5-2020

Investigation of Multiple Imputation Methods for Categorical Variables

Samantha Miranda
East Tennessee State University

Investigation of Multiple Imputation Methods for Categorical Variables

---

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

Samantha Miranda

May 2020

Christina Nicole Lewis, Ph.D., Chair

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

JeanMarie L. Hendrickson, Ph.D.

Keywords: Missing data, multiple imputation methods, categorical data

ABSTRACT

Investigation of Multiple Imputation Methods for Categorical Variables

by

Samantha Miranda

We compare different multiple imputation methods for categorical variables using the MICE package in R. We take a complete data set and remove different levels of missingness and evaluate the imputation methods for each level of missingness. Logistic regression imputation and linear discriminant analysis (LDA) are used for binary variables. Multinomial logit imputation and LDA are used for nominal variables while ordered logit imputation and LDA are used for ordinal variables. After imputation, the regression coefficients, percent deviation index (PDI) values, and relative frequency tables were found for each imputed data set for each level of missingness and compared to the complete corresponding data set. It was found that logistic regression outperformed LDA for binary variables, and LDA outperformed both multinomial logit imputation and ordered logit imputation for nominal and ordered variables. Simulations were ran to confirm the validity of the results.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Nicole Lewis. Her advice and support helped guide me not only through research but also through my time at ETSU. I would also like to thank my family and friends for there love and encouragement through my educational journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Statistical methods used in research are utilized throughout businesses, medical fields and biological fields who use the results of the analysis to make big decisions. It is important to have a complete data set to ensure that results are accurate. Complete data sets lead to the most accurate results. However, a complete data set is not always available. Although missing data is very common, it can be hard to avoid and lead to inaccurate and biased results. The sample size can be drastically reduced with even just a small amount of missing data points. A smaller sample size can cause confidence intervals to be less precise and decreased power in tests for significance. Dealing with missing data can be difficult as it requires in depth analysis to identify the type of missingness and decide what type of imputation method would be best, but it is a vital part of the data preprocessing to ensure one yields the most efficient results.

A case study from a consumer good's company describes how they have been affected by missing data. The consumer good's company uses a few survey questions on the warranty card returned by customers as their main source of information from customers. The company's marketing team wants to gain a better understanding of the demographics of their customer base to more efficiently target promotions. The customer demographics characteristics are combined into groups: gender, occupation, marital status and income. Ignoring the missing data it is shown that the largest customer group is married women at 38.4%, which is an equal split between professional and non-professional women. When exploring the missing data, it was found that the majority of the missing data was on income, about 34% of the cards returned

had skipped this question. With further investigation it is found that more women than men did not answer the income question. Using maximum likelihood estimation and imputation, a complete data set is created. After rerunning the analysis with the new complete data set, women still account for the largest amount of customers. Women now make up 46% of total customers with 26% being non-professional married women and 18% being professional married women. Though there was other smaller percentages of missing data among the groups, the largest group appeared to be non-professional married women. Although we are unaware of the exact reason why the non-professional married women had a lower response rate for income, it altered the results for the company's survey. Moving forward the company can take these results and adjust their promotions to better reach the different demographic characteristic groups [23].

From the case study, we can see that even though the missing data was largely effecting one demographic, it had a big impact on the analysis of the data. Here the missing data could have a large effect on the budget for promotions. The misleading results caused by missing data can not only effect the analysis but also alter the application of the analysis.

1.1 Proposed Work

In this study, we examine modern methods for dealing with missing data to determine the best method when working with categorical variables. Each modern method for categorical variables will be applied to the data set that have different percentages of missingness and then compared to a complete data set. Simulations will then be

run to confirm the validity of the results.

1.2 Overview of Thesis

The thesis is arranged as follows. Chapter 2 identifies and describes the three types of missing data. Chapter 3 describes methods of imputation. Section 3.1 explains traditional methods and Section 3.2 explains modern methods. Chapter 4 further explains multiple imputation for categorical variables, specifically the methods this thesis implements. Chapter 5 describes the data source. Chapter 6 describes our proposed method. Chapter 7 explains the results of our methods and illustrates the precision of our method via simulations. Chapter 8 concludes the thesis.

2 TYPES OF MISSING DATA

When working with missing data it is important to examine the data carefully to identify the pattern of missingness. Correctly identifying the pattern of missingness is the first step to choosing the appropriate method to deal with missing data. It is difficult to give a reason for the missing data. Surveys might ask for private information that some respondents may be unwilling to give, respondents may feel that questions are inapplicable and choose not to answer them, or respondents may simply forget to answer questions.

There are three different types of missing data, missing completely at random, missing at random, and missing not at random. Data is missing completely at random (MCAR) if suppose we have missing data on a particular variable Y. The data on Y are said to be MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variables in the data set. Since in MCAR the missingness is unrelated to the data, we can show this type of missingness with the observed Y values (Y_{obs}), the missing Y values (Y_{mis}), and where R is the missing data indicator either 0 (missing) or 1 (observed). We can simply state the probability of a missing value on Y as $P(R|\phi)$ where ϕ is a parameter describing the relationship between R and the data. Data that is missing completely at random can be thought of as a simple random sample. For example, let's say we were studying the main determinants of income. Now consider age as one of the main determinants. We would assume MCAR if the people that did not report their income was not related to their age. The MCAR assumption would be violated if the people that did not report their income were, on average, younger than those that reported it.

A simple way to test if the MCAR assumption is fulfilled is to split the data into two groups, people that did report their income and people that did not report their income. We then test if there is a difference in the mean ages of the two groups. Our null hypothesis would be $H_0 : \mu_1 = \mu_2$ where μ_1 is the mean age of people who did the report their income and μ_2 is the mean age of people who did not report their income. Our alternative hypothesis would be $H_1 : \mu_1 \neq \mu_2$. If we conclude that there is significant evidence against the null hypothesis, then we can assume MCAR.

A weaker assumption to MCAR is that data is missing at random (MAR). Data on Y are said to be missing at random if the probability of missing data on Y is unrelated to the value of Y, after controlling for other variables in the analysis. We can formally state this as: $P(Y_{mis}|Y, X) = P(Y_{mis}|X)$ where for the two variables X and Y, X is always observed and Y is sometimes missing. For example, the MAR assumption would be satisfied if the probability of missing data on income depended on the persons age, but within each age group the probability of missing data on income was unrelated to the persons income. Another example of the MAR assumption being satisfied would be if the probability of missing data on income depended on marital status, but within each marital status category the probability of missing data on income is unrelated to income. As of now there is no way to confirm the MAR mechanism.

If the MAR assumption is fulfilled, then the missing data is said to be ignorable. If the missing data is ignorable, there is no need to model for the missing data in the analysis of the data set. If the MAR assumption is not fulfilled the missing data is said to be nonignorable. In this case we need to have a strong grasp on

the appropriate methods of dealing with missing data because we must model for the missing data. Modeling this missing data is necessary to allow for accurate estimations of the parameters in interest.

The third and most commonly problematic type of missing data is missing not at random (MNAR). This is where the missing values do depend on the other values. Data that is missing not at random is considered nonignorable because most analysis models are not accurate with this type of missingness. The distribution of data missing not at random can be written as $P(R|Y_{obs}, Y_{mis}, \phi)$, where P is the standard probability distribution, R is the missing data indicator, Y_{obs}, and Y_{mis} are the observed and missing parts of the data and ϕ is a parameter that describes the relationship between R and the data. In other words, the probability of missing data on Y can depend on the other variables such as Y_{obs}, or even Y_{mis}. There is no way to test for data MNAR. For example, we would fulfill the MNAR assumption if the people with lower income are less likely to report their income.

Identifying the proper cause of missingness is key to an accurate analysis. Proper identification leads to the correct steps to account for the missing data during analysis. The most common type of missing data is data that is missing at random. Missing data is less often found to be MCAR and MNAR since it is difficult to reach these conclusions when testing those two situation. MAR is not as random as commonly thought. Here some may say that the word random could be interchanged with conditioned or controlled for, because once one has conditioned or controlled for the data the remaining of the missingness is random. Taking the time to properly identify the type missing data will lead to more accurate analysis in the end.

3 METHODS OF IMPUTATION

For years methodologists have been studying missing data. Techniques have been proposed and altered with some methods being more successful then others. This chapter identifies and describes certain traditional methods along with their possible flaws. It also describes modern methods of dealing with missing data. Traditional methods include different deletion methods and single imputation methods while modern methods include joint modeling and multiple imputation.

3.1 Traditional Methods

Deletion methods such as pairwise and listwise deletion are the most common of the traditional methods for handling missing data. This is because deletion is easy, convenient and often included as a package on most statistical software. Listwise deletion discards the data for any case that has one or more missing values. Pairwise deletion tries to mitigate the loss of data by eliminating cases on an analysis-by-analysis basis. The main reason for using a deletion method is convenience. The list of disadvantages here far outweighs the advantages. The main problem with the two deletion methods is that they both require that the data is missing completely at random. Even if this assumption is met, the MCAR data can still produce distorted parameter estimates. Deleting data can lead to underrepresented cases and can dramatically decease the sample size. Even if the MCAR assumption is fulfilled, eliminating the data is wasteful and can significantly reduce power.

The other traditional methods of dealing with missing data include single imputation methods. Some single imputation methods include arithmetic mean imputation,

regression imputation, and stochastic regression imputation. Arithmetic mean imputation simply fills in the missing data with the arithmetic mean of the available cases. Though appealing to have a complete data set, entering in values at the center of the data set can dramatically alter the not only the spread of the data, but also the standard deviation and variance of the data. Regression imputation replaces the missing data with the predicted values from a regression equation. Since variables tend to be correlated, it makes sense to generate imputations that shares information with the observed data. This idea of sharing information with the observed data is also seen in maximum likelihood and multiple imputation, but they go about it in a more sophisticated manner. Replacing the missing values with the predicted values leads to a high predictable bias. Other traditional single imputation methods replace the missing data with the most common value for the data or look for similar response patterns. In longitudinal studies, they might use the last observation that was recorded before the missing value to replace the missing value.

Traditional methods can alter the summary statistics of the data set and introduce high amount of bias to the model. Even with the convenience and simplistic idea, the traditional ideas are not recommended

3.2 Modern Methods

Since the traditional methods are not recommended by most researchers, modern methods have been created to obtain more accurate methods for dealing with missing data. Most modern methods can be divided into two groups: joint modeling or fully conditional specification.

3.2.1 Joint Modeling

Joint modeling is one of the main modern methods of dealing with missing data. Joint modeling begins when the data can be described as a multivariate distribution. This model can be based on any multivariate distribution but the multivariate normal happens to be the most common. Joint modeling partitions the observations into groups of identical missing data patterns and imputes the missing entries within each according to a joint model for X, Y, and R that is common to all observations. Assuming the data is missing at random, the imputations are created as draws from the fitted distribution.

When dealing with categorical data the multivariate normal model is the most appropriate model. Joint modeling with categorical variables can either use rounding or not use rounding. Several models have been proposed that use rounding. Horton, Lipsitz, and Parzen (2003) suggest that rounding off continuous imputed values in categorical data to the nearest category to preserve the distributional properties as fully as possible and to make them intelligible to the analysis. They also showed that simple rounding may introduce bias in the estimates of interest, especially the binary variables. Bernaards, Belin, and Schafer (2007) confirmed the results of Horton's simple rounding approach and proposed two possible improvements to simple rounding: coin flip and adaptive rounding. Their simulations showed that adaptive rounding seemed to provide the best performance, although its advantage over simple rounding was sometimes slight. Many more proposals have been created by other researchers. One researcher proposed a rounding method based on logistic regression and an additional drawing step that makes rounding dependent on other variables in

the imputation model. Another proposal is to model the indicator variables of the categorical variables. Since a single best rounding method has not been identified, most researchers say to try to avoid rounding if possible and focus on using methods that are specific to categorical variables.

Several joint modeling methods for categorical variables have been proposed that do not include rounding. Missing data in contingency tables can be imputed under the log-linear model. The model preservers the high order interactions and works best if the number of variables is small, no more then six. Olkin and Tate (1961) developed the general location model for imputing mixed continuous-categorical data. This model combines the log-linear and multivariate normal models by fitting a restricted normal model to each cell of the contingency table. It had been found that this model has limitations when the data set has a large number of variables, specifically when there are more than ten continuous and categorical variables. Other imputation methods based on joint modeling incorporate the k-means clustering algorithm and other two-way imputation proposals.

