5-2018

Performance Comparison of Imputation Algorithms on Missing at Random Data

Evans Dapaa Addo
East Tennessee State University

Performance Comparison of Imputation Algorithms on Missing at Random Data

---

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Evans Dapaa Addo

May 2018

Nicole Lewis, Ph.D., Chair

Robert Price, Ph.D.

JeanMarie Hendrickson, Ph.D.

Keywords: Missing data, Multiple imputation by chain equation, Multiple imputation, predictive mean matching, Bayesian linear regression, linear regression, non Bayesian.

ABSTRACT

Performance Comparison of Imputation Algorithms on Missing at Random

Data

by

Evans Dapaa Addo

Missing data continues to be an issue in any field that deals with data due to the fact that almost all the widely accepted and standard statistical methods assume complete data for all variables included in the analysis. Hence, in most studies statistical power is weakened and parameter estimates are biased, leading to weak conclusions and generalizations.

Many studies have established that multiple imputation methods are effective ways of handling missing data. This paper examines three different imputation methods (predictive mean matching; Bayesian linear regression; linear regression, non Bayesian) in the $MICE$ package in the statistical software, R, to ascertain which of the three methods imputes data that yields parameter estimates closest to the parameter estimates of a complete data given different percentages of missingness. The paper extends the analysis by generating a pseudo data of the original data to establish how the imputation methods perform under varying conditions.

DEDICATION

To Mr. Kwabena Nkansah Darfor, Mrs. Stella Nkansah Darfor, Animounyam Akosua Nkansah Darfor, N'dom Nkansah Darfor and Nyametease Nkansah Darfor

ACKNOWLEDGMENTS

I am most grateful to my advisor Dr. Nicole Lewis for giving me honest reviews and for being there for me through thick and thin. I would like to express my special appreciation and thanks for encouraging me when I thought I could not make it. Her encouragement, direction, supervision and guidance have made the completion of this study possible. She saw the potential in me and encouraged me to pursue this study.

I am very appreciative to my committee members, Dr. Robert Price and Dr. JeanMarie Hendrickson for serving as my committee members even in difficulty. I want to thank you for your brilliant suggestion and comments.

I would also like to extend a special thanks to my parents, Mr. and Mrs. Addo Dapaa, for the support and sacrifices that they have made on my behalf.

I would like to express my gratitude to George Affadu-Danful, Obed Koomson, Augustine Oppong, Theophilus Neeaquaye, Isaac Nwi-mozu, Nicholas Carney and Evelyn Fokuoh for their support and encouragement. My appreciation also goes out to Dr. Arnold Nyarambi, Mrs. Dumisa Nyarambi, Mr. Abby-Nana Boadi and Mrs. Juliana Bonsu for their advice, support and encouragement throughout my master's program. To Gloria Baffour, thank you for your inspiration and always cheering me up.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Missing data has been a serious problem in the field of statistics and other related fields for a very long time. Analyzing a data set with missing data in the same manner compared to a data set that is complete can lead to reduced power and biased results, which could potentially lead to incorrect conclusions and weak generalization. For example, assume you are conducting a research on economic growth for a certain period of time for a specific country. More specifically, you are focused on economic indicators like gross domestic product (GDP), inflation, and population but your data is incomplete. Analyzing the incomplete data set means working with a reduced sample size, which reduces the statistical power of the results hence producing biased parameter estimates. Using the results obtained can lead to drawing false conclusions and giving inaccurate recommendations, thus, affecting policy making.

Because of these problems, it is important to think about the reason for the missingness and their impact on the analysis. Rubin provided three types of missingness mechanisms [1]. They are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

Although there are some traditional methods such as listwise deletion and arithmetic mean imputation, that are widely use in analyzing missing data, those methods does not provide accurate results. Some problems associated with these methods include reduction of sample size, hence reducing statistical power and leading to biased parameter estimates. Also, the presence of outliers can make the imputed values biased toward the outliers, hence leading to biased parameter estimates. Researchers recommend methods that compute the missing data multiple times. One method is multiple imputation. Multiple imputation methods compute multiple values to fill-in each missing value. Then each imputed data set is analyzed and the results are pooled together.

The advantages of multiple imputation methods over single imputation methods like listwise deletion and arithmetic mean imputation are (1) preserving sample size and statistical power; (2) results in unbiased estimates; and (3) may be used with standard statistical software that are user friendly.

The $MICE$ package in R is designed to impute missing data using the multiple imputation method. Some of the functions in the $MICE$ package for imputing missing data multiple times are the predictive mean matching (PMM), Bayesian linear regression (norm) and linear regression, non Bayesian (norm.nob).

This study will help researchers in the field of statistics and other related fields such as economics to understand the use of MICE to impute missing data values. Also, this study will guide researchers to known which MICE procedure in R is appropriate for a given percentage of missing data. Lastly, the study will serve as a student's contribution to already existing literature on missing data analysis.

1.1 Proposed Work

The general purpose of this study is to examine the different multiple imputation by chain equation (MICE) procedures in the *R* package, MICE, for imputing data for different percentage of missing data.

Specifically, this study aims to find out which univariate method of imputation is better for a certain percentage of missing data; examine and compare data analysis results of the original data and the imputed data; simulate the data to see if we get similar conclusions as the original data when compared to the imputed data.

The basic assumption of this study is that the missing data are missing at random. That is, the data are missing due to observed variables. The assumption of the data missing at random implies that the missing data can be imputed by the observed data.

However, it should be emphasized that, this assumption might not be true for every data and in true situation. Missing data could be due to factors outside the context of the data set. Some of these factors include personal reasons, cultural believes, religion. Therefore, analyzing such data the same as data missing at random may produce misleading results. Also, this study focuses on analyzing data that have quantitative variables. Hence, the analysis in this study may not be applicable to studies involving qualitative variables or models with both types of variables.

1.2 Overview of the Thesis

The thesis is arranged as follows. Chapter 2 describes the missing data handling techniques. It explains the mechanisms of missing data and the problems associated with analyzing data with missing information. Chapter 3 presents and explains some of the widely recognized methods of handling missing data. This chapter expounds on some of the disadvantages of using the traditional methods of analyzing missing data. Chapter 4 introduces the efficient methods of handling missing data, including the advantages and disadvantages. Chapter 5 discusses in detail the research methods followed in this study. Chapter 6 provides and illustrates the results of the analyses. Chapter 7 introduces the simulation study. Section 7.1 describes the technique used in simulating the new data, and Section 7.2 provides the analyses and results of the simulated data. Chapter 8 discusses the imputation methods in reference to the analysis made and results in Chapters 6 and 7. Chapter 9 concludes the thesis.

2 MISSING DATA HANDLING TECHNIQUES

Missing data is the situation where by there is partial response or no response for one or more variables in a given data set. Missing data are values that are lost and that the availability of these values would have made the result of the analysis more meaningful [2]. Allison (2009) defines missing data as “data that are missing for some (but not all) variables and for some (but not all) cases” [3]. He noted that, if all the data of a particular variable are missing, that variable is known as a latent or unobserved variable. However, in a situation where all the data on a given observation is missing for all variables, we have what is known as unit non-response. Table 2.1 shows the various forms of missing data as described by Allison (2009). In Table 2.1, X_1, X_2, X_3 and X_4 are the variables contain in the data set and x_{11}, x_{12}, x_{13}, x_{21}, x_{23}, x_{32}, x_{41}, x_{42} are the observed data while x_{14}, x_{22}, x_{24}, x_{31}, x_{33}, x_{34}, x_{43}, x_{44}, x_{51}, x_{52}, x_{53}, x_{54} are the missing/unobserved data. Here X_4 is known as the latent variable and observation 5 is the unit non-response variable.

Table 2.1: An illustration of a missing data set.

Observation	X_1	X_2	X_3	X_4
1	x_{11}	x_{12}	x_{13}	
2	x_{21}		x_{23}	
3		x_{32}		
4	x_{41}	x_{42}		
5				

2.1 Missing Data Mechanisms

There are three mechanisms for the missingness of a data: They are missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) [1, 4]. In almost all statistical analysis, two main groups of data are encountered: observed and unobserved data.

2.1.1 Conceptual Overview of Missing Data Mechanism

Data is missing completely at random if the missingness is not due to the observed data or the unobserved data. The MCAR mechanism occurs when the probability of the missing data is independent of the observed and the unobserved data [24]. In other words, MCAR is the mechanism where the probability of missingness is unrelated to observed data and unobserved data [6]. For example, a person leaves a study because he or she relocates or an individual in a study passes away before the study is completed. Using data which is MCAR yields unbiased parameter estimates but leads to loss of statistical power.

Data is missing at random if the missingness is due to observed data but not unobserved data [1, 7]. That is, the likelihood of missingness is associated with observed data but not unobserved data or the missingness of the data is conditional on an observed variable. Since MAR is conditional on the observed variable, MAR should be called conditionally missing at random [8]. An example is when the income variable is not answered but can be predicted by the sector, status, and/or education level variables. MAR is not a serious problem because there are meaningful ways to analyze the data to obtain relatively unbiased parameter estimates.

If the probability of missingness depends on unobserved data but not observed data, the data is missing not a random (MNAR). Data is missing not a random (MNAR) when the probability of missingness is related to the missing values itself [6], such as, when big firms are more inclined not to reveal their marketing strategies. Unlike MCAR and MAR, MNAR may produces biased parameter estimates, however, the bias may be small.

2.1.2 Rubin's (1976) Theoretical Overview of Missing Data Mechanisms

Rubin (1976) defines the missingness of data in terms of a probability model. To understand Rubin's (1976) missing data mechanisms, some basic notation must be defined. Let us denote the data set by Y. The observed portion of Y is denoted by Y_{obs} and the missing/unobserved portion of Y be denoted by Y_{mis}. Also, Rubin (1976) defines a binary variable R to be an n × p matrix of indicator variables whose elements denotes whether data on a particular variable is observed or missing. That is, $R = 1$ if the data is observed and $R = 0$ if the data is missing.

Rubin (1976) viewed every case as having a pair of observations on each variable. The first portion is the data may be observed (Y_{obs}) or may be missing (Y_{mis}). The second portion is a corresponding code on the missing data indicator. Table 2.2 shows an example of a missing data set including the corresponding missing data indicator. One can see that x_{11} and x_{41} are the observed observations, and their corresponding missing data indicator value is 1 while x_{21}, x_{31}, x_{51} are the unobserved observations with their corresponding missing data indicator value is 0.

Table 2.2: An illustration of a missing data set including the missing data indicator.

Observation	X_1	Indicator
1	x_{11}	1
2		0
3		0
4	x_{41}	1
5		0

Rubin defines the probability model for the MNAR mechanism as

$$P(R|Y, \xi) = P(R|Y_{obs}, Y_{mis}, \xi) \tag{2.1}$$

where ξ is a unknown parameter describing the relationship between R and the data [24]. Equation 2.1 shows that the probability of R being either 1 or zero depends on $Y = (Y_{obs}, Y_{mis})$. That is, the probability of the missingness on Y depends on other variables in the data set (Y_{obs}) as well as other underlying variables of Y itself (Y_{mis}).

Also, Rubin defined the MAR mechanism as the probability of missingness on Y depending on other variables in the analysis model (Y_{obs}) but not on the underlying variables of Y itself [1, 24]. Consequently, Rubin defines the probability distribution of MAR as

$$P(R|Y, \xi) = P(R|Y_{obs}, \xi). \tag{2.2}$$

Equation 2.2 shows that the probability of missingness on Y is related to the observed portion of data set via some parameter ξ that relates (Y_{obs}) to R.

Finally, Rubin defined the MCAR mechanism as the probability of missingness on Y that is not related to the observed variables or on the underlying variables of Y itself. For the MCAR mechanism, the missingness is completely unrelated to the data [24]. Rubin defines the probability distribution as

$$P(R|\xi). \tag{2.3}$$

Equation 2.3 depicts that both (Y_{obs}) and (Y_{mis}) are unrelated to R.

2.2 The Problem of Missing Data

The problem of missing data has always existed [8]. Although the problem of missing data may arise due to item non-response, it also may be as a result of the design of the study. [9]. Graham argues that the challenges of missing data is mostly minimal for longitudinal research [8]. The major problem of missing data is the outdated statistical procedures used in studies that have missing data.

The problem of missing data may not only be as a result of the statistical procedures used in analyzing missing data. The problem of missing data may arise due to the statistical software that are used in analyzing missing data [3]. These archaic statistical procedures and software presumes that there is a complete response for all variables and for all cases. The default method is to delete any case with a missing response on the variable of interest, thus reducing the sample size. This is commonly known as listwise deletion or complete case analysis.

3 TRADITIONAL METHODS OF HANDLING MISSING DATA

With the understanding that missing data cannot be analyzed the same way as complete data, there was the need to develop methodologies to help overcome the problem of missing data. For years, researchers in the field of statistics have proposed and employed several techniques to solve the problem of missing data. Although some of these techniques have been widely accepted as a way of overcoming the problem of missing data, it has been seen that most of these techniques do not solve the problem of missing data entirely. The method of deleting missing data, which for a long time has been widely accepted and included in many statistical packages, is one of the worst methods for handling missing data. [10, 11]

There have been several traditional methods incorporated to handle missing data, which include but is not limited to listwise deletion, pairwise deletion, arithmetic mean imputation, regression imputation, and hot deck imputation.

Listwise deletion (also known as complete cases analysis) is when the entire data on a subject is removed if a value is missing for at least one variable. The method of listwise deletion demands that cases with missing data are deleted before any statistical analysis is done. Listwise deletion leads to a complete reduction of the sample size. The reduction of the sample size leads to the problem of loss of statistical power along with biased parameter estimates if more subjects with a particular characteristic are deleted. It should be noted that the method of listwise deletion is only applicable when the missingness is MAR. However, some researchers believe that, listwise deletion is good to use if the cases removed are small. Graham (2009) argues that listwise deletion is a desirable option if the cases removed are less than 5% [8]. He contends that, with 5% cases removed, loss of statistical power and biasness are insignificant. Also, Little (1992) asserts that using listwise deletion under any missing data mechanism can produce unbiased estimates for the regression slope if the probability of missingness is due to an independent variable but not the response variable [12].

With the method of pairwise deletion (also known as available case analysis), a correlation is calculated for any pair of variable using available data for these variables and for each pair, the case of a particular variable with missing data and the same case for the other variable (missing or not) is deleted. This implies that, there is a different sample size for each pair that is analyzed. Pairwise deletion is not limited to correlation matrix but can also be applied in regression and ANOVA analysis [13]. The method of pairwise deletion is better than listwise deletion because not all data for a particular case is deleted. Hence, pairwise deletion is desirable when the sample size is small. Nevertheless, the inconsistency in the sample size makes it difficult in computing the standard error. Although some statistical softwares uses the average sample size per variable method for computing the standard error, the method is likely to overestimate the standard error for some variables and underestimate the standard error for other variables [12]. Moreover, Graham (2009) claims that because different sample sizes are used in the method of pairwise deletion, the method can produce biased parameter estimates [8].

The method of arithmetic mean imputation is where the arithmetic mean is computed for a particular variable and the computed value is used to replace all the cases with missing values for that variable. The advantage of the arithmetic mean imputation is that it produces a complete data set. However, in a data set with outliers, the mean is biased toward the outliers. Hence, using the arithmetic mean can affect the variability of the data and affect the parameter estimates.

The regression imputation method uses predicted values from a regression model to replace the missing data. With this method, the cases with complete data is used to develop the regression equation and the predicted values from the regression equation are used to replace the missing values. Although the regression imputation is better than the arithmetic mean imputation, it still yields biased parameter estimates.

Hot deck imputation is where values drawn from the observed values are used to replace the missing values. Drawing values from the observed data is done with replacement, hence, giving equal chance to the observed datum to be selected to replace the missing values. Hot deck has a major disadvantage that drawing values from the observed data to replace the missing data can underestimate the variability of the completed data, leading to narrow intervals [14].

Other traditional techniques for calculating missing values include last observation carried forward, stochastic regression imputation and similar response pattern imputation. The last observation carried forward is a poor and lazy way of dealing with missing data. The act of assuming that scores do not change after the last observed measurement or during the intermittent period where scores are missing can lead to biased parameter estimates. A major disadvantage of the stochastic regression imputation is that it is complex to use when there are several missing data patterns in a multivariate data because each missing data pattern will require a unique regression equation. For similar response pattern imputation, although computer simulation studies suggest that this technique can produce relatively accurate parameter estimates with MCAR data, it is can produce substantial bias when the data are MAR [24]

4 EFFICIENT METHODS OF HANDLING MISSING DATA

Because of the many short falls for traditional methods of handling missing data, researchers had to develop more effective and efficient ways of handling missing data. Two general methods have been recommended by Buuren and Grothuis-Oudshoorn to handle missing data that is multivariate, which is a data set with more than one response variable [15]. These methods are joint modeling (JM) and multivariate imputation by chained equations (MICE). These methods are used when the missingness of the data is MAR.

4.1 Handling Missing Data using Joint Modeling (JM)

JM is desirable if the data can be described by a multivariate distribution. That is, a probability distribution with more than one random variable. The JM method first specifies a multivariate distribution for the missing data [15]. Then Markov chain Monte Carlo (MCMC) techniques are used to draw imputations from the conditional distribution. A conditional distribution is a probability distribution that a randomly selected element from a subset of a sample space has the one characteristic of interest [16].

4.1.1 Overview of Markov Chain Monte Carlo (MCMC)

Markov chain Monte Carlo is the technique used to draw a pseudorandom sample from a probability distribution using Markov chains [17, 18]. Monte Carlo is the term that relies on the generation of random numbers [18]. Consider the normal distribution. We can generate a series of random numbers from a normal distribution with mean, μ, and some variance, σ^2. The equation for generating a series of random numbers from a normal distributions is

$$\theta_t \sim N(\mu, \sigma^2), \tag{4.1}$$

where θ_t is the randomly selected value at time t. From Equation 4.1, the normal distribution is called the proposal distribution.

A Markov chain is a sequence of random variables in which the distribution of the current individual element depends on the value of the previous element [17]. With Equation 4.1, the equation for generating a sequence of random variables where the distribution of the current individual element depends on the previous element is

$$\theta_t \sim N(\theta_{t-1}, \sigma^2). \tag{4.2}$$

Equation 4.2 shows that each value is drawn from a normal distribution, with mean equal to the previous value randomly selected and some variance, σ^2. In other words, the distribution of the next individual element is conditional on the current element. In MCMC, Markov chains are constructed and the goal is for the sequence to converge to a stationary probability distribution. Through a process of simulating repeatedly, the steps of the Markov chain draws samples from the stationary probability distribution. The two main methods used in MCMC are Metropolis-Hasting algorithm and the Gibbs sampling.

