5-2021

Performance Comparison of Multiple Imputation Methods for Quantitative Variables for Small and Large Data with Differing Variability

Vincent Onyame
East Tennessee State University

Performance Comparison of Multiple Imputation Methods for Quantitative Variables for Small and Large Data with Differing Variability

---

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

Vincent Onyame

May 2021

Nicole Lewis, Ph.D., Chair

Robert Price, Ph.D.

JeanMarie Hendrickson, Ph.D.

Keywords: missing data, multiple imputation methods, quantitative data

ABSTRACT

Performance Comparison of Multiple Imputation Methods for Quantitative Variables for Small and Large Data with Differing Variability

by

VINCENT ONYAME

Missing data continues to be one of the main problems in data analysis as it reduces sample representativeness and consequently, causes biased estimates. Multiple imputation methods have been established as an effective method of handling missing data. In this study, we examined multiple imputation methods for quantitative variables on twelve data sets with varied sizes and variability that were pseudo generated from an original data. The multiple imputation methods examined are the predictive mean matching, Bayesian linear regression and linear regression, non-Bayesian in the MICE (Multiple Imputation Chain Equation) package in the statistical software, R. The parameter estimates generated from the linear regression on the imputed data were compared to the closest parameter estimates from the complete data across all twelve data sets.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Nicole Lewis for her guidance, support and encouragement throughout the pursuit of this thesis. I am also grateful to my committee members, Dr. Robert Price and Dr. JeanMarie Hendrickson for their suggestions and comments.

I extend my gratitude to my friends, Gaffar Solihu, Kazeem Kosebinu, Jett Burns, Makafui Azasoo, Ishmael Amartey, Dennis Quayesam and Dorothy Della for their support and encouragement.

My special thanks goes to my mum, Emelia Amaku Zotorvie and my siblings for their continue support and prayers for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Decision making is of one the spines of our existence and cuts across every facet of our lives. It relies on information, which comes collectively as data. It is worthy to note that data, which have been used one way or the other in decision making since the creation of earth has gained more prominence in the 21st Century. It has inevitably been the pivot on which many institutions, both private and government decisions are based on. The rapid growth in the field of information technology has also contributed to the demand for data in making informed decisions. This has led to many publications on the impact of data in various fields. In fact, in May 2017, the economist.com published an article titled "The world's most valuable resource is no longer oil, but data". The article explained that, data has dislodged oil as the most profitable resource on earth as the likes Facebook, Apple, Microsoft, and Amazon, which are the largest profitable companies in the world all being driven by data. However, most data used for analysis in real-life are incomplete. An incomplete data is called missing data and it occurs when some portions of the data are empty. Missing data arises due to many factors such as when an interviewer forgets to ask some questions during the interview process, when respondents do not answer some questions they perceive as personal, or unintentional data entry error. Missing data is a major problem nearly every researcher faces and it negatively impacts the outcome of the results. When an incomplete data is analyzed, it reduces the level of sample representativeness and creates unbiased estimates, consequently leading to incorrect inferences. Handling missing data is important as many statistical software used for

data analysis assume complete data even when the data is incomplete.

1.1 Proposed Work

In this study, we examined multiple imputation methods of handling missing data for quantitative variables. The study specifically, will apply the Predictive Mean Matching, Bayesian linear regression and linear regression non-Bayesian methods on incomplete small and large data sets with differing variability, and then compared to the complete data set determine the best imputation method.

1.2 Overview of Thesis

Chapter 2 describes the missing data mechanisms and introduces basic vocabulary. Chapter 3 describes the traditional methods of handling missing data along with modern methods.In Chapter 4, explanation is given on multiple imputation methods for quantitative variables. Chapter 5 describes the methodology. Chapter 6 describes our results. Chapter 8 concludes the thesis.

2 MISSING DATA MECHANISMS

Missing data manifests in two patterns. They are the monotone and non-monotone (or general). Monotone pattern occurs if for example the variables Y_j are ordered such that when Y_j is missing, then all variables Y_k with $k > j$ is also missing. An example is shown in table 1. An example of this can occur in longitudinal studies with drop-outs. If the pattern is not monotone, it is called non-monotone or general. An example is shown in table 2 where observation A_3 is called the latent variable because it is has empty data due to missingness [5].

Table 1: An Example of Monotone Missing Data Pattern

Observation	A_1	A_2	A_3
1	a_{11}	a_{21}	a_{31}
2	a_{21}	a_{22}	
3	a_{31}		
4	a_{41}		
5	a_{51}		

Table 2: An Example of Non-monotone Missing Data Pattern

Observation	A_1	A_2	A_3
1	a_{11}	a_{21}	
2	a_{21}		
3		a_{23}	
4	a_{41}		
5	a_{51}		

Missing data is divided into three types according to Rubin (1976) namely, miss-

ing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). He explained that MCAR occurs when we suppose that missing data exist on a particular variable Y and the probability of missing data on Y is not related to the value of Y itself or to the rest of the values in the data set. For MCAR, we label the observed Y values (Y_{obs}) and the missing Y values (Y_{miss}). We define R as the missing data indicator variable where 0 is missing and 1 is observed. The probability of finding a missing value on Y per MCAR is $P(R|\epsilon)$ where ϵ describes the existing relationship between R and the data. Graham (2009) added that, analyses on MCAR data results in unbiased estimates but are rather limited to loss of statistical power. An example of MCAR is when during a survey, a respondent accidentally skipped one or more of the questions.

Missing at random (MAR) is a less stringent assumption than MCAR. In this case, the data on the variable Y will be MAR if the probability of missing data on Y is not related to the value of Y after controlling other variables in the analysis. Its probability distribution is defined as $P(Y_{miss}|Y, X) = P(Y_{miss}|X)$ where X is the observed values and Y is the missing values. Here, the missing value Y depends on the observed X. An example of MAR is when people that are highly education fail to reveal their income compared to those that are not.

Missing not random (MNAR) is the common type of missing data among the three mechanisms. [8]. It occurs when the probability of missingness depends on the unobserved data. This means the probability of missingness is rather related to the missing values itself. The probability distribution is stated as $P(R|Y_{obs}, Y_{miss}), \epsilon)$.

The parameters are defined as R for missing data indicator, Y_{obs} for observed Y values, Y_{miss} for missing Y values and ϵ describing the relationship between R and the data. An example is when people with terminal illness drop out of a study.

The three missing data mechanisms are further classified into ignorable and non-ignorable. MNAR is called non-ignorable because when dealing with it, you have to include any information about the missing data while MCAR and MAR are both considered ignorable because when dealing with them, you are not required to include any information about the missing data [11].

3 HANDLING MISSING DATA

Various techniques have been applied in handling missing data. These techniques require the application of special methods to get desired accurate results. The methods are broken into two categories: traditional and modern.

3.1 Traditional Method

Several traditional methods have been applied in solving problems associated with missing data. We shall look at a few of these methods. In complete case analysis (CCA), the missing data are deleted from the variable and thereby making the data complete for analysis. It is easy to work with since statistical packages assumes complete data. Allison (2001) adds that, CCA is easy to work with since no special computation methods are required. If the data are MCAR, the reduction of sample is the random sub-sample of the original sample. That is, for any parameter of interest in the analysis, if the estimates are unbiased for the complete data set (with no missing data), then they will also be unbiased for the incomplete data set. Graham (2009) also noted that some loss of statistical power will occur after deletion because of the unused partial data. In some cases, this loss of power can be massive, making this method not the best choice. He also argued that, if the loss of cases due to missing data is small (e.g., less than about 5%), biases and loss of power are both likely to be inconsequential [8].

Pairwise deletion works by calculating the correlation for any pair of variables

using the available data. The paired data are compared and the cases are deleted whether missing or not. Graham (2009) explained that each correlation is estimated based on the cases having data for both variables but had issues with this method as different correlations (and variance estimates) are based on different subsets of cases. While the pairwise deletion is simple to use, it requires the data to be MCAR [8].

The arithmetic mean imputation replaces the missing values with the computed mean of that variable. This approach is limited in usage when outliers exist in the data set since the estimated mean will biased toward the outliers.
In hot deck imputation, one withdraws values from the observed values and uses the values to replace the values that are missing. After the values are drawn from the observed variable, a replacement is effected to enable the observed datum to be selected fairly to substitute the missing values. One huge drawback to this approach is that since values are drawn from the observed variable to replace the missing values, it may lead to variability in the variable with complete values being underestimated and will as a result lead to narrow intervals [12].

3.2 Modern Methods of Handling Missing Data

Traditional methods of handling missing data have several limitations, which consequently affects the true outcome of results. Due to this, researchers have developed more efficient ways of handling missing data. The modern methods are broken into two approaches: joint modeling (JM) and fully conditional specification (FCS). These approaches are superior to the traditional methods since they produce unbiased esti-

mates for missing data that are both MAR and MCAR.

3.2.1 Joint Modeling

The joint modeling (JM) approach to handling missing data requires that the missing data must be multivariate and thus follow a multivariate distribution. Thus, the JM works more efficiently if the missing data are multivariate normally distributed. JM is used to solve missing data on longitudinal and time-to-event data and provides appropriate estimate of treatment effects.[9]. After these assumptions are met, the Markov chain Monte Carlo (MCMC) or maximum likelihood estimation (MLE) techniques are employed to solve the missingness in the data. With JM, the probability distribution is found under MCMC and the parameters are estimated for the posterior probabilities. Schafer (1997) adds that if we repeatedly simulate steps of the chain, it simulates draws from the distribution of interest [13]. With the maximum likelihood estimation (MLE), parameter estimates of the JM can be based on the observed data [7].

3.2.2 Fully Conditional Specification

Most real-world missing data are not multivariate normal and so another approach is needed to deal with combating missing data. The fully conditional specification (FCS) also known as the multiple imputation chain equation (MICE) was first developed by Rubin (1977) as a method for handling missing data that does not require the assumption of multivariate normal (Rubin, 1977, 1978). Figure 1 illustrates the three stages of multiple imputation. The first stage called the imputation stage involves

the replacement of missing values with estimated values to create complete data. It is repeated m number of times to create m complete data. The next stage is the analysis stage where each of the m complete data in stage one is analyzed with a statistical method of interest. The final stage is called the pooling stage. The pooling combines the results analyzed in stage two to obtain a single point estimate. Multiple imputations are repeated random draws from the predictive distribution of the missing values. More precisely, multiple imputations are drawn from a posterior predictive distribution of the missing data conditional on the observed data [4]. Allison (2001) also adds that through multiple imputation (MI), one is able to solve the problems associated with single imputation by introducing another form of error based on the variation in the parameter that have been estimated across the imputation, which is called "between imputation error". It works by replacing each missing item with two or more ($m > 1$) acceptable values, representing a distribution of possibilities [3].

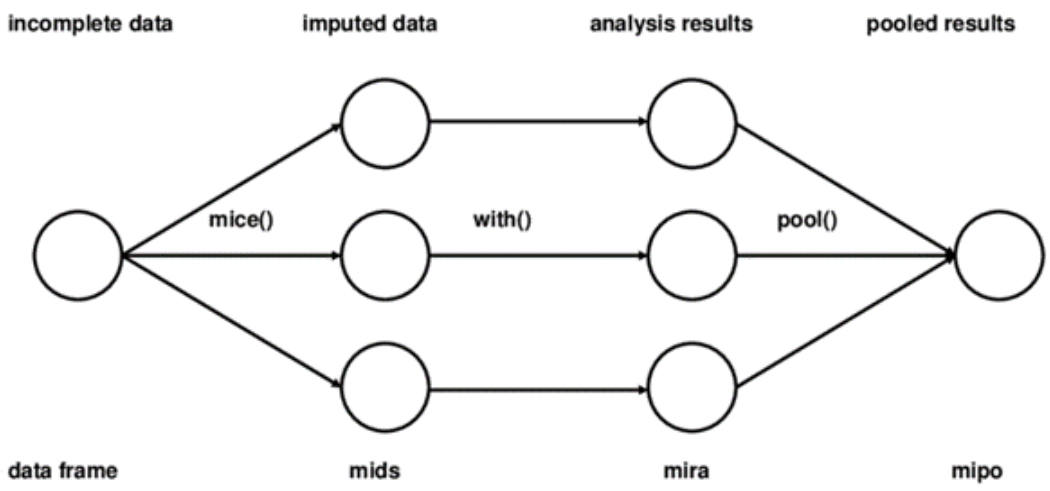

Figure 1: Illustration of the MICE procedure

4 MULTIPLE IMPUTATION FOR QUANTITATIVE VARIABLES

Multiple imputation can be used for different types of data. The focus of this thesis is on quantitative variables. There are several approaches to handling missing data for quantitative variables. They are the predictive mean matching (PMM), Bayesian linear regression, and linear regression non-Bayesian.

