2017

The Effect of a Missing at Random Missing Data Mechanism on a Single Layer Artificial Neural Network with a Sigmoidal Activation Function and the Use of Multiple Imputation as a Correction

Taron Dick

The Effect of a Missing at Random Missing Data Mechanism on a Single Layer Artificial Neural Network with a Sigmoidal Activation Function and the Use of Multiple Imputation as a Correction

Taron Dick

Thesis submitted
to the School of Public Health
at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in
Biostatistics

Michael Regier, Ph.D., Chair
Sijin Wen, Ph.D.
R. David Parker, Ph.D.

Department of Biostatistics

Morgantown, West Virginia
2017

Keywords: missing data, multiple imputation, neural networks

Abstract

The Effect of a Missing at Random Missing Data Mechanism on a Single Layer Artificial Neural Network with a Sigmoidal Activation Function and the Use of Multiple Imputation as a Correction

Taron Dick

Missing data is a common problem encountered in statistical analysis. However, little is known about how bias inducing missing at random missing data mechanisms affect predictive model performance measures such as sensitivity, specificity, error rate, ROC curves, and AUC. I investigate the effect of missing at random missing data mechanisms on a single layer artificial neural network with a sigmoidal activation function, equivalent to a binary logistic regression. Binary logistic regression is frequently used in health research and so it is a logical starting point to understand the effects of missing data on statistical learning models that could be used in health research. I then examine whether multiple imputation is a useful analytic correction for improving the predictive model performance measures relative to performing a complete case analysis.

Two simulation studies are conducted to understand how the complexity of the missing data mechanism, type of covariate missing, and rate of missing values affect the measures of interest and whether multiple imputation is robust to the various scenarios investigated. It was found that sensitivity, specificity, and error rate estimates were biased for all scenarios and the magnitude of bias increased as the missing rate increased. However, the AUC remained unbiased. Multiple imputation was observed to be an effective correction for missing values by decreasing the bias of the performance measures relative to the complete case analysis.

I conclude that missing at random missing data mechanisms do affect performance measures such as sensitivity, specificity, and error rate estimates, but multiple imputation is a useful analytic correction for reducing the bias of these measures. It is advised that caution should be taken when reporting AUC and it should be reported alongside other measures such as sensitivity and specificity.

Acknowledgements

I would first like to thank my thesis advisor, Dr. Michael Regier, for introducing me to statistical learning and missing data topics. I am appreciative of the time he has taken to mentor me in both topics as well as his guidance in successfully merging these topics together to create a meaningful study. I am grateful for his continuous support and endless insights that have influenced my statistical and writing abilities.

I would also like to thank my other committee members– Dr. Sijin Wen for his valuable insights about simulation studies and Dr. David Parker for his valuable insights on strengthening my written and verbal communication.

Finally, I would like to thank my wife Brittany. It would not have been possible to finish this thesis without her patience, understanding, and constant support and encouragement she has given me throughout this process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical learning techniques are increasingly popular for understanding data and can therefore be useful when analyzing various data sources such as EMR data. Depending upon the research question, this may be through finding associations between variables or building models to make predictions for future observations. These techniques may also be used for exploring data to generate hypotheses for new experiments, examine trends, or identify clusters of related patients within the data.

Statistical learning can be classified into two categories – unsupervised and supervised [17]. Unsupervised learning refers to finding relationships between different input variables (eg. predictors) with no corresponding output variable (eg. response) such as through clustering similar observations together. Supervised learning refers to building a statistical model for predicting an output (response) Y based on p different inputs (predictors) $X_1, X_2, ..., X_p$. This is accomplished by estimating a function $f(X)$ that maps the values of the input variables X to a value of the output variable Y. The focus of this thesis is on supervised learning.

For supervised learning, suppose a dataset has p independent variables $X_1, X_2, ..., X_p$ and a dependent variable Y. The relationship between $\mathbf{X} = (X_1, X_2, ..., X_p)$ and Y can be written as

$$\mathbf{Y} = f(\mathbf{X}) + \epsilon \tag{1.1}$$

where $f(\mathbf{X})$ is some function of $X_1, X_2, ..., X_p$ and ϵ is a random error term

independent of X. The function $f(X)$ may be estimated for inference or prediction. I will focus on prediction only so there are p predictors $X_1, X_2, ..., X_p$ and the response Y can be predicted by

$$\hat{Y} = \hat{f}(X) \tag{1.2}$$

since it is assumed the error term averages to zero (eg. E[$\epsilon \mid X$] = 0) [17].

An important measurement to take into account is the accuracy of $\hat{Y}$ because $\hat{f}(x)$ is not a perfect estimate of f. This is because Y is a function of an error term ϵ (Equation 1.1) which cannot be predicted using $\mathbf{X}$ and represents a source of irreducible error. Another type of error, reducible error, can also play a role, but this type of error can be eliminated by using appropriate statistical techniques. Therefore, the goal is to estimate f while minimizing the reducible error in order to get the most accurate prediction possible.

While EMR data may have many missing values, little is known about how this missing data affects the predictive ability of these techniques. One challenge with EMR data is not knowing whether information is missing due to lack of documentation (eg., a value is not recorded because it is not perceived to be relevant) or if it is missing due to lack of collection (eg., a test is not performed). Missing data has the potential to bias results and lead to incorrect conclusions [5].

This thesis examines the impact of missing at random missing data mechanisms on neural network prediction. The focus is on a single layer neural network with a sigmoidal activation function, which is the equivalent of a binary logistic regression. The utility of multiple imputation (MI) as a correction is also assessed. The main hypothesis is that multiple imputation will correct the prediction accuracy from machine learning classifications models relative to performing a complete case analysis, measured by sensitivity, specificity, receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC).

A simulation study is conducted to examine how missing data affects prediction and if multiple imputation improves the performance of predictive models. If the results show that a complete case analysis is sufficient, then

researchers may not have to worry about implementing more complex techniques. However, if the results show a complete case analysis is insufficient, then researchers should consider using MI techniques because many software programs already have this capability built in (eg., jomo [18], mice [25], and norm [16] in R [19]; PROC MI in SAS[24]).

This thesis is organized as follows: Chapter 1 finishes with a literature review; Chapter 2 describes common methods for handling missing data, different types of missing data mechanisms, the theoretical effect missing data has on a complete case analysis, and the process of multiple imputation; Chapter 3 describes the single layer neural network with a sigmoidal activation function equivalent, the logistic regression model, and how to assess predictive model performance; Chapter 4 describes the simulation study design and results; Chapter 5 provides a discussion of the results as well as implications for future research.

