who are affected by cancer and use cancer medications to cure this disease. These individuals were categorized according to whether they use cancer medication or not. The observations with missing data for any of this variable are excluded from the study.

3.2.2 Selection of dependent and independent variables

Independent variables are grouped into three categories. These are demographic variables, socio economic variables, and the geographic variables that have an influence in cancer-medication in South Africa.

The response variable:

The dependent variable that is the main focus of the analysis is cancer-medication intake in South Africa which is a binary random variable. A patient who is taking cancer medication is coded as 1 and a patient who is not taking cancer medication is coded as 2. This variable was used as a dependent variable in the logistic regression model and needed to be estimated in order to investigate the effect of other independent variables.

The explanatory variables in the model:

The independent predictor variables consisted of baseline demographic and geographic variables, which were collected from each household. The socio-economic variables were working for wage. Geographic variables were Area of living and demographic variables were gender, age. The variables, age, working for wage, Area of living and gender were all collected at the individual level. These predictors that were considered in this study predict the dependent variable (cancer-medication) because they are to be influential in cancer-medication intake in South Africa. The variables were used either in the logistic regression models for parameter estimation or data augmentation models of interest.
The **Table 3.1** below shows the variables of the model and number of levels for each variable.

Table 3.1 Table of predictors variables used in Cancer-medication

Variable	Levels	code
Cancer-Medication	1=Yes, 2=No	Q26bCAN
Gender	1=Male, 2= Female	Gender
Working for wage	1=Yes, 2=No	Q41awge
Area of living	1=Urban, 2=Rural	Geotype
Age	Continuous	Age

Note: Missing values are allowed to be in all variables in the model except dependent variable (Q26CAN), in order to apply data augmentation(***mix***) package in R for mixed data.

3.3 Bayesian methodology

3.3.1 Bayesian methods

Over the past 10 years, researchers in the field of social sciences and medicine are increasingly using Bayesian statistical methods to analyse the data. The classical statistical methods are to produce parameter estimates from the sample data to hypothesized population parameter estimates that are assumed to be unknown but equivalent to a fixed value. In contrast to classical statistical theory, Bayesian methods incorporate preexisting evidence and beliefs about a parameter to what we call a prior distribution. In short, under Bayesian analysis framework, the population parameter will change depending on subjective probability. The mathematical method was derived from Bayesian theorem by combining the prior beliefs about parameter with the new data to get the updated posterior distribution. If one assumes a uniform prior distribution (also referred to as a diffuse or noninformative), in which each and every value in the existing distribution is having an equally likely probability of occurring, then the likelihood exists that the observed data is asymptotically equal to the posterior distribution of the data. Under the Bayesian analysis, the missing data values are taken as additional parameters to be estimated in the data, subject to the constraints of the analysts model and the prior distributions of parameters in the model. According to Burton et al. (1998) Bayesian-based inferences are more intuitive and can lead to more appropriate decisions; than classical method of testing the null hypothesis significance within the data to make decisions. One of the main reason for preference of using the Bayesian methods is that it improves the computing power.

3.3.2 Posterior distribution

Firstly, let us define the joint distribution of variable **x** and the parameter θ as follows:

$$P(\mathbf{x}, \theta) = P(\mathbf{x} \mid \theta)P(\theta) \tag{3.1}$$

where $P(\mathbf{x} \mid \theta)$ is the sampling distribution. It can also be expressed as the likelihood function $L(\theta \mid \mathbf{x})$, when the sampling distribution is defined as a function of θ with fixed values of **x**. Bayesian statistics uses a posterior distribution to make inferences about the unknown parameter θ. To apply Bayes theorem ,under the assumption of ignorability to the model parameters of an incomplete data is considered as

$$P(\theta \mid \mathbf{x}_{ov}) = \frac{P(\mathbf{x}_{ov} \mid \theta)P(\theta)}{P(\mathbf{x}_{ov})} \propto P(\mathbf{x}_{ov} \mid \theta)P(\theta) \tag{3.2}$$

where $P(\theta)$ is the prior distribution of the model parameters and, $P(\mathbf{x}_{ov} \mid \theta)$ denote the observed-data posterior distribution. However, we can update our information about the distribution because of posterior predictive distribution of the missing part in sub-vector **x**.

3.3.3 Posterior predictive distributions

The predictive distribution in Bayesian is the distribution made by averaging future estimations over posterior distributions of unknown parameters. In Bayesian inference, it is very important to check the adequacy of the model. The predictive distribution is classified into two distributions, prior and posterior predictive distribution. For prior distribution, means that before we actually observe the data, we make our own assumptions and prior beliefs about data and the distribution of the prior distribution is defined by the following expression.

$$P(\mathbf{x}_{mv}) = \int P(\mathbf{x}_{mv} \mid \theta)P(\theta)d\theta \tag{3.3}$$

Therefore, it is useful for to check whether the choice of prior distribution does capture prior beliefs about a distribution. The posterior predictive distribution is one of the techniques which can be used in model checking by drawing samples of the replicated data and comparing these samples to the observed data. Let $\mathbf{x}_{mv}$ denote the future values or unknown observations. The aim of using predictive distribution decision is to make inferences based on future values $\mathbf{x}_{mv}$. where $\mathbf{x}_{ov}$ is the observed

data, thus the prediction distribution of $\mathbf{x}_{mv}$ given the complete data $\mathbf{x}$ is defined in Aitchison & Dunsmore (1976).

After knowing that $\mathbf{x}_{ov}$ is an observed data, it is easy to forecast $\mathbf{x}_{mv}$. The posterior predictive distribution of $\mathbf{x}_{mv}$ given data $\mathbf{x}_{ov}$ is defined as follows

$$P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}) = \int P(\mathbf{x}_{mv}, \theta \mid \mathbf{x}_{ov}) d\theta \tag{3.4}$$

$$= \int P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta) P(\theta \mid \mathbf{x}_{ov}) d\theta \tag{3.5}$$

where $\mathbf{x}_{ov}$ is observed data, $\mathbf{x}_{mv}$ represent unobserved values and θ are unknown parameters of the model.

3.4 The Basic Principles of Data augmentation

3.4.1 The Bayesian Data Augmentation algorithm

The Data augmentation (DA) algorthm was originally introduced by Tanner & Wong (1987) as the estimation procedure for dealing with the latent data or unobserved values by adding latent data on the observe data, so that will be easier to analyse data with missing values. The data augmentation can be implemented in solving or simulating missing values, so that we can be able to compute full posterior distributions of θ (Tanner & Wong, 1987). The Markov chain Monte Carlo data augmentation has the advantage over the classical EM procedure because it shows greater flexibility when the underlying distributions are unknown unlike ML model-based methods only function under the certain distributional assumptions, for example multivariate normality etc.

Under MCMC, a different set of procedures can be used to simulate random numbers. Gibbs sampling is the most commonly used because it is most available in the statistical software. This process can be called Bayesian because we are able to obtain a probability distribution known as posterior distribution. A posterior distribution can be used for parameter estimation in the Bayesian framework, to make proper inferences. The standard MCMC methods such as Gibbs sampling, Metropolis and much more are implemented to simulate values independently following the given probability distribution. The standard Monte Carlo methods produce simulated values in a Markov chain. A Markov chain is a stochastic model that describes a sequence of possible random variables in which the probability for each value depends only on the previous step. Therefore, we can conclude that MCMC process produces simulated random values that are dependent on each other since the previous outcomes influence predictions for the next experiment. The Data augmentation algorithm has become more popular in both Classical and Bayesian frameworks. This method can be used to impute the missing data or augmenting latent values in such a way that:

- Augment data with the missing values in order to make valid decisions about data.
- Easy to obtain unbiased estimates of the augmented data given the parameter θ in the analysis.

The EM algorithm (Dempster et al., 1977) also uses this tool in solving maximum likelihood problems. The disadvantage of EM algorithm is that when the likelihood

cannot be estimated by using the normal likelihood, the EM algorithm tends to underestimatestandard errors, which are critical to hypothesis testing (Allison, 2002). The common MCMC procedure has desirable features than ML model-based procedure are efficiency and flexibility (McKnight et al., 2007). The MCMC process are efficient because it allows us to estimate parameter even if the unexpressed distributions are unknown or non-normally distributed. The advantage of using software MCMC procedure (Bayesian data augmentation) is that it always finds the solutions even for most complex missing data problems. Especially when given data distributions are unknown or when it does not follow the multivariate normal distribution. The data Augmentation procedures almost follow the the same logic as the EM, have two iterative steps but Bayesian Data Augmentation procedures are unrestricted to the expectations of distribution (e.g. multivariate normal) unlike EM estimation procedure. In the iterative process, the EM is restricted by the expectation derived from a distribution in order to estimate parameters while the MCMC methods are unlimited by distributional assumptions.

Given the data with missing observations, our main purpose is to obtain unbiased parameter estimates but it is very difficult to obtain such estimates by ignoring unobserved data and using the observed data only. The augmented data consist of the set of missing values $\mathbf{x}_{mv}$ and observed values $\mathbf{x}_{ov}$, which form $\mathbf{x}$=$(\mathbf{x}_{ov}, \mathbf{x}_{mv})$. Data Augmentation procedure allows to augmenting or completing the observed data $\mathbf{x}_{ov}$ with simulated values of the missing data $\mathbf{x}_{mv}$ to manage missing data problem. The main focus is to compute the posterior distribution $P(\theta \mid \mathbf{x}_{ov})$, but unfortunately it is very difficult to compute this posterior distribution due to the presence of missing values in $\mathbf{x}$ (M. A. Tanner & Wong, 2010). If both $\mathbf{x}_{ov}$ and $\mathbf{x}_{mv}$ are given, we can calculate or sample from the augmented posterior distribution $P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv})$. Inorder to obtain posterior distribution,first find the multiple imputation of $\mathbf{x}_{mv}$ from the predictive distribution and then calculate the mean of $P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv})$ divide by the above imputations. Since the predictive distribution $P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov})$ depends on the full posterior distribution $P(\theta \mid \mathbf{x}_{ov})$, it is important to determine $P(\theta \mid \mathbf{x}_{ov})$ for iterative algorithm.

The motivation of the procedure is that the distribution in these two steps are much easier to draw from either of the posterior distributions $P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv})$ and $P(\theta \mid \mathbf{x}_{ov})$ or the joint posterior distribution $P(\theta, \mathbf{x}_{mv} \mid \mathbf{x}_{ov})$. The MCMC procedure has two iterative steps to augment the data, which is the imputation or I-step and posterior or P-step. A data augmentation algorithm provides the way of improving the quality of data and inferences, especially if we have small-sample of data to refine the EM algorithm. The data augmentation iterative steps are repeated until we reach the convergence stage of the algorithm.

The foundation of data augmentation algorithm

The original data augmentation is motivated by following the two basic integrals (Tan et al., 2009):

1.The Posterior identity

$$P(\theta \mid \mathbf{x}_{ov}) = \int_{\mathbf{x}_{mv}} P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}) P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}) d\mathbf{x}_{mv} \tag{3.6}$$

2. The predictive identity

$$P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}) = \int_{\theta} P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \varphi) P(\varphi \mid \mathbf{x}_{ov}) d\varphi \tag{3.7}$$

where $P(\theta \mid \mathbf{x}_{ov})$ represent the posterior distribution of θ given the observe data $\mathbf{x}_{ov}$, $P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov})$ represent the predictive distribution of the missing values $\mathbf{x}_{mv}$ given the observed data $\mathbf{x}_{ov}$ and $P(\theta \mid \mathbf{x}_{mv}, \mathbf{x}_{ov})$ also represent the conditional distribution of the parameter θ given the augmented data $\mathbf{x} = (\mathbf{x}_{mv}, \mathbf{x}_{ov})$.

In the above integrals identities, we substitute (3.6) into integral (3.7) and then interchange the order of integration. Thus, the posterior distribution $P(\theta \mid \mathbf{x}_{ov})$ satisfy the following integral equation

$$h(\theta) = \int A(\theta, \varphi) h(\varphi) d\varphi \tag{3.8}$$

$$P(\theta \mid \mathbf{x}_{ov}) = \int A(\theta, \varphi) P(\varphi \mid \mathbf{x}_{ov}) d\varphi \tag{3.9}$$

where the kernel function is given by

$$A(\theta, \varphi) = \int P(\theta \mid \mathbf{x}_{mv}, \mathbf{x}_{ov}) P(\mathbf{x}_{mv} \mid \varphi, \mathbf{x}_{ov}) d\mathbf{x}_{mv} \tag{3.10}$$

Let T be an integral transformation that may transforms any function g that is inte-

grable into another integrable function Tg by following integral

$$Tg(\theta) = \int A(\theta, \varphi) g(\varphi) d\varphi \tag{3.11}$$

Equation (3.11) can be solved by the method of successive substitution and these suggest a method of determining $P(\theta \mid \mathbf{x}_{ov})$. When we apply a functional analysis, then the fixed point iteration becomes

$$P_{i+1}(\theta \mid \mathbf{x}_{ov}) = \int A(\theta, \varphi) P_i(\varphi \mid \mathbf{x}_{ov}) d\varphi, \quad i \in N \tag{3.12}$$

On the equation (3.12) above we can implement the method of successive substitution to approximate for the solution.The function P_i calculated above will always converge to the posterior distribution of $P(\theta \mid \mathbf{x}_{ov})$ under mild conditions. Tanner & Wong (1987) explain the mild condition of the equation (3.12) as follows

- sufficient conditions for convergence of P_{i+1} to P in l_1-norm

Tanner & Wong (1987) define the sufficient conditions as follows

$$\int \mid P_{i+1}(\theta \mid \mathbf{x}_{ov}) - P_i(\theta \mid \mathbf{x}_{ov}) \mid d\theta \to 0 \tag{3.13}$$

as i became large i.e. $i \to \infty$:

1. The Kernel function $A(\theta, \varphi)$ must be uniformly bounded and equicontinuous in the parameter θ;
2. The starting value is any initial approximation P_0 must satifies the condition $sup_\theta \frac{P_0(\theta|\mathbf{x}_{ov})}{P(\theta|\mathbf{x}_{ov})} < \infty$.

The computation of intractable integration (3.10), (3.11) and (3.12) is not easy. There are many ways of approximating such complex integrals which are numerical integration, analytical approximations, and Monte Carlo methods. But it is always possible to use Monte Carlo methods because it performs very well (Tanner & Wong, 1987). In the seminal paper, Tanner & Wong (1987) implement the method of Monte Carlo to determine or to compute such integration in (3.9).

3.4.2 General steps of data augmentation algorithm

The data augmentation algorithm iteration process start with imputation (I) step and proceeds to posterior (P) step through the process of Markov chain. This posterior distribution $P(\theta \mid \mathbf{x}_{ov})$ may be difficult to be calculated directly. Given the updated

parameter at iteration t which is $\theta^{(t)}$ of the original θ the data augmentation algorithm simulate values using the following two steps:

Imputation (I) Step:

Simulates multiple values of missing data of independent sample $\mathbf{x}_{mv}^1, \mathbf{x}_{mv}^2, .., \mathbf{x}_{mv}^k$ from the current i^{th} approximation $P_i(\theta \mid \mathbf{x}_{ov})$ to the predictive distribution given by $P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov})$.

Posterior (P) Step:

Updating the current i^{th} approximation distribution $P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov})$ can be approximately obtained as the average of $P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv})$ over missing data that was imputed in the imputation step.

$$P_{i+1}(\theta \mid \mathbf{x}_{ov}) = \frac{1}{k} \sum_{j=1}^{k} P(\theta \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}^j) \tag{3.14}$$

The missing data $\mathbf{x}_{mv}^1, \mathbf{x}_{mv}^2, .., \mathbf{x}_{mv}^k$ are generated multiple times,then are often called *multiple imputation* (D. B. Rubin, 1987b). This iterative procedure can be shown to eventually converge to a draw from the joint distribution of $P(\mathbf{x}_{mv}, \theta|\mathbf{x}_{ov})$ as $j \to \infty$. The value of k need not be very large, in fact with $k = 1$ the DA algorithm reduces to a special case of the Gibbs sampler.

3.5 Non-Bayesian data augmentation techniques

3.5.1 Maximum likelihood method

The Maximum Likelihood (ML) method is the most used method under the model-based procedure for handling missing data problem in the dataset. Maximum likelihood is listed as one of the good methods among others for tackling missing data problem on the incomplete data in different scenarios (Graham, 2009). The literature review of the ML technique shows that most researchers use ML technique for parameter estimation of the logistic regression model. The maximum likelihood technique works well under the MAR assumption, and it yields unbiased estimates

when the sample is large enough, it is efficient (small standard errors) and the estimates of the parameters approximately follow a normal distribution when we sample repeatedly (these estimates can be used for computing confidence intervals and p-values).

Maximum likelihood (ML) estimation procedure is commonly found in statistical software packages such as SPSS, SAS, S-plus, R, AMOS, EQS and more. The advantage of using ML procedure is that it handles missing data adequately and simply. But, it holds when missing data mechanism are ignorable (i.e. MAR) and ML give unbiased estimates (Allison, 2002; Arbuckle, 1996).

The complete data model for covariates

Let **x** be a full covariate data matrix with $n \times p$ dimension, which rows of x are independent, identically distributed and $\mathbf{x}_i$ denote the i^{th} row of **x**, $i = 1, 2, 3, ..., n$. Since $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ are independent random samples and assuming that each probability distribution depends on the parameter of interest or unknown parameter given by θ. Given the independent observed values $\mathbf{x}_1 = x_1, \mathbf{x}_2 = x_2, ..., \mathbf{x}_n = x_n$ given some parameter, then the probability density function is $f(x_1, x_2, ..., x_n \mid \theta)$. The probability density of complete covariate data can be given by

$$P(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(\mathbf{x}_i \mid \theta) \tag{3.15}$$

where $\mathbf{x}_i$ are independent and identically distributed and f is the density function of a single row. This is equivalent to maximizing the log likelihood since logarithm is an increasing function

$$L(\theta) = \sum_{i=1}^{n} \log f(\mathbf{x}_i \mid \theta) \tag{3.16}$$

The distribution of the function f is classified into three classes:

1. the multivariate normal distribution, see Schafer (1997);
2. the multinomial model for categorical variables, also include loglinear models;
3. a class of models for mixed normal and categorical variables, see review (Little & Schluchter, 1985; W. J. Krzanowski, 1980; W. Krzanowski, 1982).

Since ML technique is restricted under some certain assumptions, however if these assumptions mentioned above hold, then the procedure for missing data will generate parameter estimates that have the the following properties associated with maximum likelihood : consistency, asymptotic efficiency and asymptotic normality (Allison, 2002). Consistency is defined as: the larger the samples provided, the more likely it is to obtain unbiased parameter estimates. Asymptotic efficiency means that the computed parameter estimates have minimal standard errors. Asymptotic normality is a very useful property for maximum likelihood since it allows normal to approximate the given data set in order to obtain confidence intervals and p-values for the statistical analysis. Handling missing data requires data to be normally distributed so that the maximum likelihood method can produce adequate estimates of standard errors that take into account all data observations including unobserved data.

Maximum likelihood function with missing data

The maximum likelihood with missing data always assumes that missing data mechanism is ignorable and must be an MAR process (Allison, 2002). The ML method can easily be described or defined under the above assumptions. When some data is missing from the original data, the maximum likelihood can be obtained by summing the normal likelihood method over all possible values of missing observations in the original data. The maximum likelihood (ML) technique for missing values can be described mathematically as proceeded from literature in Graham (2009). Let **x**, **y** be two discrete variables and $f(\mathbf{x}, \mathbf{y} \mid \theta)$ be a joint distribution for n independent sample of observations, where θ is the vector of unknown parameters that control the distribution of the given discrete variables. If the complete data (no missing data) and given observations are independent and identically distributed, the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} f(x_i, y_i \mid \theta) \tag{3.17}$$

To obtain good maximum likelihood estimates, we must estimate and find the values of unknown vector of parameter θ that makes the likelihood function as large as possible. Now, missingness is observed on the original data and to handle those missing values created, we can use method of maximum likelihood. Now, suppose that data are missing at random on variable **x** for the first k observations, both **x** and **y** are observed variables and the $(n - k)$ remaining observations. According to Allison (2002) only variable **y** can be measured.

To obtain marginal distribution on **x** by summing over **y** is denoted by

$$g(\mathbf{x} \mid \theta) = \sum_{\mathbf{y}} f(\mathbf{x}, \mathbf{y} \mid \theta) \tag{3.18}$$

And also, the marginal distribution on **y** by summing over **x** is denoted by

$$h(\mathbf{y} \mid \theta) = \sum_{\mathbf{x}} f(\mathbf{x}, \mathbf{y} \mid \theta) \tag{3.19}$$

If the variables missing in the original data are continuous, replace summation signs with integral signs. Therefore, the marginal distribution on **x** is obtained by integrating over **y** denoted as below :

$$g(\mathbf{x} \mid \theta) = \int_{\mathbf{y}} f(\mathbf{x}, \mathbf{y} \mid \theta) \tag{3.20}$$

And also, the marginal distribution (Likelihood function) on **y** by integrating over **x**

is denoted by

$$h(\mathbf{y} \mid \theta) = \int_{\mathbf{x}} f(\mathbf{x}, \mathbf{y} \mid \theta) \tag{3.21}$$

The likelihood function for the full data set becomes:

$$L(\theta) = \prod_{i=1}^{k} g(x_i \mid \theta) \prod_{i=k+1}^{n} h(y_i \mid \theta) \tag{3.22}$$

There are several methods available to solve this optimization problem above, to compute parameter estimates of the model and make inferences on the analysis. To find the values of the unknown vector parameter given by θ by making likelihood function (3.17) as large as possible. The maximum likelihood function (3.22) is separated into two parts (observed and missing data) that are equal to different missing data designs. The likelihood for each design can be found by summing the joint probability distributions over all expected values of the different variables with missing data. In case, where the given variables are continuous, replace summation signs with integrals signs.

To use maximum likelihood for missing data the joint distribution of the models must be known for all variables that are included in the model otherwise we cannot be able to do parameter estimation. Therefore, characteristics mentioned above is needed in order to implement maximum likelihood to handle missing data. For all categorical variables used in the model, two appropriate models are more likely to be used which is unrestricted multinomial model or a log-linear model that has some limitations depending on the given data (McKnight et al., 2007). For all continuous variables used in the model,it is typical to assume that the model is multivariate-normal (Allison, 2000). This implies that each variable in the model is normally distributed and each variable can be expressed as the linear function of the other variables, with homoscedasticity error terms and mean zero. These assumptions are commonly be found in the basis for multivariate analysis and linearstructural equation modelling in statistics. Under the assumption of multivariate normal model, the maximum likelihood function can be maximized using the expectation-maximization (EM) algorithm or direct maximum likelihood (McKnight et al., 2007).

3.5.2 Direct Maximum likelihood method

Let the matrix $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$ be $n \times k$ of the incomplete data, which is separated into two parts which is observed part $\mathbf{x}_{ov}$ and missing data part $\mathbf{x}_{mv}$. Thus, let θ be some unknown parameter belonging to the likelihood of the complete data. The log-likelihood function of the observed-data can be defined as

$$L(\theta \mid \mathbf{x}_{ov}) = \sum_{i=0}^{n} \ln f(\mathbf{x}_{ov,i} \mid \theta) \tag{3.23}$$

Solving this log-likelihood function is possible, but it requires expensive computation. However, Allison (2001) introduces the specializing software mainly in used for Structural Equation Modeling to determine parameter estimates of these functions.

In theory, a direct maximum likelihood is the more appropriate method because it provides accurate estimates for standard error and good for over-identified statistical models. However, it also requires the joint distribution to be known for all given variables with missing observations (Allison, 2001). The two-step method has a problem concerning the standard error of the estimates. The problem is that it does not give us dependable standard error estimates. The direct maximum likelihood was introduced to solve such problem associated with standard error estimates. The direct maximum likelihood is also known as "Raw" maximum likelihood or full information maximum likelihood (Arbuckle, 1996; Allison, 2003). In this procedure, it is very important to specify the linear model of interest. This approach maximizes the likelihood function directly with respect to the parameters of the model. The limitation of these approach is that it requires specified software that may have a steep learning curve to obtain estimates of the model and find standard errors, which may be time consuming and difficult.

3.5.3 Expectation Maximization (EM) method

The EM algorithm is an iterative technique that is commonly used to estimate parameters of the model by using maximum likelihood estimation when the model has missing data problem (Dempster et al., 1977). For complex dataset, ML method might have pitfalls in determining parameter estimates but EM algorithm was introduced as an effective method to handle statistical data when ML method fails (McKnight et al., 2007). The EM was introduced to be implemented by statisticians or researchers when handling the effect of missing values in the data set and reduce the problems associated with parameter estimation corresponding to a use of ML. The EM algorithm comes up with a basic idea for determining parameter estimates of the model to augment the observed data with missing data using their maximum likelihood distributions. The EM procedure is always taken as a default data augmentation method based on maximum likelihood framework.

The first attractive property for EM algorithm is that EM estimator is unbiased and efficient when the missing data mechanism is ignorable (Graham, 2003). The main advantages of using EM algorithm method is that it is simple to use it (Dempster et al., 1977), stable (Couvreur, 1996) and most of the software or packages, both free and commercial can be implemented to determine EM parameter parameter estimates. The EM algorithm is made up of two iterative steps: Expectation and Maximization. The iterative steps for EM in the case where the data completely missing, is that it obtains possible observations for the data that are missing (E-step) and impute these expected observations to the likelihood and maximizes the complete likelihood of a complete data (observe and missing data) which is M-step. These two steps are repeated in the loop as:

1. first replace missing observations by predicted values.
2. predict parameters.
3. re-estimate the missing data assuming the updated parameter estimates are adequate.
4. re-predict new parameters.

The possible observations of the missing data is determine by computing the expectation of the missing data given the observed data $E(\mathbf{x}_{mv}|\mathbf{x}_{ov})$ of the current values of the parameters (Lesaffre & Lawson, 2012). These above loops repeat several times until it reaches the converged stage solution in which the difference is very small compared to the previous solution. The general EM algorithm was in-

troduced by Dempster et al. (1977) for determining maximum likelihood estimates especially where data is incomplete by using the method of maximum likelihood estimate (MLE). The basic idea is quite simple for EM algorithm. Consider the random variable vector **x** with the joint density function $f(\mathbf{x} \mid \theta)$, where θ is the vector of parameters of interest with k-dimensional $\theta \in \boldsymbol{\theta} \subseteq \Re^k$. The vector **x** denotes complete data model which is the observed data and we are interested in finding the maximum likelihood estimate of parameter θ based on the given distribution function of **x**. The log-likelihood function of the vector **x** is given as follows (Dempster et al., 1977)

$$\log L(\theta \mid \mathbf{x}) = \log f(\mathbf{x} \mid \theta) \tag{3.24}$$

must be maximised so that we can obtain parameter estimates. Under the incomplete data, we tend to consider the observed data only in the function of the complete-data **x** vector, whereas the missing data are present in the model. The problem can be expressed by splitting **x** into two parts as $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$, where $\mathbf{x}_{ov}$ denotes the observed but data is incomplete and $\mathbf{x}_{mv}$ denote the missing data. For simplicity of description, let's assume that the missing data in the original data are missing at random(D. B. Rubin, 1976), then the joint function can be expressed as

$$f(\mathbf{x} \mid \theta) = f(\mathbf{x}_{ov}, \mathbf{x}_{mv} \mid \theta) \tag{3.25}$$

$$f(\mathbf{x} \mid \theta) = f_1(\mathbf{x}_{ov} \mid \theta) f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta) \tag{3.26}$$

where f_1 is the joint density function of the observed data and f_2 is the joint density functions of missing data given the observed data. Thus, we compute the likelihood of the observed data as follows

$$L'(\theta, \mathbf{x}_{ov}) = L(\theta, \mathbf{x}) - \log f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta) \tag{3.27}$$

where $L'(\theta, \mathbf{x}_{ov})$ is the log-likelihood of the observed data. But also by viewing each term in the expression (3.27) as a function of θ, then the likelihood function obtained is

$$L(\theta, \mathbf{x}) = L'(\theta, \mathbf{x}_{ov}) + \log f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta) + C \tag{3.28}$$

where C is an arbitrary constant.
The term $f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta)$ is the predictive distribution of the missing observations given parameter θ.