3.2.2 Fully Conditional Specification

Fully Conditional Specification (FCS) imputes multivariate missing data on a variable-by-variable basis. When using this method an imputation method must be specified for each incomplete variable, and iteratively creates imputations. FCS specifies a multivariate distribution through a set of conditional distributions. We use FCS in an attempt to define $P(Y, X, R|\phi)$ by a conditional density $P(Y_i|X, Y_i, R, \phi)$ for each Y_i. After beginning with random draws from the marginal distributions,

imputation is done by iterating over the conditionally specified imputation model. Rubin [14] broke up the imputation procedure into three parts: modeling, estimation, and imputation. Modeling determines a specific model for the data. Then estimation creates the posterior distribution given the model. Finally, the imputation takes the random draws for the missing data by drawing successively from parameter and data distributions. FCS is an appealing method because it does not restrict the conditional distribution to follow a normal distribution. This allows for flexibility when creating the multivariate models. It can also use specialized imputation methods that are difficult to formulate as a part of a multivariate density.

The main method in FCS is multiple imputation (MI). Rubin developed MI in the 1970s and it is now accepted as the best general method to deal with missing data. The idea of imputation was first thought of in the 1930s by Allen and Wishart and involved two formulas used to estimate the value of a single missing value and this value replaced the missing data point. As time went on, the advancements were made to generalize the idea to impute value for more than one missing data point. Researchers could not guarantee that the single imputed value was accurate so the idea to have multiple imputations for a single missing data point was originated. With the amount of technology available at the time, the idea of having multiple imputations for the missing values was quite ambitious. For this reason, instead of including the formulas for calculating the combined estimates, Rubins original proposal stressed the study of variation due to the uncertainty of the estimates [10] .

The principle of multiple imputation is to generate m imputed data sets to reflect the uncertainty of the imputed values. Multiple imputation can be described in three

steps. The first step involves generating and identifying initial values for the missing values for all variables $Y_1^{(0)}, ..., Y_k^{(0)}$. For a categorical variable with missing values, use the non-missing values to find the observed portion of each category, then fill in the missing value with random draws from a multinomial distribution with category probabilities equal to the observed category proportions. The second step consists of analyzing each imputed data set by a statistical method that will estimate the quantities of interest. The last step pools the m estimates into one estimate. Here we are combining the variation within and across the m imputed data sets. The three-step process can describe all multiple imputation procedures, but the imputation step can be altered to the specific data analysis which is intended by the researcher [21]. The imputation step can be altered based on the type of categorical variable that is being used in the analysis.

4 MULTIPLE IMPUTATION FOR CATEGORICAL VARIABLES

Multiple imputation (MI) is one of the most common approaches that researchers use to study a data set with missing values as a whole. When dealing with categorical variables, there are three different types of variables: binary, nominal, and ordinal. Each type of variable has an MI method that is specific to it. Binary variables have two mutually exclusive levels. An example of a binary variable would a categorical variable with the levels: yes or no, on or off, agree or disagree. For binary variables, we use logistic regression imputation. A nominal variable is a categorical variable that has more than two levels. For example, a nominal variable could be eye color that has levels blue, green, brown, hazel, etc. For nominal variables, we use multinomial logit model imputation. The last type of categorical variable is ordinal. Ordinal variables consist of more than two levels that have a natural ordering. This variable could be in the form of performance (first, second or, third place) or day of the week. For example, a researcher is interested in what factors influence medaling in Olympic swimming. Relevant predictors include training hours, diet, age, and popularity of swimming in the athlete's home country. The researcher believes that the distance between gold and silver is larger than the distance between silver and bronze. Depending on the study or research question at hand a nominal variable could be considered ordinal. Ordinal variables can use the ordered logit model also known as, proportional odds model for imputation. Linear discriminant analysis (LDA) is another imputation used for all types of categorical variables and thus, will be used as well.

4.1 Logistic Regression Imputation

Logistic regression is a statistical model that uses a logistic function to model data with a binary categorical dependent variable. The binary response variable takes on values 0 and 1 with probabilities π and $1-\pi$, respectively. The response variable Y is a Bernoulli random variable with parameter $E(Y) = \pi$. A simple logistic regression model is in the form: $Y_i = E(Y_i) + \varepsilon_i$. The distribution of the error term ε_i depends on the Bernoulli distribution of the response variable Y_i. Since Y_i are independent Bernoulli random variables with expected values $E(Y_i) = \pi_i$, the simple logistic model can be shown as: $E(Y_i) = \pi_i = \frac{e^{(\beta_0+\beta_1 X_i)}}{1+e^{(\beta_0+\beta_1 X_i)}}$, where the X_i are assumed to be a known constant. As the number of variables and coefficients increase, we add more β's on to the model. We add a β for every predictor variable. A logistic model with multiple predictor variables can be shown as: $E(Y_i) = \pi_i = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\ldots+\beta_k X_k)}}$, where $X = (X_1, ..., X_k)$ are k predictors. Logistic regression can also be fit via maximum likelihood estimation. Maximum likelihood estimation is a method of estimating the parameters of a probability function. The parameters are found by maximizing a likelihood function that is described by the model produced by the data. The likelihood function is chosen in order to make the observed data "most likely". The most common likelihood function is the log-likelihood function.

To impute the missing values using the logistic regression imputation, the variable with missing data points will be treated as the response and the other variables will be treated as the predictors. For a completed data set, a random draw is made from the posterior distribution of the parameters. A data set is said to be complete if for every observation their is a value for each predictor variable. Based on the fitted

logistic regression equation, a probability is generated for each case with missing data and a Bernoulli draw is made for the probability, producing imputed values of 0 or 1 [2]. In other words, a probability is generated for each level of the binary variable, and the imputed value is assigned a 0 or 1 depending on which probability is larger. Consider the example predicting credit default (yes or no) from balance, income, and student (yes or no). We can make a prediction by plugging in the values of the predictor variables into the model. We then predict the response Default-Yes (1) if the probability is greater 0.5 and we predict Default-No (0) if the probability is less than 0.5.

4.2 Linear Discriminant Analysis

Where logistic regression performs well when the data consists of categorical variables with two levels, when the variables have more than two levels linear discriminant analysis (LDA) is the better option. LDA can be utilized in a variety of different environments. One would be determining good credit and bad credit based on income, age, number of credit cards, and family size. Another example would be classifying alcoholics and non-alcoholics based on the activity of monoamine oxidase enzyme, activity of adenylate cyclase enzyme. It is preferred to use LDA instead of logistic regression when the categorical variables have more then two levels because logistic regression only works for categorical variables with two levels. LDA also outperforms logistic regression when the number of observations, n, is small and the distribution of the predictors X is approximately normal in each of the classes [18]. LDA attempts to express one dependent variable as a linear combination of features or measurements.

Features, often represented as feature vectors, are individual measurable properties or characteristics of a phenomenon being observed. In classification, feature vectors can be found from nearest neighbor classification, neutral networks, and statistical techniques such as Bayesian approaches. For LDA, the data must follow a multivariate normal distribution which then assumes that each individual predictor follows a univariate normal distribution. There also must be some correlation between the predictors.

During imputation, the assumptions are that the variables are normally distributed with means that vary across the levels but the covariance matrix that are constant over the levels. LDA estimates the probability that a new set of imputes belong to every class. The new imputes are assigned to the class that has the largest probability. LDA uses Bayes' theorem to estimate the probabilities. We let π_k represent the prior probability that a randomly chosen observation comes from the kth class. In other words, π_k is the probability that a given observation is associated with the kth category of the response variable Y. The notation to indicate that a p-dimensional random variable X has a multivariate normal distribution is $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $E(X) = \boldsymbol{\mu}$ is the mean of the vector X with p components, and $Cov(X) = \boldsymbol{\Sigma}$ is the p x p covariance matrix. The multivariate Gaussian density is defined as $f(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}^{1/2}|} exp\Big(- \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\Big)$. When working with multiple predictor variables, the LDA classifier assumes that the observations in the kth class are drawn from a multivariate normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_k$ is a mean vector specific to each class, and Σ is the covariance matrix for all K classes. Bayes' theorem states that $p_k(x) = P(X = x|Y = k) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$. We refer

to $p_k(x)$ as the posterior probability, that is, the probability that the observation belongs to the kth class, given the predictor value for that observation. If we plug Bayes' theorem into the multivariate density function and we can find through some algebra that the Bayes' classifier assigns an observation $X = x$ to the class which $\delta_k(x) = \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k{}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + log\pi_k$ is the largest.

4.3 Multinomial Logit Model

When the categorical variables have more then two levels that are unordered, we can use the multinomial logit model for imputation. The multinomial logit model is utilized in many field such as in business a market researcher may wish to relate a customer's choice of a product (product A, product B, product C) to the customer's age, gender, geographic location, and several other potential explanatory variables. Medical fields also utilize the multinomial logit model. The multinomial logit model performs in a similar manner as logistic regression in the way that it uses probabilities to predict the responses for the missing data values. If the data is binary, then a single threshold is applied to divide the continuous distribution. If the variable contains multiple levels, thresholds are applied to divide the continuous distribution into the same number of sections as there are levels.

Consider the example of a study that was undertaken to determine the strength of the association between several risk factors and the duration of pregnancies. The explanatory variables are mother's age, nutritional status, history of tobacco use, and history of alcohol use. The response variable is pregnancy duration which has three levels: preterm, intermediate term, and full term. We will assume that there are J

response categories and i observations. The response variable is represented as

$$Y_{ij} = \begin{cases} 1 & \text{if case } i \text{ response is category } j \\ 0 & \text{otherwise} \end{cases}.$$

We know that since only one category can be selected as the response that $\sum_{j=1}^{J} Y_{ij} = 1$. Since we are dealing with multiple levels we need to have a probability for the probability that each category is selected for the ith observation. We say $\pi_{ij} = P(Y_{ij} = 1)$ is the probability that category J is selected for the ith response. In our example, $J = 3$; however, we do not compare three probabilities. We chose one category to be our baseline category and then make out predictions based off of that. It does not matter which category we chose to be the baseline, most researchers either chose the first category or the last category. For our purposes we will chose the Jth level for our baseline, we now make $J-1$ comparisons. The logit function for the jth comparison is $\pi'_{ij} = log_e \left[\frac{\pi_{ij}}{\pi_{iJ}}\right] = \boldsymbol{X}'_i\boldsymbol{\beta}_j$ where $j = 1, 2, ..., J-1$, $\boldsymbol{\beta}_j$ is the $J-1$ parameter vectors, and $\boldsymbol{X}'_i$ is the $J-1$ variable vectors. With the logit function we can find the direct expressions for the resulting probabilities, $\pi_{ij} = \frac{exp(\boldsymbol{X}'_i\boldsymbol{\beta}_j)}{1+\sum j=1^{J-1} exp(\boldsymbol{X}'_i\boldsymbol{\beta_j})}$. When making predictions for an observation, one would classify that observation into the category with the highest probability. Recall the pregnancy data, the explanatory variables are mother's age, nutritional status, history of tobacco use, and history of alcohol use, and the response variable is pregnancy duration which has three levels: preterm, intermediate term, and full term. If we were to classify a women into one of the three response categories, we would find the probability that she could place in either category. We assume the regression equation has been fitted with pregnancy

duration as the response variable and nutritional status, age, alcohol use history, and smoking history as the predictor variables. Let's say the probabilities for category 1, 2, and 3 were .31, .58, and .82, respectively. We would classify the woman into category 3 since the probability for that category was the highest.

