5-2019

Comparison of Imputation Methods for Mixed Data Missing at Random

Kaitlyn Heidt
East Tennessee State University

Comparison of Imputation Methods for Mixed Data Missing at Random

---

A thesis

presented to

the faculty of the Department of Mathematics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Kaitlyn Heidt

May 2019

Christina Nicole Lewis, Ph.D., Chair

Robert M. Price, Jr., Ph.D.

JeanMarie L. Hendrickson, Ph.D.

Keywords: Missing data, Multiple imputation methods, Multiple imputation by chained equation, Mixed data

ABSTRACT

Comparison of Imputation Methods for Mixed Data Missing at Random

by

Kaitlyn Heidt

A statistician's job is to produce statistical models. When these models are precise and unbiased, we can relate them to new data appropriately. However, when data sets have missing values, assumptions to statistical methods are violated and produce biased results. The statistician's objective is to implement methods that produce unbiased and accurate results. Research in missing data is becoming popular as modern methods that produce unbiased and accurate results are emerging, such as MICE in R, a statistical software. Using real data, we compare four common imputation methods, in the MICE package in R, at different levels of missingness. The results were compared in terms of the regression coefficients and adjusted R^2 values using the complete data set. The CART and PMM methods consistently performed better than the OTF and RF methods. The procedures were repeated on a second sample of real data and the same conclusions were drawn.

ACKNOWLEDGMENTS

I would like to thank my friends and family for supporting me in school and in life. I would like to give and extra special thanks to Dr. Nicole Lewis not only for being my advisor, but for being a great teacher, role model, and friend throughout my years at ETSU.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Data makes up the world around us. In our continually growing society, data analysis is becoming more important every single day. As companies and businesses continue to flourish, the volume of data they collect expands. However, the expansion of data comes hand in hand with the abundance of missing data. Missing data is the data value that is not stored for a variable in the observation of interest [1]. The research behind missing data is exceptionally important as missing data can lead to many problems. Missing data is an issue across all fields including marketing, health sciences, and political science. When missing data arises, incorrect conclusions are often drawn. A glaring issue in data analysis occurs when researchers have an incomplete data set, but draw conclusions based on the assumption that the data set was complete [1]. While this simplifies the conclusions that researchers make, the conclusions are not valid and are usually incorrect. Missing data causes both bias in the estimation of parameters and a reduction in statistical power.

Chapter 2 addresses the types of missing data and the issues that consequently arise. How these issues are handled, including both traditional and modern methods, are also discussed. We discuss multiple imputation by chained equation approaches in Chapter 3. In Chapter 4, the data set used is introduced along with the analysis of the regression model. Chapter 5 reviews the model selection. Chapters 6 and 7 discuss the evaluations of the data sets chosen. The conclusion and future work sections follow in Chapters 8 and 9.

2 MISSING DATA

There are three types of missing data: missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR). When data are MAR, the missing responses depend on the set of observed responses, but are not related to the specific values expected to be obtained. This is usually the most realistic assumption, and will be the assumption made with the data used in this research. When data are MCAR, the missing responses are not related to either the specific value which is supposed to be obtained or the set of observed responses. The analysis remains unbiased if data are MCAR, but it is often difficult to assume MCAR. When data are MNAR, a problem is presented. When this happens, one must model the missing data to obtain unbiased estimates.

Suppose a survey asks students their gender, height, and weight. Since many females may choose not to answer the question regarding their weight, the values of weight that are missing are related to gender, and not necessarily the value of weight itself. This is an example of data MAR. Now consider an additional question regarding the family income of the students. The missing values are likely directly related to the value of income itself, so these missing values are MNAR. Consider one last question that asks the students to report their favorite color. Since this is not a personal question and there would be no reason for someone to not report his or her answer, missing responses to this question would be MCAR.

2.1 Traditional Methods

There are many methods that have been traditionally used to deal with missing data. However, each of them come with limitations. A major downfall is that the following methods assume that the data is MCAR.

A common method used when data are missing is complete case analysis (CCA). When implementing CCA, entire observations that are missing at least one value are dropped from the analysis [11]. A major disadvantage of using this method is that entire observations are dropped, even if all but one of the values are present and valid. In this method, the researcher assumes that the collection of complete cases is a random sample of the originally targeted sample [11]. However, in real data, this is not always the case since there are often reasons as to why data values are missing.

Mean replacement is another commonly used method to impose when missing data is present. Mean replacement involves replacing all the missing values of a variable by the mean of the values present for that variable [1]. A positive aspect to this method is that all cases can be used in the analysis, even if one or more of its associated values is missing. In this method, the researcher must assume that the mean of the observed observations is a reasonable estimate for the missing cases [1]. While this may sometimes be the case, generally values that are missing are not strictly random, thus leading to inconsistent bias with this method [1].

Pairwise deletion is often used to try to minimize the loss that occurs from CCA [12]. In pairwise deletion, correlation matrices are computed for the pairs of variables that are present. Thus, the corresponding correlation coefficients are not based on the same subjects or number of subjects. However, in this method, software uses the

average sample size across all analyses, therefore standard errors are typically either underestimated or overestimated [12].

2.2 Modern Methods

With many limitations of the traditional methods, modern methods have been developed to handle missing data while trying to alleviate the issues with the traditional methods. The modern methods used when missing data can be split into two categories: joint modeling (JM) and multiple imputation of chained equation (MICE).

2.2.1 Joint Modeling

JM is used when one is studying longitudinal and time-to-event data. JM is an attractive approach to handling the missing data in these studies because JM provides efficient estimates of treatment effects and reduces bias in these treatment effects [13]. The JM consists of a linear model with random effects and has two components: the longitudinal component and the time-to-event component [13]. JM assumes a multivariate normal distribution, as JM draws missing values simultaneously for all incomplete variables using a multivariate normal distribution [24]. It is often difficult to assume a multivariate distribution, thus our focus in this paper will be on MICE.

2.2.2 Multiple Imputation of Chained Equation

MICE is also known as fully conditional specification (FCS). Unlike JM, MICE methods impute variables one at a time from a series of univariate conditional distri-

butions [24]. A main emphasis in MICE is that multiple imputations are computed, rather than a single imputation. Multiple imputations help account for uncertainty in the imputations, since single imputations are often not precise. An advantage of MICE is that the approach is flexible to the type of data.

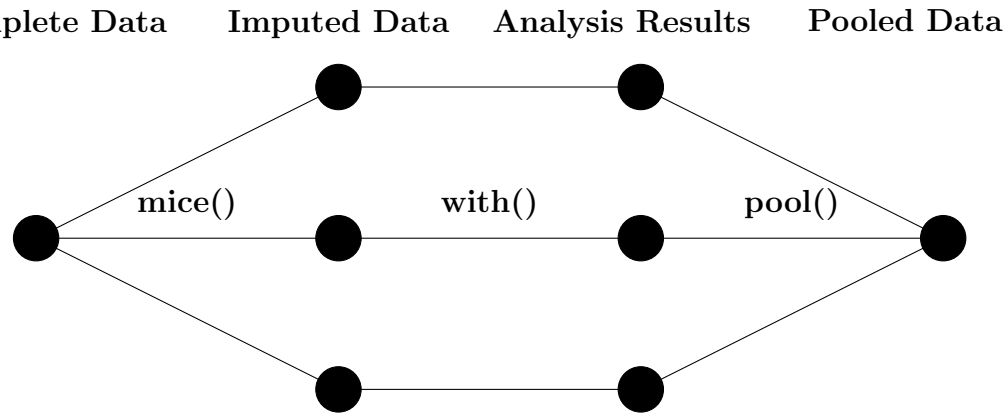

Figure 1: Diagram for MICE method [16].

In Figure 1, we observe the sequence in which incomplete data is imputed, analyzed, and then pooled. The results are given to us in R, and we can specify the number of times we want to impute our data before pooling the results.

There are many modern methods used in handling missing data that are quantitative. First, there is predictive mean matching (PMM). PMM is an attractive approach to use when quantitative variables are not normally distributed. When using PMM, the imputed values are real values that are "borrowed" from individuals with real data. The PMM method finds the complete observation with values closest to the observation with missing values, and then the missing values are replaced with the observed values from the complete case [2]. Then, we have linear regression. Lin-

ear regression involves a predictive analysis in which regression estimates are used to explain the linear relationship between one dependent variable and one or more independent variables. Next, we consider Bayesian linear regression (BLR). BLR is an alternative approach to linear regression. However, in BLR, the distribution of possible models parameters is based on the data and the prior. In BLR, linear regression is formulated using probability distributions rather than point estimates [14]. Then, there is unconditional mean imputation (UMI). When implementing UMI, missing values are replaced with the mean of the observed values for that variable.

Similarly, there are many modern methods used in handling missing data for categorical data. The first of these methods involves implementing a two-level linear model (TLLM). When producing a TLLM, one assumes hierarchical data, usually nested within groups. The response variable is only measured at the lowest level, where as the explanatory variable(s) are measured at all levels. In the TLLM, indicator variables are created for the categorical variables. In the first stage of the procedure, the predicted probabilities of belonging to each category must be computed, corresponding to each indicator variable used in the second stage [15]. The data set is completed by substituting these values in for the missing values. Next, we consider logistic regression. Logistic regression is typically applied to a binary dependent variable and explains the relationship between one dependent binary variable and one or more independent variables. In this method, complete cases are used to estimate the logistic regression model, and then this model is used to predict the missing values [16]. Similarly, one can implement a multinomial logit model (MLM). Implementing an MLM only differs from implementing logistic regression in that the

dependent variable has more than two discrete outcomes. Next, one can implement an ordered logit model (OLM). An OLM is used when the dependent variable has more than two outcomes and the values have a meaningful sequential order where the next values are higher than the previous values. This method uses frequencies to fill in missing values for ordinal variables, such as low, medium, high or a 5-star rating system [17]. Lastly, we consider linear discriminant analysis (LDA). LDA involves preprocessing the data to reduce dimensions. Then one separately analyzes multiple classes of objects by splitting the data into a training set and a testing set. The training set is composed of complete cases, and then the resulting model is used to impute the values in the testing set [18].

Now, the focus of this paper considers modern methods that can be used when working with mixed data, data that consists of both quantitative and categorical variables.

3 MICE METHODS FOR MIXED DATA

There are several methods used to impute mixed data. Mixed data is common, as it allows for both categorical and continuous variables in the data set.

3.1 Classification and Regression Trees

One can use classification and regression trees (CART), also known as decision trees, to impute missing values. In a classification tree, the predicted outcome will be a class to which the data belongs. In regression trees, the predicted outcome is a real number.

In Figure 2, we observe a classification and regression tree that predicts systolic blood pressure of the subject. At the first level, this tree involves regression, in which the subject will be moved to the following level dependent on whether the subject has diastolic blood pressure less than 93 or not. At the third level, the tree involves classification, in which the subjects will move to the fourth level if cuff size is not 1 (c1=0) and will be predicted to have systolic blood pressure of 100.6 if cuff size is 1. Variable importance decreases at each lower level of the tree, meaning diastolic blood pressure is the most important variable in this tree, followed by cuff size and blood pressure time.

