5-2017

Performance of Imputation Algorithms on Artificially Produced Missing at Random Data

Tobias O. Oketch
East Tennessee State University

Performance of Imputation Algorithms on Artificially Produced 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

Tobias Oketch

May 2017

Nicole Lewis, Ph.D., Chair

Anant Godbole, Ph.D.,

Bob Price, Ph.D.

JeanMarie Hendrickson, Ph.D.

Keywords: Multiple imputation by chained equation, Relative efficiency.

ABSTRACT

Performance of Imputation Algorithms on Artificially Produced Missing at Random Data

by

Tobias Oketch

Missing data is one of the challenges we are facing today in modeling valid statistical models. It reduces the representativeness of the data samples. Hence, population estimates, and model parameters estimated from such data are likely to be biased.

However, the missing data problem is an area under study, and alternative better statistical procedures have been presented to mitigate its shortcomings. In this paper, we review causes of missing data, and various methods of handling missing data. Our main focus is evaluating various multiple imputation (MI) methods from the multiple imputation of chained equation (MICE) package in the statistical software R. We assess how these MI methods perform with different percentages of missing data. A multiple regression model was fit on the imputed data sets and the complete data set. Statistical comparisons of the regression coefficients are made between the models using the imputed data and the complete data.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to all the members of his advisory committee who greatly contributed to the preparation of this study. Special thanks are extended to Dr. Nicole Lewis whose time and efforts were invaluable.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Missing data occur in many studies, when questions may be partially answered or be completely unanswered. This may frustrate researchers efforts to achieve objectives of their studies because the information they intended to collect for analysis would be incomplete or missing.

The impact of missing data is always negative, and the extent may be quite severe even with small amount missing data. This is because; it usually leads to loss of samples representativeness, unbiased estimates, and exaggerated variances and standard error of the estimates of the true values. Missing data also reduces the researcher's ability to make correct decisions regarding the subject matter. Limitations of missing data are further worsened by the fact that, most of the analytical software and researchers assume that the data are always complete even when they are actually missing [1]. Therefore, handling of missing data is very critical for good results and understanding the reasons or mechanisms that causes missing data is helpful in addressing the problem.

There are several methods out there that have been put forward to tackle the missing data problems. In this project our focus will be on multiple imputations by chained equations (MICE) also known as fully conditional specification (FCS) methods.

1.1 Objectives

The objective of this study is to evaluate the performance of the latest methods of imputing data for different percents of missing data. These methods are trusted to yield plausible results, imputed values that are as good as observed. In this project there are two specific objectives:

- To evaluate the performance of multiple imputation by chained equation (MICE) methods using a statistical software, the MICE package in R
- To compare data analysis results on complete and imputed data sets.

MICE package offers different imputation models that we employ, separately, on varying percentages of missing values. We are using the percentage deviation of the estimated regression coefficients using imputed values to determined the plausibility of the results. We also compare the mean and variances of the estimated parameters to evaluate the reliability of each of the imputation models in MICE package.

1.2 Limitations

MICE methods assume that the missing data are missing at random, (MAR). Suggesting that one can completely account for the missing values using available data. This assumption might not perfectly fit a true situation. This is because, missing data could be due to other personal reasons that are only known to the respondent, and could not be specified in the context of the study.

Secondly, missing data analysis is a developing area under study with limited conventional breakthroughs. For example, the basic idea of imputing missing data

is based on statistically untestable assumption of data MAR. Currently, there is no known ad hoc methods to test this assumption. However, we are using ProdNA function in R software to introduce missing values by deleting at random different amounts of complete data. Therefore, simulating and satisfying the assumption of data MAR.

1.3 Significance of the Study

With missing values, one is likely to lose very important information on variables. For example, incomplete list of telephone contacts of respondents my hinder data collection if the only way to reach them is through telephone. Therefore, suppose the study was about product development, then incomplete telephone contacts for customers can cause researchers to miss the required opinion on the product, that could have help in improving the product for customer retention and increased profits. Also, it reduces sample size, which in turn affects the representativeness of data collected. Proceeding with analysis as though there was no missing values would then distort the results and mislead on inferences drawn from such. It is therefore good practise to reconstruct the missing values using methods that would reliably replace the " missingness " with values that are as good as observed. Secondly, in some situation, there could be very little information on important variables of study with no alternative means of collecting the information again, perhaps due to limited allocation of resources, yet we still have to proceed with the study. So, imputing the missing values would be the best option to pursue.

So, embarking on studying about missing data helped us mitigate it's problems

and offered solutions that were specific to the severity of the amount of missing data.

1.4 Outline of Thesis

The thesis is arranged as follows. Chapter 2 describes the causes of missing data and both traditional and modern methods of handling missing data. Various advantages and disadvantages of these methods of handling missing data are also presented. Chapter 3 describes our proposed method. Section 3.1 provides the descriptions and sources of data, and Section 3.2 describes the procedures that are used in collecting our samples. Section 3.3 gives details of the analytical software that are employed in the project. Section 3.4 discusses in details our analysis procedures and models. Section 3.5 presents the assumptions on which the analyses are based. Chapter 4 provides the results of the analyses. Chapter 5 discusses the imputation models, namely; PMM, Bayesian linear regression, and linear regression non Bayesian in relation to the analyses. Chapter 6 concludes this thesis.

2 MISSING DATA AND METHODS OF HANDLING MISSING DATA

Missing data (MD) is defined as incomplete observations on variables that occurs when partial or no information is captured for variable(s). MD is likely to introduce more error in data analysis and results [2].

2.1 Types of Missing Data

Missing data can be classified into three main categories based on their causes. These include: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

Missing completely at random (MCAR) implies that the cause of the missing data is independent of both observed and unobserved data, and the MD just occurs completely at random. With this kind of missingness, data is still likely to be representative of the population and we may proceed with analysis and get favorable unbiased results. However, it is always hard to find this type of MD. For example, when a respondent drops out from the study due to relocation, this would be a rear cause of this type of missing data. On the other hand, missing at random (MAR) is where the missing data is due to other observed variables in the study and not the missing value itself. So, one can completely account for the missing data with observed variables. For example, when respondents from a particular race skip a question on their political believes. One can accurately predict what would have been the answer for the missing values by using answers that have been provided by this section of the respondent. Finally, missing not at random, also known as non-ignorable nonresponse, is where the cause of such missing data depends on the

unobserved data. That is to say, the missing value has a relationship with the reason why it is missing. For example, when an individual fails to answer a survey questions on depression because of their level of depression [3].

2.2 Imputation of Missing Data

In anticipation of missing data, plans should be put in place ahead of time to avoid missing data. These could include but not limited to; use of online data collection tools that would not allow skipping of questions, staying in communication with participants throughout the study period when possible, use of incentives such as gift cards to motivate participation, and study campaigns to highlight usefulness, purpose and ethical standards (of privacy and confidentiality) of the study amongst participants [3]. These steps may only reduce but not eliminate the menace of missing data. Thus, MD will still always occur, and imputing them is currently the only remedy.

Imputation is defined as a process where missing data are replaced with estimated values based on the available information. This can be done for just one specific missing value, (unit imputation), or for a whole part of data points, (item imputation). Data imputation usually helps retain the affected cases, which would have otherwise be deleted from analysis by most of the analytical software [4].

2.3 Methods of Imputations

Methods of imputing missing data can be broadly classified into two groups; traditional and modern methods.

2.3.1 Traditional Methods of Imputation

Traditional methods consists of complete case analysis, pairwise deletion, cold and hot deck methods, mean imputations, regression imputation, and stochastic regression imputation methods.

Complete case analysis, also known as list-wise deletion methods is an imputation method in which, analysis is only done on cases with complete data. Cases with missing data are omitted from the analysis. Since this method only uses part of the data, it may results into biased estimates of population parameters, and there is also loss of power in statistical tests of parameters. However, if small number of cases are lost, less than 5%, due to their omission from the analysis, then biases and loss of statistical power is likely to be inconsequential [1].

Table 1 shows a data set with cases with missing values represented by ".". Suppose logistic regression analysis is conducted to predict health status based on age and religion, under complete case analysis, all cases with missing values will be ignored. Therefore, only cases with Ids 00007 to 00010 will be included in the analysis.

Table 1: An illustration of complete case analysis and pairwise deletion methods of handling missing data.

Id	Age	Gender	Health	Religion
00001	.	Female	Good	Christianity
00002	40	Female	Bad	.
00003	20	Male	Good	.
00004	.	Male	Good	Judaism
00005	.	Female	Bad	Islam
00006	.	Male	Bad	Buddhism
00007	19	.	Bad	Christianity
00008	21	.	Good	Islam
00009	35	.	Bad	Christianity
00010	49	.	Good	Christianity

Pairwise deletion method, also referred to as available case analysis, is where cases are only deleted whenever there is missing information for the variable that is needed for a particular analysis. Otherwise such cases would be included for variables in which they have the complete information. For example, suppose gender and health status were used to predict age, then all cases are considered for analysis except for those with Ids 00007 to 00010 (see Table 1). This method is usually used in exploratory data analysis such as correlation analysis. Therefore, it is very difficult to asses its performance as the analysis does not provide the error variance of the parameters [1].

Hot deck method is where the missing value is replaced by similar measured value from the same data set. Cold deck method is similar to hot deck method except that the measured value used for imputation is obtained from another source, data set other than the one with the missing value.

Mean imputation method imputes the missing values with the mean of the complete cases of the variable involved. This method preserves the univariate sample mean, but reduces correlations between the imputed variables [4].

Regression imputation method is one that imputes the missing values by fitting the usual regression model based on the available information from other variables. The predict values are then used to imputing the missing cases. This method has its shortcomings as it may over overstate the missing values because imputed values are based on predicted values that always fall right on the regression line. As a result, imputed values do not reveal the uncertainty around the missing values [4].

Stochastic regression imputation method is an improvement of regression imputation. It works the same way as the regression imputation except that it accounts for the variance of the predicted incomplete values. This method adds an error term to the predicted value. The error term is randomly generated and follows a normal distribution with a mean of zero and a variance equal to the previous regression model. As a result, the parameter estimates from this model are unbiased. However, it is very difficult to adjust for the standard errors of the values generated by this method to compensate for the fact that the imputed values are just predictions about the true values. Therefore, the standard errors are quite small, and significance tests will have high rates of Type I error [5].

2.3.2 Modern Methods of Imputation

It is a great challenge to deal with missing values in multivariate data sets where data would be missing in a number of variables. Therefore, we need more advanced approaches to deal with such missing data because the single imputation methods are no longer sufficient.

Modern methods of imputing multivariate data can be categorised into two groups; joint modelling (JM) and fully conditional specification (FCS).

Joint modelling method assumes that the data set is multivariate normal, and the multivariate distribution of the missing data is known. Imputations are then selected from the conditional distributions of the missing values by Markov chain Monte Carlo (MCMC) techniques. MCMC simulates the entire joint posterior distribution of missing values to obtain simulated posterior true estimates [6]. Maximum likelihood estimation (MLE) falls is an example of a JM method.

Fully conditional specification (FCS), also referred to as multivariate imputation by chained equations (MICE), is a method that imputes missing data by specifying the multivariate imputation model on a variable-by-variable basis [6]. It is an alternative to JM whenever the multivariate distribution assumption is violated. An example include, multiple imputation (MI).

These modern methods of imputations are extremely superior to the traditional methods in that they produce unbiased estimates for data that are both MCAR and MAR. These methods also preserve all cases for analysis. Meaning, all cases are used in the analysis upon imputation of missing values.

Multiple imputation is an iterative imputation procedure that reproduces m com-

plete copies of the original data set, each imputed with different values. The difference in the imputed values accounts for the variability around the true values [7]. The m imputed data sets are then analyzed separately and their results combined for appropriate decision.

Multiple imputation is carried out in three stages. The first stage is called the imputation (fill-in) phase. In this stage, the missing data are substituted with estimated values to create a complete data set. The process is then repeated m times; therefore producing m data sets that are different from each other. The second stage is called the analysis phase. In this stage, each of the m complete data sets are analyzed separately using statistical methods of interest. The pooling phase is the final stage of the process. Here, the results obtained from the analysis stage are then combined together for inference. These three stages are illustrated in Figure 1.

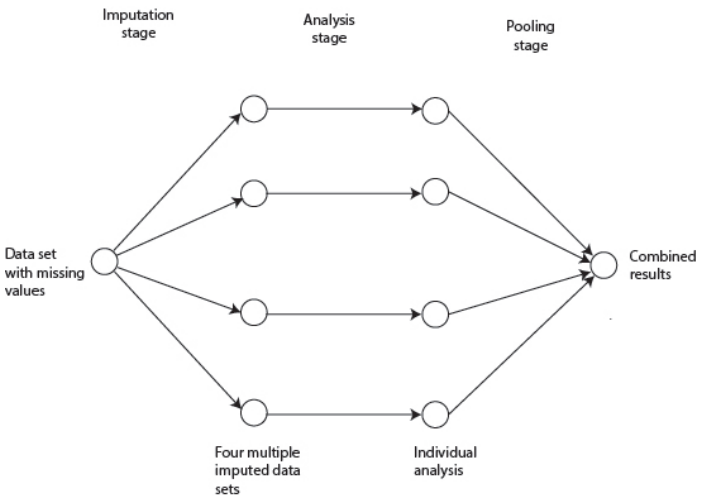

Figure 1: Steps for multiple imputation.

The multiple imputation procedure can be implemented in R using the MICE

package. R is a free software for statistical computing and graphics. It compiles on a variety of computer operating systems such as windows and UNIX.

This MICE package has built-in univariate imputation models that takes in complete predictors and returns a single imputation value for the incomplete targeted variable. The package can recognize three types of variables, namely: numeric, binary (factors with 2 levels), and categorical (factors with more than 2 levels) variables. Each variable type has a default imputation model. This package is able to check the choice of models specified versus the variable type for any mismatch. Some of the available models in the MICE package are shown in Table 2.