There are two approaches for estimating $\boldsymbol{\beta}_j$, but both use maximum likelihood estimation. The first approach carries out separate binary logistic regression for each of the $J-1$ comparisons to the baseline category. For example, to estimate $\boldsymbol{\beta}_1$, we drop out all the cases except those of $Y_{i1}=1$ and $Y_{1J}=1$ from the data set. By only picking out the two types of cases we now have a logistic regression and can apply that approach directly. This approach is beneficial to those who do not have access to software that is capable of multicategory logistic regression. The more efficient approach is the estimate the $\boldsymbol{\beta}'_J s$ simultaneously. To do this we will use a likelihood function for the entire data set. For n independent observations and J categories the likelihood function is: $P(Y_1,...,Y_n)=\prod_{i=1}^{n} P(Y_i)=\prod_{i=1}^{n}[\prod_{j=1}^{J}(\pi_{ij})^{Y_{ij}}]$. Once the model is fitted, one can use interpretations and inference to gain more information about the data [24]. This means that one can interpret the coefficients in the regression equation and do hypothesis testing and use confidence intervals to see more relationships with the variables.

4.4 Ordered Logit Model

The ordered logistic or proportional odds model is used when the response variable consists of ordered levels with no assumed spacing. Some examples of ordinal variables are the following: (1) a food product is rated by customers on a 1-10 hedonic scale,

(2) the severity of cancer is rated by stages on a 1-4 basis, and (3) in an economic study, persons are classified as either not employed, employed part time, or employed full time. The ordered logit model is similar to multinomial logit model but it takes into account the ordering of the categories. Here we can use the same pregnancy duration data as we did in multinomial logit. We can adjust the response variable to change it form nominal to ordered. Now the response variable is set up in the following way:

$\boldsymbol{Y_i}$	**Category**	$\boldsymbol{Y_i^c}$	**Cutpoint** T
1	Preterm	$0 \leq Y_i^c < 36$ weeks	T_1= 36 weeks
2	Intermediate term	36 weeks $\leq Y_i^c < 38$ weeks	$T_2 = 38$ weeks
3	Full term	38 weeks $\leq Y_i^c < \infty$	$T_3 = \infty$

The response variable is now ordinal because we are now looking at certain pregnancy delivery time intervals in each category, which puts a natural ordering to the three categories.

Unlike the multinomial logit model, the ordered logit model models the cumulative probabilities $P(Y_i \leq j)$ rather than modeling for each category probabilities $P(Y_i = j)$. For $j = 1$, the cumulative probabilities can be expressed as: $P(Y_i \leq j) = P(\varepsilon_L \leq \alpha_1 + \beta_1 X_i)$, where ε_L follows a standard logit distribution with mean zero and standard deviation $\pi/\sqrt{3}$, X_i is a predictor variable, $\alpha_1 = (T_1 - \beta_0^*)/k$, $\beta_1 = -\beta_1^*/k$, and k is a constant that satisfies $\sigma\{Y_i^c\} = k\sigma\{\varepsilon_L\} = k\frac{\pi}{\sqrt{3}}$. We assume that X is linearly related to some appropriate log odds. The log odds here for the $J-1$ cumulative logit we have, $log_e[\frac{P(Y_i \leq j)}{1-P(Y_i \leq j)}] = \alpha_j + \boldsymbol{X_i'\beta}$ for $j = 1, ..., J-1$. With a multiple regression case with J ordered categories, we let $X_i = \begin{bmatrix} X_{i1} \\ X_{i2} \\ ... \\ X_{i,p-1} \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_{p-1} \end{bmatrix}$. The cumulative probability can be expressed by using the cumulative distribution

function as $P(Y_i \leq j) = \pi_{i1} = \frac{exp(\alpha_j + X_i{}'\beta)}{1+exp(\alpha_j + X_i{}'\beta)}$. The discrete variable is assigned a certain category value if the underlying normal variable X is above a given threshold and below the next threshold.

Similar to the previous methods the coefficients are estimated with maximum likelihood estimation. The coefficients can be interpreted as the change in the cumulative odds ratio for a unit change in the predictor. The interpretation for the ordinal logistic regression model is much easier than that for the nominal logistic regression model since only a single slope vector $\boldsymbol{\beta}$ is estimated [24].

5 DATA SOURCE

With the sport of cycling gaining popularity across the country, the number of cyclists on the road increases, and thus the chance of bicycle crashes increases. North Carolina Department of Transportation has been compiling data from all bicycle crashes from 2007 to 2014, which can be accessed freely at https://catalog.data.gov /dataset/north-carolina-bicycle-crash-data#sec-dates. The information that was collected for each crash includes: county, city, crash date, crash day, crash group, crash location, crash time, crash severity, bike age group, bike alcohol detected, bike direction, bike injury, bike position, bike race, bike sex, ambulance response, driver age group, driver estimated speed, speed limit, driver alcohol detected, driver injury, driver race, driver sex, driver vehicle type, hit and run, development, light condition, locality, number of lanes, road characteristics/class/condition/
configuration, road defects/features, traffic control, crash type, and/or weather.

For our purposes, we will use driver's age as the continuous response variable for each of the three models. The binary variables that we will consider in the regression model are the following: bike sex, work zone, bike alcohol, hit run, driver sex, driver alcohol, location of accident. The nominal variables are the following: bike race, bike position, driver vehicle type, driver race, road defects, road condition, development, crash server, driver injury, road character, region, bike direction, traffic control, road configuration, number of lanes, road surface, road feature, crash group, weather, light condition. The ordinal variables are the following: locality, bike age group, crash month, crash day, driver estimated speed, driver age group.

The first step of the regression analysis is to build our model from the data.

We are investigating the different imputation methods based on the different levels of categorical variables. Thus, we will fit three models, one for binary variables, nominal variables and ordinal variables. Each model is built in the same fashion. When building the model, we start with looking at a model including all variables. Using the variable driver's age as the response variable, we set up a linear model for each type of variable. Looking at the summary of each full model, we take out the variables one-by-one that are not significant. We say that a variable is not significant if the variable's individual p-value is greater than an alpha value of 0.05. We do this until we have a set of variables that all have significant p-values. Once we see that the set of variables are all significant, we check the global F-test for each model to test if, for each model, the set of predictors are useful in predicting driver's age. The p-values for the global F-test for the binary, nominal and ordinal variables are 0.01483, 0.008813, and $< 2.2e^{-16}$, respectively. Since each of the p-values are less than $\alpha = 0.05$, we can conclude that the set of predictors are useful in their respective models. Our final models, in terms of the variables kept are:

1) Binary: driver age = bike sex + location of accident

2) Nominal: driver age = driver race + crash severe + region + road surface

3) Ordinal: driver age= locality + driver estimated speed + driver age group.

The final fitted models are:

1) Binary:

$$\hat{Y} = 47.421 - 4.347X_1 - 3.449X_2$$

2) Nominal

$$\hat{Y} = 17.009 + 14.555X_1 + 13.214X_2 + 37.890X_3 + 21.465X_4 + 18.747X_5 + 4.298X_6$$
$$+ 0.383X_7 + 8.681X_8 + 7.328X_9 - 6.048X_{10} - 1.344X_{11} - 3.640X_{12}$$
$$+ 3.941X_{13} + 0.879X_{14} + 11.245X_{15} + 4.427X_{16}$$

3) Ordinal

$$\hat{Y} = 17.514 + 1.015X_1 + 0.297X_2 + 0.620X_3 + 0.584X_4 - 0.411X_5 + 0.152X_6$$
$$+ 0.562X_7 + 0.251X_8 - 0.023X_9 - 0.098X_{10} - 1.109X_{11} - 4.059X_{12}$$
$$+ 0.295X_{13} + 3.856X_{14} + 8.840X_{15} + 16.306X_{16} + 26.768X_{17}$$
$$+ 36.411X_{18} + 46.124X_{19} + 51.944X_{20}$$

After finding the models, we need to check the model assumptions. The first assumption we will check is that the regression between the response and the predictors specified in the model is appropriate. To do this, we will look at the residual plot between the residuals versus fitted values. From Figures 1, 2, and 3, we can see that there is random scatter, and so we can assume that the chosen regression between the response and the predictor specified in the model is appropriate.

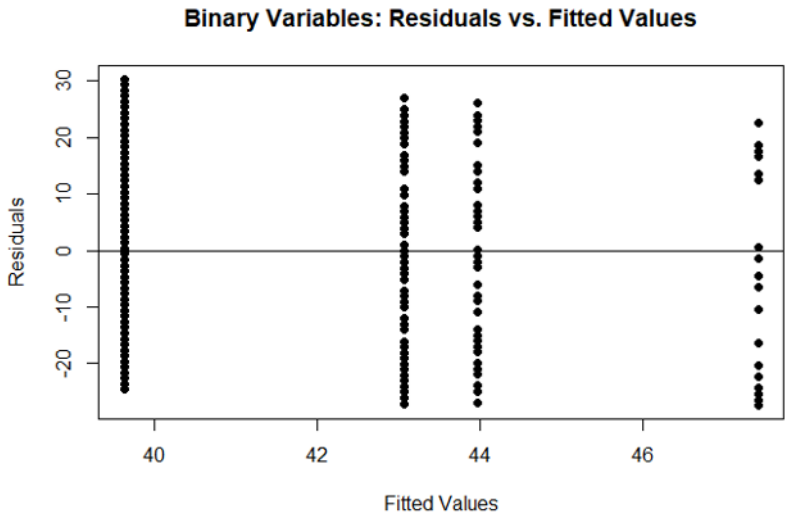

Figure 1: Residual plot between the binary residual versus the fitted values to check if the regression model is appropriate

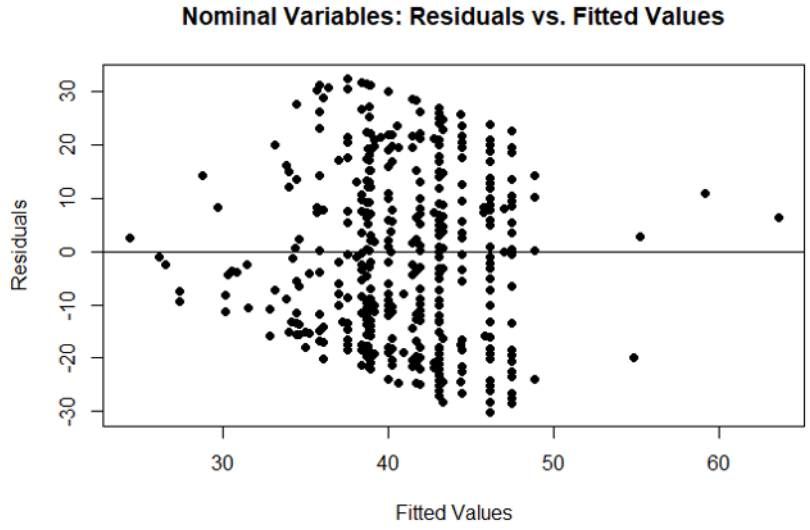

Figure 2: Residual plot between the nominal residual versus the fitted values to check if the regression model is appropriate

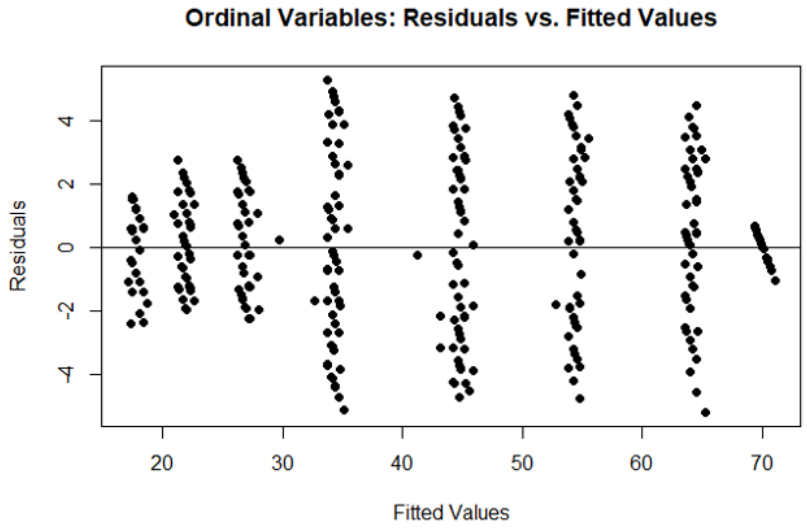

Figure 3: Residual plot between the ordinal residual versus the fitted values to check if the regression model is appropriate

By definition, the second assumption that the error terms have a mean of zero is satisfied. The assumption of constant variance is checked using the residuals versus fitted value plots. Figures 1, 2, and 3 show that there is random scatter indicating the assumption of constant variance is met. Since our sample size is large, we can assume the data is normally distributed by the central limit theorem, and thus this assumption is satisfied. The last assumption is the assumption that the errors are independent. This assumption is typically only a concern when dealing when the data is gathered over time. Since our data is gathered over time, from 2007 to 2014. We plot the residuals against the number of observations and look for random scatter. In Figures 4, 5, and 6 random scatter is present, thus this assumption is met.