The Metropolis-Hasting algorithm is used to determine which sampled value of θ selected randomly by the Markov chain to accept or discard [18]. The Metropolis-Hasting algorithm begins by calculating the posterior probability using the newly generated value of θ. The posterior probability is also calculated using the previous value of θ. In the Metropolis-Hasting algorithm, one does not need to know the functional form of the posterior distribution. To obtain the posterior distribution, we multiply the prior distribution by the likelihood function. The likelihood distribution is the distribution of the observed data and the prior distribution is our subjective feeling of the behavior of θ. Then the ratio of the posterior probability of the new value (θ_{new}) and the posterior distribution of the previous value generated (θ_{t-1}) is computed and can be defined as

$$\rho(\theta_{new}, \theta_{t-1}) = \frac{\text{posterior probability of } \theta_{new}}{\text{posterior probability of } \theta_{t-1}}$$

$$= \frac{\text{Prior } (\theta_{new}) \times \text{Likelihood } (\theta_{new})}{\text{Prior } (\theta_{t-1}) \times \text{Likelihood } (\theta_{t-1})}$$

$$= \frac{\pi(\theta_{new}) \times f(y \mid \theta_{new})}{\pi(\theta_{t-1}) \times f(y \mid \theta_{t-1})}.$$

If the posterior probability of θ_{new} is greater than the posterior probability of θ_{t-1}, then $\rho(\theta_{new}, \theta_{t-1}) > 1$ and we will always accept the new value of θ. However, if the posterior probability of θ_{new} is less than the posterior probability of θ_{t-1}, $\rho(\theta_{new}, \theta_{t-1}) < 1$, we will not necessary reject the new value of θ.

We treat the ratio of the posterior probability of θ_{new} and the posterior probability of θ_{t-1} which is less than one as an acceptance probability. The acceptance probability is

$$\alpha(\theta_{new}, \theta_{t-1}) = min[\rho(\theta_{new}, \theta_{t-1}), 1].$$

Having the acceptance probability in hand, we draw a random number from a standard uniform distribution, $u \sim uniform(0, 1)$, and keep θ_{new} if the random number from the uniform distribution is less than the acceptance probability. Thus, if $u < \alpha(\theta_{new}, \theta_{t-1})$, then $\theta_t = \theta_{new}$. Otherwise, $\theta_t = \theta_{t-1}$. This process is repeated until the sequence converges. After obtaining an estimate for θ, it is then used to impute values for the missing entries.

There are two main issues that arise with the Metropolis-Hasting algorithm. First, is the dependency of the sequence on the starting values. However, this problem can be reduce by discarding the first part of the sample. The first part of the sample is known as the burn-in period. The burn-in period is the time it takes the sequence to stabilize so that it is drifting up and down overtime [17]. The other problem is autocorrelation. The values of θ are correlated because they are generated by a Markov chain. Excessive autocorrelation may indicate problems with model specification. Nevertheless, if the model is correctly specified, thinning can be use to reduce the influence of autocorrelation [18]. Thinning is the process of increasing the MCMC sample size and drawing samples at regular intervals. For example, instead of generating 1000 samples, we can generate 5000 samples and keep every 5^{th} value to get our sample of 1000.

Gibbs sampling is a special case of the Metropolis-Hasting algorithm. In Gibbs sampling, we draw from the conditional distribution of each subvector given all the other subvectors [17]. Suppose that there is a random vector $Z = (z_1, z_2, ..., z_n)$ and we want to obtain j samples of Z from a joint distribution $P(Z) = P(z_1, z_2, ..., z_n)$, which is also the target distribution to be simulated.

Denote the t^{th} sample by $Z^t = ({z_1}^t, {z_2}^t, ..., {z_n}^t)$ and let Z^t be the initial value. This value is determined randomly or by some process such as the expectation-maximization algorithm. The next sample, which is denoted by $Z^{(t+1)} = ({z_1}^{(t+1)}, {z_2}^{(t+1)}, ..., {z_n}^{(t+1)})$, is obtained by continuously drawing from the distribution as shown below:

$$\begin{aligned}
{Z_1}^{(t+1)} &\sim P(z_1 \mid {z_2}^t, {z_3}^t, ..., {z_n}^t) \\
{Z_2}^{(t+1)} &\sim P(z_2 \mid {z_1}^{(t+1)}, {z_3}^t, ..., {z_n}^t) \\
{Z_3}^{(t+1)} &\sim P(z_3 \mid {z_1}^{(t+1)}, {z_2}^{(t+1)}, {z_4}^t, ..., {z_n}^t) \\
&\vdots \\
{Z_n}^{(t+1)} &\sim P(z_n \mid {z_1}^{(t+1)}, {z_2}^{(t+1)}, ..., z_(n-1)^{(t+1)}).
\end{aligned}$$

Particularly, one draws from the conditional distribution of $Z_1, Z_2, ..., Z_n$, conditioning each time on the current drawn values [17]. This process is repeated to obtain $z^{(t+2)}, z^{(t+3)}, z^{(t+4)}$ and so on until the sequence converges to the stationary distribution which equals to $P(Z)$ [17]. Thus, as $t \rightarrow \infty$, $Z^t \rightarrow Z$, where $t = 1, 2, 3, ...$

Schafer (1997) claims that, using Markov chains for simulation on large data set is time consuming and requires computers with fast memory and large storage capacity [17]. For more details on MCMC, see Robert and Casella, (2002), Schafer, (1997) and Gilks, Richardson and Spiegelhalter, (1998).

4.1.2 Overview of the MLE for Missing Data

The likelihood method is one of the widely used methods in the joint model literature [19]. Through maximum likelihood estimation (MLE), parameter estimates of the joint modelling can be based on the observed-data likelihood [21, 20, 19].

Given a probability density function (PDF) and the observations, the method of finding a parameter (θ) that maximizes the probability of making the observations given the parameters is called MLE. It is a well-known method of estimation in the statistical field.

The method of finding the parameter that maximizes the parameters starts by finding the joint PDF [16], a probability distribution for two or more random variable [22], of each observation present. Then, the joint PDF of each observation is multiplied together to obtain the likelihood function, which is a function of the parameter (θ). Given n independent observations and k variables, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f_i(x_{i1}, x_{i2}, ..., x_{ik}; \theta),$$

where $L(\theta)$ is the likelihood function and $f_i(x_{ij}; \theta)$ is the joint distribution function.

To obtain the parameter that is maximized, we differentiate the likelihood function, equate it to zero and then solve for the parameter θ. Since it is difficult to differentiate the likelihood function, an easy step is to take the natural logarithm of the likelihood function. Since the natural logarithm is an increasing function, it implies that the values of θ that maximize the natural logarithm is the same θ that maximizes the likelihood function [16].

4.1.3 MLE for Missing Data

Suppose for a given case, x_1 and x_2 are missing data that satisfy the assumption of MAR. The joint probability is then obtained for these cases by summing or integrating over the variables that have missing data to obtain the marginal probabilities of the variables having complete data [23]. The joint PDF for a discrete missing data will be

$$f_i{}^*(x_{i1}, x_{i2}, ..., x_{ik}; \theta) = \sum_{x_1} \sum_{x_2} f_i(x_{i1}, x_{i2}, ..., x_{ik}; \theta),$$

and for continuous missing data, the joint PDF is

$$f_i{}^*(x_{i1}, x_{i2}, ..., x_{ik}; \theta) = \int_{x_1} \int_{x_2} f_i(x_{i1}, x_{i2}, ..., x_{ik}; \theta).$$

However, if for a given data set m cases are complete and $n - m$ cases are missing data that satisfy the assumption of MCAR and MAR, then the likelihood function for the full data set is;

$$L(\theta) = \prod_{i=1}^{m} f_i(x_{i1}, x_{i2}, ..., x_{ik}; \theta) \times \prod_{i=m+1}^{n} f_i{}^*(x_{i3}, x_{i4}, ..., x_{ik}; \theta).$$

This likelihood is then used to compute the MLE for θ, which are the unknown parameters for the distribution of the missing data set. The value of θ is then used to impute the missing values. Allison (2002) asserts that the method of MLE is easy when the missing data have a monotonic pattern [23]. That is, when data is missing for a particular variable, the same data is missing for other variables in the data set.

Under a multivariate normal model, the likelihood can be maximized using the expectation-maximization (EM) algorithm, which is widely used because of its availability in a lot of statistical softwares [23]. The EM algorithm is an iterative method for finding maximum likelihood estimates [8]. In the E-step of the iteration process, the expected value and the covariance obtained from the observed data are used to build regression equations that are used to predict the missing values. In the M-step of the iteration process, a standard complete data formula is used on both the filled-in data and the observed values to obtain new estimates of the mean and covariance. The updated estimates of the mean and the covariance are used in another E-step to build the new regression equation to predict new missing values. Subsequently, the newly generated missing values and the observed values are used in another M-step to estimate another mean and covariance. These two steps are repeated until convergence (the parameter estimates remain the same for each iteration) is reached.

To illustrate how the EM algorithm works mathematically, a bivariate analysis example, as used in Enders (2010) is used. Here X represent a complete data set and Y represent an incomplete data set. Using the observed data, the formulas that generates the maximum likelihood estimates for the mean and covaraince is

$$\begin{aligned}
\hat{\mu}_Y &= \frac{\sum Y}{N}, \\
\hat{\sigma}_Y^2 &= \frac{1}{N}(\sum Y^2 - \frac{(\sum Y)^2}{N}), \text{and} \\
\hat{\sigma}_{XY} &= \frac{1}{N}(\sum XY - \frac{(\sum X \sum Y)}{N})
\end{aligned}$$

where N is the number of observed cases.

After obtaining the estimates for the mean and covariance, those estimates are then used to build a regression model using the following formulas:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}, \\
\hat{\beta}_0 &= \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu}_X, \\
\hat{\sigma}_{Y|X} &= \hat{\sigma}_Y^2 - \hat{\beta}_1^2 \hat{\sigma}_X^2, \text{and} \\
\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i.
\end{aligned}$$

The E-step of the iteration process uses the regression equation to fill the missing values. After obtaining the missing values, the M-step uses the formulas for the mean and covariance to calculate a new maximum likelihood estimates for the mean and covaraince. The new mean and covariance are use to build a new regression equation. This process is repeated until convergence.

Under the multivariate-normal model, the mean, variance and the covariances are the parameters that are estimated by the EM algorithm. According to Graham (2009) and Allison (2009), the EM algorithm estimates the appropriate parameters, but one weakness is that it does not provide standard error estimates [8, 3].

4.2 Handling Missing Data using Multiple Imputation by Chain Equation (MICE)

Sometimes it is difficult and time consuming to determine a particular multivariate model whose assumptions are satisfied by your data set. The multiple equation by chain equation method is desirable when it is inappropriate to assume a multivariate distribution using the JM method [15]. The MICE, which is also known as fully condition specification (FCS), uses a set of conditional densities for each variable with missing data to assume a multivariate imputation model on a variable by variable basis [15]. An example of the MICE method is the multiple imputation (MI).

4.2.1 Summary of the Multiple Imputation Method

One of the most recognized techniques for handing missing data is multiple imputation (MI). The first step in MI is to compute the missing values using an appropriate model that includes random variation [25, 26]. An appropriate model for computing missing data that include random variation is the linear regression [35]. Using all observed cases for all variables, the variable with missing data is regressed on the other variables with no missing data to get the predicted values. The random variation which is the product of the root mean squared error from the regression model and a random draw from a standard normal distribution is added to predicted values to get data for the missing entries.

This is done m times to obtain m different complete data sets. After this, an appropriate method is used to analyze each of the m complete data set. Thus, the dependent variable is regressed on the independent variables to obtain m different parameter estimates. An average of the m different parameters is calculated to procure a single parameter estimate(s). To obtain the standard error, the within variance is calculated, The within variance is the variations caused by differences within individual data set, by taking the average of the square standard errors of the m parameters. Then the between variance is calculated, the variation due to the interaction between the different m data sets, of each of the parameter estimates. The standard error is the square root of the sum of the within variance and the between variance.

By introducing a random error term in the model for computing the missing values, it allows the MI parameter to be unbiased [23]. Moreover, adding the random variation preserves the distribution in the filled-in data set thereby making the regression model less dependent on normality [28].

The advantage of the multiple imputation over the single imputation is that, repeating the imputation to obtain m different data sets helps to obtain parameter estimates that are efficient and close to the real parameter estimates [23].

4.2.2 Multiple Imputation by Chain Equation in R

The MICE package in R, a statistical software that can be accessed at https://www.r-project.org/, makes it easy to impute missing values using MI. The MICE package in R assumes a number of univariate imputation techniques of each incomplete variable. The univariate imputation methods take a set of complete independent variables and returns a single imputed value for each missing entry in the incomplete targeted variable [15]. Another property of the MICE package in R is that, it can detect three scales of measurements for each variable [15, 20]. These scales are numerical, binary (factors with 2 levels) and categorical (factors with more than 2 levels). With these properties, the MICE package checks the choice of a univariate imputation method assumed and the scale of measurement of a variable to avoid a mismatch. Table 4.1 presents some of the univariate imputation models, their name in R and the supported scale of measurement for the MICE package in R.

Table 4.1: A list of imputation methods in the MICE *R* package.

Name of Model	**Name in *R***	**Supported Scale Type**
Predictive mean matching	pmm	Numeric
Bayesian linear regression	norm	Numeric
Linear regression, non Bayesian	norm.nob	Numeric
Unconditional mean imputation	mean	Numeric
Two-level linear model	2L.norm	Factor, 2 levels
Logistic regression	logreg	Factor > 2 levels
Multinomial logit model	polyreg	Ordered > 2 levels
Ordered logit model	polr	factor
Linear discriminant analysis	ida	factor
Random sample from observed data	sample	Any
Classification and regression trees	cart	Any
Random forest imputation	rf	Any

The predictive mean matching (PMM) method of imputation is a general purpose semi-parametric imputation method that uses observed values to impute missing values. One advantage of the PMM is that it preserves non-linear relations even when the structural part of the imputation is wrong [15].

To impute missing values using PMM, first, the variables with missing data is regressed on the variables with no missing data to obtain some predicted values. Using a bivariate model, we illustrate this by letting Y denote the variable containing the missing variables and X denote the variables having complete data. The regression model for predicting the missing values is

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i. \tag{4.3}$$

From equation 4.3, $\hat{Y}_i$ is the predicted Y values for a given X values, $\hat{\beta}_0$ is the estimated intercept, and $\hat{\beta}_1$ is the estimated slope.

In regressing the variables with missing data on the observed variables, a random variation is added to the predicted values. This is done to preserve the distribution of the filled in data [28]. The equation that adds a random variation to predicted values is given as

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \delta\mu, \tag{4.4}$$

where δ is the root mean squared error and μ is a random draw from a standard normal distribution.

Equation 4.4 is used to generate values for all cases of the variables with missing data. Then from the values generated, a set of k cases with observed values whose predicted values are close to the predicted value for a case with missing data are identified. The choice of k is based on a trade-off between large enough to simulate the predicted distribution effectively and small enough to maintain quality of the matches [28]. With the MICE package in R, the default is $k = 5$ [29]. From the close cases identified, one is selected at random and assigned its observed value to fill the missing entry for that case [30, 29]. This is repeated until a complete data set is obtained. The steps described above is done m times to produce m complete data sets [28]. In each of the m complete data sets, the dependent variable is regressed on the independent variables to obtain m different parameter estimates. An average of the m different parameters are calculated to yield a single parameter estimate(s) [28]. One advantage of the PMM is that it imputes only eligible values. Because observed values are used to substitute missing values, it avoids imputing values outside the range of the data set. Furthermore, since the predicted mean is only used for matching, PMM is less sensitive to misspecification [28].

There are two main steps in the Bayesian linear regression (norm) approach: EM algorithm and data augmentation. In the EM algorithm step, the mean, variance and covariance of the data are obtained. These estimates are used to estimate the missing values. The estimated values are added to the data set to get a complete data set. Then, we repeat the EM algorithm to obtain a new mean, variance and covariance. The new mean, variance and covariance are used to estimate new values for the missing data. This process is repeated until convergence is reached.

In the data augmentation step, the process uses the population mean, variance and covariance from the EM algorithm (or recreates the population mean, variance and covariance) to estimates missing values. In estimating the missing values, the data augmentation procedure uses the Bayesian approach to create a likely distribution of the parameter values. In Bayesian analysis, the parameter are seen as random variables that have a distribution. The goal of the Bayesian analysis is to describe the behavior of the distribution. Thus, to determine the posterior probability of the parameter values obtained. In the Bayesian paradigm, the prior distribution of the parameter of interest $P(\theta)$ and the likelihood function $f(Y \mid \theta)$ are combined to obtain the posterior distribution $P(\theta \mid Y)$:

$$P(\theta \mid Y) = P(\theta) \times f(Y \mid \theta). \tag{4.5}$$

Then from the posterior distribution of θ, one draws an estimate at random for the mean and covariance. These new estimates for the mean and covariance are used to generate new fill-in values for the missing entries. This process is repeated until convergence. A good number of interactions before convergence in data augmentation process is greater than or equal to the number of iterations it took the EM algorithm to converge [31]. After acquiring a complete data set, the data augmentation process is repeated over and over to generate multiple complete data set. Getting three to five data sets is enough to end the data augmentation process [26].

The linear regression, non Bayesian (norm.nob), imputes missing values using the spread around the fitted linear regression line. First, the variable with missing data is regressed on the variables with no missing data using all the observed data. In a situation whereby all the variables contains missing data, the linear regression, non Bayesian method uses the observed data to regress target incomplete variables on covariate complete variables [15, 20]. In regressing the variable with missing data on the variable with no missing data, the linear regression, non Bayesian, approach uses a parametric linear regression analysis to impute the missing values [15]. Then, the spread around the fitted line is used to predict a value for each missing value.

The disadvantage of the linear regression, non Bayesian, is that, it does not incorporate sample uncertainty. Sample uncertainty is the potential variation in point estimates as a result of the fact that the estimates depends on a sample from the population. The linear regression, non Bayesian, approach is not proper because it does not include variability of the estimates of the regression coefficients, hence underestimating the variability of the imputed values for small samples [32]. However, the linear regression, non Bayesian, is suitable for data that follow a normal distribution with a large sample size where variability is not much of a concern [32].