Table 2: Examples of imputation models in MICE R package.

Name of the Model	Model	Supported Variable Type
Predictive mean matching	PMM	Numeric
Bayesian linear regression	norm	Numeric
Two-level normal imputation	2l.norm	Numeric
Classification and regression trees	cart	Any
Linear regression non Bayesian	norm.nob	Numeric
Random forest imputations	rf	Any
Logistic regression	logreg	Factors with 2 levels

Predictive mean matching (PMM) is an imputation model, in the MICE R package, that is partially parametric and combines the ordinary linear regression method and the nearest neighbor imputation approaches. This method imputes missing values from the observed data preserving the distribution of the observed data in the missing data. It is a robust method in comparison to the completely parametric linear regression approach. However, it may not work well with small samples because it does not emphasize on between imputation variability with few predictors [6].

PMM imputes missing values by regressing incomplete variables on co-variates, producing a set of coefficients ($\boldsymbol{\beta}$). The procedure then draws at random, a set of coefficients ($\boldsymbol{\beta}^*$) from the distribution of $\boldsymbol{\beta}$. These $\boldsymbol{\beta}^*$ are used as new coefficients to generate predicted values for all cases in incomplete variables. In this case, predicted values are treated as metrics to identify complete cases with observed values that are close to the predicted values of each missing case of the target incomplete variable. Observed values of such complete cases are then used to impute the missing values. Those close cases are picked at random. By default, each missing case is matched to $k = 5$ completed cases with close predicted values [8].

Norm (Bayesian approach) model fills in the missing values using Bayesian linear regression, assuming that the data follows a multivariate normal distribution with a mean μ and covariance, Σ. It also assumes that the data are MAR. The method employs the data augmentation (DA) algorithm to simulate random draws from the population. For example, if data is represented as $\boldsymbol{D} = \{D_{obs}, D_{mis}\}$ for observed and missing data, respectively, DA produces MI's for for D_{mis} in two steps for each iteration. (1) I - Step (Imputation Step) simulates missing data based on the current parameter estimate. Initial parameters are estimated using EM (expected maximum likelihood estimate) of completed cases. Suppose we have current parameter values and imputed data as $(\mu^{(i)}, \Sigma^{(i)}, D^{(i)})$ at iterations $(i = 0, 1, 2, ...)$ and $(d_{J\ (obs)}, d_{j\ (mis)})$ as observed and missing of the j^{th} case of the data respectively. The missing values are imputed independently (in each case) by simulating $d_{j\ (mis)}^{(i+1)} \sim p(d_{j\ (mis)}|d_{j\ (obs)}, \mu^{(t)}, \Sigma^{(t)})$. (2) P -Step (Posterior Step) simulates parameter estimates based on the current imputed data. Parameters $(\mu^{(i+1)}, \Sigma^{(i+1)})$ are

drawn from the corresponding conditional posterior distributions $P\ (\mu, \Sigma\ |D^{(i+1)})$. These two steps are then repeated until convergence. Hence producing a Markov chain of imputed values and parameter estimates [9].

Linear regression, non-Bayesian approach, uses parametric linear regression analysis to impute the missing data. First, it fits a regression model given the observed data, by regressing target incomplete variables on covariate complete set. Then it uses the spread around the fitted line to predict values for the missing data points. This method does not emphasize on the variability of the predicted missing values because it does not randomize the regression coefficients. However, it is well suited for large sample sizes where variability is not a big concern and on data that follow a normal distribution.

Most of the imputation models in the MICE package of R are based on joint modeling approach [6], assuming that all variables follows a specific joint distribution, multivariate normal distribution. Basically, these methods are meant for imputing item nonresponse in surveys. However, most data from surveys exhibit grouping structures such as characteristics of individuals within regions, student performance within schools and so on, and these grouping characteristics should be accounted for in imputation and analysis of results [10]. As a result, two level linear model (2l.norm method) of the MICE package was developed to support and impute nonresponse in multilevel, clustered data. Multilevel imputation is still an area under study though, and it presents many challenges. Generally, the assumption of standard multivariate distribution of data is limiting since most of the studies involves variables of different types that may not necessarily follow the assumed distribution. Some studies involves

variables that take all manner of forms; continuous, categorical, logical and skip patterns thus making it extremely difficult to model. The 2l.norm method only works with continuous variables, and it allows a maximum of two levels and one class variable. The method requires that fixed effects, random effects, and the class variable be specified, and random effect and class variable be coded as a "2", and "-2", respectively in the predictor matrix. The method uses Gibbs sampler for the linear multilevel model, allowing for the within class error variance that is necessary for imputation of the multilevel data [6].

In a general perspective, Gibbs sampler method basically uses elementary properties of Markov chains to generate random variables from their marginal distributions without actually computing their densities. Therefore, it avoids complicated calculations. Consider a simple case, a pair of random variables (X, Y). In the Gibbs sampler method, samples are drawn from the distribution of $f(x)$ by sampling from the known distributions of $f(x|y)$ and $f(y|x)$. To achieve this, a sequence is generated for the pairs:

$$(X'_0, Y'_0),\ (X'_1, Y'_1),\ (X'_2, Y'_2), \ldots\ldots\ldots (X'_n, Y'_n) \tag{1}$$

Once the initial values, $Y'_0 = y'_0$, are specified the following values are generated by computing the conditionals

$$X'_i \sim f(x|Y'_i = y'_i)\ \ \&\ \ Y'_{i+1} \sim f(y|X'_i = x'_i) \tag{2}$$

For large values of n, the distribution of X'_i converges to $f(x)$, the true marginal distribution of X [11].

3 METHODOLOGY

Two different complete data sets are used in this project, namely; commercial properties data, and $2013-2014$ body measures (BMXH) data. The latter is considered as the primary data set, and commercial properties data is for verification and contrasting the results from BMXH data.

3.1 Data Source and Description

Commercial properties data is one of the learning tools (data sets) that come with the book, "Applied Linear Statistical Models", fifth edition, by Michael H. Kutner, Christopher J. Nachtsheim, John Neter and William Li. The data is collected by a commercial real estate company in order to evaluate vacancy rates, square footage, rental rates, and operating expenses for commercial properties in a large metropolitan area in order to provide her clients with quantitative information that can be used to make rental decisions. The data is collected from 81 suburban commercial properties. These properties are in prime locations: the newest, most attractive, and expensive [12]. There are five variables in this data. These include; age of property (X_1), operating expenses and taxes (X_2), vacancy rates (X_3), total square footage (X_4), and rental rates (Y).

On the other hand, data on the $2013-2014$ Body Measures (BMXH) is obtained from the National Health and Nutrition Examination Survey (NHANES). These data are collected from observational studies conducted by the Centers for Disease Control and Prevention of the National Center for Health Statistics (NCHS). NCHS is a branch of the U.S. Public Health Service in the U.S. Department of Health and

Human Services. The original data set consisted of 26 variables and 9,813 cases and targeted the entire U.S population. That is, it included all children, adolescents and adults aged between 0 years to 150 years old. This data can be accessed from the link https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Examination.

To obtain the data set used in the study, we first considered complete cases only, of the original data set. This gave us a data set of eight (8) variables and 7,157 complete cases of all children, adolescents, and adults aged between 8 years and 150 years old. Next, we employed systematic sampling to select every 15^{th} case of the 7157 complete cases to be included in the final data set for analysis. As a result, the final data set consisted of (8) variables and 477 complete cases. Systematic sampling is chosen over simple random sampling (SRS) because it is easier to work with, and it converges to the latter for large samples.

These variables included; BMI (Body Mass Index, KgM^{-2}, Weight (Kgs), Stand-Height (Standing Height in Cm), UpLegLength (Upper Leg Length in Cm), UpArm-Length (Upper Arm Length in Cm), ArmCirc (Arm Circumference in Cm), WaistCirc (Waist Circumference in Cm), and AvSAD (Average Sagittal Abdominal Diameter in Cm).

3.2 Sample Suitability and Procedures

Body measures data are obtained from all respondents except for sagittal abdominal diameter for pregnant women and individuals who weighed 600 pounds or more that was not measured due to their conditions. Appropriate care was taken to obtain measurements from individuals in wheelchairs. Trained health technicians were em-

ployed to collect and record body measures data at the Mobile Examination Centers (MEC). Arm and leg measurements were made on the right side of the body. Respondents who could not be measured this way due to amputation, medical conditions or appliance (cast) were measured from the left hand side. Due to disclosure concern, the data does not identify persons with amputations and measurements for weight do not include those individuals with limb amputation. There are no other groups of the respondents that were excluded from all other measurements [13]

We defined anthropometrics as measurements related to human body which include; actual stature, weight, and body measurements including skinfolds, girths, and breadths. These measurements are good indicators of human growth and the distribution of body fat.

In this project, body measures data (BMI, Weight, Standing Height, Upper Leg Length, Upper Arm Length, Arm Circumference, Waist Circumference, and Average Sagittal Abdominal Diameter) are used to evaluate the size, shape and composition of the human body.

Conventionally speaking, the main key measures for evaluating ones weight and health risks include; BMI, waist circumference, and risk factors for diseases and conditions associated with obesity. BMI is a measure of ones weight in kilograms divided by the square of height in meters. BMI itself is not a measure of someones health or fatness, it is an indicator of ones weight levels that is likely to cause health problems [14].

Similarly, waist circumference also predicts weight levels that would be associated with health risks. Much concentration of body fat around the waist rather than hips

would put someone at risk for heart disease and type 2 diabetes. Waist size greater than 35 inches or greater than 40 inches would raise an alarm for women and men, respectively. In addition, the following health risk factors have been associated with obesity, and are likely to pose greater risk for heart disease and other health condition for someone who is overweight or obese; high blood pressure (hypertension), high LDL cholesterol ("bad" cholesterol), low HDL cholesterol ("good" cholesterol), high triglycerides, high blood glucose (sugar), family history of premature heart disease, physical inactivity and cigarette smoking. Individuals who are overweight or obese and suffers from at least two of the risk factors, are always advised to embark on losing weight because even small weight loss (5 -10%) of the current weight would lessen their likelihood of suffering from excessive weight (obesity) related disease [14].

3.3 Software Implementation

SAS software was used to fit the regression model. However, R was used for the main part of the analysis, evaluating the various imputation models. The MICE package for R was used to implement FCS Approach. This package imputes incomplete multivariate data by chained equations. The R function prodNA was used to randomly delete specified percentage of values in a data set.

3.4 Analysis of Interest and Imputations

Once we obtained the final complete data set, we reproduced five copies of the data set. Using the R function, prodNA, 10%, 20%, 30%, 40% and 50% of the values in the five copies of the data set were randomly removed, respectively. As a result,

five data sets with different amount of missing values were created in addition to one complete data set.

In this project, we believed that anthropometrics and body measures data explains BMI. Therefore, our interest is to establish if there exists any relationship between BMI and anthropometrics and body measures. We treated BMI as the response variable and Weight, StandHeight, UpLegLength, UpArmLength, ArmCirc, WaistCirc and AvSAD as predictors. Using the original complete data set for the body measures, we embarked on a model building process to establish the best model that defines the relationship between BMI and other anthropometric measurements of the U.S. population. Forward stepwise regression method on all the potential seven predictor variables identified a model of BMI against the predictors weight; standing height; arm circumference, and waist circumference as the best model. Therefore, dropping other variables, namely; average sagittal abdominal diameter; upper arm length, and upper leg length. The forward stepwise model selection method was done with alpha to enter value of 0.10 and alpha-to remove of 0.05. The model was checked for multicollinearity, and variable "Weight" was dropped due to its strong multicollinearity with other variables in the model. Box-Cox transformation procedure is also used to transform the response variable (BMI) to improve the performance of the model, and to obtain a more random pattern of the residual plots, possibly for a more constant variance of the residuals. Therefore, the final regression model was found to be;

$$\sqrt{\hat{Y}} = 3.394 - 0.012X_1 + 0.0616X_2 + 0.0196X_3 \tag{3}$$

Where X_1 (standing height) in cm, X_2 (Arm circumference in cm) and X_3 (waist

circumference) in cm.

From the model, we estimate that for every unit increase in ones standing height, the square-root of the mean BMI decreases by 0.012 KgM^{-2} while holding arm circumference and waist circumference of the individual constant. Similarly, for every unit increase in ones arm circumference, we estimate that the square-root of the mean of the individuals BMI increases by 0.0616 KgM^{-2} while holding standing height and waist circumference constant. Finally, for every unit increase in ones Waist circumference, we estimate that the square-root of the mean BMI of the individuals will increase by 0.0196 KgM^{-2} while holding standing height and arm circumference constant.

Statistical F- test for the significance of the model concludes that the there is a regression relationship between $\sqrt{\hat{Y}}$ (square root of the estimated BMI) and the three predictors. The selected variables X_1 (standing height) in cm; X_2 (Arm circumference in cm) and X_3 (waist circumference) are also significant predictors in the presence of other predictors.

The estimate model parameters are stable and do not have problems with multicollinearity. The model is also reasonable with a good predictive ability. This is because, the predicted residual sum of squares (PRESS) and the error sum of squares of the model are very close and comparable as shown in Table 3.

Table 3: BMI model validation.

Sum of Residuals	0
Sum of Squared Residuals	8.21126
Predicted Residual SS (PRESS)	8.37773

The model performs well. Since 95.68% of the sample variation of the square root of the mean BMI is explained by the model, adjusting for the number of model parameters indicating the model fits the data well.

