Original Research Article

Performance Comparison of Imputation Methods for Mixed Data Missing at Random with Small and Large Sample Data Set with Different Variability

Abstrac

One of the concerns in the field of statistics is the presence of missing data, which leads to bias in parameter estimation and inaccurate results. However, the multiple imputation procedure is a remedy for handling missing data. This study looked at the best multiple imputation methods used to handle mixed variable datasets with different sample sizes and variability along with different levels of missingness. The study employed the predictive mean matching, classification and regression trees, and the random forest imputation methods. For each dataset, the multiple regression parameter estimates for the complete datasets were compared to the multiple regression parameter estimates found with the imputed dataset. The results showed that the random forest imputation method was the best for mostly a sample of 150 and 500 irrespective of the variability. The classification and regression tree imputation methods worked best mostly on sample of 30 irrespective of the variability.

yusto
2022-07-31 10:39:32

Please add your Keywords here

1 INTRODUCTION

Missing data in the world of economics, medicine, business management, natural and social sciences has been of growing concern over the years. Missing data is considered as an unstored data value for a variable in observation of interest [1]. As complete data sets are needed to help firms and institutions to produce more accurate and precise results, the presence of missing data rather leads to inaccurate results, bias in parameter estimation and reduction in statistical power. Missing data invariably give rise to reduced sample size and thus, leads to a less precise confidence interval and reduced power in the tests of significance. All these pitfalls lead to incorrect conclusions and invalid recommendations.

Objective of the Study

Considering data with different sample sizes, variability, and different percentages of missingness, the handling of missing data as a part of the preprocessing step can be a tedious task that requires the use of the most appropriate imputation methods to yield accurate and unbiased results. The study assesses the best multiple imputations by chain equation (MICE) procedure for handling missing data for large and small mixed data sets with different variability and with different percentage levels of missingness. One of the fundamental assumptions made was that the missing data were missing at random.

1.1 Overview of Study

Section 2 of the study looks at the three types of missing data. Section 3 examines the methods of imputations. Section 4 explains the multiple imputation chained equation (MICE) methods for mixed data. Section 5 addresses the methodology of the study. Section 6 elaborates on the results and Section 7 highlights the conclusion and future work.

2. TYPES OF MISSING DATA

While the reason for missing data is difficult to establish in a survey with some reasons being the unwillingness on the part of respondents to answer private questions or the forgetfulness to answer certain questions, it is still imperative to carefully examine the pattern of missingness in data to set out the appropriate mechanism to handle such missing data.

According to Rubin [5]; Little and Rubin [28]; Diggle et al. [29]; Diggle and Kenward [30], there are three types of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

When the missingness of data is a result of observed and unobserved (missing) data, then the data is missing completely at random (MCAR) [3]. In this case, the probability of missingness is independent of the observed and unobserved data [4]. MCAR is considered ignorable since no information about the missing data is required. With the missing Y value (Y_{miss}) *and* observed Y value (Y_{obs}) the probability of missing Y value is given as $P\ (R|\emptyset)$ where R is an indicator function with 0 representing a missing value and 1 as an observed value; and $\emptyset$ describes the relation between the data and R. Data that is completely missing at random is considered a simple random sample. For instance, consider marital status as one of the factors that determine the salary of an individual. The assumption of MCAR is satisfied if the individuals who didn't report their salary were unrelated to their marital status. On the other hand, the assumption is being breached if individuals who didn't report their salary on average were younger than the individuals that reported it. To test for MCAR assumption, we separate the data into two categories and then, test the difference between the two groups using a two-sample t-test. If there is a significant difference between the two groups, the MCAR assumption is satisfied.

When missing data is due to observed data but not unobserved data, then the data is missing at random (MAR) [5, 6]. The missing data is conditional on the observed variable. We can denote this as $P\ (Y_{miss}\ |Y,\ X) = P\ (Y_{miss}\ |X)$ where Y is considered as a missing value, but X is always observed. For instance, the assumption of MAR is satisfied when the salary of a respondent which is missing depends on the person's educational status but within each educational status category, the probability of missing data on salary was unrelated to the person's salary. If a data is considered MAR, then some complete case analyses are valid under weaker assumption than MCAR. For instance, linear regression is unbiased if missingness is independent of the response variable but conditional on the predictors. When missing data is ignorable (any information about the missing data is not included when dealing with the missing data) and the missing data does not need to be modeled in the analysis of the dataset, then the MAR assumption is satisfied. However, when the missing data is non-ignorable (any information about the missing data is included when dealing with the missing data), then the modeling of the missing data leads to accurate parameter estimation. As of now, the MAR mechanism cannot be tested. When dealing with data that are completely missing at random, biased parameter estimates are produced and there is also a loss of statistical power.

When missing data is due to unobserved data but not observed data, then the data is missing not at random (MNAR). The probability of missingness is associated with the missing value itself [4]. The MNAR produces small and biased parameter estimates. Data which is MNAR is non-ignorable since information of the missing data is required and most models are also not precise with this form of missingness. The probability distribution of MNAR is given as $P\ (R\ |\ Y_{obs};\ Y_{mis};\ \emptyset)$, where the missing Y value is Y_{miss}; the observed Y value is Y_{obs} ; R represents the missing data indicator and $\emptyset$ describes the relation between the data and R. MNAR data cannot be tested. The assumption of MNAR would be satisfied if individuals with lesser salary do not report their salary.

3. METHODS OF IMPUTATION

Several methods have been proposed on how to handle missing data and can be broken down into two categories: traditional and modern methods. The traditional methods are comprised of the deletion methods (such as pairwise and listwise deletion) and the single imputation methods (such as arithmetic mean imputation, regression imputation, and stochastic regression imputation). The modern methods of handling missing data are further broken into two approaches: joint modeling method and multiple imputation of chained equations (MICE).

Traditional Methods

The two most common traditional methods of handling missing data are the listwise deletion and the pairwise deletion. With the listwise deletion, also known as the complete case analysis (CCA), when at least one value is missing from the entire observation, then the entire observation is dropped from the analysis [7] which is the main shortfall. With this method, there is an assumption that a random sample chosen from the originally targeted sample is collected to represent the complete case, [7] which is not the case in real data since there is often a reason why a data value might be missing.

Another traditional method of handling missing data is the pairwise deletion. The pairwise deletion method involves the removing cases on an analysis-by-analysis basis which minimizes the loss that results from the listwise deletion [8]. In pairwise deletion, variables with missing information are deleted in a specific analysis. Else, variables with complete information have their cases included in the analysis. According to Graham (2009), biased parameter estimates are produced because of the diverse sample sizes used in the pairwise deletion method.

One of the main shortfalls of the two deletion methods is that the data are missing completely at random. However, the MCAR data can lead to reduced sample size, loss of statistical power, and then generate biased parameter estimates [9] and thus, the deletion methods are not ideal in most situations.

The single imputation methods are another traditional way of handling missing data. With the arithmetic mean imputation, all cases of missing values for a particular variable are replaced with the computed arithmetic mean for that particular variable. Since the mean is biased towards outliers, the arithmetic mean method can affect the parameter estimate and variability of the data.

With the regression imputation method, a regression model predicts the missing value, and the estimated response value replaces the missing data. The regression imputation method produces biased parameter estimates even though it is a better method as compared to the arithmetic mean method.

With the deck imputation, values are randomly drawn from the observed values and these values are used to replace the missing values. In hot deck imputation, the observed values are obtained from the same dataset which contains the missing values while in cold deck imputation, the observed value used for the imputation is obtained from an external source (such as data from a previous survey) which does not contain the missing value. The replacement of the missing values with the observed values leads to a narrow interval by underestimating the variability of the completed data [10].

With the stochastic regression imputation, which is a way to improve regression imputation, accounts from the variability in the predicted incomplete values. This method adds a random error to the predicted value from the regression and able to reproduce the appropriate correlation between the missing value and observed terms. The shortfall of the stochastic regression imputation is that the complexity that arises from the several missing data in multivariate data since each missing data require a unique regression equation. With the response pattern imputation, this method can generate relatively accurate parameter estimates with MCAR data and bias estimates when dealing with MAR data [11].

The most obvious drawback of single imputation is the main assumption of considering the true value as the imputed value. This drawback underestimates of the variance, thus affects statistical tests and confidence interval [27].

Modern Methods

The shortcomings associated with the traditional methods of handling missing data led to the adoption and implementation of modern methods to handle missing data with high accuracy.

Joint Modeling

The joint modeling (JM) method of handling missing data is most appropriately used when dealing with time-to-event data (data which occur when attention is fixated on the time elapsing prior to experiencing an event) and longitudinal data since the JM gives an efficient estimate of the treatment effect hence decreases the bias in the treatment effect [12]. The time-to-event component and longitudinal component serve as the two components of the joint modeling method. JM comprises of a linear model with a random effect [12].

The model is built on a multivariate distribution. Mostly, the JM model is based commonly on the multivariate normal distribution, which is used to draw missing data simultaneously from all incomplete variables [14]. With the JM method, the missing data are partitioned into groups of identical patterns and the joint model, which is common to all the observations are used to impute the missing entries with each group of the identical missing data pattern. For more information on JM, see [35, 36].

Multiple Imputation of Chained Equation

Multiple imputation of chained equations (MICE), also known as fully conditional specification (FCS) is used for the computation of multiple imputations instead of a single imputation. The multiple imputation method resolves the impreciseness and uncertainties in single imputation. When the cause for the missing value is unknown, then the multiple imputation method aims to provide valid inference [27]. MICE is required when a multivariate distribution is inappropriate, unknown, or both unlike the JM method that requires the assumption of a known multivariate distribution [16]. Unlike the JM, the MICE method imputes variables one-by-one from series of the univariate conditional distribution. One main advantage of the MICE approach is that the method is flexible to the type of data. It can impute data for binary, categorical, and quantitative variables including data sets with mixed type of data.

The multiple imputation chained equation (MICE) process is illustrated in Figure 1. The first stage, also termed the imputation stage, involves creating a complete data set by substituting the missing values with estimated values using a multiple imputation (MI) method based on the type of variable(s). The second stage, called the analysis stage, involves analyzing the complete data in the first stage with a statistical method of interest. The pooling stage, which is the final stage, generates single point estimates for the missing observations by merging the analyzed results in the second stage.

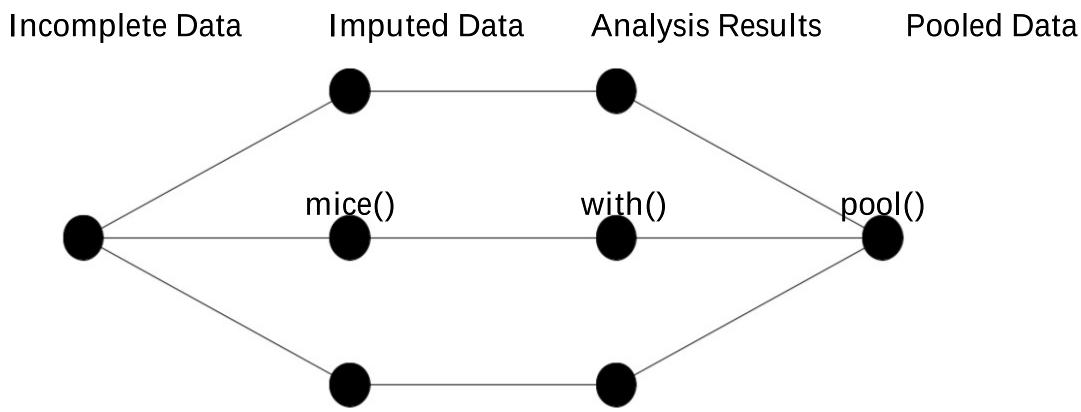


Figure 1: Illustration of MICE procedure

The table 1 indicates some of the imputation methods in the *mice* package.