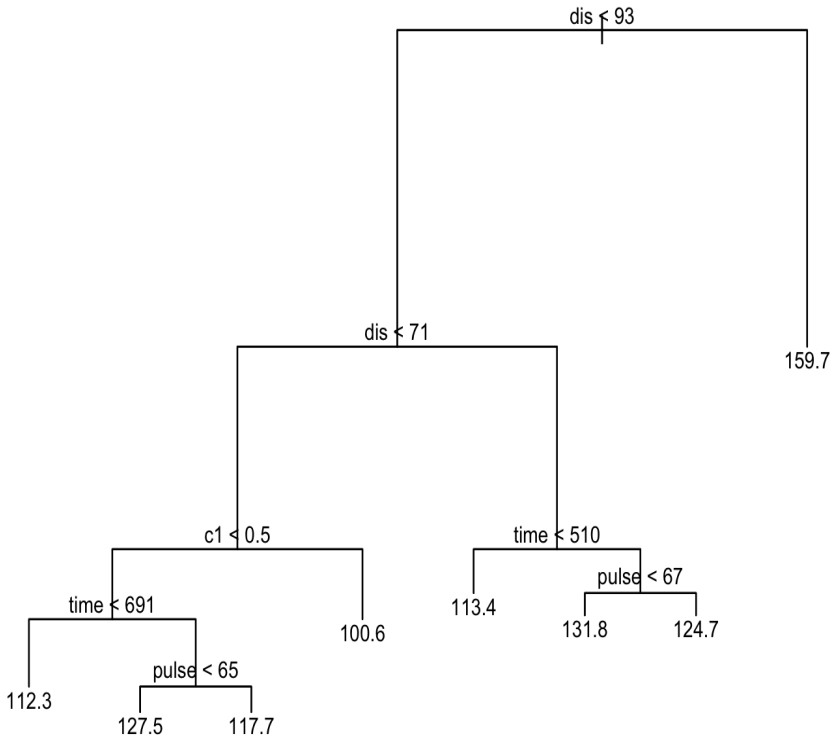

Figure 2: Diagram of Classification and Regression Tree

Imputation in CART is first implemented by fitting the classification and regression tree with the observed data. Next, the method determines which terminal node of the fitted tree each missing observation is predicted to end up in. Lastly, a random draw is made for the member in the node, and we take the observed value from that draw as the imputation [19].

3.2 Predictive Mean Matching

PMM is often used because it produces realistic values. This is because PMM takes values from individuals that were studied [2]. This method searches for the complete case most similar to the case with missing values, and replaces those missing values with the corresponding values from the complete case. The PMM approach is becoming more popular since it has become embedded in MICE software [2]. While PMM is similar to the regression approach, PMM can often be more appropriate than regression, especially when the assumption of normality is violated [4]. PMM works by replacing missing values with an observed value from the observation whose regression-predicted values are closest to the regression-predicted value for the missing value from the simulated regression model [4]. Thus, the imputed values are always realistic and representative of a value that the missing value could be. Constraints and bounds are always met, as well as discrete or continuous conditions.

3.3 Random Forest Approaches

A random forest (RF) is a collection of multiple decision trees fit with training data. For continuous variables, RF imputes the missing values by randomly drawing from independent normal distributions, centered on means predicted from RF. However, for categorical variables, RF predicts missing values trained on observed values.

3.3.1 Proximity Imputation

When using the proximity imputation method, one must pre-impute the data. This means one must use some method of imputing the missing values before a random forest model can be fit. Quantitative missing values can be imputed by taking the median of the non-missing values, while categorical missing values can be imputed by taking the most occurring non-missing value [6]. This is called strawman imputation. Once the data is pre-imputed, a random forest model is fit. A symmetric n x n proximity matrix is produced. Each *(i,j)* entry represents the fraction of trees in which elements i and j share the same terminal node [5]. One wishes to see similar observations on the same terminal nodes and dissimilar observations in separate terminal nodes. The proximity matrix is then used to impute the original missing values in the data [6].

When working with mixed data, quantitative and categorical values are imputed using different methods. The proximity weighted average over non-missing data is used to impute quantitative variables and the largest average proximity over non-missing data is used to impute categorical variables [6]. Once the missing values are imputed, a new random forest is produced and the procedure can be iterated as many times as the researcher deems sufficient.

When using RF in R, one chooses how tree nodes are to be split. The default in RF tests all possible split points for each of the potential splitting variables. However, this method can be computationally extensive when working with a large number of observations or variables. When this is the case, one can use a method of RF in which random splits are chosen. This method of random splitting is considerably

faster than deterministic splitting [6]. One can speed up the splitting process even further by using pure random splitting. In this method, no splitting rule is applied. Each tree node is split by randomly selecting both a variable and split point [6].

3.3.2 On-the-fly Imputation

Unlike proximity imputation, on-the-fly imputation (OTF) simultaneously imputes data while growing the forest [6]. This method is designed to address the weaknesses of proximity imputation, which include having biased estimates and variable importance. In OTF, split statistics are calculated using only observed data. When data is missing, a value is imputed using a random value from the in-bag observed data [6]. After each split, imputed values are reset to missing. Once the terminal nodes are reached, the missing values are imputed using out-of-bag (OOB) observed terminal node data from all the trees. The average observed value for quantitative values is used and the maximum observed value for categorical values is used.

The variables used to split each node are selected randomly. The *nsplit* function in R can be implemented to increase computational speed. This function uses random splitting. One can also implement pure random splitting to further increase computational speed. The process can then be iterated. It is important to note that during the first iteration, OOB estimates are used. For each additional iteration, in-bag estimates must be used because no OOB estimates exist [6].

3.3.3 missForest and mForest Imputation

In contrast to other imputation methods, the missForest algorithm involves a prediction problem. Missing data for the response variable is predicted by first imputing data. The data is imputed by regressing each variable against the other variables [6]. Depending on how many variables one is working with, this process can be computationally slow. If one is working with n variables, n forests will be fit for each iteration. When n is large, one may wish to implement mForest instead, which is a computationally faster version of missForest. This method involves assigning the n variables to groups, which in turn leads to less forests being fit.

Each forest is grown using multivariate splitting. Missing values in the response are excluded and the split-rule is averaged over observed responses [6]. Prediction methods are used to impute the final missing response values. In some studies, mForest has been shown to perform as well as missForest, even with less computations.

4 MODEL SELECTION

It is important that health care professionals are able to get accurate blood pressure readings from patients. Unusual blood pressure readings can often be signs of hypertension or other diseases that can adversely affect lives. The data set chosen to analyze was collected by the Centers of Disease Control and Prevention (CDC) from 2007-2008 and was first published in 2009 [7]. The CDC asked questions to and took measurements on the examinees in order to collect various statistics. The examinees were of all ages and backgrounds. We chose to predict the systolic blood pressure reading given a combination of both categorical and continuous predictors.

The response variable, systolic blood pressure, was collected on both children and adults. The first readings ranged in values from 74 mmHg to 230 mmHg. For very young children, it is healthy for this reading to be as low as 80 mmHg and as great as 120 mmHg. For most adults, a healthy reading ranges between 110 mmHg and 130 mmHg.

There are a total of 7146 observations that have complete responses on all the variables measured, including the response variable and all predictor variables. The continuous predictor variables include blood pressure time, pulse rate, maximum inflation level, and diastolic blood pressure. The categorical predictor variables include cuff size, pulse type, and responses given to questions regarding if the examinee has had food, alcohol, coffee, or a cigarette in the past thirty minutes. Of these categorical predictors, only cuff size has more than 2 levels.

4.1 Fitted Model

After analyzing the data, the final regression model found is:

$$\begin{aligned}\widehat{Y_i} = &- 13.8648 - 0.0071X_{i1} + 1.4319X_{i2} - 3.1359X_{i3} - 2.3824X_{i4} - 1.5503X_{i5} \\ &+ 0.0309X_{i6} + 0.9287X_{i7} + 0.0437X_{i8} + \varepsilon_i \text{ for i=1,...,n,}\end{aligned} \tag{1}$$

where the predictor variables include PEASCTM1, cig, cuff1, cuff2, cuff3, BPXPLS, BPXML1, and BPXDI1, respectively. The variable "PEASCTM1" measures blood pressure time recorded in seconds, where the values range from 45 to 1521. The indicator variable "cig" is equal to 1 if the examinee has smoked a cigarette in the last 30 minutes, and is equal to 0, otherwise. Indicator variables were created for cuff size, where "cuff1" equal to 1 refers to an examinee measured in a child sized cuff (9 cm by 17 cm), "cuff2" equal to 1 refers to an examinee measured in an adult sized cuff (12 cm by 22 cm), and "cuff3" equal to 1 refers to an examinee measured in a large sized cuff (15 cm by 32 cm). If "cuff1", "cuff2", and "cuff3" are all equal to zero, the examinee was measured in a thigh sized cuff (18 cm by 35 cm). The variable "BPXPLS" measures pulse rate over 60 seconds. The 30 second pulse rate was recorded and then multiplied by 2. The values range from 40 to 224. The variable "BPXML1" measures maximum inflation level in mm Hg, where the values range from 110 to 240. The variable BPXDI1 measures diastolic blood pressure in mm Hg, where the values range from 0 to 116.

4.2 Results

One of the tests performed during model selection was the global F-test. The hypotheses were:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_{14} = 0$$

$$H_1 : \text{At least one } \beta_j \text{ does not equal 0 for j=1,...,14.}$$

This test was performed on the set of 4 continuous predictor variables, as well as 8 categorical predictor variables containing 10 indicator variables. The resulting p-value was less than 0.0001, so we reject the null hypothesis and conclude that at least one predictor is significant in the model.

The global F-test was performed again after reducing the model to 4 continuous predictor variables and 2 categorical predictor variables containing 4 indicator variables, and the hypotheses were:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_8 = 0$$

$$H_1 : \text{At least one } \beta_j \text{ does not equal 0 for j=1,2,...,8.}$$

The resulting p-value once again was less than 0.0001, so we conclude that at least one predictor is significant in the model.

4.3 Variable Selection and Model Building

When fitting the model, the variable selection was fairly simple due to cooperative data. In the initial multiple linear regression model with 4 continuous predictor variables and 10 indicator variables, there were no issues. A model was fit, and the

conclusion was that at least one predictor was needed in the model. In the process of trying the reduce the model, most tests and processes yielded similar results.

Looking at the initial summary output, using $\alpha = 0.05$, the model suggests that the predictors needed include "PEASCTM1", "cig", "cuff1", "cuff2", "cuff3", "BPXPLS", "BPXML1", and "BPXDI1." One method I chose to implement was backwards stepwise regression with alpha set to 0.05. This method suggested that the best model include 7 of the possible 14 continuous and indicator variables, including "PEASCTM1", "cuff1", "cuff2", "cuff3", "BPXPLS", "BPXML1", and "BPXDI1". When alpha was set to 0.10, the same method suggested that the best model include the previous 7 continuous and indicator variables, with the addition of "cig". This variable selection is the same as the model chosen by including the variables in the multiple regression model with p-values less than 0.05. The models were evaluated using the adjusted R^2 and Mallow's Cp criteria. The Cp method showed that the best model had the same 8 predictors as the previous models, the adjusted R^2 method suggested that a model with only "BPXML1" as a predictor would have only slightly weaker predictive ability than many of the more complex models. The adjusted R^2 for the original model with all 14 predictors was 0.8512, while the adjusted R^2 for the model reduced to only 1 predictor1 was 0.8464. Thus, nearly 85% of the variability in systolic blood pressure can be explained by its linear relationship with only maximum inflation level. However, for the purpose of this research, I chose to work with the more complex model including 4 continuous predictors and 4 indicator variables. Figures 3-6 show the added variable plots for the 4 continuous predictors to ensure these variables did not need to enter in the model in a curvilinear fashion. Since

these plots did not seem to have any severe curved pattern, entering the variables in a linear fashion seems appropriate.