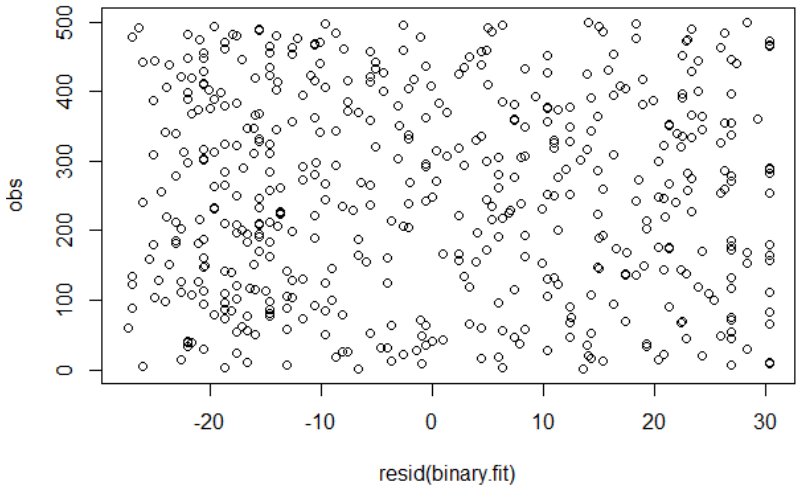

Figure 4: Time plot between residuals of the binary variables and the number of observation to check the for independent errors

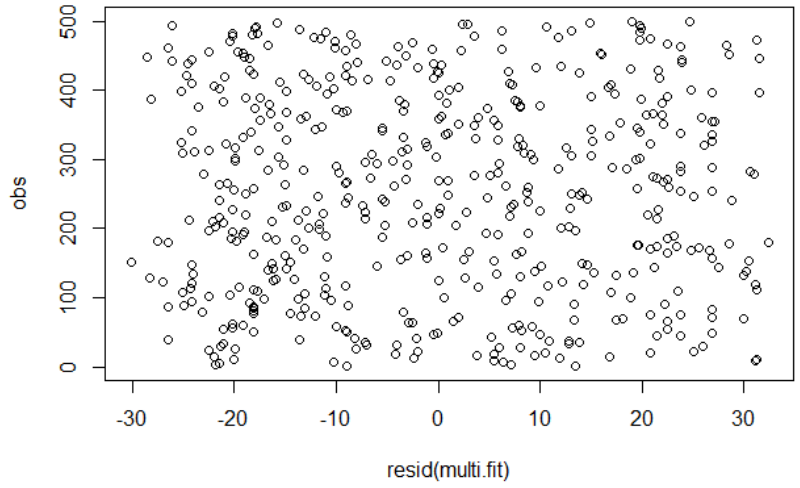

Figure 5: Time plot between residuals of the nominal variables and the number of observation to check the for independent errors

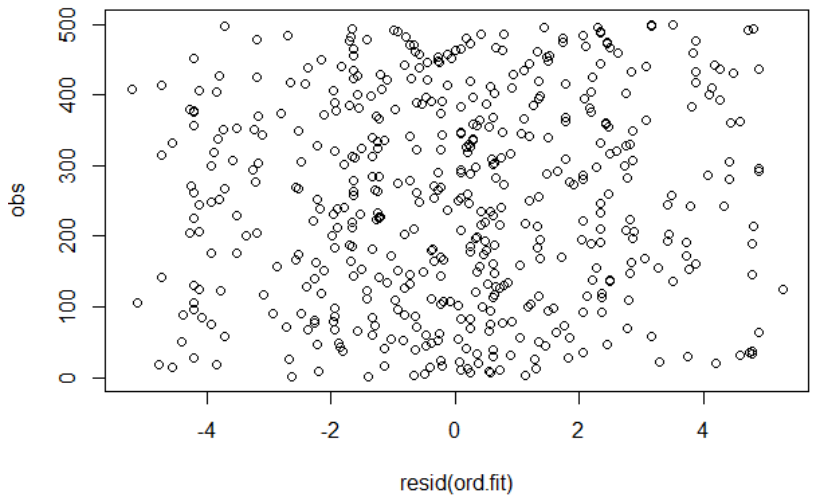

Figure 6: Time plot between residuals of the ordinal variables and the number of observation to check the for independent errors

6 METHODOLOGY

Our goal is to assess the best multiple imputation method for categorical variables. In this thesis we assume our missing data is missing at random. Since we are looking at three types of categorical variables (binary, nominal, and ordinal), we have three complete data sets. Each complete data sets consist of the categorical variables in each model. That is, the complete data set for the binary variables include bike sex, and location of accident. The complete data set for nominal variables include driver's race, crash severe, region, and road surface and the complete data set for ordinal variables include locality, driver's estimated speed, and driver's age group. For each complete data set, we created missing data sets with different levels of missingness. Using the R package, *missForest*, we remove 10%, 20%, 30%, 40%, and 50% of data randomly from each complete data set. That is done with the use of the R package, by first removing 10% out of the complete data set, then removing another 10%, and so forth until 50% of the data has been randomly removed. Thus, for each of the three complete data sets, there are a total of 5 data sets with missing values.

Each imputation method, is then applied to each of the missing data sets. For the binary data sets, logistic regression and LDA imputation methods were used to impute the missing values for each level of missingness. For the nominal data sets, the multinomial logit model and LDA imputation methods were used to impute the missing values for each level of missingness. Finally, or the ordinal data sets, the ordered logit model and LDA imputation methods were used to impute the missing values for each level of missingness. Each imputation method produced $m = 30$ iterations of the imputed data, i.e, for each imputation, 30 complete data sets were

found using that specific imputation method. For each of the 30 completed data (by imputation), the corresponding linear model (binary, nominal, or ordinal) was fit with the response variable as driver's age. In each case, the coefficients for the 30 models were recorded and then the means of the 30 coefficients for each variable were computed. To better estimate the true coefficient values, simulation performed. Each imputation method ($m = 30$) were ran for 1000 iterations. For each iteration, the mean of the 30 coefficients was stored. We then found the mean, for each variable, of the 1000 means of the coefficients. Lastly, we compared the means of the coefficients to each of the coefficients from the model of the complete data sets.

Along with comparing the coefficients, we look at the relative efficiency. The relative efficiency is used to find what the best procedure should be. That is, the procedure that produces the most accurate results. This uses the percent of missingness (10% - 50%) and the number of imputation ($m = 5$ up to $m = 50$) to find what combination will produce the most accurate imputed coefficients. The percent deviation index (PDI) is used to analyze the difference in the original coefficients and estimated imputed coefficients. We also analyze the difference of the original and imputed coefficients by implementing a simple t-test. We will be testing H_0: original coefficient = estimated mean coefficient, versus H_1: original coefficient $\neq$ estimated mean coefficient. We look at the relative frequency of each method with their respective variables. Relative frequencies show the proportion of each category that the imputation method is predicting. Lastly, we look at the success rate of each imputation method. This involves finding where the imputation method imputed the same level of the variable for each type of variable with each imputation method.

7 RESULTS

Relative efficiency is used to find the best procedure. That is, the procedure that will produce the most accurate results. It measures the difference in accuracy using various levels of missingness and number of imputations. The equation for the relative frequency is the following: $RE = \frac{1}{1+\frac{\lambda}{m}}$, where λ is the percent of missingness, and m is the number of imputations.

Table 1: Relative efficiency of the imputed models for various numbers of imputations at several amounts of missing data

Number of Imputations (m)	Percent of Missingness (λ)				
m	10%	20%	30%	40%	50%
5	0.980	0.962	0.943	0.926	0.909
10	0.990	0.980	0.971	0.962	0.952
15	0.993	0.987	0.980	0.974	0.968
20	0.995	0.990	0.985	0.980	0.976
25	0.996	0.992	0.988	0.984	0.980
30	0.997	0.993	0.990	0.987	0.984
50	0.998	0.996	0.994	0.992	0.990

The different levels of efficiency are shown in Table 1. Table 1 shows that at each number of imputations, the relative efficiency decreases as the percent of missingness increases. We see that the larger the amount of missing data, the less accurate the model will be for imputations. We can see that as the number of imputations increase, the relative efficiency increases, while holding the amount of the missing data constant. Overall, we see that when $m = 30$ the relative efficiency is 99% for levels of missingness.

7.1 Binary Variables

From the regression analysis discussed in chapter 5, the model was found to be

$$\hat{Y} = 47.421 - 4.347X_1 - 3.449X_2$$

where x_1 and X_2 represents the predictor variables. We will refer to these coefficients as the true coefficients since they are estimated from the complete data set. Table 2 displays the estimated coefficients for b_0, b_1, and b_2 based on the LDA imputation method. We see that for each percent missingness, the estimated coefficients for b_0 and b_2 are close to the coefficients from the complete data set but it is not the case for b_1. Note that as the percent of missingness increases, the estimated for b_0 gets further form the true coefficients. However, when the percent of missingness increases, the estimated coefficients b_2 gets closer to the true coefficient. This can also be seen from Table 3, which displays the PDI's for each coefficient at each different level of missingness. We see that the PDI's are roughly the same for b_0 and b_2 overall while they are extremely large for b_1.

Table 2: Estimated means of the binary variable's regression coefficients from the LDA imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	35.317	35.056	34.780	34.780	33.797	**47.421**
b_1	-0.856	-0.860	-0.320	-0.320	-0.703	**-4.347**
b_2	-4.612	-4.244	-4.360	-4.360	-3.788	**-3.499**

The estimations for b_2 are interesting. From 10% - 40% of missing data have a

Table 3: PDI values of LDA imputation model estimated regression coefficients at each level of data missingness for binary variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	25.5	26.1	26.6	26.6	28.7
b_1	80.3	80.2	92.6	92.6	83.2
b_2	31.8	21.3	24.6	24.6	8.3

consistent PDI ranging from 31.8% to 21.3%. This follows similarly to the estimation for b_0. However, at 50% of missing data where we see a PDI of 8.3%. A low PDI indicates that the estimated coefficient is relatively close to the original coefficient. The range, looking at Table 3, for the estimates for b_0 is the lowest of the three coefficients which indicates that the b_0 was the most consistent over the different amounts of missingness. Table 4 displays the p-values from the t-test which tests the following: H_0: estimated coefficient = original coefficient vs. H_a: estimated coefficient $\neq$ original coefficient. Since all of the p-values are less than $\alpha = 0.05$, we conclude that the estimated coefficients are not equal to the original coefficients.

Table 4: P-values for t-tests for each estimated regression coefficient for binary variables in the LDA imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	0
b_1	0	0	0	0	0
b_2	0	0	0	0	3.1180e-233

Table 5 shows the relative frequency for each category of the binary variables at each percent of missingness. The first two rows in the variable column represent the levels of gender of the biker: female and male while the last two rows represent the levels of location of accident: rural and urban. With 10% missingness, we can see that the proportions for each category is similar to the complete data proportions. We see that as the percent of missingness increases, the LDA imputation method predicted more males then females. At both 10% and 20% missingness, the imputed proportions are very similar to the completed data proportions. The imputed proportions for 30%, 40%, and 50% perform in a similar manner by overestimating for rural and underestimating urban. Overall, the differences for the imputed proportions are not very large.