4.1 Predictive Mean Matching (PMM)

Normally, it is statistically appropriate to transform data that is not normally distributed before performing any analysis on it but transformations do not always work. Transformations will possibly not achieve near-normality, and even if normality is achieved, bivariate relations may be affected after imputation if normality is assumed. Von Hippel (2013) cautions that the application of techniques used in making distributions of skewed variables closer to normality (e.g., censoring, transformation, truncation) may make matters worse. He argues that censoring which occurs when we round a disallowed value to the closest allowed value and truncation which happens when we redraw a disallowed value until it is within the allowed range can alter both the mean and variability in the data [15]. Stef Buuren (2012) also noted that even though the examples of Von Hippel are extreme to some extent, they do underline the fact that the attempt to use certain methods to achieve normality are limited by what they can do [14]. Allison (2015) explained that PMM was originally used under conditions such as when single variable has missing data and even more generally, when the missing data has a pattern that is monotone. He also adds that PMM is

mostly used for the reason that it produces real values as it naturally takes values from individuals that were studied. However, he cautioned that PMM is weakened by the fact that no mathematical theory exists to justify it usage even though it works well, which is also true of MICE methods more generally. As a result, we have to rely on Monte Carlo simulations, yet we know that no simulation can study all the possibilities. The PMM approach is widely used now because of its inclusion in the MICE software [2].

The PMM works by first regressing the variable that has missing values on the variables without missing values to get predicted values. If all the variables have missing values, the PMM method will use the observed data to regress the target variable with missing value on covariate complete variables. An example is assuming we have a bivariate data where Y denotes the variable with missing values and X denotes the variables with complete values. The missing values will be predicted based regression model

$$\hat{Y}_i = \beta_0 + \beta_1 X_i \tag{1}$$

where $\hat{Y}_i$ is the predicted Y value for a given X_i value, β_0 is the estimated intercept, and β_1 is the estimated slope. During the regression process, we add random variation to the values predicted in other to maintain the distribution of the data that was imputed [10].

Our new equation after the addition of random variation to the values estimated is

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + \delta\mu \tag{2}$$

where δ is the root mean squared error and μ is a random draw from a standard normal distribution. The next step is to use the equation 2 to create values for the variables that has missing values. After creating the values, k number of cases are identified from the observed values that has its predicted values close to the predicted value for case with missing data. The PMM is in the MICE package in R with default value of m=5. One of the close cases identified is selected through simple random sampling and an observed value is then assigned to serve as a replacement for the missing value for that case [2]. We then repeat the entire process until we have convergence after which the whole process is repeated m times in other to produce m complete data sets. We pick each of the m complete data sets and then fits a linear regression model get m different estimates. We finally calculate the mean of the m parameters to get one single point estimate(s) [10]. A major importance of using the PMM is that only qualified values are imputed since we used the observed values in computing the missing values.

4.2 Bayesian Linear Regression

Bayesian linear regression model works by filling the missing values using Bayesian linear regression. The Bayesian linear regression has two main steps: expectation maximization (EM) algorithm and data augmentation (DA). The expectation maximization (EM) notably uses an iterative method to find the maximum likelihood estimates. The EM step works by obtaining the mean, variance and covariance from the complete data to estimate the missing values in the incomplete data. The estimates are then added to the missing data to create a complete data. The EM

process is repeated to obtain a new set of values for mean, variance and covariance which are then use to estimate the new missing data points. Finally, the EM and DA process is repeated until we reach convergence. C. K. Enders (2010) proved the EM mathematically by using a bivariate analysis where X and Y represent complete and incomplete data sets, respectively. This he proved based on the observed dataset that the maximum likelihood estimates for the mean and covariance is

$$\hat{\mu}_Y = \sum \frac{Y}{N},$$
$$\hat{\delta_Y^2} = \frac{1}{N}(\sum Y^2 - \frac{\sum Y^2}{N}),\ and$$
$$\hat{\delta}_{XY} = \frac{1}{N}(\sum XY - \frac{\sum X \sum Y}{N}),$$

where N is the number of observed cases.

The estimates for the mean and covariance is used to build a regression model with the formulas shown below:

$$\hat{\beta}_1 = \frac{\hat{\delta}_{XY}}{\hat{\delta^2}_X},$$
$$\hat{\beta}_0 = \hat{\mu}_Y - \hat{\beta}_1 \hat{\mu_X},$$
$$\hat{\delta}_{Y|X} = \hat{\delta^2}_Y - \hat{\beta^2}_1 \hat{\delta^2}_X$$
$$\hat{Y_i} = \beta_0 + \beta_1 X_i \text{ [6]}$$

The next step is the data augmentation, which uses the population mean, variance and covariance from the EM stage to estimate the missing values. Assume the data for the analysis is represented as $K = (K_{miss}, K_{obs})$ for missing and observed data, respectively. The DA works first by using the current parameter estimate obtained from the EM stage to simulate the missing data. This process is

known as the imputation step. The imputation step is the first of the two steps Bayesian linear regression model. For example, assuming the current parameter values and imputed data are $(\mu^{(i)}, \sum^{(i)}, K^{(i)})$ with iterations (i = 0, 1, 2, . . .) and $(K_{(obs)}, K_{j(miss)})$ as the observed and missing values of the j^{th} case of the data, respectively, then the imputations for the missing values are done independently by simulating $K_j^{(i+1)}{}_{(MIS)} \sim p(K_{j(miss)}|(K_{j(obs)}, \mu^{(t)}, \sum^{(t)})$. The other step is the posteriors Step (P-Step) where the parameters are estimated from the data imputed. The parameters $(\mu^{(i+1)}, \sum^{(i+1)})$ are taken from the corresponding conditional posterior distributions $(\mu, \sum |K^{(i+1)})$. We then repeat the two steps until convergence, thereby generating imputed values and parameter estimates that are Markov Chain. Unlike the PMM method, Bayesian linear regression methods are supported mathematically using Bayes' theorem [2].

4.3 Linear Regression, Non-Bayesian

In linear regression non-Bayesian, linear regression is use in the imputation of missing values. It fits a regression model on the observed data, by regressing variables with missing data on the variables with no missing data. Finally, the spread on the fitted line is use in predicting the missing values [5]. The linear regression, non-Bayesian is limited in practice as it does not incorporate sample uncertainty which occurs when the potential variation in point estimates arises due to the fact that the estimates are dependent on the population sample. This causes the linear regression non-Bayesian, method to underestimate the variability in values imputed for small sample sizes since it does not take into account the variability from the estimates of

the regression coefficients. Because of this shortfall, the linear regression non-Bayesian works for large sample size data that follow a normal distribution [5].

5 METHODOLOGY

5.1 Data Source and Description

The Combined Cycle Power Plant data set was used for this study. It was taken from the Machine Learning Repository of the University of California Irvine. The data can be accessed from the link: http://archive.ics.uci.edu/ml/datasets/ Combined+Cycle+Power+Plant. The combined cycle power plant (CCPP), is an electrical power plant, that uses gas and steam turbine to produce together up to 50 percent more electricity from the same fuel than a traditional simple-cycle plant. The CCPP uses waste heat from the gas turbine to increase efficiency and electrical output. More on how the CCPP operates can be accessed from https://www.gepower.com/resources/knowledge-base/ombined-cycle-power-plant-how-it-works [Online; accessed January 22, 2021].

The dataset have 9568 observations that was collected in a 6 years period from 2006 to 2011. The predictor variables are the average ambient temperature (AT), ambient pressure (AP), relative humidity (RH) and exhaust vacuum (V) with energy output (EP) as the response variable. The ambient temperature (AT) values ranges from 1.81°C to 37.11°C on average whiles the ambient pressure (AP) ranges from 992.89 to 1033.30 millibar. The relative humidity (RH), exhaust vacuum (V) and net hourly electrical energy output (EP) ranges between 25.56% to 100.16%, 25.36-81.56 cm Hg and 420.26-495.76 MW respectively. In this study, the mean and variance from the normal distribution of the CCPP data set was used on the CCPP data set to generate a pseudo data set in the following sizes 500, 150, 50 and 15 while

preserving the correlation structure. We simulated these datasets to preserve the correlation structure since there was strong correlation with the response variable and each predictor variable. This was achieved using the rnorm_multi() function in R. The rnorm_multi() function in R enabled us to preserve the correlation structure while randomly generating multiple samples of the datasets to cater for different levels of variability. Each dataset generated have variability that are small, regular and large. The regular variability is the variability based on the CCPP dataset for each variable. The small variability was calculated by halving the regular variability from the CCPP dataset. The large variability was also achieved by doubling the regular variability from the CCPP dataset. A total of twelve datasets were generated with each of the four sizes having the three levels of variability i.e. small, regular and large.

5.2 ANALYSIS OF COMPLETE DATASET

The twelve complete datasets obtained are used to fit multiple linear regression with AT, AP, RH and V as the predictor variables and EP as the response variable. The estimated regression model based on the four predictors will be in form

$$\hat{EP} = \hat{\beta}_0 + \hat{\beta}_1 V + \hat{\beta}_2 AP + \hat{\beta}_3 RH + \hat{\beta}_4 AT \tag{3}$$

5.2.1 Model Building with 15 Observations

The multiple regression model for the complete dataset with 15 observations and small, regular and large variability indicates that all the predictor variables are needed

in the models in the presence of other predictors at 5% level of significance except AP. These hypotheses tests were conducted using the standard t-test. However, the predictor variable AP will not be dropped from each of the three models as its presence does not affect the significance of these models in the presence of the other predictors. A formal test to determine the normality of the residuals for the three datasets was performed using the Shapiro Wilk's test, with the outcomes all indicating the three datasets being normally distributed at 5% level of significance. The global F-test on each of the three models also shows that all the predictors are significant in predicting the net hourly electrical energy output (EP) as shown in table 6. The estimates for the three models are displayed in tables 3, 4 and 5. The residual plots for the three models showed random patterns which indicates that the assumption of constant variance is met. The residual plots for the three models are displayed in figures 2, 3 and 4. Tables 7, 8 and 9 show that the variance inflation factor (VIF) values are all less than 10, which indicates that the is no multicollinearity issue for any of the 3 models. We also determined how good each of the three models can predict the net hourly electrical energy output (EP) by comparing the predicted residual sum of squares (PRESS) to the sum of squares error (SSE). A relatively closer PRESS value to the SSE value indicates a good model predictability. For each model, the PRESS statistic was much larger than the SSE and thus the model lacks good predictive ability as shown in table 10.