4.4 Predictive Ability

The PRESS statistic was calculated as 406742.43, and it was compared to SSE = 405658. The ratio of the PRESS statistic and SSE yielded a 1.002673 value, which suggests that the model has good predictive ability, as ratios between PRESS and SSE close to 1 suggest good predictive ability and ratios between PRESS and SSE that are much larger than 1 suggest poor predictive ability.

4.5 Assumptions

Since the data set is large (over 7000 observations), the assumption of normality is met due to the Central Limit Theorem. Next, we checked for constant error variance in both the full and the reduced models. In Figure 7 and Figure 8, we see almost identical plots that show mostly random scatter, which suggest constant error variance is satisfied in both the full and reduced models.

Next, we checked for issues we may have with multicollinearity. In both the full and reduced model, multicollinearity did not seem to be an issue. The largest VIF value detected was less than 2.5, while many were near 1, which is what one wants to see. Only VIF value larger than 10 are severe and need to be checked out. I additionally graphed and calculated the pairwise correlations between the 4 continuous predictors. In Figure 9, it seems that none of the pairwise relations between the continuous predictor variables have strong linear relationships. When calculated, the

pairwise correlations were low (between -0.175 and 0.377), which results in the same conclusion about there not being an issue with multicollinearity.

4.6 Outlier Detection

Multiple methods to detect for both outliers and influential observations were employed. Out of the 7146 observations, only a few were flagged as outliers in the Y direction. However, there were quite a few observations that were flagged as outliers in the X direction using the h_{ii} rule. In Figure 10, we observe that many observations fell above the 0.0025 threshold ((2*p)/n), where $p = 9$ and $n = 7146$.

Next, both DFFITS and Cook's distance were found to check for influential observations. In Figure 11, we observe many observations falling above or below the +/- (0.071) threshold (+/- (2*sqrt(p/n))), where $p = 9$ and $n = 7146$. In Figure 12, the influential observation threshold is 0.927 (qf(0.5,p,$n-p$), where $p = 9$ and $n = 7146$) , so no observations are detected as influential. Since the reduced model yielded strong results, nothing was done to remove potentially outlying or influential observations.

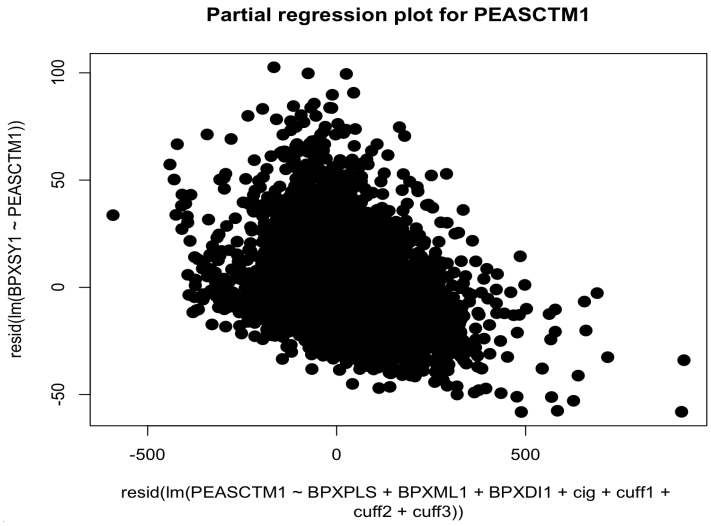

Figure 3: Partial regression plot created using the variable PEASCTM1.

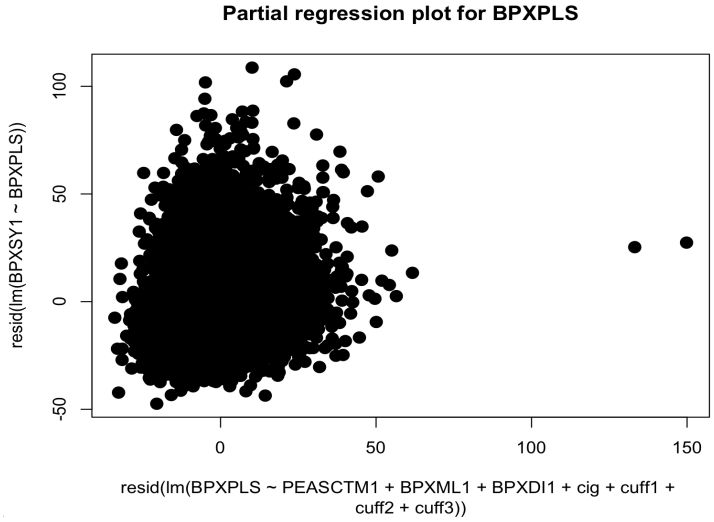

Figure 4: Partial regression plot created using the variable BPXPLS.

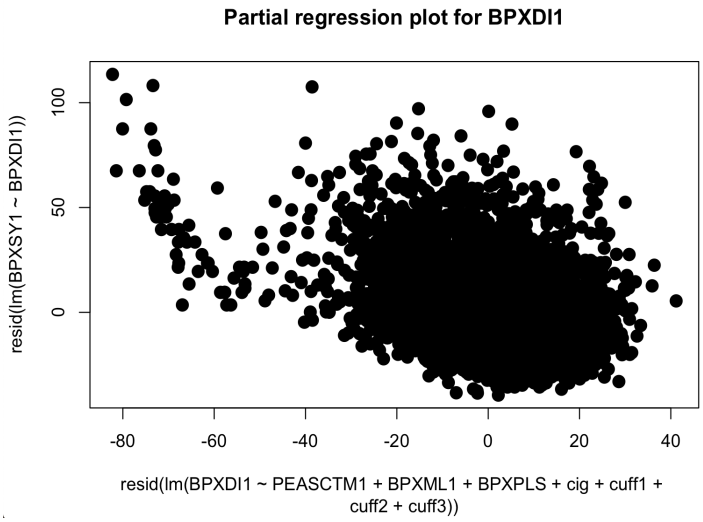

Figure 5: Partial regression plot created using the variable BPXDI1.

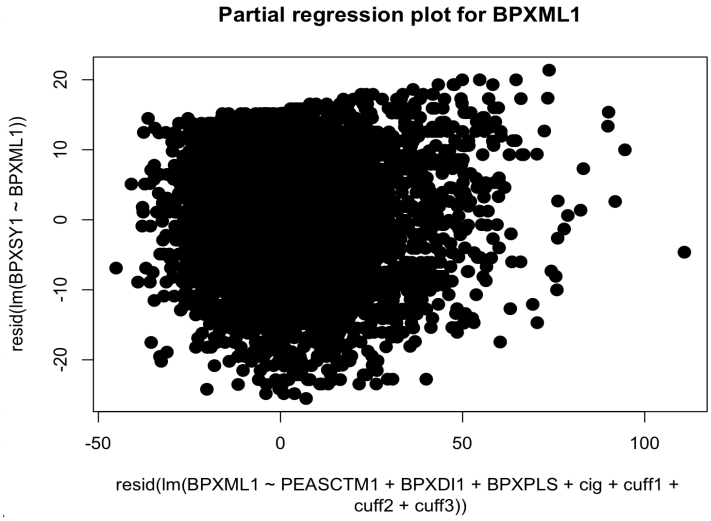

Figure 6: Partial regression plot created using the variable BPXML1.

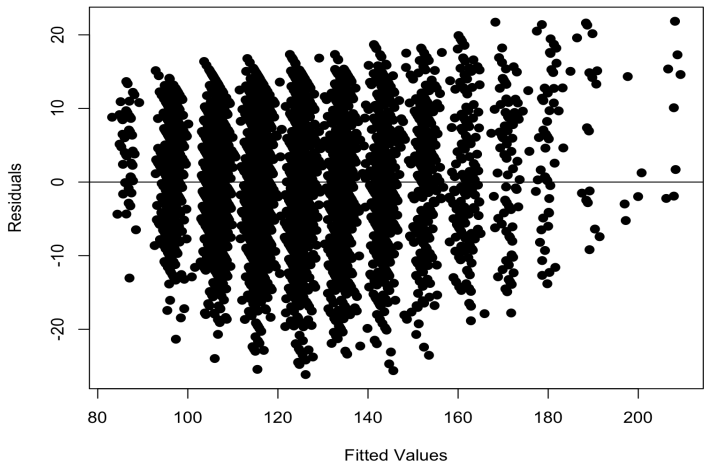

Figure 7: Residual plot created using the full model to check for constant error variance.

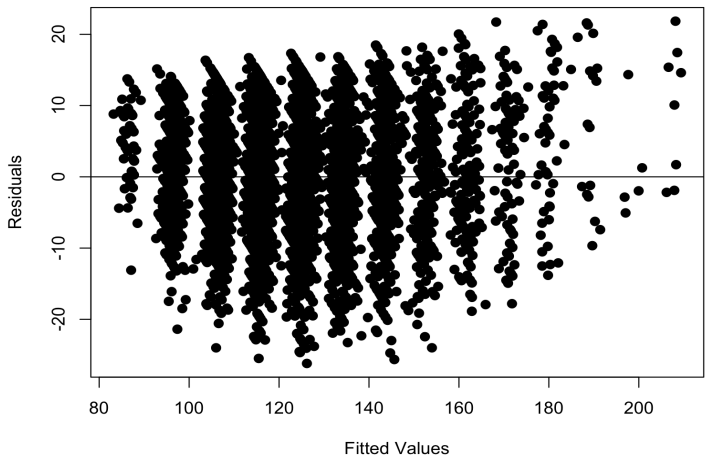

Figure 8: Residual plot created using the reduced model to check for constant error variance.

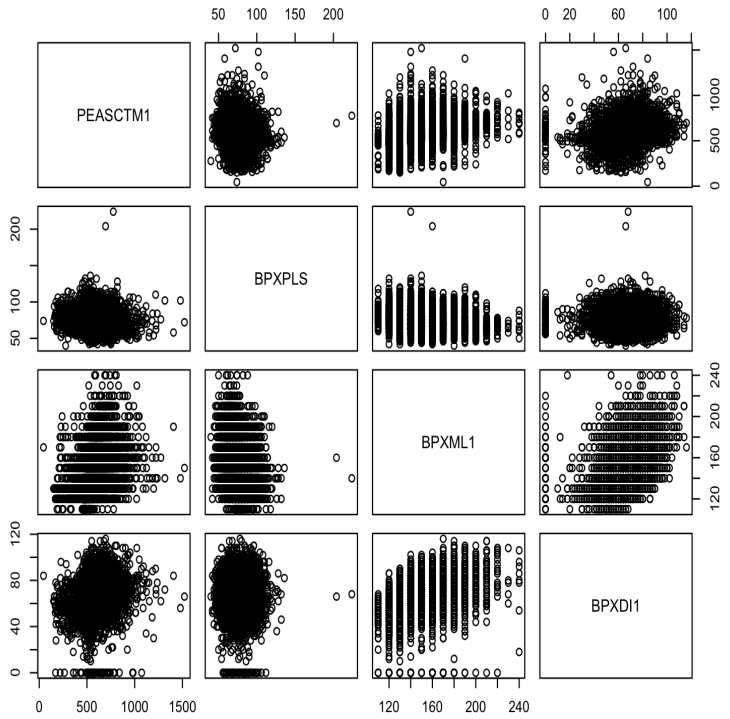

Figure 9: Pairwise correlations between the four continuous predictor variables.