This predictive distribution plays the important role in showing the link between $\mathbf{x}_{mv}$ and θ in EM. In general, the conditional predictive distribution of the missing observation cannot be compute in the data set, we usually use the current parameter

of interest $\theta = \theta^{(t)}$ to obtain

$$U(\theta \mid \theta^{(t)}) = L(\theta \mid \mathbf{x}_{ov}) + W(\theta \mid \theta^{(t)}) + C \tag{3.29}$$

where

$$W(\theta \mid \theta^{(t)}) = \int \log f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta) f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta^{(t)}) d\mathbf{x}_{mv} \tag{3.30}$$

The used of EM algorithm can be very useful to maximize the log-likelihood of the observed data, although it can be a little bit challenging to find parameter estimates but for the complete data log-likelihood (L) can be maximized very simply. However, if $\mathbf{x}$ is unobserved, the log-likelihood (L) is impossible to be assessed and maximized. The EM algorithm is trying to maximize the log-likelihood $L(\theta, \mathbf{x})$ by using iterations where the parameter estimates are updated in the loop, until the estimates converges to a certain decimal placed value (Dempster et al., 1977; Wu, 1983). This can be done by replacing it by conditional mean or expectation given the observed data $\mathbf{x}_{ov}$. The expectation is computed with respect to complete data distribution function assessed at the present estimates of the parameter of interest θ. Technically, considering $\theta^{(0)}$ as the initial value for θ, then at the first iterations we compute (Dempster et al., 1977).

$$U(\theta, \theta^{(0)}) = E_{\theta^{(0)}}[L(\theta, \mathbf{x} \mid \mathbf{x}_{ov})] \tag{3.31}$$

Then $U(\theta, \theta^{(0)})$ is maximized with respect to θ, so that, we can obtain $\theta^{(1)}$ such that

$$U(\theta^{(1)}), \theta^{(0)}) \geq U(\theta, \theta^{(0)}) \tag{3.32}$$

for all $\theta \in \Theta$. The EM algorithm is named from the two iterative steps involved under this procedure. The E-step (Expectation step) and the M-step (Maximization step) can be defined mathematically as follows according to Dempster et al. (1977). The E-step find the existence of the current expected log-likelihood of the complete-data.

E-step : Compute $U(\theta, \theta^{(t)})$, where

$$U(\theta, \theta^{(t)}) = E_{\theta^{(t)}}[L(\theta, \mathbf{x} \mid \mathbf{x}_{ov})] \tag{3.33}$$

$$U(\theta, \theta^{(t)}) = \int L(\theta \mid \mathbf{x}) f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta^{(t)}) d\mathbf{x}_{mv} \tag{3.34}$$

The M-step is the step where current expected log-likelihood of the complete-data

obtain new parameters $\theta^{(t+1)}$.

M-step : Find $\theta^{(t+1)}$ in θ, such that

$$U(\theta^{(t+1)}, \theta^{(t)}) \geq U(\theta, \theta^{(t)}) \tag{3.35}$$

for all $\theta \in \Theta$.

These two-steps, which is E-step and M-step alternates up until the convergence is achieved (Allison, 2001). This means that these two-steps alternate many times repeatedly until the difference between log likelihood functions

$$L(\theta^{(t+1)}) - L(\theta^{(t)}) \tag{3.36}$$

become less than ξ, where ξ is the prescribed small quantity. Dempster et al. (1977), shows that the definition $\theta^{(t+1)}$ as the value of θ that minimizes $U(\theta \mid \theta^{(t)})$, then the value of current parameter $\theta^{(t+1)}$ is a better estimate than the previous estimate $\theta^{(t)}$ in such a way that the observed data loglikelihood is at least higher for $\theta^{(t+1)}$ than that of $\theta^{(t)}$.

$$L(\theta^{(t+1)} \mid \mathbf{x}_{ov}) \geq L(\theta^{(t)} \mid \mathbf{x}_{ov}) \tag{3.37}$$

Thus, it can be shown by the following equation

$$L(\theta^{(t+1)} \mid \mathbf{x}_{ov}) \geq L(\theta^{(t)} \mid \mathbf{x}_{ov}) = U(\theta^{(t+1)} \mid \theta^{(t)}) - U(\theta^{(t)} \mid \theta^{(t)}) + W(\theta^{(t)} \mid \theta^{(t)}) - W(\theta^{(t+1)} \mid \theta^{(t)}). \tag{3.38}$$

where, $U(\theta^{(t+1)} \mid \theta^{(t)}) - U(\theta^{(t)} \mid \theta^{(t)})$ cannot be negative because the updated $\theta^{(t+1)}$ chosen in such way that it holds under the following restriction :

$$U(\theta^{(t+1)} \mid \theta^{(t)}) \geq U(\theta \mid \theta^{(t)}), \quad \forall \boldsymbol{\theta} \tag{3.39}$$

The remainder $W(\theta^{(t)} \mid \theta^{(t)}) - W(\theta^{(t+1)} \mid \theta^{(t)})$ can be written as

$$\int \log \left[\frac{f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta^{(t)}}{f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta^{(t+1)}} \right] f_2(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \theta^{(t)}) d\mathbf{x}_{mv} \tag{3.40}$$

and it is very easy to show that quantity (3.39) must be non-negative by using Jensen's inequality and the convexity of the $x \log x$ function.

The difference in the above log likelihood functions follows a chi-squared distribution when the log likelihoods hold under certain conditions (J. L. Schafer & Graham, 2002; Wilks, 1938). The difference in number of parameters estimated between two given models is equal to the degrees of freedom of that chi-squared statistic.

The imputation in the expectation step for the data, which assumed to be multivariate normal, is the same as running the regression on the missing observations (Allison, 2002). The EM is described as an oversimplification. The EM algorithm has software that are designed for estimation of missing data problems are available in both commercial and freely statistical software packages. The most popular statistical software packages including SPSS, SAS and S-Plus. We also have standalone software packages, which is EMCOV (Graham & Hofer, 1991), and Amelia (King et al., 2001) uses the EM procedure. These softwares run the EM algorithm steps and provide different parameter estimates based on ML. The advantages of ML and EM algorithm for managing missing data is that they allowed to be used when missing data are ignorable. Under this property, the ML parameter estimates are known to be consistent and efficient for large samples. The statistical hypothesis testing for the model-based approaches provide small assistant in this area. Especially, the EM algorithm tends to underestimate standard errors, which are critical to hypothesis testing (Allison, 2002). The underestimation of standard error has negative effects on parameter estimation and type I errors tends to be large. The greatest negativity for underestimation of standard errors is that an influence of missing data cannot be estimated and then this cannot be used to provide correct estimates of standard errors. If the analysis has incorrect standard errors, it negatively affects hypothesis testing, which leads to a greater likelihood of type I errors. Thus, we can use a Direct ML method to obtain the correct standard error estimates, so that we can read off a negative influence of low standard error estimates in the analysis.

3.5.4 The ECM algorithm

The Expectation-Conditional Maximization algorithm is the expansion of EM algorithm under the situation where M-step in the EM algorithm has no close form solution (Meng & Rubin, 1993). The purpose of proposing ECM is to solve the complicated M-step from EM algorithm by introducing a new computationally simpler Conditional maximization (CM) steps. The ECM algorithm has some disadvantages over EM. One of the disadvantages is that ECM converges more slowly than EM algorithm in terms of number of iterations needed for convergence, although it is a little bit faster in computing time over EM. The advantage of ECM algorithm is that it converges at the same or approximately the same rate as EM and also maintaining the monotone convergence property of the EM algorithm.

Let the CM step of the ECM algorithm contain a sequence of S conditional maximizations steps i which the U function is maximized as its defined in (3.71), thus, this must be done over the entire parameter θ but with the same vector function of θ. The vector function of parameter θ is denoted by $h_s(\theta)(s = 1, 2, 3, ..., S)$ is fixed at its previous value of s. The set of $s = 1, 2, 3, ..., S$ of the function $H = (h_s(\theta))$ must be selected in advance and also satisfy adequate conditions described in Meng & Rubin (1993). However, the ECM algorithm must be implemented when the all conditions above satisfied. Now the one iteration can be defined as follows.

Let $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$ be a complete data, where $\mathbf{x}_{ov}$ denote observed data and $\mathbf{x}_{mv}$ denote the missing data. Also let θ be the vector of the unknown parameters, and h be a complete data density function. The observed data log-likelihood can be expressed as

$$L'(\theta, \mathbf{x}_{ov}) = \log \int_{\mathbf{x}_{mv}} h(\mathbf{x} \mid \theta) d\mathbf{x}_{mv} \tag{3.41}$$

The EM algorithm is representing the special case of ECM algorithm by maximizing the log-likelihood function $L'(\theta, \mathbf{x}_{ov})$ using the following steps, given an initial $\theta^{(0)}$. At first, we perform the E-step using the current value of parameters $\theta^{(t)}$ to determine the $U(\theta \mid \theta^{(t)})$ as in the EM. Therefore, we proceed to find the updated parameter $\theta^{(t+1)}$ by maximizing $U(\theta \mid \theta^{(t)})$ under constraint

$$h_s(\theta) = h_s\left(\theta^{\frac{t+(s-1)}{S}}\right) \tag{3.42}$$

for $s = 1, 2, ..., S$.

E-step : Compute $U(\theta \mid \theta^{(t)})$,where

$$U(\theta \mid \theta^{(t)}) = E_{\theta^{(t)}}[L(\theta, \mathbf{x}) \mid \mathbf{x}_{ov}, \theta^{(t)})] \tag{3.43}$$

CM-step : for each $s = 1, 2, ..., S$, find parameter value $\theta^{(t+1)} = \theta^{\frac{t+s}{S}}$ which is the input in next E-step such that

$$U(\theta^{\frac{t+s}{S}} \mid \theta^{(t)}) = \max_{\theta} U(\theta \mid \theta^{(t)}) \tag{3.44}$$

under the constraint $h_s(\theta) = h_s\left(\theta^{\frac{t+(s-1)}{S}}\right)$.
This algorithm has, under regularity conditions described in Meng & Rubin (1993) for ECM, and Dempster et al. (1977) and Wu (1983) for EM, the properties that the likelihood is increased with each iteration, and the limit point of the generated parameter sequence corresponds to a stationary point of the likelihood.

3.5.5 Weighting methods

Weighting methods introduced by Flanders & Greenland (1991) and Zhao & Lipsitz (1992), are based on observed values in the datasets. In the Application of Weighting methods, we exclude all missing values from the analysis, and then we are left with only observed values in the analysis. The remaining observed values are weighted in correspondence with how their distribution predicts the complete sample or population. Most of the researchers employ weighting methods to correct for either standard errors associated with the model estimated parameters or the population variability in the data set. The huge discussion of the weighting methods is documented by Kalton & Flores-Cervantes (2003) literature. This literature provide a detailed review, and the weighting process stages involved.

To derive suitable weights, the predicted probability of each response is estimated from the data for the variable with missing values. There are several discussions by D. B. Rubin (1987a,b) based on survey data applying and estimating weighting methods. Under a suitable joint model for the values and variables, these weighting methods are, in many instances, expected to produce results similar to those of multiple imputation (J. L. Schafer & Graham, 2002). L. Robins et al. (1995) developed a new weighting method called: The Weighted Regression model under the field of Biostatistics, which requires an explicit model for the missingness but relaxes some of the parametric assumptions of the data model. The weighted regression model is an extension of the generalized estimating equations (GEE) that was proposed by

Liang & Zeger (1986). GEE method can also be used after Multiple Imputation (MI) and hence the so-called MI-GEE approach J. L. Schafer (2003). Based on the GEE method (Liang & Zeger, 1986), the new method was developed so-called weighted generalized estimating equations (WGEE) to solve the problem of biasness caused by excluding missing values.

The difference between classical GEE method and WGEE method is that, classical GEE is only valid under MCAR assumption, whereas, the WGEE method was developed to work on MAR as well as MNAR mechanisms, provided that the missing data model based on observed data or variables is correctly defined or described (D. B. Rubin, 1996). The WGEE method was improved by Birhanu et al. (2011) to a new method known as the doubly-robust estimating equations (DREEs). Some literature for further explanation about the extended method (WGEE) is detailed in D. B. Rubin (1996) and Rotnitzky et al. (1998). Recently application studies show that weighting methods can be performed in the most popular software's, such as STATA, SAS and SUDANA. There are several studies that applied weighting methods that is shown by Schluchter & Jackson (1989); Ibrahim (1990); Lipsitz & Ibrahim (1998, 1996); Horton & Laird (1999); Seaman & White (2014).

3.6 Bayesian data augmentation technique

3.6.1 Introduction

Bayesian data augmentation uses the Markov Chain Monte Carlo (MCMC) method to simulate random numbers. Firstly, define what Markov Chain Monte Carlo method is and how Gibbs sampling connects with data augmentation method. When the missing data is present, the main aim for the researchers is to yield unbiased estimates to make valid conclusions and decisions. This can be achieved by not including observed data only but also by including missing data. The benefits of MCMC method is that the observed data needs to be added with simulated values of the unobserved data to get adequate regression estimates. This section explains MCMC method idea and how to use it to simulate imputation values from the target distribution.

3.6.2 Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo is a method of generating random numbers directly and draws from a complex probability distribution via the Markov Chains procedure. A Markov chain is a stochastic model that describes a sequence of possible random values, where the probability for each value depend on previous value (dependence). Markov Chain Monte Carlo (MCMC) also uses the same idea, the previous sample observation is used to randomly simulate the next sample observation. In recent years, these techniques have become a subject of interest for Statisticians, who produced a wide range of applications and innovative theory framework. The main aim of using Markov Chain Monte Carlo is to simulate the one or many random variables **x**, where **x** is the matrix with two or more variables.

Let's consider the density function of **x** given by $P(\mathbf{x}) = f(\mathbf{x})$. This density function of **x** is also called a target distribution. Considering that the density function $f(\mathbf{x})$ is a complex function, it is very difficult to draw directly from this density function. The MCMC gives us the way to simulate values. We generate a sequence of random variables rather than to try to draw density function $f(\mathbf{x})$, such that $(\mathbf{x}_{(1)}, \mathbf{x}_{(2)}, ..., \mathbf{x}_{(t)})$ is a given sequence of random variables. Where each variable in the sequence of random variables above depends in the other and where the stationary distribution (i.e. the limiting marginal distribution of $\mathbf{x}_{(t)}$ as $t \to \infty$) is the target distribution $f(\mathbf{x})$ (J. L. Schafer, 1997). The random variable $\mathbf{x}_{(t)}$ is approximately a random draw from density function $f(\mathbf{x})$ as t become sufficiently large (J. L. Schafer, 1997).

Markov Chain Monte Carlo is used when the density function $f(\mathbf{x})$ is intractable to draw from directly; but is simple to drawing each variable in a given sequence. In general, most of Markov Chain Monte Carlo methods are classified under the Bayesian framework especially in posterior distributions simulation because most recently known applications have Bayesian framework. The two most popular methods of Markov Chain Monte Carlo (MCMC) are Gibbs sampling (Geman & Geman, 1984; A. E. Gelfand & Smith, 1990) and the Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970). In Gibbs sampler (Geman & Geman, 1984; A. E. Gelfand & Smith, 1990) procedure, it allows sampling from the conditional of each parameter given that all other parameters and observed data $\mathbf{x}$ is known. In the Metropolis-Hasting algorithm (Metropolis et al., 1953; Hastings, 1970) one draws from a probability distribution calculated to approximate the distribution of interest, and we can use a specified probability to Accept or Reject the drawn observation.

The following list of Markov chain Monte Carlo methods are statistically to proven to be most useful in the analysis of incomplete multivariate data. The resources to review this techinique in a more general setting and additional references are given by A. E. Gelfand & Smith (1990); the articles by Geyer (1992) and Smith & Roberts (1993), Gelman & Rubin (1992) with powerful discussions; and Tierney (1994). Markov chain Monte Carlo applications are mentioned and discussed by M. P. Gelfand et al. (1990); Casella & George (1992); Smith & Roberts (1993); Gilks et al. (1993), among others. A overview including theory and applications in the books by C. A. Tanner et al. (1993); Gilks et al. (1995) . In this thesis, let focus only on Gibbs sampling method (Geman & Geman, 1984) because there is link between Gibbs sampling and Bayesian data augmentation (M. A. Tanner & Wong, 1987b).

3.6.3 Gibbs Sampling

The Bayesian data augmentation (DA) (M. A. Tanner & Wong, 1987b; Li, 1988) has the connection with a Gibbs sampler (Geman & Geman, 1984; A. E. Gelfand & Smith, 1990),which is one of the popular MCMC sampling techniques in the Bayesian framework literature. Suppose that a random vector of the unknown parameters $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_k]^T$ is split into $i = 1, 2, .., k$ subvectors

$$\boldsymbol{\theta}^{(t)} = \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, .., \theta_k^{(t)} \tag{3.45}$$

Let us assume the joint probability density function θ is given $P(\theta)$. Suppose that $P(\theta)$ is very complex and $P(\theta_i \mid \theta_1, ..., \theta_{i-1}, \theta_{i+1}, .., \theta_k)$ full conditional distribution is known and easy to be simulated. The Gibbs sampling (Geman & Geman, 1984; A. E. Gelfand & Smith, 1990) draws the parameters iteratively from the conditional distribution of each parameter given all the other remaining ones. At the t iteration, the corresponding value of is denoted by

$$\boldsymbol{\theta}_k^{(t)} = \theta_1^{(t)}, \theta_2^{(t)}, \theta_3^{(t)}, .., \theta_k^{(t)} \tag{3.46}$$

Moreover, we can find the value of θ at $t + 1$ iteration by successively drawing from the following distribution

$$\boldsymbol{\theta}^{(t+1)} \sim P(\theta_k \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, .., \theta_{k-1}^{(t+1)}) \tag{3.47}$$

We can choose a starting point where $t = 0$ and $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_k^{(0)})$. The Gibbs sampler algorithm has the following iterations

- Draw $\boldsymbol{\theta}_1^{(t+1)} \sim P(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, .., \theta_k^{(t)})$
- Draw $\boldsymbol{\theta}_2^{(t+1)} \sim P(\theta_2 \mid \boldsymbol{\theta}_1^{(t+1)}, \theta_3^{(t)}, .., \theta_k^{(t)})$
- ... ...
- Draw $\boldsymbol{\theta}_k^{(t+1)} \sim P(\theta_k \mid \boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, .., \boldsymbol{\theta}_{k-1}^{(t+1)})$

If t becomes large, the random vector $\theta^{(t)}$ at t iteration must converge in the stationary distribution $P(\boldsymbol{\theta})$ (Geweke et al., 1994). According to M. P. Gelfand et al. (1990) Gibbs sampler under mild conditions, the sequence of vector $[\theta(t)]_{t=1}^{\infty}$ converge into a stationary distribution $P(.)$. Schervish & Carlin (1992) discusses a sufficient condition in which the geometric convergence is guaranteed. According to J. L. Schafer (1997), Bayesian data augmentation (DA) is closely related to Gibbs sampling method, more explanation in the literature J. L. Schafer (1997). The purpose of Gibbs sampling is to know about missing data problems, how it draws the

posterior predictive distribution of the missing values that are included in the sequence of random variables. The Bayesian is the stochastic party that link with steps in Expectation-Maximum. M. A. Tanner & Wong (1987a) shows that we can draw from the posterior predictive distribution for missing data $\mathbf{x}_{mv}$ and draw from the posterior distribution $\boldsymbol{\theta}$.

M. A. Tanner & Wong (1987a) use the Gibbs sampler to drawing for imputation (I) step

$$\mathbf{x}_{mv}^{(t+1)} \sim P(\mathbf{x}_{mv} \mid \mathbf{x}_{ov}, \boldsymbol{\theta}^{(t)}) \tag{3.48}$$

The posterior (P) step is where we draw the parameter estimates for the algorithm

$$\theta^{(t)} \sim P(\boldsymbol{\theta} \mid \mathbf{x}_{ov}, \mathbf{x}_{mv}^{(t+1)}) \tag{3.49}$$

The data augmentation can be converted from Gibbs sampling by letting $k = 1$ to reducing it into a special case of the Gibbs sampling algorithm.

3.6.4 Markov Chain Monte Carlo method in the presence of missing data

In 1987, Tanner & Wong published the seminal paper where they introduced the concept of data augmentation approach in the Bayesian framework to determine posterior distribution using the iterative procedure. The Tanner and Wong data augmentation algorithm is closely related to Gibbs sampling method. Let **x** be a random subvector that is separated into two sub-vectors $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$. where, the joint distribution of $P(\mathbf{x})$ is very difficult to compute but $P(\mathbf{x}_{ov}, \mathbf{x}_{mv}) = f(\mathbf{x}_{ov}, \mathbf{x}_{mv})$ and $P(\mathbf{x}_{ov}, \mathbf{x}_{mv}) = z(\mathbf{x}_{mv}, \mathbf{x}_{ov})$ conditional distributions can be established. However, at iteration t the vector becomes

$$\mathbf{x}^{(t)} = \left(x_1^{(t)}, x_2^{(t)}, ..., x_n^{(t)}\right) \tag{3.50}$$

$$= \left(\left(x_{1,ov}^{(t)}, x_{1,mv}^{(t)}\right), \left(x_{2,ov}^{(t)}, x_{1,mv}^{(t)}\right), ..., \left(x_{n,ov}^{(t)}, x_{n,mv}^{(t)}\right)\right) \tag{3.51}$$

where n denote the sample size from the distributions that estimate the unknown distribution $P(\mathbf{x})$. Then, we can update this sample using two important steps.

Step 1: We create the the updated vector $\mathbf{x}_{ov}$ with sample n which is denoted as below

$$\mathbf{x}_{ov}^{(t+1)} = \left(x_{1,ov}^{(t)}, x_{2,ov}^{(t)}, ..., x_{n,ov}^{(t)}\right) \tag{3.52}$$

and it is drawn from

$$x_{i,ov}^{t+1} \sim f\left(x_{ov} \mid x_{i,mv}^{(t)}\right) \tag{3.53}$$

where $i = 1, 2, 3, ..., n$.

Step 2: We create the the updated vector $\mathbf{x}_{mv}$ with sample n which is denoted as below

$$\mathbf{x}_{mv}^{(t+1)} = \left(x_{1,mv}^{(t+1)}, x_{2,mv}^{(t+1)}, ..., x_{n,mv}^{(t+1)}\right) \tag{3.54}$$

and it is drawn from the independent and identical distributed sample

$$x_{i,mv}^{t+1} \sim z\left(x_{mv} \mid x_{i,mv}^{(t)}\right) \tag{3.55}$$

where $i = 1, 2, 3, ..., n$.. Then, the weighted or average mixture of the conditional

distribution and it drawn as iid sample

$$\bar{z}\Big(x_{mv} \mid x_{i,ov}^{(t+1)}\Big) = \frac{1}{n}\sum_{i=1}^{n} z\Big(x_{mv} \mid x_{i,ov}^{(t+1)}\Big) \tag{3.56}$$

which made the new updated sample

$$\mathbf{x}_{(t+1)} = \left(\Big(x_{1,ov}^{(t+1)}, x_{1,mv}^{(t+1)}\Big), \Big(x_{2,ov}^{(t+1)}, x_{2,mv}^{(t+1)}\Big), ..., \Big(x_{n,ov}^{(t+1)}, x_{n,mv}^{(t+1)}\Big)\right) \tag{3.57}$$

In the paper of Tanner & Wong (1987) shows that the distribution of $x(t)$ converges to $P(x)$ as values of t become very large by using functional analysis. This requires the sample size of $n = 1$, then data augmentation becomes reduced to a special case of Gibbs sampler with a vector of $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$ separated into two parts which is missing and observed parts. After letting $n = 1$, we can modify Step 2 of each iteration by sampling

$$x_{mv}^{(t+1)} \sim z(x_{mv} \mid x_{i,ov}^{(t+1)}) \tag{3.58}$$

with independency for $i = 1, 2, ..., n$.

However, now we can draw from equation (3.58), but when we draw from the mixture (3.56), then this algorithm becomes n independent parallel runs of a Gibbs sampling. The Markov chain Monte Carlo data augmentation is one of the procedures which is popular for handling missing data and the advantage for this procedure is that it shows greater flexibility when the underlying distributions are unknown. While, an ML model-based methods only function under the certain distributional assumptions, for example, multivariate normality etc. Under MCMC, we have a different set of procedures for simulation of random values but a method that is most commonly used for applying MCMC is Gibbs sampling since it most available in the statistical software. The MCMC approaches are related to Bayesian estimation methods and most of the researchers view MCMC as a Bayesian simulation. In physics, MCMC was implemented to investigate equilibrium of interacting molecules distributions. In Statistics, the MCMC process is usually implemented in parameter estimation when data has missing values and when the given distribution does not hold under the assumptions of ML process. This process can be called Bayesian since; we can obtain a probability distribution known as posterior distribution. A posterior distribution can be used for parameter estimation in the Bayesian analysis so that we can make proper inferences.

A posterior distribution is a probability distribution of unobserved parameter estimates that follows the observed data and updates the statistical model by using the

information given from the data (Gill et al., 2002). The standard MCMC methods such as Gibbs sampling, Metropolis and much more are used to simulate values independently following a certain probability distribution. The standard Monte Carlo methods produce simulated values using the Markov chain procedure. A Markov chain is a stochastic model that describe a sequence of possible random variables in which the probability for each value depends on the previous value. Therefore, we conclude that the MCMC process produces simulated random values that are dependent on each other since the previous outcomes influence predictions for the next experiment outcomes. In addition, a Markov chain is characterized as a stochastic process moving from one value to another until it finds or generates the posterior distribution for parameter estimation(Gill et al., 2002). In the Bayesian statistics, the main aim is to calculate the posterior distribution for $\boldsymbol{\theta}$. Where $\boldsymbol{\theta}$ is the parameter of interest, so that we can obtain the parameter estimates and make inference in the Bayesian analysis. The data augmentation can be used to compute maximum likelihood estimates and this means that the algorithm can also be used in the computation of the posterior distribution (Tanner & Wong, 1987)..

Given the data with missing observations, our focus is to produce parameter estimates that are unbiased but this is very difficult since we are used only using the observed data and ignore unobserved data. The main aim of this process is to create a distribution of estimates and randomly select the unobserved values from the given distribution. The MCMC procedure is still a growing literature in Bayesian statistics. However, this procedure also became popular in terms of computational software, which include Bayesian methods in software packages. The MCMC software was implemented in different research papers such as MCMC in the SAS Proc MI (Multiple imputation procedures), new R/S-Plus functions and stand-alone programs such as WinBUGS (Gilks et al., 1994).