Table 1: Imputation Methods in the MICE

Model Name	Name of model in R	Variable type
Predictive mean matching	pmm	numeric
Bayesian linear regression	norm	numeric
Unconditional mean imputation	Mean	numeric
Two-level normal imputation	2l. norm	numeric
Multinomial logit model	polyreg	Ordered > 2 levels
Classification and regression trees	cart	any
Linear regression non-Bayesian	norm.nob	numeric
Ordered logit model	polr	factor
Random forest imputations	rf	any
Linear discriminant analysis	lda	factor
Random sample from observed data	sample	any
Logistic regression	logreg	Factor with 2 levels

4. MI METHODS FOR MIXED DATA

Data containing both quantitative and categorical variable (mixed data) that has missing values can be imputed using several different methods. The methods focused in this paper includes classification and regression trees, predictive mean matching, and random forest.

Classification and Regression Trees

Classification and regression trees (CART), similarly identified as decision trees, are used to impute missing values. For the classification tree, the predicted response is the class that contains the data while in the regression tree, the predicted response is a real number. The implementation of the imputation method in CART is done by first using the observed data to fit the classification and regression tree. Then the prediction of the terminal node of the fitted tree where each missing observation finally ends up is determined. Finally, the observed value which is derived from a random draw for the elements in the node is regarded as the imputation [23].

Consider a CART that aims to predict the systolic blood pressure. The CART is illustrated in figure 2. We observe at the first level that the condition under which the subject moves to the next level is conditional on whether the diastolic blood pressure (*dis*) is less than 93 or not. At the second level, the movement to the third stage is dependent on whether *dis* is less than 71 or not. At the third stage, the condition under which the subject moves to the fourth stage is dependent on whether the cu-size (*c1*) is less than 0.5 or blood pressure time (*time*) is less than 510. The classification tree at the fourth level indicates the movement of the subject to the fifth level if the if *time* is less than 691 or *pulse* is less than 67. The final stage predicts the systolic blood pressure if the *pulse* is less than 65 or not. Considering the order of importance in the tree, the diastolic blood pressure variable is most important, followed by cu_ size, blood pressure time, and then pulse.

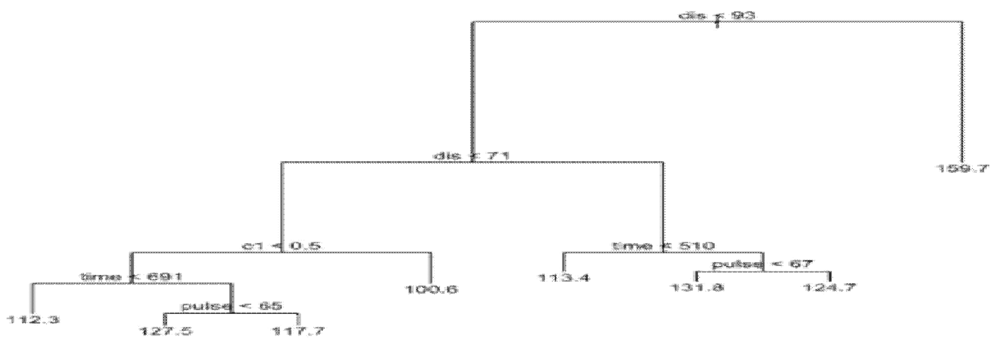


Figure 2: Diagram of Classification and Regression Tree

Predictive Mean Matching

The predicted mean matching (PMM) method takes values from observed data to impute missing values which preserves the distribution of the observed data in the missing, thus enables the PMM method to generate realistic values [2]. With PMM, corresponding values from the complete case that are most similar to the missing values replace these missing values [18]. Even when the structural part of the imputation is incorrect, the PMM preserves the non-linear relation which serves as an advantage for using the PMM method [16]. When the assumption is normality is breached, the PMM is considered more suitable than regression even though the PMM is alike to the regression approach [24]. The imputed values are mostly realistic and a good representation of the possible missing value. On the other hand, the PMM method does not work properly on small sample sizes because the PMM does not emphasize on the between imputation variability with small number of predictors [16].

With the PMM methods, missing values are imputed by regressing incomplete variables on co-variates, thus generating a set of coefficients, β. A random set of coefficients, β^*, are then drawn from the distribution of β. Predicted values for all occurrences in incomplete variables are produced using the new coefficients β^*. The predicted values are then used as a system of measurement to detect complete cases with observed values that are near to the predicted values of each missing case of the target incomplete variable. The missing values are imputed using the observed values of the complete cases. Each of the missing cases is fitted to 5 completed cases with observed values which are close to the predicted values [31].

Random Forest Approaches

The random forest is considered as the collection of several decision trees fit with training data. The random forest is used to impute missing values for continuous variables by drawing randomly from an independent normal distribution, centered on means predicted by the random forest. On the other hand, for categorical variables, the random forest predicts missing values trained on observed values.

Proximity Imputation

With the proximity imputation method, the random forest model can be fit after some method of imputing the missing values has been implemented and this process is termed the pre-imputing of data. The median of the non-missing value is imputed for the quantitative missing values whereas the most occurring non-missing value is imputed for categorical missing values [25]. This is termed as strawman imputation. An $n \times n$ proximity matrix (a square matrix that contains the distance taken pairwise between the elements of the matrix) used to detect structures in the data and symmetry is generated. For each element, i and j that share a common terminal, the (i, j) entry denotes the fraction of tress. One’s expectation is to have the same terminal nodes having similar observations and different terminal nodes having dissimilar observations. The original missing values in the data set are imputed using the proximity matrix [25].

For mixed data, the quantitative variable is imputed using the weighted averages of the non-missing observations, with the weights serving as the proximities while the categorical variable is imputed using the category with the largest mean proximity [25]. A new random forest is generated, and the process is iterated a few times [33]

On-the-fly Imputation

Contrary to the proximity imputation, data is imputed simultaneously while growing the forest when employing the on-the-fly imputation (OTF) [25]. One of the shortcomings of the proximity imputation which includes variable importance (a measure of how much including or removing a variable affect the prediction accuracy) and bias estimates is addressed using one-the-fly imputation. With OTF, observed data is used to calculate the split statistics and imputed values reset to missing after each split. When data is missing, a random value from the in-bag observed data is used to impute the value. If the terminal node is reached, the out-of-bag (OOB) observed terminal node data from all the trees is used to impute the missing values. For quantitative values, the mean observed value is used while the highest observed value is used for categorical values. There is a random selection of the variable used to split each node. There is an iteration of the process where in the first iteration, the estimates used are OOB. Then in-bag estimates are used for additional iteration since there is the non-existence of the OOB estimates [25].

missForest and mForest Imputation

The missForest is usually employed to predict missing values using a random forest trained on the observed values of a data matrix. Apart from its use in imputing mixed data, the missForest can also be used to impute complex interaction and non-linear relations [34]. Compared to the other imputation methods, there are prediction problems associated with the missForest algorithm. First imputing data by regressing each variable against the other variables helps in the prediction of the missing data of the response variable [25]. There could be slowness in computation depending on the amount of data. Considering the case of n variables, each iteration will be well fit if there are n forests. The mForest is usually employed when handling large n values that is a computationally faster form of missForest. With this method, n variables are assigned to groups hence resulting in less forest being fit.

Multivariate splitting is used to grow each forest. There is the exclusion of missing values in the response and the split-rule is averaged over observed responses [25]. Final missing response values are imputed using the prediction method. With less computation, some studies have concluded that the performance of the mForest and the missForest are at par.

5. METHODOLOGY

This section describes the data source, generation of the 9 complete datasets, analysis of the data, and the imputation implementation in the study.

DATA SOURCE AND DESCRIPTION

The data generated for this study is modeled after the 1985 Auto Import Database. This data measures the price of an automobile based on the width of an automobile, engine size, aspiration and drivetrain (denoted as drive wheels). The data can be freely accessed on the UCI Machine Learning Repository at:

https://archive.ics.uci.edu/ml/datasets/Automobile.

For this study, the response variable is the *price* of an automobile while the predictor variables were *width* of an automobile, *engine size, aspiration* and *drive wheels*. *Width* of the automobile and *engine size* are quantitative while *aspiration* and *drive wheels* are categorical. The *aspiration* is a binary variable with categorized as *4wd* and *fwd* while drive wheel is a binary variable categorized as *std* and *turbo*. The entire data set contains 795 observations. The regression model was found to be:

$$\hat{Y} = -68978.03 - 2178.85X_1 + 2208.55X_2 + 1098.56X_3 + 79.85X_4$$

where $\hat{Y}$ is the estimated price of an automobile, X_1 represents aspiration, X_2 represents drive wheels, X_3 represents, and X_4 represents engine size.

Evaluating the Model

One of the vital tests conducted during model selection is the test of the significance of the predictors in the model

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_4 = 0$$

$$H_1: \text{At least one } \beta_j \text{ does not equal 0 for } j=1\ldots,4.$$

The Global *F*-test resulted in a p-value of approximately 0 indicating that at least one predictor is significant in the model.

Since the data set is large (795 observations), the central limit theorem satisfies the assumption of normality. For more information on the central limit theorem, see [3] and [6]. Figure 3 shows the residual plot for the main model and indicates assumption Of constant error variance was met.

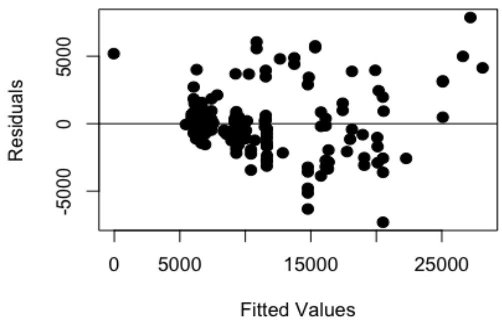


Figure 3: Residual plot for main model

All the 4 predictors have variance inflation factor (VIF) values less than 10 as shown in Table 2, which signifies that there is no serious issue of multicollinearity in the best regression model.

Table 2: VIF values for the best regression model with 4 predictors

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.255447	1.753623	3.041604	3.089541

The ratio of the PRESS statistic and SSE produces a value of 1.114822 (close to 1) which indicates that the regression model has a good predictive ability.

Based on the internally studentized residual, only a few observations had the $|r_i|$ greater than 2.5, hence there are only a few outliers in the response. Also, a few observations were flagged as outliers in X using the leverage value (h_{ii}) as shown in Figure 4 where many of the observations fell above the threshold of 0.012 computed as $(2*p)/n$, where p = 5 and n =795. The Difference in Fits (DFFITS) and Cook's distance were used to check for influential outliers. In figure 5, we noticed lots of observations falling above or below the threshold of +/- 0.158 computed as +/- *(2*sqrt(p/n)))*, where $p = 5$ and $n = 795$. In Figure 6, the threshold for influential observation is 0.8710369 (computed as *qf (0.5, p, n-p)*, and there were no influential observations detected. No action was carried out to eliminate potentially influential observation since the reduced model produced strong results.