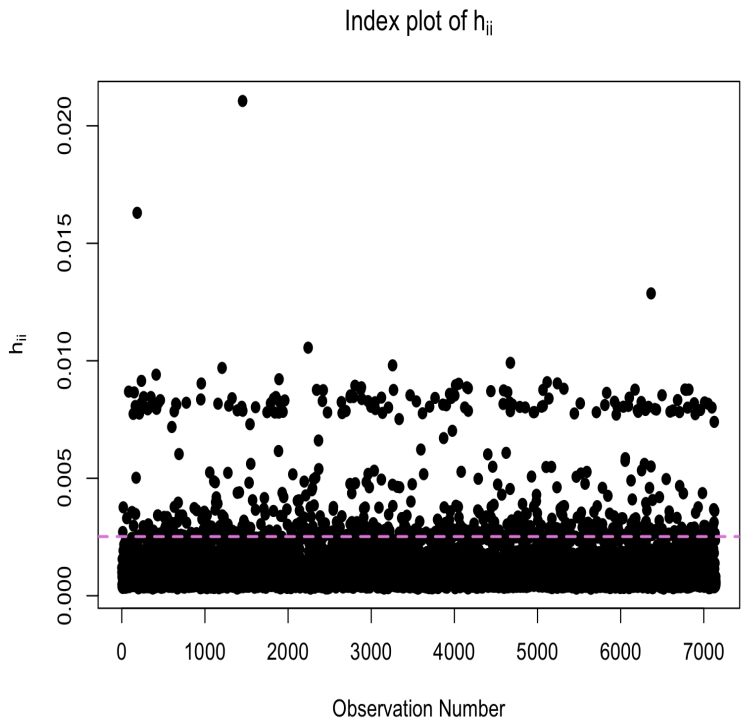

Figure 10: Outlier detection in X using h_{ii} rule.

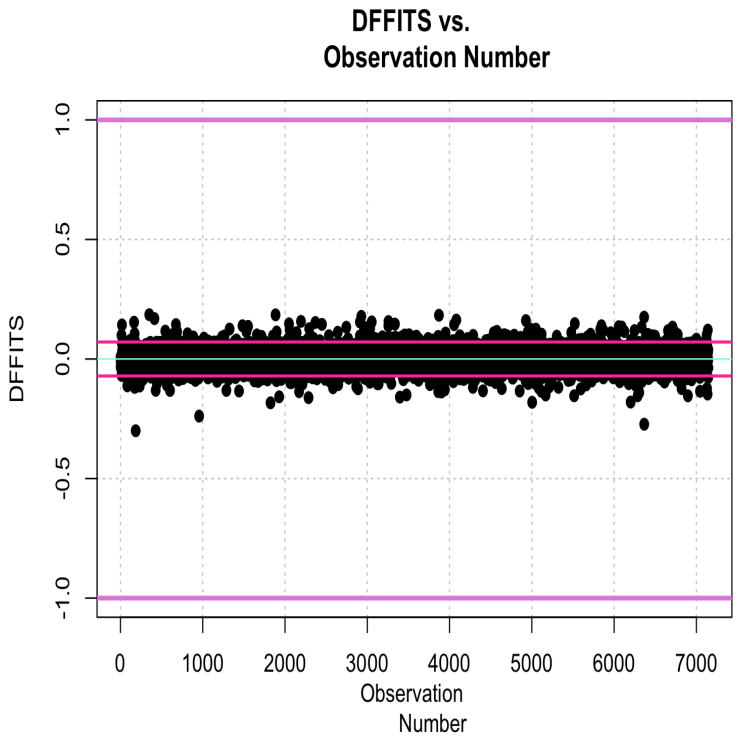

Figure 11: Checking for influential observations using dffits rule.

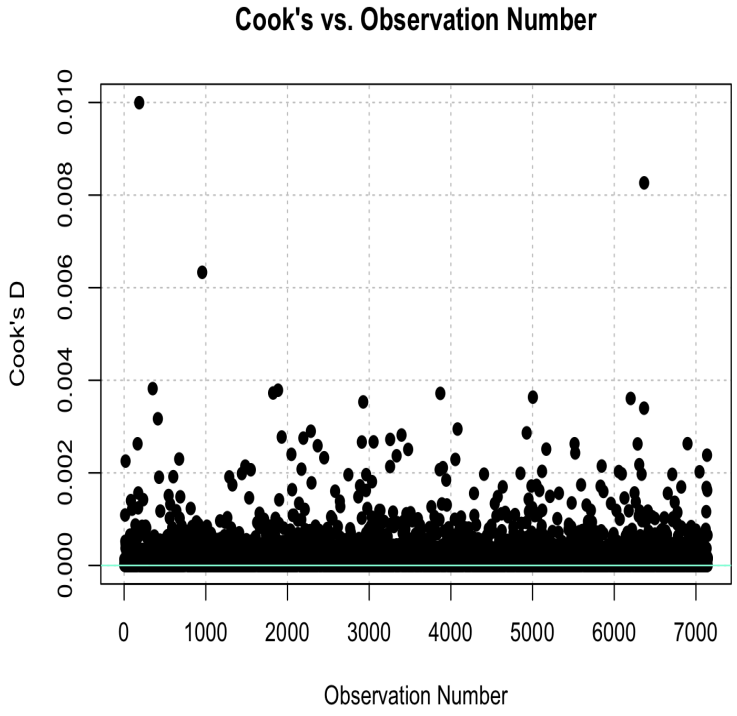

Figure 12: Checking for influential observations using Cook's rule.

5 IMPUTATION MODEL SELECTION

Now that a regression model has been fit to the complete data set, we constructed a function to randomly remove values from the complete data set at levels of 10%, 20%, 30%, 40%, 50%, and 60%. The function removed values from both the response variable and the categorical and continuous predictor variables. There are four levels for the cuff size variable, so if one of the values was missing, we made all of the values missing.

Since the data set contains both continuous and categorical variables, we are limited to which methods in the MICE package can be implemented. The four methods that will be implemented include "cart" (classification and regression trees), "sample" (on-the-fly imputation), "rf" (random forest), and "pmm" (predictive mean matching). Each of these methods will fit a model at each level of missingness to produce a total of twenty-four multiple regression models.

Once the models were fit, we compared each estimated regression coefficient to the parameters from the multiple regression model produced by the complete data set. We calculated the associated percent deviation indices (PDI) for each of the 216 estimated regression coefficients and constructed one-sample t-tests in order to determine if these values were significantly different from the multiple regression model parameters.

In addition to calculating PDI values for each estimated regression coefficient, we computed the R^2 and adjusted R^2 values for each of the twenty-four imputation models. These values were used to determine the prediction accuracy of the imputation models. We compared the prediction accuracy of each imputation model to the prediction accuracy of the multiple regression model fit with the complete data set.

This helped us determine which methods of imputation perform well at given levels of data missingness.

We computed the proportion of missing values for the response variable and each of the predictor variables at each of the six levels of data missingness. The proportions will be compared within their respective data set, as well as among each of the six data sets.

6 EVALUATION OF IMPUTATION MODELS

The relative efficiency (RE) for a model is used to determine how efficient parameter estimates will be given the number of imputations and the fraction of missing observations [3].

The equation for RE is given by

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

where λ is the proportion of missing values and m is the number of imputations [20].

In Table 1, we observe that for each value of the number of imputations, RE decreases as the proportion of missing values increases. We also notice that within each level of data missingness, RE increases as the number of imputations increases. While the calculated RE values do not vary much between the number of imputations for the smaller proportions of level missingness, we chose to use 50 imputations for each of our twenty-four models. The larger values for the number of imputations will not only make RE increase, but the standard error values will be more accurate and therefore yield for accurate p-values [3].

6.1 Estimated Regression Coefficients

In Table 2, we observe the parameter values for the multiple regression model produced by the complete data set. These values will be used as a comparison to each of the estimated regression coefficients from each of the four imputation models at levels of 10%, 20%, 30%, 40%, 50%, and 60% of data missingness. The estimated means of the regression coefficients from the CART, OTF, RF, and PMM imputation

models at each level of data missingness are given in Tables 3, 4, 5, and 6. We observe that the estimated means of the regression coefficients from the CART and PMM imputation models seem to be more similar to the parameter values than for the OTF and RF imputation methods.

6.2 Percent Deviation Index

In Table 7, we observe the ranges of each of the estimated regression coefficients for the CART, OTF, RF, and PMM imputation methods. For some of the estimated regression coefficients, we observe that the OTF and RF imputation methods have noticeably larger ranges than the CART and PMM methods. For the rest of the estimated regression coefficients, all four models seem to have relatively similar range values.

In Tables 8, 9, 10, and 11, we observe the PDI values for each of the estimated regression coefficients from the CART, OTF, RF, and PMM imputation models at each level of data missingness. The equation for PDI is given by

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

In their respective tables, we observe that many of the PDI values are large and in some cases extremely large, especially for the estimated regression coefficents in the OTF and RF imputation models. In general, we see that the PDI values increase as the proportions of data missingness increase. This does not seem unusual, as we expect that the imputation models may not perform as well as the proportions of missing values increase. However, with the CART and PMM imputation models, we

notice that some of the PDI values decrease at the 40% level of data missingness, which is interesting.

6.3 Model Accuracy

Since the PDI values were so unexpectedly large, we were concerned that each of the imputation methods may not be performing well. To test this theory, the R^2 and adjusted R^2 values were calculated for each of the four imputation models at each level of data missingness. We used the R^2 and adjusted R^2 values from the multiple regression model produced by the complete data set as a comparison. These values are 0.8513 and 0.8512, respectively. The R^2 value is the statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model [21]. The adjusted R^2 value is the R^2 value, adjusted for the number of independent variables in the model. We wish to see these values close to 1, with the adjusted R^2 value close to the R^2 value. When the adjusted R^2 is smaller than the R^2 value, this means that we have independent variables in the model that do not need to be there. This could result from multicollinearity among independent variables. In Tables 13, 14, 15, and 16, we observe the R^2 and adjusted R^2 values calculated for each of the four imputation models at each level of missingness. We notice that the R^2 and adjusted R^2 values are large and similar to the values calculated from the regression model produced by the complete data set for both the CART and PMM imputation methods. This holds true at each of the six levels of data missingness. The values do not decrease much as the proportion of data missingness increases, which means that these two

methods are performing well across all levels. Interestingly, the R^2 and adjusted R^2 values increase slightly at the 40% level of data missingness, which is similar to what we noticed about the PDI values at the same level.

However, the same can not be said for the OTF and RF imputation methods. In particular, the R^2 and adjusted R^2 values for the OTF are very poor. Even at only the 10% level of data missingness, the independent variables are barely explaining half of the variation in the dependent variable. The RF imputation performed slightly better than the OTF imputation method, but the R^2 and adjusted R^2 values decrease steeply as the proportion of data missingness increases.