Table 3: The Estimated Regression Coefficients for Data Size of 15 with Small Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Estimates	1027.6769	-0.6062	-0.4182	-0.4703	-4.1245
Test Statistic	2.020 (0.0709)	-2.342 (0.0412)	-0.852 (0.4142)	-2.254 (0.0478)	-6.352 (0.0000)

Table 4: The Estimated Regression Coefficients for Data Size of 15 with Regular Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Estimates	84.8587	-0.5175	0.4320	-0.1782	-1.3774
Test Statistic	0.421 (0.0709)	-3.775 (0.0036)	2.194 (0.0529)	-2.694 (0.0225)	-5.438 (0.0002)

Table 5: The Estimated Regression Coefficients for Data Size of 15 with Large Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Estimates	728.5205	-0.1328	-0.2283	-0.1523	-1.3031
Test Statistic	6.651 (0.0000)	-2.491 (0.0319)	-2.194 (0.0530)	-2.389 (0.0380)	-6.877 (0.0000)

Table 6: Global F-test for the three Datasets of Size 15

Data	Small	Regular	Large
F- Statistic	54.74	37.91	121.41
P-Value	9.187e-07	5.159e-06	6.457e-07

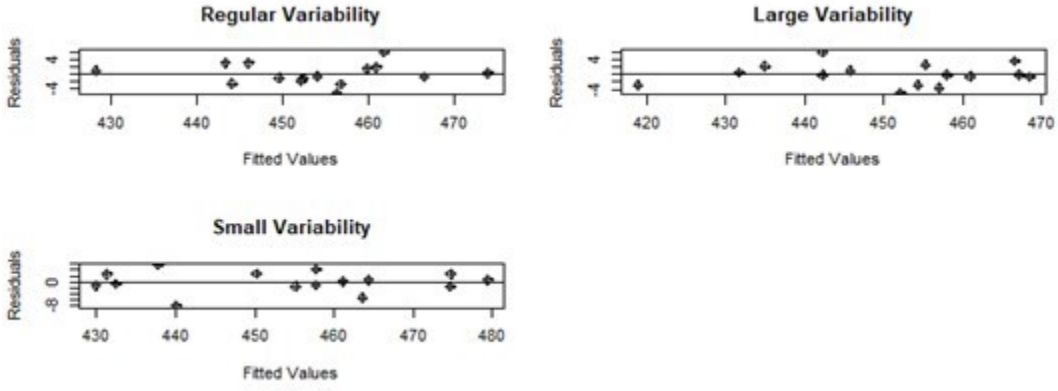

Figure 2: Residual plots for each of the three Models of Size 15

Table 7: VIF Values for the Model with 15 Observations and Small Variability

Variable	V	AP	RH	AT
VIF Value	3.106691	2.514250	1.062133	3.952052

Table 8: VIF Values for the Model with 15 Observations and Regular Variability

Variable	V	AP	RH	AT
VIF Value	1.325952	2.372849	1.413368	1.762973

Table 9: VIF Values for the Model with 15 Observations and Large Variability

Variable	V	AP	RH	AT
VIF Value	1.457617	6.986152	4.938358	1.7459683

Table 10: PRESS Statistic and SSE Values of the three Datasets of Size 15

Data	Small	Regular	Large
PRESS Statistic	477.8445	270.8041	477.8444
SSE	174.29	110.69	121.41

5.2.2 Model Building with 50 Observations

The multiple regression model for the complete dataset with 50 observations and a small, regular and large variabilities indicates that all the predictor variables are needed in the models in the presence of other predictors at 5% level of significance. These hypotheses tests were conducted using the standard t-test. A formal test to

determine the normality of the residuals for the three datasets was performed using the Shapiro Wilk's test, with the outcomes all indicating the three datasets being normally distributed at 5% level of significance. The global F-test on each of the three models also shows that all the predictors are significant in predicting the net hourly electrical energy output (EP) as shown in table 14. The estimates for the three models are displayed in tables 11, 12 and 13. The residual plots for the three models showed random patterns which indicates that the assumption of constant variance is met. The residual plots for the three models are displayed in figure 3. Tables 15, 16 and 17 shows that the variance inflation factor (VIF) values are all less than 10, which indicates that the is no multicollinearity issue for any of the 3 models. We also determined how good each of the three models can predict the net hourly electrical energy output (EP) by comparing the predicted residual sum of squares (PRESS) to the sum of squares error (SSE). A relatively closer PRESS value to the SSE value indicates a good model predictability. For each model, the PRESS statistic was much larger than the SSE and thus the model lacks good predictive ability as shown in table 18.

Table 11: The estimated regression coefficients for data size of 50 with Small Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	8.4695	-0.6974	0.5644	-0.2583	-3.5111
Test Statistic	0.035 (0.9724)	-3.661 (0.0006)	2.363 (0.0225)	-9.039 (0.0000)	-2.567 (0.0136)

Table 12: he Estimated Regression Coefficients for Data Size of 50 with Regular Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	192.38014	-0.21791	0.31357	-0.10298	-1.85033
Test Statistic	1.816 (0.0763)	-3.016 (0.0043)	3.045 (0.0039)	-2.356 (0.0231)	-11.427 (0.0000)

Table 13: The Estimated Regression Coefficients for Data Size of 50 with Large Lariability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	301.49747	-0.10142	0.18132	-0.10129	-0.94705
Test Statistic	4.983 (0.0000)	-2.505 (0.0159)	3.079 (0.0035)	-3.827 (0.0004)	-12.130 (0.0000)

Table 14: Global F-test for the three Datasets of Size 50

Data	Small	Regular	Large
F- Statistic	189.9	319.4	235.2
P-Values	2.2e-16	2.2e-16	2.2e-16

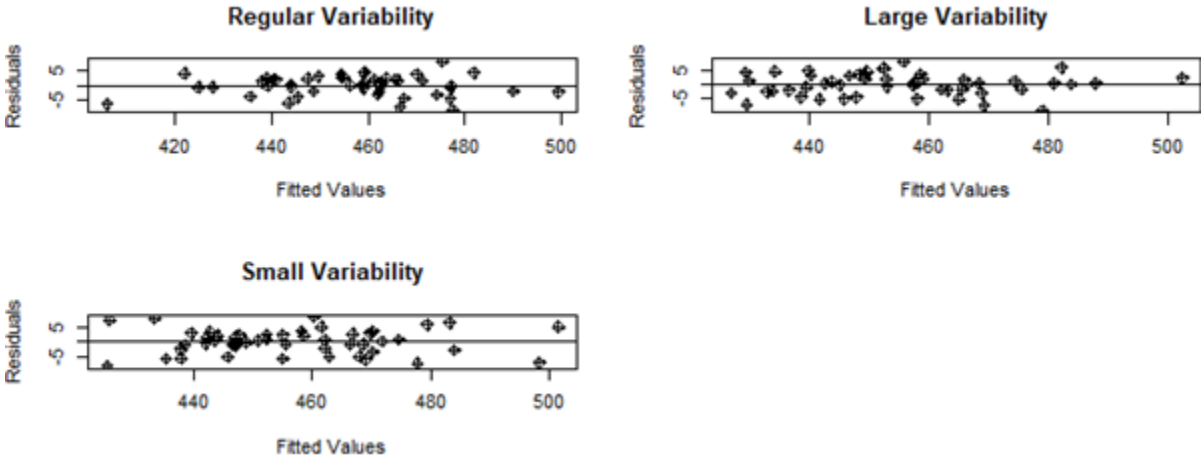

Figure 3: Residual plots for each of the three Models of Size 50

Table 15: VIF Values for the Model with 50 Observations and Small Variability

Variable	V	AP	RH	AT
VIF Value	3.856440	5.299007	1.349373	1.478244

Table 16: VIF Values for the Model with 50 Observations and Regular Variability

Variable	V	AP	RH	AT
VIF Value	4.395035	1.618554	1.702938	6.468061

Table 17: VIF Values for the Model with 50 Observations and Large Variability

Variable	V	AP	RH	AT
VIF Value	3.486297	1.337938	1.633406	4.825843

Table 18: PRESS Statistic and SSE Values of the three Datasets of Size 50

Data	Small	Regular	Large
PRESS Statistic	1073.37	671.2846	873.3458
SSE	827.5	516.8	691.0

5.2.3 Model Building with 150 Observations

The multiple regression model for the complete dataset with 150 observations and a small, regular and large variabilities indicates that all the predictor variables are needed in the models in the presence of other predictors at 5% level of significance. These hypotheses tests were conducted using the standard t-test. A formal test to

determine the normality of the residuals for the three datasets was performed using the Shapiro Wilk's test, with the outcomes all indicating the three datasets being normally distributed at 5% level of significance. The Global F-test on each of the three models also shows that all the predictors are significant in predicting the net hourly electrical energy output (EP) as shown in table 22. The estimates for the three models are displayed in table 19, table 20 and table 21. The residual plots for the three models showed random patterns which indicates that the assumption of constant variance is met. The residual plots for the three models are displayed in figure 4. Tables 22, 23 and 24 shows that the variance inflation factor (VIF) values are all less than 10, which indicates that the is no multicollinearity issue for any of the 3 models. We also determined how good each of the three models can predict the net hourly electrical energy output (EP) by comparing the predicted residual sum of squares (PRESS) to the sum of squares error (SSE). A relatively closer PRESS value to the SSE value indicates a good model predictability. For each model, the PRESS statistic was closer to the SSE and thus the models have good predictive ability as shown in table 26.

Table 19: The Estimated Regression Coefficients for Data Size of 150 with Small Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	243.31570	-0.52660	0.32615	-0.23567	-3.75328
Test Statistic	1.719 (0.0876)	-4.768 (0.0000)	2.382 (0.0185)	-3.772 (0.0002)	-16.382 (0.0000)

Table 20: The Estimated Regression Coefficients for Data Size of 150 with Regular Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	402.66101	-0.10295	0.08033	-0.06217	-0.99031
Test Statistic	9.737 (0.0000)	-3.454 (0.0007)	1.998 (0.0476)	-3.571 (0.0005)	-16.140 (0.0000)

Table 21: The Estimated Regression Coefficients for Data Size of 150 with Large Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	301.49747	-0.10142	0.18132	-0.10129	-0.94705
Test Statistic	4.983 (0.0000)	-2.505 (0.0159)	3.079 (0.0035)	-3.827 (0.0004)	-12.130 (0.0000)

Table 22: Global F-test for the three Datasets of Size 150

Data	Small	Regular	Large
F- Statistic	1556	539.1	391.9
P-Values	2.2e-16	2.2e-16	2.2e-16

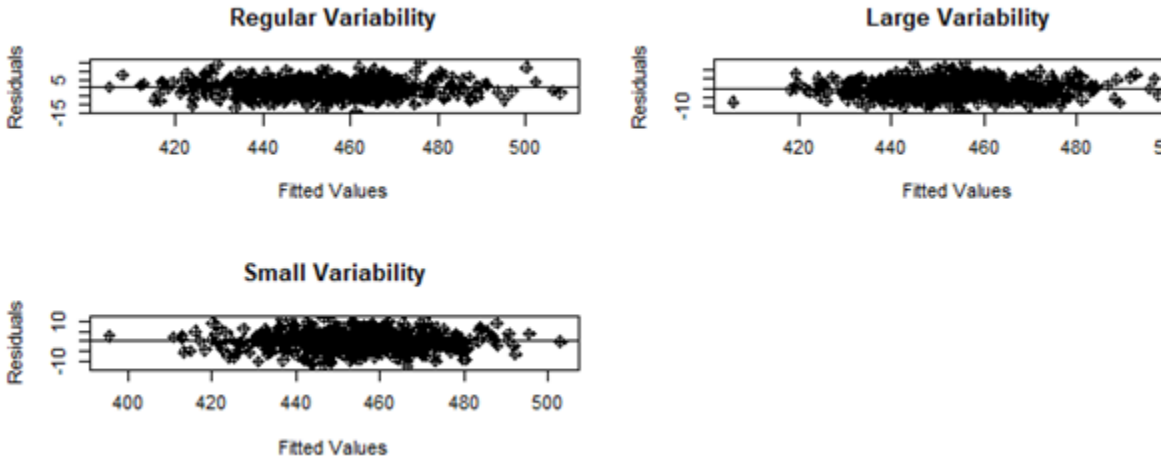

Figure 4: Residual plots for each of the three Models of Size 150

Table 23: VIF Values for the Model with 150 Observations and Small Variability

Variable	V	AP	RH	AT
VIF Value	3.856440	5.299007	1.349373	1.478244

Table 24: VIF Values for the Model with 150 Observations and Regular Variability

Variable	V	AP	RH	AT
VIF Value	4.395035	1.618554	1.702938	6.468061

Table 25: VIF Values for the Model with 150 Observations and Large Variability

Variable	V	AP	RH	AT
VIF Value	3.486297	1.337938	1.633406	4.825843

Table 26: PRESS Statistic and SSE Values of the three Datasets of Size 150

Data	Small	Regular	Large
PRESS Statistic	1073.37	671.2846	873.3458
SSE	827.5	516.8	691.0