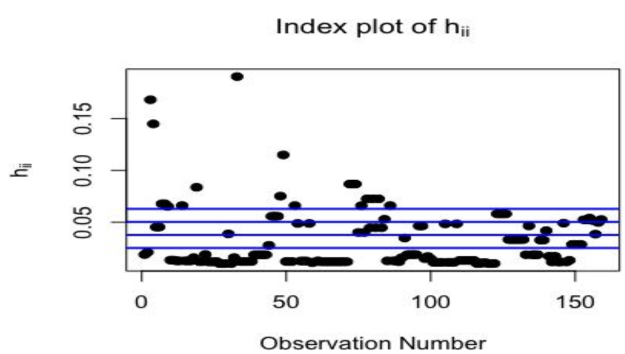
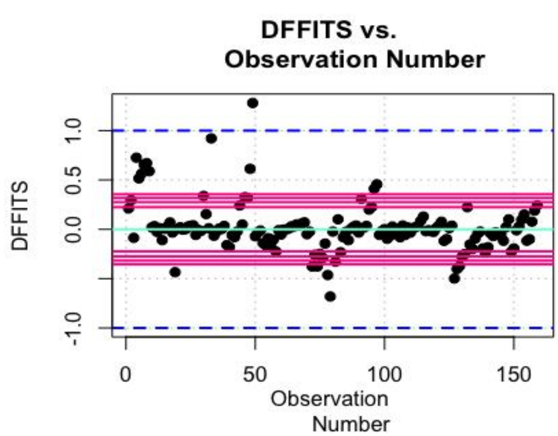


Figure 4: Index plot of h_{ii}



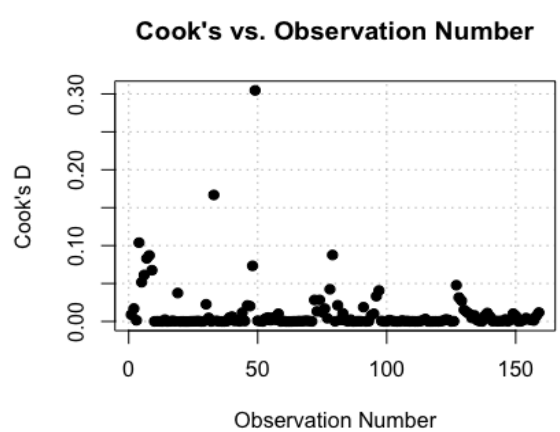


Fig.5: Influential observation by dffits rule Fig.6: Influential observation by cook's rule

GENERATION OF DATA

An R package, SimMultiCorrData, was used to generate the dataset with a specified correlation matrix simultaneously. The continuous variables (width, engine size, and price) were generated with Headrick's fifth order power method transformation using the mean and variance from auto import dataset for each variable while preserving the correlation structure. This method matches the six standardized cumulants (mean, variance, skewness, standardized kurtosis, and standardized fifth and sixth cumulants). We assumed the skewness, standardized kurtosis, and the standardized fifth and sixth cumulants were zero [37]. The categorical variables were simulated by discretizing the standard normal variables at quantiles. These quantiles were found by looking at the inverse standard normal based on the probabilities of success for each variable (aspiration and drive wheels) as described in section 5.1. [37].

A pseudo complete dataset of sample sizes of 30, 150 and 500 was generated. For each sample size, a dataset with small, regular, and large variabilities was also generated. We define the regular variability as the same variability from the auto import dataset for each variable. The small variability was obtained by halving the regular while the large variability was obtained by doubling the regular variability from the auto import dataset. A total of 9 complete datasets were produced for each of the 3 sample sizes of 30, 150 and 500 with each having small, regular, and large variabilities.

MODEL BUILDING FOR THE 9 COMPLETE DATASETS

Using price as the response variable with aspiration, drive wheels, width and engine size as the predictor variables, regression models were fitted for 9 complete datasets.

Model building for 30 observations.

Tables 3, 4, and 5 show the estimated parameters from the regression model along with the *t* test statistic and corresponding p-value for with small, regular, and large variability. All the predictors were needed in the model in the presence of other predictors except for *drive wheels* at the 5% level of significance. Based on Global *F*-test for the three distinct models as shown in table 6, the set of predictor variables were significant in predicting the *price*, hence we left *drive wheels* in the model for comparison of the other sample sizes used. The assumption of constant variance is met based on the random patterns in the residual plots for the sample size of 30 with small, regular and large variability, which is shown in figures 7, 8 and 9, respectively. All the VIF values for the predictors in the three models as indicated in table 7, 8 and 9 were less than 10, hence there is no serious multicollinearity problems.

Table 3: The estimated regression coefficients and p-values for data size of 30 with small variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-82853.22	-3789.19	1832.32	1195.46	144.56
t (P-value)	-2.395 (0.02)	-2.980 (0.00)	1.516(0.14)	2.105(0.04)	3.960(0.00)

Table 4: The estimated regression coefficients and p-values for data size of 30 with regular variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-54833.35	-3789.19	1832.32	845.32	102.22
t (P-value)	-2.232 (0.03)	-2.980 (0.01)	1.516 (0.14)	2.105 (0.04)	3.960 (0.00)

Table 5: The estimated regression coefficients and p-values for data size of 30 with large variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-35020.31	-3789.19	1832.32	597.73	72.28
t (P-value)	-2.002 (0.06)	-2.980 (0.01)	1.516 (0.14)	2.105 (0.04)	3.960 (0.00)

Table 6: Global *F*-test for data size of 30

Variability Type	Small	Regular	Large
F (P-value)	38.5 (0.0000)	38.5 (0.0000)	38.5 (0.0000)

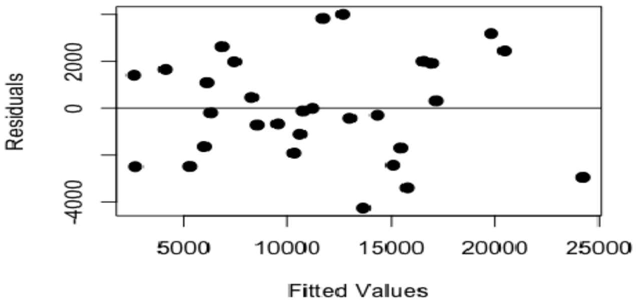


Fig. 7: Residual plot for model of size 30 with small variability

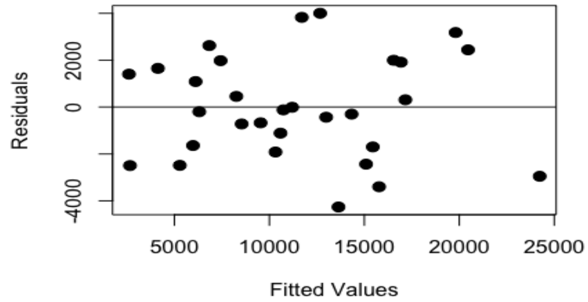


Fig. 8: Residual plot for model of size 30 with regular variability

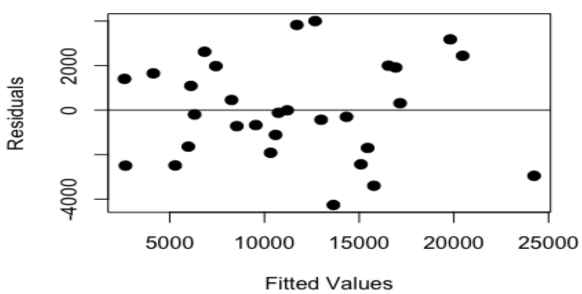


Fig. 9: Residual plot for model of size 30 with large variability

Table 7: VIF for data size of 30 with small variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.203736	1.738988	3.171044	3.203679

Table 8: VIF for data size of 30 with regular variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.203736	1.738988	3.171044	3.203679

Table 9 VIF for data size of 30 with large variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.203736	1.738988	3.171044	3.203679

Model building for 150 observations

Tables 10, 11, and 12 show the estimated parameters from the regression model along with the *t* test statistic and corresponding p-value for with small, regular, and large variability. All the predictors were needed in the model in the presence of other predictors at the 5% level of significance. The Global F-test for the three different models as shown in table 13, the set of predictor variables were significant in predicting the *price*. The assumption of constant variance is met based on the random patterns in the residual plots for the sample size of 150 with small, regular and large variability, which is shown in figures 10,11 and 12 respectively. All the VIF values for the predictors in the three models as indicated in table 14, 15 and 16 were less than 10, thus there is no serious multicollinearity problems.

Table 10: The estimated regression coefficients and p-values for data size of 150 with small variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-110813.37	-1836.14	2562.31	1640.85	103.70
t (P-value)	-7.521 (0.00)	-3.015 (0.003)	4.397 (0.00)	6.829 (0.00)	6.534 (0.00)

Table 11: The estimated regression coefficients and p-values for data size of 150 with regular variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$

Estimate	-75661.50	-1836.14	2562.31	1160.26	73.33
t (P-value)	-7.213 (0.00)	-3.015 (0.003)	4.397 (0.00)	6.829 (0.00)	6.534 (0.00)

Table 12: The estimated regression coefficients and p-values for data size of 150 with large variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-50805.367	-1836.137	2562.310	820.425	51.850
t (P-value)	-6.767 (0.00)	-3.015 (0.003)	4.397 (0.00)	6.829 (0.00)	6.534 (0.00)

Table 13: Global F-test for data size of 150

Variability Type	Small	Regular	Large
F (P-value)	189.2 (0.0000)	189.2 (0.0000)	189.2 (0.0000)

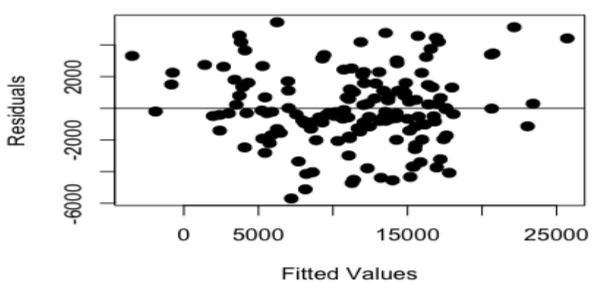


Fig. 10: Residual plot for model of size 150 with small variability

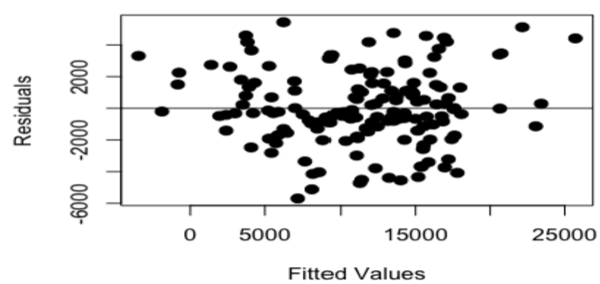


Fig. 11: Residual plot for model of size 150 with regular variability

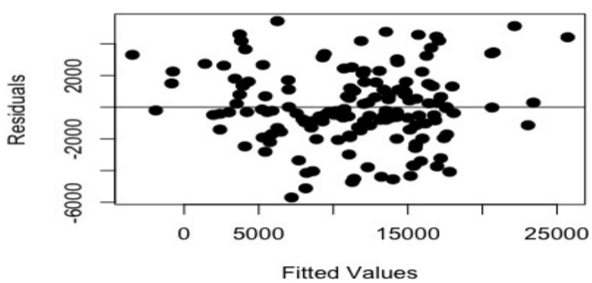


Fig. 12: Residual plot for model of size 150 with large variability

Table 14: VIF for data size of 150 with small variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.309386	1.849866	2.858691	3.049957

Table 15: VIF for data size of 150 with regular variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.309386	1.849866	2.858691	3.049957

Table 16: VIF for data size of 150 with large variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.309386	1.849866	2.858691	3.049957

Model building for 500 observations

Tables 17, 18, and 19 show the estimated parameters from the regression model along with the *t* test statistic and corresponding p-value for with small, regular, and large variability. All the predictors were needed in the model in the presence of other predictors at the 5% level of significance. Based on Global F-test for the three distinct models as shown in table 20, the set of predictor variables were significant in predicting the *price*. The assumption of constant variance is met based on the random patterns in the residual plots for the sample size of 500 with small, regular and large variability, which is shown in figures 13,14 and 15. All the VIF values for the predictors in the three models as indicated in table 21, 22 and 23 were less than 10, hence there is no serious multicollinearity problems.