6.4 Standard Deviations for Regression Coefficients

The standard deviations for each of the parameters from the multiple regression model produced by the complete data set, as well as the standard deviations from each of the regression coefficients from each imputation model at each level of data missingness are given in Tables 17, 18, 19, 20, and 21. These values are measures of how the data values for each parameter and estimated coefficient vary around the mean of that parameter or estimated coefficient. The OTF and RF imputation methods seem to yield larger standard deviations than the CART and PMM imputation methods for each of the estimated regression coefficients.

6.5 Significance of Regression Coefficients of Imputation Models

We constructed one-sample t-tests for each of the estimated regression coefficients for each of the four imputation models at each level of data missingness. The following

hypotheses are given:

$$H_0 : \hat{\beta}_i = \beta_i, \text{where i=0,1,...,8}$$

$$H_1 : \hat{\beta}_i \text{ does not equal } \beta_i \text{ for i=0,1,...,8.}$$

We set $\alpha = 0.05$. When p-value < 0.05, we reject the null hypothesis and conclude that our estimated regression coefficient is significantly different from the parameter from the multiple regression model produced by the complete data set. The significant p-values are in bold in Tables 22, 23, 24, and 25. We see that the majority of the p-values for the estimated regression coefficients are significant for the OTF and RF imputation methods, while the CART and PMM imputation methods have much fewer p-values that are significant. We notice that in general, especially for the CART and PMM imputation models, there are more significant values as the level of data missingness increases.

6.6 Significance of Regression Coefficients of Imputation Models, Adjusted for Multiple Testing

Since there are multiple t-tests being done simultaneously, we must adjust our p-values for multiple testing because the probability of committing a false statistical inferences considerably increases when more than one hypothesis is simultaneously tested [22]. We used the Benjamini-Hochberg (BH) procedure in R to adjust for multiple testing in order to help avoid Type I errors, where a Type I error is the probability of rejecting the null hypothesis, when in fact the null hypothesis is true [23]. In Tables 26, 27, 28, and 29, we observe the adjusted p-values, where the

significant p-values at the $\alpha = 0.05$ level are in bold. While the significance of most of the p-values did not change with the BH adjustment, there were some p-values that are no longer significant.

6.7 Proportion of Missing Values for Each Variable

In Tables 30, 31, 32, 33, 34, and 35, we observe the proportion of missing values for the dependent variable and each of the independent variables in the data sets where 10%, 20%, 30%, 40%, 50%, and 60% of the values are missing. We observe that at each of the levels of data missingness, all of the variables seem to have fairly equal proportions of missing values.

Table 1: Relative efficiency of the imputation models for various numbers of imputations at several levels of data missingness.

m	10%	20%	30%	40%	50%	60%
5	0.9804	0.9615	0.9434	0.9259	0.9091	0.8929
10	0.9901	0.9804	0.9709	0.9615	0.9524	0.9434
20	0.9950	0.9901	0.9852	0.9804	0.9756	0.9709
30	0.9967	0.9934	0.9901	0.9868	0.9836	0.9804
40	0.9975	0.9950	0.9926	0.9901	0.9877	0.9852
50	0.9980	0.9960	0.9940	0.9921	0.9901	0.9881

Table 2: Parameter values for the multiple regression model produced by the complete data set.

Parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Actual Values	-13.8648	-0.0071	1.4319	-3.1359	-2.3824	-1.5503	0.0309	0.9287	0.0437

Table 3: Estimated means of the regression coefficients from the CART imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-14.1396	-0.0077	1.2658	-3.1096	-2.0973	-1.2443	0.0333	0.9300	0.0433
20 %	-14.0843	-0.0076	1.6402	-2.6815	-2.0289	-1.0674	0.0312	0.9301	0.0410
30 %	-14.7466	-0.0069	1.9031	-2.0715	-1.7355	-0.7974	0.0299	0.9266	0.0512
40 %	-14.4310	-0.0062	1.7709	-1.7167	-1.0583	-0.5793	0.0262	0.9192	0.0568
50 %	-14.9377	-0.0039	1.7376	-2.1927	-0.7489	-0.4043	0.0360	0.9099	0.0506
60 %	-12.0866	-0.0032	1.7698	-1.7872	-0.8833	-0.2636	0.0102	0.9040	0.0443
Actual Parameter	-13.8648	-0.0071	1.4319	-3.1359	-2.3824	-1.5503	0.0309	0.9287	0.0437

Table 4: Estimated means of the regression coefficients from the OTF imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	8.5592	0.0013	1.7807	-3.6454	-1.7835	-0.5619	-0.0022	0.7251	0.1055
20 %	29.6312	0.0074	2.4930	-4.0480	-1.5930	-0.0623	-0.0256	0.5614	0.1195
30 %	47.0967	0.0085	1.9840	-3.8607	-1.4232	0.1364	-0.0282	0.4342	0.1314
40 %	65.1435	0.0089	2.2400	-3.4213	-0.8951	0.1402	-0.0323	0.3118	0.1268
50 %	80.8215	0.0084	2.1703	-3.3226	-0.7272	0.4466	-0.0282	0.2121	0.1069
60 %	96.4430	0.0049	1.0745	-2.1635	-0.6003	0.4249	-0.0302	0.1381	0.0684
Actual Parameter	-13.8648	-0.0071	1.4319	-3.1359	-2.3824	-1.5503	0.0309	0.9287	0.0437

Table 5: Estimated means of the regression coefficients from the RF imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-6.2500	-0.0034	1.5440	-3.5233	-2.2584	-1.1658	0.0177	0.8567	0.0652
20 %	2.2769	0.0009	2.4794	-4.2886	-2.5076	-1.0509	0.0007	0.7812	0.0851
30 %	10.0653	0.0043	2.5632	-5.0882	-2.5664	-0.9990	-0.0080	0.7079	0.1127
40 %	22.8139	0.0074	2.8822	-5.7815	-2.5759	-0.9129	-0.0180	0.6036	0.1318
50 %	33.4396	0.0114	2.6120	-7.2516	-2.3876	-0.5767	-0.0209	0.5098	0.1446
60 %	51.5729	0.0122	2.3013	-8.3311	-2.9727	-0.6102	-0.0450	0.3995	0.1375
Actual Parameter	-13.8648	-0.0071	1.4319	-3.1359	-2.3824	-1.5503	0.0309	0.9287	0.0437

Table 6: Estimated means of the regression coefficients from the PMM imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-13.7384	-0.0085	1.1616	-3.3937	-2.3707	-1.5193	0.0366	0.9326	0.0392
20 %	-13.3713	-0.0085	1.8639	-3.4760	-2.6183	-1.6533	0.0382	0.9336	0.0311
30 %	-14.8370	-0.0082	2.4208	-3.3503	-2.6047	-1.7370	0.0459	0.9315	0.0468
40 %	-13.9330	-0.0070	2.9853	-3.4241	-2.3471	-1.7884	0.0406	0.9238	0.0454
50 %	-14.0729	-0.0032	3.9299	-4.4731	-2.6540	-1.9352	0.0419	0.9103	0.0461
60 %	-11.8260	-0.0036	4.5916	-3.1585	-2.1605	-1.4474	0.0117	0.9032	0.0578
Actual Parameter	-13.8648	-0.0071	1.4319	-3.1359	-2.3824	-1.5503	0.0309	0.9287	0.0437

Table 7: Range of the estimated regression coefficients for each of the four imputation methods.

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
CART	2.8511	0.0045	0.6373	1.3929	1.3484	0.9807	0.0258	0.0261	0.0158
OTF	87.8838	0.0076	1.4185	1.8845	1.1832	1.0085	0.0301	0.5870	0.0630
RF	57.8229	0.0156	1.3382	4.8078	0.7143	0.5891	0.0627	0.4572	0.0794
PMM	3.0110	0.0053	3.4300	1.3146	0.4935	0.4878	0.0342	0.0304	0.0267

Table 8: Percent deviation indices of CART imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	MEAN
10 %	-1.9816	-8.0647	11.6039	0.8413	11.9672	19.7354	-7.7738	-0.1397	0.9823	3.0189
20 %	-1.5830	-7.0079	-14.5446	14.4927	14.8391	31.1452	-0.8785	-0.1558	6.1233	4.7145
30 %	-6.3601	2.2757	-32.9055	33.9443	27.1529	48.5655	3.0790	0.2270	-17.1597	6.5355
40 %	-4.0839	12.6202	-23.6731	45.2588	55.5782	62.6306	15.1512	1.0238	-29.8695	14.9596
50 %	-7.7379	45.1178	-21.3504	30.0775	68.5653	73.9180	-16.4142	2.0198	-15.7904	17.6006
60 %	12.8254	55.1605	-23.5924	43.0102	62.9244	82.9961	67.0489	2.6582	-1.4133	33.5131
MEAN	-1.4869	16.6836	-17.4104	27.9375	40.1712	53.1651	10.0354	0.9389	-9.5212	13.3904

Table 9: Percent deviation indices of OTF imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	MEAN
10 %	161.7332	118.9848	-24.3553	-16.2465	25.1378	63.7522	106.9972	21.9223	-141.4009	35.1694
20 %	313.7152	204.1996	-74.0999	-29.0828	33.1359	95.9784	182.8523	39.5500	-173.3590	65.8766
30 %	439.6851	219.5780	-38.5575	-23.1115	40.2638	108.8000	191.4292	53.2420	-200.7080	87.8468
40 %	569.8478	225.6640	-56.4314	-9.0993	62.4297	109.0434	204.4628	66.4209	-190.0547	109.1426
50 %	682.9252	217.6941	-51.5641	-5.9506	69.4771	128.8094	191.2213	77.1657	-144.5558	129.4691
60 %	795.5953	168.6546	24.9623	31.0097	74.8044	127.4075	197.8441	85.1300	-56.4000	161.0009
MEAN	493.9170	192.4625	-36.6743	-8.7468	50.8748	105.6318	179.1345	57.2385	-151.0797	98.0842

Table 10: Percent deviation indices of RF imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	MEAN
10 %	54.9217	51.8240	-7.8266	-12.3507	5.2079	24.8011	42.7245	7.7464	-49.1992	13.0943
20 %	116.4223	113.3714	-73.1543	-36.7561	-5.2528	32.2142	97.8279	15.8765	-94.6817	18.4297
30 %	172.5957	160.3084	-79.0006	-62.2563	-7.7213	35.5569	125.8128	23.7730	-157.7475	23.4801
40 %	264.5449	204.4831	-101.2809	-84.3609	-8.1186	41.1159	158.1677	35.0048	-201.6675	34.2098
50 %	341.1833	260.2701	-82.4103	-131.2424	-0.2158	62.8018	167.7079	45.1044	-230.7970	48.0447
60 %	471.9697	271.4514	-60.7162	-165.6647	-24.7749	60.6382	245.7914	56.9783	-214.4954	71.2420
MEAN	236.9394	176.9514	-67.3982	-65.4385	-6.8126	42.8547	139.6720	30.7472	-158.0981	34.7501

Table 11: Percent deviation indices of PMM imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	MEAN
10 %	0.9117	-20.4439	18.8793	-8.2206	0.4942	1.9992	-18.5011	-0.4231	10.2515	-1.6725
20 %	3.5598	-20.2226	-30.1680	-10.8449	-9.9012	-6.6466	-23.6428	-0.5328	28.9234	-7.7195
30 %	-7.0122	-14.8617	-69.0563	-6.8366	-9.3300	-12.0484	-48.7046	-0.3020	-7.0653	-19.4686
40 %	-0.4919	1.7358	-108.4799	-9.1901	1.4854	-15.3584	-31.5066	0.5223	-3.8693	-18.3503
50 %	-1.5011	54.8295	-174.4508	-42.6411	-11.3995	-24.8312	-35.5139	1.9721	-5.4737	-26.5566
60 %	14.7053	48.6967	-220.6555	-0.7190	9.3138	6.6361	62.2551	2.7381	-32.3176	-12.1500
MEAN	1.6953	8.2890	-97.3219	-13.0754	-3.2129	-8.3749	-15.9357	0.0051	-1.5918	-14.3196

Table 12: R^2 and Adjusted R^2 values for multiple regression model produced by the complete data set.