3.6.5 Summary advantages of data augmentation

The Markov chain Monte Carlo (MCMC) is one of the procedures that is popular for handling missing data and the advantage for this procedure is that it shows greater flexibility when the underlying distributions are unknown. The method that is most commonly used for applying MCMC is Gibbs sampling since it most available in the statistical software. Given the data with missing observations, our aim is to produce parameter estimates that are unbiased but it is very difficult when we only using the observed data and ignore unobserved data. The MCMC procedure allows for

augmenting the observed data $\mathbf{x}_{ov}$ with simulated values of the missing data $\mathbf{x}_{mv}$ to solve the issue associated with parameter estimation. The common MCMC procedures have desirable features than ML model-based procedures are efficiency and flexibility. The MCMC process is efficient because it allows us to estimate parameter even if the unexpressed distributions are unknown or non-normal distributed. The advantage of using software MCMC procedure (Bayesian data augmentation) is that it always find the solutions even for most complex missing data problems. Especially when given data distribution is unknown and does not follow the multivariate normal distribution. In the iterative process, the EM is restricted by the expectation derived from a distribution in order to estimate parameters while the MCMC methods are unlimited by distributional assumptions.

3.6.6 Disadvantages of data augmentation

There are many technical problems we face when we implement data augmentation algorithm since we use computer programs. Let the augmented data given as $\mathbf{x} = (\mathbf{x}_{ov}, \mathbf{x}_{mv})$. If the computed augmented data have more regressors or predictors than observations, then we must propose values to a much larger number of parameters than the known information. This may increase the computational load problem and it makes it harder to assess convergence because there are too many additional and unwanted parameters to justify storing in memory of the computer. Another problem is that most augmented data are correlated with each other in terms of parameters included in the model, so it makes it difficult to design adequate distributions.

3.7 Simulation of missing data procedure

In this thesis study, we assume that missing values are ignorable, which means that the missing data can be estimated based on the observed data. To create a data set with missing values, the baseline data (complete data) was used to create missingness using different missing mechanism (MAR, MCAR and MNAR). The data set for MAR was created in such a way that missing values was not related to missing data itself but related to variable of interest. Under MCAR, the data set was independent of observed and unobserved variables since the missingness was randomly sampled. The data is MNAR, when the missingness in the data cannot hold either MCAR or MAR and data is related to missing data even after including observed data. The use of these assumptions is to give insight on the performance of data augmentation algorithm under different rates or proportions of missing data and estimate the parameters of each model that corresponds to proportion rates of missingness..

3.7.1 Preparation of data for data augmentation algorithm

After creating different missing data percentages, a data augmentation algorithm was used to impute missing values in different dataset that corresponds to missing information created. The data augmentation has two iterative steps which is I-step and P-step to simulate missing values from the predictive distribution. Since the given data set for cancer medication intake in South Africa is the mixed data, which means that it contains both categorical and continuous variables. The package to use is called **mix** in R software for data augmentation algorithm under mixed data. The package **mix** in R is easy to use. The researcher are more certain about this package because it produces improper posterior when it has structural zeroes in the contingency table. Under the improper posterior, the final results cannot be computed and we are unable to do parameter estimation to make inferences about the distribution. The package **mix** used to impute the data frame with missing values (NA), where NA represent missing values in R package and produces the full imputed data (more information on **Figure** 3.2 below).

In the study, we ran a different number of datasets that are created using different missing percentages $1\%, 5\%, 10\%$ in the variables working for wage and age from the original data, and then apply data augmentation algorithm to impute those missing values. The binary logistic regression model was used to investigate whether the proportions of missingness can impact on the performance of the data augmentation used. The main goal is to compare the original data and imputed data sets to

identify which one gives the closest estimates or standard errors or AIC to those of original data. In order to find out that the method of data augmentation is improving data analysis by comparing the produced means and standard deviations for imputed data on the continuous Age variable for different percentages of missing information, thereafter, the reasonable diagnostics can be performed. The package **mix** depends on the data size and number of missing information in terms of computation time.

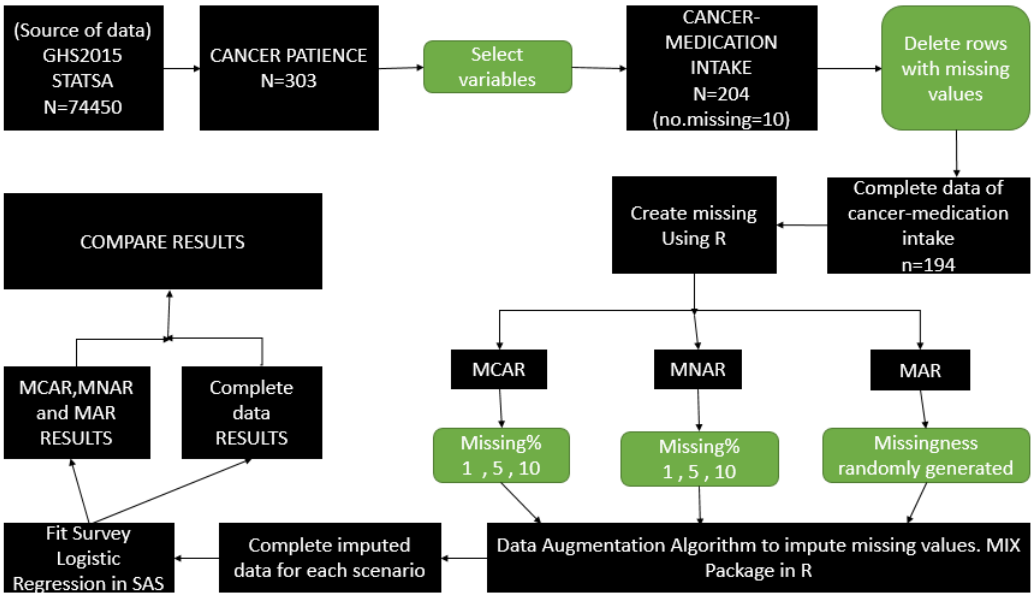

Figure 3.2 – Diagram of the data augmentation and process flow/procedure

Data preparation principles

- Only cancer patience are considered for data preparation.
- Either patience is on cancer medication or not.

3.8 Statistical tools for data analysis

This section gives brief information about the statistical tools used in the analysis. It is useful to use an appropriate statistical tools after data collection and data visualization, in order to make good decisions using that particular data. This analysis is classified into two types which is univariate analysis and bivariate analyses. The data analysis is a very important part when you are dealing with data because it is where one can extract information from the collected data. In this study, we used two methods of data analysis which is descriptive and inferential statistics methods.

3.8.1 Descriptive Statistics

The univariate analysis is used to summarize the data and it produces quick summary information and characteristics of each variable in the data set. This information can be expressed in the form of descriptive statistics and diagrams. The descriptive statistics that are used in this study are summary statistics (Means, Standard deviations, Medians) and frequency distributions. In diagramatic format, we use Bar graphs and Pie chart to visualize the cancer medication intake data.

3.8.2 Inferential Statistics

The inferential statistics used are based on the nature of the study and objectives of the study these are as follows: Chi-square, Paired t-test, Wilcoxon signed-rank test and Kappa test. This section gives the partial theoretical framework for each inferential statistics used in the study (Bivariate analyses, model fitting, model diagnostics and test).

Chi-square

There are many different types of Chi square tests, the two most often used, and look at whether there are potential associations between categorical variables are a chi-square test of independence or a chi-square test of homogeneity. A Chi-square test of independence are used to determine if two variables are related in any way or not, while Chi-square tests of homogeneity are used to determine whether the distribution of one categorical variable, is similar or different to the other, across all the levels of the second categorical variable. Let's denote a qualitative dependent variable Z with r_n categories $(Z_1, Z_2, ., Z_i, ..., Z_{r_n})$ and the explanatory variables X

with c_n categories $(X_1, X_2, ., X_j, ..., X_{c_n})$ that is sampled from the sample of n observations.

In our analysis, Chi-squared test is used for testing independence and homogeneity to investigate if there is any association between some of our categorical variables and to test whether the categorical variable, are similar or different before and after imputation. We defined the null and alternative hypotheses related to Chi-square as follows:

$\mathbf{H}_0$: No association between between two categorical variables.

$\mathbf{H}_1$: There is an association between two categorical variables

The chi square test statistics is:

$$\chi^2 = \sum_{i=1}^{r_n} \sum_{j=1}^{c_n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.59}$$

where r_n is row and c_n is column, O_{ij} - observed frequency and E_{ij} - expected frequency. The degrees of freedom is given by $(r_n - 1)(c_n - 1)$. If the χ^2 calculated value is larger than the χ^2 critical value reject null hypothesis $\mathbf{H}_0$ and support the alternative hypothesis.

Paired t-test

Most of the Researchers use the paired t-test to examine for a mean difference between matched data observations (Hsu & Lachenbruch, 2008). Suppose that you want to compare two paired samples, y and x and we calculate differences $d_i = y_i - x_i, i = 1, 2, ..., n$. Under this test the aim is to test whether the mean of the differences was statistically different from zero (David & Gunnink, 1997).

The null and Alternative Hypotheses of the paired t-test.

The basic null hypothesis of the paired t-test becomes

$$\mathbf{H}_0 : \mu_{diff} = 0 \tag{3.60}$$

with one of the alternative hypothesis options listed below

$$\mathbf{H}_{1a} : \mu_{diff} \neq 0, \tag{3.61}$$

$$\mathbf{H}_{1b} : \mu_{diff} < 0, or \tag{3.62}$$

$$\mathbf{H}_{1c} : \mu_{diff} > 0. \tag{3.63}$$

Paired t-test Assumptions

The assumptions of the paired t-test are:

- The given data set are continuous.
- The difference for paired data must be normally distributed.
- The sample must be randomly selected from the population.

The test statistic is calculated by

$$t = \frac{\bar{d}}{s_{\bar{d}}} \tag{3.64}$$

where

$$\bar{d} = \frac{1}{n}\sum d_i \quad and \quad s_{\bar{d}} = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n(n-1)}}. \tag{3.65}$$

Therefore, we can construct a $(1-\alpha)$ confidence interval (CI) for the mean μ_d. The confidence interval is found to equal to :

$$\left(\bar{d} - t_{\frac{\alpha}{2},n-1}s_{\bar{d}} \quad , \quad \bar{d} + t_{\frac{\alpha}{2},n-1}s_{\bar{d}}\right) \tag{3.66}$$

where $t_{\frac{\alpha}{2}}$ is the upper significance of $t-test$ with $n-1$ degrees of freedom. This test is valid only under assumption that d_i are independent and identically distributed (iid) and $\bar{d}$ be normal and independent of $s_{\bar{d}}$, where $(n-1)s_{\bar{d}}^2 \sim \chi_{n-1}^2\sigma_{\bar{d}}^2$.

The Wilcoxon signed-rank test

The Wilcoxon signed rank test, also known as the Wilcoxon matched pairs test, is a non-parametric statistical hypothesis test used to test the median difference in paired data. The Wilcoxon signed rank test is the non-parametric test that is used as an alternative of the paired t-test for dependent samples when the population is assumed to be not normally distributed (Lam & Longnecker, 1983). The paired t-test assumes that the data is measured on an interval or a ratio scale that follows the normal distribution. This test can be used to detect whether two dependent samples were selected from populations having the same distribution, when the assumptions of the t-test are violated (Gibbons & Chakraborti, 2011).

The distinction between parametric and non-parametric techniques is discussed by Clutton-Brock et al. (1998). The difference between parametric and non-parametric methods is that parametric methods have distributional assumptions, that data is normally distributed whereas non-parametric are not normally distributed. The distributional assumption under Wilcoxon signed rank test can be avoided because the test is based on the rank order of the differences rather than the actual value of the differences. However, it is necessary to make an assumption that the distribution of the differences is symmetric although is assumed to not be normally distributed. The Wilcoxon signed rank test is based on the magnitude of the difference between the pairs of sample values. To calculate the difference between the pairs of observations, the actual data points in the sample must be measured on an interval scale, as is required for the t-test. The Wilcoxon test statistic W_{test} is given by the sum of the positive ranks in the data set as:

$$W_{test} = \sum_{i=1}^{n"} RANK_{i(+)} \tag{3.67}$$

The null and alternative hypotheses for Wilcoxon test in our case are

$\mathbf{H}_0$: Mean difference between the two paired measurements is zero

$\mathbf{H}_1$: Mean difference between the two paired measurements is not equal to zero

Assumptions

1. Data is paired and comes from the same population.

2. Each pair is chosen randomly and independently

3. The data are measured at least on an ordinal scale (i.e., they cannot be nominal).

The Wilcoxon signed rank test sums the ranks of the positive differences and sum the ranks of the negative differences. The test statistic is the lesser of these two sums. If the p-value is greater than 0.05, then null hypothesis is not rejected which means that there was no difference, then we would expect the rank sums for both positive and negative ranks to be the same. Further literature on calculation and interpretation of the Wilcoxon signed rank test is found in Bland (1995) and Conover & Conover (1980).

The Kappa Test

A common application of the Kappa test, in a situation when a researcher needs to assess agreement on a nominal scale data, is to determine the presence or absence of some disease or conditions. The Kappa statistics is the measure of inter-variation across the cross-tabulation or it measures the inter-rater agreement between two or more raters. Under Kappa statistics, we assume each value in the cross-tabulation is called a subject. The variables, however, record frequencies with which rating was assigned.

Let N be the total number of subjects, each independently assigned to one of j categories by two separate raters. These would result in a $j \times j$ contingency table. P_{ik} denotes the percentage rate of subjects that Rater A, classified in category i in contingency table. The Rater B classified in category k in contingency table, with i, k = 1, 2,..., j. The frequencies are given by $P_{i.}$ and $P_{.k}$ are assigned into each categories i and k with respect of Rater A and Rater B. The category of each Rater must sum to 1.

Table 3.2 Kappa Test for Agreement Between Two Raters

	Rater B				
Rater A	**1**	**2**	...	**j**	Total
1	P_{11}	P_{12}	...	P_{1j}	$P_{1.}$
2	P_{21}	P_{22}	...	P_{2j}	$P_{2.}$
3	P_{31}	P_{32}	...	P_{3j}	$P_{3.}$
⋮	⋮	⋮	⋮	⋮	⋮
j	P_{j1}	P_{j2}	...	P_{jj}	$P_{j.}$
Total	$P_{.1}$	$P_{.2}$	...	$P_{.j}$	1

Source: (PASS Sample Size Software, page 1)

The percentage rates on the diagonal of the contingency table, P_{ii}, represent the percentage rates of subjects in each category for which the two raters agreed on the same thing. The main percentage rates of observed agreement are

$$P_O = \sum_{i=1}^{j} P_{ii} \tag{3.68}$$

and the main percentage rates of agreement expected to occur by chance is

$$P_E = \sum_{i=1}^{k} P_{i.} P_{.i} \tag{3.69}$$

The overall value of kappa, which measures all the degree of Rater agreement, is given by:

$$\kappa = \frac{P_O - P_E}{1 - P_E} \tag{3.70}$$

where P_0 is the proportion of observed agreements and P_E is the proportion of agreements that are expected to occur by chance. The data for paired ratings on a 2-category nominal scale are usually displayed in a 2×2 contingency table, with the notation indicated in Table 3.2.

A kappa value of 1 denotes a perfect agreement between the two raters. A kappa value of 0 indicates no more rater agreement than that expected by chance. A kappa value of -1 indicates a perfect disagreement between the two raters. If the Kappa test has a range between 0 and 1 with larger values indicating better reliability. In General, a Kappa test greater than 0.70 is considered satisfactory. But if Kappa test is less than 0.70, we can conclude that the inter-rater reliability is not satisfactory.

Interpretation of Kappa

Table 3.3 Interpretation of Kappa

Kappa-value	**Agreement**
<0	Less than chance agreement
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1	Almost perfect agreement

The real value of kappa can be predicted in the given sample by taking the observed and expected percentage rate and replace it by their sample predicted values.

$$\hat{\kappa} = \frac{\hat{P_O} - \hat{P_E}}{1 - \hat{P_E}} \tag{3.71}$$

where

$$\hat{P_O} = \sum_{i=1}^{j} \hat{P}_{ii} \tag{3.72}$$

$$\hat{P_E} = \sum_{i=1}^{j} \hat{P}_{i.}\hat{P}_{.i} \tag{3.73}$$

The possible value of $\hat{\kappa}$ depends on the marginal percentage rate.

3.8.3 Multivariable analysis

Binary Survey Logistic Regression Model

The Binary Survey logistic Regression model is applicable when the response variable is a dichotomous or binary response variable, meaning the dependent variable is always binary in nature. It is used in situations when we want to test for defaults and non-defaults, passed or failed, success or failure and it utilizes a base model (reference group) for comparison purposes. Interpretation of results from Binary Survey Logistic Regression models is based on the probability of an event occurring, or not, as a function of the values (levels) of the independent variables, which can be categorical or numerical. It estimates the probability that an event occurs, for a randomly selected observation with a profile of levels of the independent variables, versus the probability that the event does not occur (Hilbe, 2011).

The data from General Household Survey are collected using multistage sampling with complex sampling design. Therefore, in order to get valid statistical inferences it is essential to account for the complexity of sampling design as failure to do so may result in biased estimates and underestimation of the variabilities. Therefore, in this chapter, we use binary survey logistic regression models. These models offer an option for accounting for complexity of sampling design. The survey sampling design may induce correlation among observations, especially when clusters samples are drawn. To appropriately estimate standard errors associated with the model parameters and estimated odds ratios, it is very crucial to account for sampling design. The survey logistic regression models have the same theory as classical logistic regression models. The only difference is the estimation of the variance. However, when these two models are used to the data collected using simple random sampling, the results are identical.

The Survey logistic Regression Model predicts the effect of a series of variables on a binary response and classifies observations by estimating the probability that an observation has one of the two outcomes under investigation. In this case, the Survey Logistic Regression model can be chosen as the dependent variable; cancer medication chronic; is binary or dichotomous, meaning that it can assume either yes or no. The explanatory variables are Gender, working for a wage, Area of living and Age are mixed variables of both types (continuous and categorical).

Let Y_{ijh} be the response variable, with $i = 1, 2, 3, ..., m_{hj}$, $j = 1, 2, 3, ..., n_h$ and $h = 1, 2, 3, ..., H$, where h is the stratum, j is the cluster and i is the household and denote the sampling weight for ijh^{th} observation as w_{ijh} and x_{ijh} the row vector of the design matrix corresponding to the i^{th} household in j^{th} PSU, nested in h^{th} stratum.

We shall assume that Y_{hij} belongs to the exponential family of distributions with the sampling distribution defined as follows:

$$g(y_{ijh}, \theta_{ijh}, \varphi) = \exp\left[\frac{y_{ijh}\theta_{ijh} - b(\theta_{ijh})}{a(\varphi)} + c\big(y_{ijh}, \varphi\big)\right] \tag{3.74}$$

where $g(.)$ denotes the density function of y_{ijh}, θ_{ijh} is known as the natural parameter and φ is known as the dispersion parameters.

Let the response variable be $y_{ijh} = 1$ if i^{th} household is taking cancer medication and be 0 if not taking cancer medication. The link function for a binary outcome as in this study based on survey logistic regression the link function is $\eta_{ijh} = logit(\mu_{ijh})$ thus the generalized logit model can be written as

$$\text{logit}(\pi_{\text{ijh}}) = \log\big(\frac{\pi_{ijh}}{1 - \pi_{ijh}}\big) = \mathbf{x}_{ijh}^{'}\beta \tag{3.75}$$

where $\pi_{ijh} = E(y_{ijh} \mid \mathbf{x}_{ijh}^{'})$, $\mathbf{x}_{ijh}^{'}$ is a vector of explanatory variables and β is the vector of unknown parameters. When the survey data have been collected under complex sampling design, straightforward application of classical maximum likelihood estimation (MLE) is no longer convenient, for various reasons. The first one is that the probabilities of selection for the $i = 1, 2, ..., n$ sample observations are no longer equal. Sampling weights are then required to estimate the finite population values of the logistic regression model parameters. Secondly the stratification and clustering of complex sample observations violates the assumption of independence of observations that is essential to the standard MLE method (Heeringa et al., 2010).

3.8.4 Odds ratio

The odds ratio is defined as the probability the certain event will happen divided by the probability that the event will not occur (J. Zhang & Kai, 1998). In this study, the odds ratio applied in form of the probability that cancer patient is taking cancer medication divided by probability that patient is not taking cancer medication.

The odds ratio are computed using the formula;

$$Odds = \frac{P(Occur)}{P(Not \quad occur)} = \frac{\pi}{1-\pi} \tag{3.76}$$

where, π is the probability of success and $1-\pi$ is the probability of failure. The odds ratio indicate that the odds of a success and odds of failure are equally likely to happen is given by

$$OddsRatio = \frac{Odds \quad occur}{Odds \quad Not \quad occur} \tag{3.77}$$

The odds ratio of 1 shows that the odds of a successful outcome are equally likely to the odds of a failure of an outcome. In the survey logistic regression the odds ratio is equal to $\exp(\beta)$ = OR and the level of significance (p-value) associated with odds ratio estimates to determine whether significance. There is an odds ratio and significance evaluation for each category of each explanatory variable except the reference category, When the p-value < 0.05 tells us that there is a significant difference in the odds of the outcome occurring between the category of interest and reference category.

For a given category, the closer the odds ratio is to 1, the weaker the association is, or the less significant difference is in the odds. Thus; when, $\beta_i > 0$, then $(\exp(\beta_i) > 1)$, implying that people under this category are $\exp(\beta)$ times more likely (or more at risk) to face the specific event of study (use cancer-medication in our case) than those of the reference category and,

when, $\beta_i < 0$, then $(\exp(\beta_i) < 1)$, implying that the individuals of the specific category are $[1 - \exp(\beta)] \times 100$ percent less likely to face the event under study (cancer-medication intake) than those of the reference category. The the p-value of the odds ratio will help us to conclude about the degree of results of the odds ratio value.

3.8.5 Estimation Method

In this section, we introduce the Maximum Likelihood Estimation (MLE) technique used to estimate parameters in the survey logistic regression model. Due to the complex design sampling properties such as unequal probability of selection, clustering and stratification may not work properly.

Parameters estimation

In this sub-subsection we derive expressions for the maximum likelihood estimators in a typical survey logistic regression. Assuming that the outcomes variable y_{ijh} follows Bernoulli distribution with density function

$$g(Y_{ijh} = y_{ijh}) = \pi_{ijh}^{y_{ijh}}(1 - \pi_{ijh})^{1-y_{ijh}} \tag{3.78}$$

the mean and variance of y_{hij} are respectively,

$$\mu_{ijh} = \frac{\exp[\mathbf{x}_{ijh}^{'}\beta]}{1 + \exp[\mathbf{x}_{ijh}^{'}\beta]} \tag{3.79}$$

and

$$\sigma^2 = \mu_{ijh}\big(1 - \mu_{ijh}\big) \tag{3.80}$$

and $\boldsymbol{\beta} = (\beta_1, \beta_2, ...\beta_p)^{'}$ denote a vector of parameters. The log-likelihood function that forms the basis for maximum likelihood estimation is given by

$$\ell = \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} w_{ijh}\vartheta_{ijh}[y_{ijh}\log(\mu_{ijh} + (1 - y_{ijh})\log(1 - \mu_{ijh})] \tag{3.81}$$

Substituting the values of mean μ_{ijh} into this expression we obtain

$$\ell = \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} w_{ijh}\vartheta_{ijh}\left[y_{ijh}\log\left(\frac{\exp[\mathbf{x}_{ijh}^{'}\beta]}{1 + \exp[\mathbf{x}_{ijh}^{'}\beta]}\right) + (1 - y_{ijh})\log\left(1 - \frac{\exp[\mathbf{x}_{ijh}^{'}\beta]}{1 + \exp[\mathbf{x}_{ijh}^{'}\beta]}\right)\right] \tag{3.82}$$

To obtain the unknown parameters we have to differentiate the log-likelihood with respect to $\boldsymbol{\beta}$ to get the following equation

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} w_{ijh} \vartheta_{ijh} \frac{\exp[\mathbf{x}_{ijh}^{'} \beta]}{\big(1+\exp[\mathbf{x}_{ijh}^{'} \beta]\big)^2} \left[\frac{y_{ijh}}{1-\big(1+\exp[\mathbf{x}_{ijh}^{'} \beta]\big)^{-1}} - \frac{1-y_{ijh}}{1+\exp[\mathbf{x}_{ijh}^{'} \beta]} \right] \mathbf{x}_{ijh}^{'} \tag{3.83}$$

$$= \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} \vartheta_{ijh} \mathbf{A}_{ijh}^{'} \big[\sigma^2(y_{ijh})\big]^{-1} \big[y_{ijh} - \mu_{ijh}\big] \tag{3.84}$$

where $\mathbf{A}_{ijh} = \mu_{ijh}\big(1-\mu_{ijh}\big)\mathbf{x}_{ijh}^{'}$
The Fisher information matrix for the parameters of the Bernoulli model follows as

$$\omega = -E\left[\frac{\partial^2 \ell}{\partial \beta \beta'}\right], \tag{3.85}$$

$$= \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} \mathbf{D}\mathbf{x}_{ijh}^{'2} b c^2 \left[y_{ijh}\big(1-2c^{-1}\big) - \big(1-y_{ijh}\big)\big(\big(1+b\big)^{-1} - 3\big(1+b\big)^{-2} b\big) \right] \tag{3.86}$$

where $\mathbf{D} = w_{ijh}\vartheta_{ijh}$, $b = \exp[\mathbf{x}_{ijh}^{'}\beta]$ and $c = 1+\exp[\mathbf{x}_{ijh}^{'}\beta]$.

After simplifying the previous equation we get the following equation which is referred to as the Fisher Information

$$\omega = \sum_{i=1}^{m_{hj}} \sum_{j=1}^{n_h} \sum_{h=1}^{H} w_{ijh} \vartheta_{ijh} \mathbf{A}_{ijh}^{'} \big[\sigma^2(y_{ijh})\big]^{-1} \mathbf{A}_{ijh} \tag{3.87}$$

Test for goodness of fit

The Likelihood Ratio Test

The likelihood ratio (LR) test evaluates the significance of the joint effect of all the variables in the Survey logistic regression procedure. Likelihood ratio is used to compare the significance of the model with multiple parameters to just the intercept model. Suppose the model contains s explanatory effects. For the i^{th} observation, let $\hat{\pi}_i$ be the estimated probability of the observed response. The statistic of -2loglikelihood is given by:

$$-2\log L = -2\sum_i w_i f_i \log(\hat{\pi}_i \tag{3.88}$$

where, w_i and f_i are weight and frequency values, respectively, of the i^{th} observations. For binary response models that use the events/trials, this is equivalent to

$$-2\log L = -2\sum_i w_i f_i \big[r_i \log(\hat{\pi}_i) + (n_i - r_i)\log(1-\hat{\pi}_i) \tag{3.89}$$

where r_i is the number of events, n_i is the number of trials, and $\hat{\pi}_i$ is the estimated event probability. The likelihood ratio comparing the log likelihood of the two models and if the difference is statistically significant, then the restricted model works better than the full model. The significance of the likelihood ratio test means that the joint of the variables in the full model is more significant than just the intercept model. Under the global null and alternative likelihood ratio tests has the following hypothesis :

$\mathbf{H}_0$: $\boldsymbol{\beta}_i = 0 \quad , for \quad i = 1, 2, ..., p$

$\mathbf{H}_1$: Not all $\boldsymbol{\beta}_i = 0 \quad , for \quad i = 1, 2, ..., p$

The test statistic of the likelihood ratio test follows a chi square distribution with p degrees of freedom according to Prempeh (2009). The log-likelihood ratio is defined as the difference between the deviance of the null model and model with explanatory variable(s).

$$Loglikelihood - Ratio = D^v_{null} - D^v_{p-1} \tag{3.90}$$

where D^v_{null} is the deviance of the model with just the constant and D^v_{p-1} is the deviance of the model with $p-1$ parameters.