Table 17: The estimated regression coefficients and p-values for data size of 500 with small variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-103800	-2730	1940	1548	113.7

t (P-value)	-12.814(0.00)	-8.572(0.00)	6.322(0.00)	11.729(0.00)	13.160(0.00)

Table 18: The estimated regression coefficients and p-values for data size of 500 with regular variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-70116.578	-2730.389	1940.197	1094.866	80.427
t (P-value)	-12.154 (0.00)	-8.572 (0.00)	6.322 (0.00)	11.729 (0.00)	13.160(0.00)

Table 19: The estimated regression coefficients and p-values for data size of 500 with large variability.

Regression Coefficient	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Estimate	-46269.082	-2730.389	1940.197	774.187	56.870
t (P-value)	-11.215 (0.00)	-8.572 (0.00)	6.322 (0.00)	11.729 (0.00)	13.160(0.00)

Table 20: Global F-test for data size of 500

Variability Type	Small	Regular	Large
F (P-value)	641.7 (0.0000)	641.7 (0.0000)	641.7 (0.0000)

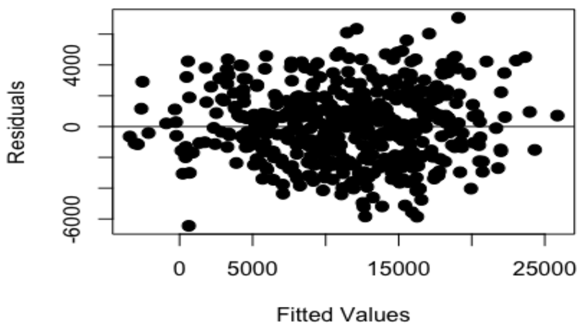


Fig. 13: Residual plot for model of size 500 with small variability

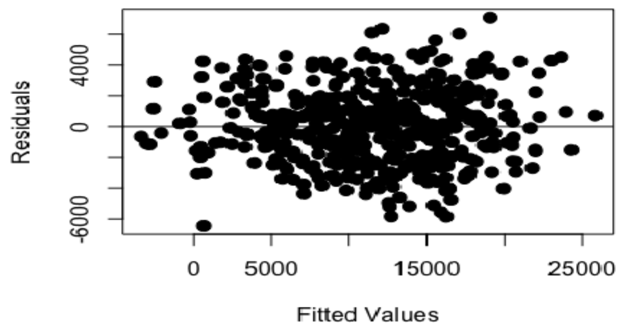


Fig. 14: Residual plot for model of size 500 with regular variability

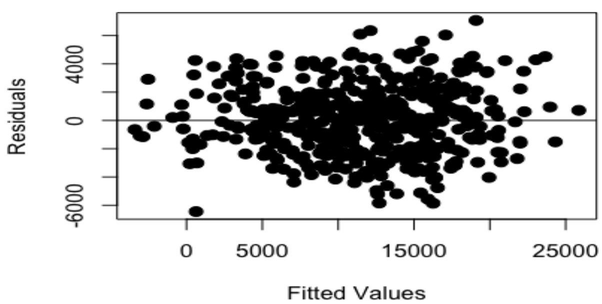


Fig. 15: Residual plot for model of size 150 with large variability

Table 21: VIF for data size of 500 with small variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.247094	1.707726	2.929908	3.070909

Table 22: VIF for data size of 500 with regular variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.247094	1.707726	2.929908	3.070909

Table 23: VIF for data size of 500 with large variability

Variable	Aspiration	Drive Wheels	Width	Engine Size
VIF Value	1.247094	1.707726	2.929908	3.070909

RELATIVE EFFICIENCY

Given the number of imputations, *m*, and the fraction of missingness (FMI), the relative efficiency (RE) determines the best imputation procedure that produces the most precise results based on the measure of the differences in accuracy [31]. The RE is defined as

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

where λ is the fraction of missingness. For each imputation value, as the fraction of missingness increase, the RE tends to decrease accordingly as shown in table 24. For the purpose of this paper, we used an *m* value of 50 because the large value of *m* tends to yield more precise standard error and p-values [31,32].

Table 24: Relative efficiency for different levels of FMI and m

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

IMPUTATION IMPLEMENTATION

For each of the 9 complete data sets of sample sizes 30, 150 and 500 and variabilities of small, regular, and large, the first level of missingness was achieved by removing 10% of the observations from the predictor variables using the R function *prodNA*. The next 20%

level of missingness was achieved by removing 10% level of missingness from the initial 10% removed. The next 30% level of missingness was also achieved by removing 10% level of missingness from the previous 20% and this continued in that sequence till 50% level of missing was attained. This produced a total of 45 missing datasets. Each of the three imputation methods for mixed dataset namely, the predictive mean matching (PMM), classification and regression tree (CART) and the random forest (RF) imputation methods were applied on the 45 missing datasets. For each imputation method, m = 50 imputed data sets were created. We then fit a regression model (as described in 5.1) for each of the 50 imputed datasets. The regression coefficient estimates ($\hat{\beta}0$ to $\hat{\beta}4$) from the 50 imputed data sets were then pooled together and stored. This is repeated for 1000 iterations and the average of each of the 1000 regression coefficients for each variable were computed and compared to the coefficients of the complete data set found in 5.1.

ANALYSIS OF INTEREST

The best imputation method for imputing the missing data for a specified percentage of missingness is the one that produces the average regression coefficient from the imputed data, which is closest to the corresponding regression coefficient from the complete data. To evaluate this comparison, we compute the percentage deviation index (PDI), which is a measure of how far the average of the estimated regression coefficient from the imputed data is away from the regression coefficient estimates from the complete data. The PDI is calculated as:

$$PDI = \frac{Mean\ of\ Estimated\ Regression\ Coefficient - Original\ Regression\ Coefficient}{Original\ Regression\ \ Coefficient} * 100.$$

For each of the complete datasets, the best imputation method is the one with the PDI closest to zero. The R^2 value measures the prediction accuracy for a regression model and was computed for each of the 45 datasets.

6. RESULTS

This section of the study evaluates the analysis on the 45 multiple imputed datasets compared to the 9 complete data sets using the methods described in section 5.6.

Analysis for Sample size of 30 with small variability

We see from table 25 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}0$, $\hat{\beta}1$ and $\hat{\beta}2$were closest in value to the mean regression coefficient from the complete dataset. At 20% and 40% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}3$ and $\hat{\beta}4$ were closest in values to the mean regression coefficients from the complete dataset, respectively. Generally, the estimated mean regression coefficients decreased as the level of missingness increased from 10% t0 40% and then increased from 40% to 50% level of missingness for the PMM method.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}1$, $\hat{\beta}2$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness. At 30% level of missingness, the estimated mean regression coefficients for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the CART methods as shown in table 27.

Considering the imputed dataset for the RF method as shown in table 29, the estimated mean regression coefficients for $\hat{\beta}1$, $\hat{\beta}2$were closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness. At 20% level of missingness, the estimated mean regression coefficients for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset and finally at 30%

level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 20% and then decreased from 20% to 30% level of missingness and then increased from 30%-50% level of missingness for the RF method.

As indicated in tables 26,28 and 30, the PDI of the CART method is closest to zero among the three imputation methods which implied that the PMM is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 25. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with small variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-113860.7	-2580.082	2135.510	1696.407	112.207	0.8323
20%	-156668.1	-1611.734	2334.245	2389.058	77.746	0.8450
30%	-133906.7	-1747.793	2339.214	2008.822	101.513	0.7879
40%	-145104.5	-1998.508	1652.049	2193.880	104.674	0.7488
50%	-108826.2	-3152.522	1689.462	1620.482	128.020	0.6980
Actual Parameter from complete data set	-82853.22	-3789.19	1832.32	1195.46	144.56	0.838

Table 26. PDI for the estimated regression coefficients for sample size of 30 with small variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	0.3742	-0.3191	0.1655	0.4190	-0.2238	0.0832
20%	0.8909	-0.5746	0.2739	0.9984	-0.4622	0.2253
30%	0.6162	-0.5387	0.2766	0.6804	-0.2978	0.1473
40%	0.7513	-0.4726	-0.0984	0.8352	-0.2759	0.1479
50%	0.3135	-0.1680	-0.0780	0.3555	-0.1144	0.0617
Mean	0.5892	-0.4146	0.1079	0.6577	-0.2748	0.1331

Table 27. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with small variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-99337.21	-2300.507	2277.565	1475.387	109.7196	0.8061
20%	-101935.9	-1880.861	3214.774	1539.970	79.2240	0.7684
30%	-93821.36	-1435.086	3755.843	1427.984	61.9492	0.6861
40%	-54746.93	-1745.642	3510.737	799.4550	85.2806	0.6148
50%	-44128.60	-2197.659	2989.593	658.9630	85.8674	0.5236
Actual Parameter from complete data set	-82853.22	-3789.19	1832.32	1195.46	144.56	0.838

Table 28. PDI for the estimated regression coefficients for sample size of 30 with small variability from the CART model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.1989	-0.3929	0.2430	0.2341	-0.2410	0.0844
20%	0.2303	-0.5036	0.75448	0.2881	-0.4519	0.0635
30%	0.1324	-0.6213	1.04977	0.1945	-0.5714	0.0368
40%	-0.3392	-0.5393	0.9160	-0.3312	-0.4100	-0.0141
50%	-0.4674	-0.4200	0.6315	-0.4487	-0.4060	-0.222
Mean	-0.0489	-0.4954	0.7189	-0.0126	-0.4161	-0.0508

Table 29. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with small variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-96548.73	-4196.755	1540.982	1404.631	154.8746	0.8254
20%	-91545.67	-310.2089	196.2865	1328.216	138.9847	0.7661
30%	-110879.9	-1896.349	2704.616	1641.696	104.2328	0.6897
40%	-102228.0	-2204.240	3366.193	1513.572	97.1247	0.6059
50%	-100531.4	-802.0412	4489.046	1513.686	52.7117	0.5316
Actual Parameter from complete data set	-82853.22	-3789.19	1832.32	1195.46	144.56	0.838

Table 30. PDI for the estimated regression coefficients for sample size of 30 with small variability from the RF model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.1652	0.1075	-0.1589	0.1749	0.0713	0.0720
20%	0.1049	-0.9181	-0.8928	0.1110	-0.0385	-0.327
30%	0.3382	-0.4995	0.4760	0.3732	-0.2789	0.0818
40%	0.2338	-0.4182	0.8371	0.2661	-0.3281	0.118
50%	0.2133	-0.7883	1.4499	0.2661	-0.6353	0.0101
Mean	0.2111	-0.5033	0.3422	0.2383	-0.2419	0.00928

Analysis for Sample size of 30 with regular variability

We see from table 31 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_4$ were closest in value to the mean regression coefficient from the complete dataset. At 30% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}_2$ was closest in values to the mean regression coefficients from the complete dataset and finally at 50% level of missingness, the estimated mean regression coefficient for $\hat{\beta}_1$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness decreased from 10% t0 30% and then increased from 30% to 50% level of missingness for the PMM method.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}$, $\hat{\beta}_3$ and $\hat{\beta}_4$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the CART methods as shown in table 33.