5.2.4 Model Building with 500 Observations

The multiple regression model for the complete dataset with 500 observations and a small, regular and large variabilities indicates that all the predictor variables are needed in the models in the presence of other predictors at 5% level of significance. These hypotheses tests were conducted using the standard t-test. A formal test to

determine the normality for the three datasets with 50 observations were performed using the Shapiro Wilk's test, with the outcomes all indicating the three datasets being normally distributed at 5% level of significance. The global F-test on each of the three models also shows that all the predictors are significant in predicting the net hourly electrical energy output (EP) as shown in table 30. The estimates for the three models are displayed in tables 27, 28 and 29 below. The residual plots for the three models showed random patterns which indicates that the assumption of constant variance is met. The residual plots for the three models are displayed in figure 5. Tables 31, 31 and 33 shows that the variance inflation factor (VIF) values are all less than 10, which indicates that the is no multicollinearity issue for any of the 3 models. We also determined how good each of the three models can predict the net hourly electrical energy output (EP) by comparing the predicted residual sum of squares (PRESS) to the sum of squares error (SSE). A relatively closer PRESS value to the SSE value indicates a good model predictability. For each model, the PRESS statistic was closer to the SSE and thus the models have good predictive ability as shown in table 34.

Table 27: The Estimated Regression Coefficients for Data Size of 500 with Small Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	2388.06821	- 0.50340	0.18840	- 0.28470	-3.88987
Test Statistic	4.412 (0.0000)	-8.111 (0.0000)	2.208 (0.0277)	-7.206 (0.0000)	-29.626 (0.0000)

Table 28: The Estimated Regression Coefficients for Data Size of 500 with Regular Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	425.71258	-0.20880	0.09110	-0.16635	-2.0358
Test Statistic	10.462 (0.0000)	-6.500 (0.0000)	2.309 (0.0214)	-8.637 (0.0000)	-30.431 (0.0000)

Table 29: The Estimated Regression Coefficients for Data Size of 500 with Large Variability. The p-value is given in parentheses

Parameter	β_0	β_1	β_2	β_3	β_4
Values	301.49747	-0.10142	0.18132	-0.10129	-0.94705
Test Statistic	9.737 (0.0000)	-3.454 (0.0007)	1.998 (0.0476)	-3.571 (0.0005)	-16.140 (0.0000)

Table 30: Global F-test for the three Datasets of Size 500

Data	Small	Regular	Large
F- Statistic	1468	1645	1417
P-Values	2.2e-16	2.2e-16	2.2e-16

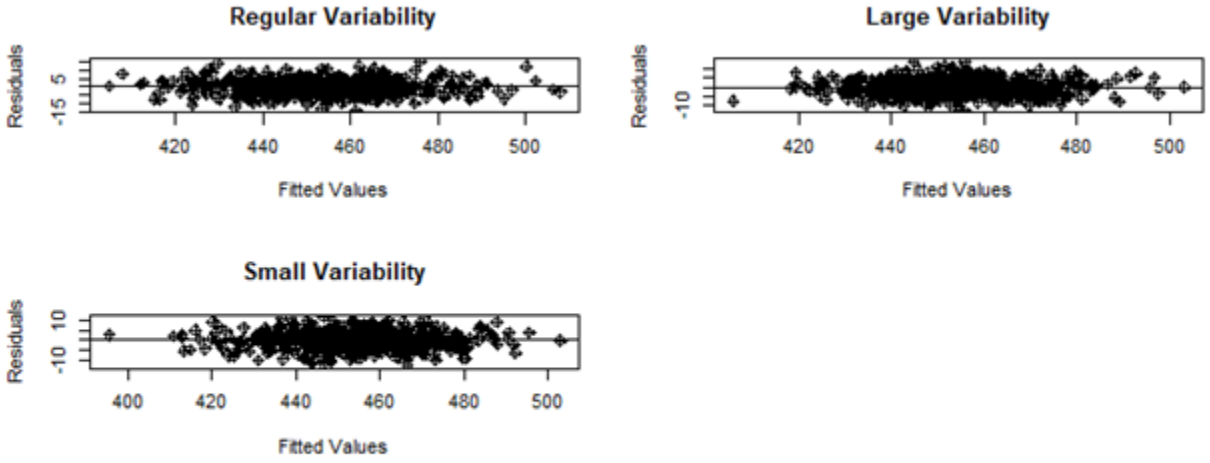

Figure 5: Residual plots for each of the three Models of Size 500

Table 31: VIF Values for the Model with 500 Observations and Small Variability

Variable	V	AP	RH	AT
VIF Value	3.242313	1.480283	1.788695	5.230153

Table 32: VIF Values for the Model with 500 observations and Regular Variability

Variable	V	AP	RH	AT
VIF Value	4.049733	1.387367	1.582342	5.796745

Table 33: VIF Values for the Model with 500 Observations and Large Variability

Variable	V	AP	RH	AT
VIF Value	3.542432	1.489708	1.977970	5.937348

Table 34: PRESS Statistic and SSE Values of the three Datasets of Size 500

Data	Small	Regular	Large
PRESS Statistic	10667.8114	10807.240	3599.276
SSE	10453.000	10588.000	3339.200

5.3 Relative Efficiency

The relative efficiency (RE) describes the best multiple imputation procedure that produces the most accurate result. It depends on the amount of missing information and the number of imputations performed. Relative efficiency (RE) is computed with 50 imputations since we have high amount of missing information that requires appropriate estimation of standard error. It is calculated as

$$RE = \frac{1}{1 + \frac{\lambda}{m}} \tag{4}$$

where λ is the fraction of missing information (FMI) and m the number of imputations [11]. Table 35 shows that as the relative efficiency increases across all the fraction of missing information, then the number of imputation increases.

Table 35: The Relative Efficiency for Different Levels of m and FMI

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	0.8000
3	0.9677	0.9375	0.9091	0.8824	0.8571
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
25	0.9961	0.9920	0.9881	0.9840	0.9801
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

5.4 Imputation Implementation

The main objective of this study is to assess the best multiple imputation method for quantitative variables in different sizes with varied variability. Using a function in R, we create different percentage of missingness for all the complete datasets except the 15 observation data in the proportions of 10%, 20%, 30%, 40% and 50%. The missingness was achieved by first removing 10% of each of the 50, 150 and 500 complete data set. The 20% missingness was done by removing another 10% on the previous 10% removed, and it was continued in that order until 50% of the data was removed randomly. The percentage of missingness for the 15-observation data was generated in the proportions of 10%, 20%, 30%. We applied each of the three multiple imputation methods, PMM, Bayesian linear regression, and Linear regression, non-Bayesian on each of the 60 missing data sets. The imputations were entered 50 times at each stage of the imputation. i.e. for each imputation stage, m=50 at each cycle

of imputation. This process was then simulated for 1000 iterations to enable us to better estimate the true coefficient values. For each of the 50 imputed data sets, the linear regression models described in sections 5.2 – 5.5 was fitted and the coefficients of the model where then pooled together and stored. Performing 50 imputations will reduce the sample variability in the dataset since each estimate stored is computed based on the mean of 50 estimates. Paul Allison (2012) adds that performing more than one imputation on a missing data provides important information to compute standard error estimates that will correctly reflect the uncertainty around the missing value. This process was repeated for 1000 iterations. The mean, for each of the 1000 coefficients for each variable, was computed and compared to each of the coefficients from the model of the complete data sets found in sections 5.2 – 5.5. The best method to impute the missing data for a specific percentage of missingness is the one that produces the mean from the imputed data, closest to the corresponding coefficient from the complete data. To evaluate the mean estimate that is closer to the true mean estimate from the complete data, we compute the percentage deviation index (PDI). The PDI measures how far the mean of the estimated regression coefficient from the imputed data is away from the estimates from the complete data. The PDI is calculated as

$$PDI = \frac{\text{Orig. regres. coefficient - Mean of estimated regres. coefficient}}{\text{Orig. regres. coefficient}} * 100.$$

To determine the statistical significant difference between the original coefficients of the estimated parameters and the mean of the estimated regression coefficients, one-sample t-tests were employed.

6 RESULTS

The multiple imputed data sets for each of the four datasets simulated are analyzed separately using regression models established in 5.1. We evaluate the performance of the three imputations models (PMM, Bayesian linear regression, and linear regression non-Bayesian) with the results obtained from the regression analysis.

6.1 Analysis of the 15 Sample Size Dataset

Regression analysis was performed on the imputed data sets to obtain the mean, and the results are compared to corresponding coefficients from the complete data. For sample size 15 data, we generated the fraction of missing information(FMI) of imputed values up to 30%. This is because, the imputed values at 40% and 50% missingness for all three datasets of sample szie 15 were linearly dependent with the response variable in the data. This occurs since when imputing missing values with higher percentage of missingness, already imputed values are reused in the imputation process to complete the imputed data. The mean of the estimated regression coefficients decreases as the fraction of missing information(FMI) increases from 10% to 30%. This is the same across the three imputation models; PMM, Bayesian linear regression, and linear regression non-Bayesian models as shown in tables 36, 39, 42, 45, 48, 51, 54, 57 and 60. Also, across all the three imputation models, the estimated mean of the regression coefficient from β_1 to β_4, using the imputed data, decreases as the percentage of missingness decreases from 10% to 30%, as shown in tables 36, 39, 42, 45, 48, 51, 54, 57 and 60. This is the same for datasets of sample sizes 15

with regular and large variability. The closest values to the mean of the parameter estimates for complete data are at 10% level of missingness for all the three data set of size 15. This means, smaller data sets having few missing values will have its imputed data closer to the actual values of a complete data. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 38, 41, 44, 47, 50, 53, 56, 59 and 62 show the p-values from the one sample *t*-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 36: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 15 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	210.5527	-0.5975	0.3859	-0.6542	-3.4079
20%	-200.1169	-0.5167	0.7878	-0.6525	-3.5007
30%	-1538.0591	-0.9516	2.0163	0.4970	-1.8224
Actual Parameter	**1027.6769**	**-0.6062**	**-0.4182**	**-0.4703**	**-4.1245**

Table 37: PDI of the Regression Coefficients of PMM Model for 15 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.7951	0.0144	1.9228	-0.3910	0.1737	0.503
20%	1.1947	0.1476	2.8838	-0.3874	0.1512	0.79797
30%	2.4966	-0.5698	5.8214	2.0568	0.5582	0.2047
Mean	1.4954	-0.1359	3.5426	0.4261	0.2943	**0.2052**

Table 38: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 15 Observations and Small Variability

FMI \Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 39: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 15 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	365.5543	-0.6508	0.2328	-0.6005	-3.4416
20%	680.76788	-0.6199	-0.0923	-0.5260	-3.0984
30%	-1552.2371	-0.8270	2.0392	0.2947	-1.8977
Actual Parameter	**1027.6769**	**-0.6062**	**-0.4182**	**-0.4703**	**-4.1245**