5 METHODOLOGY

As indicated in chapter 1, the purpose of this study is to examine the different multiple imputation by chain equation (MICE) procedures in the R package, MICE, for imputing data for different percentage of missing data. This chapter provides a description of how the data used in this study is analyze.

5.1 Data Source and Description

The data employed for this study is the Combined Cycle Power Plant data Set from the UCI Machine Learning Repository. This data can be accessed from the link:

http://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant#.

A combined cycle power plant (CCPP), is an electrical power plant, which uses both a gas turbine and a steam turbine to produce more electrical energy from the same fuel than would be possible from a single traditional cycle power plant [33]. It is assumed that the CCPP produces 50 percent more electric energy than a traditional cycle power plant [33]. The CCPP works by using the gas turbine to compress air and mix it with fuel that is heated to a very high temperature. The mixture of hot air and fuel moves through the gas turbine blades, making them spin. The fast-spinning turbine drives a generator that converts a portion of the spinning energy into electricity.

Then, the exhaust heat from the gas turbine is captured by a Heat Recovery Steam Generator (HRSG). The HRSG creates steam from the exhaust heat from the gas turbine and delivers it to a steam turbine. Lastly, the steam turbine sends its energy to the generator drive shaft, where it is converted into additional electricity [36].

Predicting the electric power generated hourly based on the ambient variables enables one to evaluate whether the generated power will be sufficient to meet the growing consumer demands. The entire data set contains 9568 observations collected from a Combined Cycle Power Plant over 6 years, from 2006 to 2011, when the power plant was set to work with full load. The hourly average Ambient Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) are used as the predictor variables to predict the net hourly electrical energy output (EP) of the combined cycle power plant.

A simple random sample was employed on the entire CCPP data set yielding a smaller complete data set with 500 observations. The simple random sample method is used because it is easy to employ and it gives all the observations an equal chance of been selected.

After obtaining the complete data made up of 500 observations, T, AP, RH and V are used as the predictor variables to fit a multiple linear regression model with EP as the response variable. The estimated regression model was found to be

$$\widehat{EP} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 V + \hat{\beta}_3 AP + \hat{\beta}_4 RH. \tag{5.1}$$

Table 5.1 displays the estimated coefficients for each predictor variable. All predictor variables are needed in the model (in the presence of all the variables) except AP, using a 5% level of significance.

Table 5.1: The estimated regression coefficients where (***) indicates the variable is needed in the model at the 5% level of significance.

Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Value	444.4881^{***}	-1.9035^{***}	-0.2337^{***}	0.0681	-0.1246^{***}

Since AP is not significant, it is excluded from the model. Therefore, only T, V and RH are employed as the predictor variables in the CCPP model with EP as the response variable in this study.

5.2 Software Implementation

R is the software used in the analysis of this study. It is the software used to evaluate the various imputation models. Specifically, the prodNA function in *R* was used to randomly delete specified percentage of values in a data set and the *MICE* package in *R* was used to implement the multiple imputation by chain equation approach.

5.3 Analysis of Interest and Imputations

From Section 5.1, we observed that AP was not significant in predicting EP. The final regression model used for this analysis is

$$\widehat{EP} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 V + \hat{\beta}_3 RH. \tag{5.2}$$

Table 5.2 displays the estimated coefficients for each predictor variable. All predictor variables are needed in the model (in the presence of all the variables), using a 5% level of significance. Table 5.2 shows that after the removal of AP from the CCPP model, all the remaining variables are significant in predicting EP. As temperature increases by 1 degree celcius, the estimated EP decreases by 1.96 millwatt (MW), holding all other variables constant. A 1 centimetres of mercury (cmHg) increase in exhaust vaccum (V) will result to a 0.22 MW decrease in EP, holding all other variables constant. Finally, a percent increase in relative humidity (RH), will lead to a 0.134 MW decrease in EP, holding all other variables constant. The adjusted R^2 value is 0.9126, implying that 91.26%, of the variation in EP, is explain by the linear relationship with temperature, exhaust vaccum, and relative humidity, adjusted for the number of variables in the model. With 500 observations, We are assured normality is met due to the central limit theorem.

Table 5.2: The estimated regression coefficients where (***) indicates the variable is needed in the model at the 5% level of significance.

Parameter	β_0	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Actual Values	514.71418^{***}	-1.95560^{***}	-0.22471^{***}	-0.13393^{***}

Table 5.3 contains the variance inflation factors (VIF) values for the CCPP model with T, V and RH variables. From Table 5.3, all the three VIF values are less than 10, which indicates that there is no serious multicollinearity problem in the regression model 5.2. A predicted residual sum of squares (PRESS) value reasonably close to the sum of squares error (SSE) supports the validity of a fitted regression model and indicates the predictive capability of a regression model [37]. The PRESS values of 12682.75 is relatively similar to the SSE value of 12467.55. Hence, the regression model 5.2 does have a good predictive capability.

Table 5.3: Variance inflation factor of the CCPP model 5.2.

Variable	T	V	RH
VIF	5.424620	4.200737	1.640421

The ProdNA function in R was used to introduced certain amount of missingness in the complete data. Using the ProdNA function in R, 5 incomplete data sets with different percentage of missingness (fraction of missing information) were produced with each one having 10%, 20%, 30%, 40% and 50% missingness respectively. It should be noted that a higher percentage of missingness contains the same missingness of the pervious percentage of missingness. Thus, a higher percentage of missingness is build-up on a lower percentage of missingness.

The $MICE$ package in R was adopted to impute the missing values in the 5 incomplete data sets. Table 4.1 displayed various models in the $MICE$ package in R employed to impute missing values, however, due to time constraint, only three of these methods are used to impute the missing values. These three methods are the pmm, norm and norm.nob, which are the methods used for quantitative data. Using the three imputation methods separately, fifty complete data sets were produced for each fraction of missing information.

Applying the CCPP model in Equation 5.2, multiple linear regression analysis is performed on each of the fifty complete data sets imputed for each percentage of missingness using the separate imputation methods. By performing the regression analysis on each of the fifty imputed data sets for each percentage of missingness using the three different imputation models, a sampling distribution of fifty estimated regression coefficients are generated at each percentage of missingness. Now, the estimated parameters are considered as variables and each of the estimated regression coefficient are treated as data points for each of the variables (estimated parameters). We obtain the mean, the variance, the range and the percentage deviation index (PDI) of the estimated regression coefficients and the best model for imputing missing data for a specific percentage of missingness is the model in which the imputed missing values has the smallest variances, range, and percent deviation index (PDI).

PDI is a way of expressing the difference between the original regression coefficient and the mean of the estimated regression coefficients by designating the original regression coefficient as the base. Mathematically, the PDI is expressed as:

$$PDI = \frac{Original\hat{\beta}_i - \mu_{\hat{\beta}_i}}{Original\hat{\beta}_i} \times 100 \tag{5.3}$$

where Original $\hat{\beta}_i$ is the original regression coefficient, $\mu_{\hat{\beta}_i}$ is the mean of the estimated regression coefficients for $i = 0, 1, 2, 3$. Each PDI reflects the percentage difference of the mean of a given estimated regression coefficients and its corresponding original regression coefficients.

Since the size of the sampling distribution is quite large, the variables (estimated parameters) are considered to be normally distributed. Therefore, to determine how significant the difference between the original coefficients of the estimated parameters and the mean of the estimated regression coefficients are, the Student's t test statistic is computed and then used.

Mathematically, the Student's t test statistic is given as:

$$t_{stat_i} = \frac{\mu_{\hat{\beta}_i} - Original\hat{\beta}_i}{\sigma_{\hat{\beta}_i}} \sim t_{n-1}, \tag{5.4}$$

where Original $\hat{\beta}_i$ is the original regression coefficient, $\mu_{\hat{\beta}_i}$ is the mean of the estimated regression coefficients, $\sigma_{\hat{\beta}_i}$ is the estimated standard deviation of the regression coefficients for $i = 0, 1, 2, 3$. Here t_{stat_i} follows a student's t-distribution with $n - 1$ degrees of freedom.

5.4 Relative Efficiency (RE)

Relative efficiency describes the efficiency in the point estimates in estimating the original regression coefficients given the number of imputations and the fraction of missing information (percentage of missingness). The equation for the relative efficiency as given by Rubin (1987) is

$$R.E = \frac{1}{1 + \frac{\lambda}{m}} \tag{5.5}$$

where λ is the fraction of missing information and m is the number of imputation [26].

Using Equation 5.5, Rubin concluded that, with λ less than 20%, $m = 2$ is sufficient to produce point estimates that estimate the original regression coefficients accurately. Also, with λ equal to 50%, $m = 3$ is enough to produce point estimates that estimate the original regression coefficients accurately. "Unless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations" [34].

However, it should be noted that, the number of imputations that is good to produce efficiency in estimating the original regression coefficients may not be necessarily good for estimating the standard error, variance, confidence and P-values [35]. Estimating the variance, standard error and P-values using just 5 observations ($m = 5$) or less may give unstable results. Thus, repeating the whole process of imputation may yield different estimates of the variance, standard error and p-value.

In this study, fifty repeated imputations was used for each of the fraction of missing information (FMI). Fifty repeated imputations were used for each of the five fractions of missing information to aid comparison across the five fractions of missing information and comparison across the three models of imputation. Moreover, fifty repeated imputations are used to give large enough observations that follows a normal distribution by the central limit theorem. Table 5.4 shows that as the number of imputation increases, the relative efficiency increases across the five fractions of missing information.

Table 5.4: Relative Efficiency of the percentage of missingness.

m\FMI	**10%**	**20%**	**30%**	**40%**	**50%**
1	0.9091	0.8333	0.7692	0.7143	0.6667
2	0.9524	0.9091	0.8696	0.8333	0.8
3	0.9677	0.9375	0.9091	0.8824	0.8571
4	0.9756	0.9524	0.9302	0.9091	0.8889
5	0.9804	0.9615	0.9434	0.9259	0.9091
10	0.9901	0.9804	0.9709	0.9615	0.9524
15	0.9934	0.9868	0.9804	0.9740	0.9677
20	0.9950	0.9901	0.9852	0.9804	0.9756
30	0.9967	0.9934	0.9901	0.9868	0.9836
40	0.9975	0.9950	0.9926	0.9901	0.9877
50	0.9980	0.9960	0.9940	0.9921	0.9901

6 RESULTS

Here we compare how the three methods of imputation (predictive mean matching, Bayesian linear regression and linear regression, non Bayesian) accurately impute missing values for a certain fraction of missing information/percentage of missingness.

6.1 Estimated Mean and Variance

The results from Tables 6.2, 6.4 and 6.6 shows the estimated mean of the regression coefficient of $\hat{\beta}_0$, using the imputed data, increases as the percentage of missingness increases from 10% to 30%, then it reduces from 30% to 50%. This observation is true for all the three methods of imputations. For the predictive mean matching method, the estimated mean of the regression coefficient, $\hat{\beta}_1$, using the imputed data, decreases as the percentage of missingness increases. However, using the Bayesian linear regression and linear regression, non Bayesian methods, the estimated mean of the regression coefficient, $\hat{\beta}_1$, using the imputed data, decreases as the percentage of missingness increases from 10% to 40%, then it increases as the percentage of missingness increases to 50%.

For the predictive mean matching method, the estimated mean of the regression coefficient, $\hat{\beta}_2$, using the imputed data, increases as the percentage of missingness increases. However, when applying the Bayesian linear regression and linear regression, non Bayesian methods, the estimated mean of the regression coefficient, $\hat{\beta}_2$, using the imputed data, increases as the percentage of missingness increases from 10% to 40%, then it decreases as the percentage of missingness increases to 50%. Using the Predictive mean matching, Bayesian linear regression and linear regression, non Bayesian methods, the estimated mean of the regression coefficient, $\hat{\beta}_3$, using the imputed data, decreases as the percentage of missingness increases from 10% to 30%, then it increases for 40%. Then it decreases at 50% for the predictive mean matching method, but increases for the Bayesian linear regression and linear regression, non Bayesian at 50%. From Tables 6.3, 6.5 and 6.7, the variance of all the regression coefficients for the predictive mean matching, Bayesian linear regression and linear regression, non Bayesian methods increases as the percentage of missingness increases.

Table 6.1: Estimated mean of the regression coefficients with the predictive mean matching method.

FMI\ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	514.0866	-1.9925	-0.2049	-0.1293
20%	514.5529	-2.0539	-0.1726	-0.1420
30%	515.7961	-2.1176	-0.1498	-0.1567
40%	514.3423	-2.1422	-0.1360	-0.1421
50%	513.5434	-2.1549	-0.1095	-0.1454
Actual Parameter	514.7142	-1.9556	-0.2247	-0.1339

Table 6.2: Estimated Variance of the regression coefficients with the predictive mean matching method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.7297	0.0008	0.0003	0.0001
20%	1.6963	0.0028	0.0007	0.0002
30%	2.5464	0.0037	0.0010	0.0003
40%	4.2519	0.0070	0.0022	0.0005
50%	7.1126	0.0106	0.0029	0.0009

Table 6.3: Estimated mean of the regression coefficients with the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	513.7087	-1.9743	-0.2088	-0.1264
20%	513.8615	-2.0484	-0.1670	-0.1373
30%	514.5835	-2.1151	-0.1396	-0.1478
40%	513.3098	-2.1580	-0.1186	-0.1354
50%	512.5321	-2.1193	-0.1227	-0.1320
Actual Parameter	514.7142	-1.9556	-0.2247	-0.1339

Table 6.4: Estimated Variance of the regression coefficients with the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.6958	0.0014	0.0004	0.00009
20%	1.4403	0.0024	0.0006	0.0002
30%	2.8297	0.0030	0.0008	0.0003
40%	5.0919	0.0076	0.0028	0.0006
50%	8.8064	0.0139	0.0042	0.0014

Table 6.5: Estimated mean of the regression coefficients with the linear regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	513.9165	-1.9715	-0.2118	-0.1280
20%	514.0613	-2.0335	-0.1767	-0.1366
30%	514.1945	-2.1133	-0.1352	-0.1474
40%	513.5283	-2.1730	-0.1094	-0.1419
50%	512.2221	-2.1185	-0.1192	-0.1301
Actual Parameter	514.7142	-1.9556	-0.2247	-0.1339

Table 6.6: Estimated Variance of the regression coefficients with the linear regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.6676	0.0013	0.0004	0.0001
20%	1.3870	0.0024	0.0007	0.0002
30%	2.1496	0.0034	0.0012	0.0003
40%	2.8960	0.0048	0.0015	0.0004
50%	5.9650	0.0109	0.0029	0.0011

In summary, using the predictive mean matching, the Bayesian linear regression and the linear regression, non Bayesian to impute missing data, the mean of the estimated regression coefficients tends to increase for data with large amount of imputed values. Moreover, the variances of these regression coefficients increases as the amount of imputed data increases.

6.2 Range and Percentage Deviation Index

Tables 6.8 and 6.10 show that when using the predictive mean matching and Bayesian linear regression methods to impute missing values, the range increases for all the parameters as the fraction of missing information increases. Using the linear regression, non Bayesian method, the range increases as the percentage of missingness increases for estimated parameters $\hat{\beta}_1$ and $\hat{\beta}_2$, but the range increases as the fraction of missing information increases from 10% to 30%, then reduces fairly at 40% missingness and increases again at 50% missingness for estimated parameter $\hat{\beta}_0$ and $\hat{\beta}_3$ as shown in Table 6.12. Comparing Tables 6.8, 6.10 and 6.12, the linear regression, non Bayesian method produces the smallest range for all estimated parameters.

Comparing Tables 6.9, 6.11 and 6.13, the predictive mean matching method of imputing missing values produces the smallest overall PDI of 4.3553%. The PDI for the linear regression, non Bayesian and Bayesian linear regression methods are 6.1670% and 6.7523% respectively. This suggests that juxtaposing the three imputation methods, the predictive mean matching method imputed data with relatively small variation from the original data. As shown in the tables, $\hat{\beta}_1$ produces the smallest overall PDI, $\hat{\beta}_2$ produces the largest overall PDI for the three methods of imputing missing data. Moreover, the PDI varies by the fraction of missing information under each of the three imputation models. Table 6.9 shows that, using the predictive mean matching methods, the smallest PDI of 1.9671% is for 30% imputed values, and the largest PDI of 8.1767% is for 50% imputed values. Applying the Bayesian linear regression method, Table 6.11 shows that the smallest PDI of 2.9874% is for 10% imputed values, and the largest PDI of 11.5672% is for 40% imputed values. Applying the linear regression, non Bayesian method, Table 6.13 shows that the smallest PDI of 2.3775% is for 10% imputed values, and the largest PDI of 10.4949% is for 50% imputed values. It follows that the variation between the original data and the imputed data tends to be smallest with small amount of imputed data and tends to be large for large amount of imputed data. This is evident across the three imputation methods.

Table 6.7: Range of the regression coefficients with the predictive mean matching method.

FMI\Estimated Parameter	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$
10%	3.7656	0.1492	0.0885	0.0491
20%	5.63	0.2591	0.1099	0.0621
30%	6.6080	0.3234	0.1567	0.0898
40%	9.8568	0.3389	0.1921	0.0973
50%	12.5502	0.5313	0.2247	0.1343

Table 6.8: PDI of the regression coefficients with the predictive mean matching method.