Considering the imputed dataset for RF method as shown in table 35, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness. Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 20% and then decreased from 20% to 30% level of missingness and then increased from 30%-50% level of missingness for the RF method.

As indicated in tables 32, 34 and 36, the PDI of the CART method is closest to zero among the three imputation methods which implied that the CART is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 31. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with regular variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-65531.45	-3038.283	2443.910	1025.445	78.6124	0.8102
20%	-79177.26	-2968.759	2297.999	1222.981	85.9305	0.7941
30%	-95356.30	-1878.115	2204.837	1488.819	65.9907	0.7214
40%	-84585.33	-3184.238	2647.687	1339.289	65.1588	0.6191
50%	-83597.78	-3226.548	2690.413	1326.301	63.4450	0.5991
Actual Parameter from complete data set	-54833.35	-3789.19	1832.32	845.32	102.22	0.838

Table 32. PDI for the estimated regression coefficients for sample size of 30 with regular variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	0.1951	-0.1981	0.3337	0.2130	-0.2309	0.0626
20%	0.4439	-0.2165	0.2541	0.4467	-0.1593	0.154
30%	0.7390	-0.5043	0.2033	0.76124	-0.3544	0.169
40%	0.5425	-0.1596	0.4449	0.5843	-0.3625	0.210

50%	0.5245	-0.1484	0.4683	0.5689	-0.3793	0.207
Mean	0.4890	-0.2454	0.3409	0.5148	-0.2973	0.160

Table 33. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with regular variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-50560.78	-3942.212	2059.869	763.1116	113.1268	0.8231
20%	-81491.53	-2217.607	844.8390	1269.244	94.4400	0.7310
30%	-71488.29	-1164.968	2168.806	1102.110	73.4001	0.5554
40%	-65110.92	-1713.514	3629.544	998.3172	56.6923	0.4941
50%	-48822.49	-1604.709	4610.970	744.3413	45.9696	0.4610
Actual Parameter from complete data set	-54833.35	-3789.19	1832.32	845.32	102.22	0.838

Table 34. PDI for the estimated regression coefficients for sample size of 30 with regular variability from the CART model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0779	0.0403	0.1241	-0.0972	0.1066	0.0192
20%	0.4861	-0.4147	-0.5389	0.5014	-0.0761	-0.008.4
30%	0.3037	-0.6925	0.1836	0.3037	-0.2819	-0.03.67
40%	0.1874	-0.5477	0.980	0.1809	-0.4453	0.0712
50%	-0.1096	-0.5765	1.5164	-0.1194	-0.5502	0.0321
Mean	0.1579	-0.4382	0.4532	0.1539	-0.2494	0.0155

Table 35. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with regular variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-64149.01	-4196.755	1540.982	993.2241	109.5129	0.8253
20%	-61169.23	-3102.089	1962.865	939.1905	98.2770	0.7661
30%	-75693.21	-1896.349	2704.616	1160.854	73.7037	0.6897
40%	-69751.49	-2204.240	3366.193	1070.257	68.67754	0.6059
50%	-69673.64	-829.6779	4482.022	1071.819	37.3873	0.5316
Actual Parameter from complete data set	-54833.35	-3789.19	1832.32	845.32	102.22	0.838

Table 36. PDI for the estimated regression coefficients for sample size of 30 with regular variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	0.1698	0.1075	-0.1589	0.1749	0.0713	0.0730
20%	0.1155	-0.1813	0.0712	0.1110	-0.0385	0.0156
30%	0.3804	-0.4995	0.4760	0.3732	-0.2789	0.0902
40%	0.2720	-0.4182	0.8371	0.2660	-0.3281	0.126
50%	0.2706	-0.7810	1.4460	0.2679	-0.6342	0.114
Mean	0.2417	-0.3545	0.5343	0.2386	-0.2417	0.0837

Analysis for Sample size of 30 with large variability

We see from table 37 that at the 20% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}0$, $\hat{\beta}2$ and $\hat{\beta}3$were closest in value to the mean regression coefficient from the complete dataset. At 30% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}3$ was closest in values to the mean regression coefficients from the complete dataset and finally at 10% level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increased from 10% t0 40% and then increased from 40% to 50% level of missingness for the PMM method.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missing. At 30% level of missingness, the estimated mean regression coefficients for $\hat{\beta}2$ and $\hat{\beta}3$ was closest in value to the mean regression coefficients from the complete dataset and at 40% level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 30% and then increased from 30% to 50% level of missingness for the CART methods as shown in table 39.

Considering the imputed dataset for RF method as shown in table 41, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness. Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 20% and then decreased from 20% to 30% level of missingness and then increased from 30%-50% level of missingness for the RF methods.

As indicated in tables 38,40 and 42, the PDI of the CART method is closest to zero among the three imputation methods which implied that the CART is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 37. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with large variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-22117.53	-4860.589	2570.559	390.3774	76.7728	0.8572
20%	-33252.45	-4341.491	2186.129	582.6866	60.7146	0.7681
30%	-45248.03	-3807.438	1299.328	785.1202	55.9785	0.7536
40%	-46400.55	-3520.290	1376.959	809.2406	49.5510	0.6960
50%	-41112.64	-5470.598	1209.332e	758.7642	51.5799	0.6.663
Actual Parameter from complete data set	-35020.31	-3789.19	1832.32	597.73	72.28	0.838

Table 38. PDI for the estimated regression coefficients for sample size of 30 with large variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.3684	0.2827	0.4028	-0.3469	0.0621	0.0065
20%	-0.0504	0.1457	0.1930	-0.0251	-0.1600	0.0206
30%	0.2920	0.0048	-0.2908	0.3135	-0.2255	0.0188
40%	0.3249	-0.0709	-0.2485	0.3538	-0.3144	0.0089
50%	0.1739	0.4437	-0.3399	0.2694	-0.2863	0.0521
Mean	0.0744	0.1612	-0.0566	0.1129	-0.1848	0.0214

Table 39. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with large variability from the CART method

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-31946.36	-3942.212	2059.86	539.6014	79.9927	0.8231
20%	-53975.19	-2245.837	759.7	901.7992	67.3925	0.7319
30%	-56282.21	-2852.966	-1100.05	972.6104	79.5794	0.6652
40%	-54289.70	-3553.769	387.86	932.2304	69.5446	0.6848
50%	-38753.6	-4100.374	1433.102	667.9877	76.5142	0.6181
Actual Parameter from complete data set	-35020.31	-3789.19	1832.32	597.73	72.28	0.838

Table 40. PDI for the estimated regression coefficients for sample size of 30 with large variability from the CART model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0877	0.0403	0.1241	-0.0972	0.1067	0.0173
20%	0.5412	-0.4073	-0.5853	0.50870	-0.0676	-0.00206
30%	0.6071	-0.2470	-1.6003	0.6271	0.1009	-0.102
40%	0.5502	-0.0621	-0.7883	0.5596	-0.0378	0.0443
50%	0.1066	0.0821	-0.2178	0.1175	0.0585	0.0294
Mean	0.3434	-0.1188	-0.6135	0.3431	0.0321	-0.00271

Table 41. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 30 with large variability from the RF method

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-41238.96	-4196.755	1540.982	702.3155	77.4373	0.8253
20%	-39689.85	-3102.089	1962.865	664.1080	69.4923	0.7661
30%	-50812.45	-1896.349	2704.616	820.8479	52.1164	0.6896

40%	-46787.16	-2204.240	3366.193	756.7860	48.5623	0.6059
50%	-47821.45	-827.1917	4467.267	759.0041	26.4521	0.5316
Actual Parameter from complete data set	-35020.31	-3789.19	1832.32	597.73	72.28	0.838

Table 42. PDI for the estimated regression coefficients for sample size of 30 with large variability from the RF model

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	0.1775	0.1075	-0.1589	0.1749	0.0713	0.0745
20%	0.1333	-0.1813	0.0712	0.1110	-0.0385	0.0191
30%	0.4509	-0.4995	0.4760	0.3732	-0.2789	0.104
40%	0.3360	-0.4182	0.8371	0.2661	-0.3281	0.139
50%	0.3655	-0.7816	1.4380	0.2698	-0.6340	0.132
Mean	0.2926	-0.3546	0.5326	0.2390	-0.2416	0.0936

Analysis for Sample size of 150 with small variability

We see from table 43 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}1$, $\hat{\beta}2$ and $\hat{\beta}4$ were closest in value to the mean regression coefficient from the complete dataset. At 30% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}3$ was closest in values to the mean regression coefficients from the complete dataset and finally at 40% level of missingness, the estimated mean regression coefficient for $\hat{\beta}0$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 30% and then decreased from 30% to 40% level of missingness and increased as the level of missingness increased from 40% to 50%for the PMM method.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 40% level of missing. At 30% level of missingness, the estimated mean regression coefficient for $\hat{\beta}2$ was closest in value to the mean regression coefficients from the complete dataset. At 20% level of missingness, the estimated mean regression coefficients for $\hat{\beta}1$ was closest in value to the mean regression coefficients from the complete dataset and at 10% level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 30% and then decreased from 30% to 40% level of missingness and increased as the level of missingness increased from 40% to 50% for the CART methods as shown in table 45.

Considering the imputed dataset for RF method as shown in table 47, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 40% level of missingness. Generally, the estimated mean regression coefficients decreased as the level of missingness increased from 10% t0 20% and then increased from 20% to 30% level of missingness for the RF method.