Table 40: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 15 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.6443	-0.0736	1.5567	-0.2768	0.1656	0.4032
20%	0.3376	-0.0226	0.7793	-0.1184	0.2488	0.2449
30%	2.5104	-0.3642	5.8761	1.6266	0.5399	2.0377
Mean	1.1641	-0.1534	2.7373	0.4104	0.3181	**0.8953**

Table 41: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 15 Observations and Small Variability

FMI \Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 42: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 15 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	1075.1848	-1.2417	-0.4591	-0.4302	-2.7220
20%	712.9075	-0.4752	-0.1021	-0.7863	-3.66074
30%	-934.6284	-1.0406	1.4321	0.4871	-2.1064
Actual Parameter	**1027.6769**	**-0.6062**	**-0.4182**	**-0.4703**	**-4.1245**

Table 43: PDI of the Regression Coefficients of the Linear Regression, non-Bayesian Model for 15 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.0462	-1.0483	-0.0978	0.0853	0.3400	-0.1534
20%	0.3063	0.2161	0.7559	-0.6719	0.1124	0.1437
30%	1.9095	-0.7166	4.4244	2.0357	0.4893	1.6284
Mean	0.7232	-0.5162	0.8217	0.4830	0.3139	**0.3651**

Table 44: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 15 Observations and Small Variability

FMI \Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 45: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 15 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	194.7044	-0.6064	0.3198	-0.0932	-1.2668
20%	71.4044	-0.7164	0.4373	-0.0542	-0.8849
30%	-110.4624	-0.6300	0.6120	0.0257	-1.1855
Actual Parameter	**84.85872**	**-0.51753**	**0.43203**	**-0.17829**	**-1.3774**

Table 46: PDI of the Regression Coefficients of PMM Model for 15 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-1.2945	-0.1718	0.2598	0.4770	0.0803	-0.1298
20%	0.1585	-0.3843	-0.0122	0.6958	0.3576	0.1631
30%	2.301	-0.2174	-0.4166	1.1442	0.1393	0.59017
Mean	0.3885	-0.2578	-0.0563	0.7723	0.1924	**0.2078**

Table 47: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 15 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 48: Estimated Means of the Regression Coefficients from the the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 15 and Regular variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	150.1618	-0.5531	0.49041	-0.0737	-1.15087
20%	-14.4925	-0.6714	0.5180	-0.0561	-0.8038
30%	-212.3770	-0.6508	0.7152	-0.0446	-0.9529
Actual Parameter	**84.85872**	**-0.51753**	**0.43203**	**-0.17829**	**-1.3774**

Table 49: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 15 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.7696	-0.0687	-0.1351	0.5864	0.1645	-0.0445
20%	1.1708	-0.2973	-0.1990	0.6852	0.4164	0.3552
30%	3.5027	-0.2575	-0.6554	0.7497	0.3082	0.7295
Mean	1.3013	-0.2078	-0.3298	0.6737	0.2963	**0.4348**

Table 50: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 15 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 51: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 15 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	175.6555	-0.6037	0.3382	-0.0900	-1.2601
20%	23.1083	-0.6285	0.4783	-0.0383	-0.8502
30%	-423.3290	-0.4390	0.9267	-0.0904	-1.5304
Actual Parameter	**84.85872**	**-0.51753**	**0.43203**	**-0.17829**	**-1.3774**

Table 52: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 15 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-1.0700	-0.1665	0.2172	0.4949	0.4600	-0.01288
20%	0.7277	-0.2144	-0.1071	0.7851	-0.0685	0.22456
30%	5.9886	0.1517	-1.1450	0.4927	0.4905	1.1957
Mean	1.8821	-0.0764	-0.3449	0.5909	0.294	**0.4691**

Table 53: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 15 Observations and regular variability

FMI \Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 54: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 15 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	460.7451	-0.1234	0.0304	-0.1265	-1.2185
20%	278.7648	-0.1613	0.2019	-0.0668	-1.0091
30%	-85.3102	-0.2612	0.5855	-0.1099	-1.5341
Actual Parameter	**728.5205**	**-0.13285**	**-0.2283**	**-0.17829**	**-1.3030**

Table 55: PDI of the Regression Coefficients of PMM Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.3676	0.0711	1.1332	0.1694	0.0649	0.36124
20%	0.6174	-0.2142	1.8844	0.5614	0.2256	0.61492
30%	1.1171	-0.9661	3.5646	0.2784	-0.1774	0.7633
Mean	0.7007	-0.3697	2.1940	0.3364	0.0377	**0.5798**

Table 56: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 57: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 15 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
%	19.5999	-0.5767	0.49041	-0.1092	-1.1508
20%	127.0950	-0.2997	0.3806	-0.15407	-1.5451
30%	53.0344	-0.1873	0.4399	-0.0482	-1.5641
Actual Parameter	**728.5205**	**-0.13285**	**-0.2283**	**-0.17829**	**-1.3030**

Table 58: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.9731	-3.3410	3.1481	0.0808	0.4600	0.1836
20%	0.8255	-1.2559	2.6671	0.3870	-0.0685	0.5134
30%	0.9272	-0.4099	2.9269	0.0081	0.4905	0.6076
Mean	0.9086	-1.6689	2.9140	0.1586	0.294	**0.4348**

Table 59: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 60: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 15 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	505.9966	-0.1043	-0.0107	-0.1554	-1.3173
20%	139.6584	-0.2830	0.3695	-0.1808	-1.5632
30%	-20.4359	-0.26803	0.5259	-0.14154	-1.6559
Actual Parameter	**728.5205**	**-0.13285**	**-0.2283**	**-0.17829**	**-1.3030**

Table 61: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.9880	-3.2514	0.7766	0.0808	0.4600	0.1836
20%	0.8083	-1.1302	0.5201	0.3870	-0.0685	0.5134
30%	1.0281	-1.0175	1.1686	0.0081	0.4905	0.6076
Mean	0.9414	-1.7997	0.8217	0.1586	0.294	**0.4348**

Table 62: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 15 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000

6.2 Analysis of Sample Size 50 and Small Variability Data

Using the imputed data, the mean of the estimated regression coefficients for β_0 are relatively far away from its corresponding mean estimates of the complete data across all the three imputation models as evident in tables 63, 66 and 69. However, the rest of the mean estimates of the regression coefficients have their closest values to their corresponding mean estimates of the complete data at 10% level of missingness for all the three imputation models. All the estimates decrease at 20% level of missingness and then increases at 30% level of missingness before decreasing at 50% level of missingness for all the three imputation models as shown in tables 63, 66 and 69.

The PDI for the Bayesian linear regression model is the lowest among the three methods as shown in tables 64, 67 and 70. This shows the Bayesian Linear regression imputation model works best for this type of data. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 65, 68 and 71 shows the p-values from the one sample *t*-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 63: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 50 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	-21.6705	-0.9519	0.588	-0.146	-2.949
20%	-340.9283	-0.8573	0.8898	-0.15188	-2.8721
30%	-285.9567	-1.1201	0.8458	-0.0995	-2.4497
40%	-79.1101	-1.0136	0.6343	-0.0368	-2.6053
50%	-366.3245	-0.2734	0.9073	0.0678	-2.1516
Actual Parameter	**8.4695**	**-0.6974**	**0.5644**	**-0.2583**	**-3.5111**

Table 64: PDI of the Regression Coefficients of PMM Model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	3.5586	0.2425	-0.0517	0.25087	0.1600	0.8320
20%	41.2536	0.0265	-0.5483	0.76035	0.1819	8.33488
30%	34.7631	-0.1274	-0.0983	0.42315	0.3022	7.0525
40%	10.3405	0.2792	0.39174	0.54742	0.2579	2.3633
50%	44.2521	-0.0469	-0.2888	1.16027	0.3872	9.0927
Mean	26.8336	0.0748	-0.1191	0.6284	0.2578	**5.5351**

Table 65: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 66: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 50 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	-25.375	-0.8966	0.5932	-0.1730	-3.0643
20%	-332.2087	-0.8123	0.8778	-0.0613	-2.7636
30%	37.5640	-0.6137	0.5147	-0.1672	-2.9887
40%	240.4518	-0.4645	0.3052	-0.1186	-3.1087
50%	-136.2795	-0.7417	0.6652	0.0299	-2.3248
Actual Parameter	**8.4695**	**-0.6974**	**0.5644**	**-0.2583**	**-3.5111**

Table 67: PDI of the Regression coefficients of Bayesian Linear Regression model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	3.9960	-0.2856	-0.0510	0.3302	0.1272	0.8233
20%	40.2241	-0.1647	-0.5552	0.7626	0.2128	8.0959
30%	-3.4352	0.1200	0.0880	0.3526	0.1487	-0.54515
40%	-27.3903	0.3339	0.4592	0.5408	0.1146	-5.1883
50%	17.0906	-0.0635	-0.1785	1.1157	0.3378	3.6604
Mean	6.0970	-0.0119	-0.0475	0.6204	0.1882	**1.374**

Table 68: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 69: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 50 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	-23.6007	-0.8665	0.5936	-0.1935	-3.1765
20%	-328.4321	-0.7159	0.8739	-0.0619	-3.0159
30%	-66.2566	-0.7863	0.6199	-0.1490	-2.7351
40%	203.0107	-0.5027	0.3433	-0.1169	-3.0630
50%	-202.2604	-0.73014	0.7274	0.0414	-2.2438
Actual Parameter	**8.4695**	**-0.6974**	**0.5644**	**-0.2583**	**-3.5111**

Table 70: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	3.7865	-0.8665	0.0517	0.2508	0.0952	0.66354
20%	39.7782	-0.7159	-0.5483	0.7603	0.1410	7.8830
30%	8.8229	-0.7863	-0.0983	0.4231	0.2210	1.71648
40%	-22.9696	-0.5027	0.3917	0.5474	0.1276	-4.4811
50%	24.8810	-0.7301	-0.2888	1.1602	0.3609	5.0766
Mean	10.8598	-0.7203	-0.0984	0.6283	0.1891	**2.1717**

Table 71: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 50 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.3 Analysis of Sample Size 50 and Regular Variability Data

Using the imputed data, the mean of the estimated regression coefficients closest to the actual parameter occurs at 20% and 30% level of missingness for β_0 while for $\beta_2$and β_4 it decreases as the percentage of missingness increases from 10% to 20% then it increases for 40% and decreases at 50%. The β_1 and β_2 have the closest value to their corresponding means from the complete data at 20% and 50% level of missingness respectively. This is shown in tables 72, 75 and 78 for all three imputation

methods. The PDI for the linear regression, non-Bayesian model method has mean of -0.2018, which is the lowest among the three methods as shown in tables 73, 76 and 79. This shows the linear regression, non-Bayesian model imputation method is the best for this type of data. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 74, 77 and 80 shows the p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 72: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 50 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	53.8659	-0.5047	0.4448	0.0133	-1.2438
20%	211.325	-0.4772	0.2958	-0.0240	-1.4870
30%	172.3285	-0.5570	0.3404	-0.0786	-1.3617
40%	124.9674	-0.5559	0.38707	-0.0818	-1.3332
50%	324.4995	-0.5522	0.1917	-0.0860	-1.4180
Actual Parameter	**192.3801**	**-0.21791**	**0.31357**	**-0.10298**	**-1.85033**

Table 73: PDI of the Regression Coefficients of PMM Model for 50 observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.2909	-0.7939	-0.1656	0.7300	0.1820	0.0486
20%	-0.0416	-0.9008	0.0296	0.7310	0.1774	-0.0008
30%	-0.3506	-0.7590	0.2040	0.1387	0.1622	-0.1209
40%	-0.1416	-1.1422	0.0729	0.0445	-0.5105	-0.3353
50%	-0.6384	-1.0403	0.3989	0.5766	0.1820	-0.1042
Mean	-0.17626	-0.9272	0.1079	0.4442	0.00277	**-0.1025**