The MI models in the MICE package of R software is used to impute the missing values in the five incomplete data sets. There are different MI models in the MICE package for imputing missing values. However, three of them that support continuous data type were employed, namely; predictive mean matching (PMM), Bayesian linear regression (norm), and linear regression non Bayesian (norm.nob). The incomplete data sets, with different fractions of missing information (FMI), were imputed separately using the three imputation models to obtain fifty multiple imputed sets. Regression analysis, assuming the functional form of Equation (3), was fit and model parameters estimated for each of the fifty completed sets.

MI models are then evaluated how they perform with the different FMI. This done by comparing the regression coefficients obtained from the originally complete data set by those obtained from data imputed by MI models. By fitting a regression model as mentioned above, on each of the fifty imputed data sets for each of the three MI models (PMM, norm, and norm.nob), we generate a sampling distribution of fifty estimated regression coefficients at each FMI. So, treating estimated parameters as variables and each of the estimated regression coefficient as a data point for each of the variables, we can obtain the mean and the variance of those estimated regression coefficients as classified by FMI and MI models. The best performing model is presented as one that imputed data from which we obtained the least values for; variances, range, and percent deviation index (PDI).

The Equation for the PDI is given by

$$PDI = \left(\frac{Original\ reg\ coef - Mean\ of\ estimated\ reg\ coef}{Original\ reg\ coef} \right) * 100 \qquad (4)$$

where "$Original\ reg\ coef$" is the value of the regression coefficient obtained from the original complete data set and "$Mean\ of\ estimated\ reg\ coef$" is the mean of the distribution of regression coefficients estimated from the imputed sets that are now complete due to the imputed values.

One sample $t-$tests are used to make statistical comparison between the mean of the estimated parameters and the corresponding original parameter values.

Using the commercial properties data, similar procedure was followed to develop a model for rental rates as the response variable and the other variables (age, operating, vacancy and total) as predictor variables.

The final model was established to be

$$\hat{Y} = 12.243 - 0.126X_1 + 0.401X_2 \qquad (5)$$

where X_1 is age of property and X_2 operating expenses and taxes.

Furthermore, similar analysis was performed on the commercial properties data, where additional five data sets were created by removing varying amounts of values ($10\%, 20\%, 30\%, 40\%$ and 50%) from the original data set. The three MI models (PMM, norm, and norm.nob) are then used to complete fifty multiple imputed sets that are now complete due to the imputed values. Finally, the MI models were evaluated by comparing the original regression coefficients in Equation (5) with those estimated by fitting the same regression Equation (5) on the imputed data sets.

3.4.1 Relative Efficiency and Imputations

Relative efficiency (RE) of an imputation is defined as how best the true population parameters are estimated by an imputation model. It depends on both the amount of missing information and the number (m) of imputations performed.

The equation for RE is given by

$$RE = \frac{1}{1 + \frac{\lambda}{m}} \tag{6}$$

where λ is the fraction of missing information and m the number of imputations [15].

For very small FMI then five imputations would be enough for a good RE. However, several imputations would be required for large FMI, in order to achieve adequate RE. In our case, same imputation power is desired for our models so as to provide a good basis for comparing the accuracy with which each of models can estimate the missing values. Therefore, same relative efficiency (RE) of 99% was established by producing fifty multiple imputations for all the imputation models across the five FMI. We see from Table 4 that the RE increases across the FMI as the number of imputations increases, and fifty multiple imputations produces adequate RE that minimizes the biasness of the imputed values. Fifty multiple imputations also produces sets of regression coefficients that are large enough and follows a normal distribution by central limit theorem (CLT).

3.5 Assumptions

We assumed that the missing data were MAR. We could account for the missing data using other observed variables. For the regression model, we assumed that

Table 4: Relative efficiency of the imputation models.

m/ FMI	10%	20%	30%	40%	50%
5	0.9804	0.9615	0.9434	0.9259	0.9091
10	0.9901	0.9804	0.9709	0.9615	0.9524
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	0.99800	0.9960	0.9940	0.9921	0.9901

the residuals were normally distributed with a mean of zero and constant variance. Consequently, responses for BMI and rentals rates followed a normal distribution.

We also assumed that the samples of the estimated model parameters from each imputation model, were independent, large enough, and followed a distribution that is normal. Measurements of the parameters were also on a continuous scale.

4 RESULTS

The multiple imputed data sets for body measures data, and commercial properties data are analysed separately using regression models of the functional forms established in Equations (3) and (5), respectively. Results obtained from the regression analysis are used to evaluate the performance of the three imputations models, namely; PMM, Bayesian linear regression, and linear regression non Bayesian.

4.1 Analysis of the Body Measures Data

The mean and variance of the sampling distribution of the estimated regression coefficients are obtained by performing regression analysis on the imputed data sets. One sample $t-$test are also performed on the sample distribution of the estimated regression coefficients. Results are then observed and compared to corresponding coefficients from the original data.

4.1.1 Evaluation and Indexing of Imputation Models

Following the regression analysis on the complete body measures data, the model parameters were established as shown below in Table 5. These are considered as the unbiased parameter estimates, and forms the benchmark against which we evaluate our imputation models of the MICE package.

Table 5: Parameters values of BMI model .

PARAMETERS	β_0	β_1	β_2	β_3
ACTUAL VALUES	3.394	-0.012	0.0616	0.0196

The mean of the estimated regression coefficients using the imputed data tends to be larger for data with large amount of imputed values while variances of these regression coefficients increases as the amount of imputed data increases. This is evident across the three imputation models; pmm, norm and norm.nob models as shown in Tables 6 to 11. Data with the smallest amount (10%) of the imputed values has smaller variance of the estimated model parameters compared to those with the largest amount (50%) of imputed values. There are cases when the mean of the regression coefficients are smaller for large amount of imputed values, such cases are bolded in Tables 6, 8 and 10. $\hat{\beta}_2$ has a negative relationship with the response variable (BMI) and its corresponding value from the imputed data decreases as the amount of imputed values increases. This is clear in all our imputation models.

Table 6: Estimated mean of the regression coefficient with PMM model. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using PMM model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.4368	−0.0123	0.0630	0.0193
20 %	3.5126	**−0.0127**	0.0604	0.0199
30%	3.5154	**−0.0130**	0.0612	0.0202
40 %	3.8332	**−0.0142**	0.0617	**0.0187**
50%	**3.6331**	**−0.0133**	**0.0612**	**0.0194**
ACT. PARAM	3.3937	−0.0120	0.0616	0.0196

Table 7: Estimated variance of the regression coefficient with PMM model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0012	0.0000001	0.0000014	0.0000001
20 %	0.0025	0.0000001	0.0000027	0.0000003
30%	0.0109	0.0000005	0.0000082	0.0000005
40 %	0.0154	0.0000009	0.0000079	0.0000007
50%	0.0184	0.0000009	0.0000133	0.0000014

Table 8: Estimated mean of the regression coefficient with norm model. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using norm model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.368	-0.0120	0.0627	0.0194
20 %	3.4466	$\mathbf{-0.0122}$	**0.0598**	0.0199
30%	3.4488	$\mathbf{-0.0127}$	0.0612	0.0205
40 %	3.5104	$\mathbf{-0.0122}$	0.06171	**0.0188**
50%	**3.4058**	$\mathbf{-0.0122}$	**0.0590**	0.0206
ACT. PARAM	3.3937	-0.0120	0.0616	0.0196

Table 9: Estimated variance of the regression coefficient with norm model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0008	0.0000001	0.0000008	0.0000001
20 %	0.0033	0.0000002	0.0000032	0.0000002
30%	0.0065	0.0000003	0.0000048	0.0000005
40 %	0.0172	0.0000009	0.0000101	0.0000007
50%	0.0266	0.0000012	0.0000089	0.0000012

Table 10: Estimated mean of the regression coefficient with norm.nob model. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using norm.nob model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	3.3713	-0.0112	0.0627	0.0194
20 %	3.4378	**−0.0121**	**0.0599**	0.0199
30%	**3.4281**	**−0.0126**	**0.0607**	0.0207
40 %	3.4875	**−0.0122**	**0.0618**	**0.0189**
50%	**3.3726**	**−0.0120**	**0.0597**	**0.0204**
ACT. PARAM	3.3937	-0.0120	0.0616	0.0196

Table 11: Estimated variance of the regression coefficient with norm.nob model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0014	0.0000001	0.000001	0.0000001
20 %	0.0021	0.0000001	0.0000025	0.0000002
30%	0.0065	0.0000003	0.0000044	0.0000004
40 %	0.0093	0.0000005	0.0000075	0.0000007
50%	0.0147	0.0000007	0.0000119	0.000001

Percentage deviation of the estimated regression coefficients is computed for every amount of the impute data. Largest overall deviation index of 3.5% is realised for regression coefficients estimated from data imputed by PMM model. Conversely, norm and norm.nob models produce coefficient estimates with the least overall deviation index of 0.8%, and 0.6%, respectively as shown in Tables 13, 14, and 15.

The deviation index also varies by the amount of imputed values under each of the imputation models. Table 13 shows the model that has small mean deviation index of 1.0% and 2.3% for 10% and 20% imputed values, respectively, and higher mean deviation index of 6.7% and 4.0% for 40% and 50% imputed data, respectively.

Generally, norm and norm.nob models produces relatively smaller mean deviation index for all the five amounts of imputed data as shown in Tables 14 and 15, respectively.

PMM tends to produce data with wider range of the estimated regression coefficients cross the variables, compared to those for norm, and norm.nob models as illustrated in Table 12.

Table 12: Range of the estimated model parameters, body measures data.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
PMM	0.3964	0.001848	0.00259	0.001486
norm	0.1424	0.000758	0.003586	0.001842
norm.nob	0.1162	0.000624	0.002986	0.001802

Table 13: Percent deviation index of PMM imputation model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	MEAN
10	1.3	2.6	2.4	-2.1	1.0
20	3.5	5.7	-1.8	1.8	2.3
30	3.6	8.3	-0.7	3.2	3.6
40	12.9	18.0	0.2	-4.4	6.7
50	7.1	10.7	-0.7	-1.0	4.0
MEAN	5.7	9.0	-0.1	-0.5	3.5

Table 14: Percent deviation index of norm imputation model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	MEAN
10	-0.8	-0.4	1.7	-0.8	-0.1
20	1.6	1.2	-3.0	1.8	0.4
30	1.6	5.9	-0.7	4.5	2.8
40	3.4	1.9	0.2	-4.2	0.3
50	0.4	1.5	-4.2	5.2	0.7
MEAN	1.2	2.0	-1.2	1.3	0.8

Table 15: Percent deviation index of norm.nob imputation model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	MEAN
10	-0.7	-0.1	1.8	-0.8	0.1
20	1.3	0.7	-2.7	1.5	0.2
30	1.0	5.1	-1.4	5.5	2.6
40	2.8	1.3	0.3	-3.7	0.2
50	-0.6	0.1	-3.0	4.3	0.2
MEAN	0.8	1.4	-1.0	1.4	0.6

4.1.2 Normality Tests of the Parameters

We verify the assumptions of the one sample $t-$tests. A set of fifty parameter estimates are generated by fitting the functional form of the of Equation (3) to the fifty multiple imputed data in each amount of the missing data. Therefore, by the central limit theorem, each set of estimated model parameters follows a distribution that is normal. The data was measured in a continuous scale and are independent from each other since each data was imputed from different data sets with varying amounts of missing values.

Q - Q plots for each distribution of the estimated parameters also fitted a straight line, a further proof that the distributions are normal. As a result all the assumptions of the one sample t test were satisfied.

There is strong evidence of normal distribution of the estimated regression parameters across the different amount of data imputed by PMM model see Figures 2, 3, 4 and 5. The normal Q - Q plots fit a straight line against the theoretical normal quantiles of these model parameters. Some points on the normality plots for $\hat{\beta}_1$ and $\hat{\beta}_3$, from data filled in by PMM, tends to lie off the fitted straight line as shown in Figures 3 and 5. However, this does not affect the normality of the respective data. The formal test of normality using the Shapiro-Wilk's test yields p-values greater than the significance level of 0.05 as seen in Tables 16, 17 and 18. Figures 6-13 are Q - Q plots for the regression coefficients estimated from data imputed by the norm and norm.nob models and one can see the assumption of normality is satisfied.