1.1 Literature Review

Prior research examines the effect of missing data mechanisms on logistic regression complete case analysis with Bartlett, Harel, and Carpenter [2] showing a logistic regression complete case analysis can provide unbiased estimates of the exposure odds ratio expressed through the model coefficients β under a wide range of missing data mechanisms. They found that 1) when missingness depends on the outcome only, the intercept β_0 is biased while β_x, the coefficients of the regressors, are unbiased; 2) when missingness depends on the covariates only, no β are biased; 3) when the missingness depends on both the outcome and the covariates, both β_0 and β_x are biased. However, these results are not known to be extended to a prediction setting nor is the utility of multiple imputation as a correction within that setting known.

However, other research does examine the utility of multiple imputation (MI) as a correction for missing data in logistic regression models. Numerous clinical studies have shown that MI produces less biased estimates than a complete case analysis. Choi, Nam, and Kwak [6] discuss how varying rates of missing data (10-50$\%$) in a clinical dataset affect logistic regression coefficient esti-

mates. It was found that the bias of parameter estimates when performing a complete case analysis increases as the rate of missing data increased but multiple imputation as a correction can reduce the bias of the parameter estimates. It is further shown that multiple imputation reduces the bias more than single imputation. However, this study does not examine the effect of missing data in the prediction setting.

Hallgren et al. [11] used logistic regression to estimate the effect of a drug on heavy drinking outcomes. They performed a complete case analysis and compared it with results from imputation techniques such as last observation carried forward, worst case scenario, and multiple imputation. They found MI yielded the least biased estimates and suggested this method to correct for missing data when analyzing binary outcomes for alcohol clinical trials.

van der Heijden et al. [37] investigated handling missing data via complete case analysis, missing indicator method, single imputation, and multiple imputation finding that complete case analysis and missing indicator methods should be avoided in multivariate diagnostic research. They did not find MI to be more effective than single imputation due to the low number of missing values but acknowledge that MI is often superior based on previous research.

Bounthavong, Watanabe, and Sullivan [4] examined MI use for correcting missingness in EMR data assuming values for covariates were missing at random (MAR). Unlike the previous studies, MI did not alter the results compared to performing a complete case analysis despite having missing data for 22% of the subjects. Furthermore, this study does not examine how missingness in electronic health records data affects predictive performance.

Some studies have examined missingness in a prediction setting but have not assessed the effect on sensitivity, specificity, ROC curves, and AUC through simulation studies with mechanisms of varying complexity, different types of missing covariates, and various missing rates nor has the utility of multiple imputation across various scenarios been examined. For example, Peng, Lei, and Naijun [24] found that the prediction accuracy is affected when >20% of data is missing but do not examine the utility of MI as a correction. Williams et al. [38] examined incomplete data classification for logistic regression but use

an estimated conditional density function and claim it is better than standard imputation techniques. Baneshi and Talei [1] examined 4 categorical covariates and used Cox regression models to compare how different imputation methods affected modeling of breast cancer specific death. They found that the multiple imputation by chained equations (MICE) model produced the highest sensitivity and specificity over median, regression, and EM imputation. Finally, Masconi et al. [21] examined imputation techniques for missing data in undiagnosed diabetes risk prediction using a specific subset of study patients and found that deletion methods resulted in the lowest concordance statistic while single imputation yielded similar results as multiple imputation.

Chapter 2

Missing Data

2.1 Common Methods for Handling Missing Data

There are several common methods for handling missing data including deletion methods and single imputation. The two following sections provide a brief summary of deletion methods and single imputation techniques to emphasize reasons for MI consideration. Missing data mechanisms are then discussed before introducing the multiple imputation procedure in greater detail.

2.1.1 Deletion Methods

Deletion refers to removing subjects with missing data from the analysis data set. One deletion method is listwise deletion or *complete case analysis* in which any subject with a missing value in any variable of interest is removed from the dataset and excluded from all analyses [7, 39]. The advantage is that it is simple and the same data set is used across all analyses. However, a disadvantage is a reduced sample size and statistical power. Furthermore, there is a loss of information from the other variables that were not missing for the subject. The impact of listwise deletion depends on why data is missing. If data are not missing completely at random (MCAR) as defined in section 2.2.1, a complete case analysis could result in biased estimators [20]. This is important because complete case analysis is the most popular method of handling missing data and often is the default option in statistical software packages. Therefore, if a data analyst ignores the missing data mechanism

and blindly runs a complete case analysis, it is possible that the results will be biased.

Another deletion method is pairwise deletion [7,39]. Pairwise deletion is also known as *available case analysis* because with this method an observation or subject is deleted when it is missing a variable required for a particular analysis. However, this observation may be included in another analysis when all required variables for that particular analysis are present. The advantage of available case analysis over complete case analysis is that each analysis will have as many cases as possible. It attempts to maximize the available information to be used for each separate analysis. However, the disadvantage of available case analysis is that a different subsample is used for each analysis so analyses may not be comparable.

2.1.2 Single Imputation

Rather than removing observations with missing data, single imputation is an alternative method for dealing with missing data by replacing a missing value with a well chosen value. For example, mean imputation replaces a missing value with the mean of the nonmissing values for that variable [7,28]. This produces a complete dataset for analysis but ignores the relationship between variables so the covariance and correlation estimates in the data are underestimated. Furthermore, the distribution of the mean is distorted so the underestimated standard errors can lead to incorrect inferences [13,33].

Regression mean imputation uses the complete cases to estimate a regression equation [7,10,39]. For each variable with missing values, the complete cases are regressed on the other variables in the data set. Now the missing value for an incomplete case can be predicted by using the non-missing information for that incomplete case. The advantage is using information from other observed data so estimates of means will vary rather than the same value being repeated as in mean imputation. However, this method does not account for the variability surrounding the predicted mean so standard errors are still underestimated [8]. A form of single imputation for longitudinal studies is last observation carried forward where missing values are replaced by the previous value observed, but this type of imputation also introduces error [7,8,39].

2.2 Missing Data Mechanisms

The notation used for missing data follows that of Carpenter and Kenward [5]. Suppose there is a sample of n units (subjects) from a population that is used to make inferences about a set of p population parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$. Let $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2}, ..., Y_{i,p})^T$ represent the p variables collected on subject i, $i = 1, ..., n$. The subset of p variables observed for each subject i is denoted $\mathbf{Y}_{i,O}$ and the subset of p variables missing for each subject i is denoted $\mathbf{Y}_{i,M}$, hence $\mathbf{Y}_i = (\mathbf{Y}_{i,O}, \mathbf{Y}_{i,M})$. Let $\mathbf{R}_i = (R_{i,1}, ..., R_{i,p})^T$ be an indicator for item response such that, for each individual $i = 1, ..., n$ and variable $j = 1, ..., p$, $R_{i,j} = 1$ if $Y_{i,j}$ is observed and $R_{i,j} = 0$ if $Y_{i,j}$ is missing.