Wald Test

The Wald test is the additional test that also can be used to evaluate the significance of the individual parameters in the Surverylogistic Regression procedure. The expression for calculating the Wald statistic is given by

$$Wald = \frac{\hat{\beta}_i}{S_E(\hat{\beta}_i)} \tag{3.91}$$

where, $\hat{\beta}_i$ is the regression coefficient estimates of the explanatory variables and $S_E(\hat{\beta}_i)$ is the standard error of the corresponding regression coefficient estimate. According to Rana et al. (2010), a squared value of the Wald statistics is chi-square

distributed with one degree of freedom.

$$(Wald)^2 = \frac{\hat{\beta}_i^{\,2}}{\left[S_E(\hat{\beta}_i)\right]^2} \tag{3.92}$$

The Wald statistic test has the following null and alternative hypotheses:

$\mathbf{H}_0$: $\boldsymbol{\beta}_i = 0 \quad , for \quad i = 1, 2, ..., p$

$\mathbf{H}_1$: Not all $\boldsymbol{\beta}_i = 0 \quad , for \quad i = 1, 2, ..., p$

If testing one parameter at a time, Wald statistic follows the χ^2 distribution with 1 degree of freedom. The null hypothesis is rejected if the p-value $< 0.05 = \alpha$, where α is the level of significance. A regression coefficient estimates with a p-value of the Wald statistic < 0.05 implies that the variable is important in the current model.

Akaikes Information Criterion (AIC)

Akaike's Information Criterion (AIC) measures the quality of each model fit in the survey logistic regression, compared to other models based on the final model. The AIC can be used to select the best model. Akaike's Information Criterion (AIC) is useless when it is used to the isolated model but the best way to use it is to compare different models. The formula for calculating AIC is:

$$AIC = -2L(\beta) + 2p \tag{3.93}$$

where β is the number of parameters in the model, L is the maximum value of the likelihood function and p is the number of parameter in the model.For generalized logit model, $p = k(m+1)$, where k is the total number of response levels subtracting one and m is the number of explanatory effects. The AIC method is used to select the best model in different set of data. A model with the smallest AIC value will be the most preferable model to use for analysis.

3.9 Statistical Computation Packages in R

All the imputations of the current research were done using ***mix*** package available in **R** software. In particular, the package **mix** by (J. L. Schafer, 1997) was used to impute missing values in St.Louis Risk Research Project to assess the effect of parental psychological disorders on child development. The package **mix** is used because the data set for cancer-medication has mixed (continuous and categorical) variables. In this section, explain how the **mix** package works in **R** to impute missing values.

- The function $\rightarrow$ prelim.mix(**x** , p) performs preliminary data manipulations for **x** - a matrix of incomplete mixed data. In which the continuous variables is centered, scaled, and sorted by missingness patterns and categorical covariates will be grouped and sorted in that data set. In fact this produces summary list of different features in the given complete data matrix. The list produced above will be used by defined functions like em.mix, ecm.mix, da.mix, imp.mix, etc.

- The arguments are : **x** - is the data matrix containing missing values and p represent the number of categorical variables in matrix **x**. The rows of **x** correspond to units of the observation for each variable, and the columns to variables. Missing values are denoted by NA inside the given dataset. Under data augmentation algorithm **mix** package, categorical variables of matrix **x** must be coded with consecutive positive integers starting with 1 in the data. In the simple example, a binary variable must be coded as 1,2 not as 0,1.

More on data augmentation in R

1. By commanding function prelim.norm for multivariate normal models, and constrained loglinear models, prelim.cat for saturated multinomial models or prelim.mix for restricted and unrestricted general location models, to determine preliminary manipulations in the dataset, such as centering, scaling, and sorting by missingness patterns on a matrix of incomplete data **x**.

2. Command function em.norm for multivariate normal models, em.cat for multinomial models, ecm.cat for constrained loglinear models, em.mix for unrestricted general location models, and ecm.mix for restricted general location models, to obtain the maximum-likelihood estimates of the parameters with the incomplete data using the EM or ECM algorithm depending on the model of interest. These parameter estimators of cell probabilities (if categorical vari-

ables are present), means, and variance-covariances, will usually be used as starting values of parameters for the iterative simulation functions da.norm, da.cat, dabipf.cat, da.mix and dabipf.mix.

3. By commanding function da.norm for multivariate normal models, da.cat for saturated multinomial models, dabipf.cat for constrained loglinear models, da.mix for unrestricted general location models, or dabipf.mix for restricted general location models, to reproduce single (one) or multiple (more than one) iterations of a single Markov chain in a normal-inverted Wishart prior. The functions listed in these step, pick parameter estimates from the posterior distribution of interest. The parameter estimates generated in this step will be used by step (4) below to simulate missing values for imputations.

4. By commanding function imp.norm for multivariate normal models, imp.cat for saturated multinomial models and constrained loglinear models, or imp.mix for general location models (both restricted and unrestricted models), to impute the missing observations of the data matrix **x** and its use the parameter estimates obtain from previous step (3). The functions in these step will produce the complete data dataset with imputed missing values in it.

5. The multiple imputations can be performed using steps above (3) and (4) multiple times to simulate missing values.

3.10 Mixed data methods of data augmentation

3.10.1 Introduction

In general, most of statistical analyses in practice contain both continuous and categorical typeS of variables in the data set. For example, especially in analysis of variance, analysis of covariance, logistic regression with independent continuous variables etc (J. L. Schafer, 1997; Allison, 2001). This chapter develops mathematical tools for incomplete multivariate matrix data containing both continuous and categorical type of variables.

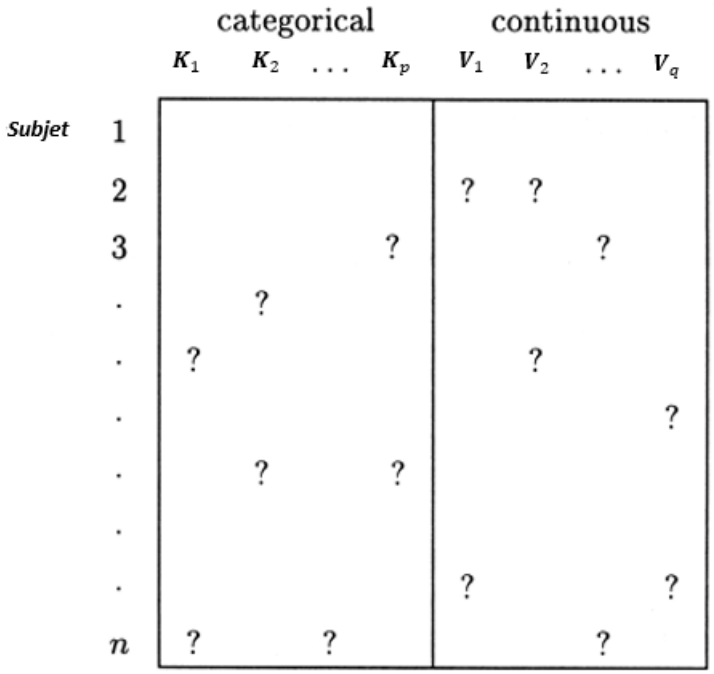

Figure 3.3 – The mixed dataset matrix with incomplete data

Missing values are denoted by question marks (?) on the above matrix dataset shown in **Figure** 3.3 above. The multiple imputation algorithm for mixed data (**MIX**) hold under MAR missing mechanism based on the general location model (J. L. Schafer, 1997). This multiple imputation is formed by EM algorithm and Data Augmentation process.

3.10.2 The general location model

Definition

For the matrix covariate vector $\mathbf{W} = (\mathbf{K}, \mathbf{V})$, which is recorded in n subjects is equal to $n \times (p+q)$ matrix. Let $K_1, K_2, , K_p$ denote a set of categorical independent variables and $V_1, V_2, .., V_q$ denote a set continuous covariates and k_i and v_i denote the values that respond to a vector $\mathbf{K}$ and $\mathbf{V}$ respectively, for subject i, where $i = 1, 2, 3, ..., n$. The categorical data components of the vector $\mathbf{K}$ may be given in terms of p dimensional contingency table with $C = \prod_{j=1}^{p} c_j$ cells, where the possible values are $I_j = 1, 2, ., c_j$ that K_j can take. The contingency table cells can be arranged according to a linear order index by $c = 1, 2, ..., C$. The cell units can be expressed as $\mathbf{x} = x_c : c = 1, 2, ..., C$, where vector $\mathbf{x}$ will be viewed as a multidimensional array. Let $\mathbf{D}$ be an $n \times C$ matrix dimension with rows d_i^T, where $i = 1, 2, ..., n$ and "T" denote the transpose of a vector or matrix. However, $d_i^T = E_c$ is a $1 \times C$-vector indicator containing a 1 if the units i belong into cell c contingency table, and 0 elsewhere. Hence each row of $\mathbf{D}$ is missing unless all categorical variables are observed, , which means it must contain a single 1 and $\mathbf{D}^T\mathbf{D}$ is a $C \times C$ matrix with $\mathbf{x} = [x_1, x_2, ..., x_c]$ in the diagonals.

$$\mathbf{D}^T\mathbf{D} = diag(\mathbf{x}) \begin{bmatrix} x_1 & 0 & \cdots & 0 \\ 0 & x_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_c \end{bmatrix}$$

This is true, since the sample units are assumed to be independent and identically distributed. All relevant statistical information in matrix $\mathbf{K}$ is contained in $\mathbf{x}$,$\mathbf{D}$ or $\mathbf{D}^T\mathbf{D}$.

The general location model is defined for both continuous and categorical variables mixed in the model by Olkin & Tate (1961). This is given by following equation

$$(\mathbf{x} \mid \boldsymbol{\pi}) \sim S(n, \boldsymbol{\pi}) \tag{3.94}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, ..., \pi_c)^T$.

Let E_c be a C-vector matrix containing a 1 in the position of C and zeroes elsewhere and conditional distribution of the certain row of $\mathbf{V}$, given $d_i = E_c$, which is assumed

to be $ST(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$. Denoted as

$$(v_i \mid d_i = E_c, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}) \sim ST(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}) \tag{3.95}$$

The marginal distribution of **K** is given by the equation (3.100) is a multinomial distribution which is represented by cell counts $\mathbf{x}_c$ given cell probabilities $\boldsymbol{\pi}_c = P(\mu_i = E_c)$. Where the sum of cell probabilities is equal to 1 for any $i = 1, 2, 3, ..., n$ and $c = 1, 2, 3, .., C$. The given equation above (3.101) denotes the conditional distribution of matrix **V** given **K** which is a multivariate normal distribution with a mean matrix of $C \times q$, which Is given by $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_c)^T$. The mean $\boldsymbol{\mu}_c$ is a vector that correspond to cell c and $\boldsymbol{\Sigma}$ is a $q \times q$ covariance matrix that corresponds to continuous variables in the model. The free number parameter is given by $(C-1) + Cp + \frac{p(p+1)}{2}$ in the unrestricted model. Thus, the model of **K** given **V** can be classified as the multivariate regression which is given by following expression

$$\mathbf{V} = \mathbf{D}\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3.96}$$

where a matrix $\boldsymbol{\epsilon}$ is a $n \times q$ dimension that represent errors and the matrix with rows that are independently distributed as $N(0, \boldsymbol{\Sigma})$. In general, the parameters of the general location model are

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.97}$$

The likelihood function

The likelihood function of the general location model can be written as the product of the complete-likelihood of the multinomial and normal likelihood

$$L(\boldsymbol{\theta} \mid \mathbf{V}) \propto L(\boldsymbol{\pi} \mid \mathbf{K}) L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{K}, \mathbf{V}) \tag{3.98}$$

where $L(\boldsymbol{\pi} \mid \mathbf{K}) \propto \prod_{c=1}^{C} \boldsymbol{\pi}_c^{\mathbf{x}_c}$

and

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{K}, \mathbf{V}) = |\boldsymbol{\Sigma}|^{\frac{-n}{2}} \exp\left\{ \frac{-1}{2} \sum_{c=1}^{C} \sum_{i \in F_c} (\mathbf{v}_i - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (\mathbf{v}_i - \boldsymbol{\mu}_c) \right\} \tag{3.99}$$

where F_c is the subject that fall into the cell c and

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \mathbf{K}, \mathbf{V}) = |\boldsymbol{\Sigma}|^{\frac{-n}{2}} \exp\left\{ \frac{-1}{2} tr \Sigma^{-1} \mathbf{V}^T \mathbf{V} + tr \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^T \mathbf{D}^T \mathbf{V} - \frac{1}{2} tr \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^T \mathbf{D}^T \mathbf{D}_{\boldsymbol{\mu}} \right\} \tag{3.100}$$

So that the likelihood can be denoted as the linear in the sufficient statistics $\mathbf{Z}_1 = \mathbf{V}^T\mathbf{V}$,$\mathbf{Z}_2 = \mathbf{D}^T\mathbf{V}$ and $\mathbf{Z}_3 = \mathbf{D}^T\mathbf{D}$.

The information about prior distribution

The application of a Bayesian method is convenient to simplify the problem of ML estimates in the model. Then with the assumptions of independent prior distributions for $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can be able to apply Bayesian method.

We use Dirichlet prior in the general location model, which is a conjugate prior for the multinomial distribution, can be applied to the cell probabilities,

$$P(\boldsymbol{\pi}) \propto G(\boldsymbol{\gamma}) \tag{3.101}$$

where $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, .., \gamma_C\}$ is a vector of hyperparameters that can be defined before estimation process. Noninformative priors can be used for both parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. If an uniform prior is applied to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ a standard noninformative prior to the covariance matrix, then

$$P(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{\frac{-(q+1)}{2}} \tag{3.102}$$

However, the posterior distribution represents the product of independent multivariate normal distributions for independent means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ..., \boldsymbol{\mu}_C$ given $\boldsymbol{\Sigma}$ and an inverted-Wishart distribution $(\mathbf{q}^{-1})$ for $\boldsymbol{\Sigma}$. Moreover, a multivariate normal distribution for $\boldsymbol{\mu}$ and an inverted-Wishart distribution for covariance matrix can be applied as informative priors as well. Further discussions about prior information are shown in (J. L. Schafer, 1997).

The multiple imputation for missing in mixed variables.

The predictive distribution of the missing data given the observed data is used when there are missing data in categorical variables $\mathbf{K}$ and continuous variables $\mathbf{V}$, since mix package consider both variable types. The categorical variables can be split into two parts, which is $\mathbf{k}_{i,ov}$ and $\mathbf{k}_{i,mv}$. These vectors represent the values of the observed and missing observations for subject i and also $\mathbf{v}_{i,ov}$ and $\mathbf{v}_{i,mv}$ denote the vectors of values of the observed and missing that correspond to continuous variables for units i.

Let $\mathbf{O}v_i(k)$ represent the subset of the observed data that correspond to categorical variables, and $\mathbf{M}v_i(k)$ represent the subset of the missing values that correspond to

categorical variables in the model. However, the predictive probability of falling cell k given the observed values is

$$P(\mathbf{u}_i = \mathbf{E}_k \mid \mathbf{k}_{i,ov}, \mathbf{v}_{i,ov}, \boldsymbol{\theta}) = \frac{\exp(\varepsilon^*_{k,i})}{\sum_{\mathbf{M}v_i(k)} \exp(\varepsilon^*_{k,i})} \tag{3.103}$$

Over the cells k that the observed part that correspond to categorical variables for subject i ($\mathbf{k}_{i,ov}$) belong to the element of $\mathbf{O}v_i(k)$. Where $\varepsilon^*_{k,i}$ represent the value of linear discriminant function of $\mathbf{v}_{i,ov}$ with respect to $\boldsymbol{\mu}_{k,i,ov}$,

$$\boldsymbol{\varepsilon}^*_{k,i} = \frac{-1}{2}\boldsymbol{\mu}^T_{k,i,ov}\boldsymbol{\mu}^{-1}_{k,i,ov} + \sum_{j \in \mathbf{O}v_i(v)} \mu_{k,i,ov} v_{i,j} + \log\pi_k \tag{3.104}$$

where $\boldsymbol{\mu}_{k,i,ov}$ and $\boldsymbol{\Sigma}_{k,i,ov}$ are the subvector of mean and sub-matrix of covariance in cell k of the continuous variables $\mathbf{v}_{i,ov}$ for subject i,$\boldsymbol{\mu}_{k,i,ov}$ is the k, jth element of $\boldsymbol{\mu}_{k,i,ov}$, and $\mathbf{O}v_i(v)$ is the subset of $(1, ..., k)$ corresponding to the variables in $\mathbf{v}_{i,ov}$.

The predictive distribution and sweep

Little & Schluchter (1985) show that the discriminant $\boldsymbol{\varepsilon}^*_{k,i}$ and the multivariate regression parameters of $\mathbf{v}_{i,mv}$ on $\mathbf{v}_{i,ov}$ can be determined by the method of a single application sweep operator. In the general location model the parameters can be arranged into a matrix form,

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\mu}^T \\ \boldsymbol{\mu} & \mathbf{P} \end{bmatrix}$$

, where $\mathbf{P}$ is a $C \times C$ dimension matrix with some elements inside which are denoted as $p_k = 2\log\pi_k$ on the diagonal and $0's$ everywhere. If we sweep this unknown parameter $\boldsymbol{\theta}$ matrix that correspond to $\mathbf{v}_{i,ov}$ in the location of $\boldsymbol{\Sigma}$, then the matrix can transformed as the matrix of parameters as follows,

$$\boldsymbol{\theta}_{ov} = \begin{bmatrix} \boldsymbol{\Sigma}_{ov} & \boldsymbol{\mu}^T_{ov} \\ \boldsymbol{\mu}_{ov} & \mathbf{P}_{ov} \end{bmatrix}$$

, where $p_{k,ov} = -\boldsymbol{\mu}^T_{k,i,ov}\boldsymbol{\Sigma}^{-1}_{k,i,ov} + 2\log\pi_k$ are the diagonal elements of $\mathbf{P}_{ov}$ that corresponding to cell k.

The EM algorithm: The EM algorithm is a method used to compute maximum likelihood estimates $(MLEs)$ in the general location model with both categorical and continuous variables. However, we general use $MLEs$ that are obtained from using

EM algorithm as starting values of parameters for data augmentation technique. As it is shown above, the complete-data loglikelihood is a linear function of the sufficient statistics, $\mathbf{Z}_1 = \mathbf{V}^T\mathbf{V}$,$\mathbf{Z}_2 = \mathbf{D}^T\mathbf{V}$ and $\mathbf{Z}_3 = \mathbf{D}^T\mathbf{D}$

The $MLEs$ of the unknown parameters $\boldsymbol{\theta} = (\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ under the unrestricted model are define as follows

$$\hat{\boldsymbol{\pi}} = n^{-1}\mathbf{x} \tag{3.105}$$

$$\hat{\boldsymbol{\mu}} = \mathbf{Z}_3^{-1}\mathbf{Z}_2 \tag{3.106}$$

$$\hat{\boldsymbol{\Sigma}} = n^{-1}(\mathbf{Z}_1 - \mathbf{Z}_2^T\mathbf{Z}_3^{-1}\mathbf{Z}_2) \tag{3.107}$$

The M-step can be determined by simply calculating the above equations (3.111) to (3.113) by using the mean versions of $\mathbf{Z}_1$,$\mathbf{Z}_2$ and $\mathbf{Z}_3$. This means that there is no need to for sufficient statistics themselves. Under the E-step of the algorithm, the only complicated part is, where we have to find the conditional expectation of $\mathbf{Z}_1$,$\mathbf{Z}_2$ and $\mathbf{Z}_3$ given $\mathbf{x}_{ov}$ and the parameter θ.

The E-step:

Step 1: We first consider the expectation of the diagonal elements of $\mathbf{Z}_3$. Notice that the elements of $\mathbf{d}_i$ are Bernoulli indicator of $\mathbf{d}_i = E_k$, for all cells in k. However, their expectations are predictive probabilities given by equation (3.109). Then, the expectation of $\mathbf{d}_i$ can be determined by using the following steps.

1. Firstly, we sweep the updated matrix θ^t that corresponds to positions of $v_{i,ov}$ in order to obtain the updated parameters of the observed data at t^{th} iteration θ^t.
2. We calculate the discriminant $\varepsilon^*_{k,i}$ for all cells k for which $\mathbf{k}_{i,ov} \in \mathbf{O}v_i(k)$ given $\mathbf{v}_{i,ov}$ and θ^t_{ov}.
3. The terms $\exp(\varepsilon^*_{k,i})$ is normalized for the cells to obtain the predictive probabilities given by:

$$\pi_{k,i,ov} = \frac{\exp(\varepsilon^*_{k,i})}{\sum_{\mathbf{M}v_i(k)} \exp(\varepsilon^*_{k,i})} \tag{3.108}$$

The equation (3.114) gives predictive probabilities, which are more useful in determining the expectation of $\mathbf{Z}_2$.

Step 2: Compute the expectation of $\mathbf{Z}_2$ based on the predictive probabilities and row k of sufficient statistics $\mathbf{Z}_2$ is given by $\sum_{i=1}^{n} d_{k,i} v_i^T$. The expectation of $\mathbf{Z}_2$ can be defined as follows

$$E(\mathbf{d}_{k,i} v_i \mid \mathbf{W}_{ov}, \theta) = \pi_{k,i,ov} v_{k,i,ov} \tag{3.109}$$

where

$$\mathbf{d}_{k,i} = \begin{cases} 1 & \text{if } unit \quad i \text{ fall into cell k} \\ 0 & \text{if } unit \quad i \text{ does not fall into cell k} \end{cases}$$

and $v_{k,i,ov}$ is the predicted mean of v_i given $v_{i,ov}$ given the $unit\ i$ falls in cell k. The $v_{k,i}$ is separated into two parts, which is observed and missing data part. However, the parts that correspond to observed part $(v_{i,ov})$ are the same $v_{i,ov}$, whereas the missing part $v_{i,mv}$ are the values that are predicted from the multivariate regression of $v_{ij,mv}$ conditional to $v_{i,ov}$ within cell k.

$$\mathbf{v}_{k,ij} = \begin{cases} v_{ij} & \text{if } j \text{ fall into subset of } \boldsymbol{O}_{v_i}(k) \text{ within cell k} \\ \mu_{k,j,ov} + \sum_{l \in O_{v_i}(k)} \sigma_{jl,ov} v_{il} & \text{if } j \text{ fall into subset of } \boldsymbol{M}_{v_i}(k) \text{ within cell k} \end{cases}$$

where $\sigma_{jl,ov}$ is the $(j,l)th$ element of $\mathbf{\Sigma}_{ov}$.

Step 3: Computation of the expectation of the sufficient statistics $\mathbf{Z}_1$ based on the predictive probabilities can found by finding the expectation of the sum of squares and cross product matrix of $\mathbf{Z}_1 = \mathbf{V}^T\mathbf{V} = \sum_{i=1}^{n} v_i v_i^T$. Consider the matrix of the sufficient statistics $\mathbf{Z}_1$, then the $(j,l)th$ elements in the matrix $\mathbf{Z}_1$ is given by $\sum_{i=1}^{n} v_{ij} v_{il}$. The single element $v_{ij}v_{il}$ can be written as

$$v_{ij} v_{il} = \sum_{k} d_{k,i} v_{ij} v_{il}, \tag{3.110}$$

then,the expectation of single element $v_{ij}v_{il}$ is expressed as follows

$$E(v_{ij} v_{il} \mid \mathbf{x}_{ov}, \theta) = \sum_{\mathbf{M}_{v_i}(k)} \pi_{k,i,ov} E(v_{ij} v_{il} \mid W_{ov}, \theta, d_{k,i} = 1), \tag{3.111}$$

where the sum is taken over cells k for which $\mathbf{O}_{v_i}(k)$ concur with $k_{i,ov}$. The expectation of sufficient statistics $\mathbf{Z}_1$

$$E(v_{ij}v_{il} \mid W_{ov}, \theta, d_{k,i} = 1) = \begin{cases} v_{ij}v_{il} & \text{if both } v_{ij} \text{ and } v_{il} \text{ observed} \\ v_{ij}v_{k,il,ov} & \text{if } v_{il} \text{ is missing and } v_{ij} \text{ is observed} \\ v_{k,ij,ov}v_{k,il,obs} + \sigma_{jl,ov} & \text{if } v_{ij} \text{ and } v_{ij} \text{ are both missing} \end{cases}$$

M-Step: The maximization step is achieved after obtaining the expectations of the each sufficient statistics $\mathbf{Z}_1, \mathbf{Z}_2$ and $\mathbf{Z}_2$ given the observed variables and updated parameters θ^t in the Expectation step. The maximization step is performed by using (3.111)-(3.113) to compute the updated estimate of θ^{t+1}.

The data augmentation algorithm: The data augmentation is a Markov Chain Monte Carlo technique for reproducing posterior distribution to draws a general location model parameters, given the matrix of both categorical and continuous variables with missing data. In the imputation step (I-step), we first use the predictive distribution of the missing values given observed data to generate missing values, then the values are drawn from the predictive distribution are used to impute missing values. This allows us to compute the complete sufficient statistics $(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3)$. Once we have complete data, then we can do the following step which is called posterior (P-step). In the P-step, an updated parameter $\boldsymbol{\theta}$ is drawn from posterior distribution given complete data from previous I-step. The first step, which is called I-step includes two steps. We first draw missing data $\mathbf{x}_{i,mv}^{t+1}$ for unit i from the predictive distribution of missing values $\mathbf{x}_{i,mv}$ given $\mathbf{x}_{i,ov}$ and the updated estimate of $\boldsymbol{\theta}$ is given by $\boldsymbol{\theta}^t$.

First step: we draw $\mathbf{d}_i^{t+1}$ from predictive distribution given vectors of the values of observed data $\mathbf{k}_{i,ov}$,which is the element of the observed parts of categorical variables in the model $\mathbf{K}_i(k)$. This distribution follows the multinomial distribution with cell probabilities given by equation (3.109).

Second step: we drawn $\mathbf{v}_{i,mv}^{t+1}$ given $\mathbf{d}_i^{t+1}$ and the vector of values of observed data in the observed parts of the continuous variables $\mathbf{v}_{i,ov}$ is actually based on the multivariate regression of $\mathbf{v}_{i,mv}$ regression on $\mathbf{v}_{i,ov}$. The conditional distribution of $\mathbf{v}_{i,mv}$ given observed data ,$\mathbf{d}_i$ and updated parameter θ^t is given as a multivariate normal distribution with means

$$v_{k,ij,ov} = \mu_{k,j,ov} + \sum_{l \in O_i(v)} \sigma_{jl,ov} v_{il} \tag{3.112}$$

According to Schafer (1997), the covariances of the model is determined by Cholesky factorization which means that $\mathbf{v}_{k,ij,ov}$ is simulated from the $\mathbf{v}_{i,mv}^{t+1}$.