Table 16: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the body measures data sets under PMM model. The significance level of each of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10	0.9836	0.2456	0.7905	0.8786
20	0.4752	0.0909	0.5101	0.1492
30	0.4024	0.9884	0.6309	0.5944
40	0.4688	0.5294	0.5294	0.3972
50	0.6153	0.4379	0.7837	0.6118

Table 17: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the body measures data sets under norm model. The significance level of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10	0.608	0.1499	0.4318	0.3977
20	0.3073	0.08151	0.5711	0.7267
30	0.5251	0.4138	0.7206	0.1786
40	0.6258	0.3206	0.7085	0.4037
50	0.7402	0.4634	0.5171	0.668

Table 18: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the body measures data sets under norm.nob model. The significance level of each of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10	0.7514	0.4728	0.7277	0.473
20	0.3545	0.3564	0.3636	0.40309
30	0.4008	0.7858	0.3699	0.4258
40	0.7194	0.5528	0.9097	0.4959
50	0.7405	0.5097	0.332	0.3665

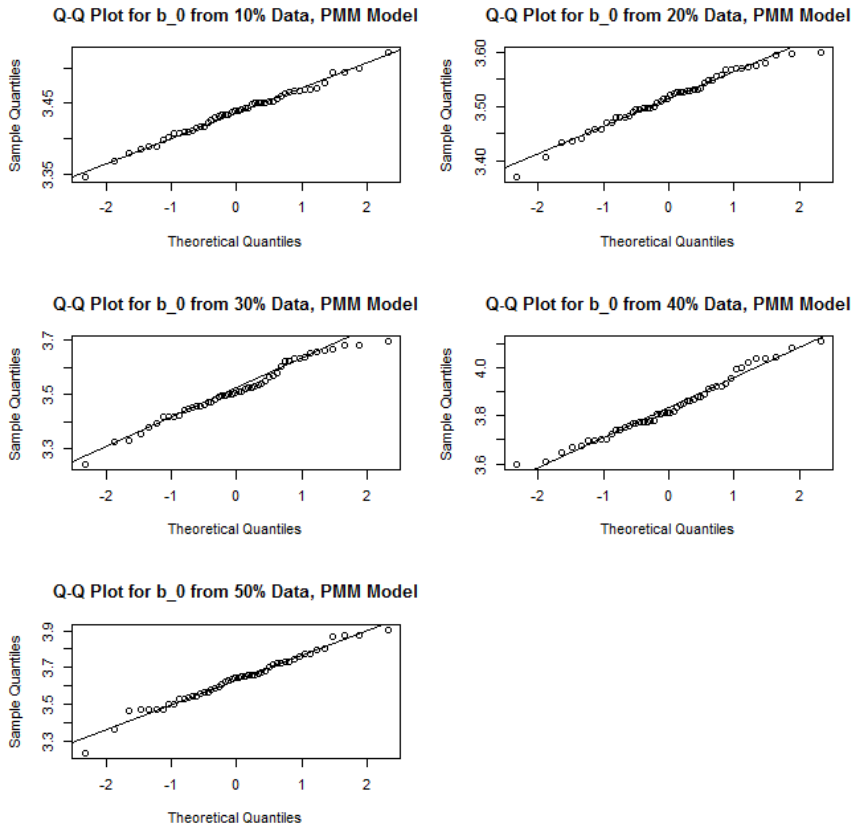

Figure 2: Normality plots of $\hat{\beta}_0$ from BMXH data imputed by PMM model.

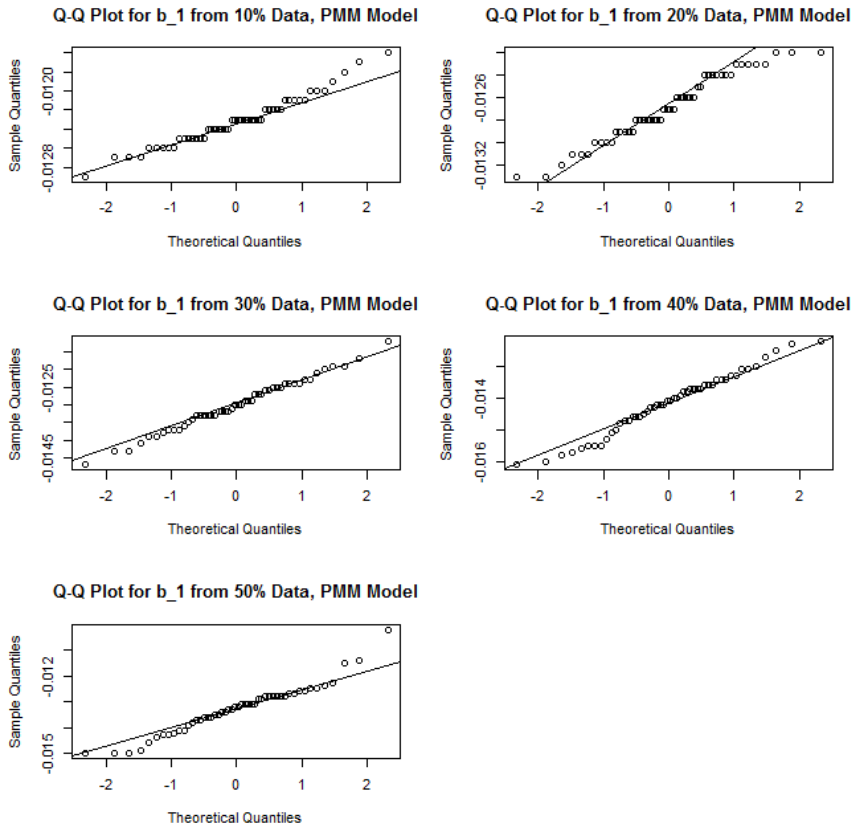

Figure 3: Normality plots of $\hat{\beta}_1$ from BMXH data imputed by PMM model.

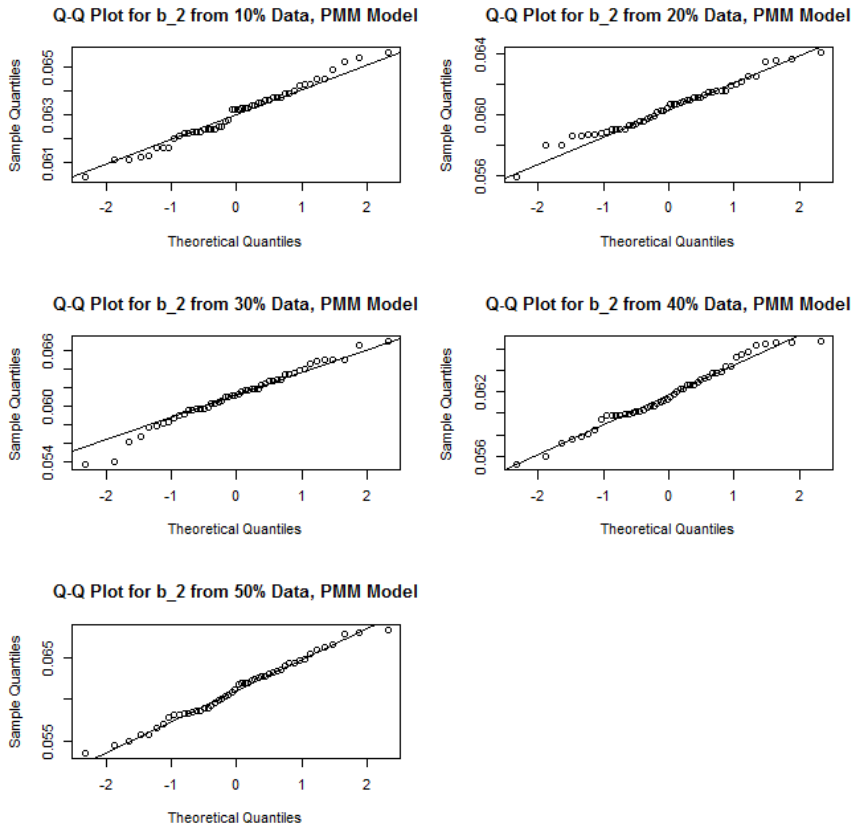

Figure 4: Normality plots of $\hat{\beta}_2$ from BMXH data imputed by PMM model.

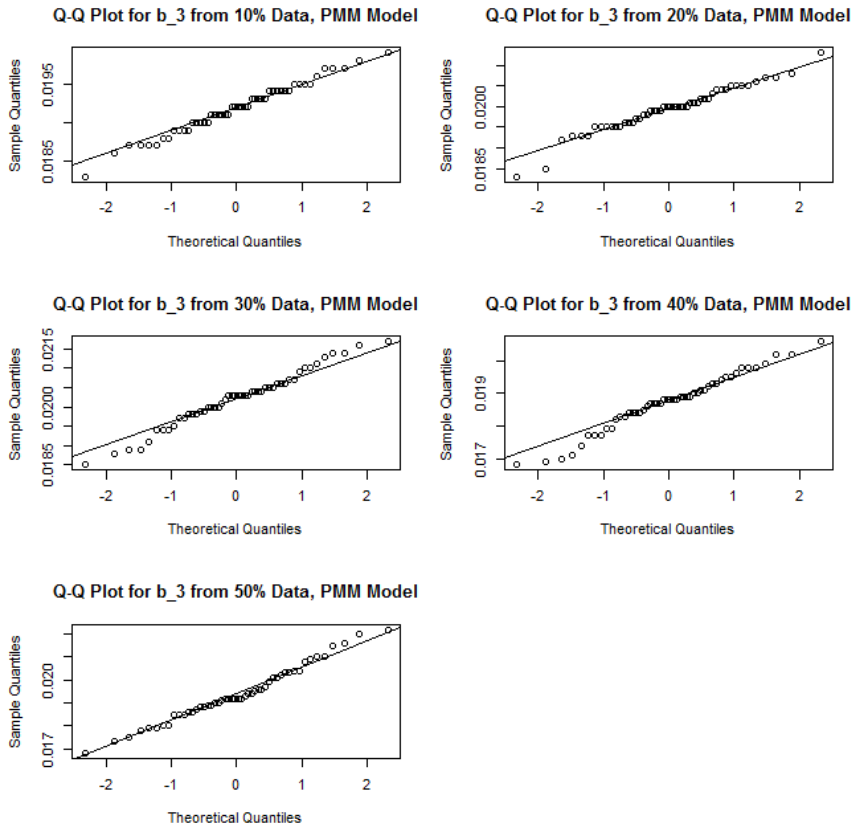

Figure 5: Normality plots of $\hat{\beta}_3$ from BMXH data imputed by PMM model.

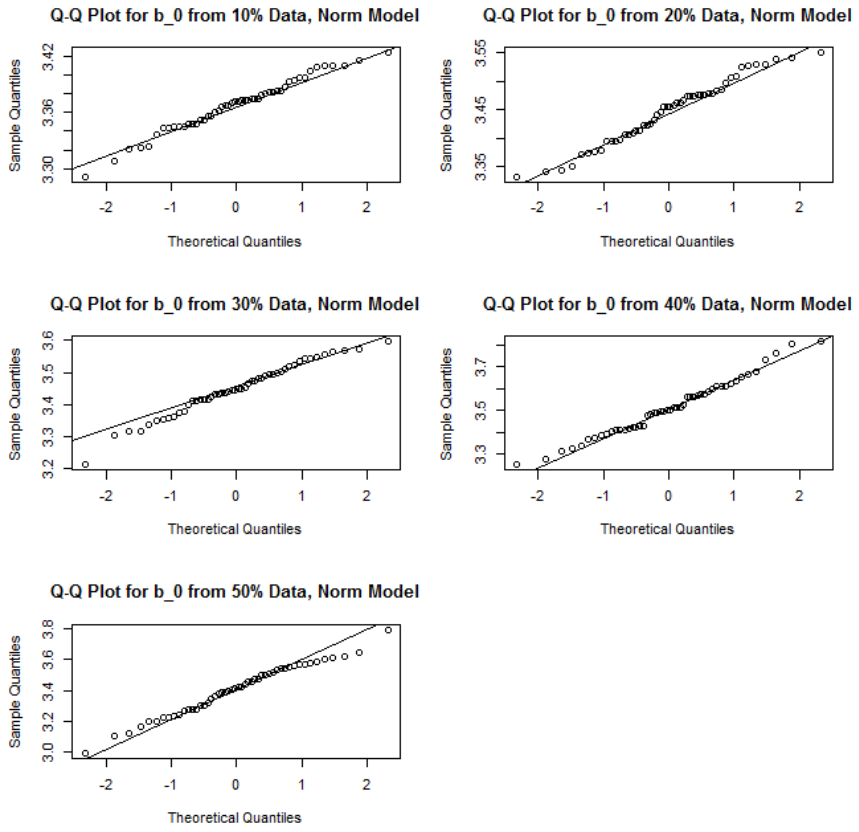

Figure 6: Normality plots of $\hat{\beta}_0$ from BMXH data imputed by norm model.

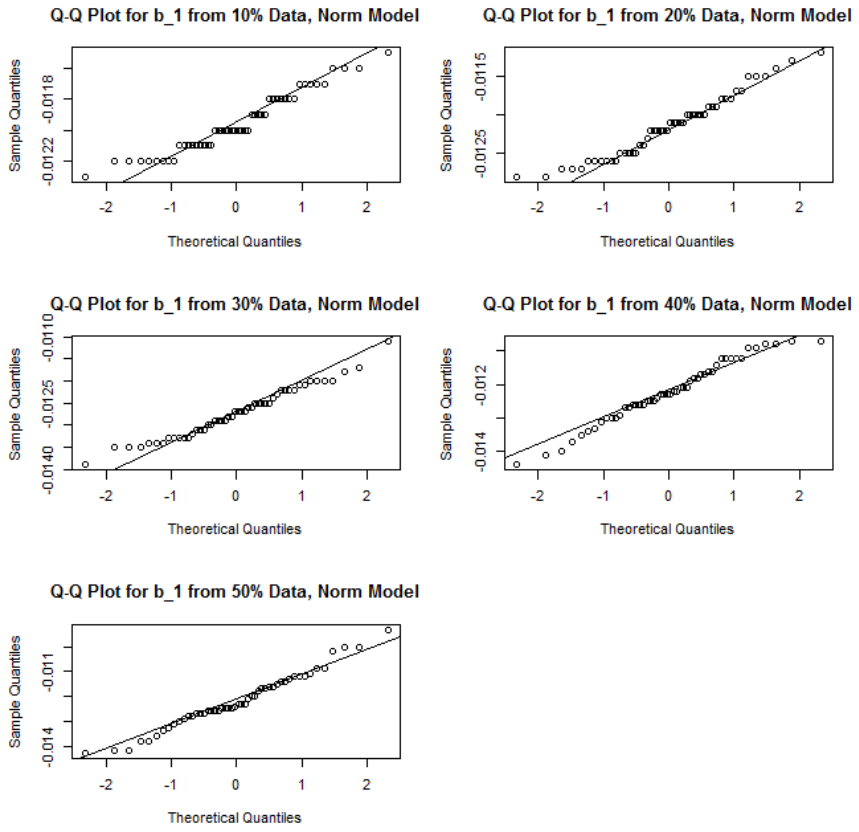

Figure 7: Normality plots of $\hat{\beta}_1$ from BMXH data imputed by norm model.