The missing data mechanism is then defined as the probability of observing data for subject i given the values of $\mathbf{Y}_i$,

$$\Pr(\mathbf{R}_i \mid \mathbf{Y}_i). \tag{2.1}$$

This probability statement differs depending on the cause of the missing data. Rubin (1976) defines three common missing data mechanisms: missing completely at random, missing at random, and missing not at random [20].

2.2.1 Missing Completely at Random

Data are missing completely at random (MCAR) if the probability of missingness does not depend on any observed or unobserved data for that subject,

$$\Pr(\mathbf{R}_i \mid \mathbf{Y}_i) = \Pr(\mathbf{R}_i). \tag{2.2}$$

When data are MCAR, the observed data are a representative subset of the population so a complete case analysis will not bias estimates. However, information has still been lost; there is a loss of efficiency and standard errors of estimates will be larger [5,15,28].

MCAR is a strong assumption that cannot always be validated since there is often no obvious way of determining whether the probability of observing a variable depends on the value of the unobserved variable. Despite the exis-

tence of an MCAR test [19], this remains a challenge as Schafer showed some mechanisms may be untestable [34].

2.2.2 Missing at Random

It is often more reasonable to assume data are missing at random where the missingness can be accounted for by a subset of variables where there is complete information [5,15,28]. That is, given the observed data, $\mathbf{Y}_{i,O}$, the probability of missingness does not depend on the unobserved data, $\mathbf{Y}_{i,M}$, written as

$$\Pr(\mathbf{R}_i \mid \mathbf{Y}_i) = \Pr(\mathbf{R}_i \mid \mathbf{Y}_{i,O}). \tag{2.3}$$

For example, let $Y_{i,1}$ be a continuous response of total cholesterol and $Y_{i,2}$ be a covariate of sex that is always observed such that $R_{i,2} = 1$ for all i subjects. Now assume males have higher total cholesterol than females and males are less likely to report their total cholesterol. However, within males and females the probability of observing the total cholesterol does not depend on the value of total cholesterol. Thus, within categories of sex, the total cholesterol is MCAR. This implies that total cholesterol is MAR dependent on sex and the probability of observing the total cholesterol given sex is expressed as

$$\Pr(R_{i,1} = 1 \mid Y_{i,1}, Y_{i,2}) = \Pr(R_{i,1} = 1 \mid Y_{i,2}). \tag{2.4}$$

Now this can be rearranged to find the distribution of total cholesterol given sex can be expressed as

$$\begin{aligned} \Pr(Y_{i,1} \mid Y_{i,2}, R_{i,1} = 1) &= \frac{\Pr(Y_{i,1}, Y_{i,2}, R_{i,1} = 1)}{\Pr(Y_{i,2}, R_{i,1} = 1)} \\ &= \frac{\Pr(R_{i,1} = 1 \mid Y_{i,1}, Y_{i,2})\Pr(Y_{i,1}, Y_{i,2})}{\Pr(R_{i,1} = 1 \mid Y_{i,2})\Pr(Y_{i,2})} \\ &= \Pr(Y_{i,1} \mid Y_{i,2}) \end{aligned} \tag{2.5}$$

by using the definition of conditional probability, $\Pr(B \mid A) = \frac{\Pr(A|B)Pr(B)}{\Pr(A)}$. Thus, 2.5 verifies the distribution of total cholesterol within sex categories is MCAR since it is the same in the population, observed data, and unobserved data.

The MAR mechanism assumption is untestable [34], thus we should have reasonable justification before using multiple imputation as a correction. First,

we need the conditional distributions of partially observed variables given fully observed variables to be the same in subjects who have data observed and subjects who do not have data observed. Second, we need to be able to piece together the marginal distributions of the observed patterns to estimate the joint distribution of the data [5]. If these two conditions are satisfied then it is reasonable to proceed under the assumption that these data are MAR.

2.2.3 Missing Not at Random

The final missing data mechanism is missing not at random where the missingness is related to the values of the unobserved data [5]. That is, the probability of a missing value depends on the underlying value. Thus, unlike MAR, the dependence still remains given the observed data, such that

$$\Pr(\mathbf{R}_i \mid \mathbf{Y}_i) \neq \Pr(\mathbf{R}_i \mid \mathbf{Y}_{i,O}). \tag{2.6}$$

Carpenter et. al [5] indicate that multiple imputation is further complicated since we must explicitly specify the joint distribution of subject i's variables and the response indicator for observing those variables, $\Pr(\mathbf{R}_i, \mathbf{Y}_i)$, as either a *selection model*

$$\Pr(\mathbf{R}_i \mid \mathbf{Y}_i)\Pr(\mathbf{Y}_i) \tag{2.7}$$

or a *pattern mixture model*

$$\Pr(\mathbf{Y}_i \mid \mathbf{R}_i)\Pr(\mathbf{R}_i). \tag{2.8}$$

As 2.7 shows, a selection model specifies the marginal distribution of $\mathbf{Y}_i$ and the conditional distribution of $\mathbf{R}_i$ given $\mathbf{Y}_i$. On the other hand, 2.8 shows a pattern mixture model specifies the marginal distribution of $\mathbf{R}_i$ and the conditional distribution of $\mathbf{Y}_i$ given $\mathbf{R}_i$ [5]. It is important to remember that the MCAR, MAR, and MNAR assumptions are made for specific analyses and not a characteristic of the dataset itself [5].

2.3 Consequences of Missing Data Mechanisms for Regression Analyses

I present the effects of missing data mechanisms on parameter estimates in terms of bias and loss of information for three different situations: missing

response only, missing covariates only, or missing both response and covariates. If a complete case analysis produces valid estimates then the missing data mechanism is considered ignorable and does not need to be included in the model. Situations where estimates are biased require the mechanism to be specified and it is considered nonignorable. I will present these observations within the context of linear regression, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

2.3.1 Partially Observed Response

Suppose there is a fully observed variable X but the response Y is only partially observed. The contribution from unit i to the likelihood for the parameters, $\boldsymbol{\beta}$, of a linear regression model is [5]

$$L_i(\boldsymbol{\beta}, \theta \mid Y_i, X_i, R_i) = \Pr(R_i, Y_i \mid X_i, \theta) = \Pr(R_i \mid Y_i, X_i, \theta)\Pr(Y_i \mid X_i, \boldsymbol{\beta}). \tag{2.9}$$

The first term on the right hand side of 2.9 is the missing data mechanism model and contains information about parameters θ while the second term is the outcome model containing information about $\boldsymbol{\beta}$.