R^2	Adjusted R^2
0.8513	0.8512

Table 13: R^2 and Adjusted R^2 values for CART imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.8514	0.8526	0.8520	0.8526	0.8451	0.8252
Adjusted R^2	0.8512	0.8524	0.8519	0.8525	0.8450	0.8249

Table 14: R^2 and Adjusted R^2 values for OTF imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.5646	0.3609	0.2244	0.1307	0.0659	0.0290
Adjusted R^2	0.5641	0.3602	0.2235	0.1297	0.0649	0.0279

Table 15: R^2 and Adjusted R^2 values for RF imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.7398	0.6423	0.5436	0.4368	0.3376	0.2303
Adjusted R^2	0.7395	0.6419	0.5431	0.4362	0.3368	0.2294

Table 16: R^2 and Adjusted R^2 values for PMM imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.8525	0.8537	0.8541	0.8558	0.8521	0.8382
Adjusted R^2	0.8528	0.8535	0.8539	0.8556	0.8520	0.8380

Table 17: Standard deviations for each of the parameters from the multiple regression model produced by the complete data set.

Parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
SD	7.5573	0.0056	4.7164	3.2816	2.1042	1.9734	0.0504	0.0380	0.0484

Table 18: Standard deviations of each of the regression coefficients in CART imputation model for each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	8.7614	0.0067	5.0410	3.7318	2.3835	2.3451	0.0581	0.0428	0.0571
20 %	9.8419	0.0075	5.9941	4.0468	2.7319	2.5356	0.0667	0.0448	0.0686
30 %	11.6598	0.0097	6.4213	4.5089	3.0754	2.7050	0.0831	0.0535	0.0772
40 %	12.5472	0.0099	7.1611	5.1555	3.2640	2.8417	0.0877	0.0627	0.0867
50 %	15.0012	0.0108	7.4797	6.4303	4.0407	3.4955	0.1063	0.0828	0.1269
60 %	19.7725	0.0183	8.2586	8.2734	4.5819	3.7400	0.1438	0.1106	0.1461

Table 19: Standard deviations of each of the regression coefficients in OTF imputation model for each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	15.3178	0.0116	9.7187	5.9322	3.7102	3.4102	0.1021	0.0896	0.0972
20 %	18.8632	0.0139	12.6812	7.7163	4.5408	4.3467	0.1248	0.1061	0.1267
30 %	22.9013	0.0159	14.1290	8.6609	5.0141	4.5638	0.1512	0.1181	0.1373
40 %	24.1497	0.0160	14.3644	8.8214	4.9340	4.8302	0.1583	0.1171	0.1328
50 %	26.7741	0.0179	16.9659	9.1476	4.9822	4.5843	0.1644	0.1212	0.1626
60 %	26.8073	0.0173	18.1518	9.0625	5.4622	5.0773	0.1645	0.1230	0.1631

Table 20: Standard deviations of each of the regression coefficients in RF imputation model for each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	17.8671	0.0118	7.2655	5.5834	3.4847	3.0829	0.0803	0.1454	0.1084
20 %	33.1883	0.0194	9.2462	8.0467	4.7383	4.0242	0.1113	0.2875	0.1461
30 %	35.0566	0.0187	10.7215	9.3582	5.0808	4.5206	0.1278	0.2579	0.1708
40 %	48.1074	0.0218	13.4155	10.8423	6.9788	5.9353	0.1419	0.3151	0.1891
50 %	54.6657	0.0253	15.8391	13.0502	7.3136	6.3725	0.1480	0.4065	0.2274
60 %	64.1971	0.0272	16.6795	17.5084	9.1621	7.1307	0.1636	0.4097	0.2755

Table 21: Standard deviations of each of the regression coefficients in PMM imputation model for each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	8.7071	0.0066	5.4790	3.7292	2.4265	2.2380	0.0581	0.0436	0.0565
20 %	10.9920	0.0085	7.2964	4.7928	3.1147	2.6445	0.0690	0.0474	0.0726
30 %	11.5203	0.0092	7.8844	5.7398	3.8392	3.5224	0.0855	0.0547	0.0817
40 %	14.0913	0.0124	9.5704	6.9552	4.6837	4.3136	0.1128	0.0620	0.1103
50 %	22.2983	0.0158	11.6566	8.6021	5.8480	5.8895	0.1341	0.0895	0.1661
60 %	26.1469	0.0213	12.8420	12.6639	8.7749	7.8529	0.1923	0.0887	0.1588

Table 22: P-values for two-sided one-sample t-tests for each estimated regression coefficient in the CART imputation model at each level of data missingness. The p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.7971	0.4743	0.8033	0.9547	0.3380	0.2730	0.7363	0.8094	0.9500
20 %	0.8373	0.5341	0.7549	0.3274	0.2348	0.0836	0.9696	0.7880	0.6960
30 %	0.4093	0.8400	0.8319	**0.0218**	**0.0297**	**0.0070**	0.8939	0.6953	0.2735
40 %	0.5963	0.2628	0.6113	**0.0022**	**<0.0001**	**0.0005**	0.5117	0.0772	0.0567
50 %	0.3155	**<0.0001**	0.6467	**0.0421**	**<0.0001**	**<0.0001**	0.4771	**0.0005**	0.3136
60 %	0.0961	**<0.0001**	0.6125	**0.0037**	**<0.0001**	**<0.0001**	**0.0037**	**<0.0001**	0.9281

Table 23: P-values for two-sided one-sample t-tests for each estimated regression coefficient in the OTF imputation model at each level of data missingness. The p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**<0.0001**	**<0.0001**	0.6011	0.2723	**0.0442**	**0.0004**	**<0.0001**	**<0.0001**	**<0.0001**
20 %	**<0.0001**	**<0.0001**	0.1117	**0.0494**	**0.0080**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
30 %	**<0.0001**	**<0.0001**	0.4078	0.1184	**0.0013**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	**<0.0001**	0.2257	0.5387	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	**<0.0001**	0.2683	0.6876	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
60 %	**<0.0001**	**<0.0001**	0.5920	**0.0361**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0003**

Table 24: P-values for two-sided one-sample t-tests for each estimated regression coefficient in the RF imputation model at each level of data missingness. The p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**<0.0001**	**<0.0001**	0.8666	0.4040	0.6767	0.1683	0.0642	**<0.0001**	**0.0017**
20 %	**<0.0001**	**<0.0001**	0.1163	**0.0130**	0.6741	0.0735	**<0.0001**	**<0.0001**	**<0.0001**
30 %	**<0.0001**	**<0.0001**	0.0899	**<0.0001**	0.5365	**0.0482**	**<0.0001**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	**<0.0001**	**0.0297**	**<0.0001**	0.5157	**0.0224**	**<0.0001**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	**<0.0001**	0.0769	**<0.0001**	0.9862	**0.0005**	**<0.0001**	**<0.0001**	**<0.0001**
60 %	**<0.0001**	**<0.0001**	0.1924	**<0.0001**	**0.0473**	**0.0008**	**<0.0001**	**<0.0001**	**<0.0001**

Table 25: P-values for two-sided one-sample t-tests for each estimated regression coefficient in the PMM imputation model at each level of data missingness. The p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.9059	0.0697	0.6853	0.5786	0.9684	0.9116	0.4229	0.4653	0.5130
20 %	0.6442	0.0728	0.5172	0.4637	0.4280	0.7120	0.3058	0.3578	0.0649
30 %	0.3630	0.1873	0.1382	0.6441	0.4551	0.5033	**0.0348**	0.6022	0.6521
40 %	0.9491	0.8776	**0.0199**	0.5346	0.9053	0.3936	0.3385	0.3673	0.8050
50 %	0.8456	**<0.0001**	**0.0002**	**0.0040**	0.3614	0.1678	0.1240	**0.0007**	0.7269
60 %	0.0564	**<0.0001**	**<0.0001**	0.9613	0.4559	0.7124	**0.0070**	**<0.0001**	**0.0392**

Table 26: P-values for two-sided one-sample t-tests for each regression coefficient in the CART imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.8826	0.6361	0.8826	0.9727	0.5008	0.4220	0.8284	0.8830	0.9725
20 %	0.9027	0.6804	0.8448	0.4912	0.3757	0.1480	0.9742	0.8773	0.7997
30 %	0.5741	0.9027	0.9027	**0.0466**	0.0062	**0.0016**	0.9420	0.7997	0.4220
40 %	0.7391	0.4174	0.7433	**0.0005**	**<0.0001**	**0.0001**	0.6690	0.1379	0.1073
50 %	0.4765	**0.0002**	0.7718	0.0084	**<0.0001**	**0.0001**	0.6361	**0.0012**	0.4765
60 %	0.1675	**<0.0001**	0.7433	**0.0085**	**<0.0001**	**<0.0001**	**0.0085**	**<0.0001**	0.9592

Table 27: P-values for two-sided one-sample t-tests for each regression coefficient in the OTF imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**<0.0001**	**<0.0001**	0.7391	0.4220	0.0875	**0.0010**	**<0.0001**	**<0.0001**	**<0.0001**
20 %	**<0.0001**	**<0.0001**	0.1929	0.0953	**0.0176**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
30 %	**<0.0001**	**<0.0001**	0.5741	0.2013	**0.0030**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	**<0.0001**	0.3638	0.6804	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	**<0.0001**	0.4220	0.7985	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
60 %	**<0.0001**	**<0.0001**	0.7391	0.0736	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**	**0.0008**

Table 28: P-values for two-sided one-sample t-tests for each regression coefficient in the RF imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**<0.0001**	**<0.0001**	0.9221	0.5741	0.7944	0.2775	0.1206	**<0.0001**	**0.0040**
20 %	**<0.0001**	**<0.0001**	0.1994	**0.0284**	0.7944	0.1335	**<0.0001**	**<0.0001**	**<0.0001**
30 %	**<0.0001**	**<0.0001**	0.1578	**<0.0001**	0.6804	0.0939	**<0.0001**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	**<0.0001**	0.0617	**<0.0001**	0.6690	**0.0474**	**<0.0001**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	**<0.0001**	0.1379	**<0.0001**	0.9862	**0.0012**	**<0.0001**	**<0.0001**	**<0.0001**
60 %	**<0.0001**	**<0.0001**	0.3125	**<0.0001**	0.0929	**0.0018**	**<0.0001**	**<0.0001**	**<0.0001**

Table 29: P-values for two-sided one-sample t-tests for each regression coefficient in the PMM imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.9452	0.1286	0.7985	0.7266	0.9742	0.9466	0.5894	0.6281	0.6690
20 %	0.7718	0.1332	0.6690	0.6281	0.5926	0.8099	0.4685	0.5257	0.1209
30 %	0.5262	0.3065	0.2314	0.7718	0.6232	0.6670	**0.0072**	0.7391	0.7739
40 %	0.9725	0.9292	**0.0043**	0.6804	0.9452	0.5630	0.5008	0.5290	0.8826
50 %	0.9042	**<0.0001**	**0.0005**	**0.0090**	0.5262	0.2775	0.2092	**0.0016**	0.8220
60 %	0.1073	**<0.0001**	**<0.0001**	0.9742	0.6232	0.8099	**0.0156**	**<0.0001**	0.0791