As indicated in tables 44, 46 and 48, the PDI of the RF method is closest to zero among the three imputation methods which implied that the RF is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM and CART methods are closest in value to the R^2 value of the complete dataset while the R^2 values for the RF was closest in value to the R^2 value of the complete dataset at 20% level of missingness. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 43. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with small variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-118078.9	-1429.335	3013.996	1760.631	87.9979	0.7946
20%	-115893.9	-1254.654	3098.745	1733.170	82.3694	0.7526
30%	-107945.8	-1072.670	3220.420	1608.224	81.5652	0.7103
40%	-111749.1	-1097.122	3077.236	1699.409	66.7475	0.6694
50%	-104524.2	-863.6152	3492.913	1571.097	67.6798	0.6143
Actual Parameter from complete data set	-110813.37	-1836.14	2562.31	1640.85	103.70	0.8348

Table 44. PDI for the estimated regression coefficients for sample size of 150 with small variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0655	-0.2215	0.1762	0.0729	-0.1514	-0.0116
20%	0.0458	-0.3166	0.2093	0.0562	-0.2056	-0.0422
30%	-0.0258	-0.4158	0.2568	-0.0198	-0.2134	-0.0836
40%	0.0084	-0.4024	0.2009	0.0356	-0.3563	-0.103
50%	-0.0567	-0.5296	0.3631	-0.0425	-0.3473	-0.123
Mean	0.0074	-0.3772	0.2413	0.0205	-0.2548	-0.0726

Table 45. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with small variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-102907.7	-1628.733	2982.426	1516.744	97.2257	0.8155
20%	-100714.5	-1780.314	3134.771	1485.533	96.0357	0.7787
30%	-95351.90	-2612.785	2525.731	1437.250	96.1608	0.7038
40%	-108170.6	-1544.029	2773.228	1630.390	83.6000	0.6657
50%	-82833.52	-1999.472	3301.893	1246.614	79.0813	0.58025
Actual Parameter from complete data set	-110813.37	-1836.14	2562.31	1640.85	103.70	0.8348

Table 46. PDI for the estimated regression coefficients for sample size of 150 with small variability from the CART model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0713	-0.1129	0.1639	-0.0756	-0.0624	-0.0317
20%	-0.0911	-0.0304	0.2234	-0.0946	-0.0739	-0.0133

30%	-0.1395	0.4229	-0.0142	-0.1240	-0.0727	0.0145
40%	-0.0238	-0.1590	0.0823	-0.0063	-0.1938	-0.0602
50%	-0.2524	0.0889	0.2886	-0.2402	-0.2374	-0.0705
Mean	-0.1156	0.0418	0.1488	-0.1082	-0.1280	-0.0322

Table 47. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with small variability from the RF method

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-918036.5	-1202.312	2330.547	1356.271	80.3034	0.65714
20%	-1.23067.1	-1075.600	2645.771	1838.881	89.2696	0.7383
30%	-115982.6	-1435.003	2149.948	1736.514	96.7736	0.68097
40%	-112277.9	-1661.330	2486.606	1661.201	10.4163	0.62659
50%	-76741.67	-915.8124	2124.596	1150.838	55.1839	0.37593
Actual Parameter from complete data set	-110813.37	-1836.14	2562.31	1640.85	103.70	0.8348

Table 48. PDI for the estimated regression coefficients for sample size of 150 with small variability from the RF model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.1715	-0.3451	-0.0904	-0.1734	-0.2256	-0.201
20%	0.1105	-0.4142	0.0325	0.1206	-0.1391	-0.579
30%	0.0466	-0.2184	-0.1609	0.0583	-0.0667	-0.682
40%	0.0132	-0.0952	-0.0295	0.0124	0.0044	-0.189
50%	-0.3074	-0.5012	-0.1708	-0.2986	-0.4678	-0.349
Mean	-0.0617	-0.3148	-0.0838	-0.0561	-0.1789	-0.139

Analysis for Sample size of 150 with regular variability

We see from table 49 that at the 50% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}_1$, and $\hat{\beta}_4$ were closest in value to the mean regression coefficient from the complete dataset. At 10% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}_0$ and $\hat{\beta}_3$ were closest in values to the mean regression coefficients from the complete dataset and at 20% level of missingness, the estimated mean regression coefficient for $\hat{\beta}_2$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 20% and then decreased from 20% to 40% level of missingness and increased as the level of missingness increased from 40% to 50% for the PMM methods.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}2$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missing. At 20% level of missingness, the estimated mean regression coefficient for $\hat{\beta}1$ was closest in value to the mean regression coefficients from the complete dataset. At 30% level of missingness, the estimated mean regression coefficients for $\hat{\beta}0$ was closest in value to the mean regression coefficients from the complete dataset and at 10% level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 20% and then decreased from 20% to 30% level of missingness and increased as the level of missingness increased from 30% to 50% for the CART methods as shown in table 51.

Considering the imputed dataset for RF method as shown in table 53, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}2$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 40% level of missingness and the estimated mean regression coefficients for $\hat{\beta}4$ was closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness. Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 30% and then decreased from 30% to 40% level of missingness and then increased from 40% to 50% for the RF method.

As indicated in tables 50, 52 and 54, the PDI of the RF method is closest to zero among the three imputation methods which implied that the RF is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 49. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with regular variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-68418.93	-1715.239	2869.865	1043.860	71.0455	0.8137
20%	-53680.58	-2487.864	2857.182	826.0395	75.9775	0.7607
30%	-63845.87	-1921.492	2154.049	991.3947	74.8979	0.7207
40%	-65795.56	-1995.437	2010.302	1036.340	68.3756	0.6820
50%	-58605.98	-1816.970	1965.771	915.5013	72.1129	0.6338
Actual Parameter from complete data set	-75661.50	-1836.14	2562.31	1160.26	73.33	0.8348

Table 50. PDI for the estimated regression coefficients for sample size of 150 with regular variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0957	-0.0658	0.1200	-0.1003	-0.0311	-0.0346
20%	-0.2905	0.3549	0.1150	-0.2880	0.0361	-0.0145
30%	-0.1561	0.0464	-0.1593	-0.1455	0.0213	-0.078
40%	-0.1303	0.0867	-0.2154	-0.1068	-0.0675	-0.0867
50%	-0.2254	-0.0104	-0.2328	-0.2109	-0.0165	-0.139
Mean	-0.1796	0.0823	-0.0744	-0.1703	-0.0115	-0.0707

Table 51. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with regular variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-70366.74	-1628.733	2982.426	1072.500	68.7489	0.8155
20%	-68735.21	-1785.593	3135.918	1049.239	67.9262	0.7787
30%	-70875.99	-1162.476	4094.712	1076.384	51.9657	0.6909

40%	-63791.61	-1356.133	3879.111	97.9725	50.1526	0.6331
50%	-61430.02	-1489.916	3255.575	959.2759	51.5242	0.5748
Actual Parameter from complete data set	-75661.50	-1836.14	2562.31	1160.26	73.33	0.8348

Table 52. PDI for the estimated regression coefficients for sample size of 150 with regular variability from the CART model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0699	-0.1129	0.1639	-0.0756	-0.0624	-0.0314
20%	-0.0915	-0.0275	0.2238	-0.0956	-0.0736	-0.0129
30%	-0.0632	-0.3668	0.5980	-0.0722	-0.2913	-0.0391
40%	-0.1568	-0.2614	0.5139	-0.1555	-0.3160	-0.0752
50%	-0.1880	-0.1885	0.2705	-0.1732	-0.2973	-0.115
Mean	-0.1139	-0.1914	0.3540	-0.1144	-0.2081	-0.0548

Table 53. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with regular variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-77032.25	-1472.155	2853.521	1173.842	69.4731	0.8043
20%	-71390.21	-1013.830	3081.882	1064.546e	74.6474	0.7884
30%	-67221.43	-1032.892	2901.799	998.9650	78.5339	0.7420
40%	-70032.56	-966.6632	3077.508	1051.257	70.3241	0.7047
50%	-41941.60	-852.8873	1842.064	633.8131	45.1345	0.3903
Actual Parameter from complete data set	-75661.50	-1836.14	2562.31	1160.26	73.33	0.8348

Table 54. PDI for the estimated regression coefficients for sample size of 150 with regular variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0181	-0.1982	0.1136	0.0117	-0.0525	-0.0215
20%	-0.056	-0.4478	0.2027	-0.0824	0.0179	-0.0732
30%	-0.1115	-0.4374	0.1324	-0.1390	0.0709	-0.0969
40%	-0.0743	-0.4735	0.2010	-0.0939	-0.0409	-0.0964
50%	-0.4456	-0.5354	-0.2810	-0.4537	-0.3845	-0.420
Mean	-0.1339	-0.4185	0.0737	-0.1514	-0.0778	-0.142

Analysis for Sample size of 150 with large variability

We see from table 55 that at the 50% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}1$, and $\hat{\beta}4$ were closest in value to the mean regression coefficient from the complete dataset. At 10% levels of missingness, the estimated mean regression coefficient for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in values to the mean regression coefficients from the complete dataset and at 20% level of missingness, the estimated mean regression coefficient for $\hat{\beta}2$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 20% and then decreased from 20% to 40% level of missingness and increased as the level of missingness increased from 40% to 50% for the PMM methods.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 40% level of missing. At 50% level of missingness, the estimated mean regression coefficient for $\hat{\beta}1$ was closest in value to the mean regression coefficients from the complete dataset. At 30% level of missingness, the estimated mean regression coefficients for $\hat{\beta}2$ was closest in value to the mean regression coefficients from the complete dataset and at 10% level of missingness, the estimated mean regression coefficient for $\hat{\beta}4$ was closest in values to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 20% and then decreased from 20% to 40% level of missingness and increased as the level of missingness increased from 40% to 50% for the CART methods as shown in table 57.

Considering the imputed dataset for RF method as shown in table 59, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness and the estimated mean regression coefficients for $\hat{\beta}4$ was closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness and the estimated mean regression coefficients for $\hat{\beta}2$ was closest in value to the mean regression coefficients from the complete dataset at 40% level of missingness Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 30% and then decreased from 30% to 40% level of missingness and then increased from 40% to 50% for the RF method.