If Y_i is MAR given X_i then the contribution for an individual with a missing response is

$$\begin{aligned} &\int_{Y_M} \prod_{i=1}^{n} \Pr(R_i \mid Y_{i,O}, X_i, \theta)\Pr(Y_i \mid X_i, \boldsymbol{\beta}) dY_{i,M} \\ &= \prod_{i=1}^{n} \Pr(R_i \mid Y_{i,O}, X_i, \theta) \int_{Y_M} \Pr(Y_i \mid X_i, \boldsymbol{\beta}) dY_{i,M} \\ &= \prod_{i=1}^{n} \Pr(R_i \mid Y_{i,O}, X_i, \theta). \end{aligned} \tag{2.10}$$

Thus, units with missing response only contribute 1 to the likelihood and do not affect the MLE of $\boldsymbol{\beta}$. This is because when integrating 2.9 the first term becomes a constant in the log-likelihood function when optimizing for β since it only includes information about θ and the second term integrated over the range of all possible values for Y_i given X_i is 1 when data is missing. Therefore, performing a complete case analysis when a response is MAR given a fully observed covariate will not bias parameter estimates and is considered valid [5].

However, this is not the case when the response Y_i in MNAR. If Y_i is MNAR then the contribution from unit i is

$$\int \prod_{i=1}^{n} \Pr(R_i \mid Y_{i,O}, Y_{i,M}, X_i, \theta) \Pr(Y_i \mid X_i, \boldsymbol{\beta}) dY_{i,M}. \tag{2.11}$$

In this instance the likelihood contribution for $\boldsymbol{\beta}$ is caught up with the missing data mechanism. This suggests a complete case analysis will result in biased inference for $\boldsymbol{\beta}$.

2.3.2 Missing Covariates

Now suppose the response Y is fully observed but the covariate X is only partially observed. Carpenter et. al [5] show that the distribution of the response Y for a regression using only the complete records is written as

$$\begin{aligned} \Pr(Y_i \mid X_i, R_i = 1) &= \frac{\Pr(Y_i, X_i, R_i = 1)}{\Pr(X_i, R_i = 1)} \\ &= \frac{\Pr(R_i = 1 \mid Y_i, X_i) \Pr(Y_i, X_i)}{\Pr(R_i = 1 \mid x_i) \Pr(X_i)} \\ &= \left\{ \frac{\Pr(R_i = 1 \mid Y_i, X_i)}{\Pr(R_i = 1 \mid X_i)} \right\} \Pr(Y_i \mid X_i). \end{aligned} \tag{2.12}$$

The final form of 2.12 suggests that a complete case analysis will result in biased estimates when the missing data mechanism involves the response Y. Thus, if X is MAR given Y or X is MNAR depending on both X and Y then a complete case analysis will yield a biased estimator for β. However, if X is MNAR and only depends on X then a complete case analysis will yield an unbiased, but inefficient estimator for β [5,20].

2.3.3 Missing Covariates and Response with Auxiliary Information

Finally, suppose there is a partially observed covariate X, partially observed response Y, and fully observed covariate Z where Y and X are MAR given Z. If Y is regressed on X and Z then units with missing X and Y contribute

$$\int \Pr(Y \mid \beta; X, Z) dY = 1 \tag{2.13}$$

to the likelihood and a complete case analysis will be unbiased [5].

2.3.4 Summary of Effects on Parameter Estimation

This section summarizes the main results of the previous sections. In general, the common theme seen throughout this chapter is that if the missing data mechanism depends on the response then a complete case analysis will result in biased parameter estimates for the intercept coefficient. Further, if the mechanism depends on both the response and a covariate then the intercept and corresponding covariate parameters will be biased. Table 2.1 provides more details about which specific parameter estimates will be biased when performing a complete case analysis where the missing data mechanism depends on different combinations of variables. It should be noted that the bias does not depend on which variable has missing data but instead on the variable the mechanism depends on. However, the variable with missing values will be important when determining an appropriate approach for handling missing data [5].

Table 2.1: Summary of missing data mechanisms and bias of coefficient estimates when performing a complete case analysis using linear and logistic regression

Variable mechanism depends on	Biased Coefficients	
	Linear Regression	Logistic Regression
Y	$\beta_0, \beta_x, \beta_z$	β_0
X	-	-
Z	-	-
Y, X	$\beta_0, \beta_x, \beta_z$	β_0, β_x
Y, Z	$\beta_0, \beta_x, \beta_z$	β_0, β_z
X, Z	-	-
Y, X, Z	$\beta_0, \beta_x, \beta_z$	$\beta_0, \beta_x, \beta_z$

X = covariate, Y = response
Z = fully observed variable correlated with Y

2.4 Multiple Imputation

Imputation is an approach used to fill in missing values in a dataset. This is useful as imputation maintains the full sample size and eliminates the bias caused by complete case analysis. However, imputation methods may introduce other kinds of bias. The standard error of estimates are typically too low when single imputation techniques are used. This is because the imputed values are treated as if they are the actual values although there is always uncertainty. Thus, multiple imputation can be used to eliminate this potential bias. The following steps for conducting multiple imputation were created by Rubin in 1987 [21]:

1. Imputation – Impute the missing entries K times to create K complete data sets.
2. Analysis – Analyze each of the K complete data sets using standard procedures.
3. Pooling – Combine all K results to produce a single MI estimator and to draw inferences.

For step 3, let $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$ denote the estimate of $\boldsymbol{\beta}$ for each k^{th} completed data set, k ϵ (1,...,K). The MI estimate of $\boldsymbol{\beta}$ is the average of these estimates,

$$\widehat{\boldsymbol{\beta}}_{\text{MI}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\boldsymbol{\beta}}_k. \tag{2.14}$$

The variance of $\widehat{\boldsymbol{\beta}}_{\text{MI}}$ takes into account both between and within imputation variance. Let $\widehat{\boldsymbol{V}}_k$ denote the covariance matrix of the k^{th} completed data set. The average within-imputation covariance matrix is the average of these estimates

$$\widehat{\mathbf{W}} = \frac{1}{K}\sum_{k=1}^{K}\widehat{\boldsymbol{V}}_k \tag{2.15}$$

and the between-imputation covariance matrix is given by

$$\widehat{\mathbf{B}} = \frac{1}{K-1}\sum_{k=1}^{K}(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{\text{MI}})(\widehat{\boldsymbol{\beta}}_k - \widehat{\boldsymbol{\beta}}_{\text{MI}})^T. \tag{2.16}$$

An estimate of the covariance of $\widehat{\boldsymbol{\beta}}_{\text{MI}}$ is

$$\widehat{\mathbf{V}}_{\text{MI}} = \widehat{\mathbf{W}} + \left(1 + \frac{1}{K}\right)\widehat{\mathbf{B}}. \tag{2.17}$$

An advantage of MI is that variability is more accurate for each missing value because it considers variability due to sampling and variability due to imputation [5]. Thus, there is less uncertainty in the estimates and standard errors. A historical disadvantage of MI was that it could require extensive programming and computational resources, but many statistical software programs now have packages built in with this capability (eg., jomo [26], mice [36], and norm [23] in R [27]; PROC MI in SAS [32]) [13,14].