Table 5: Relative frequency for the binary variables with LDA imputation at each level of missingness

Variable	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
Female	0.4305	0.4197	0.4245	0.4245	0.4169	**0.426**
Male	0.5695	0.5803	0.5755	0.5755	0.5831	**0.574**
Rural	0.2997	0.3032	0.3129	0.3129	0.3117	**0.3**
Urban	0.7003	0.6968	0.6871	0.6871	0.6883	**0.7**

Table 7 displays the estimated coefficients for b_0, b_1, and b_2 based on the logistic regression imputation method. One can see that the estimated coefficients are similar to those of LDA. For each percent of missingness, the estimated coefficients for b_0 and b_2 are close to the coefficients from the complete data set but it is not the case for b_1. Once again, we see that as the percent of missingness increases, the

estimated coefficients for b_0 gets further from the true coefficient. However, when the percent of missingness increases, the estimated coefficients for b_2 gets closer to the true coefficient. This can also be seen from Table 8, which displays the PDI's for each coefficient at each different level of missingness. We see the PDI's are roughly the same for b_0 and b_2 overall while they are extremely large for b_1.

Table 6: Success rate of classifications for the binary variables with LDA imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Driver's Sex	95.72	90.74	87.60	87.60	81.08
Location of Accident	95.30	91.92	87.82	87.82	81.73

To further explore the imputations we can look at the success rates in Table 6. Since the estimated coefficients for b_1 are so different from the true coefficients we would expect to see it's success rates to be low. For both divers sex and location of accident we see that as the level of missingness increases, the success rate decreases. Overall, the success rate for the LDA imputations for binary variables are high.

Table 7: Estimated means of the binary variable's regression coefficients from the logistic regression imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	35.310	35.044	34.773	34.104	33.795	**47.421**
b_1	-0.858	-0.850	-0.308	-0.065	-0.699	**-4.347**
b_2	-4.603	-4.238	-4.362	-3.715	-3.786	**-3.499**

The range of PDI's is the lowest at 3.2 for b_0 which indicates that the logistic regression is consistent when estimating b_0. The range of PDI's for b_2 is the largest at 25.4 indicating that logistic regression imputation was not consistent when estimating b_2. Overall the larger the amount of missing data, the larger the PDI. We see similar behaviour with b_2 with not only 50% missingness but also 40% missingness. The estimates for b_2 are the most accurate estimated coefficients. When looking at the p-values from the *t*-test in Table 9, we see that all of the estimated coefficients are significantly different than the original coefficients. This was the same conclusion we saw when using the LDA imputation method.

Table 8: PDI values of logistic regression imputation model estimated regression coefficients at each level of data missingness for binary variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	25.5	26.1	26.7	28.1	28.7
b_1	80.3	80.4	92.9	98.5	83.9
b_2	31.6	21.2	24.7	6.2	8.2

Table 10 contains the relative frequency for each binary variable for each level of missingness using the logistic regression imputation method. For the variable biker's sex, we can see that the most accurate imputed proportion occurs with 30% missingness. With 10% missingness, logistic regression imputation method over estimated females. At all other levels of missingness (20%, 40%, 50%), the logistic regression imputation method overestimated males and underestimated females. Looking at the second variable, location of accident, we see that at 10% and 20% missingness the imputed proportions are very similar to the original proportions. As the level

Table 9: P-values for t-tests for each estimated regression coefficient for binary variables in the logistic regression imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	0
b_1	0	0	0	0	0
b_2	0	0	0	1.4302e-191	1.5141e-234

of missing data increases, the logistic imputation method overestimates rural and underestimates urban.

Table 10: Relative frequency for the binary variables with logistic regression imputation at each level of missingness

Variable	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
Female	0.4306	0.4198	0.4247	0.4237	0.4172	**0.426**
Male	0.5694	0.5802	0.5753	0.5763	0.5828	**0.574**
Rural	0.2999	0.3035	0.3134	0.3162	0.3129	**0.3**
Urban	0.7001	0.6965	0.6866	0.6838	0.6871	**0.7**

Table 11 shows the success rate of the logistic imputations for the binary variables at each level of missingness. We see success rates that are similar to LDA imputation. For both variables we can see that as the level of missingness increases the success rate decreases. Overall, both binary variables have a high success rate.

Table 11: Success rate of classifications for the binary variables with logistic regression imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Driver's Sex	95.73	90.74	87.59	84.33	81.07
Location of Accident	95.28	91.91	87.80	84.47	81.71

7.2 Nominal Variables

From the regression analysis discussed in chapter 5, the model was found to be

$$\hat{Y} = 17.009 + 14.555X_1 + 13.214X_2 + 37.890X_3 + 21.465X_4 + 18.747X_5 + 4.298X_6 + 0.383X_7 + 8.681X_8 + 7.328X_9 - 6.048X_{10} - 1.344X_{11} - 3.640X_{12} + 3.941X_{13} + 0.879X_{14} + 11.245X_{15} + 4.427X_{16}$$

where the X_i's represent the indicator variables for the predictor variables. We will refer to these coefficients since they are estimated from the complete data set. Table 12 displays the estimated coefficients the nominal variables using the LDA imputation method. We can see that overall the estimated coefficients vary from the true coefficients.

Looking at Table 12 and 13, we can see that the estimated coefficients are the most accurate for b_0, b_1, and b_6. Table 13 shows us the PDI's for the nominal variables using the LDA imputation method. The coefficients for b_0 are the most similar, with a PDI of 8.1% at 10% missingness and reaches a maximum of 17.1% at 40% missingness. Interestingly, at 50% missingness, the PDI for b_0 is lowest at 0.3%. We also see the phenomena with b_9. Notice that the smallest PDI occurs when the levels

Table 12: Estimated means of the nominal variable's regression coefficients from the LDA imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	18.318	18.724	18.728	19.926	17.058	**17.009**
b_1	11.298	10.123	9.749	8.855	7.866	**14.555**
b_2	4.784	6.252	7.675	7.214	6.248	**13.214**
b_3	12.985	13.839	17.278	15.701	19.483	**37.890**
b_4	15.133	11.184	11.846	9.809	8.783	**21.465**
b_5	13.282	13.175	12.942	12.224	11.149	**18.747**
b_6	4.086	4.225	4.212	2.813	7.038	**4.298**
b_7	12.748	10.783	7.975	7.572	9.927	**0.383**
b_8	-3.402	-2.245	-2.726	-4.696	-0.517	**8.681**
b_9	2.223	2.229	2.274	1.635	5.999	**7.328**
b_{10}	-4.436	-5.083	-4.042	-2.445	-3.216	**-6.048**
b_{11}	-3.692	-3.276	-3.313	-2.738	-2.737	**-1.344**
b_{12}	2.189	3.036	4.204	5.250	6.102	**-3.640**
b_{13}	0.473	-1.348	-0.096	-4.562	-4.208	**3.941**
b_{14}	-1.700	-3.681	6.878	13.817	12.804	**0.879**
b_{15}	0.507	0.063	0.654	0.291	-0.082	**4.427**

of missingness is at 50%. Overall, the estimated regression coefficients were far from the true coefficient values.

Table 14 shows the p-values for the t-test for each variable. We see that the estimated coefficient for b_0 at 50% missingness is the only estimated coefficient that is not significantly different than the original coefficient. This is not surprising since the PDI value was 0.3% for that coefficient. All other variables at all levels of missingness are less than $\alpha = 0.05$.

Based on the estimated coefficients, PDI's and t-test, one would think that the LDA imputation method did not perform well. However, Table 15 may suggest oth-

Table 13: PDI values of the LDA imputation model estimated regression coefficients at each level of data missingness for nominal variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	8.1	10.1	10.1	17.1	0.3
b_1	22.4	30.5	33.0	39.2	46.0
b_2	63.8	52.7	41.9	45.4	52.7
b_3	65.7	63.5	54.4	58.6	48.6
b_4	29.5	47.9	44.8	54.3	59.1
b_5	29.2	29.7	31.0	34.8	40.5
b_6	4.9	1.7	2.0	34.6	63.8
b_7	3228.5	2715.4	1982.2	1877.0	2491.9
b_8	139.2	125.9	131.4	154.1	106.0
b_9	69.7	69.6	69.0	77.7	18.1
b_{10}	26.7	16.0	33.2	59.6	46.8
b_{11}	174.7	143.8	146.5	103.7	103.6
b_{12}	160.1	183.4	215.5	244.2	267.6
b_{13}	88.0	134.2	102.4	215.8	206.8
b_{14}	190.5	518.8	682.5	1471.9	1356.7
b_{15}	88.5	98.6	85.2	93.4	101.9

erwise. Table 15 shows the relative frequency table for each variable at each level of missingness. In Table 15, we see that the imputed proportions for Asian and Native Americans are the most inaccurate compared to the true proportions. Asians differ from the true by about 0.07 for each level of missingness. The LDA imputation method, underrepresented the Asian category when compared to the complete data set. Native Americans differ from the true proportions by about 0.002 across the levels of missingness. The LDA imputation method, overrepresented the Native American category when compared to the complete data set. Black, Hispanic, and White's imputed proportions are relatively consistent with the true proportions. In

Table 14: P-values for t-tests for each estimated regression coefficient for nominal variables in the LDA imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	**0.1528**
b_1	0	0	0	0	0
b_2	0	0	0	0	0
b_3	0	0	0	0	0
b_4	0	0	0	0	0
b_5	0	0	0	0	0
b_6	0	6.9166e-55	1.2986e-14	0	0
b_7	0	0	0	0	0
b_8	0	0	0	0	0
b_9	0	0	0	0	2.5963e-320
b_{10}	0	0	0	0	0
b_{11}	0	0	0	0	0
b_{12}	0	0	0	0	0
b_{13}	0	0	0	0	0
b_{14}	0	0	0	0	0
b_{15}	0	0	0	0	0

Table 15, we see that the imputed estimated proportions for the categories killed, no injury, and possible injury are relative consistent with the true proportions. The LDA imputation method underestimated the disabling injury category, but overestimated the evident injury category. The LDA imputation method produced estimated proportions similar to the true proportions for the three levels of Region. The imputed proportions for course asphalt and smooth asphalt are relatively consistent with the true proportions. The LDA imputation method overestimated the proportions for

the concrete, gravel, and grooved concrete categories. Since the Other category had such a low true proportion to start, the LDA imputation method did not classify any missing value as Other.

Table 15: Relative frequency for the nominal variables with LDA imputation at each level of missingness

Variable		Percent of Missingness					
		10%	20%	30%	40%	50%	**Complete**
Driver's Race	Asian	0.0093	0.0103	0.0115	0.0135	0.0148	**0.08**
	Black	0.2667	0.2609	0.2627	0.2581	0.2663	**0.284**
	Hispanic	0.0244	0.0228	0.0241	0.0236	0.0259	**0.024**
	Native American	0.0166	0.0191	0.0178	0.0189	0.0232	**0.008**
	Other	0.0154	0.0156	0.0181	0.0195	0.0220	**0.014**
	White	0.6676	0.6712	0.6657	0.6664	0.6480	**0.662**
Crash Severity	Disabling Injury	0.0388	0.0374	0.0406	0.0383	0.0388	**0.044**
	Evident Injury	0.4607	0.4580	0.4570	0.4626	0.4771	**0.444**
	Killed	0.0346	0.0328	0.0334	0.0351	0.0313	**0.036**
	No Injury	0.0653	0.0649	0.0687	0.0635	0.0517	**0.066**
	Possible Injury	0.4005	0.4069	0.4002	0.4005	0.4012	**0.410**
Region	Coastal	0.3347	0.3465	0.3290	0.3281	0.3371	**0.330**
	Mountains	0.1064	0.1013	0.1006	0.1102	0.1134	**0.104**
	Piedmont	0.5589	0.5522	0.5704	0.5616	0.5495	**0.566**
Road Surface	Course Asphalt	0.3253	0.3233	0.3180	0.3143	0.3149	**0.324**
	Concrete	0.0266	0.0275	0.0338	0.0348	0.0408	**0.024**
	Gravel	0.0050	0.0056	0.0067	0.0061	0.0068	**0.004**
	Grooved Concrete	0.0122	0.0113	0.0095	0.0096	0.0069	**0.006**
	Other	0	0	0	0	0	**0.006**
	Smooth Asphalt	0.6309	0.6323	0.6321	0.6351	0.6306	**0.640**

In Table 16, we see the first evidence in where LDA imputation differed from the complete data set. Overall, the variables driver's race, crash severity, and region have high success rates at each level of missingness. Road surface has low success rates.