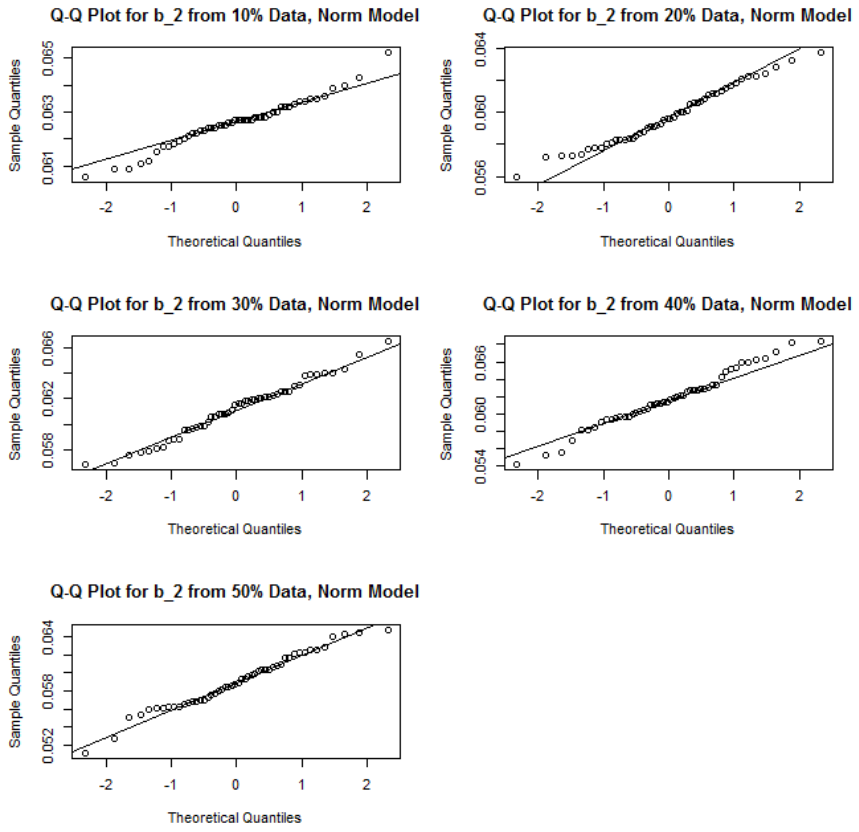

Figure 8: Normality plots of $\hat{\beta}_2$ from BMXH data imputed by norm model.

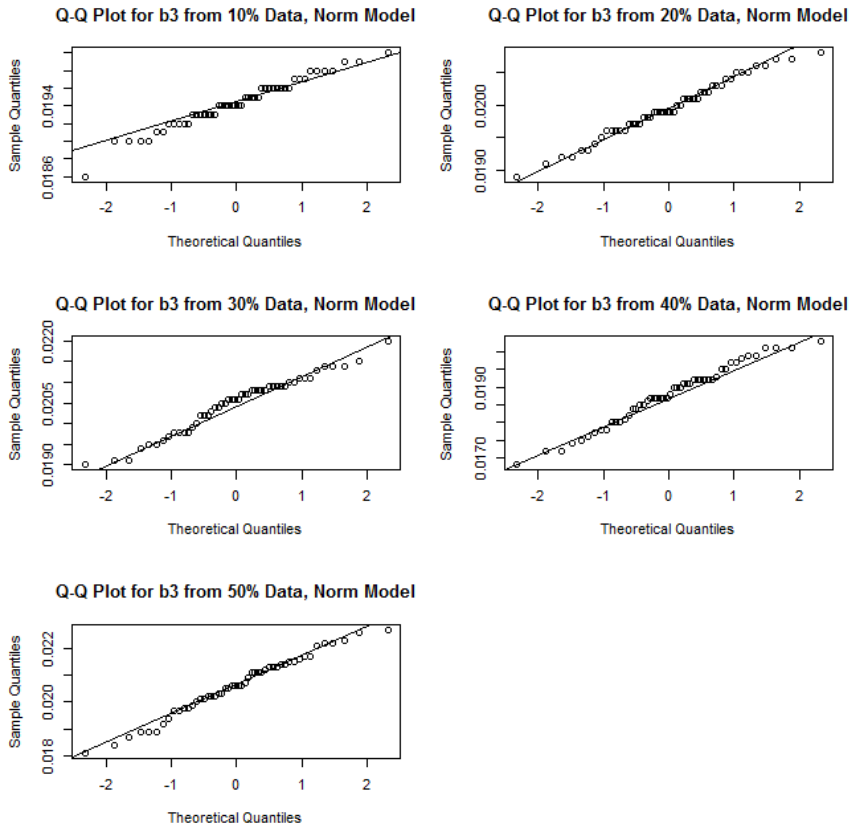

Figure 9: Normality plots of $\hat{\beta}_3$ from BMXH data imputed by norm model.

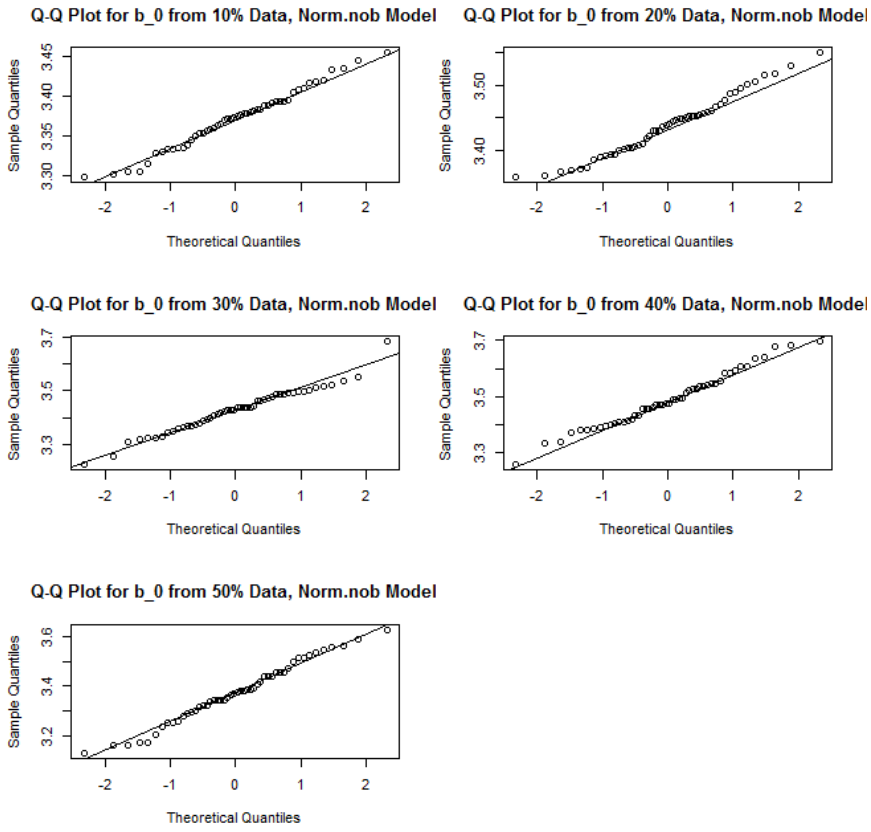

Figure 10: Normality plots of $\hat{\beta}_0$ from BMXH data imputed by norm.nob model.

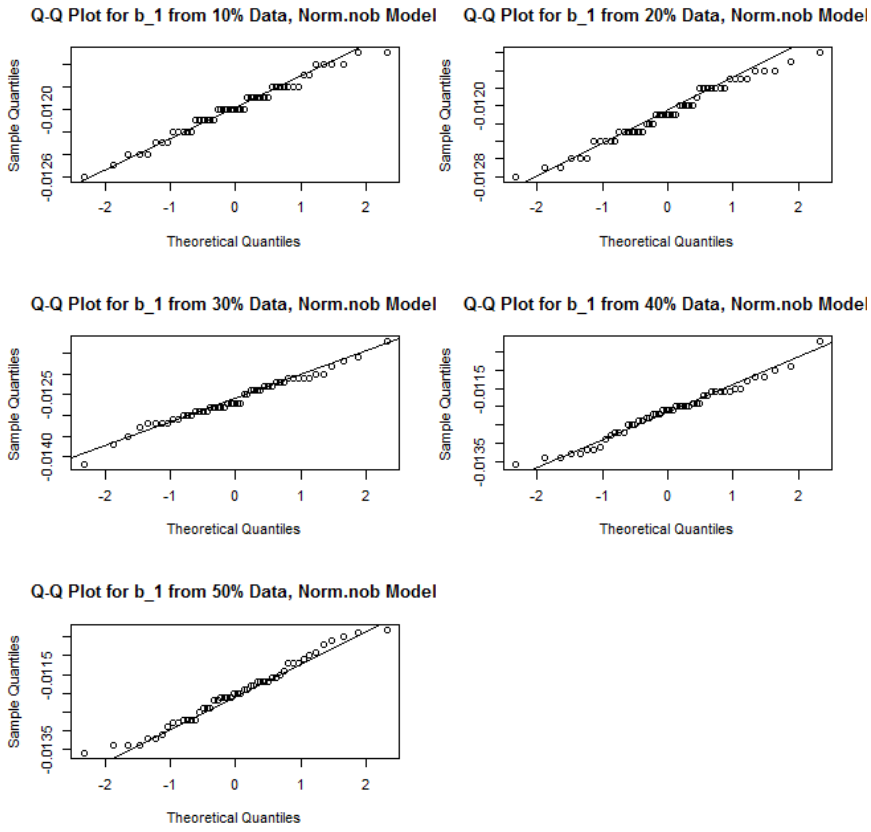

Figure 11: Normality plots of $\hat{\beta}_1$ from BMXH data imputed by norm.nob model.

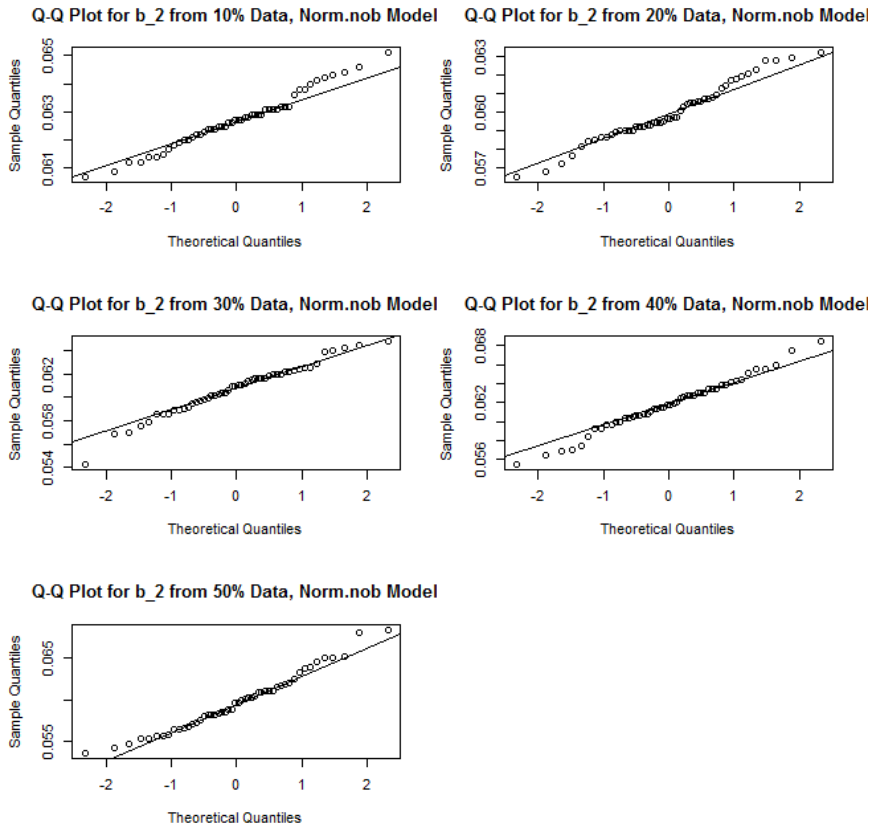

Figure 12: Normality plots of $\hat{\beta}_2$ from BMXH data imputed by norm.nob model.

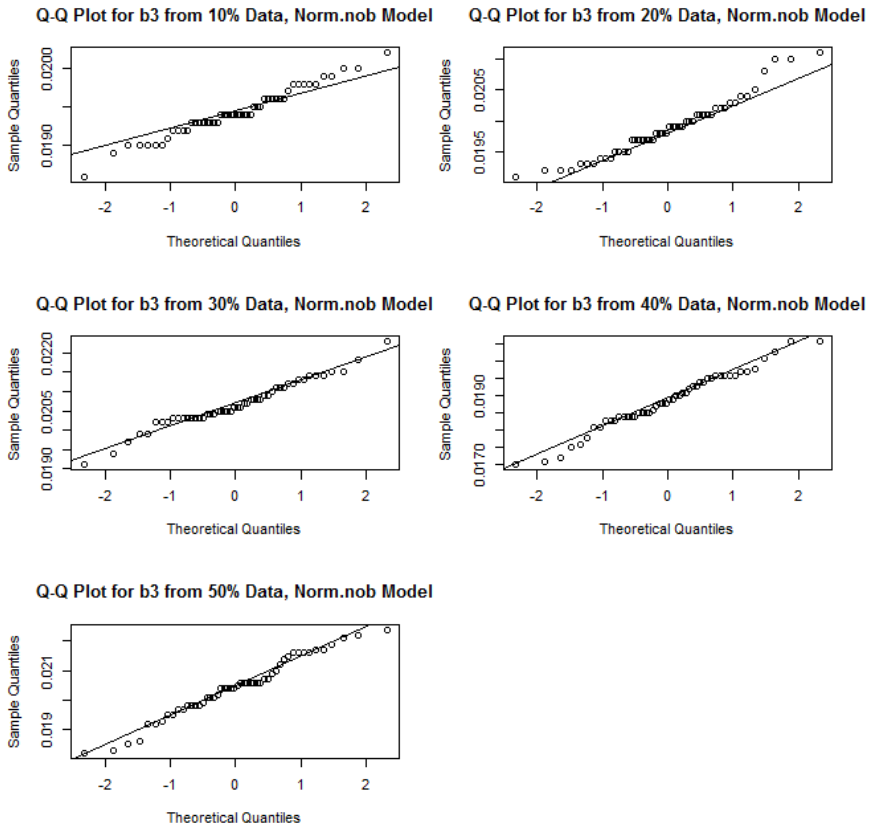

Figure 13: Normality plots of $\hat{\beta}_3$ from BMXH data imputed by norm.nob model.