FMI\Estimated Parameter	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$	**Mean**
10%	0.12%	-1.89%	8.79%	3.42%	2.61%
20%	0.03%	-5.03%	23.21%	-6.03%	3.05%
30%	-0.21%	-8.29%	33.33%	-16.97%	1.97%
40%	0.07%	-9.54%	39.49%	-6.13%	5.97%
50%	0.23%	-10.19%	51.26%	-8.59%	8.18%
Mean	0.05%	-6.99%	31.22%	-6.86%	4.36%

Table 6.9: Range of the regression coefficients with the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.7278	0.1744	0.0800	0.0427
20%	5.4089	0.2542	0.1107	0.0619
30%	7.5408	0.2527	0.1198	0.0753
40%	10.6018	0.3635	0.2293	0.1325
50%	14.3545	0.5174	0.2526	0.2064

Table 6.10: PDI of the regression coefficients with the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	**Mean**
10%	0.20%	-0.95%	7.08%	5.63%	2.99%
20%	0.17%	-4.75%	25.68%	-2.52%	4.64%
30%	0.03%	-8.16%	37.87%	-10.35%	4.85%
40%	0.27%	-0.08%	47.21%	-1.13%	11.57%
50%	0.42%	-8.37%	45.40%	-1.40%	9.72%
Mean	0.22%	-4.46%	32.65%	-1.39%	6.75%

Table 6.11: Range of the regression coefficients with the linear regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.7213	0.1607	0.0999	0.0425
20%	4.6781	0.2296	0.1097	0.0576
30%	7.8379	0.2612	0.1691	0.0915
40%	7.4003	0.3588	0.1796	0.0852
50%	9.6888	0.4438	0.2228	0.1630

Table 6.12: PDI of the regression coefficients with the linear regression regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	**Mean**
10%	0.16%	-0.81%	5.74%	4.42%	2.38%
20%	0.13%	-3.98%	21.39%	-1.99%	3.89%
30%	-0.10%	-8.07%	39.82%	-10.02%	5.46%
40%	0.23%	-11.12%	51.30%	-5.93%	8.62%
50%	0.48%	-8.33%	46.93%	2.89%	10.49%
Mean	0.22%	-6.46%	33.04%	-2.13%	6.17%

6.3 Test for Normality of the Parameter Estimates

Since the number of observations for each of the parameter estimates is equal to fifty, by the central limit theorem (CLT), each sampling distribution of the parameter estimates follow an approximate normal distribution. Furthermore, Q-Q plots are produced to provide visual support of normality. Figures 6.1-6.15 show in most cases the points fall on a straight line indicating the assumption of normality is satisfied. While there are a few plots that show some curvature, we are assured normality is met due to the CLT since we have a large sample size, 50.

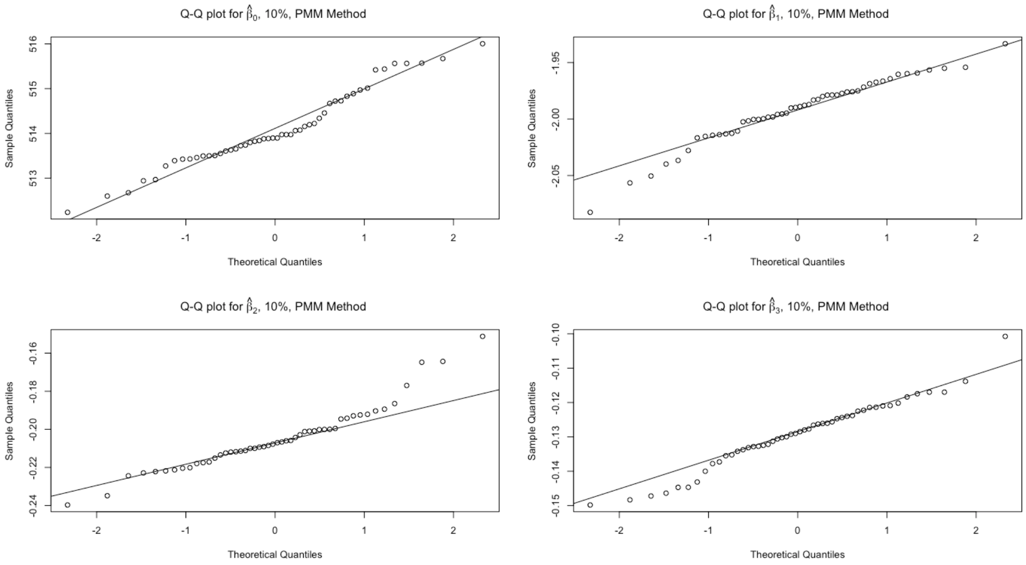


Figure 6.1: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 10% missingness.

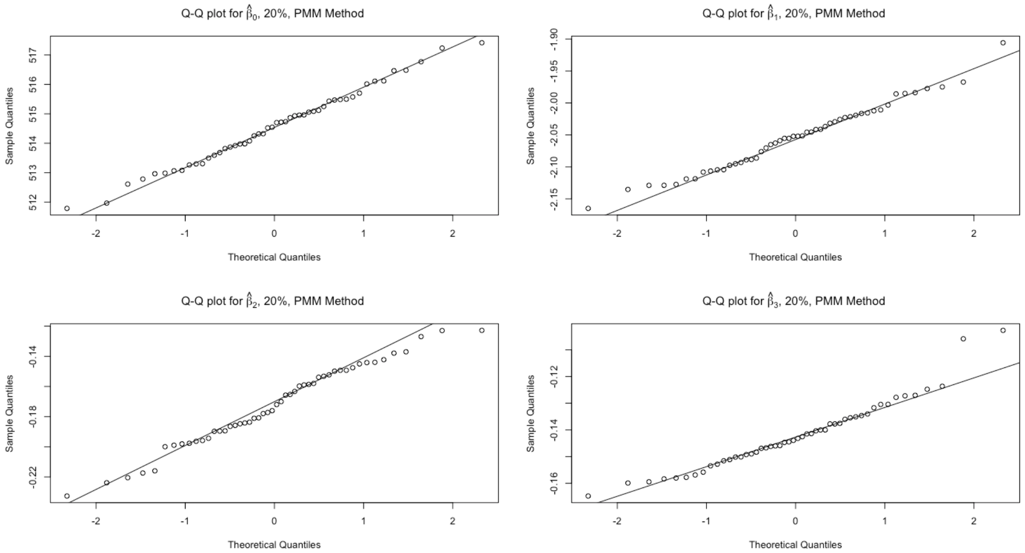


Figure 6.2: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 20% missingness.

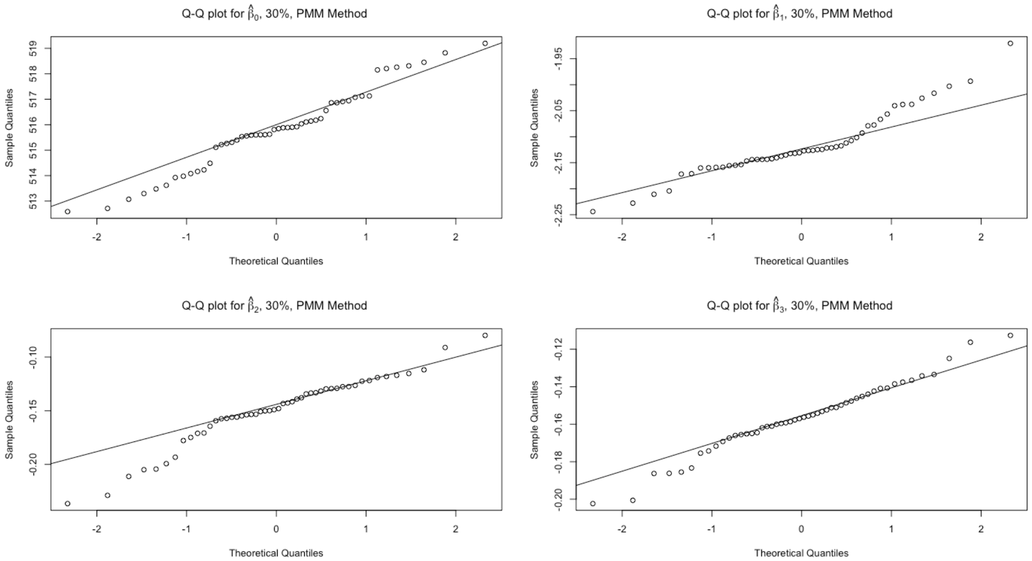


Figure 6.3: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 30% missingness.

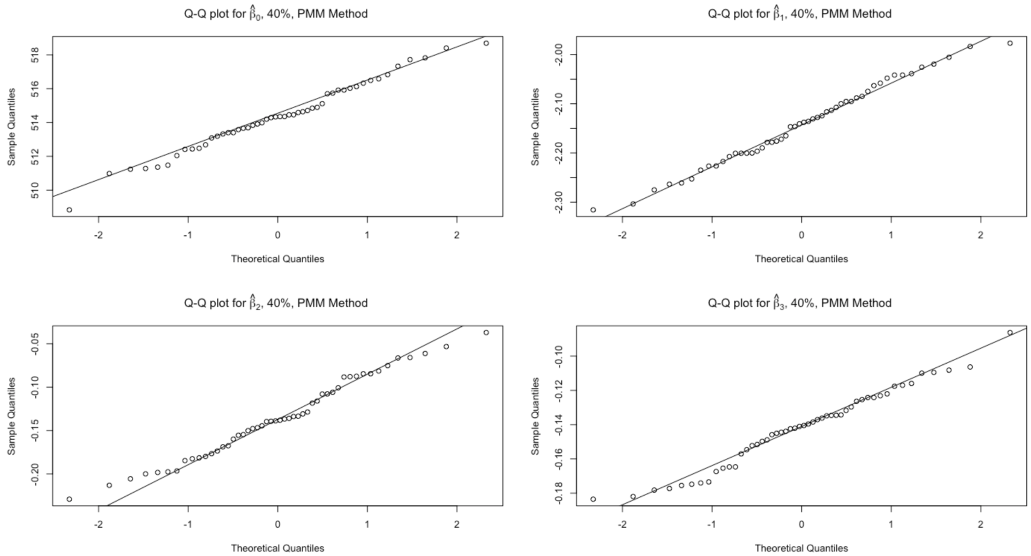


Figure 6.4: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 40% missingness.

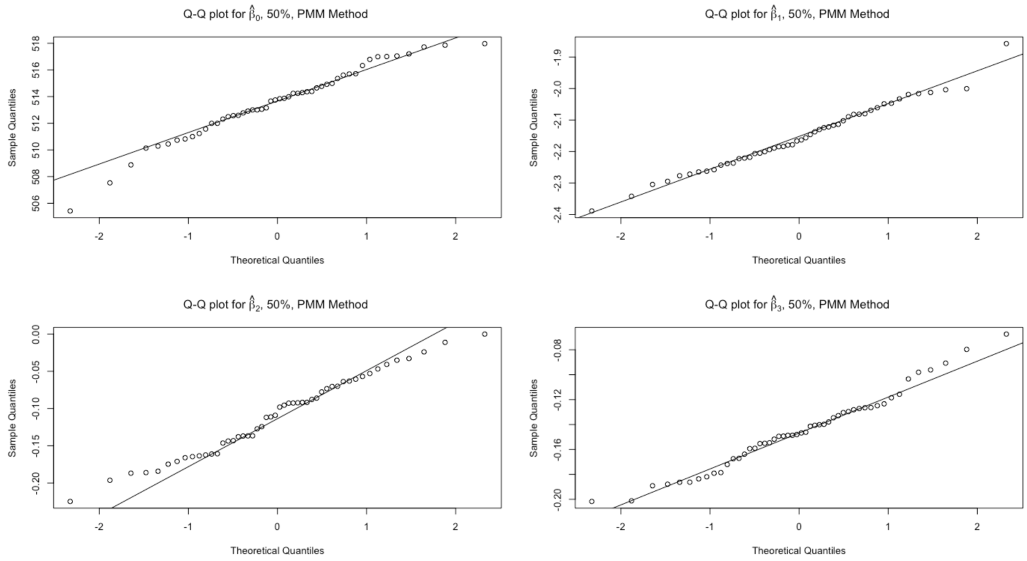
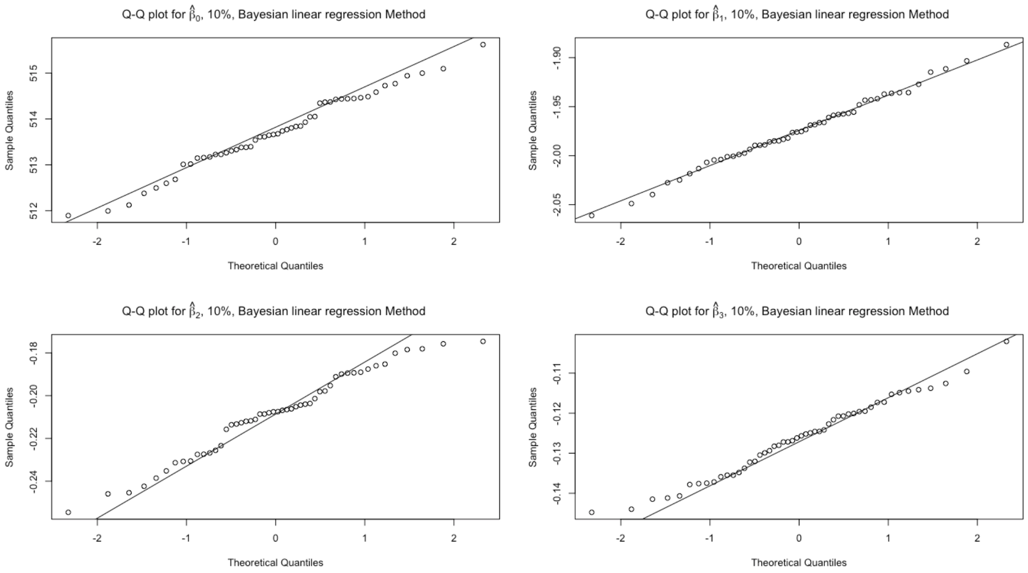


Figure 6.5: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 50% missingness.

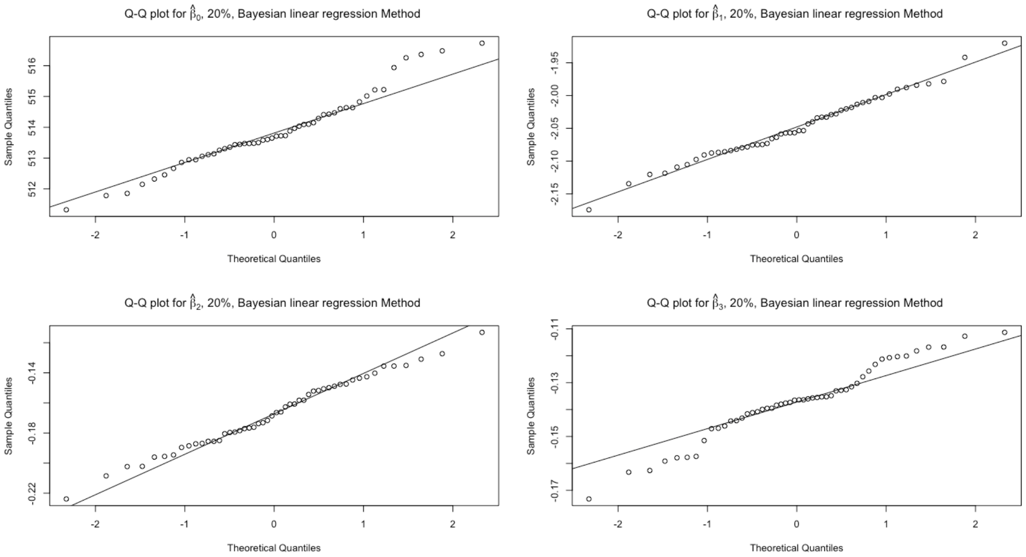

Q-Q plot for $\hat{\beta}_0$, 10%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_1$, 10%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_2$, 10%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_3$, 10%, Bayesian linear regression Method

Figure 6.6: Normality plots of the sampling distributions of the regression coefficients estimated using the Bayesian linear regression method at 10% missingness.



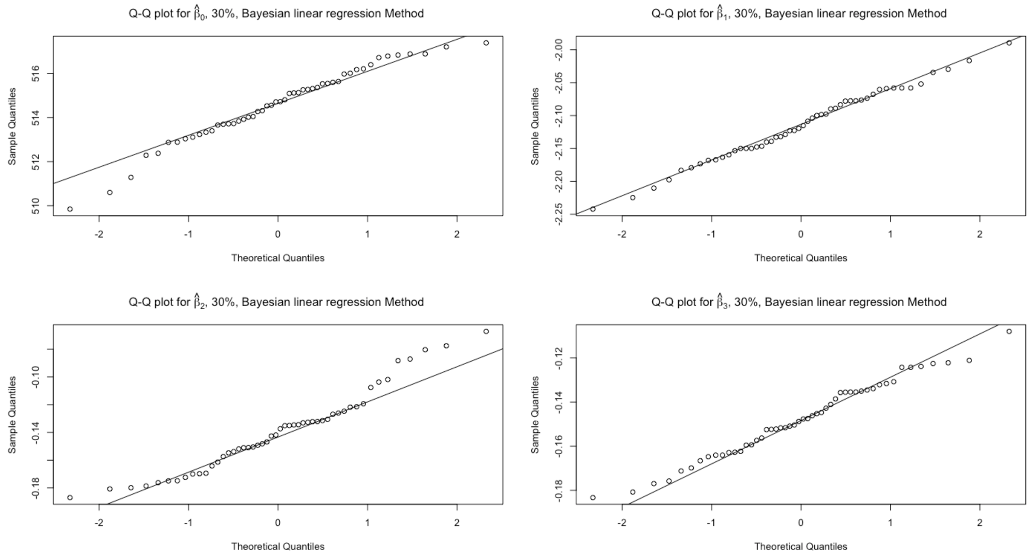
Q-Q plot for $\hat{\beta}_0$, 20%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_1$, 20%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_2$, 20%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_3$, 20%, Bayesian linear regression Method

Figure 6.7: Normality plots of the sampling distributions of the regression coefficients estimated using the Bayesian linear regression method at 20% missingness.



Q-Q plot for $\hat{\beta}_0$, 30%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_1$, 30%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_2$, 30%, Bayesian linear regression Method

Q-Q plot for $\hat{\beta}_3$, 30%, Bayesian linear regression Method

Figure 6.8: Normality plots of the sampling distributions of the regression coefficients estimated using the Bayesian linear regression method at 30% missingness.