2.4.1 Continuous Variable Imputation

Assuming data are missing at random and there is not a pattern of missingness, the imputation model is the multivariate normal model

$$\mathbf{Y} \sim \text{N}(\boldsymbol{\beta}, \boldsymbol{\Omega}). \tag{2.18}$$

A Gibbs sampler is used to estimate the joint model by sampling from the conditional distribution of each variable and to impute the missing data [5]. As before, partition the response $\mathbf{Y}$ into an observed vector,$\mathbf{Y}_O$, and missing vector, $\mathbf{Y}_M$. The Gibbs sampler is initialized by estimating values for $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ using the observed data. Further, a value for each missing variable $\mathbf{Y}_M$ is drawn by sampling with replacement from the observed values of the corresponding variable. These initial values are denoted by $\boldsymbol{\beta}^0$, $\boldsymbol{\Omega}^0$, and $\mathbf{Y}_M^0$. Next, calculate the sample mean, $\overline{\mathbf{Y}}^0$, and variance, $\mathbf{S}^0$, using $\mathbf{Y}_M^0$ and $\mathbf{Y}_O$ [5].

At iteration r of the Gibbs sampler [5],

1. Draw $\boldsymbol{\Omega}^{-1,r} \sim \text{W}\{n + v, (S_p^{-1} + S^{r-1})^{-1}\}$, where W denotes a Wishart distribution.
2. Draw $\boldsymbol{\beta}^r \sim N(\overline{\mathbf{Y}}^{r-1}, n^{-1}\boldsymbol{\Omega}^r)$;
3. Draw $\mathbf{Y}_M^r \sim f(\mathbf{Y}_M|\boldsymbol{\beta}^r, \boldsymbol{\Omega}^r, \mathbf{Y}_O)$. See Appendix B for more details.
4. Update $\overline{\mathbf{Y}}^r$ and $\mathbf{S}^r$ using $\mathbf{Y}_M^r, \mathbf{Y}_O$. This completes iteration r.

5. Repeat steps 1-4.

Since the sampler is initialized with biased estimates, the samples generated in the beginning may not be representative of the true posterior distribution. Therefore, the samples from this 'burn in' period should be discarded. This burn in period allows the sample to converge to a good approximation of the sampling distribution from which samples may be drawn. Further, a Markov chain of samples is generated since each sample is based on approximations from the previous sample. This means each sample is correlated with nearby samples. As a result, there is typically a 'burn between' period if independent samples are desired.

For example, a Gibbs sampler with a burn in of 5000 and burn between of 1000 works as follows:

1. 5000 $\mathbf{Y}_M$ samples are generated but are discarded (burn in).
2. The next $\mathbf{Y}_M$ sample generated is then combined with $\mathbf{Y}_O$ to form imputed dataset $\mathbf{Y}^1$.
3. 1000 new $\mathbf{Y}_M$ samples are generated but are discarded (burn between).
4. The next $\mathbf{Y}_M$ sample generated is then combined with $\mathbf{Y}_O$ to form imputed dataset $\mathbf{Y}^2$
5. Steps 3-4 are repeated until the desired number of imputed datasets are created, $\mathbf{Y}^k, k = 3, ..., K$.

2.4.2 Binary Variable Imputation

Assume there is a nonmonotone missingness pattern and missing binary covariates. A latent variable approach can be used in order to use a multivariate normal model as the imputation model as described in section 2.4.1 [5]. A latent normal variable Z_i is defined such that

$$\begin{aligned} Z_i > 0 &\iff Y_i = 1 \\ Z_i \leq 0 &\iff Y_i = 0, \end{aligned} \tag{2.19}$$

where $Z_i \sim N(\beta, 1)$. The general steps for binary variable imputation are similar to continuous variable imputation. The sampler is initialized with values using the complete data and then a series of succesive draws are made to update these values. See Appendix B for detailed steps.

Chapter 3

Statistical Learning

Binary outcomes are common in health research where outcomes are often classified as “diseased” or “not diseased”, or "event" or "no event". Further, EMR data can be useful in predicting whether a patient will be classified into a diseased (or event) category based on demographic and clinical characteristics and medical test results.

The statistical learning model of interest is a neural network, the focus of which is a single layer neural network with a sigmoidal activation function. This model is chosen because there is a general lack of knowledge about the impact of bias inducing MAR mechanisms on prediction models and their associated metrics (eg., sensitivity and specificity). As this model is equivalent to a binary logistic regression, it is a logical starting point to understand the effects of missing data on statistical learning models that could be used in health research.

3.1 Modeling Binary Outcomes

3.1.1 Generalized Linear Model Overview

Recall the multiple linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is used to model the linear relationship between a continuous variable Y and p predictors $X_1, ..., X_p$. An issue with the linear model is that it does not model nonnormal responses well. Therefore, a class of models known as generalized linear models (GLM)

were developed to represent other response types from the exponential family such as categorical or binary responses. A GLM is made up of three components [22]:

- Random component - the probability distribution of the response Y. For linear regression, $Y \sim N(\mu, \sigma^2)$.
- Systematic component, η - a linear combination of $X_1, ..., X_p$ known as the linear predictor. For linear regression, the linear predictor is $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- Link function, $g(\mu)$ - a function that links the random and systematic components by describing how the mean of the response, $E(Y) = \mu$, relates to the linear predictor. Linear regression has the simplest link function, $g(\mu_i) = \eta_i$, the identity link function.

For GLM, it is assumed the response Y is a member of the exponential family with the general form

$$f(y \mid \theta, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right] \tag{3.1}$$

where θ is known as the canonical parameter representing location and ϕ is the dispersion parameter representing scale. Members of the exponential family are defined through the functions $a(\phi), b(\theta)$, and $c(y, \phi)$ [22]. The mean and variance for a GLM are found by

$$\begin{aligned} \mathrm{E}(Y) &= \mu = b'(\theta) \\ \mathrm{Var}(Y) &= b''(\theta)a(\phi). \end{aligned} \tag{3.2}$$

Furthermore, parameters of the GLM models can be fit to data using a form of iteratively reweighted least squares. See Appendix B for more details.