4.1.3 Hypothesis Testing

The main question of interest is, "Are the estimated population parameters the same as the corresponding true parameters?" In order to do this, we are carrying out one sample $t-$tests on each distribution of the estimated parameters, where we compare the mean value of the estimated parameters and the actual value from the population. This is a case of multiple testing, where we are conducting a total of sixty, one sample $t-$tests, twenty from each of the three imputation models, namely; PMM, norm and norm.nob models. We are using 0.05 family level of significance, and the holm method is used to adjust for multiple testing, therefore, controlling the error rate for the family tests. Our test results show that all the estimated parameters are different from the true values except for a few estimated values that are not statistically significant. In most cases, our test is yielding adjusted p -values that are less than the 0.05 family level of significance. This is leading us to reject our belief that the estimated and actual values of the parameters are the same. Thus, the conclusion that the mean of the estimated regression coefficients are different from the corresponding actual coefficients.

For PMM model, only four of the twenty tests turn out to affirm our null hypothesis, while the remaining sixteen tests are significant at 0.05 family level of significance. The p-values of the tests that are not significant are bolded as shown in Table 19.

Table 19: Adjusted p-values under PMM model, BMXH data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0000000009	0.0000000017	0.000000001	0.00000000047
20%	0.00000000000	0.00000000000	0.00025655	0.00173232
30%	0.0000000034	0.00000000003	**1**	0.000007378
40%	0.00000000000	0.00000000000	**1**	0.000000158
50%	0.00000000000	0.000000000029	**1**	**1**

Similarly, only seven of the twenty tests under norm imputation model are not significantly different from the actual parameter values. See Table 20 for p-values of the regression coefficients that are not significant.

Table 20: Adjusted p-values under norm model, BMXH data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.0000017172	**1**	0.00000000801	0.00178342
20%	0.0000014689	**0.22185**	0.00000009922	0.000061451
30%	0.00039832	0.0000000007	**1**	0.0000000002
40%	0.0000030192	**1**	**1**	0.0000003055
50%	**1**	**1**	0.00000528	0.000001736

Finally, same parameters estimated from the data imputed by norm model, were also significantly different under norm.nob imputation model, except for the β_2 of the 30% imputed data. This parameter is significant under norm.nob imputation model.

Table 21: Adjusted p-values under norm.nob model, BMXH data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
10%	0.018975	**1**	0.0001318	0.010168
20%	0.0000001	**1**	0.000000000	0.0021924
30%	0.029016	0.0000000	0.00178342	0.00000000
40%	0.0000001	**1**	**1**	0.000004125
50%	**1**	**1**	0.0006305	0.0000123

4.2 Analysis of the Commercial Data

Commercial properties data is used to compare and contrast results obtained from the primary data, body measures data. Therefore, we conduct similar analyses as in the case of body measures data. Regression analysis is performed on the multiple imputed commercial properties data sets to obtain sample distribution of the estimated coefficients. The mean and variances of those distributions are observe and compared to the corresponding results of the body measures data.

4.2.1 Evaluation and Indexing of Imputation Models

We are using analysis results from commercial data to verify our findings in body measures data. We are drawing comparison of results in these two data sets. There-

fore, assuming that the commercial data was a complete enumeration of the targeted population of the newest suburban commercial properties, the parameters of the fitted regression model for the rental rates are indicated below in Table 22.

Table 22: Parameters values of the rental rates model .

PARAMETERS	β_0	β_1	β_2
ACTUAL VALUES	12.243	-0.126	0.401

When we introduce different amounts of missing values in the complete commercial data as described in Section 3.4, and impute the missing values using the same imputation models (PMM, norm, and norm.nob models) used in BMHX data, we see very similar patterns in the variation and values of the estimated parameters as established in the case of body measures data, refer to Tables 23-29. The mean of the estimated regression coefficients using imputed data, tends to be larger for data with large amount of imputed values. Variance of the estimated parameters increase with an increase in the amount of imputed values, except for those estimates for 40% imputed data which are smaller as shown in bolds in Tables 24, 26, and 28. Data with the smallest amount (10%) of imputed values has smaller variances of the estimated parameters compared to those with the largest amount of (50%) imputed values.

However, there are instances when the mean of the estimated coefficients are smaller for large amount of imputed values as shown in bolds in Tables 23, 25 and 27.

Rental rates model obtained from the commercial data has three parameters (β_0, β_1 and β_2). PMM model imputes data with a wider range of values for all the esti-

mated parameters when compared to the other models, namely; norm and norm.nob. This suggests a wider variance of parameters estimated from data imputed by PMM model compared to other models. β_1 estimated from data imputed by norm model has a wider range than those estimated from norm.nob model. β_0 and β_2 estimated from norm.nob model also have a wider range of values than for those estimated from norm model as shown in Table 29.

Table 23: Estimated mean of the regression coefficients with PMM model, commercial data. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using PMM model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	12.16928	-0.13726	0.422022
20 %	**12.10622**	-0.10595	0.403084
30%	**12.97448**	-0.06765	**0.277784**
40 %	13.60998	**−0.11894**	**0.251202**
50%	14.12138	**−0.08340**	**0.164752**

Table 24: Estimated variances of the regression coefficients with PMM model, commercial data. Variances that are represented in bolds have shown a decreased for large amount of imputed values.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	0.021337	0.000111	0.000276
20 %	0.037009	0.000123	0.000442
30%	0.333584	0.000734	0.005503
40 %	0.403421	**0.000683**	**0.003586**
50%	1.126236	0.001048	0.01553

Table 25: Estimated mean of the regression coefficients with norm model, commercial data. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using norm model.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	12.216694	-0.143662	0.422894
20 %	**11.998276**	-0.105292	**0.415206**
30%	12.674066	-0.075150	**0.321596**
40 %	13.754832	**−0.119122**	**0.240182**
50%	13.449068	**−0.114170**	**0.254226**

Table 26: Estimated variances of the regression coefficients with norm model, commercial data. Variances that are represented in bolds have shown a decreased for large amount of imputed values.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	0.066118	0.0001625	0.000892
20 %	0.072752	0.0002167	0.000912
30%	0.511143	0.0011833	0.009680
40 %	**0.387161**	**0.0011612**	**0.006029**
50%	1.659539	0.0017837	0.020153

Table 27: Estimated mean of the regression coefficients with norm.nob model, commercial data. Values that are in bold, represent the mean of the regression coefficients that are smaller for large amount of imputed values using norm.nob model

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	12.147182	-0.144922	0.430788
20 %	**11.98205**	-0.109018	**0.419564**
30%	12.58615	-0.079434	**0.334718**
40 %	13.860112	**−0.132242**	**0.24147**
50%	**13.381148**	-0.113166	**0.254308**

Table 28: Estimated variances of the regression coefficients with norm.nob model, commercial data. Variances that are represented in bolds have shown a decreased for large amount of imputed values.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	0.034467	0.000076	0.000429
20 %	0.044898	0.000137	0.000718
30%	0.37157	0.000725	0.006184
40 %	**0.303025**	**0.000547**	**0.003858**
50%	1.335885	0.001812	0.018512

Table 29: Range of the estimated model parameters, commercial data.

MI model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
PMM	2.015154	0.069608	0.25727
Norm	1.756556	0.068512	0.182712
Norm.nob	1.878062	0.065488	0.189318

The three imputation models; PMM, norm, and norm.nob, underestimated the missing values introduced in the commercial data by 12.2%, 8.1%, and 6.6%, respectively as shown in Tables 30, 31, and 32. Tables 30 and 31 shows MI models with smaller deviation from the true values for data with small amounts of (10 -20%) missing values. Table 30 shows that PMM model has a deviation index of 4.5%, and -5.5% for data with 10% and 20% imputed values, respectively.

Norm, and norm.nob model have relatively small deviations for data with large amounts (40 - 50%) missing values. Table 31 shows that norm model has an index of 11.1%, and 12% for 40% and 50% missing values, respectively. Similarly, Table 32 shows that norm.nob model has an index of 7.2% and 12.5% for 40% and 50% missing values, respectively.

Table 30: Percent deviation index of PMM imputation model using commercial data.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MEAN
10	-0.6	8.9	5.2	4.5
20	-1.1	-15.9	0.5	-5.5
30	6.0	-46.3	-30.7	-23.7
40	11.2	-5.6	-37.4	-10.6
50	15.3	-33.8	-58.9	-25.8
MEAN	6.2	-18.5	-24.2	-12.2

Table 31: Percent deviation index of norm imputation model using commercial data.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MEAN
10	-0.2	14.0	5.5	6.4
20	-2.0	-16.4	3.5	-5.0
30	3.5	-40.4	-19.8	-18.9
40	12.3	-5.5	-40.1	-11.1
50	9.9	-9.4	-36.6	-12.0
MEAN	4.7	-11.5	-17.5	-8.1

Table 32: Percent deviation index of norm.nob imputation model using commercial data.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	MEAN
10	-0.8	15.0	7.4	7.2
20	-2.1	-13.5	4.6	-3.7
30	2.8	-37.0	-16.5	-16.9
40	13.2	5.0	-39.8	-7.2
50	9.3	-10.2	-36.6	-12.5
MEAN	4.5	-8.1	-16.2	-6.6

4.2.2 Normality Tests of the Parameters

We have three parameters for the model of the rental rates of the commercial data. A set of fifty parameter estimates are generated under each model parameter for the five different amounts of imputed data. Therefore, by central limit theorem, the distribution of each set of parameter estimates follows a normal distribution. The parameter estimates are measured on a continuous scale, and each of the five amounts of imputed data of 10%, 20%, 30%, 40%, and 50% missing values are independent of each other. Therefore, our data satisfies the assumptions of one sample $t-$test.

We are constructing Q - Q plots to verify the assumptions of normality of each distribution of the parameter estimates. Figures 14-22 shows the Q-Q plots that fit a straight line indicating that our data is normal. Some of the Q - Q plots appear curved. However, the Shapiro Wilk's tests of normality yields p-values greater than the significance level of 0.05 as shown in Tables 33, 34 and 35. Therefore, confirming that in deed those values follow a normal distribution.

Table 33: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the commercial properties data sets under PMM model. The significance level of each of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10	0.0935	0.7661	0.06524
20	0.1221	0.07374	0.8352
30	0.5334	0.1585	0.5869
40	0.06991	0.9112	0.633
50	0.1638	0.2136	0.6124

Table 34: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the commercial properties data sets under norm model. The significance level of each of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10	0.1129	0.9154	0.2859
20	0.2176	0.1873	0.2692
30	0.4683	0.249	0.5067
40	0.3272	0.7018	0.9535
50	0.9015	0.0805	0.8415

Table 35: Shapiro-Wilk tests for normality of the sampling distributions of regression coefficients estimated from the commercial properties data sets under norm.nob model. The significance level of each of the individual tests is 0.05.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10	0.6639	0.3037	0.2044
20	0.3867	0.3208	0.08871
30	0.5826	0.6009	0.3268
40	0.1512	0.984	0.5789
50	0.2587	0.1073	0.123

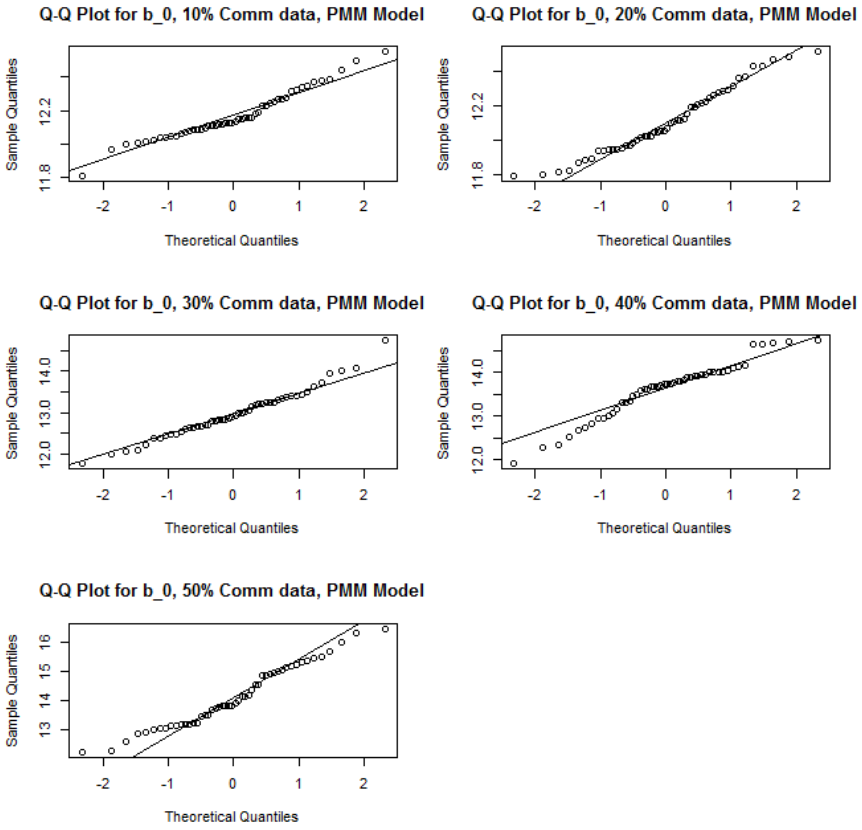

Figure 14: Normality plots for $\hat{\beta}_0$, commercial data, PMM model.