Table 16: Success rate of classifications for the nominal variables with LDA imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Driver's Race	94.78	89.52	86.26	83.17	79.73
Crash Severity	93.65	87.93	82.14	77.80	72.53
Region	95.04	90.36	85.93	82.48	78.71
Road Surface	34.14	32.42	30.52	28.21	27.10

After some brief investigation we can see that LDA struggled with the levels course asphalt and smooth asphalt. LDA imputed course asphalt for smooth asphalt and vise versa. We would need further investigation of the imputed data sets to see if there are other levels that are misclassified.

The estimated coefficients for the the nominal variables imputed with the multinomial logit model are shown in Table 17. By looking at Table 17, it is easy to see that the estimated coefficients differ form the original coefficients. The estimated coefficients for b_0 are the most accurate. The estimated coefficients for b_7 are the most inaccurate. The estimated coefficients for b_8 and b_{13} are the negative but the original coefficients are positive while the opposite occurs with b_{12}. Multinomial logit imputation underestimates the coefficients for b_1, b_2, b_3, b_4, b_5, b_6, b_9, and b_{16}. For b_0, b_7, and b_{14} multinomial logit imputation overestimated the coefficients.

To see a more specific view of how the estimated coefficients differ from the original coefficients, we can look at the percent deviation index in Table 18. Overall, we see that as the level of missingness increases, the PDI values also increase. This implies that the multinomial logit imputation method performs worse as the level of

Table 17: Estimated means of the nominal variable's regression coefficients from the multinomial logit imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	19.101	21.122	22.341	25.215	25.364	**17.009**
b_1	11.206	9.100	8.458	7.336	6.605	**14.555**
b_2	5.027	5.681	5.629	5.797	5.725	**13.214**
b_3	15.063	13.746	13.334	11.632	11.323	**37.890**
b_4	14.954	10.503	10.129	8.216	7.513	**21.465**
b_5	13.229	12.219	11.674	10.600	9.817	**18.747**
b_6	3.263	2.645	1.586	-0.823	0.048	**4.298**
b_7	10.124	6.779	4.735	2.743	2.143	**0.383**
b_8	-4.148	-3.773	-4.893	-7.282	-5.287	**8.681**
b_9	1.566	0.747	0.053	-2.085	-1.058	**7.328**
b_{10}	-4.153	-4.697	-4.012	-2.941	-3.144	**-6.048**
b_{11}	-3.695	-3.216	-3.364	-2.729	-2.769	**-1.344**
b_{12}	2.095	2.901	2.996	3.355	2.753	**-3.640**
b_{13}	-2.873	-3.225	-2.309	-3.263	-2.950	**3.941**
b_{14}	4.494	4.877	6.857	12.311	10.204	**0.879**
b_{15}	0.455	0.042	0.547	0.246	-0.082	**4.427**

missingness increases.

As expected, we see the p-values for the t-test in Table 19, for testing the difference in the estimated coefficients to the true coefficients, are all significant at the 5% level of significance.

Table 20 shows the relative frequencies for the nominal variables imputed with multinomial logit imputation. For the diver's race, the Asian category's estimated proportions are lower than the true proportion. The estimated proportions for Native Americans are larger than the true proportions. The estimated proportions for Black, Hispanic, Other, and White categories are similar to the true proportions.

Table 18: PDI values of the multinomial logit imputation model estimated regression coefficients at each level of data missingness for nominal variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	12.3	24.2	31.3	48.2	49.2
b_1	23.0	37.5	41.9	49.6	54.6
b_2	62.0	57.0	57.4	56.1	56.7
b_3	60.2	63.7	64.8	69.3	70.1
b_4	30.3	51.1	52.8	61.7	65.0
b_5	29.4	34.8	37.7	43.5	47.6
b_6	24.1	38.5	63.1	119.1	98.9
b_7	2543.3	1670.0	1136.3	616.2	459.5
b_8	144.8	143.5	156.4	183.9	160.9
b_9	78.6	89.8	99.3	128.5	114.4
b_{10}	31.3	22.3	33.7	51.4	48.0
b_{11}	174.9	139.3	150.3	103.1	106.0
b_{12}	157.6	179.7	182.3	193.2	175.6
b_{13}	172.9	181.8	158.5	182.8	174.9
b_{14}	411.3	454.8	680.1	1300.6	1060.9
b_{15}	89.7	99.1	87.6	93.4	101.9

The variable crash severity's estimated proportions are overall similar to the true proportions for each level. However, we do see for the categories, disabling injury, no injury, and possible injury are underestimated and as the level of missingness while the categories evident injury and killed are overestimated. For the region variable, coastal stays consist in terms of the estimated proportions compared to the true proportions for all levels of missingness. The mountain regions have similar proportions for the lower levels of missingness but are slightly overestimated for higher levels of missingness. For the Piedmont region, the estimated proportions are similar to the true proportions for all levels of missingness except for the 50% level. Here we see

Table 19: P-values for t-tests for each estimated regression coefficient for nominal variables in the multinomial logit imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	0
b_1	0	0	0	0	0
b_2	0	0	0	0	0
b_3	0	0	0	0	0
b_4	0	0	0	0	0
b_5	0	0	0	0	0
b_6	0	0	0	0	0
b_7	0	0	0	0	0
b_8	0	0	0	0	0
b_9	0	0	0	0	0
b_{10}	0	0	0	0	0
b_{11}	0	0	0	0	0
b_{12}	0	0	0	0	0
b_{13}	0	0	0	0	0
b_{14}	0	0	0	0	0
b_{15}	0	0	0	0	0

a drop in proportion compared to the truth. For both course asphalt and smooth asphalt, we see similar estimated proportions compared to the truth for each level of missingness with smooth asphalt being slightly underestimated. As we saw with the LDA imputation method, the other category in road surface was estimated for a category for a missing value. This is probably due to the already small proportion (0.6%) of that category in road surface. For the other three categories of road surface (concrete, gravel, and grooved concrete), we see that they are overestimated using

the imputation methods.

Table 20: Relative frequency for the nominal variables with multinomial logit imputation at each level of missingness

Variable		Percent of Missingness					
		10%	20%	30%	40%	50%	**Complete**
Driver's Race	Asian	0.0093	0.0111	0.0128	0.0146	0.0161	**0.08**
	Black	0.2689	0.2653	0.2683	0.2637	0.2727	**0.284**
	Hispanic	0.0248	0.0240	0.0239	0.0233	0.0267	**0.024**
	Native American	0.0100	0.0108	0.0118	0.0129	0.0147	**0.008**
	Other	0.0157	0.0161	0.0187	0.0220	0.0243	**0.014**
	White	0.6713	0.6727	0.6645	0.6635	0.6456	**0.662**
Crash Severity	Disabling Injury	0.0405	0.0416	0.0435	0.0409	0.0423	**0.044**
	Evident Injury	0.4577	0.4520	0.4541	0.4618	0.4677	**0.444**
	Killed	0.0382	0.0386	0.0370	0.0390	0.0372	**0.036**
	No Injury	0.0653	0.0650	0.0694	0.0593	0.0572	**0.066**
	Possible Injury	0.3982	0.4028	0.3960	0.3990	0.3956	**0.410**
Region	Coastal	0.3346	0.3456	0.3289	0.3280	0.3398	**0.330**
	Mountains	0.1077	0.1043	0.1052	0.1140	0.1169	**0.104**
	Piedmont	0.5576	0.5501	0.5659	0.5579	0.5433	**0.566**
Road Surface	Course Asphalt	0.3273	0.3249	0.3209	0.3175	0.3181	**0.324**
	Concrete	0.0270	0.0272	0.0311	0.0332	0.0364	**0.024**
	Gravel	0.0047	0.0056	0.0066	0.0079	0.0088	**0.004**
	Grooved Concrete	0.0075	0.0082	0.0095	0.0084	0.0061	**0.006**
	Other	0	0	0	0	0	**0.006**
	Smooth Asphalt	0.6334	0.6341	0.6320	0.6331	0.6307	**0.640**

In Table 21, we can see the success rates for the nominal variables with multinomial logit imputation at each level of missingness. We can see the same pattern here as we did in the success rates of nominal variables with LDA imputation. The success rate for road surface is the lowest of all nominal variables. Again, with some brief investigation we can see that multinomial logit imputation also struggled to impute

Table 21: Success rate classifications for the nominal variables with multinomial logit imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Driver's Race	95.14	89.79	89.43	83.30	79.90
Crash Severity	93.56	87.59	81.95	77.73	71.99
Region	94.78	90.00	85.55	81.99	78.25
Road Surface	34.20	32.41	30.64	28.31	27.23

course asphalt and smooth asphalt correctly.

7.3 Ordinal Variables

From the regression analysis discussed in chapter 5, the model was found to be

$$\begin{aligned}\hat{Y} = 17.514 &+ 1.015X_1 + 0.297X_2 + 0.620X_3 + 0.584X_4 - 0.411X_5 + 0.152X_6 \\ &+ 0.562X_7 + 0.251X_8 - 0.023X_9 - 0.098X_{10} - 1.109X_{11} - 4.059X_{12} \\ &+ 0.295X_{13} + 3.856X_{14} + 8.840X_{15} + 16.306X_{16} + 26.768X_{17} \\ &+ 36.411X_{18} + 46.124X_{19} + 51.944X_{20}\end{aligned}$$

where the X_i's represent the indicator variables for the predictor variables. We will refer to these coefficients as the true coefficients since they are estimated from the complete data set. The estimated coefficients for the ordinal variables using LDA imputation are displayed in Table 22. Here we are estimating 20 coefficients with different levels of missingness. Overall the estimated coefficients differ greatly from the true coefficients. The estimated coefficients for b_2, b_3, b_4, b_6, b_{13}, b_{14}, b_{15}, b_{17}, and b_{19} are negative while their respective true coefficients are positive. The estimated

coefficients for b_{10}, b_{11}, and b_{12} are positive while their respective original coefficients are negative. LDA imputation estimated coefficients greater then the original coefficients for b_0, b_8, b_{10}, b_{11}, and b_{12}. The rest of the estimated coefficients are less then their respective original coefficients.

Table 22: Estimated means of the ordinal variable's regression coefficients from the LDA imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	33.611	34.457	33.921	34.715	38.025	**17.514**
b_1	1.354	-0.218	-0.998	-3.552	-5.465	**1.015**
b_2	-1.073	-1.198	-0.052	-0.811	-0.647	**0.297**
b_3	-0.572	-0.301	-1.206	0.267	-0.716	**0.620**
b_4	-2.583	-2.575	-2.845	-3.477	-1.570	**0.584**
b_5	-6.433	-5.898	-5.535	-4.706	-3.985	**-0.411**
b_6	-6.448	-4.775	-6.543	-6.550	-7.515	**0.152**
b_7	-4.548	-3.351	-2.708	-2.559	-4.507	**0.562**
b_8	1.672	2.862	4.261	4.980	5.600	**0.251**
b_9	-4.902	-3.503	-3.142	-2.644	-2.222	**-0.023**
b_{10}	6.137	6.451	4.758	4.328	4.127	**-0.098**
b_{11}	9.185	9.835	10.073	11.368	11.702	**-1.109**
b_{12}	12.884	13.929	15.938	16.246	16.232	**-4.059**
b_{13}	-1.472	-1.138	-2.820	-2.823	-3.008	**0.295**
b_{14}	-1.397	-2.356	-1.965	-1.626	-4.799	**3.856**
b_{15}	-3.254	-2.371	-2.489	-2.068	-5.351	**8.840**
b_{16}	1.654	0.057	-0.441	0.034	-3.684	**16.306**
b_{17}	-1.069	-2.551	-1.836	-2.359	-6.078	**26.768**
b_{18}	1.535	0.159	-0.548	-2.241	-5.714	**36.411**
b_{19}	-0.845	-2.353	-2.296	-2.588	-5.850	**46.124**
b_{20}	3.183	0.238	-0.947	-1.529	-2.982	**51.944**

To get a more detailed view of the estimated coefficients, we can look at the PDI for each coefficient at each level of missingness displayed in Table 23. The most accurate

estimated coefficient occurs with b_1 at 10% missingness with a PDI of 33.4%. The second most accurate estimated coefficient occurs with b_3 at 50% missingness with a PDI of 56.9%. Overall, the estimated coefficients are not accurate as we can see from the large PDI's displayed in Table 23.