3.1.2 Logistic Regression

One GLM for modeling the relationship between a binary response Y, $Y \epsilon$ {0,1}, where Y follows a binomial distribution

$$Y \sim \text{Binomial}(n, \pi), \tag{3.3}$$

and regressors $X_1, ..., X_p$ is known as logistic regression. This relationship is measured in terms of the probability that $Y = 1$ given X, $\pi = \Pr(Y = 1 \mid X)$. While a normal linear regression model with a dummy variable approach (coded as 0/1) would produce an estimate of this probability, it would also be possible for π to be less than 0 or greater than 1, complicating the interpretation of these values as probabilities. To ensure the estimated probabilities are restricted to the range [0,1], the logistic function is used [17] and takes the form

$$\pi = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \tag{3.4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$, which is equivalent to the sigmoidal function

$$\pi = \frac{1}{1 + e^{-\mathbf{X}\boldsymbol{\beta}}}. \tag{3.5}$$

It is seen from Equation 3.4 that π approaches 0 as $\mathbf{X}\boldsymbol{\beta}$ approaches negative infinity and π approaches 1 as $\mathbf{X}\boldsymbol{\beta}$ approaches positive infinity. Thus, the logistic function successfully limits the range of π to [0,1]. Equation 3.4 can be manipulated to find

$$\frac{\pi}{1 - \pi} = e^{\mathbf{X}\boldsymbol{\beta}} \tag{3.6}$$

where $[\pi/(1 - \pi)]$ represents the odds of an event occurring. Odds can range from [0, ∞], with values close to 0 indicating a low relative probability of an event occurring and values approaching ∞ indicating a high relative probability of an event occurring [17]. Equation 3.6 can be further manipulated by taking the logarithm of both sides to find

$$log\left(\frac{\pi}{1 - \pi}\right) = \mathbf{X}\boldsymbol{\beta}. \tag{3.7}$$

The left side of Equation 3.7 is the link function for logistic regression referred to as the logit. The logit link models the log-odds of the mean, π. It is seen by Equation 3.7 that the logit is linear in X. Thus, an example interpretation for

this model is that a one-unit increase in X_1 while holding $X_2, ..., X_p$ constant changes the log-odds by β_1 [17].

Since the binomial distribution is a member of the exponential family, this can be written in exponential family notation,

$$\begin{aligned} f(y \mid \theta, \phi) &= \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \exp\left[y\log(\pi) + (n-y)\log(1-\pi) + \log\binom{n}{y} \right] \\ &= \exp\left[y\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi) + \log\binom{n}{y} \right]. \end{aligned} \tag{3.8}$$

From 3.8 we see that $\theta = log\left(\frac{\pi}{1-\pi}\right), a(\phi) = 1, b(\theta) = -n\log(1-\pi)$, and $c(y, \phi) = \log\binom{n}{y}$. Thus, we now see the canonical link has g such that $\eta = g(\mu) = \theta$. Also, using Equation 3.2 the variance can be expressed as $\pi(1-\pi)$.

The coefficients $\beta_0, \beta_1, ..., \beta_p$ are estimated using the maximum likelihood approach. See Appendix B for more details. For logistic regression, the parameters of interest for this process are

$$\begin{aligned} \eta &= log\left(\frac{\pi}{1-\pi}\right) \\ \frac{d\eta}{d\pi} &= \frac{1}{\pi(1-\pi)} \\ V(\mu) &= \pi(1-\pi) \\ w &= \pi(1-\pi). \end{aligned} \tag{3.9}$$

In general, there are three assumptions of the logistic regression model. The first assumption was introduced in the beginning of this section; that is, the response Y_i has a binomial distribution $Y_i \sim B(n_i, \pi_i)$. Second, there should not be any outliers in the data. Third, the model should have little or no multicollinearity among the predictors. These are important assumptions that are taken into account when generating data for the simulation studies [35].

3.2 Assessing Model Performance

Measures of classification performance can be found by looking at a confusion matrix as shown in Table 3.1. The possible outcomes are correctly predicting a positive outcome (eg., $Y = 1$) (true positive), correctly predicting a negative outcome (eg., $Y = 0$) (true negative), incorrectly predicting positive (false positive), and incorrectly predicting negative (false negative) [17].

Table 3.1: Confusion matrix depicting four possible classification outcomes

		Actual	
		+	-
Predicted	+	True Positive (TP)	False Positive (FP)
	-	False Negative (FN)	True Negative (TN)

Sensitivity is a measure of the true positive rate calculated as the proportion of actual positives that are correctly predicted as positive [17],

$$\frac{TP}{TP + FN}. \tag{3.10}$$

Specificity is a measure of the true negative rate calculated as the proportion of actual negatives that are correctly predicted as negative [17],

$$\frac{TN}{TN + FP}. \tag{3.11}$$

The classification accuracy is calculated as the proportion of predicted outcomes that are correctly predicted [17],

$$\frac{TP + TN}{TP + FP + TN + FN}, \tag{3.12}$$

and likewise the error rate is calculated as the proportion of predicted outcomes that are incorrectly predicted [17],

$$\frac{FP + FN}{TP + FP + TN + FN}. \tag{3.13}$$

A method of visualizing the performance of the classifier graphically is a receiving operating characteristic (ROC) curve by plotting the true positive rate

(sensitivity) against the false positive rate (1 - specificity) for all possible classification thresholds. For each classification threshold, if the probability of response is greater than the threshold then the predicted response is classified as an event that occurred and if the probability is lower than the threshold then the predicted response is classified as an event that did not occur [17]. For example,

$$\begin{aligned} \hat{\pi} > c \rightarrow \hat{Y} = 1 \\ \hat{\pi} < c \rightarrow \hat{Y} = 0 \end{aligned} \tag{3.14}$$

for all $c \, \epsilon \, [0, 1]$, where c is the threshold for classification.

The area under the curve (AUC) is a measure of the ability to correctly classify positive and negative outcomes and is the integral of the ROC curve over the false positive rate. An AUC close to 1 is indicative of near perfect prediction whereas an AUC near 0.50 is no better than a random coin flip at correctly predicting an outcome [17].

Chapter 4

Simulation Studies

The first goal of the simulation study is to determine whether missing values affect prediction accuracy, sensitivity, specificity, ROC curves, and AUC. This is accomplished by analyzing a dataset with no missing values, applying a missing data mechanism, and then analyzing the dataset with missing values to compare the corresponding results. The second goal is to determine if multiple imputation is a useful analytic correction for improving the prediction accuracy, sensitivity, specificity, ROC curves, and AUC relative to these values from the missing values dataset. I use two simulation studies to explore both goals with a simple and more complex missing data mechanism.