Table 30: Proportion of missing values for the dependent variable and each of the independent variables where 10% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.0985	0.0985	0.0999	0.1031	0.1031	0.1031	0.1038	0.1002	0.0959
Proportion not Missing	0.9015	0.9015	0.9001	0.8969	0.8969	0.8969	0.8962	0.8998	0.9041

Table 31: Proportion of missing values for the dependent variable and each of the independent variables where 20% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.2004	0.1962	0.2015	0.2015	0.2015	0.2015	0.2047	0.2028	0.1933
Proportion not Missing	0.7996	0.8038	0.7985	0.7985	0.7985	0.7985	0.7953	0.7972	0.8067

Table 32: Proportion of missing values for the dependent variable and each of the independent variables where 30% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.2990	0.3009	0.2996	0.2950	0.2950	0.2950	0.3059	0.3042	0.2962
Proportion not Missing	0.7010	0.6991	0.7004	0.7050	0.7050	0.7050	0.6941	0.6958	0.7038

Table 33: Proportion of missing values for the dependent variable and each of the independent variables where 40% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.4002	0.4013	0.3988	0.3999	0.3999	0.3999	0.4071	0.4034	0.3959
Proportion not Missing	0.5998	0.5987	0.6012	0.6001	0.6001	0.6001	0.5929	0.5966	0.6041

Table 34: Proportion of missing values for the dependent variable and each of the independent variables where 50% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.4990	0.5048	0.5028	0.5000	0.5000	0.5000	0.5049	0.4979	0.4973
Proportion not Missing	0.5010	0.4952	0.4972	0.5000	0.5000	0.5000	0.4951	0.5021	0.5027

Table 35: Proportion of missing values for the dependent variable and each of the independent variables where 60% of the total values in the data set are missing.

Variable	BPXSY1	PEASCTM1	cig	cuff1	cuff2	cuff3	BPXPLS	BPXML1	BPXDI1
Proportion Missing	0.5996	0.6061	0.6006	0.5953	0.5953	0.5953	0.6041	0.6052	0.5973
Proportion not Missing	0.4004	0.3939	0.3994	0.4047	0.4047	0.4047	0.3959	0.3948	0.4027

7 EVALUATION OF SECOND DATA SET

A second data set was investigated to determine whether or not we come to the same conclusion about which imputation methods perform the best. The data set chosen contains data collected on the same variables as the original data set collected by the CDC from 2011-2012 [8]. A multiple linear regression model with the same predictor and response variables was fit from 500 randomly selected observations and the procedures were repeated.

7.1 Fitted Model

The reduced multiple linear regression model for the sample data we chose to analyze is:

$$\begin{aligned}\widehat{Y}_i = &- 20.8300 - 0.0001X_{i1} - 4.0160X_{i2} - 2.1030X_{i3} - 0.9129X_{i4} - 0.9303X_{i5} \\ &+ 0.0307X_{i6} + 0.9397X_{i7} + 0.0533X_{i8} + \varepsilon_i \text{ for i=1,...,n,}\end{aligned} \tag{2}$$

where the predictor variables include PEASCTM1, cig, cuff1, cuff2, cuff3, BPXPLS, BPXML1, and BPXDI1, respectively. These are the values to which the estimated regression coefficients from the imputation models will be compared.

7.2 Estimated Regression Coefficients

In Tables 36-40, we observe that the CART and PMM methods seem to produce estimated regression coefficients most similar to the regression model fit with the complete sample of 500 observations. This holds true across all levels of data missingness. In Tables 41-44, we observe the PDI values calculated for each of the estimated re-

gression coefficients. As in the evaluation of the original model, we observe many PDI values that are very large and warrant further investigation into the prediction accuracies of the imputation models.

7.3 Model Accuracy

In Tables 45-49, we observe the R^2 and adjusted R^2 values for the multiple regression model produced by the complete sample data set, as well as the four imputation models. The adjusted R^2 value in Table 45 will serve as the baseline comparison. Across all levels of data missingness, we conclude that the CART and PMM imputations perform better than both the OTF and RF imputation methods. The adjusted R^2 values for the OTF and RF imputation methods drastically decrease at each increased level of data missingness. While these values still decrease at each increased level of data missingness for the CART and PMM imputation methods, they decrease less steeply and are much closer to the baseline adjusted R^2 value, especially at low levels of data missingness.

7.4 Significance of Regression Coefficients of Imputation Models, Adjusted for Multiple Testing

In Tables 50-53, we observe the p-values associated with each estimated regression coefficient for each imputation method at every level of data missingness. The estimated regression coefficients are compared to the parameter values for the multiple regression model produced by the complete sample data set. The p-values are adjusted for multiple testing using the BH procedure discussed in Chapter 6. We con-

clude that the OTF and RF imputation methods produce more estimated regression coefficients that are significantly different from the parameter values.

Table 36: Parameter values for the multiple regression model produced by the complete sample data set.

Parameters	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
Actual Values	-20.8300	-0.0001	-4.0160	-2.1030	-0.9129	-0.9303	0.0307	0.9397	0.0533

Table 37: Estimated means of the regression coefficients from the CART imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-21.2605	-0.0027	-4.6277	-1.7124	-0.7439	-1.0069	0.0335	0.9544	0.0523
20 %	-15.6301	0.0007	-4.0122	-2.6625	-0.2332	-1.3764	-0.0118	0.9382	0.0159
30 %	-14.3024	0.0006	-2.8831	0.7332	1.5796	0.2031	-0.0495	0.9175	0.0569
40 %	-14.9107	-0.0021	-0.0553	1.4948	0.3010	0.3224	-0.0301	0.9208	0.0712
50 %	-16.4168	-0.0035	-1.0626	0.1845	1.5938	0.2832	-0.0338	0.9369	0.0874
60 %	-8.6102	0.0005	-0.4473	0.2893	0.2770	0.1947	-0.0411	0.8193	0.1882
Actual Parameter	-20.8300	-0.0001	-4.0160	-2.1030	-0.9129	-0.9303	0.0307	0.9397	0.0533

Table 38: Estimated means of the regression coefficients from the OTF imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	7.8708	0.0070	-1.3036	-3.0349	-2.7811	-1.9303	-0.0404	0.6990	0.1675
20 %	21.9121	0.0095	-0.8006	-3.7804	-0.3938	0.2656	-0.0271	0.5815	0.1464
30 %	35.9125	0.0099	0.0006	-2.1485	0.7587	1.5384	-0.0660	0.4686	0.1997
40 %	56.5333	0.0038	1.3003	-3.4527	0.5749	1.0020	-0.0809	0.3607	0.1940
50 %	73.8394	0.0018	1.0494	-3.9028	1.0360	0.7028	-0.0543	0.2508	0.1684
60 %	81.0483	0.0021	0.4980	-2.4112	0.6743	0.0511	-0.0282	0.1833	0.1700
Actual Parameter	-20.8300	-0.0001	-4.0160	-2.1030	-0.9129	-0.9303	0.0307	0.9397	0.0533

Table 39: Estimated means of the regression coefficients from the RF imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-6.2951	0.0026	-3.7291	-2.7034	-2.6189	-1.9671	-0.0099	0.8217	0.1250
20 %	-1.1701	0.0060	-2.6174	-3.7397	-1.4013	-1.5096	-0.0327	0.7841	0.1121
30 %	9.0895	0.0074	-2.2824	-0.8898	1.3834	1.4233	-0.0549	0.6699	0.1725
40 %	26.2822	0.0035	0.5323	-2.5455	1.0459	1.2953	-0.0762	0.5653	0.1988
50 %	37.1410	0.0041	0.7218	-3.5494	2.0315	0.7251	-0.0621	0.4685	0.2332
60 %	50.0012	0.0039	-0.0420	-3.3224	1.0054	0.6064	-0.0651	0.3599	0.2745
Actual Parameter	-20.8300	-0.0001	-4.0160	-2.1030	-0.9129	-0.9303	0.0307	0.9397	0.0533

Table 40: Estimated means of the regression coefficients from the PMM imputation model at each level of missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-21.0005	-0.0027	-5.0906	-1.9556	-1.7401	-1.7896	0.0290	0.9504	0.0720
20 %	-20.0858	-0.0002	-4.2426	-3.1011	-1.6131	-2.2965	-0.0056	0.9807	0.0120
30 %	-16.6595	-0.0016	-5.0627	2.1404	0.7299	-0.9080	-0.0578	0.9421	0.0842
40 %	-16.9009	-0.0061	1.4802	2.4704	0.1653	0.0128	-0.0295	0.9659	0.0429
50 %	-13.5214	-0.0072	2.5102	3.0438	1.8492	0.2591	-0.0769	0.9345	0.1265
60 %	-7.9787	-0.0012	0.2329	4.1580	1.1737	-0.3556	-0.1095	0.8463	0.2048
Actual Parameter	-20.8300	-0.0001	-4.0160	-2.1030	-0.9129	-0.9303	0.0307	0.9397	0.0533

Table 41: Percent deviation indices of CART imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-2.0669	-1865.0660	-15.2316	18.5735	18.5124	-8.2339	-9.1205	-1.5643	1.8025
20 %	24.9634	609.4614	0.0946	-26.6049	74.4550	-47.9523	138.4365	0.1596	70.1465
30 %	31.3377	536.6812	28.2097	134.8645	273.0310	121.8317	261.2378	2.3625	-6.8344
40 %	28.4173	-1428.3840	98.6230	171.0794	132.9718	134.6555	198.0456	2.0113	-33.6838
50 %	21.1866	-2447.3070	73.5408	108.7732	274.5865	130.4418	210.0977	0.2980	-64.1006
60 %	58.6643	463.9010	88.8621	113.7565	130.3429	120.9287	233.8762	12.8126	-253.3609
MEAN	27.0837	-688.4522	45.6831	86.7404	150.6499	75.2786	172.0956	2.6800	-47.6718
Previous MEAN	-1.4869	16.6836	-17.4104	27.9375	40.1712	53.1651	10.0354	0.9389	-9.5212

Table 42: Percent deviation indices of OTF imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	137.8292	5194.6140	67.5398	-44.3129	-204.6445	-107.4922	231.5961	25.6146	-214.4949
20 %	205.1950	7014.1190	80.0647	-79.7622	56.8627	128.5499	188.2736	38.1186	-174.8780
30 %	272.4077	7305.2400	100.0149	-2.1636	183.1088	265.3660	314.9837	50.1330	-274.9531
40 %	371.4034	2865.6480	132.3780	-64.1797	162.9751	207.7072	363.5179	61.6154	-264.2508
50 %	454.4859	1410.0440	126.1305	-85.5825	213.4845	175.5455	276.8730	73.3106	-216.1848
60 %	489.0940	1628.3840	112.4004	-14.6553	173.8635	105.4929	191.8567	80.4938	-219.1889
MEAN	321.7359	4236.3415	103.0881	-48.4427	97.6084	129.1948	261.1835	54.8810	-227.3251
Previous MEAN	493.9170	192.4625	-36.6743	-8.7468	50.8748	105.6318	179.1345	57.2385	-151.0797