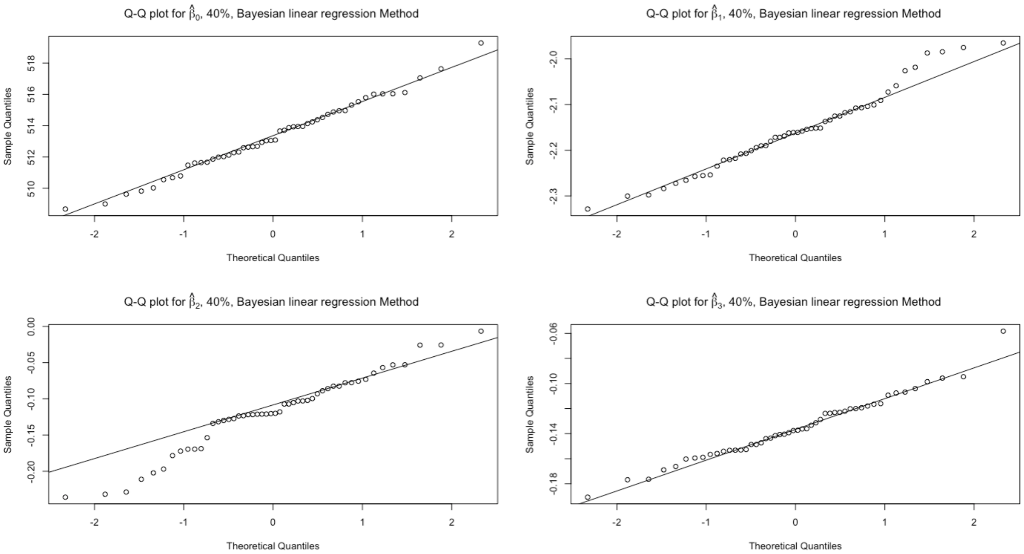


Figure 6.9: Normality plots of the sampling distributions of the regression coefficients estimated using the Bayesian linear regression method at 40% missingness.

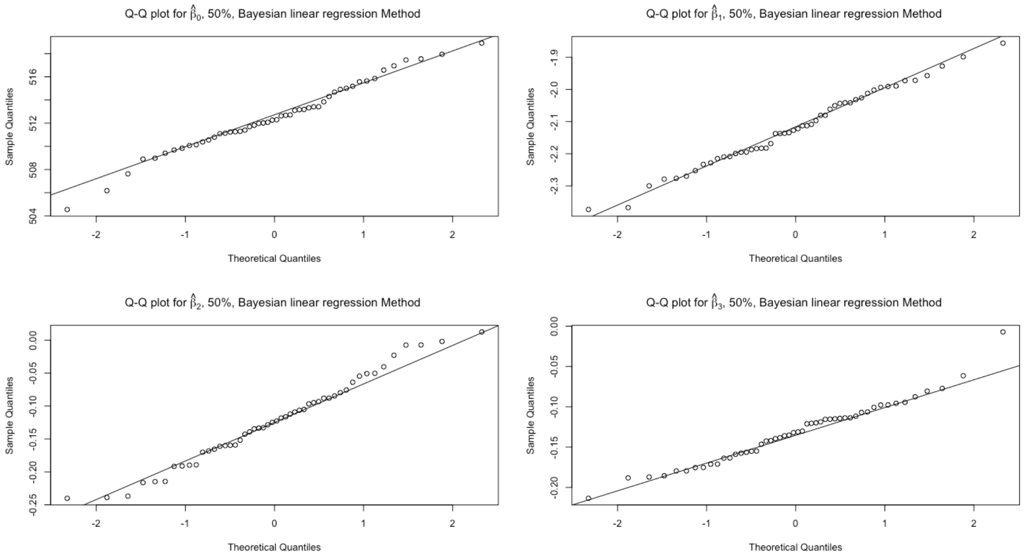


Figure 6.10: Normality plots of the sampling distributions of the regression coefficients estimated using the Bayesian linear regression method at 50% missingness.

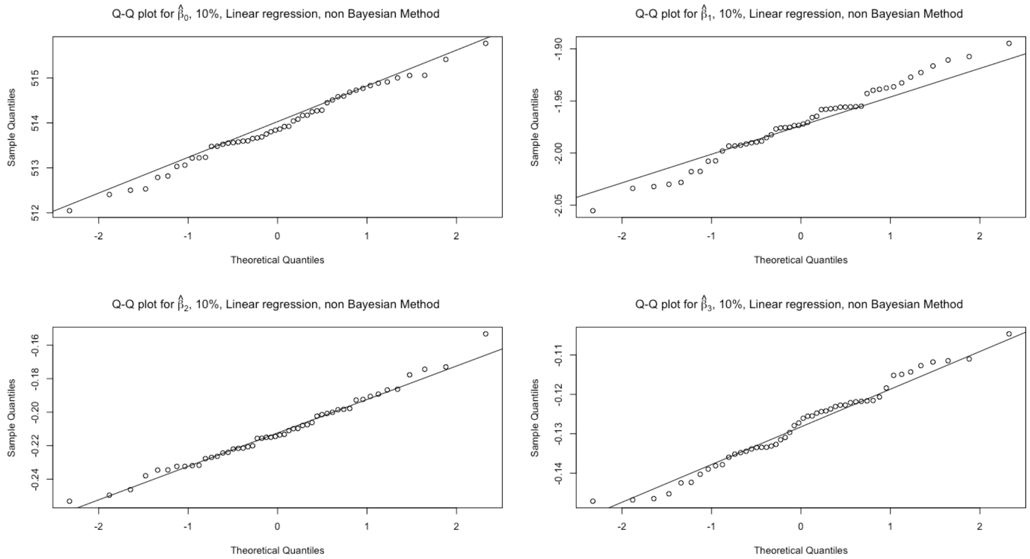


Figure 6.11: Normality plots of the sampling distributions of the regression coefficients estimated using the linear regression, non Bayesian method at 10% missingness.

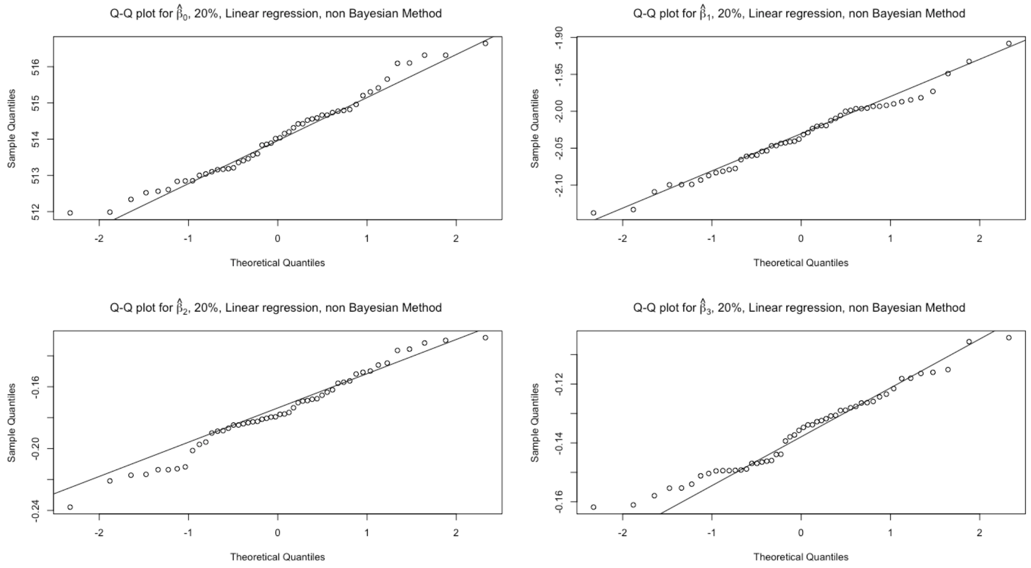


Figure 6.12: Normality plots of the sampling distributions of the regression coefficients estimated using the linear regression, non Bayesian method at 20% missingness.

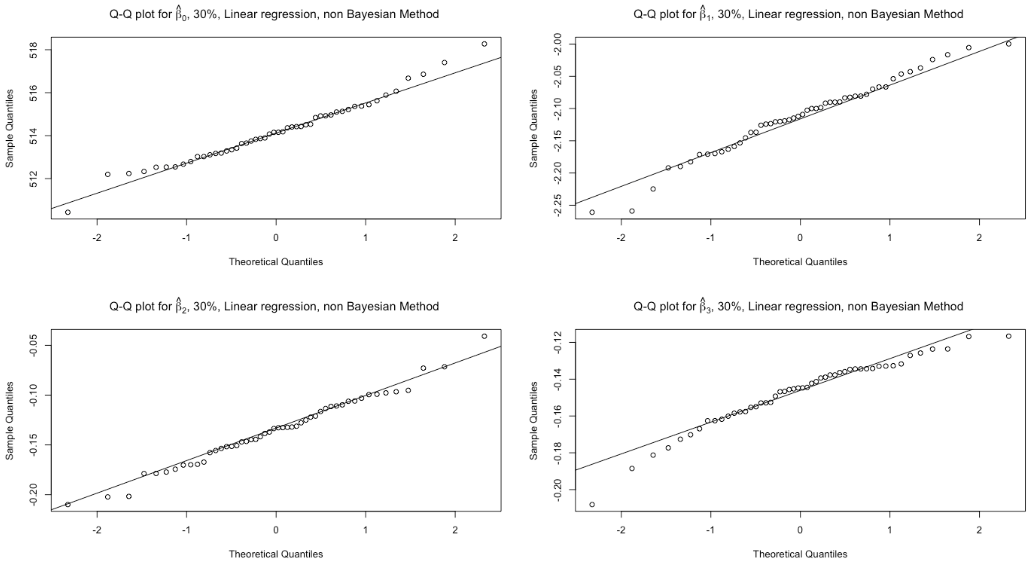


Figure 6.13: Normality plots of the sampling distributions of the regression coefficients estimated using the linear regression, non Bayesian method at 30% missingness.

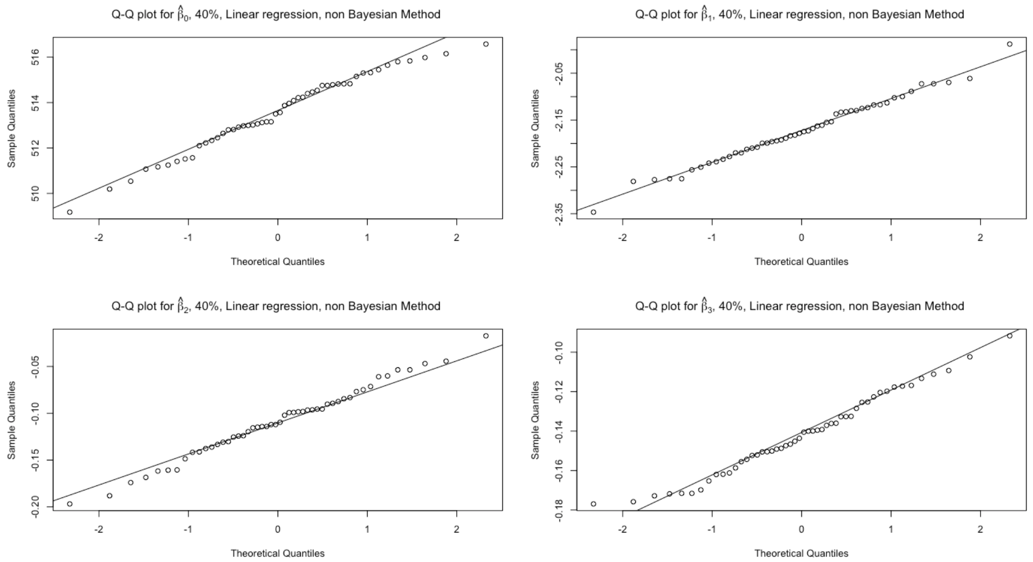


Figure 6.14: Normality plots of the sampling distributions of the regression coefficients estimated using the linear regression, non Bayesian method at 40% missingness.

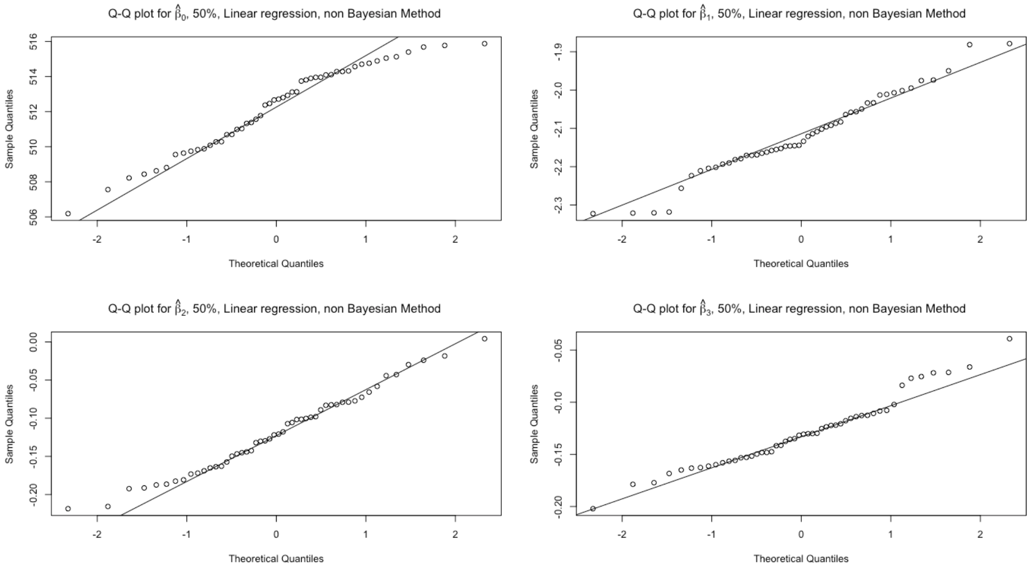


Figure 6.15: Normality plots of the sampling distributions of the regression coefficients estimated using the linear regression, non Bayesian method at 50% missingness.

6.4 Hypothesis Test Using Student's t Test Statistic

After obtaining the mean of the estimated regression coefficients, it is important to check whether there is enough evidence to show that the mean of the estimated regression coefficients are the same as the actual unbiased parameter estimated. The results from Table 6.17-6.19 indicate that almost all the estimated parameters from the imputed values are different from the actual unbiased parameter estimates. The P-values from Table 6.17 show that, using the predictive mean matching method, only the estimated parameters for $\hat{\beta}_0$ at 20% and 40% missingness are statistically insignificant. That is, statistically, the estimated parameters of $\hat{\beta}_0$ at 20% and 40% missingness are the same as the true parameter estimate of $\hat{\beta}_0$. From Table 6.18, using the Bayesian linear regression method for imputing missing values, the estimated parameter for $\hat{\beta}_0$ is statistically insignificant at 30% missingness. Also, the estimated parameters $\hat{\beta}_3$ at 20%, 40% and 50% missingness are statistically the same as the the actual parameter estimate for $\hat{\beta}_3$. Similarly, the results from Table 6.19 show that, using the linear regression, non Bayesian method, only the parameter estimates for $\hat{\beta}_3$ at 20% and 50% missingness are statistically insignificant. That is, there is enough evidence that the estimated parameters for $\hat{\beta}_0$ at 20% and 50% missingness are the same as the true values of the estimated parameters. The P-values in bold indicate the results were not significant. The family level of significance for the hypothesis test is 0.05.

Table 6.13: P-Values of the *t* statistic of the sampling distributions of regression coefficients estimated using the predictive mean matching method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	< 0.0001	< 0.0001	< 0.0001	0.0018
20%	**0.3856**	< 0.0001	< 0.0001	< 0.0001
30%	< 0.0001	< 0.0001	< 0.0001	< 0.0001
40%	**0.2082**	< 0.0001	< 0.0001	0.0133
50%	0.0032	< 0.0001	< 0.0001	0.0105

Table 6.14: P-Values of the t statistic of the sampling distributions of regression coefficients estimated using the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	< 0.0001	0.0008	< 0.0001	< 0.0001
20%	< 0.0001	< 0.0001	< 0.0001	**0.0877**
30%	**0.5853**	< 0.0001	< 0.0001	< 0.0001
40%	< 0.0001	< 0.0001	< 0.0001	**0.6679**
50%	< 0.0001	< 0.0001	< 0.0001	**0.7282**

Table 6.15: P-Values of the t statistic of the sampling distributions of regression coefficients estimated using the linear regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	< 0.0001	0.0027	< 0.0001	0.0002
20%	0.0003	< 0.0001	< 0.0001	**0.193**
30%	0.0156	< 0.0001	< 0.0001	< 0.0001
40%	< 0.0001	< 0.0001	< 0.0001	0.0089
50%	< 0.0001	< 0.0001	< 0.0001	**0.407**

7 SIMULATION STUDY

Creating a pseudo data via simulation, in modern research, is often a powerful tool to test the effectiveness of a model under various situations. Here we simulate data similar to the data previously used and then apply the imputation models on the simulated data with the aim of getting an idea of how the imputation models perform under varying conditions.

By assuming that the underlying probability distribution for each of the variables follows a normal distribution, we generate a multivariate data set that follows a normal distribution.

7.1 Multivariate Normal Distribution

The multivariate normal distribution is one of the most useful multivariate distributions. The parameters for the multivariate normal distribution are a mean vector and a covariance matrix. Using μ and Σ as the true or parametric matrix for the center point and dispersion of the multivariate distribution respectively, we write $X \sim MVN(\mu, \Sigma)$ to refer to a column vector that is drawn from the multivariate normal distribution. Using the mean, μ, and the covariance, Σ, for the variables in the CCPP data, we can write

$$X = \begin{bmatrix} T \\ V \\ RH \\ EP \end{bmatrix} \sim MVN\left(\begin{bmatrix} 20.069 \\ 54.744 \\ 72.286 \\ 453.484 \end{bmatrix}, \begin{bmatrix} 55.928 & 80.91 & -59.271 & -119.615 \\ 80.91 & 161.18 & -56.616 & -186.864 \\ -59.271 & -56.616 & 210.157 & 100.487 \\ -119.615 & -186.864 & 100.487 & 287.438 \end{bmatrix} \right),$$

where

$$\mu = \begin{bmatrix} 20.069 \\ 54.744 \\ 72.286 \\ 453.484 \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} 55.928 & 80.91 & -59.271 & -119.615 \\ 80.91 & 161.18 & -56.616 & -186.864 \\ -59.271 & -56.616 & 210.157 & 100.487 \\ -119.615 & -186.864 & 100.487 & 287.438 \end{bmatrix}$$

In simulating the multivariate normal data with R, the *mvrnorm* function in the MASS package is used [38]. The required parameters needed to use the *mvnorm* function are (1) the number of draws required, n, (2) the mean vector μ that contains p elements, and (3) the variance matrix Σ which is a $p \times p$ matrix. The desired result is an $n \times p$ matrix in which each row is a draw from MVN(μ, Σ). Here n is the number of observations and p is the number of variables.