As indicated in tables 56, 58 and 60, the PDI of the RF method is closest to zero among the three imputation methods which implied that the RF is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 55. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with large variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-45884.07	-1715.676	2869.556	738.2494	50.2121	0.8136
20%	-35161.10	-2481.368	285.6624	584.0148	53.7719	0.7608
30%	-42104.09	-1919.455	2155.048	699.6423	53.0512	0.7207
40%	-43593.34	-1990.290	2013.114	734.4150	48.2105	0.6818
50%	-38459.31	-1818.529	1967.988	646.7120	51.0173	0.6337
Actual Parameter from complete data set	-50805.367	-1836.137	2562.310	820.425	51.850	0.8348

Table 56. PDI for the estimated regression coefficients for sample size of 150 with large variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0968	-0.0656	0.1199	-0.1001	-0.0315	-0.0349
20%	-0.3079	0.3514	0.1148	-0.2881	0.0370	-0.0185
30%	-0.1712	0.0453	-0.1589	-0.1472	0.0231	-0.0818
40%	-0.1419	0.0839	-0.2143	-0.1048	-0.0701	-0.0895
50%	-0.2430	-0.0095	-0.2319	-0.2117	-0.0160	-0.142
Mean	-0.1922	0.0811	-0.0740	-0.1704	-0.0115	-0.0734

Table 57. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with large variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-47356.82	-1628.733	2982.426	758.3721	48.6128	0.8155
20%	-46249.62	-1782.709	3137.464	742.6380	48.0028	0.7787
30%	-48010.87	-1784.518	2898.325	777.3723	47.0297	0.7261
40%	-49414.50	-1220.237	3588.942	786.7089	38.5480	0.6658
50%	-40660.05	-1884.456	3068.734	675.9959	39.4289	0.6075
Actual Parameter from complete data set	-50805.367	-1836.137	2562.310	820.425	51.850	0.8348

Table 58. PDI for the estimated regression coefficients for sample size of 150 with large variability from the CART model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0678	-0.1129	0.1639	-0.0756	-0.0624	-0.0310
20%	-0.0896	-0.0290	0.2244	-0.0948	-0.0741	-0.0127
30%	-0.0550	-0.0281	0.1311	-0.0524	-0.0929	-0.0195
40%	-0.0273	-0.3354	0.4006	-0.0410	-0.2565	-0.0520
50%	-0.1996	0.0263	0.1976	-0.1760	-0.2395	-0.0783
Mean	-0.0879	-0.0958	0.2235	-0.0880	-0.1451	-0.0387

Table 59. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 150 with large variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-52049.64	-1472.155	2853.521	830.0320	49.1249	0.8043
20%	-48327.16	-1013.830	3081.882	752.7475	52.7837	0.7884
30%	-45296.77	-1035.526	2895.720	706.7186	55.5795	0.7422

40%	-45412.53	-785.6756	2655.283	715.3982	53.5476	0.6911
50%	-31214.19	-395.5668	2217.603	486.3431	29.8140	0.4196
Actual Parameter from complete data set	-50805.367	-1836.137	2562.310	820.425	51.850	0.8348

Table 60. PDI for the estimated regression coefficients for sample size of 150 with large variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	0.0244	-0.1982	0.1137	0.0117	-0.0526	-0.0202
20%	-0.0487	-0.4478	0.2028	-0.0825	0.0180	-0.0717
30%	-0.1084	-0.4360	0.1301	-0.1386	0.0719	-0.0962
40%	-0.1061	-0.5721	0.0363	-0.1280	0.0327	-0.1470
50%	-0.3856	-0.7846	-0.1345	-0.4072	-0.4250	-0.4270
Mean	-0.1249	-0.4878	0.0697	-0.1489	-0.0710	-0.1530

Analysis for Sample size of 500 with small variability

We see from table 61 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}2$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the PMM methods.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}1$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missing. At 50% level of missingness, the estimated mean regression coefficients for $\hat{\beta}3$ was closest in value to the mean regression coefficients from the complete dataset. At 40% level of missingness, the estimated mean regression coefficient for $\hat{\beta}2$ was closest in value to the mean regression coefficients from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0

20% and then increased from 20% to 50% level of missingness for the CART methods as shown in table 63.

Considering the imputed dataset for RF method as shown in table 65, the estimated mean regression coefficients for $\hat{\beta}2$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness and the estimated mean regression coefficients for $\hat{\beta}1$ was closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness and the estimated mean regression coefficients for $\hat{\beta}0$ was closest in value to the mean regression coefficients from the complete dataset at 40% level of missingness. Generally, the estimated mean regression coefficients decreased as the level of missingness increased from 10% t0 20% and then increased from 20% to 30% level of missingness and then decreased from 30% to 40% and then increased from 40% to 50% for the RF method.

As indicated in tables 62, 64 and 66, the PDI of the RF method is closest to zero among the three imputation methods which implied that the RF is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 61. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with small variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-103893.00	-2462.6620	1957.6970	1553.9660	108.1973	0.7916
20%	-107215.70	-2251.3880	1961.5730	1617.7560	99.1899	0.7551
30%	-105920.50	-2213.5760	1820.1220	1610.0520	94.0290	0.7032
40%	-100686.90	-2082.3870	1833.8440	1531.9900	91.5546	0.6480
50%	-90633.97	-1991.9760	1914.2280	1378.6420	89.5345	0.5882

Actual Parameter from complete data set	-103800.00	-2730.0000	1940.00	1548.00	113.7000	0.8370

Table 62. PDI for the estimated regression coefficients for sample size of 500 with small variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0009	-0.0979	0.0091	0.0039	-0.0484	-0.0265
20%	0.0329	-0.1753	0.0111	0.0451	-0.1276	-0.0428
30%	0.0204	-0.1892	-0.0618	0.0401	-0.1730	-0.0727
40%	-0.0300	-0.2372	-0.0547	-0.0103	-0.1948	-0.1050
50%	-0.1268	-0.2703	-0.0133	-0.1094	-0.2125	-0.1460
Mean	-0.0205	-0.1940	-0.0219	-0.0061	-0.1513	-0.0788

Table 63. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with small variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-105194.70	-2428.8030	2006.8170	1575.6590	106.2510	0.7949
20%	-112654.4	-2355.077	2034.343	1712.408	92.2291	0.7413
30%	-109857.3	-2012.258	2049.193	1671.442	87.8637	0.6935
40%	-100686.10	-2362.6750	1989.8700	1545.3710	83.7254	0.6301
50%	-100038.40	-2089.1280	1900.2290	1546.0220	76.5229	0.5626
Actual Parameter from complete data set	-103800.00	-2730.0000	1940.0000	1548.000	113.7000	0.8370

Table 64. PDI for the estimated regression coefficients for sample size of 500 with small variability from the CART model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0134	-0.1103	0.0344	0.0179	-0.0655	-0.0220
20%	0.0853	-0.1373	0.0486	0.1062	-0.1888	-0.0172
30%	0.0584	-0.2629	0.0563	0.0797	-0.2272	-0.0592
40%	-0.0300	-0.1346	0.0257	-0.0017	-0.2636	-0.0808
50%	-0.0362	-0.2348	-0.0205	-0.0013	-0.3270	-0.1240
Mean	0.0182	-0.1760	0.0289	0.0402	-0.2144	-0.0606

Table 65. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with small variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-95357.64	-2688.8900	1888.4290	1400.0080	123.9874	0.7937
20%	-101278.70	-2452.4290	1967.3700	1500.1890	114.9021	0.7444
30%	-99413.51	-2315.8420	2106.6770	1473.0740	110.5008	0.6914
40%	-100110.30	-2205.1490	2158.1920	1493.0530	103.3730	0.6175
50%	-95610.29	-2124.4950	2400.9050	1430.2050	95.3939	0.5485
Actual Parameter from complete data set	-103800.00	-2730.0000	1940.0000	1548.0000	113.7000	0.8370

Table 66. PDI for the estimated regression coefficients for sample size of 500 with small variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	-0.0813	-0.0151	-0.0266	-0.0956	0.0905	-0.0256
20%	-0.0243	-0.1017	0.0141	-0.0309	0.0106	-0.0264
30%	-0.0423	-0.1517	0.0859	-0.0484	-0.0281	-0.0369
40%	-0.0355	-0.1923	0.1125	-0.0355	-0.0908	-0.0483
50%	-0.0789	-0.2218	0.2376	-0.0761	-0.1610	-0.0600
Mean	-0.0525	-0.1365	0.0847	-0.0573	-0.0358	-0.0395

Analysis for Sample size of 500 with regular variability

We see from table 67 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}0$, $\hat{\beta}1$, $\hat{\beta}2$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the PMM methods.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}2$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 50% level of missing. At 10% level of missingness, the estimated mean regression coefficients for $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset. At 40% level of missingness, the estimated mean regression coefficient for $\hat{\beta}0$ was closest in value to the mean regression coefficients from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the CART methods as shown in table 69.

Considering the imputed dataset for RF method as shown in table 71, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}2$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset at 20% level of missingness and the estimated mean regression coefficients for $\hat{\beta}1$ was closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness and the estimated mean regression coefficients for $\hat{\beta}4$ was closest in value to the mean regression coefficients from the complete dataset at 40% level of missingness. Generally, the estimated mean regression coefficients decreased as the level of missingness increased from 10% t0 20% and then increased from 20% to 50% level of missingness for the RF method.

As indicated in tables 68, 70 and 72, the PDI of the RF method is closest to zero among the three imputation methods which implied that the RF is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 67. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with regular variability from the PMM method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-70244.06	-2461.573	1958.300	1098.599	76.5293	0.7915

20%	-72592.93	-2253.0830	1964.1520	1142.7950	70.1314	0.7549
30%	-71740.61	-2212.2250	1819.5860	1139.1700	66.4620	0.7033
40%	-67961.05	-2085.6120	1831.4270	1081.8520	64.8342	0.6480
50%	-61096.22	-1986.0450	1914.1730	975.9748	63.3122	0.5883
Actual Parameter from complete data set	-70116.58	-2730.3890	1940.1970	1094.866	80.4270	0.8370

Table 68. PDI for the estimated regression coefficients for sample size of 500 with regular variability from the PMM model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0018	-0.0985	0.0093	0.0034	-0.0485	-0.0265
20%	0.0353	-0.1748	0.0123	0.0438	-0.1280	-0.0423
30%	0.0232	-0.1898	-0.0622	0.0405	-0.1736	-0.0724
40%	-0.0307	-0.2361	-0.0561	-0.0119	-0.1939	-0.1060
50%	-0.1286	-0.2726	-0.0134	-0.1086	-0.2128	-0.1470
Mean	-0.0198	-0.1944	-0.0220	-0.0066	-0.1514	-0.0788

Table 69. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with regular variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-71206.46	-2428.8020	2006.8170	1114.1590	75.1308	0.7949
20%	-76576.54	-2352.2960	2035.2870	1211.5850	65.1795	0.7415
30%	-71194.95	-2477.070	2069.316	1129.807	65.0326	0.7032
40%	-69535.82	-2182.6100	1876.5640	1105.6910	63.9851	0.6350
50%	-60128.91	-2302.2250	1909.2660	959.1289	66.6857	0.5858
Actual Parameter from complete data set	-70116.58	-2730.3890	1940.1970	1094.8660	80.4270	0.8370

Table 70. PDI for the estimated regression coefficients for sample size of 500 with regular variability from the CART model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0155	-0.1105	0.0343	0.0176	-0.0659	-0.0218
20%	0.0921	-0.1385	0.0490	0.1066	-0.1896	-0.0161
30%	0.0154	-0.0928	0.0665	0.0319	-0.1914	-0.0341
40%	-0.0083	-0.2006	-0.0328	0.0099	-0.2044	-0.0872
50%	-0.1424	-0.1568	-0.0159	-0.1240	-0.1709	-0.1220
Mean	-0.0055	-0.1398	0.0202	0.0084	-0.1644	-0.0562

Table 71. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with regular variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	R^2 value
10%	-64111.75	-2688.9250	1888.6320	989.7214	87.6866	0.7936
20%	-65815.03	-2592.9720	1915.2900	1021.3500	83.4956	0.7434
30%	-64042.88	-2451.8920	2002.1750	994.2339	81.1726	0.6831
40%	-62631.29	-2319.6130	2081.6430	971.0694	79.0294	0.6260
50%	-59706.33	-2407.291	2174.290	932.3030	75.3107	0.5602
Actual Parameter from complete data set	-70116.58	-2730.3890	1940.1970	1094.8660	80.4270	0.8370

Table 72. PDI for the estimated regression coefficients for sample size of 500 with regular variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}0$	$\hat{\beta}1$	$\hat{\beta}2$	$\hat{\beta}3$	$\hat{\beta}4$	Mean
10%	-0.0856	-0.0152	-0.0266	-0.0960	0.0903	-0.0266
20%	-0.0613	-0.0503	-0.0128	-0.0671	0.0382	-0.0307
30%	-0.0866	-0.1020	0.0319	-0.0919	0.0093	-0.0479
40%	-0.1068	-0.1504	0.0729	-0.1131	-0.0174	-0.0629
50%	-0.1485	-0.1183	0.1207	-0.1485	-0.0636	-0.0716
Mean	-0.0978	-0.0873	0.0372	-0.1033	0.0113	-0.0480

Analysis for Sample size of 500 with large variability

We see from table 73 that at the 10% level of missingness for the PMM method, the estimated mean regression coefficient for $\hat{\beta}1$, $\hat{\beta}2$ and $\hat{\beta}4$ were closest in value to the mean regression coefficient from the complete dataset. At 30% level of missingness, $\hat{\beta}0$ was closest in value to the mean regression coefficient from the complete dataset and at 40% level of missingness, $\hat{\beta}3$ was closest in value to the mean regression coefficient from the complete dataset. Generally, the estimated mean regression coefficients increased as the level of missingness increases from 10% t0 50% for the PMM methods.