The second step, which is called P-step includes three steps to draw the updated estimate $\boldsymbol{\theta}^{t+1} = (\boldsymbol{\pi}^{t+1}, \boldsymbol{\Sigma}^{t+1}, \boldsymbol{\mu}^{t+1})$. This estimates are drawn from their posterior distribution, given the complete version of sufficient statistics $\mathbf{Z}_1, \mathbf{Z}_2$ and $\mathbf{Z}_3$ from imputation step. Under noninformative prior joint distribution is given by

$$P(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \left(\prod_k \pi_k^{(\gamma_k - 1)} \right) \mid \boldsymbol{\Sigma} \mid^{(\frac{q+1}{2})} \tag{3.113}$$

Therefore, the posterior distribution of parameter are given following distributions

$$\boldsymbol{\pi} \mid \boldsymbol{W} \propto F(\boldsymbol{\gamma} + \mathbf{x}) \tag{3.114}$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{\pi}, \mathbf{W} \propto Q^{-1}(n - C, (\hat{\varepsilon}^T \hat{\varepsilon})^{-1}) \tag{3.115}$$

$$\boldsymbol{\mu}_k \mid \boldsymbol{\pi}, \boldsymbol{\Sigma}, \mathbf{W} \propto N(\hat{\boldsymbol{\mu}}_k, x_k^{-1} \boldsymbol{\Sigma}) \tag{3.116}$$

for $c = 1, 2, 3, ..., C$; where

$$\hat{\boldsymbol{\mu}} = (\mathbf{D}^T \mathbf{D})^{-1} \mathbf{D}^T V \tag{3.117}$$

$$\hat{\varepsilon} = V - \mathbf{D}\hat{\boldsymbol{\mu}} \tag{3.118}$$

If any cell count $\mathbf{x}_c$ is zero, the matrices $\mathbf{D}$ and $\mathbf{D}^T\mathbf{D}$ will have deficient rank, and (3.123) will not be defined. In this case, the posterior distribution will be improper due to the inestimability of $\boldsymbol{\mu}_k$. When this occurs, an analysis under this prior may proceed by omitting the inestimable parameters $\boldsymbol{\mu}_c$ from the model.

Step 1: We first start by drawing new π_k^{t+1} for each cell k from the standard gamma distribution with shape parameters $x_k + \gamma_k$, where $\boldsymbol{\gamma} = (\gamma_k)$ is an array of hyperparameters that can be specified; x_k is the diagonal element of $\mathbf{Z}_3$.

Step 2: We draw the updated $\boldsymbol{\Sigma}^{t+1}$ from an inverted-Wishart distribution with $(n - C)$ parameters in the model and $(\hat{\varepsilon}^T \hat{\varepsilon})^{-1}$; where $\hat{\varepsilon}^T \hat{\varepsilon}$ is equal to $\mathbf{Z}_1 - \mathbf{Z}_2^T \mathbf{Z}_3^{-1} \mathbf{Z}_2$.

Step 3: Then, we draw the updated parameter of mean $\boldsymbol{\mu}_k^{t+1}$ from normal distribution with mean $\hat{\mu}_k = \mathbf{Z}_3^{-1} \mathbf{Z}_2$ and variance $x_k^{-1} \boldsymbol{\Sigma}$. The algorithm iterate until it reach

convergence.

3.10.3 The Restricted general location model

The restricted general location model is the one of the models in Bayesian data augmentation, used when the number of categorical variables p grows in the model, since C and $C \times q$ become very large as p variables added in the model. The unrestricted general location model only works adequately if the number of subjects n in the model is large compared to the number of free parameters given by $(C - 1) + Cq + \frac{q(q+1)}{2}$ according (J. L. Schafer, 1997). Under the mixed data problem in application, the restricted model is more desirable since most of the variables are categorical in practice. Under the restricted models, the model is restricted using the loglinear models for cell probabilities and to linear models within-cell means .

The loglinear models for cell probabilities

The loglinear constraints is applied to the cell probabilities in order to put constraints within cell probabilities $\boldsymbol{\pi}$. The loglinear models for putting constraints within the cell probabilities $\boldsymbol{\pi}$ is given by

$$\log\boldsymbol{\pi} = \mathbf{H}\lambda \tag{3.119}$$

where $\mathbf{H}$ denote a user-specific matrix. The contingency table of the categorical variables is cross-classified by $k_1, k_2, ..., k_p$, and the user-specified matrix $\mathbf{H}$ that will show the structure containing main the effect for categorical variables and interactions within them. The first components of λ which is the intercept is not a free parameter especially when the first column of the user-specified matrix $\mathbf{H}$. If that is the case, the normalizing constant that scales cell probabilities $\boldsymbol{\pi}$ is summing to 1. The total number of the free parameters of the loglinear models has a rank $(\mathbf{H}) - 1$.

The linear models within-cell means

The basic idea here is that, we can discuss how to put constraints on the within-cell means $\boldsymbol{\mu}$ of the continuous variables such that matrices $\mathbf{D}$ and $\mathbf{D}^T\mathbf{D}$ will have adequate rank and posterior distribution will be proper. Under the unrestricted general location models, the conditional distribution of $\mathbf{V}$ given $\mathbf{K}$ is stated by the multivariate regression model given by

$$\mathbf{V} = \mathbf{D}\boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{3.120}$$

where $\mathbf{D}$ is $n \times C$ matrix of dummy indicators that shows the cell location $c = 1, 2, 3, ..., C$. Now we can impose the restriction within cell means $\boldsymbol{\mu}$,then $\boldsymbol{\mu}$ become

$$\boldsymbol{\mu} = \mathbf{B}\hat{\boldsymbol{\beta}} \tag{3.121}$$

for some free parameter $\boldsymbol{\beta}$. Let $\mathbf{B}$ be a $C \times r$ design matrix and $\mathbf{u} = \mathbf{B} \times \boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ denote $r \times q$ matrix under the assumption that the $rank(\mathbf{B}) = r \neq C$. Now this only means that we can only estimate $r \times q$ dimension of the matrix $\boldsymbol{\Gamma}$ instead of $C \times q$ dimension of the mean $\boldsymbol{\mu}$. The constrained general location model also allows the means $\boldsymbol{\mu}_c$ to move freely from cell to cell, but the only change is that each column in the continuous variables of the matrix $\boldsymbol{\mu}$ and is bounded in the linear subspace with r-dimensional around $\mathbf{R}^c$ spanned by columns of matrix $\mathbf{B}$. The new regression model becomes

$$\mathbf{V} = \mathbf{DB}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.122}$$

with the reduced number of regression coefficients in the free parameter $\boldsymbol{\beta}$. The special case of general location model can be obtained by saturating the loglinear model for Dirichlet distribution with hyperparameters and letting matrix $\mathbf{B} = \mathbf{I}_{C \times C}$ (identity matrix). The regression coefficients are estimable if the contingency tables for categorical variables contains no random zeroes but if it does contain zeroes, it may still be estimable just because estimability depends on the rank of $\mathbf{UB}$ instead of $\mathbf{U}$ itself. These only holds under the assumption that

$$rank(\mathbf{B}) = rank(\mathbf{DB}) = r \tag{3.123}$$

The likelihood inference of the restricted models

Under restricted models,we have two types of restrictions that we can apply on the models. The first restriction is the loglinear restriction on cell probabilities $\boldsymbol{\pi}$ and the linear restrictions within-cell means $\boldsymbol{\mu}$. Let the joint unknown parameter space for $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and the individual space for the product of $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ can be obtained on the above restrictions. The problem with the joint likelihood for parameter $\boldsymbol{\theta}$ is that maximization and the estimate for maximum likelihood for cell probabilities can be determined by using conventional IPF.

In the restricted model, we usually apply the marginal distribution to cell probabilities which allows us to separate factors in the full likelihood in the given model. The estimate for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be obtained from the least-squares regression for the reduced model $\mathbf{V} = \mathbf{DB}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, which generates the estimates for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$. Thus,

$$\hat{\boldsymbol{\beta}} = (\mathbf{B}^T\mathbf{D}^T\mathbf{D}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{D}^T\mathbf{V} \tag{3.124}$$

$$= (\mathbf{B}^T\mathbf{Z}_3\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{Z}_2); \tag{3.125}$$

and

$$n\hat{\boldsymbol{\Sigma}} = (\mathbf{V} - \mathbf{D}\mathbf{B}\hat{\boldsymbol{\beta}})^T(\mathbf{V} - \mathbf{D}\mathbf{B}\hat{\boldsymbol{\beta}}) \tag{3.126}$$

$$\mathbf{Z}_1 - \mathbf{Z}_2^T\mathbf{B}(\mathbf{B}^T\mathbf{Z}_3\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{Z}_2) \tag{3.127}$$

The ML estimate of the cell means $\boldsymbol{\mu}$ is given by $\hat{\boldsymbol{\mu}} = \mathbf{B}\hat{\boldsymbol{\beta}}$ and the covariance matrix will uses the unbiased estimate $n(n-r)^{-1}\hat{\boldsymbol{\Sigma}}$ instead of $\hat{\boldsymbol{\Sigma}}$.

Under the Bayesian inferences the restricted model may be implemented to the independent product of prior distributions for unknown parameter set $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. These independent parameter set $\boldsymbol{\pi}$ and $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ remain independent even under the complete-data posterior distribution. General location model can be obtained by saturating the loglinear model for Dirichlet distribution with hyperparameters. However, we can define the proir distribution to be a constrained Dirichlet prior to be components of parameter $\boldsymbol{\pi}$ with the density function

$$P(\boldsymbol{\pi}) \propto \prod_{c=1}^{C} \boldsymbol{\pi}_c^{\gamma-1} \tag{3.128}$$

where values of π hold under the loglinear constraints.The posterior density for the full data can be constrained using Dirichlet with the new hyperparameters $\prime\gamma_c = \gamma_c + x_c$.

The inferences for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ under a noninformative prior

The multivariate regression model under the Bayesian inference is given by $f(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{W})$. The likelihood function for $\boldsymbol{\Sigma}$ and $\boldsymbol{\beta}$ is given by

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{W}) \propto |\boldsymbol{\Sigma}|^{\frac{-n}{2}} \exp\left\{ \frac{-1}{2} \mathrm{tr}\boldsymbol{\Sigma}^{-1}(\mathbf{V} - \mathbf{D}\mathbf{B}\boldsymbol{\beta})^T(\mathbf{V} - \mathbf{B}\mathbf{B}\boldsymbol{\beta}) \right\} \tag{3.129}$$

then, the likelihood function also be rewritten in terms of least-squares estimate

$$\propto |\boldsymbol{\Sigma}|^{\frac{-n}{2}} \exp\left\{ \frac{-1}{2} \mathrm{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}} - \frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\Sigma} \otimes \mathbf{Y})^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right\} \tag{3.130}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficients matrix, $\mathbf{Y} = (\mathbf{B}^T\mathbf{D}^T\mathbf{D}\mathbf{B})^{-1}$ and $\hat{\boldsymbol{\varepsilon}} = \mathbf{V} - \mathbf{D}\mathbf{B}\hat{\boldsymbol{\beta}}$ is the estimated residuals. The following symbol $\otimes$ defines the Kronecker product;

$$\boldsymbol{\Sigma} \otimes \mathbf{Y} = \begin{pmatrix} \sigma_{11}Y & \sigma_{12}Y & \cdots & \sigma_{1q}Y \\ \sigma_{21}Y & \sigma_{22}Y & \cdots & \sigma_{2q}Y \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q1}Y & \sigma_{q2}Y & \cdots & \sigma_{q,q}Y \end{pmatrix}$$

The quantity $(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\boldsymbol{\Sigma} \otimes \mathbf{Y})^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ will be meaningful if the columns of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are stacked to form a vectors with the length of rq. We continue to apply an improper uniform prior to $\boldsymbol{\beta}$ and Jeffreys prior for parameters given $\boldsymbol{\Sigma}$ that is ,

$$P(\boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\frac{(q+1)}{2}} \tag{3.131}$$

The combination between the joint prior density (3.365) with the likelihood function (3.263) and also use the Kronecker products between $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$,

$$|\boldsymbol{\Sigma} \otimes \mathbf{Y}| = |\boldsymbol{\Sigma}|^r|\mathbf{Y}|^q, \tag{3.132}$$

Therefore, we can get the updated posterior density function which is expressed as follows

$$P(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{W}) \propto |\boldsymbol{\Sigma}|^{\frac{-(n-r+q+1)}{2}} \exp\left\{\frac{-1}{2}\mathrm{tr}\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}}\right\} \times |\boldsymbol{\Sigma}\otimes\mathbf{Y}|^{\frac{-1}{2}} \exp\left\{\frac{1}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{\Sigma}\otimes\mathbf{Y})^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\} \tag{3.133}$$

and the following posterior distribution of the product of multivariate normal density for $\boldsymbol{\beta}$ given $\boldsymbol{\Sigma}$ and an inverted-Wishart density for $\boldsymbol{\Sigma}$,

$$P(\boldsymbol{\Sigma} \mid \mathbf{K}, \mathbf{V}) = K^{-1}(n-4, (\hat{\boldsymbol{\varepsilon}}^T\hat{\boldsymbol{\varepsilon}})^{-1}) \tag{3.134}$$

$$P(\boldsymbol{\beta} \mid \boldsymbol{\Sigma}, \mathbf{K}, \mathbf{V}) = N(\hat{\boldsymbol{\beta}}, \boldsymbol{\Sigma} \otimes \mathbf{Y}), \tag{3.135}$$

and $\boldsymbol{\Sigma}$ and $\mathbf{Y}$ is the Kronecker product of $\boldsymbol{\Sigma}$ and $\mathbf{Y}$ (M. Anderson, 2010).

According to (J. L. Schafer, 1997), data augmentation can be used in different multiple imputation softwares such as a multivariate normal models, multinomial models and general location models to impute missing observations for different types of variables. J. L. Schafer (1997) shows that these algorithms can be implemented under continuous variables, categorical variables and mixed (continuous and categorical) variables respectively and also prove that this model only holds under the

assumptions of ignorability; that is, missing observations occur at random (MAR). The descriptions of this model can be found in J. L. Schafer (1997). In this study, we applied the unrestricted general location model since our variables are both categorical and continuous variables.

To apply unrestricted general location model to incomplete mixed data above, the software started by using the EM algorithm to compute the maximum likelihood estimates for each cell probabilities, the cell means, and the covariances. EM algorithm may be used as starting points simulation step iteration in the algorithm. Thus, the software starts to apply the iterative simulation technique in the loop to reproduce one or more iterations of a single Markov chain. The iterations simulated consists two steps which is I-step and P-step. In the I-step the random imputation for both missing in categorical and continuous data are drawn from the predicted multinomial distribution and multivariate normal distribution, respectively, with the current estimate of the parameter in the model. The restricted general location model is very useful when n is large compared to number cells in the given data set. The unrestricted general location model has $(C-1)+Cp+\frac{p(p+1)}{2}$ free parameters and it became difficult to compute $C \times p$ estimate as a number of categorical variables p increase in the model. The Bayesian Iterative Proportional Fitting (BIPF) algorithm is used to reduce the number of the parameters needed to estimated in the general location model (J. L. Schafer, 1997) . In application, the package ***mix*** is used in **R** software that was developed by J. L. Schafer (1997), which can be downloaded from cran. In the ***mix*** library, the general model location data augmentation uses the function ***da.mix*** and ***BIPF*** algorithm for the restricted general location model use ***dabipf.mix*** function.

3.11 Imputation Analysis

3.11.1 Missing values imputation for mixed data

All the data augmentation that exist in statistics holds under the assumption that data are MAR, according to (Allison, 2001; McKnight et al., 2007), other missing data mechanisms such as MCAR can also be assumed if the objective is to compare the performance of data augmentation algorithm under different rates proportion of missingness. This study aimed to determine the performance of data augmentation algorithm under different rates proportion of missingness in terms of bias in the estimated regression coefficients and standard errors of the regression coefficients,

when data is missing at random, missing completely at random or missing not at random on the covariates in the regression models. The MCMC data augmentation algorithm simulates pseudo-random values and replaces each missing value by adding missing data on the observed data, so that will be easier to analyze data with missing values. The data augmentation can be implemented in solving or simulating missing values so that we can be able to compute full posterior distributions of θ (Tanner & Wong, 1987). There are many statistical software packages on which data augmentation analysis can have done such as R, SPSS, STATA, and SAS. The unrestricted general location model to incomplete mixed data can be applied to cancer medication data since it consists of categorical and continuous variables. Then, the software package mix in R applies the iterative simulation method to reproduce one or more iterations of a single Markov chain algorithm. The iteration is made up two steps, which is I-step and P-step(J. L. Schafer, 1997). In I-step the random imputation for both missing categorical and continuous data are drawn from the predicted multinomial distribution and multivariate normal distribution, respectively, using the current parameter estimate in the model.

The data augmentation model for mixed data including the analysis model covariates and dependent variables of interest, the variables were associated with missingness on the explanatory variables to be imputed and the dependent variables are not allowed to have any missing values (dependent variable must be complete). The binary regression models with no missing data were fitted and the results were compared to the models estimated using data sets with imputed missing data and the completed (observed + imputed) data sets using data augmentation method.

To impute this, explanatory variables with missing data problems using data augmentation algorithm, the Tanner & Wong (1987) method for mixed data set was utilized. That is, the dependent variable must be completed and they must be coded with consecutive positive integers starting with 1. For example, a binary variable must be coded as 1,2 rather than 0,1 before being imputed. Thus, the levels variables of Working for a wage (Yes, No) were included in the imputation model, treating No as a reference category. Assuming that WW denotes the levels of Working for a wage to be imputed, the following logistic regression model was estimated for each working for wage dichotomised category:

$$WW = \beta_0 + \beta_1 D + \beta_2 E + \beta_3 G + \epsilon \tag{3.136}$$

where β_i denote the survey logistic regression estimates and D, E, G and ϵ denote

gender, living area, age and error term respectively. The WW can be imputed include the effect of gender, living Area, age. For variable Age was included in the imputation model, this explanatory variable is a continuous variable. Let that AG denote the Age to be imputed, the following logistic regression model was estimated to determine the parameter estimates:

$$AG = \beta_0 + \beta_1 H + \beta_2 E + \epsilon \tag{3.137}$$

where β_i denote the survey logistic regression estimates and H, E and ϵ denote gender,living area and error term respectively. The MAR assumption, the missing values were created such that missingness depended on whether a cancer patient was using a cancer medication or not. Although this variable is excluded from the analysis model but is related to missingness such as Gender, Living Area, and Age are included in data augmentation model as missing auxiliary variables. When data under the data augmentation method were defined as MCAR, then all these variables are related to the dependent variable (cancer medication intake) that was used in the model.

In the above we mentioned that, this study was utilized by using a South African (GHS2015) data set, which is a survey that is complicated with a complicated sampling design and weighting procedure of the data involved, that must be taken into consideration during the analysis, since it has some effect on the results. Generally, in survey sampling, all the units do not have the same probabilities to be included in the sample. With complicated survey data sets, these probabilities are computed and used to calculate the sample weights. As an example, consider a population for which the estimator of the total V is as follows:

$$\hat{v_w} = \sum_{i=1}^{n} w_i v_i \tag{3.138}$$

where v_i represent the observed values of V and w_i is the weight that depends on the probability that the unit i will be included in the sample. The weight, in this case, is the inverse selection probability and it represents the number of individuals in the target population represented by sample unit i (Levy and Lemeshow, 2013). As noted by Levy and Lemeshow (2013), ignoring the sample weights during the analysis results in more biased estimators than weighted estimators.

The sample has a small portion of the respondent population in a survey of inter-

est. The weighted sample data, it improved estimators of the model to be more closer to the target population estimators. There are several studies that reveal that when survey datasets contain weight variables, weighted sampled outputs are chosen mostly since they produce fewer bias estimates than unweighted outputs (Korn & Graubard, 1995).). The outputs of the survey logistic regression model in this study were based on the weighted data sets to give parameters of the model that has unequal probabilities of selection for each sample unit in the cancer medication data. The data augmentation algorithm was performed using the **R**, with the package called **mix** for mixed data set was used to perform imputation of missing data. The other software such as SAS 9.4 and R were used to do data preparation like creating the proportion rates of missingness and imputation analysis. Data analysis was done by SAS 9.4 for model fitting and graphics were determined in Power BI, R and SAS 9.4.

3.11.2 Measurement of model performance

The survey logistic regression models with the baseline data set were first estimated to get the values of the survey logistic regression coefficients (true coefficients of the complete data) and their corresponding standard errors to each variable in the model. The outcomes from the baseline data regression model were considered as true results that are considered as a benchmark of outcomes from the imputed data sets and data sets with missing values. Then survey logistic regression models with the data sets with missing values and imputed datasets using Data Augmentation algorithm were estimated and the results (in terms of bias and standard errors estimates) were recorded. To judge the performance of the data augmentation algorithm, these estimates were considered in the results section. They were compared for each data set to assess the performance of the Complete data and imputed missing values on the different rates of missingness when data was arbitrary MAR, MCAR or MNAR on working for wage and age variables, and treated as explanatory variables from the cancer medication data set. However, according to these Tanner & Wong (1987), larger numbers of imputations are required if the objective is to compare different proportion rates of missingness imputation models that correspond to the specific proportion or to obtain stable and less unbiased estimates. To obtain sufficient accuracy while comparing different proportion rates of missingness, this study used 50 imputations for each data set with missing values, which resulted in 50 different imputed simulated versions of complete data sets. Each imputed data set was analyzed separately using data augmentation algorithm, and the estimates of standard errors were produced using binary survey logistic regression

model under each data set.

Relevant performance measures in the missing data evaluation

When comparing performance of estimates of missing data mechanism (MCAR, MAR and MNAR) under $1\%, 5\%,$ and 10% of missing percentage of observations. Within each condition, the performance of the model was evaluated by computing raw bias, Cox and Snell R^2, and root-mean-square error (RMSE). The proportion of significant effects was examined for estimates that were zero in the population; this is the observed Type I error rate. Raw bias is simply the true value (θ) subtracted from the corresponding estimate $(\hat{\theta})$

$$Parameter \quad Bias = \hat{\theta} - \theta \tag{3.139}$$

The bias of standard deviation is given by following calculation

$$Standard \quad DeviationBias = \sigma_\theta - s_{\theta_i} \tag{3.140}$$

Cox and Snell(1989,pp. 208209) propose the following generalization of the coefficient of determination to a more general linear model:

$$R^2 = 1 - \left(\frac{L(\theta)}{L(\hat{\theta})}\right)^{\frac{2}{n}} \tag{3.141}$$

where $L(\theta)$ is the likelihood of the intercept-only model, $L(\hat{\theta})$ is the likelihood of the specified model, and n is the sample size. and is simply raw bias scaled as a percentage of the population parameter. In addition, the mean square error of the parameter estimate bias for each missing percentage. RMSE is computed as

$$RMSE = \sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{N}} \tag{3.142}$$

and because it is in the same metric as the data, it can be interpreted as representative of the size of a typical error. RMSE is not strictly a measure of bias; rather it takes into account the variance of the errors and the mean error, so RMSE will not necessarily be zero when a parameter estimate is unbiased.

Chapter 4

Results

4.1 Introduction

This chapter presents the results based on the cancer-medication data sets for the complete data set and imputed datasets under different proportion rates of missing data, when data are missing at random, missing completely at random or missing not at random. Three sections are covered under this chapter. The first section of this chapter (Section 4.1) provides an overview of the results of this chapter. The second section (Section 4.2) presents the results of the analysis of complete data (without missing data) such as descriptive statistics, fitting the logistic regression model, data visuals and some tests (Normality test, Chi-square test). The third section (section 4.3) represented the different models when the proportion rate of missing data is 1%, 5% and 10% , when data are missing at random, completely at random or missing not at random scenarios on two covariates (Working for a wage, Age). The estimates of bias and standard errors in the regression coefficients obtained using the complete data or Bayesian data augmentation under different proportion rate of missing data, when the data is missing at random, completely at random or not missing at random are presented. The results model diagnostics obtained in the survey logistic regression model and comparison of the results of complete data and imputed models are also provided in this section. The relevant performance measure calculations were performed such as Raw bias, coefficient of determination(R^2) and Root Mean Square Error (RMSE) to evaluate missing data problem.

4.2 Analysis of complete-data

4.2.1 Descriptive statistics for complete-data

Of the target population, 194 cancer patients were randomly selected in the study. The graph below shows that the majority (90.21%) cancer-patients uses cancer medication while only (9.79%) dont use cancer medication. The descriptive bar graph is given below (**Figure** 4.1).

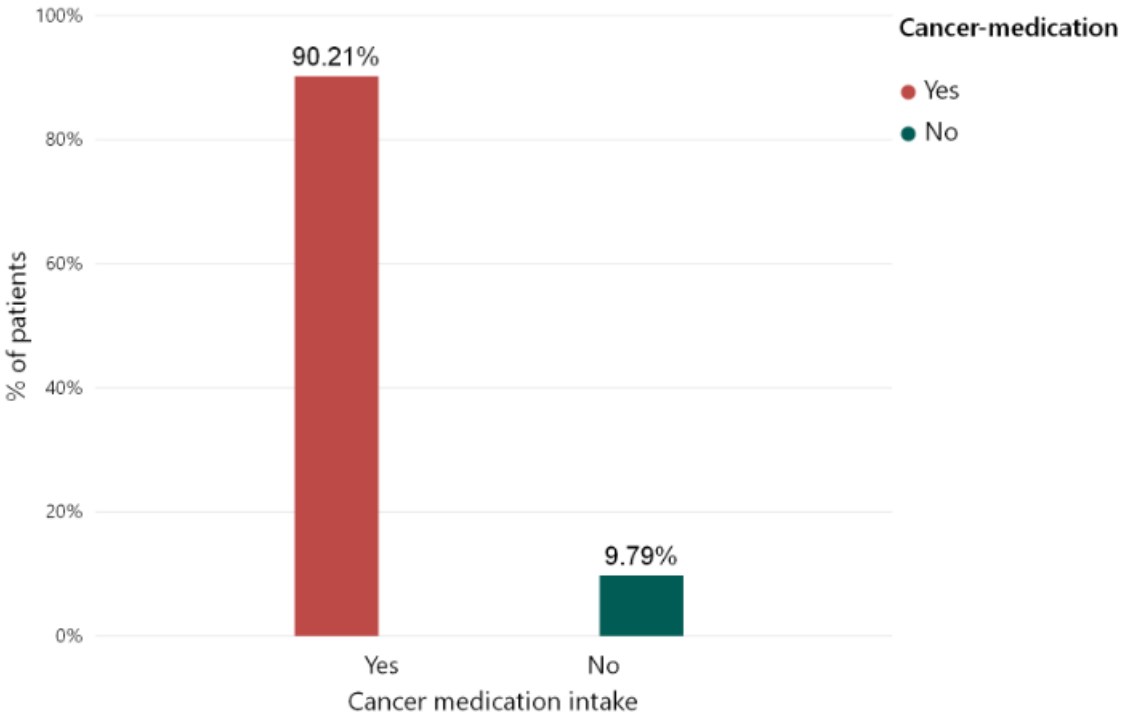

Figure 4.1 – The Bar chart of the intake of cancer-medication in South Africa, 2015

Summary of the descriptive statistics from cancer-medication data

The following table indicates descriptive statistics and summary statistics of variables related with the dependent variable.