7.2 Analysis of the Simulated Data

Here we are apply the same methodology as discussed in chapter 5 on the simulated data to verify our findings in using the CCPP data. We are drawing comparison of the results in the CCPP data set and the simulated data set to ascertain how best the methods (predictive mean matching, Bayesian linear regression and the linear regression, non Bayesian) perform on different data sets. The parameters of the fitted regression model 5.2 are indicated below in Table 7.1.

Table 7.1: Results of the CCPP model 5.2. using the simulated data

Parameter	β_0	β_1	β_2	β_3
Actual Values	517.08875^{***}	-1.94623^{***}	-0.23456^{***}	-0.15715^{***}

Using the simulated data, Table 7.1 shows that all the variables are significant in predicting EP. Table 7.1 demonstrates that, holding all other variables constant and varying only one variable at a time, EP is predicted to decrease by 1.94623 MW when temperature (T) goes up by one degrees Celsius, decrease by 0.23456 MW when exhaust vacuum (V) goes up by one cmHg, decrease by 0.15715 MW when relative humidity (RH) increases by one percent, and is predicted to be 517.08875 when T, V and RH are zero simultaneously. The adjusted R^2 value is 0.913, implying that 91.3%, of the variation in EP, is explain by the linear relationship with temperature, exhaust vaccum, and relative humidity, adjusted for the number of variables in the model. With 500 observations, we are assured normality is met due to the central limit theorem. Table 5.2 contains the VIF values for the CCPP model using the simulated data. Table 7.2, shows that all the three VIF values are less than 10, which indicates that there is no serious multicollinearity problem in the regression model 5.2 to fit the simulated data. The PRESS value of 12744.43 is relatively close to the SSE value of 12541.28. This implies that, using the simulated data, the regression model 5.2 does have a good predictive capability.

Table 7.2: Variance inflation factor of the CCPP model 5.2 using the simulated data.

Variable	T	V	RH
VIF	5.803135	4.561447	1.630341

After introducing different amounts of missingness in the complete simulated data as described in Section 5.3, and imputing the missing values using the same imputation methods (predictive mean matching, Bayesian linear regression, and linear regression, non Bayesian), we see very similar patterns in the variation and values of the estimated parametrs.

Table 7.3 shows that, applying the predictive mean matching method, the estimated mean of the regression coefficient, $\hat{\beta}_0$ using imputed data for the simulated data, tends to increase from 10% to 30%. The mean decreases at 40%, then it increases at 50%. There is an overall decrease in the estimated mean of the regression coefficients, $\hat{\beta}_1$ from 10% to 50%. Also, there is an increase in the estimated mean of the regression coefficients, $\hat{\beta}_2$ from 10% to 50%. The estimated mean of the regression coefficient, $\hat{\beta}_3$, using the imputed data, decreases from 10% to 30%, increases at 40% and then decreases at 50%.

Using the Bayesian linear regression, Table 7.5 shows that the estimated mean of the regression coefficient, $\hat{\beta}_0$, increases and decreases alternatively as the percentage of missingness increases from 10% to 50%. Furthermore, the estimated mean of the regression coefficient, $\hat{\beta}_1$, increases as the percentage of missingness increases from 10% to 20%, it decreases from 20% to 40%, then it increases from at 50%. However, the estimated mean of the regression coefficient, $\hat{\beta}_2$, decreases as the percentage of missingness increases from 10% to 20%, it increases from from 20% to 40%, and then it decreases at 50%. The regression coefficient, $\hat{\beta}_3$, increases as the percentage of missingness increases from 10% to 40%, then it decreases at 50%.

Table 7.7 specifies that, using linear regression, non Bayesian methods, the estimated mean of the regression coefficient, $\hat{\beta}_0$, alternates from increasing to decreasing as the percentage of missingness increases from 10% to 50%. The estimated mean of the regression coefficient, $\hat{\beta}_1$, increases from 10% to 20%, it decreases from 20% to 40%, and increases at 50%. Also, the estimated mean of the regression coefficient, $\hat{\beta}_2$ using the imputed data, decreases as the percentage of missingness increases from 10% to 20%, then it increases as the percentage of missingness increases from 20% to 50%. For the estimated mean of the regression coefficient, $\hat{\beta}_3$, it alternates from decreasing to increasing as the percentage of missingness increases from 10% to 50%.

From Table 7.4, variance of all the regression coefficients for the predictive mean matching method increases as the percentage of missingness increases.

Table 7.6 shows that, using the Bayesian linear regression method, the variance of the regression coefficient, $\hat{\beta_0}$ increases as the percentage of missingness moves from 10% to 50%. Also, the variance of the regression coefficient, $\hat{\beta_1}$ increases from 10% to 30% missingness, decreases at 40% and then increases for 50%. The variance of the regression coefficient, $\hat{\beta_2}$ increases from 10% to 20%, it decreases at 30%, and then it increases from 40% to 50%. For the variance of the regression coefficient, $\hat{\beta_3}$, it increases from 10% to 40% missingness, and then it decreases for 50%. Table 7.8, indicates that, the variance of the regression coefficients $\hat{\beta_0}$ and $\hat{\beta_3}$ for the linear regression, non Bayesian method increases as the percentage of missingness increases. The variance of the regression coefficients $\hat{\beta_1}$ decreases from 10% to 20%, and then it increases from 20% to 50%. Lastly, the variance of the regression coefficients $\hat{\beta_2}$ increases from 10% to 40%, and then it decreases at 50% missingness.

Table 7.3: Estimated mean of the regression coefficients with the predictive mean matching method for the simulated data.

FMI\ Estimated Parameter	$\hat{\beta_0}$	$\hat{\beta_1}$	$\hat{\beta_2}$	$\hat{\beta_3}$
10%	514.0866	-1.9925	-0.2049	-0.1293
20%	514.5529	-2.0539	-0.1726	-0.1420
30%	515.7961	-2.1176	-0.1498	-0.1567
40%	514.3423	-2.1549	-0.1095	-0.1454
50%	513.5434	-2.1549	-0.1095	-0.1454
Actual Parameter	517.0888	-1.9462	-0.2346	-0.1572

Table 7.4: Estimated Variance of the regression coefficients with the predictive mean matching method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.7297	0.0008	0.0003	0.0001
20%	1.6963	0.0028	0.0007	0.0002
30%	2.5464	0.0037	0.0010	0.0003
40%	4.2519	0.0070	0.0022	0.0005
50%	7.1126	0.0106	0.0029	0.0009

Table 7.5: Estimated mean of the regression coefficients with the Bayesian linear regression method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	517.6831	-1.9424	-0.2360	-0.1656
20%	517.5025	-1.9174	-0.2447	-0.1619
30%	514.5835	-2.0284	-0.1394	-0.1478
40%	513.3195	-2.1581	-0.1190	-0.1287
50%	512.4919	-2.1118	-0.1269	-0.1303
Actual Parameter	517.0888	-1.9462	-0.2346	-0.1571

Table 7.6: Estimated Variance of the regression coefficients with the Bayesian linear regression method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.7104	0.0015	0.0005	0.0001
20%	1.7777	0.0032	0.0010	0.0003
30%	2.8297	0.3636	0.0008	0.0003
40%	5.1404	0.0076	0.0028	0.0023
50%	8.8153	0.0140	0.0040	0.0014

Table 7.7: Estimated mean of the regression coefficients with the linear regression, non Bayesian method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	517.5752	-1.9295	-0.2415	-0.1640
20%	518.0179	-1.8360	-0.2539	-0.1641
30%	516.4696	-1.8995	-0.2351	-0.1626
40%	516.9906	-2.0344	-0.1923	-0.1933
50%	516.7151	-1.9751	-0.1697	-0.1904
Actual Parameter	517.0888	-1.9462	-0.2346	-0.1572

Table 7.8: Estimated Variance of the regression coefficients with the linear regression, non Bayesian method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.7273	0.0016	0.0005	0.0001
20%	1.3908	0.3012	0.0010	0.0002
30%	3.4215	0.0040	0.0013	0.0004
40%	4.4574	0.0076	0.0838	0.0008
50%	6.7807	0.0134	0.0041	0.0010

We conclude that using the predictive mean matching, the mean and variance of the estimated regression coefficients tend to increase as the percentage of missingness increases. Using the Bayesian linear regression and the linear regression, non Bayesian to impute missing data, although the overall variance of the regression coefficients tends to increase as the percentage of missingness increases, there is no clear direction of the estimated mean of the regression coefficients.

7.3 Range and Percentage Deviation Index for the Simulated Data.

Tables 7.9 shows that the range increases for all the parameters as the fraction of missing information increases when using the predictive mean matching method. Using the Bayesian linear regression method to impute missing values, Table 7.11 indicates that the range increases as the percentage of missingness increases for estimated parameter $\hat{\beta}_0$. For the estimated parameter $\hat{\beta}_1$, the range increases as the percentage of missingness increases from 10% to 30%, it reduces at 40%, and then it increases at 50%. The range for the estimated parameter, $\hat{\beta}_2$, increases as the percentage of missingness increases from 10% to 20%, it reduces at 30%, and then it increases from 40% to 50%. The range for the estimated parameter, $\hat{\beta}_3$, increases as the percentage of missingness increases from 10% to 20%, it reduces at 30%, it increases for 40%, and then it reduces for 50%. From Table 7.13, the range increases as the percentage of missingness increases from 10% to 40% for estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_3$ and then it reduces at 50%. The range for the estimated parameter, $\hat{\beta}_1$, increases sharply as the percentage of missingness increases from 10% to 20%, it reduces at 30% and then it increases from 30% to 50%. The range for the estimated parameter, $\hat{\beta}_2$, increases as the fraction of missing information increases from 10% to 20%, it reduces fairly at 30% missingness and increases again from 30% to 50%.

Comparing Tables 7.10, 7.12 and 7.14, the predictive mean matching method of imputing missing values for the simulated data produces the largest overall PDI of 9.011%. The PDI produces 7.173% and -0.968% for the Bayesian linear regression and linear regression, non Bayesian methods respectively. This implies that when using the simulated data, the predictive mean matching method imputed data with relatively large variation from the original data hence making it less effective to impute missing data. On the contrary, the linear regression, non Bayesian method underestimates the imputed data produced. Table 7.14 specifies that, using the linear regression, non Bayesian method, most of the missing values introduced in the simulated data are underestimated. Additionally, Table 7.14 shows the linear regression, non Bayesian method imputes data with smaller deviation. However, while the predictive mean matching imputes data with large deviation index for 10% and 20% missingness, the Bayesian linear regression method produces imputed data with large deviation for 40% and 50% missingness.

Table 7.9: Range of the regression coefficients with the predictive mean matching method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.7656	0.1492	0.0885	0.0491
20%	5.63	0.2591	0.1099	0.0621
30%	6.6080	0.3234	0.1567	0.0898
40%	9.8568	0.3389	0.1921	0.093
50%	12.5502	0.5313	0.2247	0.1343

Table 7.10: PDI of the regression coefficients with the predictive mean matching method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	**Mean**
10%	0.58%	-2.38%	12.62%	17.69%	7.13%
20%	0.49%	-5.53%	26.44%	9.64%	7.76%
30%	0.24%	-8.81%	36.13%	0.31%	6.97%
40%	0.53%	-10.07%	42.03%	9.55%	10.51%
50%	0.69%	-10.72%	53.31%	7.46%	12.68%
Mean	0.51%	-7.50%	34.11%	8.93%	9.01%

Table 7.11: Range of the regression coefficients with the Bayesian linear regression method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	4.0071	0.1492	0.0846	0.0443
20%	6.3430	0.2933	0.1513	0.0831
30%	7.5408	4.3750	0.1198	0.0753
40%	10.6018	0.3635	0.2293	0.3510
50%	14.3545	0.5174	0.2526	0.2064

Table 7.12: PDI of the regression coefficients with the Bayesian linear regression method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	**Mean**
10%	-0.11%	0.20%	-0.61%	-5.40%	-1.48%
20%	-0.08%	1.48%	-4.34%	-3.03%	-1.49%
30%	0.48%	-4.22%	40.56%	5.95%	10.69%
40%	0.73%	-10.89%	49.26%	18.08%	14.30%
50%	0.89%	-8.51%	45.91%	17.10%	13.85%
Mean	0.38%	-4.39%	26.16%	6.54%	7.17%

Table 7.13: Range of the regression coefficients with the linear regression, non Bayesian method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	4.0953	0.1916	0.1237	0.0529
20%	5.6811	4.0190	0.1594	0.0706
30%	9.9696	0.2701	0.1512	0.0927
40%	10.3836	0.3983	2.1136	0.1550
50%	10.1803	0.5220	0.2890	0.1464

Table 7.14: PDI of the regression coefficients with the linear regression regression, non Bayesian method for the simulated data.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	**Mean**
10%	-0.09%	0.86%	-2.95%	-4.38%	-1.64%
20%	-0.18%	5.66%	-8.25%	-4.42%	-1.80%
30%	0.12%	2.40%	-0.24%	-3.50%	-0.30%
40%	0.02%	-4.53%	18.02%	-22.98%	-2.37%
50%	0.07%	-1.48%	27.63%	-21.15%	1.27%
Mean	-0.01%	0.58%	6.84%	-11.28%	-0.97%

7.4 Test for Normality of the Parameter Estimates using the Simulated Data.

With fifty observations for each of the parameter estimates, by the central limit theorem (CLT), each sampling distribution of the parameter estimates follows an approximate normal distribution. To verify the assumption of normality, Q-Q plots are constructed to provide visual support of normality. Figures 7.1-7.15 show in most cases the points are close to a straight line, demonstrating the assumption of normality is satisfied. Some of the plots appears to be curved, but with fifty observations, we are assured of normality because of CLT.

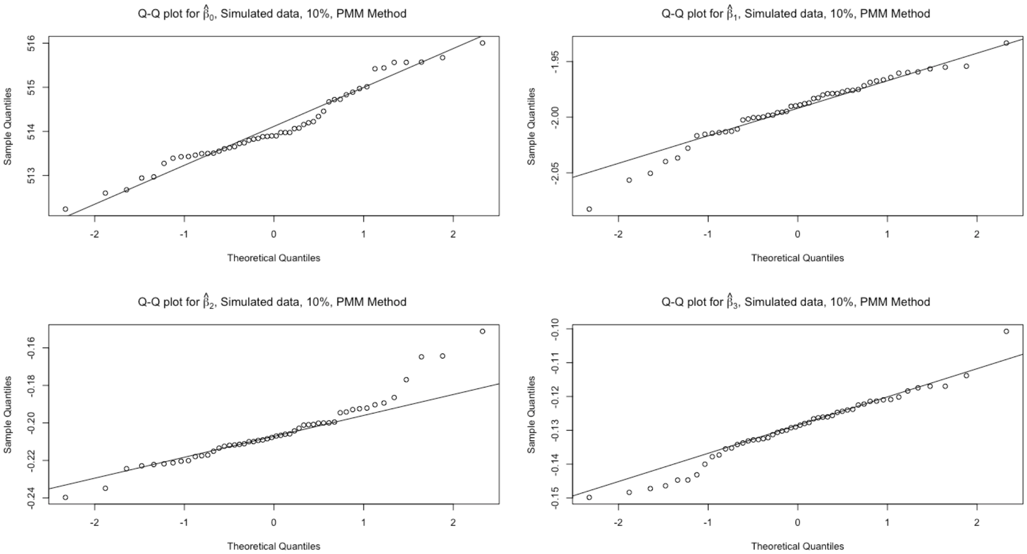


Figure 7.1: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the predictive mean matching method at 10% missingness.

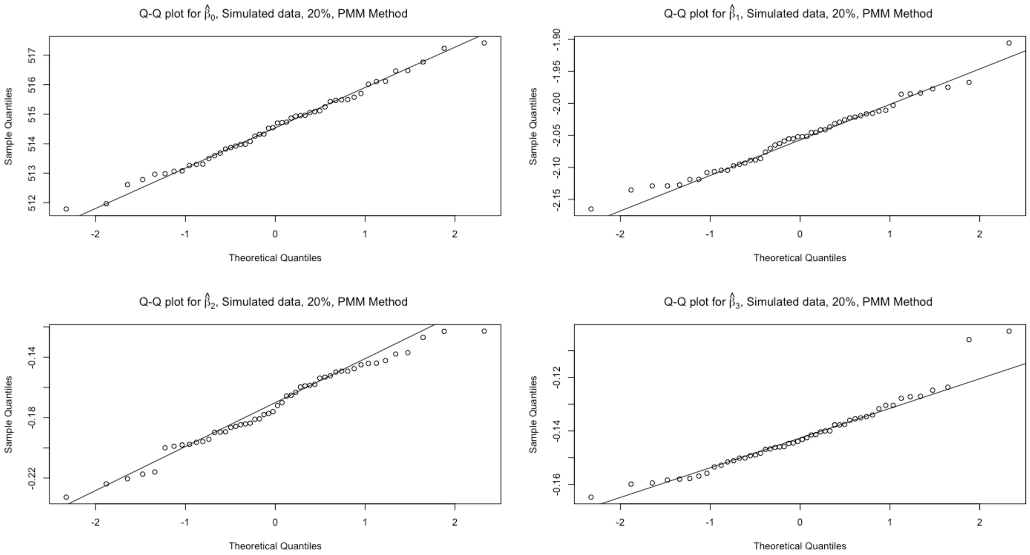


Figure 7.2: Normality plots of the sampling distributions of the regression coefficients estimated using the predictive mean matching method at 20% missingness.

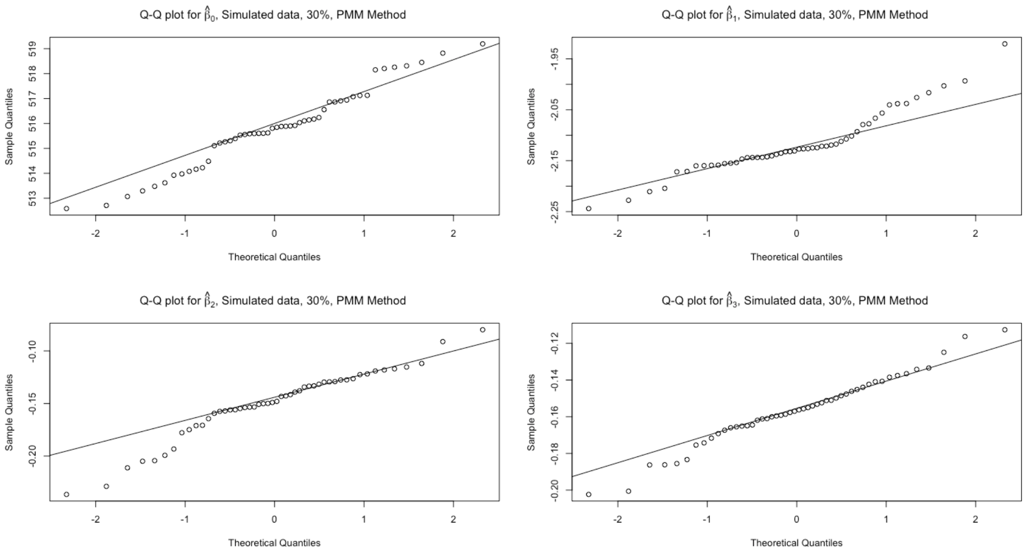


Figure 7.3: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the predictive mean matching method at 30% missingness.