Table 74: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 50 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 75: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 50 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	136.4159	-0.3909	0.3655	-0.0278	-1.5136
20%	200.3835	-0.4142	0.3043	-0.0277	-1.522
30%	259.8199	-0.3833	0.2496	-0.0887	-1.5502
40%	219.6256	-0.4668	0.2907	-0.0984	-1.5440
50%	315.1978	-0.4446	0.1885	-0.0436	-1.2510
Actual Parameter	**192.3801**	**-0.21791**	**0.31357**	**-0.10298**	**-1.85033**

Table 76: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 50 observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.2909	-0.7939	-0.1656	0.7300	0.1820	0.0486
20%	-0.0416	-0.9008	0.0296	0.7310	0.1774	-0.0008
30%	-0.3506	-0.7590	0.2040	0.1387	0.1622	-0.1209
40%	-0.1416	-1.1422	0.0729	0.0445	-0.5105	-0.3353
50%	-0.6384	-1.0403	0.3989	0.5766	0.1820	-0.1042
Mean	-0.17626	-0.9272	0.1079	0.4442	0.00277	**-0.1025**

Table 77: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 50 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 78: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 50 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	120.4030	-0.3791	0.3819	-0.0358	-1.5419
20%	126.1197	-0.4628	0.3773	-0.0181	-1.405
30%	202.2480	-0.4161	0.308	-0.0943	-1.5288
40%	173.1805	-0.4907	0.3359	-0.0827	-1.3173
50%	313.3487	-0.36178	0.1917	-0.0833	-1.3706
Actual Parameter	**192.38014**	**-0.21791**	**0.31357**	**-0.10298**	**-1.85033**

Table 79: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 50 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.3741	-12.6588	-0.2179	0.6524	0.1667	-2.3367
20%	0.3444	-0.0234	-0.2032	0.8242	0.2407	0.2365
30%	-0.0513	1.6048	0.0178	0.0843	0.1738	0.3658
40%	0.0998	1.5562	-0.0712	0.1969	0.2881	0.4139
50%	-0.6288	1.3456	0.3887	0.1911	0.2593	0.3112
Mean	0.02764	-1.6351	-0.0171	0.3398	0.22572	**-0.2018**

Table 80: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 50 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.4 Analysis of Sample Size 50 and Large Variability Data

The mean of the estimated regression coefficients for the PMM model have the closest values to the complete data at 10% level missingness apart from β_4 whiles the Bayesian linear regression has the closest value to the complete data at 20% level of missingness as shown in tables 81 and 84, respectively. The linear regression, non-Bayesian model interestingly has its closest value to the complete data at 50% level of

missingness, which is shown in table 87. The PDI for the Bayesian linear regression and linear regression, non-Bayesian are close and lower than the PDI of the PMM model as evident in tables 82, 85 and 88. This shows both imputation models can be used for this type of data. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 83, 86 and 89 shows the p-values from the one sample *t*-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 81: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 50 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	283.3466	-0.1831	-0.1831	0.0110	0.5473
20%	346.4983	-0.1521	0.2698	-0.0474	-0.7190
30%	221.6233	-0.2727	0.2536	-0.0041	-0.5577
40%	245.1259	-0.2918	0.2293	0.0145	-0.5216
50%	244.1290	-0.2483	0.2284	0.0109	-0.5393
Actual Parameter	**301.49747**	**-0.10142**	**0.18132**	**-0.10129**	**-0.94705**

Table 82: PDI of the Regression Coefficients of PMM Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0602	-0.8054	-0.0385	1.1086	1.5779	0.38056
20%	-0.1493	-0.4997	-0.4880	0.5320	0.2408	-0.07284
30%	0.2649	-1.6888	-0.3986	0.9595	0.4111	-0.09038
40%	0.1870	-1.8771	-0.2646	1.1432	0.4492	-0.0724
50%	0.1903	-1.4482	-0.2597	1.1076	0.4305	0.0041
Mean	0.11062	-1.2638	-0.2898	0.97018	0.6219	**0.0298**

Table 83: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 84: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 50 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	221.2642	-0.1574	0.2580	-0.0817	-0.7653
20%	281.3053	-0.2527	0.2238	-0.0829	-1.5395
30%	223.5400	-0.24441	0.2514	-0.0168	-1.6057
40%	261.0283	-0.2415	0.2115	0.0093	-0.5284
50%	227.3895	-0.1782	0.2412	0.0109	-0.5393
Actual Parameter	**301.49747**	**-0.10142**	**0.18132**	**-0.10129**	**-0.94705**

Table 85: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.2661	-0.5520	-0.4229	0.1934	0.1919	-0.0647
20%	0.0670	-1.4916	-0.2343	0.1816	-0.6256	-0.4205
30%	0.2586	-1.4099	-0.3865	0.8341	-0.6955	-0.2798
40%	0.1342	-1.3812	-0.1664	1.0918	0.4421	0.0241
50%	0.2458	-0.7570	-0.3302	-0.3302	0.4520	0.1377
Mean	0.1943	-1.1183	-0.3081	0.6758	-0.0470	**-0.1206**

Table 86: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 87: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 50 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	231.3800	-0.1474	0.2489	-0.09049	-0.7928
20%	214.6902	-0.1499	0.2626	-0.0227	-0.6059
30%	216.3324	-0.2222	0.2586	-0.1201	-1.8037
40%	328.8503	-0.2242	0.40481	-0.0076	-0.5849
50%	192.3801	-0.2179	0.3135	-0.00232	-1.8503
Actual Parameter	**301.49747**	**-0.10142**	**0.18132**	**-0.10129**	**-0.94705**

Table 88: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.2326	-0.4534	-0.3727	0.1066	0.1629	-0.0648
20%	0.2879	-0.4780	-0.4483	0.7759	0.3602	0.0995
30%	0.2825	-1.1909	-0.4262	-0.1857	-0.9045	-0.4849
40%	-0.0907	-1.2106	-1.2326	0.9250	0.3824	-0.2453
50%	0.3619	-1.1485	-0.7290	0.9771	0.953	0.0829
Mean	0.2148	-0.89628	-0.64176	0.51978	0.1908	**-0.1225**

Table 89: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 50 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.5 Analysis of Sample Size 150 and Small Variability Data

The mean of the estimated regression coefficients for the three methods have the closest values to the complete data at 50% level of missingness as shown in tables 90, 93 and 96. In general, the mean of the estimated regression coefficients decreases as the percentage of missingness increase for all the three methods as evident in tables 90, 93 and 96. The PDI for the Bayesian linear regression and linear regression, non-

Bayesian are close and lower than the PDI of the PMM model as shown in tables 91, 92 and 93. This shows both imputation models can be used for this type of data. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 92,95 and 98 shows the p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 90: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 150 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	383.9666	-0.1299	0.0967	-0.0402	-0.8830
20%	328.9693	-0.2922	0.1496	-0.0324	-0.7871
30%	339.078	-0.1396	0.1386	-0.0282	-0.7925
40%	310.7136	-0.1569	0.1645	-0.0056	-0.6930
50%	275.4081	-0.1425	0.1990	-0.0195	-0.6535
Actual Parameter	**243.31570**	**-0.52660**	**0.32615**	**-0.23567**	**-3.7533**

Table 91: PDI of the Regression Coefficients of PMM Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.5781	0.7533	0.7035	0.8294	0.7647	0.4945
20%	0.3520	0.4451	0.5413	0.8625	0.7903	0.5982
30%	-0.3936	0.7349	0.5750	0.8803	0.7889	0.5171
40%	-0.2770	0.7021	0.4956	0.9762	0.8154	0.5424
50%	-0.1319	0.7294	0.3899	0.9173	0.8259	0.5461
Mean	-0.2057	0.6730	0.5411	0.8931	0.7970	**0.5397**

Table 92: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 93: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 150 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	374.6431	-0.1315	0.1057	-0.03879	-0.8669
20%	302.2398	-0.1746	0.1753	-0.0263	-0.7022
30%	315.6801	-0.1527	0.1610	-0.0262	-0.7155
40%	306.9224	-0.1419	0.1679	-0.0146	-0.6823
50%	236.6761	-0.1425	0.2354	-0.0162	-0.5953
Actual Parameter	**243.31570**	**-0.52660**	**0.32615**	**-0.23567**	**-3.7533**

Table 94: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.5397	0.7503	0.6759	0.8354	0.7690	0.4982
20%	-0.2422	0.6684	0.4625	0.8884	0.8129	0.5180
30%	-0.2974	0.7100	0.5064	0.8888	0.8094	0.5234
40%	-0.2614	0.7305	0.4852	0.9380	0.8182	0.5421
50%	0.0273	0.7294	0.2782	0.9313	0.8414	0.5615
Mean	-0.2627	0.7177	0.4816	0.8964	0.8102	**0.5286**

Table 95: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 96: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 150 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	372.4906	-0.1303	0.1076	-0.0373	-0.8668
20%	311.6313	-0.1551	70.1664	-0.0341	-0.7484
30%	324.4659	-0.14563	0.15234	-0.02980	-0.7293
40%	313.7159	-0.1519	0.1608	-0.0119	-0.6486
50%	235.7427	-0.1483	0.2364	-0.0176	-0.5852
Actual Parameter	**243.31570**	**-0.52660**	**0.32615**	**-0.23567**	**-3.7533**

Table 97: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.5309	7 0.7526	0.6701	0.8417	0.7691	0.5005
20%	-0.2808	0.7055	0.4898	0.8553	0.8006	0.5140
30%	-0.3335	0.7235	0.5329	0.8736	0.8057	0.5204
40%	-0.2893	0.7115	0.5070	0.9495	0.8272	0.5411
50%	0.0311	0.7184	0.2752	0.9253	0.8441	0.5588
Mean	-0.2807	0.7223	0.4950	0.8891	0.8093	**0.5270**

Table 98: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 150 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.6 Analysis of Sample Size 150 and Regular Variability Data

The mean of the estimated regression coefficients for the Bayesian linear regression and linear regression, non-Bayesian methods have the closest values to the complete data at 10% level of missingness as shown in tables 102 and 105. However, the PMM method have most of its estimates closer to the complete data at 30% and 40% level of missingness as shown in table 99. In general, the mean of the estimates for the PMM model decreases as the percentage of missingness increase from 10% to 20% for β_0,

then increases at 30% level of missingness, then decreases at 40% level of missingness and increases at 50% level of missingness with the reverse for β_2 . The mean estimates of the coefficients for β_1 and β_4 decreases as the percentage of missingness increase from 10% to 50% for all the three imputation methods as shown in tables 99, 102 and 105. The mean estimates of β_0 for Bayesian linear regression method in table 102 decreases at 20% level of missingness and increases at 30% before decreasing at 50%. Also, the mean estimates of β_0 for linear regression non-Bayesian method decreases as the percentage of missingness increases from 10% to 30% and increases at 40% level of missingness, then it decreases at 50% level of missingness as shown in table 105. The PDI for the PMM model is the lowest among the three methods at 8.04% which indicates that the PMM model is the best for this data as shown in table 99 for PMM, table 102 for Bayesian linear regression and table 105 for linear regression non-Bayesian model. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 101, 104 and 107 shows the p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 99: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 150 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	322.8130	-0.3630	0.1920	-0.1180	-1.7350
20%	278.5230	-0.3905	0.2280	-0.0426	-1.5428
30%	328.6398	-0.4172	0.1768	-0.0265	-1.4350
40%	307.8475	-0.4524	0.1970	-0.0377	-1.2933
50%	331.5107	-0.4691	0.1745	-0.0507	-1.2484
Actual Parameter	**354.18249**	**-0.29570**	**0.16326**	**-0.15987**	**-1.89436**

Table 100: PDI of the Regression Coefficients of PMM Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0886	-0.2276	-0.1760	0.2619	0.0841	0.0062
20%	0.2136	-0.3206	-0.3965	0.7335	0.1856	0.0831
30%	0.0721	-0.4109	-0.0829	0.8342	0.2425	0.1310
40%	0.1308	-0.5299	-0.2067	0.7642	0.3173	0.0951
50%	0.0640	-0.5864	-0.0688	0.6829	0.3410	0.08654
Mean	0.1138	-0.4151	-0.1862	0.6553	0.2341	**0.0804**