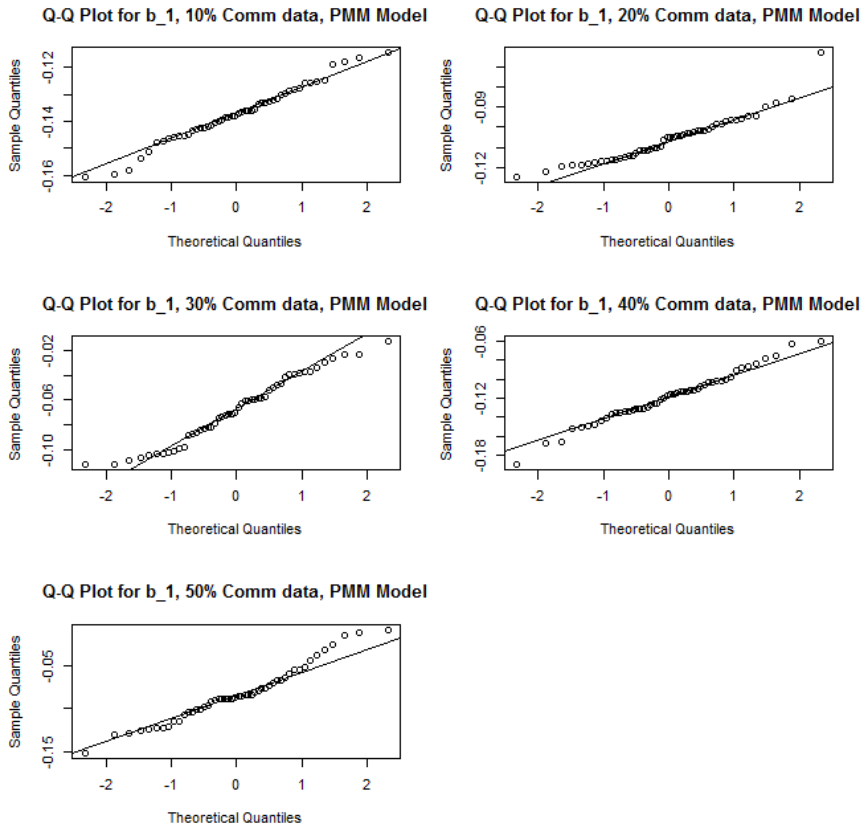

Figure 15: Normality plots for $\hat{\beta}_1$, commercial data, PMM model.

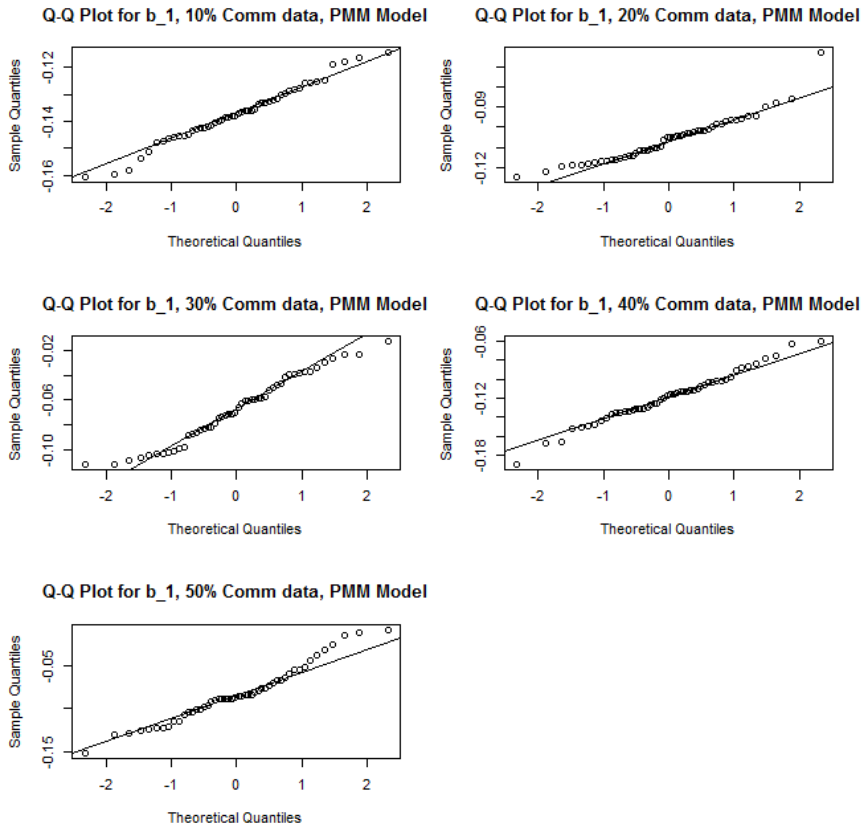

Figure 16: Normality plots for $\hat{\beta}_2$, commercial data, PMM model.

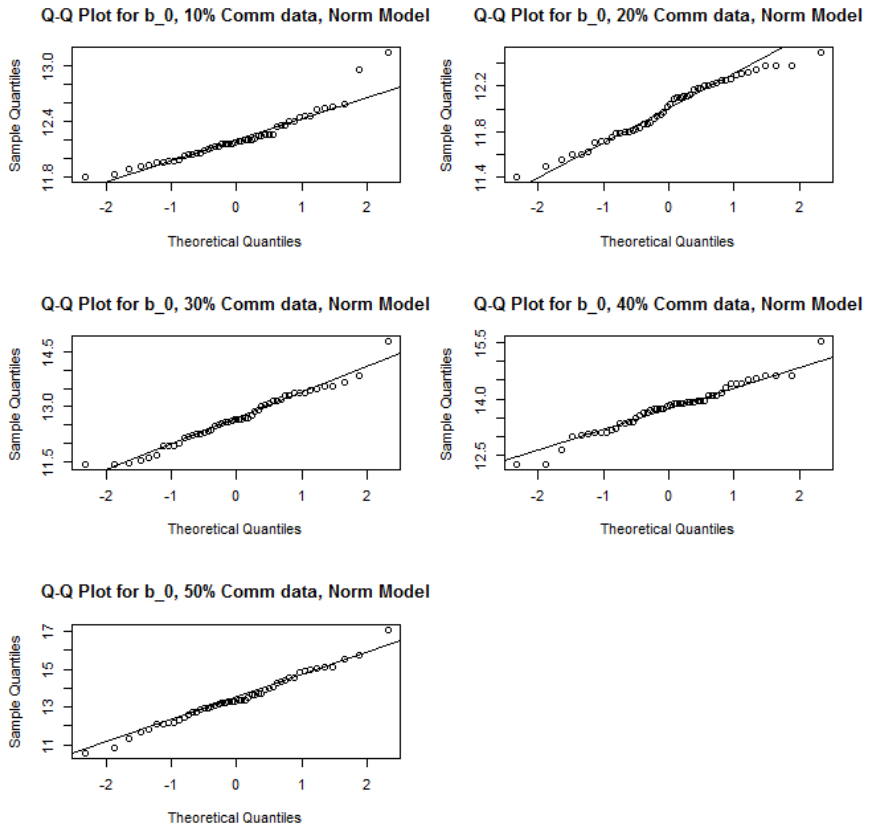

Figure 17: Normality plots for $\hat{\beta}_0$, commercial data, Norm model.

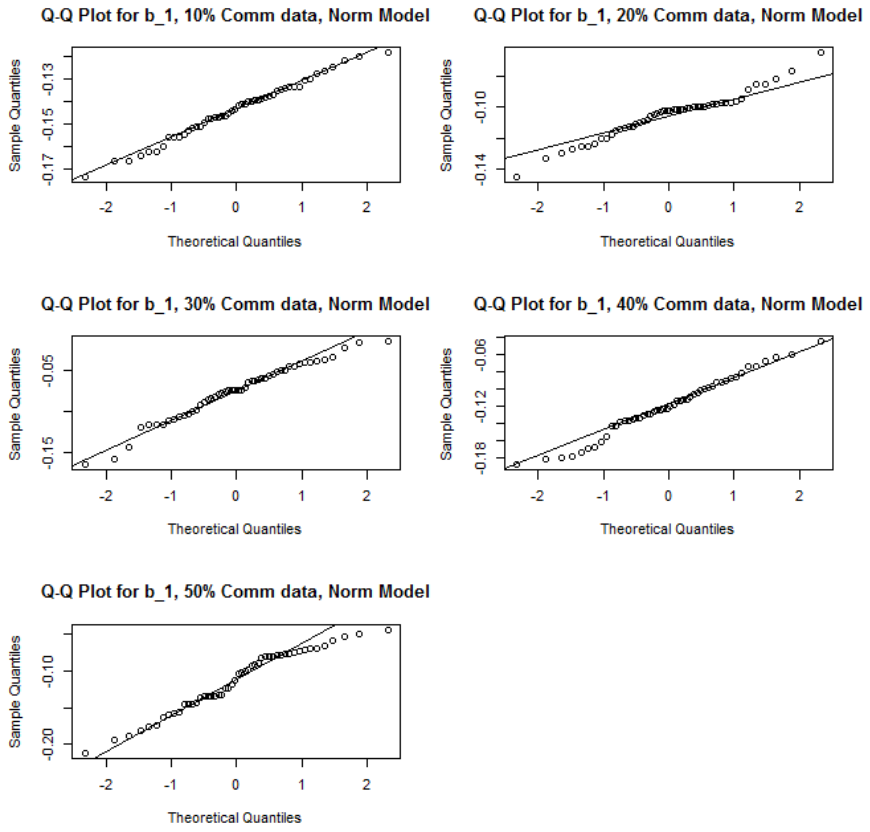

Figure 18: Normality plots for $\hat{\beta}_1$, commercial data, Norm model.

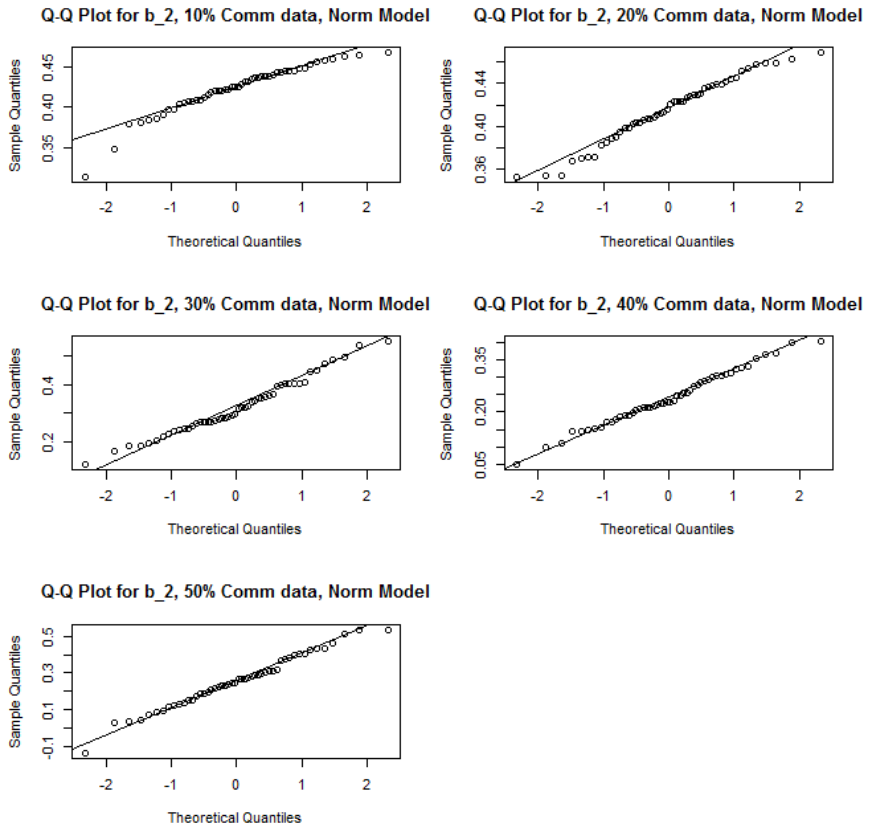

Figure 19: Normality plots for $\hat{\beta}_2$, commercial data, Norm model.

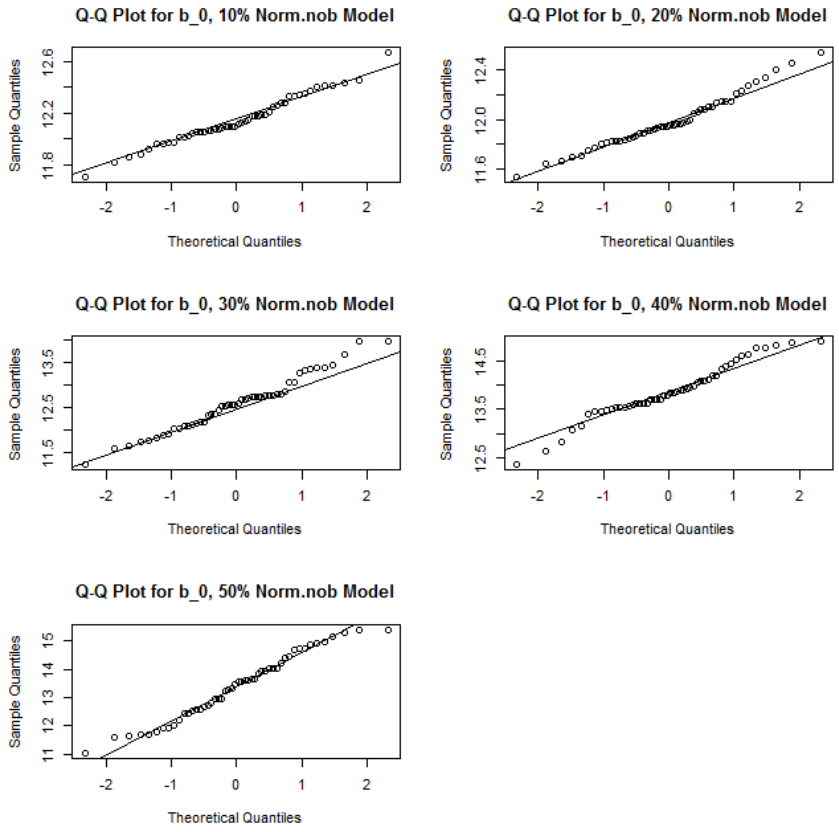

Figure 20: Normality plots for $\hat{\beta}_0$, commercial data, Norm.nob model.