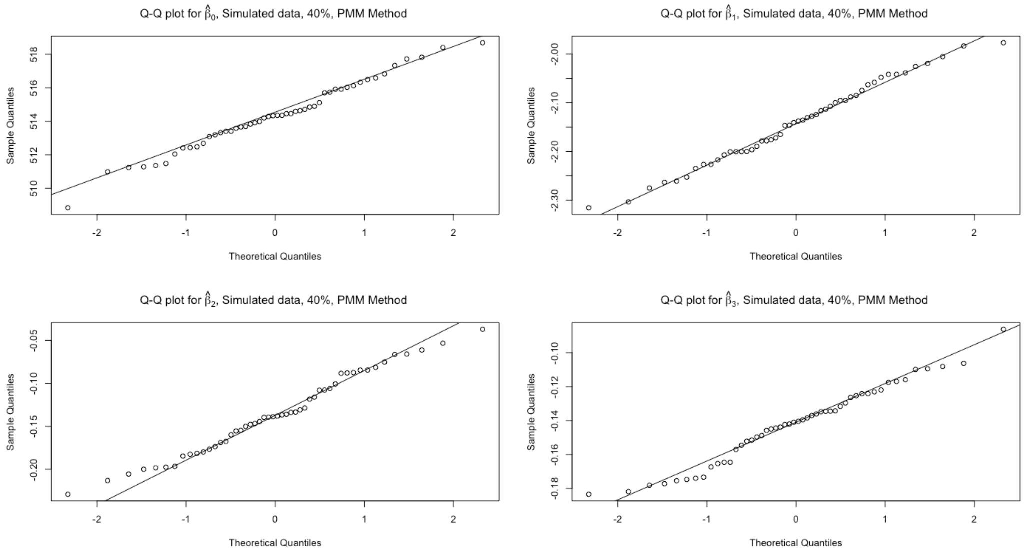


Figure 7.4: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the predictive mean matching method at 40% missingness.

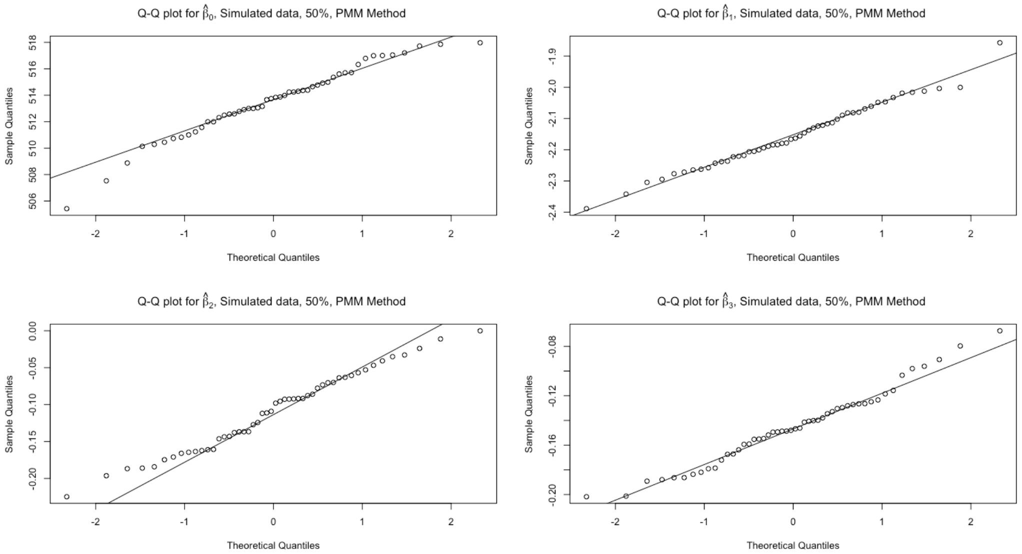


Figure 7.5: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the predictive mean matching method at 50% missingness.

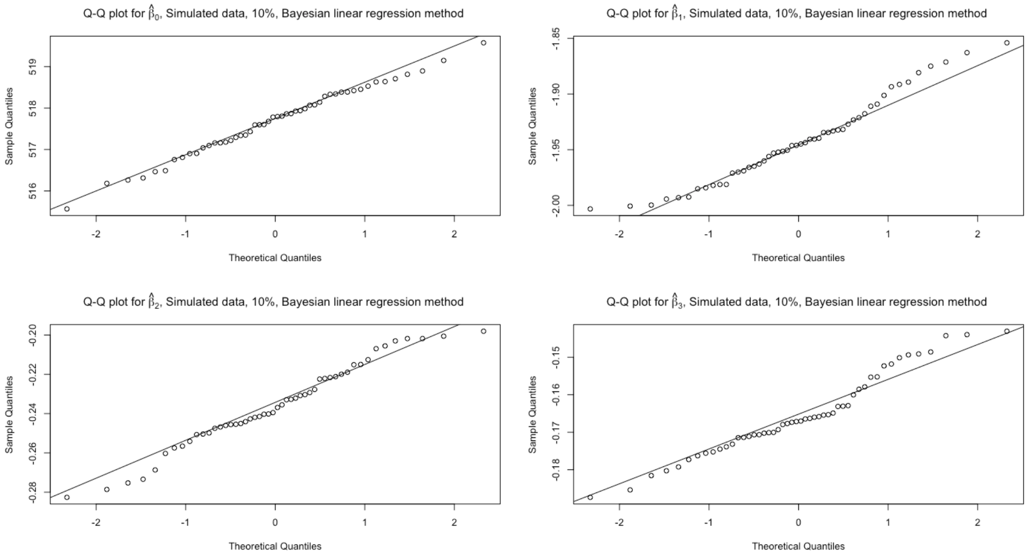


Figure 7.6: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the Bayesian linear regression method at 10% missingness.

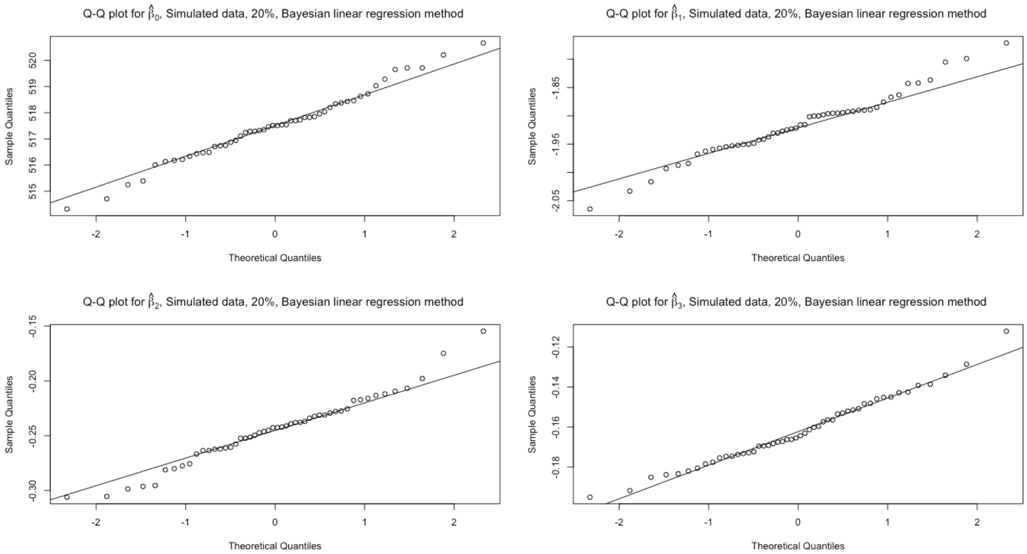


Figure 7.7: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the Bayesian linear regression method at 20% missingness.

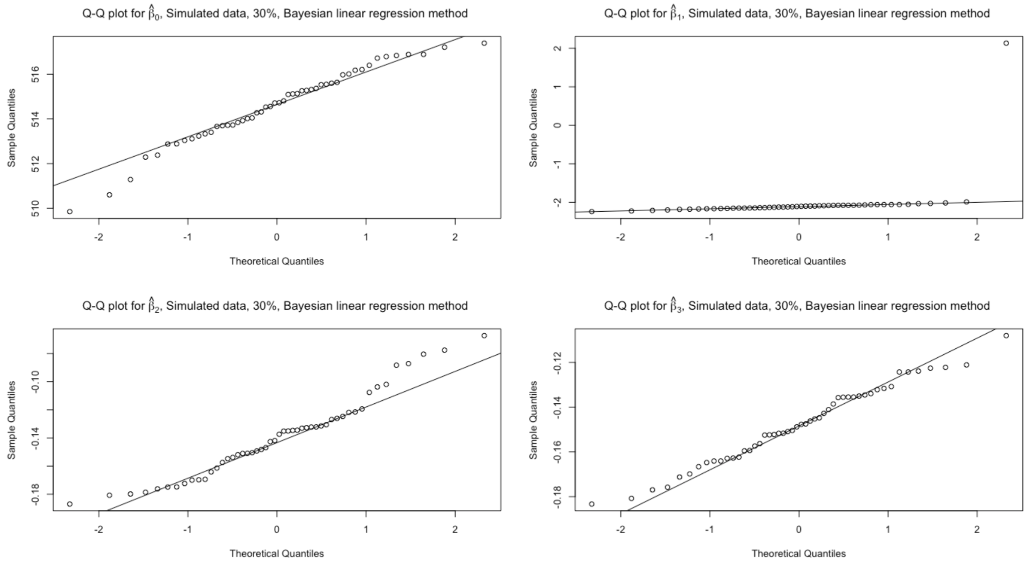


Figure 7.8: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the Bayesian linear regression method at 30% missingness.

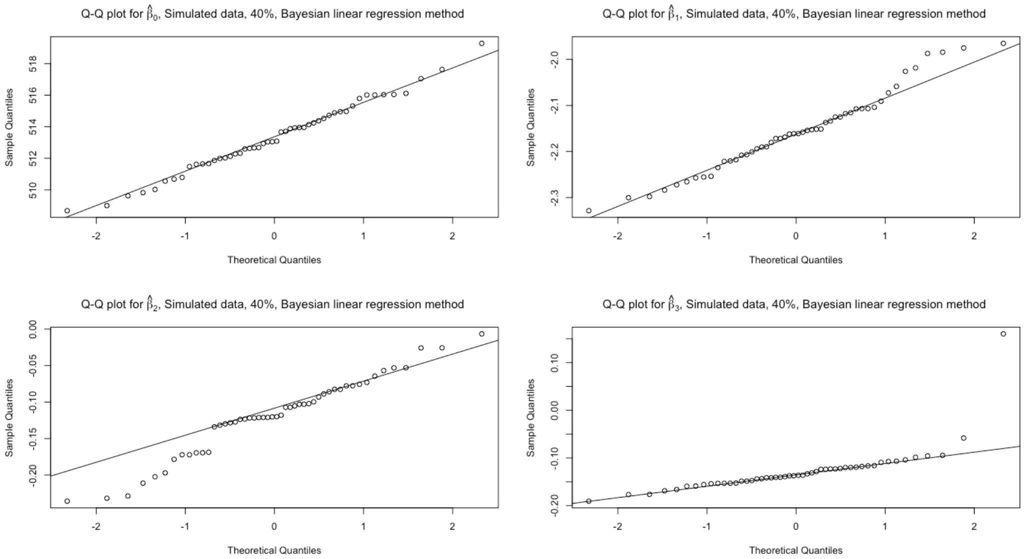


Figure 7.9: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the Bayesian linear regression method at 40% missingness.

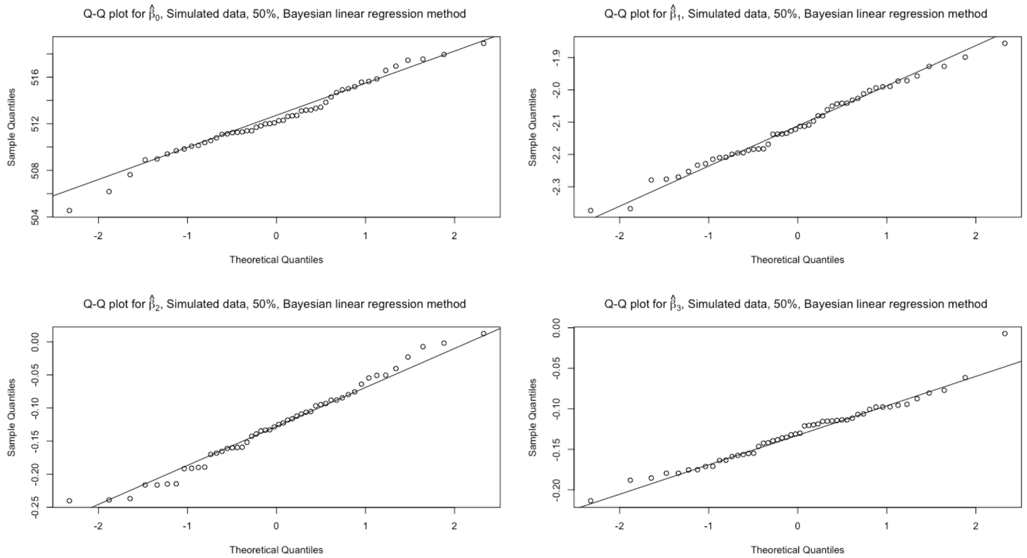


Figure 7.10: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the Bayesian linear regression method at 50% missingness.

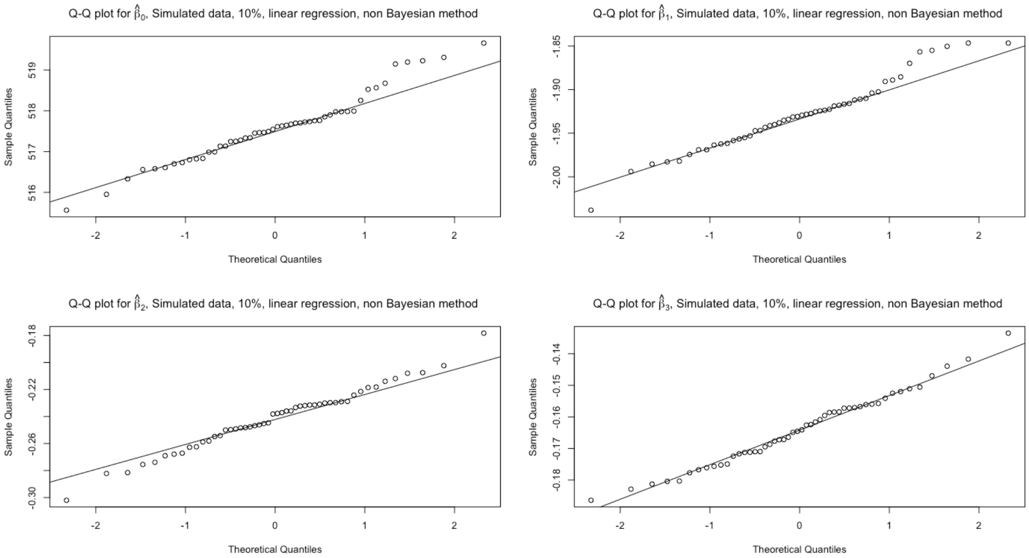


Figure 7.11: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the linear regression, non Bayesian method at 10% missingness.

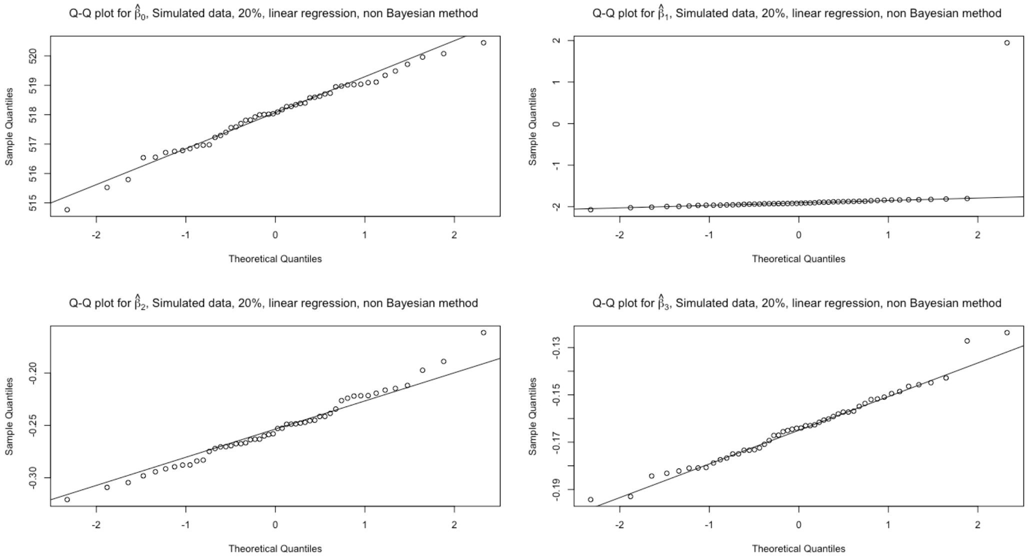


Figure 7.12: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the linear regression, non Bayesian method at 20% missingness.