Table 4.1 Descriptive statistics and marginal distribution for complete-data

Variables	Count	Cell%
Dependent Variable:		
Cancer Medication		
1. Yes	175	90.21%
2. No	19	9.79%
Independent Variables:		
Wage		
1. Yes	42	21.65%
2. No	152	78.35%
Gender		
1. Male	79	40.72%
2. Female	115	59.28%
Area of living		
1. Urban	150	77.32%
2. Rural	44	22.68%

The **Table** 4.1 above shows the descriptive statistics of the original data of the cancer medication. The respondents response on cancer medication is whether a person uses cancer medication or not. The table which shows the respondents categorized according to gender, indicates 59.28% are female and 40.72% are males.Thus the majority of respondents are female.The working for wage is also an important factor that affects the use of cancer medication. Based on the above table, 21.65% are working for wage. While 78.35% are not working for wage. The above table also shows that most of the people who responded on the survey, 77.32%, came from urban areas and 22.68% came from rural areas.

Table 4.2 Summary statistics for original data Age

Variables	Mean	Standard deviation	Median	IQR
Age	57.49	13.605	58	19.75

The variable Age on original data set has mean of 57.49, median is equal to 58, interquartile range is 19.75 and standard deviation is equal to 13.605 under the observed data of sample size $n = 195$. This estimate shows values has a sufficiently strong central tendency, that is, a tendency to cluster around some particular values or not.

Table 4.3 Frequency distribution of cancer-medication in household survey of South Africa 2015

Variables	Category	Yes		No		Total
		Count	Row N%	Count	Row N%	Count
Gender	Male	68	86.08%	11	13.92%	79
	Female	107	93.04%	8	6.96%	115
Area of living	Urban	135	90.00%	15	10.00%	150
	Rural	40	90.91%	4	9.09%	44
Working for wage	Yes	38	90.48%	4	9.52%	42
	No	137	90.13%	15	9.87%	152

The Table 4.3 above shows descriptive statistics and summary statistics of Gender, Area of living and Working for wage variables related to the cancer medication intake (dependent) variable. Respondents response are categorized according to gender, (93.04%) of the female are using cancer medication, whereas, (6.96%) of females are not on cancer medication. The results in contingency table also show that (86.08%) of males are using cancer medication and (13.92%) are not using cancer medication. Area of living is also considered as an important factor that affects intake of cancer medication in S.A. Based on the above table, (90.00%) of cancer patients who were interviewed come from the urban areas are using cancer medication, whereas (10.00%) are not using cancer medication. While (90.91%) of those who are from rural areas are using cancer medication and (9.09%) are not using cancer medication. For working status, (90.48%) of cancer patients who are working for a wage is using cancer medication and (9.52%) are not using cancer medication. While (90.13%) of people who are not working for the wage is using cancer medication to

heal cancer chronic, whereas (9.87%) of those are not using cancer medication.

4.2.2 Bivariate analysis of cancer-medication intake

The use of bivariate analysis is the main feature of the descriptive analysis in this study. Bivariate analysis was used to describe the relationship between cancer medication intake and gender of each individual in the dataset and other variables. The bivariate analyses incorporated with the hypotheses testing performed helps to provide a preliminary explanation of cancer medication intake in South Africa.

Table 4.4 Results of test association between predictors and the Intake cancer-mediation

Variables	Value	DF	p-value
Gender	1353.064	1	<.0001
Working for wage	145.809	1	<.0001
Area of living	125.738	1	<.0001

From the **Table** 4.4, we observe that the predictors are associated with intake of cancer medication.

Gender

The study shows that the gender is an important risk factor in answering the questions about cancer and cancer medication intake. Table 4.4 summarizes results of the differential analysis of cancer medication according to gender. The Chi-square test $\chi^2 = 1353.064$(p-value $<.0001 < 0.05$) reveals a statistically significant association between gender and cancer medication intake.

Working for wage

The chi-square test $\chi^2 = 145.809$ (p-value $<.0001 < 0.05$) brought evidence of a statistical significant association between cancer-medication intake and Working for wage.

Area of living

The chi-square test $\chi^2 = 125.738$ (p-value $<.0001 < 0.05$) brought evidence of a statistical significant association between cancer-medication intake and Area of living.

In conclusion : It is observed from the table above gender, working for wage and area of living are individually significant associated with the intake of cancer medication at 5% level of significance comparing with p-values in Table 4.4 above and each of the other factors have not be controlled for. It is important to note this as the significance between working for wage and intake of cancer medication changes in the multivariable analysis after controlling for all the factors.

4.2.3 Multivariable statistics for complete-data

The Survey logistic Regression Analysis

The survey logistic regression was fitted to cancer medication data. The analysis was done using the proc survey logistic procedure in SAS®software version 9.4 (SAS Institute Inc, 2009). The result of the binary survey logistic regression model represented below in **Table 4.5** for original data estimation of cancer medication intake in South Africa. The patients that taking cancer medication assigned a value of 1 if the patients do not use cancer medication is assigned a value of 2. The Survey logistic regression model was used to modelling the probability of a patient taking cancer medication. The Since the data is binary we use reference category to compare the values with the other categories of the given variables. Our interest is to estimate the regression estimates for fitted model for each dataset.

Table 4.5 Survey logistic regression that predicts original data for cancer-medication chronic

Parameter		**DF**	**Estimate**	**Standard error**	**Pr** $>\|t\|$	**exp(Estimate)**
Intercept			1.2889	1.2970	0.3216	3.629
Gender (ref: Female)						
	Male	1	-0.7891	0.5110	0.1242	0.454
Wage (ref: No)						
	Yes	1	0.0178	0.6958	0.9796	1.018
Area of living (ref: Rural)						
	Urban	1	-0.1559	0.6759	0.8178	0.856
Age		1	0.0257	0.0210	0.2210	1.026

Test	**F-Value**	**df**	**p-value**
Overall model evaluation			
Likelihood	0.99	4	0.4138
Score test	1.26	4	0.2864
Wald test	1.21	4	0.3083

Reporting and Interpreting of Survey logistic Regression Results

The parameter estimates are obtained by using maximum likelihood estimation method with 95% Walds confidence limits as is shown in Table 4.5 above. There is a negative estimate for the variable Area of living and Gender which means that this variable negatively influences the probability of Cancer medication intake in South Africa according to the reference category. Since the response variable of interest is dichotomous (taking cancer medication or not), a survey logistic procedure was used to identify the factors that has an effects on the patient's cancer medication intake and modelling the probability of taking cancer medication. Survey logistic regression procedure describes the relationship between the binary response variable and a set of explanatory variables. The interpretation of results is given in the form of odds ratios. For continuous explanatory variables, the odds ratio is for a 1-unit increase of the corresponding variable, for categorical variables the odds ratio is between the corresponding category and the reference category. The highest odds ratios were obtained for Wage and Age. The most of the odds ratios were approximately to or greater than 1 except the variable Gender.

Focusing on Gender, where Females has been used as a baseline for making a comparison, Gender (being a Female) seems to have an effect on the cancer medication intake. The effect of gender is insignificant. The odds of Male taking cancer medication were 0.454 times less compared to Female (Coefficient= -0.8958, OR=0.454 and p-value 0.1242). The odds for people who working for wage increases by 1.018 times more as compared to those who are not working for wage (Coefficient= 0.0178, OR=1.018 and p-value=0.9796). This shows that the effect of working for wage is insignificant since the p-value$>$ 0.05. The association between Wage and Cancer-medication is insignificant since (p-value=0.9796 $>$0.05) and (95% CI: -1.3546,1.3902). The odds of cancer-medication intake people to came from Urban is 0.856 times less likely than person come from urban compare to rural area. The effect of the Area of living on Cancer-medication is insignificant since (p-value=0.8178$>$ 0.05). Based on the model, the factor change in odds of Age=1.026. An increase of year as the Age of an patient was recorded in years, the odds of using cancer-medication change the factor of 1.026, holding all other covariates constant (Coefficient=0.0257, OR=1.026, p-value= 0.2210). The effect of the Age on Cancer-medication is insignificant since (p-value=0.2210$>$ 0.05)

Tests of individual predictors of the model

The **Likelihood Ratio test** was used to evaluates the significance of the joint effect of all the variables in the Survey logistic regression procedure. Since (p-value$>$ 0.05) of the Likelihood ratios of the original model, and insignificance of the likelihood ratio test means that the joint of the variables in the full model is insignificant than just the intercept model.

The **Wald test** is used to evaluates the significance of the joint effect of all the variables or individual variable in predicting the probability of taking and not taking cancer medication in the Survey logistic regression procedure. Since (p-value$>$ 0.05) of the Wald statistics of the original model, and insignificance of the Wald test means that the joint of the variables in the full model is insignificant than just the intercept model.

4.2.4 Assessment of normality in the variable (Age)

Q-Q plot and Histogram

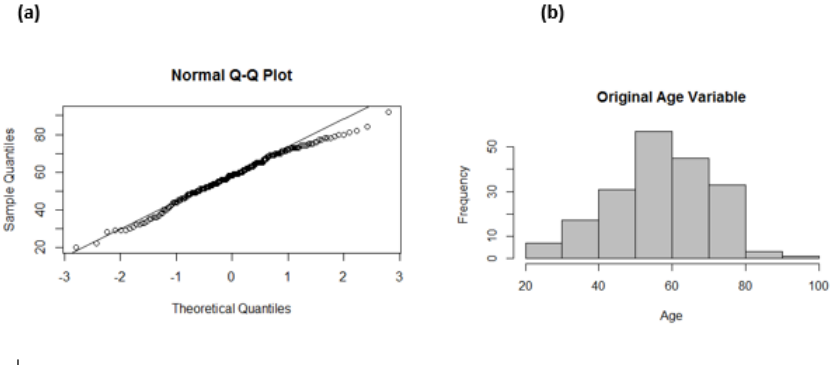

Figure 4.2 – The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age) in the data is quite symmetric as it shown in the histogram above and Q-Q plot indicate some few outliers on the tails.

Shapiro-Wilk test

The Shapiro-Wilk test for normality of the distribution especially for continuous variables. The null hypothesis for Shapiro-Wilk test is that the data is normally distributed and against the alternative that the data is not normally distributed. Under the assumption that the level of significance $\alpha = 0.05$ and then if the *p*-value $< \alpha = 0.05$, then we reject the null hypothesis. If the *p*-value $> \alpha = 0.05$, then we do not reject the null hypothesis.

Table 4.6 Shapiro-Wilk normality test for Age

w	***p*-value**
0.98647	0.6031

The **Table** 4.6 above shows that the value of $W = 0.98647$ is closer to 1. Since the level of significance is assumed to be $\alpha = 0.05$ but the p-value is greater than $\alpha = 0.05(0.6031 > 0.05)$. We do not reject the null hypothesis that the data is normally distributed.There is insufficient evidence to reject H_0, then we can conclude that the data came from the normally distributed population.

4.3 Analysis of missing data imputation

4.3.1 Assessment of normality of the variable (Age)

Normal probability plot (Q-Q plot)

The Normal probability plot is the graphical method for assessing whether a data set is approximately normally distributed. The data is plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. If the points depart from this straight line it indicates departures from normality.

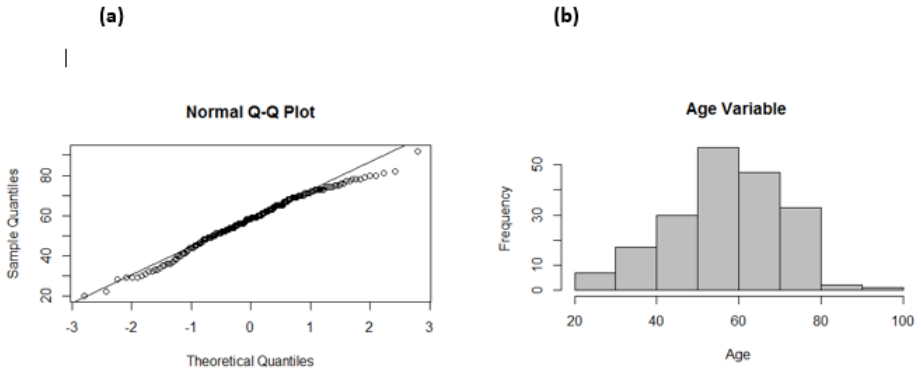

Figure 4.3 – 1% MCAR Imputation: The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age)under 1% MCAR imputed in the data is quite symmetric as it shown in the histogram above and Q-Q plot indicate some few outliers on the tails.

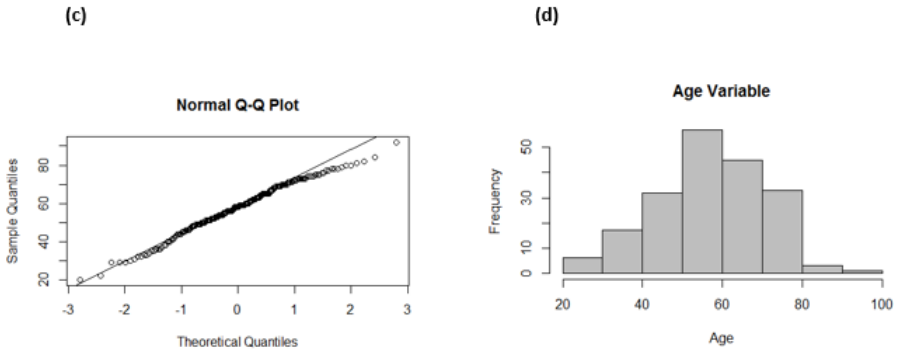

Figure 4.4 – 1% MNAR Imputation: The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age) under 1% MNAR imputed in the data is quite symmetric as it shown by histogram above and Q-Q suggest that the data point fit to the straight line except in the tails.

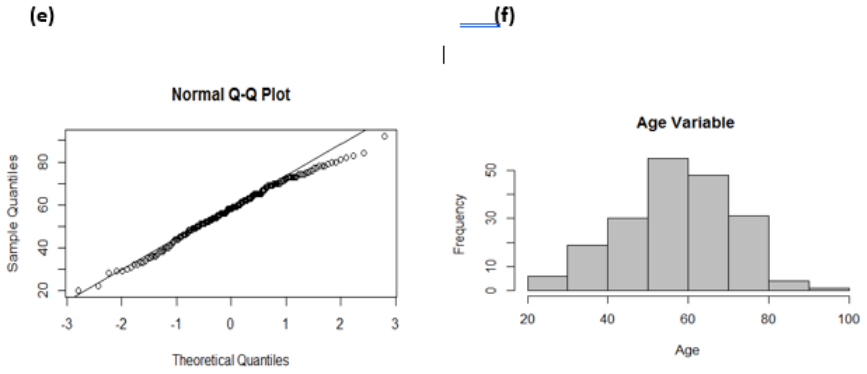

Figure 4.5 – 5% MCAR Imputation: The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age)under 5% MCAR imputed in the data is quite symmetric if we look at the the histogram above and the points in the Q-Q plot fall on the straight line through first quantile, second quantile which indicate that data might be normal except on the upper tail of the distribution.

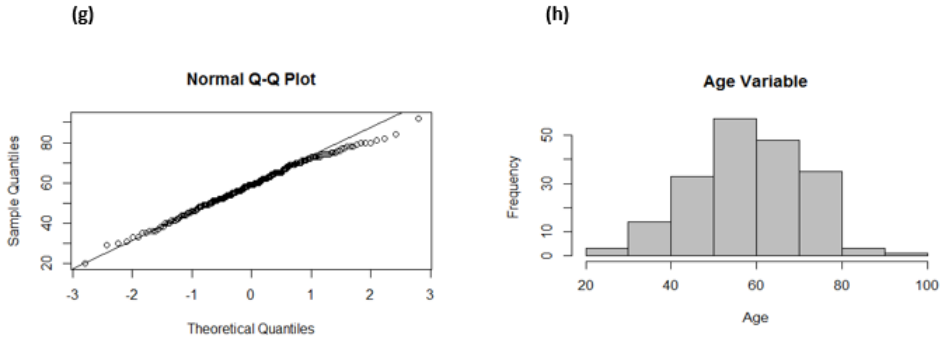

Figure 4.6 – 5% MNAR Imputation: The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age) under 5% MNAR imputed in the data is quite symmetric as it shown by histogram above and Q-Q suggest that the data point fit to the straight line except in the tails move away to the line.

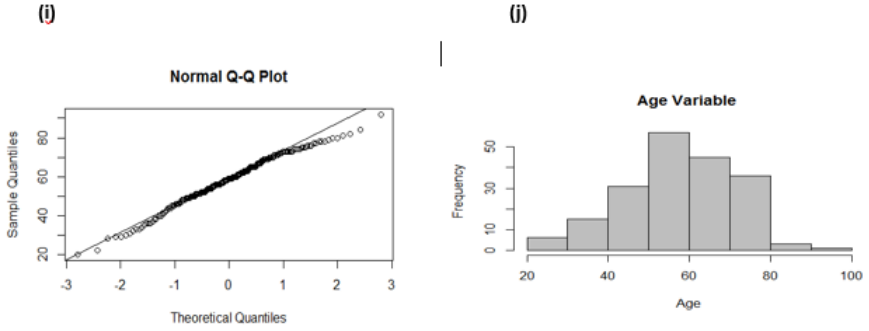

Figure 4.7 – 10% MCAR Imputation: The Q-Q Plot and Histogram for variable Age

The plot of residuals for variable (Age) suggest that the observed circles all lie quite close to the line , close enough to conclude that this data comes from normal distribution. The distribution of the variable (Age) under 10% MCAR imputed in the data is quite symmetric as it shown by histogram above and Q-Q suggest that the data point fit to the straight line except in the tails move away to the line.

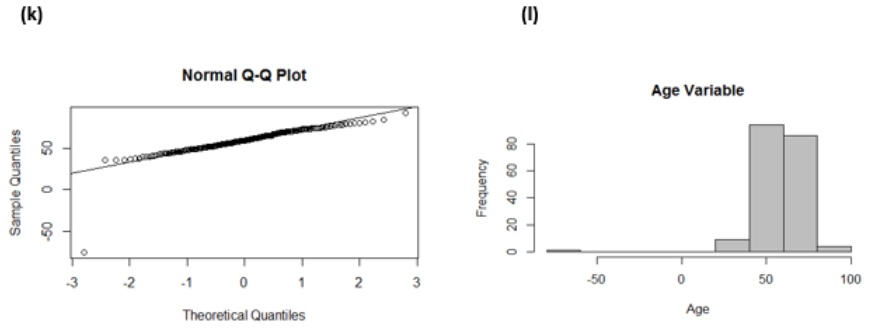

Figure 4.8 – 10% MNAR Imputation: The Q-Q Plot and Histogram for variable Age

The distribution of the variable (Age) under 10% MNAR imputed in the data is not symmetric as it shown by histogram above and the Q-Q suggest that the data point fit to the straight line and the distribution of Age might be not normally distributed at 10% MNAR data imputation.

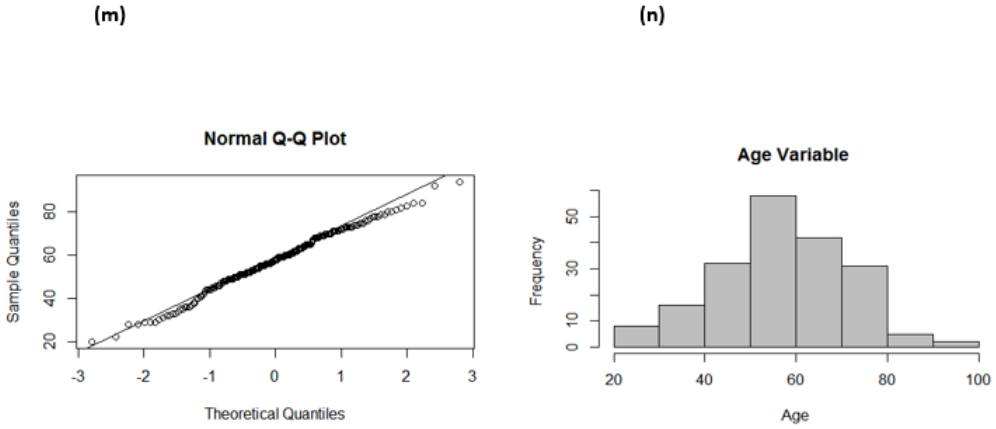

Figure 4.9 – MAR Imputation data: The Q-Q Plot and Histogram for variable Age

The plot of residuals for variable (Age) suggest that the observed circles all lie quite close to the line, close enough to conclude that this data comes from a normal distribution. The distribution of the variable (Age) under MAR imputed in the data is quite symmetric as it shown by histogram above and Q-Q suggest that the data point fit to the straight line except in the tails move away to the line.

Shapiro-Wilk Test for Normality

Table 4.7, below provides the the formal test for normality based on the imputed data set using data augmentation.

The specific null and alternative hypothesis for the Normality test is given by:

$\mathbf{H}_0$: residuals are normally distributed

$\mathbf{H}_1$: residuals are not normally distributed

Table 4.7 Shapiro test for imputed data

	Shapiro	**p-value**
MCAR1%	0.98353	0.0225
MNAR1%	0.98771	0.09158
MCAR5%	0.98864	0.1247
MNAR5%	0.99169	0.3326
MCAR10%	0.98597	0.051
MNAR10%	0.77143	4.245 e-16
MAR	0.99159	0.3223

The level of significance $\alpha = 0.05$ and Rejection Region : reject $\mathbf{H}_0$ if p-value $<$ α=0.05.

Conclusion:

Under MNAR: p-value =0.98771, we fail to reject H_0 at 5% level of significance and conclude that the residuals are normally distributed. Under MCAR: p-value=0.0225 less than 0.05 level of significance, hence we can conclude that variable Age is not not normally distributed. Under MAR: p-value=0.3223 is greater than 0.05 level of sig-

nificance, hence we can conclude that variable Age is normally distributed. Under MCAR: p-value =0.1247, we fail to reject H_0 at 5% level of significance and conclude that the residuals are normally distributed. Under MNAR: p-value=0.3326 is greater than 0.05 level of significance, we do not reject the null hypothesis and conclude that the variable Age is normally distributed. Under MAR: pvalue=0.3223 is greater than 0.05 level of significance, so we can conclude that variable Age is normally distributed. Under MCAR: reject H_0 if p-value $< \alpha = 0.05$. Since p-value=0.051, we fail to reject H_0 at 5% level of significance and conclude that the residuals are normally distributed. Under MNAR: p-value= 4.245e-16 less than 0.05 level of significance, so we can conclude that variable Age is not not normally distributed. Under MAR: p-value=0.3223 is greater than 0.05 level of significance, so we can conclude that variable Age is normally distributed.

Summary Statistics for imputed datasets

A summarized table below provides the different missing mechanisms and summary statistics for imputed and original models.

Table 4.8 Summary Statistics for imputed missing Age under different missing mechanisms

		1% IMP			*5% IMP*			*10%IMP*		
	Original	**MCAR**	**MNAR**	**MAR**	**MCAR**	**MNAR**	**MAR**	**MCAR**	**MNAR**	**MAR**
MEAN	57.49	57.479	57.572	57.521	57.572	58.644	57.521	58.196	59.129	57.521
MEDIAN	58.00	58.500	58.00	58.00	58.00	59.00	58.00	59.00	59.00	58.00
S.D	13.605	13.523	13.449	14.099	13.654	12.733	14.099	13.324	14.909	14.099
IQR	19.75	19	19.75	20.0	19.75	19	20.0	19	18	20.0

Comparisons of the original and imputed data using summary statistics above

For the descriptive statistics above, Mean, Median and Standard deviation were computed for original data and compared with the simulated missing percentage data sets for 1%, 5% and 10%. Descriptive statistics table above, shows that the Mean increase as we increase the missing percentages in the data set. Table 4.8 above shows that at 1% MCAR, the variable Age is not normally distributed then we can use median to compare with median of complete data. The median of 1% MCAR obtained is close to the median obtained from complete data set. For IQR in the above table shows that at 1% MCAR is equal to 19, which is bit closer to 19.75 of

the original data standard deviation. For the 1% MNAR data set, the variable Age is normally distributed. The mean 57.479 of 1% MNAR is slightly different results compare to complete data mean of 57.49. The standard deviation of 1% MNAR is 13.532 is slightly different of complete data. At 5% MCAR, the variable Age is normally distributed as it shown in **Table** 4.8 Shapiro test result. Comparing the mean and standard deviation of 5% MCAR to the mean and standard deviation of complete data, we observed that mean of 5% MCAR data is little bit closer to the mean of complete data. The 5% MCAR standard deviation of 13.645 is closer to 13.605 of complete data.

The results of the Shapiro Wilk test of imputed data shown in Table 4.8 shows that for 5% MNAR, the variable is still normally distributed. For the 5% MNAR data, the mean and standard deviation were completely different from the descriptive statistics of complete data without missing data. The results of the Shapiro test of imputed data shown in Table 4.8 shows that for 10% MCAR the variable is normally distributed. For the 10% MCAR data, the mean and standard deviation were completely different from the descriptive statistics of complete data without missing data. Table 4.8 shows that for 10% MNAR the variable is not normally distributed, then at 10% MNAR the median is equal to 59 but different to the median of the original data set. For the 10% MNAR the Interquartile range (IQR) is equal to 18, which is different to the original data of 19.75 IQR.

Since the variable Age is normally distributed under MAR process, we must compare standard deviation and mean of the MAR and Original data. From the Table above shows that the standard deviation of MAR process is equal to 14.099 whereas in Original data is equal to 13.605. Hence, the standard deviation of MAR process is quite larger than of the original data. The mean of the MAR process is equal to 57.521 and the original data mean is 57.49 by comparing the means of this data set, we have found that a mean of MAR data is larger than mean of the original data.

4.3.2 Hypothesis Testing

Table 4.9 Test Statistics : Difference in means of Original Age and imputed Age

TEST		MCAR 1%	MNAR 1%	MAR
T-test	*t*	-	-1	0.36
	p-value	-	0.3186	0.7200
	estimate	-	-01134021	0.1134
Wilcoxon-test	*V*	3	-	-
	p-value	1	-	-

Paired t-test for difference in means

The dependent-samples t-test was also conducted to identify variables whose pattern of missing values might be influencing the continuous variables of interest. In this case, age in completed years were considered. The means of age in completed years for missingness and completeness were calculated.

The specific null and alternative hypothesis of the paired-t test is given by:

$\mathbf{H}_0 : \mu_d = 0$ (Mean difference between the two paired measurements is zero)

$\mathbf{H}_1 : \mu_d \neq 0$ (Mean difference between the two paired measurements is not equal to zero)

The paired-t test assumes that the differences between the paired measures are normally distributed. Since the 1% MNAR above is normally distributed as it is shown by formal test which is Shapiro-Wilk test, we can use paired t-test to check for significance of the difference in means. From this row we observe the t statistic, t = -1, and p-value = 0.3186; i.e., a very high probability of this result occurring by chance, under the null hypothesis of difference in means. The null hypothesis is not rejected, since p-value> 0.05 (in fact p-value = 0.3186).Therefore, we can conclude that the paired observations are not significantly different and that the imputed data mean is close to the original mean under MNAR assumption. The differences between the paired measures are normally distributed shown by Table 4.9. Since the MAR above is normally distributed, shown using Shapiro-Wilk test, we use paired t-test to check

for significance of the difference in means. From the above Table we observe that the t statistic, t =0.36, and p-value = 0.7200; i.e., a very high probability of this result occurring by chance, under the null hypothesis of difference in means. The null hypothesis is not rejected, since p-value> 0.05 (in fact p-value = 0.7200). Therefore,we can conclude that the paired observations are not significantly different and that the imputed data mean is close to the original mean under MAR assumption.

Wilcoxon-test for location shift in observations

The null and alternative hypothesis for Wilcoxon signed rank test is given by

$\mathbf{H}_0$: difference between the pairs follows a symmetric distribution around zero

$\mathbf{H}_1$: difference between the pairs does not follow a symmetric distribution around zero.