Table 101: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 102: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 150 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	352.2778	-0.3311	0.1626	-0.1227	-1.7856
20%	261.2722	-0.3422	0.2450	-0.0263	-1.6151
30%	278.5606	-0.4018	0.2250	-0.0329	-1.3967
40%	345.7724	-0.3989	0.1786	-0.0284	-1.3898
50%	323.1877	-0.4297	0.2354	-0.0162	-1.2192
Actual Parameter	**354.18249**	**-0.29570**	**0.16326**	**-0.15987**	**-1.89436**

Table 103: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0054	-0.1197	0.0040	0.2325	0.0574	0.0359
20%	0.2623	-0.1573	0.5007	0.8355	0.1474	0.3177
30%	0.2135	-0.3588	-0.3782	0.7942	0.2627	0.1066
40%	0.0237	-0.3490	-0.0940	0.8224	0.2663	0.13388
50%	0.0875	-0.4532	-0.4419	0.8987	0.3564	0.0895
Mean	0.1185	-0.2876	-0.0819	0.7167	0.2180	**0.1367**

Table 104: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 105: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 150 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	357.4404	-0.3548	0.1579	-0.1191	-1.7563
20%	291.8717	-0.3256	0.2151	-0.0633	-1.655
30%	254.0877	-0.4136	0.2495	-0.0328	-1.3835
40%	329.7442	-0.4641	0.1753	-0.0387	-1.2506
50%	286.0532	-0.1483	0.2364	-0.0176	-0.5852
Actual Parameter	**354.18249**	**-0.29570**	**0.16326**	**-0.15987**	**-1.89436**

Table 106: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	-0.0092	-0.1999	0.0328	0.2550	0.0729	0.0332
20%	0.1759	-0.1011	-0.3175	0.6041	0.1264	0.1148
30%	0.2826	-0.3987	-0.5282	0.7948	0.2697	0.0702
40%	0.0690	-0.5695	-0.0737	0.7579	0.3398	0.0956
50%	0.1924	0.4985	-0.4480	0.8899	0.6911	0.3438
Mean	0.1421	-0.1541	-0.2669	0.6603	0.3000	**0.1315**

Table 107: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 150 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.7 Analysis of Sample Size 150 and Large Variability Data

The mean of the estimated regression coefficients for the smaller fraction of missing information (FMI) at 10% have the closest values to the mean of the estimated regression coefficients from the complete data across all the three imputation methods as shown in tables 108, 111 and 114. Also, apart from the linear regression, non-Bayesian method, the PMM and Bayesian linear regression have their lowest values closer to the mean estimates of the complete data at 50% missingness indicating that as the percentage of missingness increases, the less accurate these models will work for this data. The PDI for the linear regression, non-Bayesian model is the lowest among the three models with a mean of -0.19.57 indicating that the Bayesian linear regression method is the best model for this data as shown in table 109, table 112 and table 115. The estimates from the imputed data satisfies the normality assumption by the central limit theorem (CLT) since the number of each estimates is fifty based on based the number imputations done. Tables 110, 113 and 116 shows the p-values from the one sample t-test, which tested if there was difference in estimated mean

coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant.

Table 108: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 150 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	383.9666	-0.1299	0.0967	-0.0402	-0.8830
20%	328.9693	-0.2922	0.1496	-0.0324	-0.7871
30%	339.078	-0.1396	0.1386	-0.0282	-0.7925
40%	310.7136	-0.1569	0.1645	-0.0056	-0.69303
50%	275.4081	-0.1425	0.1990	-0.0195	-0.6535
Actual Parameter	**402.6610**	**-0.10295**	**0.0803**	**-0.0621**	**-0.9903**

Table 109: PDI of the Regression Coefficients of PMM Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0464	-0.2618	-0.2038	0.3534	0.1084	0.0085
20%	0.1830	-1.8383	-0.8623	0.4788	0.2052	-0.3667
30%	0.1579	-0.3560	-0.7254	0.5464	0.1997	-0.0354
40%	0.2283	-0.5240	-1.0478	0.9099	0.3002	-0.0266
50%	0.3160	-0.3842	-1.4773	0.6863	0.3401	-0.1038
Mean	0.1863	-0.5530	-0.8633	0.5950	0.2307	**-0.0808**

Table 110: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 111: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 150 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	374.6431	-0.1315	0.1057	-0.03879	-0.8669
20%	302.2398	-0.1746	0.1753	-0.0263	-0.7022
30%	315.6801	-0.1527	0.1610	-0.0262	-0.7155
40%	306.9224	-0.1419	0.1679	-0.0146	-0.6823
50%	236.6761	-0.1425	0.2354	-0.0162	-0.5953
Actual Parameter	**402.6610**	**-0.10295**	**0.0803**	**-0.0621**	**-0.9903**

Table 112: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0696	-0.2773	-0.3158	0.3761	0.1246	-0.0045
20%	0.2494	-0.6960	-1.1822	0.5770	0.2909	-0.1522
30%	0.2160	-0.4832	-1.0042	0.5786	0.2775	-0.0830
40%	0.2378	-0.3783	-1.0901	0.7652	0.3110	-0.0308
50%	0.4122	-0.3842	-1.9304	0.7394	0.3989	-0.1528
Mean	0.2370	-0.4438	-1.1045	0.6073	0.2806	**-0.0846**

Table 113: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 114: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 150 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	372.4906	-0.1303	0.1076	-0.0373	-0.8668
20%	311.6313	-0.1551	0.1664	-0.0341	-0.7484
30%	324.4659	-0.14563	0.15234	-0.0298	-0.7293
40%	313.7159	-0.1519	0.1608	-0.0119	-0.6486
50%	351.4801	-0.3718	0.1530	-0.0639	-1.3663
Actual Parameter	**402.6610**	**-0.10295**	**0.0803**	**-0.0621**	**-0.9903**

Table 115: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.0749	-0.2657	-0.3395	0.4000	0.1247	-0.0011
20%	0.2261	-0.5066	-1.0715	0.4515	0.2443	-0.13124
30%	0.1942	-0.4143	-0.8964	0.5207	0.2636	-0.0664
40%	0.2209	-0.4755	-1.0017	0.8086	0.3451	-0.0205
50%	0.1271	-2.6115	-0.9046	-0.0278	-0.3797	-0.0205
Mean	0.1686	-0.8547	-0.8427	0.4306	0.1196	**-0.1957**

Table 116: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 150 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.8 Analysis of Sample Size 500 and Small Variability Data

The mean of the estimated regression coefficients decreases as the percentage of missing values increases across all the three imputation methods as shown in tables 117, 120 and 123. The closest value to the mean of the estimated regression coefficients from the complete data is at the 10% level of missingness across all the three imputation methods as shown tables 117, 120 and 123. The PDI for the Bayesian linear regression method in table 121 is the lowest among the three methods with a mean of -0.0338, which indicates that the Bayesian linear regression model is the best for this data. The PDI values for PMM and linear regression non-Bayesian are displayed table 118 and table 124, respectively. The estimates are also normally distributed based on the central limit theorem. The p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant. The p-values are shown in tables 122, 125 and 128.

Table 117: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 500 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	193.677	-0.6165	0.3332	-0.1699	-3.4104
20%	143.6567	-0.6877	0.4134	-0.1313	-3.1272
30%	-49.7714	-0.7116	0.5940	-0.0520	-2.8212
40%	-125.3945	-0.7145	0.6580	0.0201	-2.5510
50%	-175.6273	-0.7226	0.6987	0.0797	-2.2752
Actual Parameter	**388.0682**	**- 0.5034**	**0.1884**	**- 0.2847**	**-3.8898**

Table 118: PDI of the Regression Coefficients of PMM model for 500 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.5009	4.2723	-0.7686	0.4032	0.1232	0.9062
20%	0.6298	4.6502	-1.1943	0.5388	0.1961	0.9641
30%	1.1283	4.7771	-2.1529	0.8174	0.2747	0.9689
40%	1.3231	4.7925	-2.4926	1.0706	0.3442	1.0075
50%	1.4526	4.8355	-2.7086	1.2799	0.4151	1.0549
Mean	1.00694	4.6655	-1.8634	0.8219	0.2706	**0.9803**

Table 119: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 500 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 120: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 500 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	202.1656	-0.6234	0.3613	-0.1824	-3.3975
20%	102.0945	-0.6833	0.4515	-0.1117	-3.0636
30%	-11.5946	-0.6889	0.5554	-0.0513	-2.8390
40%	-174.0286	-0.7299	0.7040	0.0335	-2.4493
50%	-182.9214	-0.7067	0.7059	0.0737	-2.3004
Actual Parameter	**402.6610**	**-0.10295**	**0.0803**	**-0.0621**	**-0.9903**

Table 121: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 500 observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.4790	-0.2384	-0.9177	0.3593	0.1266	-0.0382
20%	0.7369	-0.3574	-1.3965	0.6077	0.2124	-0.0394
30%	1.0299	-0.3685	-1.9480	0.8198	0.2701	-0.0393
40%	1.4484	-0.4499	-2.7367	1.1177	0.3703	-0.0500
50%	1.4714	-0.4039	-2.7468	1.2589	0.4086	-0.0023
Mean	1.0331	-0.3636	-1.9491	0.8326	0.2776	**-0.0338**

Table 122: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 500 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 123: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 500 and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	372.4906	-0.1303	0.1076	-0.0373	-0.8668
20%	311.6313	-0.1551	70.1664	-0.0341	-0.7484
30%	324.4659	-0.14563	0.15234	-0.02980	-0.7293
40%	313.7159	-0.1519	0.1608	-0.0119	-0.6486
50%	235.7427	-0.1483	0.2364	-0.0176	-0.5852
Actual Parameter	**243.31570**	**-0.52660**	**0.32615**	**-0.23567**	**-3.7533**

Table 124: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 500 observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.4089	-0.2191	-0.7818	0.3246	0.1127	-0.0309
20%	0.6644	-0.3649	-1.2511	0.6182	0.2082	-0.0250
30%	0.9552	-0.4068	-1.8052	0.7938	0.2682	-0.0389
40%	1.4131	-0.4309	-2.6699	1.0945	0.3577	-0.0471
50%	1.5911	-0.3544	-2.9973	1.1602	0.3986	-0.0403
Mean	1.0065	-0.3552	-1.9010	0.7982	0.2691	**-0.0365**

Table 125: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 500 Observations and Small Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.9 Analysis of Sample Size 500 and Regular Variability Data

The mean of the estimated regression coefficients decreases as the percentage of missing values increases across all the three methods as shown in tables 126, 129 and 132. The closest value to the mean of the estimated regression coefficients from the complete data is at the 10% level of missingness across all the three imputation methods as evident in tables 126, 129 and 132. The PDI for the Bayesian linear regression method is the lowest among the three methods with a mean of -0.1956 which indicates that the Bayesian linear regression method is the best for this data. The PDI values for PMM, Bayesian linear regression and linear regression non-Bayesian are displayed in tables 127, 130 and 133. The estimates are also normally distributed based on the central limit theorem. The p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the completed data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant. This is shown in tables 128, 131 and 134.