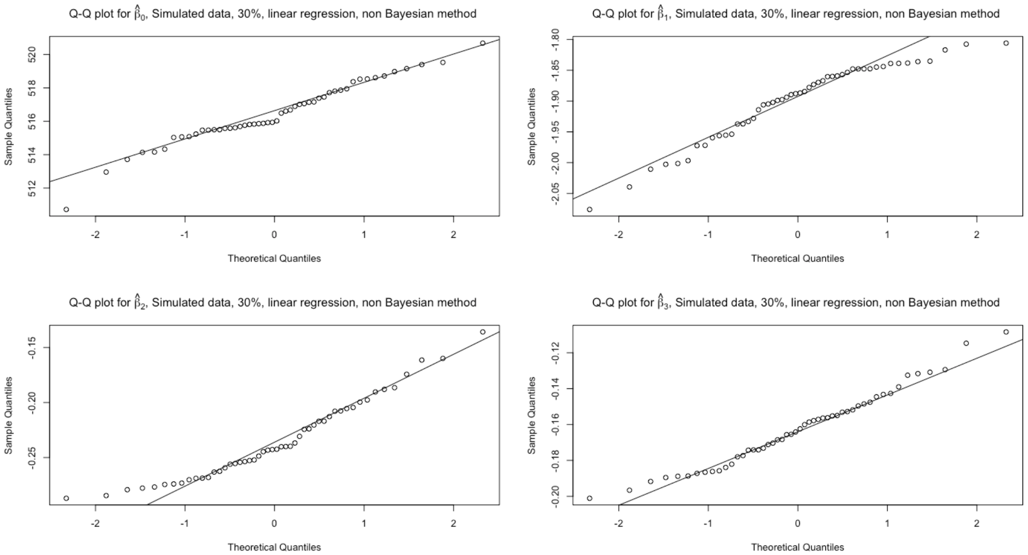


Figure 7.13: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the linear regression, non Bayesian method at 30% missingness.

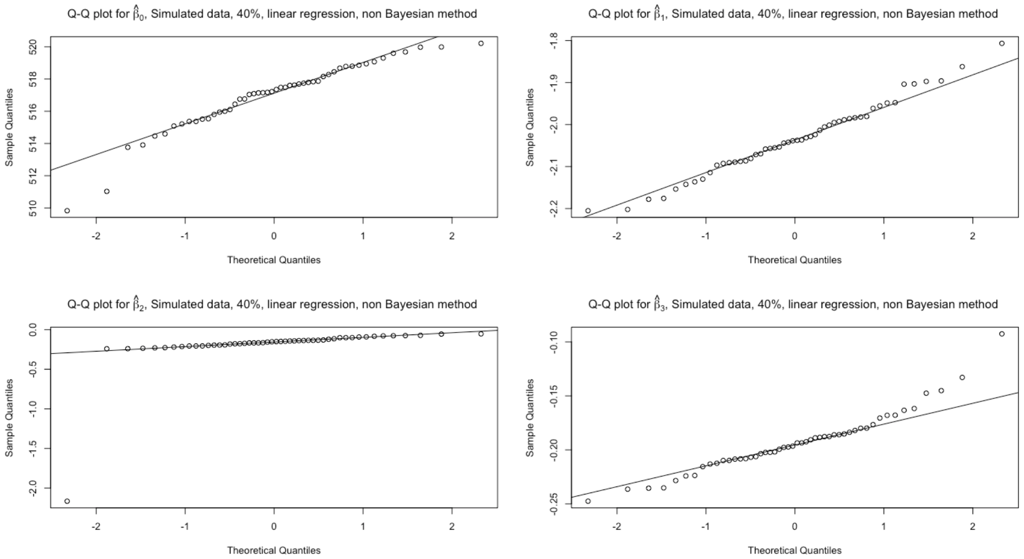


Figure 7.14: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the linear regression, non Bayesian method at 40% missingness.

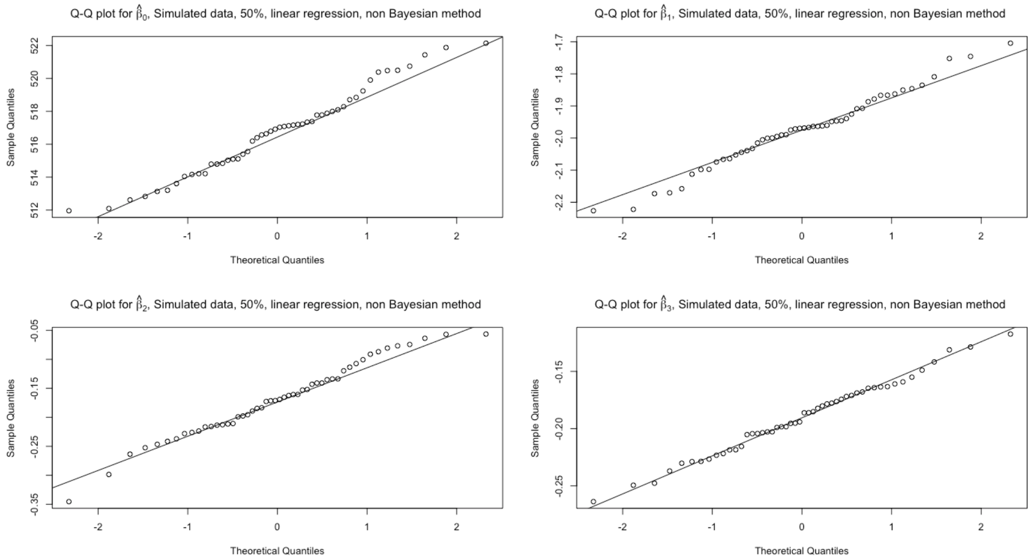


Figure 7.15: Normality plots of the sampling distributions of the regression coefficients estimated for the simulated data using the linear regression, non Bayesian method at 50% missingness.

7.5 Hypothesis Test Using Student's t Test Statistic

Now, we establish if there is evidence that the estimated regression coefficient for the simulated data is the same as the actual regression unbiased parameter estimates of the simulated data. Tables 7.15-7.17 show that only a few of the estimated parameters from the imputed values for the simulated data are significantly equal to the actual unbiased parameter estimates. Using the predictive mean matching method, Table 7.15 indicates that only the estimated parameter for $\hat{\beta}_3$ at 30% is statistically equal to the actual parameter. From Table 7.16, using the Bayesian linear regression method for imputing missing values, the estimated parameter, $\hat{\beta}_1$ is statistically insignificant at 10% and 30% missingness. Also, the estimated parameter, $\hat{\beta}_2$ is statistically equal to the true parameter at 10% missingness. Statistically, the estimated parameter $\hat{\beta}_3$ is the same as the the actual parameter estimate at 20% missingness. Table 7.17 shows that, using the linear regression, non Bayesian method, the parameter estimate for $\hat{\beta}_0$ is statistically insignificant at 40% and 50% missingness. The parameter estimate $\hat{\beta}_1$ is statistically the same as the true parameter at 20% and 50% missingness. There is enough evidence that the estimated parameter for $\hat{\beta}_2$ at 30% and 40% missingness are the same as the true values of the estimated parameters. The estimated parameters for $\hat{\beta}_3$ at 30% is statistically equal to the actual parameter. The P-values in bold indicate the results were not significant. The family level of significance for the hypothesis test is 0.05.

Table 7.15: P-Values of the *t* statistic of the sampling distributions of regression coefficients estimated for the simulated data using the predictive mean matching method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	< 0.0001	< 0.0001	< 0.0001	< 0.0001
20%	< 0.0001	< 0.0001	< 0.0001	< 0.0001
30%	< 0.0001	< 0.0001	< 0.0001	**0.8524**
40%	< 0.0001	< 0.0001	< 0.0001	< 0.0001
50%	< 0.0001	< 0.0001	< 0.0001	0.0092

Table 7.16: P-Values of the t statistic of the sampling distributions of regression coefficients estimated for the simulated data using the Bayesian linear regression method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	< 0.0001	**0.4862**	**0.6357**	< 0.0001
20%	0.0330	0.0007	0.0246	**0.0528**
30%	< 0.0001	**0.3398**	< 0.0001	0.0004
40%	< 0.0001	< 0.0001	< 0.0001	0.0001
50%	< 0.0001	< 0.0001	< 0.0001	< 0.0001

Table 7.17: P-Values of the t statistic of the sampling distributions of regression coefficients estimated for the simulated data using the linear regression, non Bayesian method.

FMI\Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0002	0.0051	0.0388	< 0.0001
20%	< 0.0001	**0.162**	< 0.0001	0.0017
30%	0.0219	< 0.0001	**0.9137**	**0.07216**
40%	**0.7437**	< 0.0001	**0.307**	< 0.0001
50%	**0.3152**	**0.0839**	< 0.0001	< 0.0001

8 DISCUSSION

The results from the CCPP and the simulated data show that, in almost all the cases, the absolute value of the estimated mean and the variance increases as the percentage of missingness increases for all the three imputation models. Oketch, 2017, argues that data with small amount of missingness contains more of the actual information than data with large missingness [20]. Thus, the variation in data with a small fraction of missing information is less than the variation in data with a large amount of missing information. According to Oketch, 2017, for a large amount of missingness, the same imputed values are revisited and used to fill-in the missing values at different positions, hence widening the variation between the imputed data and the actual data. [20].

One question that needs to be answered is, for all the imputation methods, how close are the estimated means to the actual means, given the percentage of missingness? We observe that, for the CCPP data, the predictive mean matching produces parameter estimates that are close to the actual parameter than the Bayesian linear regression and the linear regression, non Bayesian methods. This could be due to the fact that the imputed values are chosen from the observed values, therefore keeping the variation between the imputed data low. This is confirmed by the small overall PDI and the range of the estimated regression coefficients. However, for the simulated data, the linear regression, non Bayesian method imputed missing values that produce estimated regression coefficients that are closest to the actual parameters for the simulated data. This is evidenced by the lower overall PDI of the estimated regression coefficients. This can be attributed to the fact that the simulated data comes from a multivariate normal distribution, hence re-enforcing the claim made by Buuren and Groothuis-Oudshoorn (2000) that the linear regression, non Bayesian performs better for data that follow a normal distribution with a large sample size, where variability is not much of an issue [32]. In summary, we can say that, since we did not know the actual distribution of the CCPP data, the predictive mean matching works better for a nonparametric data and the linear regression, non Bayesian produces better results for a multivariate normal data. One interesting observation is that, with the predictive mean matching method, the estimated regression coefficient, the variance and the range for both the CCPP data and the simulated data are the same.

From the one sample t-test, most estimated regression coefficients for the CCPP and the simulated data are significantly different from the corresponding actual parameters. Yet, comparing the three imputation methods, the linear regression, non Bayesian method produces relatively more estimated regression coefficients (using the simulated data) that are significantly equal to the actual parameter. This affirms the point that, for a multivariate normal distribution, the linear regression, non Bayesian method generates better imputations values.

9 CONCLUSION

This paper discusses three imputation methods, namely predictive mean matching, Bayesian linear regression method and the linear regression method, non Bayesian, and evaluates how these methods perform at certain percentages of missingness.

We conclude that the predictive mean matching produces better imputed data for nonparametric data than the Bayesian linear regression and the linear regression non Bayesian method. With a non parametric data, the predictive mean matching produces better results for all the percentages of missingness. Considering the three imputation methods, with a data that is approximatly multivariate normal, the linear regression, non Bayesian method imputes accurate data that yields better results. This is true for all the percentages of missing information.

9.1 Future Work

In our quest to solve the problem of missing data, it is important to identify the actual distribution of each of the variables in the actual data set, simulate data from the actual distribution of each of the variable and extend this analysis on the simulated data.

In addition, as indicated in Table 4.1, there are other functions in the $MICE$ of R that can be used to impute missing data. Extending the analysis to incorporate other scales of measurements is vital to ensure an overall understanding of the imputation methods and to determine which methods work best for different situations.

BIBLIOGRAPHY

[1] D. B. Rubin. Inferences and Missing Data. *Biometrika*. Dec., 1976, Volume 63, Issue 3, 581-592.

[2] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T Farrar, et al. The Prevention and Treatment of Missing Data in Clinical Trials. N Engl J Med. 2012;367:1355–60. `http://dx.doi.org/10.1056/NEJMsr1203730`, [Online; accessed August 29, 2017]

[3] P. D. Allison. Missing Data. In Roger E. Millsap and Alberto Maydeu-Olivare (Eds), *The Sage Handbook of Quantitative Methods in Psychology*, Thousand Oaks, California, 2009.

[4] R. J. Little, and D. B Rubin. Statistical Analysis with Missing Data. 2nd ed. New York: John Wiley and Sons, 2002

[5] C. K. Enders. Applied Missing Data Analysis. The Guilford press, New York, 2012

[6] Y. Dong, and C.-Y. J. Peng. Principled Missing Data Methods for Researchers. Springer Plus. 2013; 2, 222, `https://doi.org/10.1186/2193-1801-2-222`, [Online; accessed August 29,2017].

[7] J. L. Schafer, and J. W. Graham. Missing Data: Our View of the State of the Art. *Psychological Methods*. 2002; 7, 147-177, http://dx.doi.org/10.1037/1082-989X.7.2.147, [Online; accessed August 29,2017].

[8] J. W. Graham. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*. 2009; Volume 60. https://doi.org/10.1146/annurev.psych.58.110405.085530, [Online; accessed August 29, 2017]

[9] W. A. Kamakura, and M. Wedel. Factor Analysis and Missing Data. *Journal of Marketing Research*. 2000; 37(4) 490-98.

[10] Wilkinson Task Force on Statistical Inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*. 1999; 54(8), 594604, 1999.

[11] A. N. Baraldi, and C. K. Enders. An introduction to modern missing data analyses. *Journal of School Psychology*. 2010; 48(1), 5-37. DOI:10.1016/j.jsp.2009.10.001.

[12] R. J. Little. Regression with missing X's: a review. *Journal of the American Statistical Association*. 1992; 87: 1227–1237.

[13] J. L. Peugh, and C. K. Enders. Missing data in educational research : A review of reporting practices and suggestions for improvement. *In: Review of Educational Research.* 2004 ; Vol. 74, No. 4. pp. 525-556

[14] D. B. Rubin, and N. Schenker. Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association.* 1986; 81, 366-374.

[15] S. Van Buuren, and C. G. M. Groothuis-Oudshoorn. Mice: Multivariate Imputation by Chained Equations in R. *Journal of statistical software.* 2011; 45(3).

[16] R. V. Hogg, J. W. McKean, and A. T. Craig. Introduction to Mathematical Statistics. 7th ed. Boston: Pearson, 2013.

[17] J. L. Schafer. Analysis of incomplete multivariate data. London, UK: Chapman and Hall. 1997.

[18] W. R. Gilks, R. Sylvia, and D. J. Spiegelhalter. Markov Chain Monte Carlo in Practice. Boca Raton, Fla. [u.a.: Chapman Hall, 1998.

[19] W. Lang, L. Wei, Y. Y. Grace, and H. YangxinAN. *Analysis of Longitudinal and Survival Data: Joint Modeling, Inference Methods, and Issues* Journal of Probability and Statistics, vol. 2012, Article ID 640153, 2012. doi:10.1155/2012/640153, [Online; accessed October 12, 2017].

[20] T. O. Oketch. Performance of Imputation Algorithms on Artificially Produced Missing at Random Data (2017). Electronic Theses and Dissertations. Paper 3217. `http://dc.etsu.edu/etd/3217`

[21] M. H. Hof, J. Z. Musoro, R. B. Geskus, G. H. Struijk, I. J. M. Ten Berge, and A. H. Zwinderman. Simulated maximum likelihood estimation in joint models for multiple longitudinal markers and recurrent events of multiple types, in the presence of a terminal event. *Journal of Applied Statistics*, 2017, 2756-2777, `DOI:10.1080/02664763.2016.1262336`, [Online; accessed October 12, 2017]

[22] S. Ross. Introduction to Probability and Statistics for Engineers and Scientists. Academic Press, 2004, 9780080470313.

[23] P. D. Allison, Missing Data. Sage University papers series on Quantitative Application in the Social Sciences, 07-136. Thousand Oak, CA: Sage, 2002.

[24] C. K Enders. Applied Missing Data Analysis. The Guilford Press, New York, NY 10012, 2010.

[25] D. B. Rubin. Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association*, 1977, 72:538-543

[26] D. B. Rubin. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, Inc., 1987.

[27] P. D. Allison. Handling Missing Data by Maximum Likelihood. SAS Global Forum 2012,2012, Paper 312-2012.

[28] R. J. Little A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association.* 1988. 83(404), 1198-1202. DOI:10.1080/01621459.1988.10478722, [Online; accessed October 15,2017]

[29] P. D. Allison. Imputation by Predictive Mean Matching: Promise and Peril. March 5, 2015. http://statisticalhorizons.com/predictive-mean-matching. [Online; accessed September 17, 2017].

[30] D. B RUBIN, AND N. SCHENKER, *Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse.* Journal of the American Statistical Association. 1986. 81, 366-374.

[31] J. L. Schafer, and M. K. Olsen. Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective. Multivariate Behav. Res. 1998; 33(4): 545–571. doi:10.1207/s15327906mbr3304_5.

[32] S. Van Buuren, and C. G. M. Groothuis-Oudshoorn. Multivariate Imputation by Chained Equations: MICE V1.0 User's manual. TNO Report PG/VGZ/00.038. 2000. Leiden: TNO Preventie en Gezondheid.

[33] M. Lichman. UCI Machine Learning Repository. 2013. University of California, Irvine, School of Information and Computer Sciences. `http://archive.ics.uci.edu/ml`, [Online; accessed October 10, 2017]

[34] J. L. Schafer. Multiple imputation: a primer. Statistical Methods in Medical Research 8: 3-15.

[35] P. D. Allison. Why You Probably Need More Imputations Than You Think. `https://statisticalhorizons.com/more-imputations`. November 9, 2012.

[36] Combined Cycle Power Plant: How it Works. `https://www.gepower.com/resources/knowledge-base/combined-cycle-power-plant-how-it-works` [Online; accessed January 10, 2018].

[37] M. H. Kutner, C. J. Nachtsheim, J. Neter, and W. Li. Applied linear statisitcal models. 5th ed. New Delhi: McGraw Hill Education, 2005.

[38] W. N. Venables, and B. D. Ripley. Modern Applied Statistics with S. New York: Springer, fourth edition. 2002. ISBN 0-387-95457-0.

VITA

EVANS DAPAA ADDO

Education:	B.A. Social Sciences (Economics), University of Cape Coast, Cape Coast, Ghana 2014 M.S. Mathematical Sciences (Statistics), East Tennessee State University Johnson City, Tennessee 2018
Professional Experience:	Teaching Assistant, University of Cape Coast, Cape Coast, Ghana, 2014–2015 Graduate Teaching Assistant, East Tennessee State University Johnson City, Tennessee, 2016–2018