Table 43: Percent deviation indices of RF imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	69.7785	1992.2850	7.1439	-28.5497	-186.8770	-111.4479	132.2476	12.5572	-134.6977
20 %	94.3826	4466.8120	34.8257	-77.8269	-53.4998	-62.2702	206.5147	16.5585	-110.4769
30 %	143.6367	5485.7350	43.1673	57.6890	251.5391	252.9937	278.8274	28.7113	-223.8828
40 %	226.1749	2647.3070	113.2545	-21.4693	214.5690	239.2347	348.2085	39.8425	-273.2632
50 %	278.3054	3083.9880	117.9731	-68.7779	322.5326	177.9426	302.2801	50.1437	-337.8520
60 %	340.0443	2938.4280	98.9542	-57.9838	210.1325	165.1833	312.0521	61.7005	-415.3962
MEAN	192.0537	3435.7592	69.2198	-32.8198	126.3994	110.2727	263.3551	34.9190	-249.2615
Previous MEAN	236.9394	176.9514	-67.3982	-65.4385	-6.8126	42.8547	139.6720	30.7472	-158.0981

Table 44: Percent deviation indices of PMM imputation model estimated regression coefficients at each level of data missingness.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	-0.8183	-1865.0660	-26.7580	6.9615	-90.6123	-92.3681	5.5375	-1.1387	-35.1859
20 %	3.9435	-45.5604	-5.6424	-47.4608	-76.7006	-146.8559	118.2410	-4.3631	77.4690
30 %	20.0217	-1064.4830	-26.0633	201.7784	179.9540	2.3971	288.2736	-0.2554	-58.0924
40 %	18.8628	-4339.5920	136.8576	217.4703	118.1071	101.3759	196.0912	-2.7881	19.4518
50 %	35.0867	-5140.1750	162.5050	244.7361	302.5633	127.8512	350.4886	0.5537	-137.5141
60 %	61.6961	-773.3624	105.7993	297.7175	228.5683	61.7758	456.6775	9.9393	-284.5287
MEAN	23.1321	-2204.7065	57.7830	153.5338	120.3133	9.0293	235.8849	0.3218	-69.7334
Previous MEAN	1.6953	8.2890	-97.3219	-13.0754	-3.2129	-8.3749	-15.9357	0.0051	-1.5918

Table 45: R^2 and Adjusted R^2 values for multiple regression model produced by the complete sample data set.

R^2	Adjusted R^2
0.8298	0.8271

Table 46: R^2 and Adjusted R^2 values for CART imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.8289	0.8179	0.8133	0.7856	0.7587	0.7411
Adjusted R^2	0.8261	0.8149	0.8103	0.7821	0.7547	0.7369
Previous Adj. R^2	0.8512	0.8524	0.8519	0.8525	0.8450	0.8249

Table 47: R^2 and Adjusted R^2 values for OTF imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.5543	0.3760	0.2797	0.1836	0.0889	0.0621
Adjusted R^2	0.5470	0.3658	0.2680	0.1702	0.0737	0.0463
Previous Adj. R^2	0.5641	0.3602	0.2235	0.1297	0.0649	0.0279

Table 48: R^2 and Adjusted R^2 values for RF imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.6962	0.6075	0.5156	0.4088	0.2735	0.2162
Adjusted R^2	0.6912	0.6011	0.5077	0.3991	0.2616	0.2032
Previous Adj. R^2	0.7395	0.6419	0.5431	0.4362	0.3368	0.2294

Table 49: R^2 and Adjusted R^2 values for PMM imputation models at each level of data missingness.

% Imputed	10%	20%	30%	40%	50%	60%
R^2	0.8293	0.8258	0.8167	0.8096	0.7899	0.7935
Adjusted R^2	0.8265	0.8229	0.8138	0.8065	0.7864	0.7902
Previous Adj. R^2	0.8528	0.8535	0.8539	0.8556	0.8520	0.8380

Table 50: P-values for two-sided one-sample t-tests for each regression coefficient in the CART imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.9712	0.3366	0.9092	0.9092	0.9516	0.9777	0.9740	0.6585	0.9875
20 %	0.3330	0.7906	0.9989	0.8453	0.6649	0.7895	0.2740	0.9800	0.2913
30 %	0.2189	0.8159	0.7906	0.2249	0.0663	0.4158	**0.0234**	0.4706	0.9625
40 %	0.2670	0.4677	0.2677	0.1096	0.4102	0.3759	0.1061	0.5508	0.6396
50 %	0.4158	0.2021	0.4158	0.3330	0.0648	0.3922	0.0829	0.9709	0.3366
60 %	**0.0096**	0.8453	0.3284	0.3082	0.4158	0.4183	**0.0488**	**<0.0001**	**<0.0001**

Table 51: P-values for two-sided one-sample t-tests for each regression coefficient in the OTF imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**<0.0001**	**0.0018**	0.4635	0.7253	0.1883	0.4706	0.0510	**<0.0001**	**0.0001**
20 %	**<0.0001**	**<0.0001**	0.3834	0.4738	0.7603	0.3994	0.1232	**<0.0001**	**0.0023**
30 %	**<0.0001**	**<0.0001**	0.2670	0.9875	0.2451	0.0552	**0.0044**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	0.1232	0.1232	0.5893	0.3069	0.1517	**0.0011**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	0.4706	0.1488	0.4527	0.1671	0.2350	**0.0146**	**<0.0001**	**0.0001**
60 %	**<0.0001**	0.4102	0.2039	0.9398	0.2677	0.4770	0.1166	**<0.0001**	**<0.0001**

Table 52: P-values for two-sided one-sample t-tests for each regression coefficient in the RF imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	**0.0015**	0.3070	0.9712	0.8343	0.2350	0.4638	0.3037	**<0.0001**	**0.0240**
20 %	**<0.0001**	**0.0094**	0.7316	0.4839	0.7759	0.7138	0.0898	**<0.0001**	0.0791
30 %	**<0.0001**	**0.0011**	0.6585	0.6327	0.0960	0.0710	**0.0138**	**<0.0001**	**<0.0001**
40 %	**<0.0001**	0.1637	0.2021	0.8933	0.1657	0.0923	**0.0015**	**<0.0001**	**<0.0001**
50 %	**<0.0001**	0.0960	0.1783	0.5508	**0.0240**	0.2314	**0.0071**	**<0.0001**	**<0.0001**
60 %	**<0.0001**	0.1149	0.2677	0.6327	0.1739	0.2670	**0.0052**	**<0.0001**	**<0.0001**

Table 53: P-values for two-sided one-sample t-tests for each regression coefficient in the PMM imputation model at each level of data missingness, adjusted for multiple testing. The adjusted p-values that are in bold are for one-sample t-tests that are significant at $\alpha = 0.05$ family level of significance.

% Imputed	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$
10 %	0.9875	0.3366	0.8037	0.9745	0.5894	0.5508	0.9813	0.7675	0.6264
20 %	0.9189	0.9875	0.9745	0.7101	0.6585	0.3287	0.3603	0.1706	0.2402
30 %	0.4431	0.6113	0.8039	0.0510	0.2557	0.9875	**0.0106**	0.9712	0.3922
40 %	0.4677	**0.0116**	0.1116	**0.0324**	0.4638	0.4971	0.1096	0.4015	0.8039
50 %	0.1622	**0.0018**	0.0510	**0.0136**	**0.0372**	0.4006	**0.0015**	0.9092	**0.0212**
60 %	**0.0059**	0.7253	0.2350	**0.0018**	0.1345	0.7138	**<0.0001**	**0.0003**	**<0.0001**

8 CONCLUSION

Modern MICE methods used to impute missing data values were described and four methods used for mixed data were implemented to compare the different methods. The PDI values for each of the estimated regression coefficients were compared to the parameters produced by the multiple regression model using the complete data set. We determined that the CART and PMM imputation methods yielded estimated regression coefficients much more similar to the parameters produced by the multiple regression model than the OTF and RF imputations did. We confirmed this once again by performing one-sample t-tests for each of the estimated regression coefficients. While some of the p-values from the t-tests were significant for the CART and PMM imputation methods, many more of the p-values from the t-tests were significant for the OTF and RF imputation methods, with well over half of them being significant. These observations led us to analyze the R^2 and adjusted R^2 values for each of the imputation methods at each level of data missingness. We concluded that the OTF and RF imputation methods do not perform well, especially as the proportion of missing values increases. However, the CART and PMM imputation methods both performed well at all levels of data missingness, all while producing R^2 and adjusted R^2 values close to the original values from the multiple regression model produced from the complete data set. One interesting observation to note is that at the 40% level of data missingness, the adjusted R^2 values for the CART and PMM imputation methods slightly increased from the 30% level of data missingness.

The procedures were repeated using a second sample of data, and similar conclusions were reached. Across all levels of data missingness, the CART and PMM

imputation methods performed better than the OTF and RF imputation methods. However, the notable increase in the adjusted R^2 values for the CART and PMM imputation methods from the original data set was not present. We recommend that either the CART or PMM imputation method be used in any situation where any level of missing data is present.

9 FUTURE WORK

In the future, we would like to compare our findings to simulated data. A simulation can serve as a controlled experiment in which we test how varying certain parameters affects other parameter estimates [9]. With a simulation, we can test observe whether or not our conclusions about the CART and PMM imputation methods hold true. However, in simulation studies, one must be careful to properly design and analyze the experiment [10]. Since simulated data is produced from random samples, one must be careful not to regard the regression estimates as true values.

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] *Why You Probably Need More Imputations Than You Think*, by Paul Allison, Published by Statistical Horizons, 2012.

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

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

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

[7] *National Health and Nutrition Examination Survey, 2007-2008 Data Documentation, Codebook, and Frequencies*, Published by Centers for Disease Control and Prevention, 2009.

[8] *National Health and Nutrition Examination Survey, 2011-2012 Data Documentation, Codebook, and Frequencies*, Published by Centers for Disease Control and Prevention, 2013.

[9] *Applied Hierarchical Modeling in Ecology*, by Marc Kry and J. Andrew Royle, Published by Science Direct, 2016.

[10] *Statistical Analysis of Simulation Output Data: The Practical State of the Art*, by Averill M. Law Published by Averill M. Law & Associates, 2010.

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

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

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

[15] *Efficient estimation with missing data in multilevel models*, by Harvey Goldstein and Geoffrey Woodhouse, Published by Institute of Education in London U.K., 1996.

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

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

[18] *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.

[19] *Imputation By Classification And Regression Trees* , by Stef van Buuren, Published by RDocumentation, 2018.

[20] *Multiple Imputation for Missing Data: Concepts and New Development*, by Yang C. Yuan, Published by SAS Institute Inc., 2016.

[21] *R-Squared Definition*, by Adam Hayes, Published by Investopedia, 2019.

[22] *A general introduction to adjustment for multiple comparisons*, by Shi-Yi Chen, Zhe Feng, and Xiaolian Yi, Published by Journal of Thoracic Disease, 2017.

[23] *Benjamini-Hochberg Procedure*, by Stephanie, Statistics How To, 2015.

[24] *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.

VITA

KAITLYN HEIDT

Education: B.S. Mathematics, East Tennessee State University,
Johnson City, Tennessee 2017
M.S. Mathematical Sciences, East Tennessee State University,
Johnson City, Tennessee 2019