Table 126: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 500 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	304.5357	-0.6134	0.2624	-0.1953	-3.4904
20%	131.6377	-0.7570	0.3368	-0.1154	-3.0052
30%	146.874	-0.7808	0.4029	-0.0434	-2.7989
40%	51.05292	-0.7636	0.4906	-0.02957	-2.5413
50%	91.7104	-0.1425	0.4456	-0.0241	-2.3091
Actual Parameter	**425.7125**	**-0.20880**	**0.09110**	**-0.16635**	**-2.0353**

Table 127: PDI of the Regression Coefficients of PMM Model for 500 observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.2846	-1.9377	-1.8804	-0.1740	-0.7149	-0.8844
20%	0.6908	-2.6255	-2.6970	0.3063	-0.4765	-0.9603
30%	0.6550	-2.7395	-3.4226	0.7391	-0.3752	0.5801
40%	0.8801	-2.6571	-4.3853	0.8227	-0.2486	-1.1176
50%	0.7846	0.3175	-3.8913	0.8551	-0.1345	-0.4137
Mean	0.1604	-0.3689	1.9642	0.8489	0.2545	**0.5718**

Table 128: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 500 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 129: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 500 and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	313.179	-0.5904	0.2550	-0.2109	-3.5566
20%	182.0300	-0.7002	0.3711	-0.0692	-3.0926
30%	125.6478	-0.8030	0.4219	-0.0323	-2.7061
40%	40.4557	-0.1419	-0.7258	0.4982	-0.0242
50%	-8.3924	-0.7853	0.5393	0.0144	-2.1426
Actual Parameter	**425.7126**	**-0.20880**	**0.09110**	**-0.16635**	**-2.03538**

Table 130: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 500 observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.2643	-1.8276	-1.7991	-0.2678	-0.7475	-0.8755
20%	0.5724	-2.3534	-3.0735	0.5840	-0.5195	-0.9580
30%	0.7049	-2.8458	-3.6312	0.8058	-0.3296	-1.0591
40%	0.9050	0.3204	8.9671	3.9949	0.9881	3.0351
50%	1.0197	-2.7610	-4.9199	1.0866	-0.0527	-1.1254
Mean	0.6932	-1.8934	-0.8913	1.2407	-0.1322	**-0.1965**

Table 131: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 500 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 132: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 500 and Regular Variability

FMI \Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	312.9665	-0.6244	0.2553	-0.2010	-3.4994
20%	153.0623	-0.688	0.3990	-0.0721	-3.0826
30%	181.8841	-0.7457	0.3691	-0.0680	-2.8624
40%	342.8777	-0.1519	0.1608	-0.0119	-0.6486
50%	25.6593	-0.8255	0.5085	0.0126	-2.1709
Actual Parameter	**425.7126**	**-0.2088**	**0.0911**	**-0.16635**	**-2.03538**

Table 133: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 500 observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.2648	-1.9904	-1.8024	-0.2083	-0.7194	-0.8911
20%	0.6405	-2.2950	-3.3798	0.5666	-0.5146	-0.9964
30%	0.5728	-2.5714	-3.0516	0.5912	-0.4064	-0.9730
40%	0.8993	0.2725	-0.7651	0.9285	0.6813	0.4033
50%	0.9397	-2.9535	-4.5818	1.0757	-0.0666	-1.1173
Mean	0.6634	-1.9075	-2.7161	0.5907	-0.2051	**-0.7149**

Table 134: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 500 Observations and Regular Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

6.10 Analysis of Sample Size 500 and Large Variability Data

The mean of the estimated regression coefficients for the smallest fraction of missing information (FMI) at 10% have the closest values to the mean of the estimated regression coefficients from the complete data across all the three imputation methods as shown in tables 135, 138 and 141. Also, apart from the PMM method, the Bayesian linear regression and linear regression, non-Bayesian methods have the smallest values closer to the mean estimate of the complete data at 50% level missingness indicating that, as the percentage of missing values increases, the less accurate these models will be. The PDI for the Bayesian linear regression method is the lowest among the three imputation methods with an overall mean of 0.5610 indicating that, the Bayesian linear regression method is the best imputation method for this data. The PDI values for PMM, Bayesian linear regression and Linear regression non-Bayesian are displayed in tables 136, 139 and 141, respectively. The estimates are normally distributed based on the central limit theorem. The p-values from the one sample t-test, which tested if there was difference in estimated mean coefficient and the coefficient from the com-

pleted data set. All the p-values are less than $\alpha = 0.05$ indicating that all differences are significant. The p-values are shown in tables 137, 140 and 143 for PMM, Bayesian linear regression and linear regression non-Bayesian methods, respectively.

Table 135: Estimated Means of the Regression Coefficients from the PMM Model at each Percentage of Missingness for Sample Size 500 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	364.4357	-0.1721	0.1158	-0.0316	-0.7938
20%	362.4600	-0.1742	0.115	-0.0123	-0.7298
30%	339.078	-0.1396	0.1386	-0.0282	-0.7925
40%	357.3081	-0.1883	0.1175	0.0071	-0.6196
50%	350.9477	-0.1903	0.1220	0.0172	-0.5650
Actual Parameter	**422.6556**	**-0.1263**	**0.0610**	**-0.0633**	**-0.9392**

Table 136: PDI of the Regression Coefficients of PMM Model for 500 observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.1377	-0.3626	1.9169	0.5008	0.1548	0.46952
20%	0.1424	-0.3793	1.9105	0.8057	0.2230	0.54046
30%	0.1977	-0.1053	2.0974	0.5545	0.1562	0.5801
40%	0.1546	-0.4909	1.9303	1.1122	0.3403	0.6093
50%	0.1697	-0.5067	1.9660	1.2717	0.3984	0.6598
Mean	0.1604	-0.3689	1.9642	0.8489	0.2545	**0.5718**

Table 137: P-values for One-sample t-test for each Estimated Regression Coefficient of PMM Model for 500 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 138: Estimated Means of the Regression Coefficients from the Bayesian Linear Regression Model at each Percentage of Missingness for Sample Size 500 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	372.4379	-0.1655	0.1081	-0.03879	-0.8669
20%	360.6591	-0.1644	0.1170	-0.0174	-0.7473
30%	369.539	-0.1779	0.11647	-0.0013	-0.6652
40%	359.2788	-0.1910	0.1054	0.0092	-0.6080
50%	342.3959	-0.2004	0.1303	0.0155	-0.5289
Actual Parameter	**422.6556**	**-0.1263**	**0.0610**	**-0.0633**	**-0.9392**

Table 139: PDI of the Regression Coefficients of Bayesian Linear Regression Model for 500 observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.1188	-0.3104	1.8559	0.3872	0.0770	0.4257
20%	0.1467	-0.3017	1.9264	0.7251	0.2043	0.5401
30%	0.1499	-0.4086	1.9216	0.9795	0.2917	0.5868
40%	0.1257	-0.5123	1.8345	1.1453	0.3526	0.5891
50%	0.1899	-0.5867	2.0317	1.2449	0.4369	0.6633
Mean	0.1462	-0.42394	1.91402	0.8964	0.2725	**0.5610**

Table 140: P-values for One-sample t-test for each Estimated Regression Coefficient of Bayesian Linear Regression Model for 500 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

Table 141: Estimated Means of the Regression Coefficients from the Linear Regression, non-Bayesian Model at each Percentage of Missingness for Sample Size 500 and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	372.4906	-0.1303	0.1076	-0.0373	-0.8668
20%	363.3100	-0.1651	0.1144	-0.0166	-0.7456
30%	363.5194	-0.1792	0.1124	-0.0026	-0.6640
40%	348.8168	-0.1800	0.1257	0.005	-0.6241
50%	335.9822	-0.1831	0.1358	0.0194	-0.5465
Actual Parameter	**422.6556**	**-0.1263**	0.0610	**-0.0633**	**-0.9392**

Table 142: PDI of the Regression Coefficients of Linear Regression, non-Bayesian Model for 500 observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Mean**
10%	0.1404	-0.0317	1.8519	0.4107	0.0771	0.4896
20%	0.1399	-0.3072	1.9058	0.7378	0.2061	0.5364
30%	0.1747	-0.4188	1.8899	0.9589	0.2930	0.5795
40%	0.2051	-0.4252	1.9952	1.0790	0.3355	0.6379
50%	0.1404	-0.4497	2.0752	1.3065	0.4181	0.6981
Mean	0.1601	-0.32652	1.9436	0.89858	0.26596	0.5883

Table 143: P-values for One-sample t-test for each Estimated Regression Coefficient of Linear Regression, non-Bayesian Model for 500 Observations and Large Variability

FMI \Est Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
10%	0.00000	0.00000	0.00000	0.00000	0.00000
20%	0.00000	0.00000	0.00000	0.00000	0.00000
30%	0.00000	0.00000	0.00000	0.00000	0.00000
40%	0.00000	0.00000	0.00000	0.00000	0.00000
50%	0.00000	0.00000	0.00000	0.00000	0.00000

7 CONCLUSION AND FUTURE RESEARCH

After analyzing the performance of each of the three multiple imputation methods on all the twelve datasets, the predictive mean matching (PMM) imputation method works best for dataset of sample size 15 and small variability while the Bayesian linear regression, works best for sample size 15 and regular variability. Also, the linear regression non-Bayesian method works best for sample size 15 and regular variability.

With the remaining nine datasets, the mean of the estimated regression coefficients decreases as the percentage of missingness increases across all the three imputation methods. For the sample sizes of 50 with large and regular variability, the Bayesian linear regression and linear regression, non-Bayesian methods were proven to be the best imputation models. Their overall percentage deviation index (PDI) was low as compared with the other methods. The sample size of 50 with small variability data worked best for the linear regression, non-Bayesian method. The sample sizes of 150 and 500 have the linear regression non-Bayesian model as the best imputation method among the three imputation methods. This affirms the study of Addo(2018) that the linear non-Bayesian model works best for data set with large sample when variability is not an issue since data comes from a multivariate normal distribution [1]. Overall, the Bayesian linear regression models produced estimates that are closer to the actual estimates of the complete data while the PMM estimates are far away from the coefficients of the complete data. In summary, while the default multiple imputation method in R is PMM, the Bayesian linear regression method works best

for datasets with large sample sizes regardless of the variability in the data. For future work, multiple imputation methods should be studied on mixed data and categorical data with small and large sample sizes with variability to ascertain which imputation method will work best.

References

[1] Evans Dapaa Addo. Performance comparison of imputation algorithms on missing at random data. 2018.

[2] Paul Allison. Imputation by predictive mean matching: Promise & peril. *Statistical Horizons*, 2015. [Online; accessed November 7, 2020].

[3] Paul D Allison. *Missing data*. Sage publications, 2001.

[4] Donia Smaali Bouhlila and Fethi Sellaouti. Multiple imputation using chained equations for missing data in timss: a case study. *Large-scale Assessments in Education*, 1(1):1–33, 2013.

[5] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68, 2010.

[6] Craig K Enders. *Applied missing data analysis*. Guilford press, 2010.

[7] Walter R Gilks. Markov chain monte carlo. *Encyclopedia of biostatistics*, 4, 2005.

[8] John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

[9] Joseph G Ibrahim, Haitao Chu, and Liddy M Chen. Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, 28(16):2796, 2010.

[10] Roderick JA Little. A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, 83(404):1198–1202, 1988.

[11] Donald B Rubin. Multiple imputation for survey nonresponse, 1987.

[12] Donald B Rubin and Nathaniel Schenker. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American statistical Association*, 81(394):366–374, 1986.

[13] Joseph L Schafer. *Analysis of incomplete multivariate data*. CRC press, 1997.

[14] Stef Van Buuren. *Flexible imputation of missing data*. CRC press, 2018.

[15] Paul T von Hippel. Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1):105–138, 2013.

VITA

VINCENT ONYAME

Education:	B.S. Actuarial Science, University for Development Studies, Tamale, Ghana, 2014
	M.S. Mathematical Sciences, East Tennessee State University Johnson City, Tennessee 2021
Professional Experience:	Assistant Underwriter, Provident Insurance Company Limited Tema, Ghana, 2014–2018
	Tutor, Osino Presbyterian High School Osino, Ghana, 2018–2019
	Graduate Assistant, East Tennessee State University Johnson City, Tennessee , 2019–2021