Table 23: PDI values of the LDA imputation model estimated regression coefficients at each level of data missingness for ordinal variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	91.6	96.7	93.7	98.2	117.1
b_1	33.4	121.5	198.3	450.0	638.4
b_2	461.3	503.4	117.5	373.1	317.8
b_3	191.9	148.5	294.5	56.9	215.5
b_4	545.3	575.3	590.5	699.5	370.7
b_5	1465.2	589.8	1246.7	1245.0	869.6
b_6	4342.1	3241.4	4404.6	4409.2	5044.1
b_7	909.3	696.3	581.9	555.3	902.0
b_8	566.1	1040.2	1597.6	1884.1	2131.1
b_9	21213.0	15130.4	13560.9	11395.7	9560.9
b_{10}	6362.2	6682.7	4955.1	4516.3	4311.2
b_{11}	928.2	986.8	1008.3	1125.1	1155.2
b_{12}	417.4	443.2	492.7	500.2	499.9
b_{13}	599.0	485.8	1055.9	1056.9	1119.7
b_{14}	136.2	161.1	151.0	142.2	224.5
b_{15}	136.8	126.8	12 8.2	123.4	160.5
b_{16}	89.9	99.7	102.7	99.8	122.6
b_{17}	104.0	109.5	106.9	108.8	122.7
b_{18}	95.8	99.6	102.5	106.2	115.7
b_{19}	101.8	105.1	105.0	105.6	112.7
b_{20}	93.9	99.5	101.8	102.9	105.7

In Table 24, we can see the p-values from the t-test for the estimated coefficients for with LDA imputation. We are testing the if the there is a significant difference

Table 24: P-values for t-tests for each estimated regression coefficient for ordinal variables in the LDA imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	0
b_1	0	0	0	0	0
b_2	0	0	1.2881e-277	0	0
b_3	0	0	0	7.3864e-150	0
b_4	0	0	0	0	0
b_5	0	0	0	0	0
b_6	0	0	0	0	0
b_7	0	0	0	0	0
b_8	0	0	0	0	0
b_9	0	0	0	0	0
b_{10}	0	0	0	0	0
b_{11}	0	0	0	0	0
b_{12}	0	0	0	0	0
b_{13}	0	0	0	0	0
b_{14}	0	0	0	0	0
b_{15}	0	0	0	0	0
b_{16}	0	0	0	0	0
b_{17}	0	0	0	0	0
b_{18}	0	0	0	0	0
b_{19}	0	0	0	0	0
b_{20}	0	0	0	0	0

in the estimated coefficients and the true coefficients. We are comparing the p-values to an α value of 0.05. Since each of the p-values are less than $\alpha = 0.05$, we can conclude that each of the estimated coefficients are significantly different from the original coefficients.

To examine the LDA imputation, we can compare the proportion of each category estimated for each percent of missingness. The three ordinal variables we have are locality, driver estimated speed, and driver age group. The relative frequencies for each ordinal variable using LDA imputation are shown in Table 25. The variable locality has three categories: mixed, rural, urban. The estimated proportions for mixed and urban are similar to the true proportions with 10% and 20% missingness but with the larger amounts of missingness their estimated proportions become greater then the true proportions. The estimated proportions for rural is less then the true proportion and becomes increasingly lower as the level of missingness increases.

Looking at the drivers estimated speed, we can see that the majority of the accidents occur between 0-5 mph since the proportion for 0-5 mph is the largest. The estimated proportions for the drivers speed ranges of 6-10 mph, 11-15 mph, 21-25 mph, and 26-30 mph are relatively consistent with the true proportions. The estimated proportions for driver speeds 31-35 mph and 36-40 mph are close to the true proportions but as the percent of missingness increases their estimated proportions become lower than the true proportions. The estimated proportions for the driver speeds of 46-50 mph, 51-55 mph, and 56-60 mph are similar to the true proportions, but as the level of missingness increases the estimated proportions become greater than the true coefficients. For the final ordinal variable, driver age group, we can see that LDA imputation consistently underestimates the age groups 20-24 and 25-29. The age groups 30-39 is underestimated at 10% and 20% and then at 30% missingness and above they are overestimated.

In Table 26, we can see that, overall, the success rates for each ordinal variable is

Table 25: Relative frequency for the ordinal variables with LDA imputation at each level of missingness

Variable		Percent of Missingness					
		10%	20%	30%	40%	50%	**Complete**
Locality	Mixed	0.1641	0.1663	0.1792	0.1811	0.1777	**0.166**
	Rural	0.1563	0.1453	0.1288	0.1325	0.1345	**0.160**
	Urban	0.6796	0.6884	0.6920	0.6864	0.6879	**0.674**
Driver's Estimated Speed	0-5mph	0.2707	0.2703	0.2718	0.2562	0.2471	**0.278**
	6-10 mph	0.1162	0.1080	0.1131	0.1024	0.1061	**0.112**
	11-15 mph	0.0745	0.0777	0.0690	0.0684	0.0711	**0.076**
	16-20 mph	0.0615	0.0589	0.0676	0.0728	0.0700	**0.062**
	21-25 mph	0.0845	0.0865	0.0812	0.0889	0.0888	**0.084**
	26-30 mph	0.0637	0.0631	0.0612	0.0665	0.0661	**0.060**
	31-35 mph	0.1160	0.1161	0.1209	0.1205	0.1272	**0.116**
	36-40 mph	0.0407	0.0445	0.0340	0.0338	0.0326	**0.042**
	41-45 mph	0.0811	0.0764	0.0701	0.0707	0.0707	**0.088**
	46-50 mph	0.0195	0.0212	0.0201	0.0238	0.0270	**0.018**
	51-55 mph	0.0673	0.0729	0.0855	0.0903	0.0848	**0.060**
	56-60 mph	0.0043	0.0045	0.0055	0.0056	0.0085	**0.004**
Driver's Age Group	0-19	0.0858	0.0843	0.0862	0.0879	0.0902	**0.080**
	20-24	0.1455	0.1378	0.1426	0.1394	0.1366	**0.152**
	25-29	0.1126	0.1173	0.1160	0.1177	0.1235	**0.118**
	30-39	0.1385	0.1382	0.1435	0.1465	0.1508	**0.142**
	40-49	0.1615	0.1679	0.1629	0.1550	0.1390	**0.154**
	50-59	0.1502	0.1527	0.1482	0.1474	0.1435	**0.150**
	60-69	0.1181	0.1168	0.1141	0.1177	0.1155	**0.122**
	70+	0.0877	0.0850	0.0865	0.0883	0.1009	**0.082**

Table 26: Success rate classifications for the ordinal variables with LDA imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Locality	97.03	93.06	89.53	87.04	83.44
Driver's Estimated Speed	92.85	86.77	78.85	73.49	68.00
Driver's Age Group	89.89	81.11	74.24	67.12	61.50

high. As the level of missingness increases, the success rates decrease. Driver's age group has the lowest success rates of the ordinal variables, while locality and diver's estimated speed have relatively high success rates.

Ordered logit model imputation was also used to impute the missing data. The estimated coefficients for the ordinal variables with ordered logit model imputation can be found in Table 27. Similar to LDA imputation, the estimated coefficients differ from the true coefficients. The estimated coefficients for b_2, b_3, b_4, b_6, b_{13}, b_{14}, b_{15}, b_{17}, and b_{19} are negative while their respective original coefficients are positive. The estimated coefficients for b_{10}, b_{11}, and b_{12} are positive while their respective true coefficients are negative. The estimated coefficients are greater than the true coefficients for b_0, b_8, b_{11}, and b_{12}. The estimated coefficients are significantly less then the true coefficients for b_{15}, b_{16}, b_{17}, b_{18}, b_{19}, and b_{20}.

The PDI values for the ordered variables imputed with the ordered logit model imputation are shown in Table 28. The most accurate estimated coefficient occurs with b_1 at 30% missingness with the smallest PDI value of 19.8%. With such large PDI values it is simple to see that ordered logit imputation did not impute the missing data points well. The range of the PDI's for b_0 is the smallest at 6.5 showing that ordinal logit imputation produced the most consistent imputations over the different levels of missingness. It is not consistent across the variables that as the percent of missingness increases, the estimated coefficients are less accurate. With some of the variables the lowest PDI's occur with either 40% or 50% of missingness.

Table 29 displays the p-values from the t-test testing if the estimated coefficients and the originals coefficients differ. Similar to LDA imputation the p-values are

Table 27: Estimated means of the ordinal variable's regression coefficients from the ordered logit imputation model at each level of missingness

Coefficients	Percent of Missingness					
	10%	20%	30%	40%	50%	**Complete**
b_0	33.396	34.531	33.651	34.075	33.638	**17.514**
b_1	1.723	0.531	1.216	0.089	-0.731	**1.015**
b_2	-0.596	-1.162	-0.280	-1.071	-0.588	**0.297**
b_3	-0.596	-0.571	-1.317	0.076	-1.095	**0.620**
b_4	-2.554	-2.544	-2.823	-3.465	-1.985	**0.584**
b_5	-6.432	-5.827	-5.624	-4.897	-4.203	**-0.411**
b_6	-6.593	-4.737	-6.367	-6.352	-5.998	**0.152**
b_7	-4.534	-3.333	-2.757	-2.324	-2.012	**0.562**
b_8	1.414	3.072	3.944	3.865	6.390	**0.251**
b_9	-4.873	-3.423	-2.976	-2.387	-2.097	**-0.023**
b_{10}	5.776	5.937	2.944	2.608	2.500	**-0.098**
b_{11}	9.094	10.393	9.029	8.495	8.215	**-1.109**
b_{12}	12.375	12.454	12.373	11.714	10.894	**-4.059**
b_{13}	-1.466	-1.114	-2.796	-2.760	-2.887	**0.295**
b_{14}	-1.156	-1.815	-1.175	-0.476	-0.591	**3.856**
b_{15}	-3.107	-2.173	-1.955	-0.998	-0.382	**8.840**
b_{16}	1.781	0.459	-0.061	0.778	0.614	**16.306**
b_{17}	-0.917	-2.302	-1.400	-1.326	-1.036	**26.768**
b_{18}	1.726	0.488	-0.106	-1.197	-1.373	**36.411**
b_{19}	-0.678	-1.951	-1.724	-1.757	-1.550	**46.124**
b_{20}	3.433	0.904	-0.466	-1.585	-1.321	**51.944**

compared to $\alpha = 0.05$. Since each p-value is less then $\alpha = 0.05$, we conclude that the estimated coefficients are significantly different than the original coefficients.