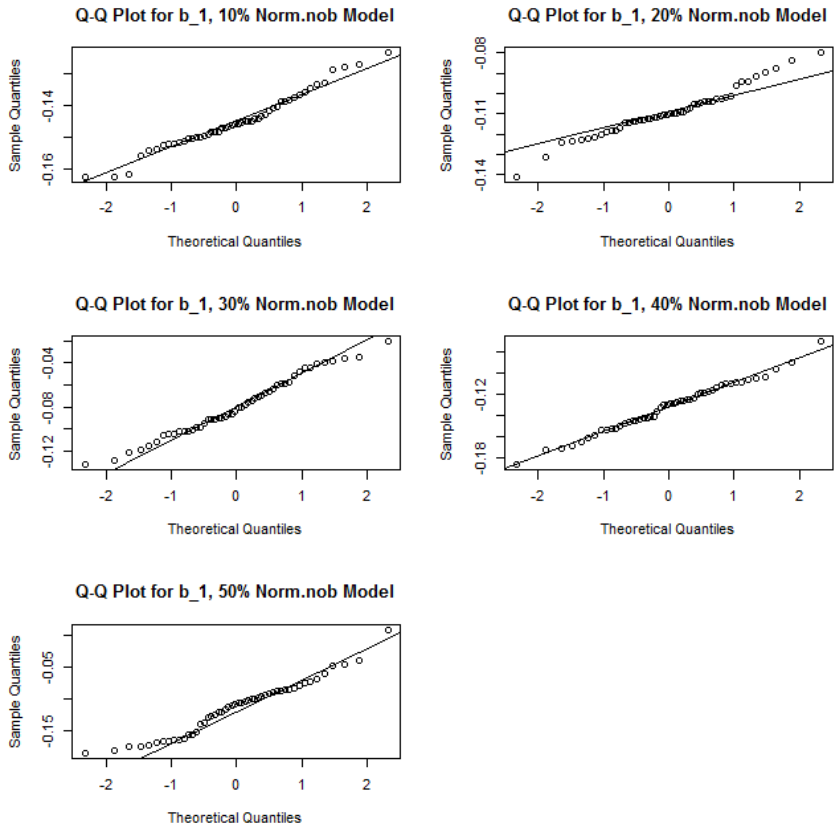

Figure 21: Normality plots for $\hat{\beta}_1$, commercial data, Norm.nob model.

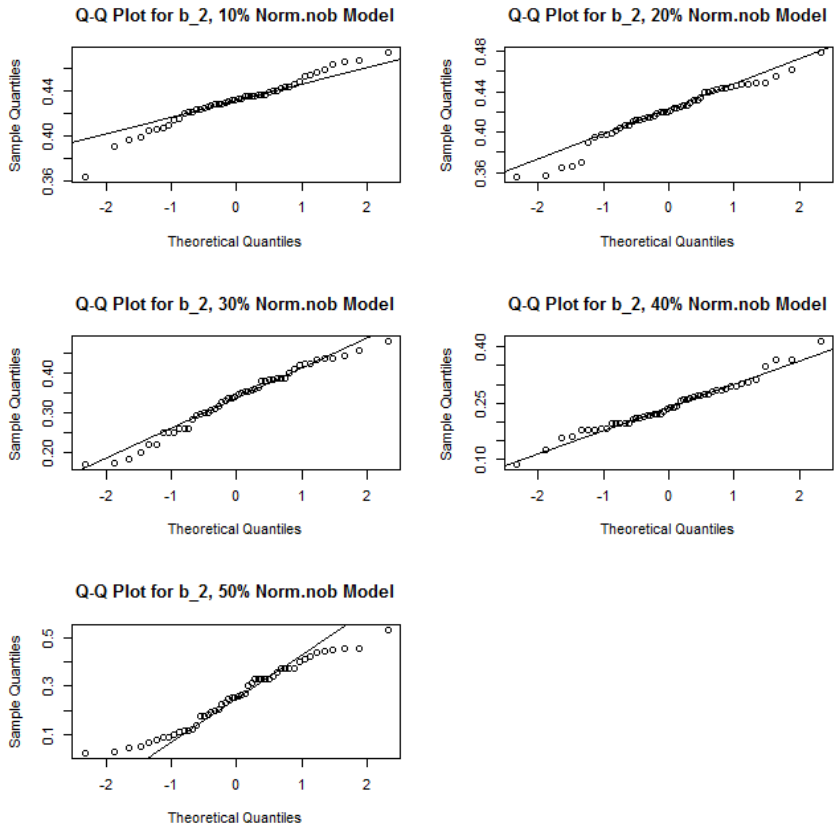

Figure 22: Normality plots for $\hat{\beta}_2$, commercial data, Norm.nob model.

4.2.3 Hypothesis Testing

Considering the commercial data, we are still interested in establishing if the mean of the estimated model parameters are the same as the actual parameters. We proceed by performing one sample $t-$tests on the distributions of the estimated model parameters. We have three parameters on the commercial data, namely; β_0, β_1 and β_2. So, we perform a total of 45 one sample $t-$tests. The p-values are adjusted for multiple testing once again using Holm's method. At the 0.05 family level of significance, only a handful of tests are not significantly different from the actual parameters. For example, for data imputed by PMM model, only two tests out of the fifteen tests on data are not significant as shown in bold in Table 36.

Table 36: Adjusted p-values under PMM model, commercial data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	0.0073323	0.00000002	0.0000000
20%	0.00009862	0.00000000	**0.9458**
30%	0.00000000	0.00000000	0.00000000
40%	0.000000000	**0.31956**	0.000000000
50%	0.000000000	0.00000000	0.00000000

We have three tests and two tests for data imputed by norm model and norm.nob model, respectively, that are not significant at 0.05 family level of significance as shown in bold in Tables 37 and 38.

Table 37: Adjusted p-values under norm model, commercial data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	**0.9458**	0.0000000	0.0000619
20%	0.00000096	0.00000000	0.0133920
30%	0.00109656	0.00000000	0.0000106
40%	0.00000000	**0.4797**	0.0000000
50%	0.00000049	**0.31956**	0.0000001

Table 38: Adjusted p-values under norm.nob model, commercial data. The adjusted $p-$values that are in bold are for one sample $t-$tests that are not significant at $\alpha = 0.05$ family level of significance.

% IMPUTED	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
10%	0.00637	0.00000000	0.00000000
20%	0.00000000	0.00000000	0.00014248
30%	0.002497	0.00000000	0.00000459
40%	0.00000000	**0.31956**	0.00000000
50%	0.00000015	**0.26656**	0.00000002

5 DISCUSSION

It is observable, in both body measures and commercial properties data, for all the imputation models, namely; PMM, norm, and norm.nob, that in most cases, the regression coefficient are estimated with larger mean and variances as the amount of the imputed data increases. This could be attributed to the fact that data with small amount of missing values are rich in actual information than those with large amount of missing values, that are used to accurately predict the missing members. The problem with data with large amount of missing values is that during the imputation process, the already imputed values are revisited many times, and used again to complete the yet to be imputed values. Therefore, many imputed values are used again to complete the imputation process. This may compromise the accuracy of the imputation process, especially when the imputed values that are re-introduced in the imputation process were not accurate.

Predictive mean matching model imputed values that are more way off the true missing values compared to the other two imputation models, namely; Bayesian linear regression and linear regression non Bayesian. This is evidence by the higher overall percentage deviation index of the regression coefficients estimated from the data imputed by this model. However, PMM impute much more accurate values for data with small amount (10%) of imputed values than for data with large amount ($> 10\%$) of imputed values. Therefore, it's overall percentage deviation index is the largest due to higher deviation for large amount of imputed values. Overally, PMM also imputed values with large variances and range.

Performances of the imputation models is also influenced by the nature and di-

mension of the data sets. All the models were much more stable and imputed values much more accurately for BMX data than for commercial data. This is due to the difference in the dimensions of these data sets. BMXH data has a total of 8 variables each with 477 data points compared to 5 variables and 81 data points for commercial data. All the models predicted higher values for BMXH data and lower values for commercial data. This is because BMX data offered much more information than commercial data.

Norm and norm.nob models produced better estimates for all amounts of imputed data with BMXH data than commercial data. Regression coefficients estimated from BMXH data filled in by these two models have smaller percentage deviation index, variances and range. Following the one sample $t-$test, norm and norm.nob models also have more estimated regression coefficients in BMXH data than commercial data, that are not significantly different from the corresponding true coefficients. Therefore, these models are much more likely to generate better imputations with larger data sets than with smaller data set.

6 CONCLUSION

Appropriateness of an imputation model depends on the nature and dimension of the data set, and amount of missing information. Bayesian linear regression and linear regression non Bayesian methods produce better results with larger data sets for all amounts of missing data of up to 50%. These two models also produces stable results for for larger amounts of missing value for low dimensional data.

Predictive mean matching model produces better results than the two models mentioned above for low dimensional data with small amounts of missing values. The model also had good results for high dimensional data with small amount of missing values.

As part of our effort to solve the problems of missing data, we look forward to extending similar analysis to data with different variable types such as binary (factors with 2 levels), and categorical (factors with more than 2 levels) variables. We also intend to develop a metric chart that would determine the exact amount of missing values versus dimensions of data within which each of the MI models in MICE package produces optimal results. This would guide the choice of appropriate MI models in MICE package for different amount of missing data.

In addition, there are other packages in R that are used for imputing missing data. These include; missForest, Hmisc, Amelia, and mi. Therefore, we are interested in comparing their performance against each other in the future.

By the end of the study, we identified predictive mean matching model of the in MICE package as most appropriate for imputing small amounts of missing values in both low and high dimensional data sets. Bayesian linear regression and Linear

regression models were also identified as most appropriate for large amount of missing values. We hope these findings will inform future handling of missing data.

BIBLIOGRAPHY

[1] John W. Graham. Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, 60:1-741, 2009.

[2] SPSS. Missing data, the hidden problem. *Whitepaper Central*, page 8, 2009.

[3] Wikipedia. Missing data. `https://en.wikipedia.org/wiki/Missing_data`, December 29, 2016. [Online; accessed October 17, 2016].

[4] Wikipedia. Imputation (statistics). `https://en.wikipedia.org/wiki/Imputation_(statistics)`, March 28, 2017. [Online; accessed September 17, 2016].

[5] Amanda N. Baraldi and Craig K. Enders. An introduction to modern missing data analyses. *Journal of School Psychology*, 48, Issue 1:537, February 2010.

[6] S. van Buuren and K. G. Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45, Issue 3, December 2011.

[7] Nicholas J. Horton and Stuart R. Lipsitz. Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *Taylor and Francis, Ltd. on behalf of the American Statistical Association*, 55(3):244–254, August 2001.

[8] Paul Allison. Imputation by Predictive Mean Matching: Promise and Peril. `http://statisticalhorizons.com/predictive-mean-matching`, March 5, 2015. [Online; accessed September 17, 2016].

[9] Yang C. Yuan. Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc*, March 2016. Rockville, MD.

[10] J. Drechsler. Multiple imputation of multilevel missing data rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, 40, 1:69–95, February 2015.

[11] George Casella and Edward I. George. Explaining the gibbs sampler. *Published by Taylor and Francis, Ltd. on behalf of the American Statistical Association*, 46(3):167–174, August 1992.

[12] John Neter Michael H. Kutner, Christopher J. Nachtsheim and William Li. *Applied Linear Statistical Models*. by McGraw-Hill/Irwin, Inc., 5th edition, 2005.

[13] Centers for Disease Control and Prevention. 2013 - 2014 Data Documentation, Codebook, and Frequencies, National Health and Nutrition Examination Survey. `https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/BMX_H.htm`, October 2015. [Online; accessed October 25, 2016].

[14] Aim for a healthy weight. *U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. National Institutes of Health, National Heart, Lung, and Blood Institute*, (NIH Publication No. 05-5213), August 2005.

[15] UCLA: Statistical Consulting Group. Introduction to SAS. `http://stats.idre.ucla.edu/sas/seminars/multiple-imputation-in-sas/mi_new_1/`. [Online; accessed March 13, 2017].

VITA

TOBIAS O OKETCH

Education:	B.S. Applied Statistics, Maseno University, Maseno, Kenya 2008 M.S. Mathematical Science, East Tennessee State University Johnson City, Tennessee 2017
Professional Experience:	Business Data Analyst Supervisor/ Productivity Monitor, Credit Reference Bureau of Africa Nairobi, Kenya, 2010–2014 Graduate Assistant, Center of Excellence in Mathematics and Science Education, East Tennessee State University, Johnson City, Tennessee, 2015–present
Skills:	Proficiency in analytical software such as R, SAS, SPSS, Minitab, GLPK, and Python Multivariate data analysis, Student behavior management and motivation