For the CART method, the estimated mean regression coefficients for $\hat{\beta}1$, $\hat{\beta}2$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missing. At 40% level of missingness, the estimated mean regression coefficients for $\hat{\beta}0$ and $\hat{\beta}3$ were closest in value to the mean regression coefficients from the complete dataset. Generally, the estimated mean regression coefficients decreased as the level of missingness increases from 10% t0 20% and then increased from 20% to 50% level of missingness for the CART methods as shown in table 75.

Considering the imputed dataset for RF method as shown in table 77, the estimated mean regression coefficients for $\hat{\beta}0$, $\hat{\beta}3$ and $\hat{\beta}4$ were closest in value to the mean regression coefficients from the complete dataset at 30% level of missingness and the estimated mean regression coefficients for $\hat{\beta}1$ and $\hat{\beta}2$ were closest in value to the mean regression coefficients from the complete dataset at 10% level of missingness. Generally, the estimated mean regression coefficients increased as the level of missingness increased from 10% t0 20% and then decreased from 20% to 30% level of missingness and increased from 30% to 50% for the RF method.

As indicated in tables 74, 76 and 78, the PDI of the CART method is closest to zero among the three imputation methods which implied that the CART is the best imputation method when considering this type of data.

At 10% level of missingness, the R^2 values for the PMM, CART and RF methods are closest in value to the R^2 value of the complete dataset. The R^2 values decreased as the level of missingness increased from 10% to 50%.

Table 73. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with large variability from the PMM method

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-48385.21	-2420.7740	1931.9760	811.5718	51.4061	0.8006
20%	-46815.54	-2366.1490	2017.0950	788.2721	49.4639	0.7672
30%	-46809.00	-2166.7340	2113.5000	787.2274	46.3549	0.7187
40%	-45842.28	-2162.6730	2031.9550	782.0615	42.5703	0.6789
50%	-43988.68	-1922.4720	2324.3070	748.3461	38.5242	0.6095
Actual Parameter from complete data set	-46269.08	-2730.3890	1940.1970	774.1870	56.8700	0.8370

Table 74. PDI for the estimated regression coefficients for sample size of 500 with large variability from the PMM model

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0457	-0.1134	-0.0042	0.0483	-0.0961	-0.0239
20%	0.0118	-0.1334	0.0396	0.0182	-0.1302	-0.0388
30%	0.0117	-0.2064	0.0893	0.0168	-0.1849	-0.0547
40%	-0.0092	-0.2079	0.0473	0.0102	-0.2514	-0.0822
50%	-0.0493	-0.2959	0.1980	-0.0334	-0.3226	-0.1010
Mean	0.0021	-0.1914	0.0740	0.0120	-0.1970	-0.0601

Table 75. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with large variability from the CART method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-47173.18	-2428.8040	2006.8170	787.8292	53.1255	0.7949
20%	-50997.24	-2354.0570	2034.3330	856.4394	46.0923	0.7414
30%	-49159.38	-2336.4330	2155.6480	829.4511	43.7683	0.7028
40%	-45739.94	-2252.973	2157.513	775.1587	44.3250	0.6638
50%	-41715.28	-2265.8550	2118.6030	718.0899	42.8254	0.5901
Actual Parameter from complete data set	-46269.08	-2730.3890	1940.1970	774.1870	56.8700	0.8370

Table 76. PDI for the estimated regression coefficients for sample size of 500 with large variability from the CART model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	0.0195	-0.1105	0.0343	0.0176	-0.0658	-0.0210

20%	0.1022	-0.1378	0.0485	0.1062	-0.1895	-0.0141
30%	0.0625	-0.1443	0.1110	0.0714	-0.2304	-0.0260
40%	-0.0114	-0.1749	0.1120	0.0013	-0.2206	-0.0587
50%	-0.0984	-0.1701	0.0920	-0.0725	-0.2470	-0.0992
Mean	0.0149	-0.1475	0.0796	0.0248	-0.1907	-0.0438

Table 77. Estimated mean of regression coefficients for each percentage of missingness for a sample size of 500 with large variability from the RF method.

FMI/ Estimated Parameter/ R^2 value	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	R^2 value
10%	-42023.05	-2690.2220	1887.9330	699.7680	61.9947	0.7937
20%	-41432.82	-2547.8700	2189.3830	681.0770	1.0374	0.7392
30%	-42045.18	-2608.2730	2249.2190	701.5647	55.1047	0.6605
40%	-38799.95	-2608.5330	2156.3560	655.2298	54.4928	0.6083
50%	-38809.45	-2493.2390	2045.1260	661.4966	51.5938	0.5430
Actual Parameter from complete data set	-46269.08	-2730.3890	1940.1970	774.1870	56.8700	0.8370

Table 78. PDI for the estimated regression coefficients for sample size of 500 with large variability from the RF model.

FMI/ Estimated Parameter	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	Mean
10%	-0.0918	-0.0147	-0.0269	-0.0961	0.0901	-0.0279
20%	-0.1045	-0.0668	0.1284	-0.1203	-0.9818	-0.2290
30%	-0.0913	-0.0447	0.1593	-0.0938	-0.0310	-0.0203
40%	-0.1614	-0.0446	0.1114	-0.1537	-0.0418	-0.0580
50%	-0.1612	-0.0869	0.0541	-0.1456	-0.0928	-0.0865
Mean	-0.1220	-0.0516	0.0853	-0.1219	-0.2115	-0.0843

7. CONCLUSION

A performance analysis on the 45 mixed datasets based on the PDI's of the three different imputation methods showed that the CART method was the best imputation method for dataset with sample size of 30 with small, regular and large variabilities as well as datasets with sample size of 50o with large variability. On the other hand, the RF method was found to the best imputation method for datasets with sample size 150 with small, regular and large variabilities. Also, the RF method was the best imputation method for datasets with sample size of 500 with small and regular variabilities.

Even though, the PMM method is considered as the default imputation method in the R package, the RF methods worked best mostly on a sample size of 150 and 500 datasets irrespective of the variability. The classification and regression tree imputation methods worked best mostly on sample size of 30 irrespective of the variability.

For future works, studies should look at the best imputation methods for mixed dataset with a different statistic for measuring categorial variables (such as, the point biserial) and also look at the variability in the response variable. One could also look at different sample sizes as well.

yusto
2022-07-31 10:41:21

Please state here who provided financial or other support for this research.

BIBLIOGRAPHY

[1] The prevention and handling of the missing data, by Hyun Kang, Published by Korean Journal of Anesthesiology, 2013.

[2] Imputation by Predictive Mean Matching: Promise & Peril, by Paul Allison, Published by Statistical Horizons, 2015.

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

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

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

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

[7] A Review of Methods for Missing Data, by Therese D. Pigott, Published by Swets and Zeitlinger, 2001.

[8] Missing Data: Listwise vs. Pairwise, Published by Statistics Solutions, 2019.

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

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

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

[12] Basic Concepts and Methods for Joint Models of Longitudinal and Survival Data, by Joseph G. Ibrahim, Haitao Chu, and Liddy M. Chen, Published by Journal of Clinical Oncology, 2010.

[14] A Comparison of Joint Model and Fully Conditional Specification Imputation for Multilevel Missing Data, by Stephen A. Mistler and Craig K. Enders, Published by Sage Journals, 2017.

[15] How to Handle Missing Data, by Alvira Swalin, Published by Towards Data Science, 2018.

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

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

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

[19] Introduction to Bayesian Linear Regression, by Will Koehrsen, Published by Towards Data Science, 2018.

[20] How to Handle Missing Data, by Alvira Swalin, Published by Towards Data Science, 2018.

[21] Missing Data Part II: Multiple Imputation, by Richard Williams, Published by University of Notre Dame, 2015.

[22] Discriminant Analysis When a Block of Observations is Missing, by Hie-Choon Chung and Chien-Pai Han, Published by Annals of the Institute of Statistical Mathematics, 2000.

[23] Imputation by Classification and Regression Trees, by Stef van Buuren, Published by RDocumentation, 2018.

[24] How do I perform multiple imputations using predictive mean matching in R?, Published by UCLA Institute for Digital Research and Education, 2019.

[25] Random Forest Missing Data Algorithms, by Fei Tang and Hemant Ishwaran, Published by Division of Biostatistics, University of Miami, 2017.

[26] Imputation of missing values for semi-supervised data using the proximity in random forests, by Tsunenori Ishioka, Published by Semantic Scholar, 2012.

[27] O. Harel, and X.H.A. Zhou. Multiple imputation. Review of theory, implementation and software. J. Wiley & Sons, New York, 2005.

[28] R.J.A Little and D.B. Rubin. Statistical Analysis with Missing Data. J. Wiley & Sons, New York, 1987.

[29] P. Diggle, K.-Y. Liang, and S.L. Zeger. Analysis of longitudinal data. Oxford University Press, 1994.

[30] P. Diggle and M.G. Kenward. Informative drop-out in longitudinal data analysis. Applied Statistics, 43(1):49-93, 1994.

[31] Imputation by Predictive Mean Matching: Promise & Peril, by Paul Allison, published by Statistical Horizons, 2015.

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

[33] Tsunenori Ishioka. Imputation of Missing Values for Unsupervised Data Using the Proximity in Random Forests. The Fifth International Conference on Mobile, Hybrid, and On-line Learning, 2013

[34] CRAN, https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

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

[36] 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.

[37] CRAN, https://cran.r-project.org/web/packages/ SimMultiCorrData/SimMultiCorrData.pdf

yusto
2022-07-31 11:23:57

SPECIAL COMMENT
This manuscript has been well written following the way of writing an Original Research Article. All references have been properly and correctly cited. The results and discussions have been presented properly according to the rules of good and correct