4.1 Study Design

4.1.1 Data Generating Mechanism

The general format of the simulation study begins by first generating a dataset with no missing values designated as the "full dataset". The simulation study uses a dataset of sample size 1000 units with two continuous predictors, two binary predictors, and one binary outcome. Data for the continuous predictors X_1, X_2 are generated from a multivariate normal distribution with mean 0, variance 1, and a low level of correlation,

$$X_1, X_2 \sim N_2(0, \Sigma), \Sigma = \begin{bmatrix} 1.00 & 0.05 \\ 0.05 & 1.00 \end{bmatrix} \tag{4.1}$$

The continuous variable X_1 represents a demographic variable such as age and X_2 represents a nearly uncorrelated continuous baseline measurement. The binary variable X_3 represents a binary treatment assignment or exposure that was randomly applied to the subjects. X_3 is generated from a binomial distribution with probability $\pi = 0.5$. The binary variable X_4 represents a binary baseline measurement and is generated from a binomial distribution with probability $\pi = 0.3$. Hence, X_4 emulates an unconditional prevalence rate. For both simulation studies, I set the coefficients for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ to be (0, 3, 2, 4, 1). Equation 3.4 is used to generate estimated probabilities π_i. The binary response variable Y_i is generated as a Bernoulli random variable taking the value 1 with probability π_i.

Full Data

Using the full dataset with no missing values, I randomly split the data 80/20 into training and test datasets, respectively. For simplicity, I use a single split rather than implement bootstrap or methods of data re-use, and for consistency in methodology as outlined in the following sections. The training set observations are used to build a predictive model and then the test set observations are applied to this model to make predictions. This is completed by fitting a logistic regression model regressing $\mathbf{Y}_{train}$ on $\mathbf{X}_{train}$ to estimate $\hat{\boldsymbol{\beta}}_{train}$. The test set covariates $\mathbf{X}_{test}$ are then applied to this model to make predictions for $\hat{\mathbf{Y}}_{test}$. This allows us to create a confusion matrix by comparing the known values of $\mathbf{Y}_{test}$ to the predicted values of $\hat{\mathbf{Y}}_{test}$. This confusion matrix is used to estimate accuracy, sensitivity, and specificity for a naive split for being classified as having the outcome (eg., Pr($Y = 1 \mid X) \geq 0.50 \rightarrow \hat{Y}_{test} = 1$). An ROC curve and the AUC are obtained.

Imperfect and Complete Case Data

Next, a missing data mechanism is applied to the full dataset to create the imperfect dataset. The imperfect dataset is randomly split 80/20 into training and test datasets. Some studies split this dataset prior to the introduction of missing values into the training dataset only, but this implies that the data

testing the model is from a different population than the population whose data trains the model [24].

The training set builds the predictive model and the test set makes predictions. However, due to missing values, there is now less information available to train the model as well as test the model since a complete case analysis is performed. Once again, the prediction error, sensitivity, specificity, ROC curve, and AUC are calculated. These results are compared to the corresponding results from the full dataset to determine if missing values have affected any of the measures of interest.

Missing Data Mechanisms

There are two missing data mechanisms investigated through the simulation study. The first missing data mechanism examined is a covariate MAR on the response and is designed such that the probability of the current covariate of interest being missing for those with the outcome of interest (Y=1) is two times the probability of missingness for those who do not have the outcome of interest (Y=0). The probability of missingness for each missing rate is shown in Table 4.1.

Table 4.1: Missing Data Mechanism 1 Design - Probability of X_2, X_4, and X_2 and X_4 being missing for both levels of the outcome for each missing rate.

Target Missing Rate	$(Y = 0)$	$(Y = 1)$
10%	0.06	0.12
30%	0.18	0.36
50%	0.30	0.60

The second missing data mechanism examined is a covariate MAR on the response as well as the other covariates X_1 and X_3. This mechanism is applied through a logistic regression process by creating a combination of Y, X_1, and X_3 as shown in Table 4.2. The logit function is then used to create a probability of a value being observed. These probabilities are then applied to a binomial distribution to determine whether a value is observed (1) or missing (0).

Table 4.2: Missing Data Mechanism 2 Design - Equations for creating missing rates of 10%, 30%, and 50% and the probability of X_2, X_4, and X_2 and X_4 being missing for both levels of the outcome when $X_1 = 0$ and $X_3 = 0$.

Target Missing Rate	Equation	$X_1 = 0, X_3 = 0$	
		$Y = 0$	$Y = 1$
10%	2.00 + 2Y + 0.50X_1 - X_3 + 0.25X_1X_3	0.12	0.02
30%	0.25 + 2Y + 0.50X_1 - X_3 + 0.25X_1X_3	0.44	0.10
50%	-0.80 + 2Y + 0.50X_1 - X_3 + 0.25X_1X_3	0.69	0.23

Within each missing data mechanism, there are three missing covariate situations examined:

1. Only continuous covariate X_2 missing
2. Only binary covariate X_4 missing
3. Both X_2 and X_4 missing

Further, for each of these three situations, there are also three varying missing rates examined to determine whether the amount of missingness is important: 10%, 30%, and 50%. All 18 possible scenarios are depicted by the tree in Figure 4.1 where the first split displays the two missing data mechanisms, the 3 splits for each of the missing data mechanisms represents the variable that is missing, and then the 3 splits for each missing variable represents the percent of that variable that is missing.

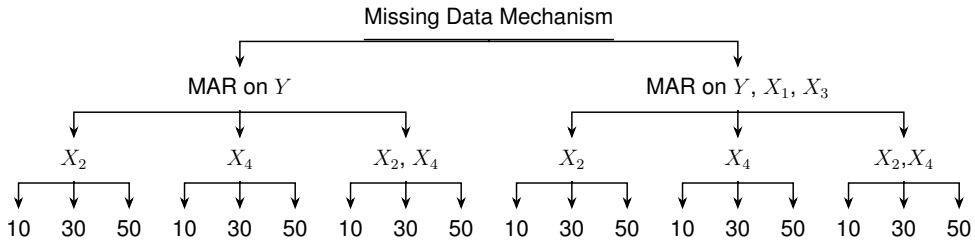

Figure 4.1: Tree depicting all 18 scenarios investigated through the simulation studies.

Multiple Imputation Data

Next, multiple imputation is applied to the imperfect dataset to determine if using this method improves the results from the missing data situation. The jomo function in the 'jomo' package [26] in R [27] is used for performing multiple imputation. This approach handles both continuous covariates and categorical covariates as long as they are binary indicator variables. Based on parameters shown to be efficient in *Multiple Imputation and its Application*, there are 5 imputations performed with a burn in of 5000 samples and burn between of 1000 samples [5]. The results of these five imputations are combined using Rubin's Rules.

This process from generating a full dataset, applying missing data mechanisms to create the imperfect and complete case datasets, and then using MI to create an imputed dataset completes one simulation. This is repeated 500 times. The results of all 500 simulations are combined to obtain simulation study averages. For the ROC curves, the final curve for each scenario is created by taking the vertical average of all 500 curves for that specific scenario [9].