The ordered logit relative frequency values are shown in Table 30. For the variable locality, we see that for the categories of rural and urban the estimated proportions are similar to the true proportions. The mixed category was underestimated with the ordinal logit imputation at the 10% and 20% levels of missingness. For 30%, 40%,

Table 28: PDI values of the ordered logit imputation model estimated regression coefficients at each level of data missingness for ordinal variables

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	90.7	97.2	92.1	94.6	92.1
b_1	69.8	47.7	19.8	91.2	157.9
b_2	230.7	490.6	194.3	460.6	298.0
b_3	196.1	192.1	312.4	87.4	276.6
b_4	537.3	535.6	583.4	693.3	429.9
b_5	1465.0	1317.8	1268.4	1091.5	922.6
b_6	4437.5	3216.4	4288.8	4278.9	4046.1
b_7	906.8	693.1	590.6	513.5	458.0
b_8	463.3	1123.9	1471.3	1439.8	2445.8
b_9	21087.0	14782.6	12839.1	10278.3	9017.4
b_{10}	5993.9	6158.2	3104.1	2761.2	2651.0
b_{11}	920.0	1037.2	919.8	866.0	840.8
b_{12}	404.9	406.8	404.8	388.6	368.4
b_{13}	596.9	477.6	1047.8	1035.6	1078.6
b_{14}	130.0	147.1	130.5	112.3	115.3
b_{15}	135.1	124.6	122.1	111.3	143.2
b_{16}	89.1	97.2	100.4	95.2	96.2
b_{17}	103.4	108.6	105.2	105.0	103.9
b_{18}	95.3	98.7	100.3	103.3	103.8
b_{19}	101.5	104.2	103.7	103.8	103.4
b_{20}	93.4	98.6	100.9	103.1	102.5

and 50% the levels of missingness, the mixed category is overestimated compared to the true proportion. The driver's estimated speed variable consists of various speed ranges from 0 mph to 70+ mph. Some of the speed ranges have estimated proportions that are similar to the true proportions. These categories include 0-5 mph, 6-10 mph, and 16-20 mph. The estimated proportions for speed ranges 11-15 mph, 31-35 mph, 36-40 mph, and 41-45 mph are lower than the true proportions. For the speed ranges

Table 29: P-values for t-tests for each estimated regression coefficient for ordinal variables in the ordered logit imputation model at each level of data missingness. The p-values that are in bold are for t-tests that are not significant at $\alpha = 0.05$ significance level

Coefficients	Percent of Missingness				
	10%	20%	30%	40%	50%
b_0	0	0	0	0	0
b_1	0	7.3271e-298	6.5080e-64	0	0
b_2	0	0	0	0	0
b_3	0	0	0	5.0194e-268	0
b_4	0	0	0	0	0
b_5	0	0	0	0	0
b_6	0	0	0	0	0
b_7	0	0	0	0	0
b_8	0	0	0	0	0
b_9	0	0	0	0	0
b_{10}	0	0	0	0	0
b_{11}	0	0	0	0	0
b_{12}	0	0	0	0	0
b_{13}	0	0	0	0	0
b_{14}	0	0	0	0	0
b_{15}	0	0	0	0	0
b_{16}	0	0	0	0	0
b_{17}	0	0	0	0	0
b_{18}	0	0	0	0	0
b_{19}	0	0	0	0	0
b_{20}	0	0	0	0	0

of 21-25 mph, 26-30 mph, 46-50 mph, 51-55 mph, and 56-60 mph their estimated proportions are greater than the true proportions. As the percent of missingness increases, the difference in the estimated and true proportions increases.

Looking at the final variable of driver age group, we can see that the groups 0-19,

Table 30: Relative frequency for the ordinal variables with ordinal imputation at each level of missingness

Variable		Percent of Missingness					
		10%	20%	30%	40%	50%	**Complete**
Locality	Mixed	0.1627	0.1636	0.1718	0.1757	0.1793	**0.166**
	Rural	0.1623	0.1544	0.1517	0.1544	0.1534	**0.160**
	Urban	0.6750	0.6819	0.6765	0.6699	0.6673	**0.674**
Driver's Estimated Speed	0-5mph	0.2710	0.2717	0.2758	0.2617	0.2622	**0.278**
	6-10 mph	0.1158	0.1088	0.1156	0.1028	0.1054	**0.112**
	11-15 mph	0.0744	0.0764	0.0695	0.0679	0.0678	**0.076**
	16-20 mph	0.0614	0.0593	0.0670	0.0683	0.0615	**0.062**
	21-25 mph	0.0857	0.0870	0.0835	0.0908	0.0908	**0.084**
	26-30 mph	0.0639	0.0642	0.0652	0.0680	0.0722	**0.060**
	31-35 mph	0.1167	0.1190	0.1247	0.1240	0.1330	**0.116**
	36-40 mph	0.0419	0.0432	0.0356	0.0382	0.0296	**0.042**
	41-45 mph	0.0814	0.0798	0.0733	0.0779	0.0803	**0.088**
	46-50 mph	0.0199	0.0222	0.0194	0.0216	0.0232	**0.018**
	51-55 mph	0.0634	0.0633	0.0645	0.0722	0.0669	**0.060**
	56-60 mph	0.0045	0.0052	0.0058	0.0066	0.0071	**0.004**
Driver's Age Group	0-19	0.0863	0.0839	0.0862	0.0845	0.0785	**0.080**
	20-24	0.1451	0.1374	0.1424	0.1427	0.1387	**0.152**
	25-29	0.1134	0.1174	0.1168	0.1190	0.1307	**0.118**
	30-39	0.1382	0.1386	0.1420	0.1503	0.1544	**0.142**
	40-49	0.1608	0.1670	0.1646	0.1568	0.1469	**0.154**
	50-59	0.1496	0.1514	0.1451	0.1420	0.1410	**0.150**
	60-69	0.1181	0.1161	0.1143	0.1182	0.1174	**0.122**
	70+	0.0885	0.0882	0.0886	0.0864	0.0924	**0.082**

25-29, and 60-69 have a similar relative frequency values compared to the complete data. The age groups of 20-24, 50-59, and 60-69 have a lower relative frequency with ordered logit imputation compared to the complete data. The age group 30-39 has an estimated proportion is lower than the true proportion for the 10% and 20% levels of missingness and at 40% and 50% levels of missingness the estimated proportion

is greater then the true proportion. One of the most accurate estimated proportions occurs in the age group 30-39 with 30% missingness. The estimated proportions for the age group 70+ is the greater than the true proportions.

Table 31: Success rate classifications for the ordinal variables with ordered logit imputation at each level of missingness

Variable	Percent of Missingness				
	10%	20%	30%	40%	50%
Locality	96.03	91.40	88.81	86.03	80.52
Driver's Estimated Speed	92.64	86.47	78.51	73.36	67.50
Driver's Age Group	89.79	81.06	74.13	67.04	60.96

In Table 31, we can see the same pattern as in the success rates for the ordinal variables with LDA imputation. Locality has the highest success rate of the ordinal variables with ordered logit imputation. Overall, as the level of missingness increases, the success rate of ordered logit imputation decreases.

8 CONCLUSION AND FUTURE RESEARCH

After examining the performance of each imputation method with their respective type of variables, we must conclude which imputation method performs the best for each type of categorical variable. Starting with binary variables, we compared LDA and logistic regression imputation. LDA and logistic imputation had a fairly similar performance. The relative frequencies for LDA and logistic regression imputation are also similar. The success rates for both LDA and logistic imputation were very similar. The software run time for each imputation is about the same for each. Since logistic regression imputation produced more estimated coefficients that are closer to the true coefficients, we would prefer to use logistic regression imputation for binary variables.

When dealing with the nominal variables, we compared LDA imputation and multinomial logit imputation. LDA imputation with the nominal variables produced the only significantly accurate estimated coefficient of the study. Multinomial logit imputation did not produce any accurate estimated coefficients. The relative frequencies for LDA is marginally better then those for multinomial logit. The success rates for both LDA and multinomial logit imputation were about the same. The software run time for LDA was significantly less than the run time for multinomial logit imputation. Due to the higher accuracy and lower run time we conclude that LDA imputation is best for nominal variables.

The last type of variable explored was ordinal variables. With the ordinal variables, we compared LDA imputation and ordered logit imputation. Neither LDA imputation nor ordered logit imputation performed particularly well. They did not

produce accurate estimated coefficients. The relative frequencies for LDA and ordered logit imputation were similar. The success rates for LDA and ordered logit imputation had no noticeable deviations from one another. The main difference between the two was the software run time. LDA took about a couple hours to run while ordered logit imputation took about 12 hours to run the loop of 1000 iterations. Due to the similar performance accuracy and the difference in run time, we would prefer to use LDA imputation with ordinal variables.

Overall, we see that in general logistic regression imputation performs best with categorical variables with two levels. For categorical variables with two or more levels both nominal and ordered LDA imputation outperforms the other imputation methods. Researchers, when dealing with machine learning and other branches of statistics, often choose to use LDA over the multinomial logit model in terms of regression analysis. This builds our confidence for choosing the LDA imputation method for nominal and ordinal variables.

Since completing this research, there are some areas and ideas that could use further research. It would be interesting to see how much of an effect there is when using categorical variables with various different amount of levels. For example, if we had a group of categorical variables with the same amount of levels or if we have a group of categorical variables with a different amount of levels. These variables would ideally contain more then two levels. I think this would be an interesting research idea because some of the variables in the data set used in this thesis had quite a few levels, some with up to 12. One could look into how many observations occurred at each level. Maybe if there is only a couple of observations in some of the levels the

estimated coefficients for those variables are not as accurate then if there was more observations in each level.

Looking at the relative frequency tables brought light to another area of future interest. After seeing that the estimated proportions for each variable were relatively similar, with mild deviation, to the true proportions, it would be interesting to investigate why the estimated coefficients were so far off from the true coefficients. The success rates gave us some insight with the nominal variable road surface. With further investigation of the imputed data sets we may find more misclassifications with the imputation methods.

BIBLIOGRAPHY

[1] *Missing data, Quantitative applications in the social sciences.*, Allison, P., 2001, Thousand Oaks, CA: Sage. Vol. 136.

[2] *Imputation of Categorical Variables with PROC MI* Allison, Paul D. Paper 113-30.

[3] *Review: A gentle introduction to imputation of missing values.,* Donders, A. R. T., van der Hijden, G. J. M. G., Stijnen, T., and Moons, K. G. M. (2006). Journal of Clinical Epidemiology, 10871091.

[4] *Applied missing data analysis*, Enders, C. K. (2010) New York: Guilford Publications.

[5] *Missing Data Analysis: Making It Work in the Real World*, Graham, J. W. (2009), Annual Review of Psychology, 60(1), 549-576.

[6] *Missing data: Our view of the state of the art*, Schafer, J. L.,and Graham, J. W. (2002), Psychological Methods, 7(2), 147177.

[7] *Categorical Data Analysis*, Agresti, A. (2002). Hoboken: John Wiley and Sons.

[8] *An introduction to modern missing data analysis*, Baraldi, A. N., and Enders, C. K. (2009), Journal of School Psychology, 48, 537.

[9] *Robustness of a Multivariate Normal Approximation for Imputation of Incomplete Binary Data*, Bernaards, C. A., T. R. Belin, and J. L. Schafer. (2007), Statistics in Medicine 26 (6): 136882.

[10] *Flexible Imputation of Missing Data*, Buuren, Stef van. Second ed., CRC Press, 2018.

[11] *Multiple imputation of discrete and continuous data by fully conditional specification*, Buuren, S. V. (2007), Statistical Methods in Medical Research, 16(3), 219242.

[12] *A Potential for Bias When Rounding in Multiple Imputation*, Horton, N. J., S. R. Lipsitz, and M. Parzen. (2003), The American Statistician 57 (4): 22932.

[13] *Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation*, Lee, K. J., and Carlin, J. B. (2010), . American Journal of Epidemiology, 171(5), 624632.

[14] Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, Rubin, D. B. (1974), Journal of Educational Psychology 66 (5): 688701. 1987b. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons.

[15] *Dealing With Missing Data*, Sainani, K. L. (2015), *Pm&r*, 7(9), 990994.

[16] *Multivariate Correlation Models with Discrete and Continuous Variables*, Olkin, I., and R. F. Tate. (1961), Annals of Mathematical Statistics 32 (2): 44865.

[17] *Modern Applied Statistics with S* Fourth ed. Venables, W. N., and Brian D. Ripley. Springer, 2002.

[18] *An Introduction to Statistical Learning with Applications in R*, James, Gareth, Witten, Hastie, Tibshirani, Springer, 2017.

[19] *An Empirical Comparison of Multiple Imputation Methods for Categorical Data*, Akande, Olanrewaju, Li, Fan, and Reiter, Jerome.

[20] *Mice: Multivariate Imputation by Chained Equations InR*, Buuren, Stef Van, and Karin Groothuis-Oudshoorn. Journal of Statistical Software, vol. 45, no. 3, 2011

[21] *MIMCA: Multiple Imputation for Categorical Variables with Multiple Correspondence Analysis*, Audigier, Vincent, Husson, Francois, and Josse, Julie.

[22] *Inference and Missing Data*, Rubin, D. (1976). Biometrika, 63(3), 581-592.

[23] *Missing data: the hidden problem*, SPSS White Paper

[24] *Applied Linear Statistical Models* Kutner, Michael H., Nachtsheim, Christopher J., Neter, John, Li, William. 2005 . McGrawHill Education.

VITA

SAMANTHA MIRANDA

Education: B.A. Mathematics, Brevard College,
Brevard, North Carolina 2018
M.S. Mathematical Sciences, East Tennessee State University
Johnson City, Tennessee 2020