Since p-value=1 is greater than level of significance $\alpha = 0.05$, we do not reject null hypothesis. Therefore, we do not have statistically significant evidence at $\alpha = 0.05$, to show that the median difference in original Age and imputed Age is not zero.

Table 4.10 Test Statistics : Difference in means of Original Age and imputed Age

TEST		MCAR 5%	MNAR 5%	MAR
T-test	*t*	-0.22779	-2.4836	0.36
	p-value	0.82	0.01386	0.7200
	estimate	-0.0773195	-1.149485	0.1134
Wilcoxon-test	*V*	-	-	-
	p-value	-	-	-

Paired t-test for difference in means

The paired-t test assumes that the differences between the paired measures are normally distributed. Since both MNAR and MCAR above is normally distributed as shown by the formal Shapiro-Wilk test, we can use paired t-test to check for significance of the difference in means. The observed t statistic for MCAR, t = -0.22779,

and p-value =0.82 $>$ 0.05. The null hypothesis is not rejected, since p-value$>$ 0.05. Therefore, we can conclude that the paired observations are not significantly different and that the imputed data mean is close to the original mean.

The observed t statistic for MCAR, t = -2.4836, and p-value = 0.01386 is less than level of significance $\alpha = 0.05$. The null hypothesis is rejected, since p-value$<$ 0.05. Therefore, we can conclude that the paired observations are not significantly different and that the imputed data mean is not close to the original mean.

The differences between the paired measures are normally distributed shown by Table 4.10. Since the MAR above is normally distributed shown using Shapiro-Wilk test, we used the paired t-test to check for significance of the difference in means. From the above Table we observe that the t statistic, t =0.36, and p-value = 0.7200; i.e., a very high probability of this result occurring by chance, under the null hypothesis of difference in means. The null hypothesis is not rejected, since p-value$>$ 0.05 (in fact p-value = 0.7200). We can conclude that the paired observations are not significantly different and that the imputed data mean is close to the original mean.

Table 4.11 Test Statistics : Difference in means of Original Age and imputed Age

TEST		**MCAR 10%**	**MNAR 10%**	**MAR**
T-test	*t*	-1.7573		0.36
	p-value	0.8045	-	0.7200
	estimate	-0.7010309	-	0.1134
Wilcoxon-test	*V*	-	17	-
	p-value	-	0.005204	-

Paired t-test for difference in means

The paired-t test assumes that the differences between the paired measures are normally distributed. Since the 10% MCAR above is normally distributed as it is shown by formal test which is Shapiro-Wilk test, so we can use paired t-test to check for significance of the difference in means. Since the observed t statistic, t = 1.7573, and p-value = 0.8045. The null hypothesis is not rejected, since p-value$>$ 0.05 (in fact p-value = 0.8045). Therefore, we can conclude that the paired observations are not

significantly different and that the imputed data mean is close to the original mean under MCAR assumption.

The differences between the paired measures are normally distributed as shown by Table 4.9. Since the MAR above is normally distributed shown using Shapiro-Wilk test, use paired t-test to check for significance of the difference in means. From the above Table we observe that the t statistic, t =0.36, and p-value = 0.7200; i.e., a very high probability of this result occurring by chance, under the null hypothesis of difference in means. The null hypothesis is not rejected, since p-value> 0.05 (in fact p-value = 0.7200). We can conclude that the paired observations are not significantly different and that the imputed data mean is close to the original mean under MAR assumption.

Wilcoxon-test for location shift in observations

The null and alternative hypothesis for Wilcoxon signed rank test is given by

$\mathbf{H}_0$: difference between the pairs follows a symmetric distribution around zero

$\mathbf{H}_1$: difference between the pairs does not follow a symmetric distribution around zero.

Since p-value=0.005204 is less than level of significance $\alpha = 0.05$, we reject null hypothesis. Therefore, we have statistically significant evidence at $\alpha = 0.05$, to show that the median difference in original Age is closed to imputed Age.

4.3.3 Comparing Estimates and Standard errors of the Survey Logistic Regression Results

In Table 4.12 parameter estimates, standard errors for variables gender, working for wage, Area of living and Age obtained with different missing data mechanism and proportions of missing data are presented, for intercept p-values are not presented on the results. Intercepts parameter estimates were mostly within the 2 of the original result and the standard error's change was minimal. While parameter estimates and standard errors for gender, working for wage, Area of living and Age were either larger or smaller depending on the missing data mechanism and proportions of missing data. All variables stayed statistically insignificant for all approaches. Standard errors for MCAR missing mechanisms were the closer to the original with complete case analysis and MNAR and MAR mechanisms overestimate or underestimate standard error compare to original results.

MCAR produces good estimator of the model for cancer-medication data under 1% of missing percentage. The model summary shows that the -2logLikelihood statistic in MCAR1% of the analysis generates the closest estimates compare to MNAR1% and MAR because values of -2logLikelihood are quite close to the -2logLikelihood statistic of the complete data set. Our results conclude that MCAR missing mechanism generates the close -2logLikelihood statistic values that are much closer to the true -2logLikelihood statistic values, then the MCAR appears to perform better. The Akaike Information Criterion (AIC) is the most common to estimate the quality of each model fit in the logistic regression, compared to other models based on the final model. These statistics measure how poor the model predicts the intake of cancer medication in South Africa, the smaller the statistic the better the model. In this study, we compare AIC for different models to the complete data set of cancer medication. For MCAR1% imputed model, the AIC of the analysis under these missing proportion generate slightly closer results compare to complete data set. The AIC also suggest that the better model is MCAR.

Table 4.12 Regression Table: Estimates of Covariates (and standard errors) using Complete case model and 1% propotions for MCAR, MNAR and MAR.

	Original	MCAR1%	MNAR1%	MAR%
Gender	-0.7891	-0.7812	-0.7849	0.1707
	(0.5110)	(0.5115)	(0.5103)	(0.7041)
Wage	0.0178	0.0299	-0.1353	-0.7231
	(0.6958)	(0.7000)	(0.6770)	(0.5189)
Area of living	-0.1559	-0.1622	-0.1315	-0.1965
	(0.6759)	(0.6770)	(0.6750)	(0.6666)
Age	0.0257	0.0260	0.0258	0.0167
	(0.0210)	(0.0208)	(0.0211)	(0.0217)
N	194	194	194	194
LogLik	87627.166	87573.762	87480.517	88486.250
AIC	87627.166	87583.762	87490.517	88486.250

Dependent Variable: Cancer-medication intake (1 = Yes; 0 otherwise).
Standard errors in parenthesis. Significance levels: * p<0.05

The Root MSE calculated with 1% missing data were MCAR mechanism the closest to the original, MNAR underestimate while MAR overestimated it. Root MSE and coefficient of determination were closest to the original with MCAR mechanisms of missing data method as can be seen from Table 4.13.

Table 4.13 Root MSE and Cox and SnellR^2 for different approaches with 1% missing data

Missing percentage and mechanism	**Root MSE**	**Cox and SnellR^2**
Original	8.050011	0.020367
MCAR1%	8.048562	0.020720
MNAR1%	8.046173	0.021301
MAR	8.074974	0.014282

In Table 4.14 parameter estimates, standard errors for variables gender, working for wage, Area of living and Age obtained with different missing data mechanism and proportions of missing data are presented, for intercept p-values are not presented on the results. Intercepts parameter estimates were mostly within the 2 of the original result and the standard error's change was minimal. While parameter estimates and standard errors for gender, working for wage, Area of living and Age were either larger or smaller depending on the missing data mechanism and proportions of missing data. All variables stayed statistically insignificant for all approaches. Standard errors for MCAR missing mechanisms were the closer to the original with complete case analysis and MNAR and MAR mechanisms overestimate or underestimate standard error compare to original results.

The results of our simulation shown in **Table** 4.14 suggest that when the missing mechanisms are MCAR5%, the model produce closest than other mechanisms (MNAR5%, MAR). Thus, MCAR produces good estimator of the model for cancer-medication data under 5% of missing percentage. The model summary shows that the -2logLikelihood statistic in MCAR5% of the analysis generates the fewer bias estimates compare to MNAR5% and MAR because values of -2logLikelihood are quite close to the -2logLikelihood statistic of the complete data set. Our results conclude that MCAR missing mechanism generates the close -2logLikelihood statistic values that are much closer to the true -2logLikelihood statistic values, then the MCAR appears to perform better. The Akaike Information Criterion (AIC) is the most common to estimate the quality of each model fit in the logistic regression, compared to other models based on the final model. These statistics measure how poor the model predicts the intake of cancer medication in South Africa, the smaller the statistic the better the model. In this study, we compare AIC for different models to the complete data set of cancer medication. For MCAR5% imputed model, the AIC of the analysis under these missing proportion generate slightly closer results compare to complete data set. The AIC also suggest that the better model is MCAR.

Table 4.14 Regression Table: Estimates of Covariates (and standard errors) using Complete case model and 5% propotion for MCAR, MNAR and MAR.

	Original	MCAR5%	MNAR5%	MAR
Gender	-0.7891	-0.7870	-0.7033	0.1707
	(0.5110)	(0.4939)	(0.5192)	(0.7041)
Wage	0.0178	-0.1514	-0.2154	-0.7231
	(0.6958)	(0.5958)	(0.7152)	(0.5189)
Area of living	-0.1559	-0.1163	-0.0938	-0.1965
	(0.6759)	(0.6713)	(0.6620)	(0.6666)
Age	0.0257	0.0237	0.0152	0.0167
	(0.0210)	(0.0200)	(0.0237)	(0.0217)
N	194	194	194	194
LogLik	87617.166	87592.356	88360.642	88486.250
AIC	87627.166	87602.356	88370.642	88486.250

Dependent Variable: Cancer-medication intake (1 = Yes; 0 otherwise).
Standard errors in parenthesis. Significance levels: * p<0.05

Table 4.15 Root MSE and Cox and SnellR^2 for different approaches with 5% missing data

Missing percentage and mechanism	**Root MSE**	**Cox and SnellR^2**
Original	8.050011	0.020367
MCAR5%	8.049384	0.020520
MNAR5%	8.071041	0.015242
MAR	8.074974	0.014282

As seen above, The Root MSE calculated with 5% missing data for MCAR mechanism the closest to the original, MNAR and MAR overestimated it. The most similar coefficient of determination with original data was with MCAR mechanisms of missing data method and other missing mechanisms underestimates it as can be seen from Table 4.15.

In Table 4.16 parameter estimates, standard errors for variables gender, working for wage, Area of living and Age obtained with different missing data mechanism and proportions of missing data are presented, for intercept p-values are not presented on the results. Intercepts parameter estimates were mostly within the 2 of the original result and the standard error's change was minimal. While parameter estimates and standard errors for gender, working for wage, Area of living and Age were either larger or smaller depending on the missing data mechanism and proportions

of missing data. All variables stayed statistically insignificant for all approaches. Standard errors for MCAR missing mechanisms were the closer to the original with complete case analysis and MNAR and MAR mechanisms overestimate or underestimate standard error compare to original results.

The model summary shows that the -2logLikelihood statistic in MCAR10% of the analysis generates the fewer bias estimates compare to MNAR10% and MAR because values of -2logLikelihood are quite close to the -2logLikelihood statistic of the complete data set. Our results conclude that MCAR missing mechanism generates the close -2logLikelihood statistic values that are much closer to the true -2logLikelihood statistic values, then the MCAR appears to perform better. The Akaike Information Criterion (AIC) is the most common to estimate the quality of each model fit in the survey logistic regression, compared to other models based on the final model. These statistics measure how poor the model predicts the intake of cancer medication in South Africa, the smaller the statistic the better the model. In this study, we compare AIC for different models to the complete data set of cancer medication. For MCAR10% imputed model, the AIC of the analysis under these missing proportion generate slightly closer results compare to complete data set. The AIC also suggest that the better model is MCAR.

Table 4.16 Regression Table: Estimates of Covariates (and standard errors) using Complete case model and 10% propotions for MCAR, MNAR and MAR.

	Original	MCAR10%	MNAR10%	MAR
Gender	-0.7891	-0.6736	-0.8958	0.1707
	(0.5110)	(0.4720)	(0.5276)	(0.7041)
Wage	0.0178	-0.1170	0.3260	-0.7231
	(0.6958)	(0.6619)	(0.8248)	(0.0118)
Area of living	-0.1559	-0.1548	-0.1289	-0.0982*
	(0.6759)	(0.6610)	(0.6802)	(0.0130)
Age	0.0257	0.00512	0.0274	0.0167*
	(0.0210)	(0.0216)	(0.0130)	(0.000699)
N	194	194	194	194
LogLik	87617.166	88960.573	85152.547	88486.250
AIC	87627.166	88970.573	85162.547	88486.250

Dependent Variable: Cancer-medication intake (1 = Yes; 0 otherwise).
Standard errors in parenthesis. Significance levels: * p<0.05

The Root MSE calculated with 10% missing data were MCAR mechanism the closest to the original, MNAR underestimate while MAR overestimated it. The most similar coefficient of determination with original data was with MCAR mechanisms of missing data method as can be seen from Table 4.17.

Table 4.17 Root MSE and Cox and SnellR^2 for different approaches with 10% missing data

Missing percentage and mechanism	**Root MSE**	**Cox and SnellR^2**
Original	8.050011	0.020367
MCAR10%	8.049384	0.020520
MNAR10%	7.934878	0.048188
MAR	8.074974	0.014282

Table 4.18 Root MSE and Cox and SnellR^2 for different approaches with 1% ,5% ,10% missing data

Missing percentage and mechanism	**Root MSE**	**Cox and SnellR^2**
Original	8.050011	0.020367
MCAR1%	8.048562	0.020720
MNAR1%	8.046173	0.021301
MCAR5%	8.049384	0.020520
MNAR5%	8.071041	0.015242
MCAR10%	8.049384	0.020520
MNAR10%	7.934878	0.048188
MAR	8.074974	0.014282

As seen above, the RMSE for MCAR under 1%, 5%, 10% were consistently the smallest out of the MNAR and MAR missing data mechanisms methods. This suggest that MCAR is more accurate than other methods. The Root MSE and coefficient of determination were closest to the original with with MCAR mechanisms under 1%, 5% and 10% of missing data method as shown in the above Table: 4.18.

Parameter estimates and standard errors visualizations

The visualizations of regression coefficients and standard errors are used to check the biasness of estimates from the true estimates and overestimation or underestimation of standard errors in the models.

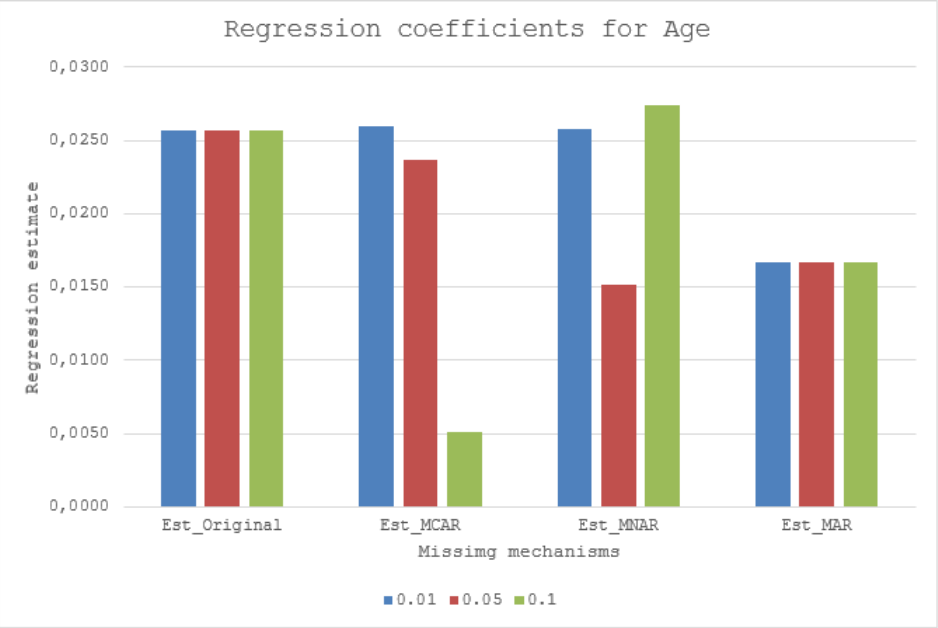

Figure 4.10 – Estimates for Age when the MCAR, MAR, and MNAR methods are used at different rates of missingness

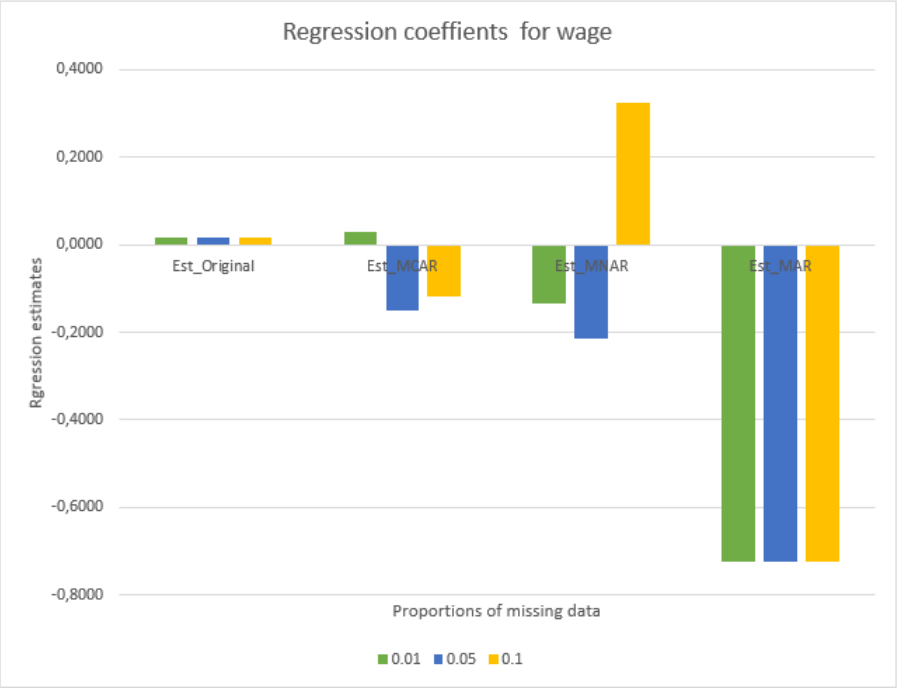

Figure 4.11 – Estimates for Working for Wage when the MCAR, MAR, and MNAR methods are used at different rates of missingness

Figure 4.12 – Estimates of standard errors for Age when the MCAR, MAR, and MNAR methods are used at different rates of missingness

Figure 4.13 – Estimates of standard errors for Working for Wage when the MCAR, MAR, and MNAR methods are used at different rates of missingness

Conclusion

In general, the visuals above show that the higher the proportions of missing data, the more imputed data regression estimates become deviated from the original data estimates for both imputed Age and Wage. It also indicates that regression coefficients for variable Age under MCAR is quite closer to the original coefficient estimates. We can conclude that the MCAR, yields less biased estimates than MNAR and MAR for Age variable. Furthermore, the results indicate that MNAR method yields negative estimates whereas other methods produce positive estimates. The regression coefficients for wage under different rates of missingness also suggest that MCAR is good method than other methods because it yields much less biased estimates compare to MAR and MNAR.

Observing the Figure 4.12 and 4.13, the standard errors behaviour shows that MCAR standard errors for different rates of missingness are quite closer to the standard error for complete data. For the MNAR, we observed that the standard errors are much deviated from the complete data analysis standard error for variable Age and Wage. The investigation of behaviour of MCAR, MNAR and MAR for different rates of missingness (1%, 5%, 10%) shows that the MCAR yields less biased estimates for both variable Age and Wage followed by MNAR and it biased for MAR. Furthermore, MCAR produces close standard error to those of complete data than other methods.

4.3.4 Assessment of adequacy of the models

Table 4.19 Evaluations of the Survey Logistic Regression Model

Missing Mechanism(%)	Test	F-value	p-value
Original	Likelihood ratio test	0.99	0.4138
	Wald test	1.21	0.3083
MCAR1%	Likelihood ratio test	1.01	0.4058
	Wald test	1.25	0.2926
MCAR5%	Likelihood ratio test	1.00	0.4092
	Wald test	1.49	0.2062
MCAR10%	Likelihood ratio test	0.53	0.7174
	Wald test	0.93	0.4503
MNAR1%	Likelihood ratio test	1.04	0.3888
	Wald test	1.24	0.2958
MNAR5%	Likelihood ratio test	0.73	0.5705
	Wald test	0.79	0.5318
MNAR10%	Likelihood ratio test	1.84	0.1219
	Wald test	1.73	0.1451
MAR	Likelihood ratio test	0.69	0.6001
	Wald test	0.93	0.4456

Statistical tests of individual predictors of the model

The **Likelihood Ratio test** was used to evaluates the significance of the joint effect of all the variables in the Survey logistic regression procedure. Since (p-value$>$ 0.05) of the Likelihood ratios of of different missing percentages and mechanisms was greater than 0.05, Hence, we can conclude that likelihood ratio test are insignificance which means that the joint of the variables in the full model of cancer medication is insignificant than just the intercept model.

The **Wald test** is used to evaluates the significance of the joint effect of all the variables or individual variable in predicting the probability of taking and not taking cancer medication in the Survey logistic regression procedure. Since (p-value$>$ 0.05) of the Wald statistics for all different missing percentages and mechanisms, Hence the insignificance of the Wald test means that the joint of the variables in the full model is insignificant than just the intercept model.

Comparing the results of imputed and complete data

Table 4.18 above, shows the results of the assessment of the adequacy of the models for cancer medication data that is affected by missing data problem. The test depends on the nature of missingness and proportions of missing data. The aim is to select the model that works best in fitting this data by comparing the imputed data sets with complete data. To evaluate the adequacy, we compare the likelihood ratio test, Wald test of the imputed data sets with likelihood ratio test, Wald test and from the analysis of the complete data set. For the different missing percentages of missing data under MCAR, we observed that for MCAR the likelihood ratio test, Wald test for imputed data gave the slightly different results compared to complete data set. The likelihood ratio test, Wald test values obtained from different imputed data set under MCAR were a bit closer to the test obtained from complete data. For both MNAR and MAR data, the likelihood ratio test, Wald test were completely different from the likelihood ratio test, Wald test of complete data without missing data, but they lead to the same results.

Table 4.20 Survey Logistic Regression Model Summary

Data set	Model	-2log likelihood	AIC
Original	Null Model	90475.406	90477.406
	Final Model	87617.166	87627.166
MCAR1%	Null Model	90475.406	90477.406
	Final Model	87573.762	87583.762
MCAR5%	Null Model	90475.406	90477.406
	Final Model	87592.356	87602.356
MCAR10%	Null Model	90475.406	90477.406
	Final Model	88960.573	88970.573
MNAR1%	Null Model	90475.406	90477.406
	Final Model	87480.517	87490.517
MNAR5%	Null Model	90475.406	90477.406
	Final Model	88360.642	88370.642
MNAR10%	Null Model	90475.406	90477.406
	Final Model	85152.547	85162.547
MAR	Null Model	90475.406	90477.406
	Final Model	88486.250	88486.250

The Likelihood Ratio Test is the most common assessment of the overall model fit in the logistic regression, which is the difference in -2log likelihood statistic estimate of the null model (i.e only constant) and the model containing the predictors. This statistic measures how poor the model predicts the intake of cancer medication in South Africa, the smaller the statistic the better the model. The model summary shows that the -2logLikelihood statistic in MCAR1% and MCAR5% of the analysis are also little bit closer to the -2logLikelihood statistic of the complete data set. We can conclude that MCAR missing mechanism gives the closest -2logLikelihood statistic values, to the true -2logLikelihood statistic values. The Akaike Information Criterion (AIC) is the most common to estimate the quality of each model fit in the logistic regression, compared to other models based on the final model. This statistic measures how poor the model predicts the intake of cancer medication in South Africa, the smaller the statistic the better the model. In this study, we compare AIC for different models to the complete data set of cancer medication. For MCAR

imputed models, the AIC of the analysis under different missing proportions gave slightly different results compared to complete data set. Compare to MNAR and MAR imputed data sets, the AIC for MCAR is little bit closer to the AIC obtained from complete data set.

The Akaike Information Criterion (AIC) visualizations

The visualizations of Akaike Information Criterion (AIC) is used to measures the quality of each model fit in the Survey logistic regression model. The selection was done by comparing the original model (AIC) and Akaike Information Criterion (AIC) for each model under different missing percentages and missing mechanisms (MAR, MCAR or MNAR). Thus, the closer the AIC value of the model to those of original, the model is taken as a good model.

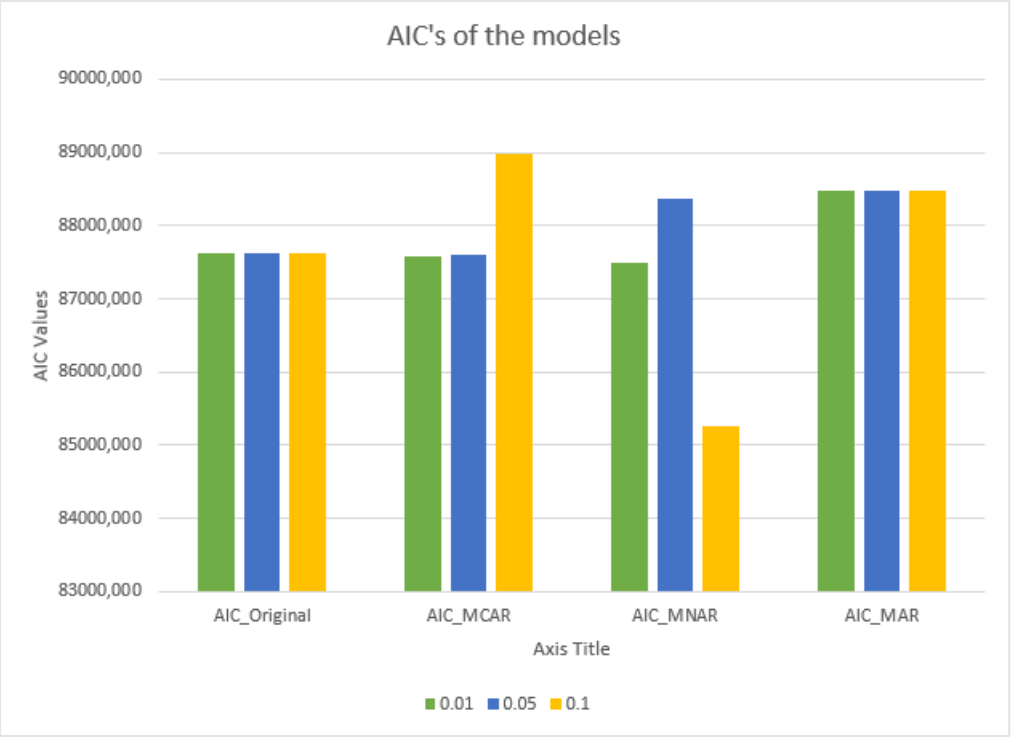

Figure 4.14 – Comparisons of values of Akaikes Information Criterion (AIC) when the MCAR, MAR, and MNAR methods are used at different rates of missingness

Conclusion: The MCAR was selected since values of imputed models were slightly closer to the Akaike Information Criterion (AIC) of the original model compare to other models.

4.3.5 Kappa Test for agreement in wage

The Kappa statistics is the measure of inter variation across the cross tabulation or it is a measure of the inter-rater agreement between two or more ratters. Under Kappa statistics we assume each observation in the cross tabulation is called subject. The variables, however, record frequencies with which rating were assigned.

Interpretation of Kappa:

If the Kappa test has a range between 0 and 1 with larger values indicating better reliability. In General, if Kappa test if greater than 0.70 is considered satisfactory. But if Kappa test is less than 0.70, we can conclude that the inter-rater reliability is not satisfactory.

Table 4.21 The Kappa test of agreement between observed and imputed wage variable

Variable	Missing(%)	MCAR	MNAR	MAR
Wage	1%	1	0.9532	0.7311
	5%	0.9104	0.9206	0.7311
	10%	0.9247	0.8337	0.7311

The Kappa test agreement shows that under different missing mechanism and missing percentages 1%, 5% and 10% the Kappa values are greater than 0.70. This means that the agreement between imputed wage and original are satisfactory.

Refer to Table 4.20 above.

Under MCAR Mechanism:

At 1% missing percentage, the Kappa test value is equal to 1, which means that it almost perfectly agrees between imputed wage and original wage. Under 5% and 10% missing percentage the Kappa test value is included into a range of 0.81-1, we can conclude that the agreement between observed and imputed wage are almost perfect under MCAR missing mechanism.

Under MNAR missing mechanism:

All Kappa test values are almost perfectly agree because fall in the range of 0.81 and 1.

Under MAR missing mechanism:

Kappa test value is 0.7311, shows the Substantial agreement because fall in the range of 0.61 and 0.80. We can conclude that the agreement between observed and imputed wage is substantial agreement under MAR missing mechanism. This visual repre-

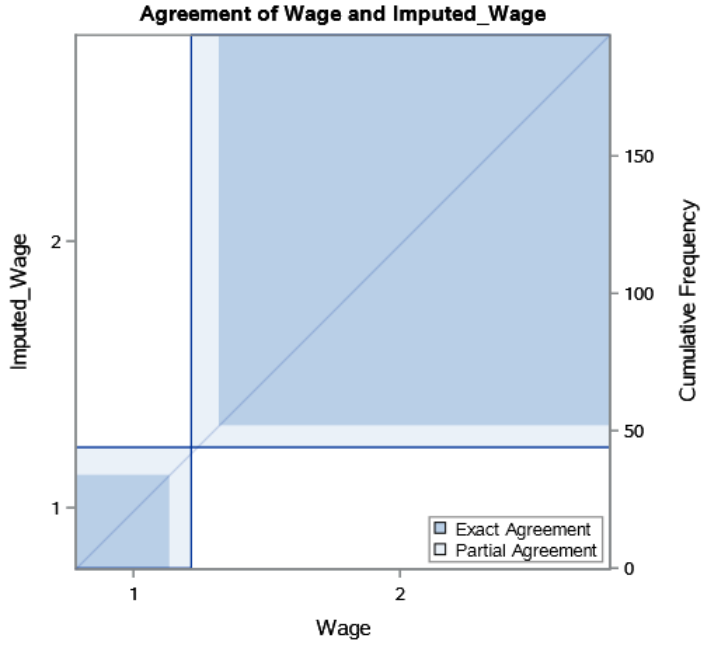

Figure 4.15 – The plot of both the agreement and distribution of working for wage under MAR

sentation of the agreement shows that there was a large amount of exact agreement (dark blue shading) for No (Not working for a wage), Wage 2, with a small percentage partial agreement. With 3 categories, only exact agreement or partial agreement is possible for the middle category. Two other takeaway points from this plot are that agreement is lower for Yes (Working for a wage) and partially Yes (Working for a wage) than not working for wage patients ones and that the distribution is skewed, with a large proportion of not working for wage patients. Because it is adjusted for chance agreement, Kappa is affected by the distribution among these categories.

Chapter 5

Discussion and conclusion

5.1 Introduction

This chapter presents a summary of all the preceding chapters in this study. Furthermore, potential advantages and shortcomings of the study and directions for future research are also given in this chapter.

5.2 Summary

Missing data is a common problem that arise in many fields of study and empirical research. This problem may lead to a small sample size and that may cause serious impact on the estimates to become more biased and that lead to incorrect decisions, especially when data set consist of many missing data values. Thus, it is important to adequately handle them for us to obtain reliable results all the time. The study has focused on the imputation method called Bayesian data augmentation algorithm for analysing cancer medication data affected by missing data problem. More specifically, a cancer-medication dataset studied contains a some proportion rates of missing observations. In order to avoid problems with missing data in the model building, such as loss of information, bias and reduced prediction power, an appropriate method of tackling missing data is needed to identified in order to solve such problem. This work has focused on statistical methods aimed at modelling the cancer-medication data. In recent years, the Tanner & Wong introduced the method of data augmentation to handle the problem of missing data in 1987. The data augmentation in these years gained the popularity as one of the best techniques to handle missing data especially if dealing with small sample sized data sets

(Allison, 2001; McKnight et al., 2007).

The Data augmentation (DA) algorthm is originally introduced by M. A. Tanner & Wong (1987a) as the estimation procedure for dealing with the latent data or unobserved values by adding latent data on the observed data, so that will be easier to analyse data with missing values. The data augmentation can be implemented in solving or simulating missing values, so that we can be able to compute full posterior distributions of $\boldsymbol{\theta}$ (Tanner & Wong, 1987). The EM and Data augmentation has the similar idea that rather than do complicated optimization calculations, the observed data can be added missing data so that to simulate unknown values is easy. The idea is that the DA is the stochastic version of EM algorithm. The Markov chain Monte Carlo data augmentation has the advantage over the classical EM procedure because it shows greater flexibility when the underlying distributions are unknown unlike ML model-based methods which only function under the certain distributional assumptions. These techniques impute missing data by replacing missing data with random drawn values from the predictive distribution based on the observed data set (J. L. Schafer & Graham, 2002). The MCMC procedure has two iterative steps to augment the data, which is the imputation (I-step) and posterior (P-step). In the imputation(I) step, the missing observations are imputed for each value independently given the observed data. In the posterior(P) step, usually use the complete data set (the observed and simulated missing data) to determine the samples of the posterior distribution for the unknown parameters, based on the imputed data in the imputation step above. The iteration process for MCMC start with imputation (I) and proceeds to posterior(P) step to create the process of Markov chain. The iterative process is done multiple times until it reach the convergence state.

Schafers multiple imputation software (1991) uses multivariate normal models, multinomial models and general location models to impute missing values for continuous variables, categorical variables and mixed variables, respectively. All models assume that the missing mechanism is ignorable; that is, missing values occur at random. Brief descriptions of the three types of models are found in Shafer (1991). The important difference between Data Augmentation and other imputation methods is that the dependent variable should not have any missing values and only independent variables are allowed to contain missing values in it. Data Augmentation has advantages over other techniques, since it impute both categorical and continuous independent variables at the same time with missing data problems. According to J. L. Schafer (1997), avoid to use more categorical variables in the model since it may lead to structural zeroes and that may lead to an improper posterior distribution ob-

tained in the analysis. In this thesis we focus on the unrestricted general location model since the independent variables for cancer medication with missing data are both continuous and categorical (working for wage, age) variables.

To apply unrestricted general location model to incomplete mixed data above, the software starts by using the EM algorithm to compute the maximum likelihood estimates for each cell probabilities, the cell means and the covariances. EM algorithm may be used as starting points simulation step iteration in the algorithm. Thus, the software starts to apply the iterative simulation technique in the loop to reproduce one or more iterations of a single Markov chain. The iterations simulated consists of two steps which is I-step and P-step. In the I-step the random imputation for both missing in categorical and continuous data are drawn from the predicted multinomial distribution and multivariate normal distribution, respectively, with current estimate of the parameter in the model (J. L. Schafer, 1997).

The restricted general location model is very useful when n is large compared to number cells in the given data set. The unrestricted general location model has $(C-1)+Cp+\frac{p(p+1)}{2}$ free parameters and it became difficult to compute $C \times p$ estimate as number of categorical variables p increase in the model. The Bayesian Iterative Proportional Fitting (BIPF) algorithm is used to reduce number of the parameters needed to be estimated in the general location model(J. L. Schafer, 1997) . In application, the package ***mix*** is used in **R** software that was developed by J. L. Schafer (1997), which can be downloaded from cran. In the ***mix*** library, the general model location data augmentation uses the function ***da.mix*** and ***BIPF*** algorithm for the restricted general location model use ***dabipf.mix*** function.

The primary objective of this study was to examine the behaviour of data augmentation, that is it improving the quality of data analysis under different missing proportions and missing mechanisms (MCAR, MAR and MNAR). To identify under which mechanism the method works adequately using the cancer-medication data set. The purpose of the study is to determine whether or not the Bayesian data augmentation improve the performance under different proportion rates of missing data, when the data is missing at random, completely missing at random or not missing at random. The conclusion concerning to the specific objective 2, we explore the behaviour of Bayesian data augmentation by looking at the impact of different proportions rates of missing data. The rate of missing data was 1%, 5% and 10% were considered in the analysis. We observe that models under different proportion rates of missing

change the behaviour of the Bayesian data augmentation. The lower missing data percentages Bayesian data augmentation works better than in higher rates of missing data.

Concluding based on the specific objective 2 and 3, models with different missing percentages on the variables of interest (using MCAR, MAR and MNAR) were estimated and results were compared to the results of complete data and models fitted after Bayesian data augmentation was used. However, the MAR may not appropriate for this particular data set of cancer medication since the estimates are different from the estimates of the original data and standard errors are overestimated all time. The MAR data mislead the conclusions of the analysis because of biased estimates and overestimated standard errors for each and every variable in the model. under MAR the best of assumption to use data augmentation according to Schafer(1997), they typically yield biased parameter estimates, biased standard estimates for cancer-medication data. Our findings indicate that the MCAR models under different missing proportions yields better estimates results regardless of missingness proportions it do better than MNAR and MAR is very poor. The MCAR give the acceptable performance of the parameter estimates, since the regression coefficients are close to estimates of complete data set.

Concluding based on the specific objective, models with different missing percentages on the variables of interest (using MCAR, MAR and MNAR) were estimated and results were compared to the results of complete data and models fitted after Bayesian data augmentation was used. The evaluation of the adequacy findings show that the MCAR is also more adequate closer to the original data by comparing Likelihood ratio test, Wald test and Hosmer and Lemeshow test values. In terms of the goodness of fits, the Hosmer-Lemeshow for MCAR yields closer Chi-square values to those of complete data set of cancer medication. The MNAR in this study works better than MAR, it follows the MCAR in terms of performance. Assessing the overall model fit of the Survey logistic regression models for each missingness rate specified and under specified nature missing, the loglikelihood ratio is used. The loglikelihood ratio test emphasis that MCAR models perform very well to imputed missing value over MNAR and MAR. The -2loglikelihood for Final model are closer to the actual values from the complete data set. The AIC values of the MCAR data models also quite close to the AICs of the original data, which means that quality of these models are more similar to complete data.

On the other hand, if the sample is small or if the proportion of cases with missing data is large, the additional variation can make a noticeable difference. According to (McKnight et al., 2007), the data augmentation is good to be used when the sample size is small. In situation of small data set particularly in cancer related survey, it can be advisable to use data augmentation algorithm to impute such data set. The study shows that the data augmentation is good in MCAR followed by MNAR and poor under MAR. As seen above, the RMSE for MCAR under 1%, 5%, 10% were consistently the smallest out of the MNAR and MAR missing data mechanisms methods. This suggest that MCAR is more accurate than other methods. The Root MSE and coefficient of determination were closest to the original with with MCAR mechanisms under 1%, 5% and 10% of missing data method followed by MNAR and MAR methods.That is why this study highly recommend use of data augmentation for situation of missing completely at random. However, it can also be used in general for small sample sizes. But, generalizing this results to any small data set in every discipline should be done with caution because it depends on data set.

5.2.1 Limitations and Recommendations for Future Research

The Bayesian data augmentation methods were applied to handling missing data problems by using different models such as multivariate normal model etc. This method has its strengths and weaknesses. The limitations of using Bayesian data augmentation for mixed data set is listed as follows :

- it does not allow the dependent variable to have any missing data (it must be complete).
- increasing the categorical variable in the model, may lead to structural zeros in the model.
- structural zeros, may lead to improper posterior distribution.
- there are many factors that can affect whether or not cancer patients medication (e.g. the stage of illness, the type of cancer etc.) but the information of these factors was not collected and therefore is unavailable to be used in this study.

One of the future directions of this thesis is to compare Bayesian data augmentation with Non-Bayesian data augmentation methods such as ML, EM and others to evaluate which one works better.

References

Afifi, A., & Elashoff, R. (1966). Missing observations in multivariate statistics i. review of the literature. *Journal of the American Statistical Association*, *61*(315), 595–604.

Aitchison, J., & Dunsmore, I. (1976). Statistical prediction analysis. *Bulletin of the American Mathematical Society*, *82*, 683–688.

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, *88*(422), 669–679.

Allison, P. D. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological methods & research*, *28*(3), 301–309.

Allison, P. D. (2001). Missing data: Sage university papers series on quantitative applications in the social sciences (07–136). *Thousand Oaks, CA*.

Allison, P. D. (2002). Missing data: Quantitative applications in the social sciences. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 193–196.

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology*, *112*(4), 545.

Anderson, A. B., Basilevsky, A., & Hum, D. P. (1983). Measurement: Theory and techniques. *Handbook of survey research*, 231–287.

Anderson, A. B., Basilevsky, A., Hum, D. P., et al. (1983). Missing data: A review of the literature. *Handbook of survey research*, *4*, 415–494.

Anderson, M. (2010). *1984*. JSTOR.

Andridge, R. R., & Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International statistical review*, *78*(1), 40–64.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. *Advanced structural equation modeling: Issues and techniques*, *243*, 277.

Baker, S. G. (1992). A simple method for computing the observed information matrix when using the em algorithm with categorical data. *Journal of Computational and Graphical Statistics*, *1*(1), 63–76.

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand journal of public health*, *25*(5), 464–469.

Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American Sociological Review*, 386–398.

Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, *34*(3), 277–313.

Besag, J., & Green, P. J. (1993). Spatial statistics and bayesian computation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25–37.

Birhanu, T., Molenberghs, G., Sotto, C., & Kenward, M. G. (2011). Doubly robust and multiple-imputation-based generalized estimating equations. *Journal of biopharmaceutical statistics*, *21*(2), 202–225.

Bland, L. (1995). *Banishing the beast: English feminism and sexual morality, 1885-1914*. JSTOR.

Blenkner, M., Bloom, M., & Nielsen, M. (1971). Pro-tective services for older people: Report on a controlled demonstration. *in Social Casework*.

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.

Casella, G., & George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, *46*(3), 167–174.

Clayton, D., & Hills, M. (2013). *Statistical models in epidemiology*. OUP Oxford.

Clutton-Brock, T., Brotherton, P., Smith, R., McIlrath, G., Kansky, R., Gaynor, D., ... Skinner, J. (1998). Infanticide and expulsion of females in a cooperative mammal. *Proceedings of the Royal Society of London B: Biological Sciences*, *265*(1412), 2291–2295.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (1983). Applied multiple regression/-correlation for the behavioral sciences.

Collins, P. G., Arnold, M. S., & Avouris, P. (2001). Engineering carbon nanotubes and nanotube circuits using electrical breakdown. *science*, *292*(5517), 706–709.

Conover, W. J., & Conover, W. J. (1980). Practical nonparametric statistics.

Couvreur, C. (1996). Hidden markov models and their mixtures. *Dept. Math., Université Catholique de Louvain, Louvain, Belgium*.

Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. CRC Press.

Craig, C. L., Marshall, A. L., Sjöström, M., Bauman, A. E., Booth, M. L., Ainsworth, B. E., ... others (2003). International physical activity questionnaire: 12-country reliability and validity. *Medicine & science in sports & exercise*, *35*(8), 1381–1395.

Damlen, P., Wakefield, J., & Walker, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *61*(2), 331–344.

David, H. A., & Gunnink, J. L. (1997). The paired t test under artificial pairing. *The American Statistician*, *51*(1), 9–12.

de Leeuw, D., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics*, *19*(2), 153.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Derrida, B., Bray, A., & Godreche, C. (1994). Non-trivial exponents in the zero temperature dynamics of the 1d ising and potts models. *Journal of Physics A: Mathematical and General*, *27*(11), L357.

Donner, A. (1982). The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. *The American Statistician*, *36*(4), 378–381.

Fayers, P. M., Curran, D., & Machin, D. (1998). Incomplete quality of life data in randomized trials: missing items. *Statistics in medicine*, *17*(5-7), 679–696.

Flanders, W. D., & Greenland, S. (1991). Analytic methods for two-stage case-control studies and other stratified designs. *Statistics in Medicine*, *10*(5), 739–747.

Fowler, J. W. (1988). *Faith development and pastoral care*. Fortress press.

Friedman, L. M., Furberg, C., DeMets, D. L., Reboussin, D. M., & Granger, C. B. (1998). *Fundamentals of clinical trials* (Vol. 3). Springer.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., & Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing*, *19*(4), 479–492.

Fuchs, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *Journal of the American Statistical Association*, *77*(378), 270–278.

Gelfand, A. E., & Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, *85*(410), 398–409.

Gelfand, M. P., Singh, R. R., & Huse, D. A. (1990). Perturbation expansions for quantum many-body systems. *Journal of Statistical Physics*, *59*(5-6), 1093–1142.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 457–472.

Geman, S., & Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*(6), 721–741.

Geweke, J., et al. (1994). *Bayesian comparison of econometric models* (Tech. Rep.). Working Paper, Federal Reserve Bank of Minneapolis, Minnesota.

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical science*, 473–483.

Gibbons, J. D., & Chakraborti, S. (2011). Nonparametric statistical inference. In *International encyclopedia of statistical science* (pp. 977–979). Springer.

Gilks, W. R. (1996). Full conditional distributions. *Markov chain Monte Carlo in practice*, 75–88.

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., & McNeil, A. J. (1993). Modelling complexity: applications of gibbs sampling in medicine. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39–52.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. (1995). *Markov chain monte carlo in practice*. CRC press.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex bayesian modelling. *The Statistician*, 169–177.

Gill, M., et al. (2002). A high-performance nonvolatile memory technology for standalone memory and embedded applications.

Gilley, O. W., & Leone, R. P. (1991). A two-stage imputation procedure for item nonresponse in surveys. *Journal of Business Research*, *22*(4), 281–291.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, *60*, 549–576.

Graham, J. W., & Hofer, S. M. (1991). Emcov. exe users guide. *Computer software manual]. Unpublished manuscript, University of Southern California, Los Angeles*.

Green, A., Purdie, D., Bain, C., Siskind, V., Russell, P., Quinn, M., & Ward, B. (1997). Tubal sterilisation, hysterectomy and decreased risk of ovarian cancer. *International journal of cancer*, *71*(6), 948–951.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67–82.

Hartley, H., & Hocking, R. (1971). The analysis of incomplete data. *Biometrics*, 783–823.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4* (pp. 475–492). NBER.

Higdon, D. M. (1998). Auxiliary variable methods for markov chain monte carlo with applications. *Journal of the American Statistical Association*, *93*(442), 585–595.

Hilbe, J. M. (2011). Logistic regression. In *International encyclopedia of statistical science* (pp. 755–758). Springer.

Holmes, C. S., Swift, E. E., Chen, R., & Hershberger, A. (2006). Demographic risk factors, mediators, and moderators in youths diabetes metabolic control. *Annals of Behavioral Medicine*, *32*(1), 39–49.

Horton, N. J., & Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, *8*(1), 37–50.

Hsu, H., & Lachenbruch, P. A. (2008). Paired t test. *Wiley Encyclopedia of Clinical Trials*.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765–769.

Institute, S. (2005). *Sas 9.1. 3*. SAS Brasil.

Kahn, H. A., & Sempos, C. T. (1989). *Statistical methods in epidemiology* (Vol. 12). Monographs in Epidemiology & B.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81.

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

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American political science review*, *95*(1), 49–69.

Korn, E. L., & Graubard, B. I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, *49*(3), 291–295.

Krishnan, T., & McLachlan, G. (1997). The em algorithm and extensions. *Wiley*, *1*(997), 58–60.

Krzanowski, W. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. *Biometrics*, 991–1002.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 493–499.

Lam, F., & Longnecker, M. (1983). A modified wilcoxon rank sum test for paired data. *Biometrika*, *70*(2), 510–513.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Li, K. (1988). Ivy: A shared virtual memory system for parallel computing. *ICPP (2)*, *88*, 94.

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

Lilienfeld, D. E., & Stolley, P. D. (1994). *Foundations of epidemiology*. Oxford University Press, USA.

Lipsitz, S. R., & Ibrahim, J. G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, *83*(4), 916–922.

Lipsitz, S. R., & Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the cox model. *Biometrics*, 1002–1013.

Little, R. J. (1988a). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287–296.

Little, R. J. (1988b). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202.

Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.

Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research*, *18*(2-3), 292–326.

Little, R. J., & Rubin, D. B. (2000). Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annual review of public health*, *21*(1), 121–145.

Little, R. J., & Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical Analysis with Missing Data, Second Edition*, 200–220.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Little, R. J., & Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika*, *72*(3), 497–512.

Lohr, S. L. (2009). Multiple-frame surveys. *Handbook of statistics*, *29*, 71–88.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

Malhotra, N. K. (1987). Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research*, 74–84.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

McLachlan, G. J., & Krishnan, T. (1997). Wiley series in probability and statistics. *The EM Algorithm and Extensions, Second Edition*, 361–369.

Meng, X.-L., & Rubin, D. B. (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association*, *86*(416), 899–909.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, *80*(2), 267–278.

Meng, X.-L., & Van Dyk, D. (1997). The em algorithman old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(3), 511–567.

Meng, X.-L., & Van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, *86*(2), 301–320.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Mira, A., & Tierney, L. (1997). On the use of auxiliary variables in markov chain monte carlo sampling. *Scandinavian Journal of Statistics*.

Molenberghs, G., & Kenward, M. (2007). *Missing data in clinical studies* (Vol. 61). John Wiley & Sons.

Molnar, F. J., Hutton, B., & Fergusson, D. (2008). Does analysis using last observation carried forward introduce bias in dementia research? *Canadian Medical Association Journal*, *179*(8), 751–753.

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*, *24*(7), 815–829.

Olkin, I., & Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics*, 448–465.

Peng, H., & Zhu, S. (2007). Handling of incomplete data sets using ica and som in data mining. *Neural Computing and Applications*, *16*(2), 167–172.

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

Piantadosi, S. (1997). Treatment allocation. *Clinical Trials: A Methodologic Perspective, Second Edition*, 331–353.

Prempeh, E. A. (2009). *Comparative study of the logistic regression analysis and the discriminant analysis* (Unpublished doctoral dissertation). University of Cape Coast.

Raghunathan, T. E. (2004). What do we do with missing data? some options for analysis of incomplete data. *Annu. Rev. Public Health*, *25*, 99–117.

Rana, T., Bera, A. K., Das, S., Bhattacharya, D., Bandyopadhyay, S., Pan, D., & Das, S. K. (2010). Effect of chronic intake of arsenic-contaminated water on blood oxidative stress indices in cattle in an arsenic-affected zone. *Ecotoxicology and environmental safety*, *73*(6), 1327–1332.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation & the health professions*, *9*(4), 395–420.

Regoeczi, W. C., & Riedel, M. (2003). The application of missing data estimation models to the problem of unknown victim/offender relationships in homicide cases. *Journal of Quantitative Criminology*, *19*(2), 155–183.

Rja, L., & Rubin, D. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons.

Roberts, G., Rosenthal, J., et al. (1997). Geometric ergodicity and hybrid markov chains. *Electronic Communications in Probability*, *2*, 13–25.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, *89*(427), 846–866.

Robins, L., Cottler, L., Bucholz, K., & Compton, W. (1995). *Diagnostic interview schedule for dsm-iv (dis-iv)*.

Rotnitzky, A., Robins, J. M., & Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association*, *93*(444), 1321–1339.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592.

Rubin, D. B. (1987a). The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, *82*(398), 543–546.

Rubin, D. B. (1987b). Comment. *Journal of the American Statistical Association*, *82*(398), 543–546.

Rubin, D. B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 1213–1234.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, *91*(434), 473–489.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of internal medicine*, *127*(8_Part_2), 757–763.

Rubin, J. S., Bottaro, D. P., Chedid, M., Miki, T., Ron, D., Cheon, H.-G., … others (1995). Keratinocyte growth factor. *Cell biology international*, *19*(5), 399–411.

Rudas, T. (2005). Mixture models of missing data. *Quality & quantity*, *39*(1), 19–36.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. CRC press.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical methods in medical research*, *8*(1), 3–15.

Schafer, J. L. (2003). Multiple imputation in multivariate problems when the imputation and analysis models differ. *Statistica Neerlandica*, *57*(1), 19–35.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147.

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate behavioral research*, *33*(4), 545–571.

Schafer, K., Yang, B., DiMauro, L., & Kulander, K. (1993). Above threshold ionization beyond the high harmonic cutoff. *Physical review letters*, *70*(11), 1599.

Schervish, M. J., & Carlin, B. P. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical statistics*, *1*(2), 111–127.

Schluchter, M. D., & Jackson, K. L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association*, *84*(405), 42–52.

Seaman, S., & White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics-Theory and Methods*, *43*(16), 3499–3515.

Selvin, D. F. (1996). *A terrible anger: The 1934 waterfront and general strikes in san francisco*. Wayne State University Press.

Siddiqui, O., & Ali, M. W. (1998). A comparison of the random-effects pattern mixture model with last-observation-carried-forward (locf) analysis in longitudinal clinical trials with dropouts. *Journal of biopharmaceutical statistics*, *8*(4), 545–563.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological methods*, *6*(4), 317.

Smith, A. F., & Roberts, G. O. (1993). Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–23.

Song, Q., & Shepperd, M. (2007). A new imputation method for small software project data sets. *Journal of Systems and Software*, *80*(1), 51–62.

Stumpf, S. A. (1978). A note on handling missing data. *Journal of Management*, *4*(1), 65–73.

Swendsen, R. H., & Wang, J.-S. (1987). Nonuniversal critical dynamics in monte carlo simulations. *Physical review letters*, *58*(2), 86.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). Using multivariate statistics.

Tan, M. T., Tian, G.-L., & Ng, K. W. (2009). *Bayesian missing data problems: Em, data augmentation and noniterative computation*. CRC Press.

Tanner, C. A., Benner, P., Chesla, C., & Gordon, D. R. (1993). The phenomenology of knowing the patient. *Journal of Nursing Scholarship*, *25*(4), 273–280.

Tanner, M. A., & Wong, W. H. (1987a). An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, *29*(1), 23–32.

Tanner, M. A., & Wong, W. H. (1987b). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, *82*(398), 528–540.

Tanner, M. A., & Wong, W. H. (2010). From em to data augmentation: the emergence of mcmc bayesian computation in the 1980s. *Statistical science*, 506–516.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, *24*(1), 53–62.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*(1), 60–62.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 95–103.

Zhang, J., & Kai, F. Y. (1998). What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama*, *280*(19), 1690–1691.

Zhang, W.-x. (2003). Nanoscale iron particles for environmental remediation: an overview. *Journal of nanoparticle Research*, *5*(3), 323–332.

Zhao, L., & Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in medicine*, *11*(6), 769–782.

(Meng & Rubin, 1991; M. A. Tanner & Wong, 1987b)