4.2 Simulation Study Assessment

The estimated coefficients of the predictive model for the full data and complete case data are calculated as the average of the coefficients over all simulations,

$$\hat{\boldsymbol{\beta}} = \frac{1}{S}\sum_{i=1}^{S}\tilde{\boldsymbol{\beta}}_i. \tag{4.2}$$

For this study, the coefficients from each of the five imputations are combined using Equation 2.15 to generate $\tilde{\boldsymbol{\beta}}_{MI,i}$. The simulation study coefficients are then calculated as the average of the $\tilde{\boldsymbol{\beta}}_{MI}$ coefficients over all simulations,

$$\hat{\boldsymbol{\beta}} = \frac{1}{S}\sum_{i=1}^{S}\tilde{\boldsymbol{\beta}}_{MI,i}. \tag{4.3}$$

A more convenient way of viewing these results is by examining the bias of $\hat{\beta}_0$,

$$\text{Bias}(\hat{\beta}_0) = \text{E}[\hat{\beta}_0] - \beta_0, \tag{4.4}$$

and the percent relative bias of $\hat{\boldsymbol{\beta}}_j$, for $j \in \{1, 2, 3, 4\}$, relative to the true coefficient values,

$$\% \text{ relative bias}(\hat{\boldsymbol{\beta}}_j) = \frac{\hat{\boldsymbol{\beta}}_{\boldsymbol{j}} - \boldsymbol{\beta}_{\boldsymbol{j}}}{\boldsymbol{\beta}_{\boldsymbol{j}}} \times 100. \tag{4.5}$$

The sensitivity, specificity, and error rate for each simulation iteration are calculated using Equations 3.14, 3.15, and 3.17, respectively, to obtain values for $\widetilde{\text{Sensitivity}}_i$, $\widetilde{\text{Specificity}}_i$, and $\widetilde{\text{Error}}_i$. The AUC of each ROC curve is also found, $\widetilde{\text{AUC}}_i$. The final results for each of these measures are then the average value over all simulation iterations,

$$\begin{aligned}
\widehat{\text{Sensitivity}} &= \frac{1}{S}\sum_{i=1}^{S} \widetilde{\text{Sensitivity}}_i \\
\widehat{\text{Specificity}} &= \frac{1}{S}\sum_{i=1}^{S} \widetilde{\text{Specificity}}_i \\
\widehat{\text{Error}} &= \frac{1}{S}\sum_{i=1}^{S} \widetilde{\text{Error}}_i \\
\widehat{\text{AUC}} &= \frac{1}{S}\sum_{i=1}^{S} \widetilde{\text{AUC}}_i
\end{aligned} \tag{4.6}$$

The percent relative bias of each of these measures for the complete case data and imputed data are calculated relative to the full data,

$$\begin{aligned}
\% \text{ relative bias}(\widehat{\text{Sensitivity}}_M) &= \frac{\widehat{\text{Sensitivity}}_M - \widehat{\text{Sensitivity}}_{full}}{\widehat{\text{Sensitivity}}_{full}} \times 100 \\
\% \text{ relative bias}(\widehat{\text{Specificity}}_M) &= \frac{\widehat{\text{Specificity}}_M - \widehat{\text{Specificity}}_{full}}{\widehat{\text{Specificity}}_{full}} \times 100 \\
\% \text{ relative bias}(\widehat{\text{Error}}_M) &= \frac{\widehat{\text{Error}}_M - \widehat{\text{Error}}_{full}}{\widehat{\text{Error}}_{full}} \times 100 \\
\% \text{ relative bias}(\widehat{\text{AUC}}_M) &= \frac{\widehat{\text{AUC}}_M - \widehat{\text{AUC}}_{full}}{\widehat{\text{AUC}}_{full}} \times 100, \\
&\text{for M = \{Complete Case, Imputed\}.}
\end{aligned} \tag{4.7}$$

4.3 Results

4.3.1 Missing Data Mechanism 1

The estimated coefficients of the predictive model for each missing variable and missing rate scenario investigated for missing data mechanism 1 are found in Table A.1 in Appendix A.

Table 4.3 displays the bias of $\hat{\beta}_0$ and the percent relative bias of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ for each of these scenarios relative to the true values. It is shown that $\hat{\beta}_0$ of a complete case analysis is biased for each missing variable scenario. Further, the bias is shown to increase as the missing rate increases. When X_2 is missing, β_4 has the largest increase in bias as the missing rate increases. When X_4 is missing, the bias of each coefficient increases as the missing rate increases, but $\hat{\beta}_4$ is consistently the most biased coefficient. Finally, when X_2 and X_4 are missing, the bias consistently increases as the missing rate increases for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ but not for $\hat{\beta}_4$. When multiple imputation is used as a correction for the missing data, the bias of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$ decreases for each missing variable and missing rate scenario investigated. However, there are 3 instances where the bias of $\hat{\beta}_4$ actually increases when multiple imputation is used: 10% X_2 missing, 10% X_2, X_4 missing, and 50% X_2, X_4 missing.

Table 4.4 summarizes the sensitivity, specificity, error rate, and AUC when the test data is applied to the predictive model. For the values of the confusion matrices used to calculate the sensitivity and specificity, see Table A.2 in Appendix A. Table 4.5 displays the percent relative bias of these measures for the complete case dataset and imputed dataset relative to the full dataset. Regardless of which variable is missing, the sensitivity from a complete case analysis is attenuated relative to the full dataset. The magnitude of this effect increases as the missing rate increases such that the more missing values there are, the greater the sensitivity is underestimated.

The opposite effect is seen with specificity. The specificity from a complete case analysis is augmented relative to the full dataset and the magnitude of the effect increases with the missing rate such that the specificity is overestimated more as there are more missing values present in the dataset. The classification

error rate from a complete case analysis is always greater than the error rate of the full dataset but there is no trend as the missing rate increases. Finally, the AUC is unchanged relative to the full dataset when a complete case analysis is performed. When multiple imputation is applied, the bias of sensitivity and the error rate is always decreased. The same is shown for specificity except for one scenario where 10% X_4 is missing.

The ROC curves for each simulation are combined by taking the vertical average of the curves. This is done by choosing fixed false positive rates and then averaging the corresponding true positive rates [8]. The full data, complete case, and imputed data ROC curves for 10%, 30%, and 50% X_2 MAR on Y are shown by Figures 4.2-4.4. Figures 4.5-4.7 display the full data, complete case, and imputed data ROC curves for 10%, 30%, and 50% X_2, X_4 MAR on Y. It is seen that as the missing rate increases, the full data and complete case curves separate farther apart. Further, when both X_2 and X_4 are missing, the distance between the curves is greater than when only one variable is missing. The ROC curves for X_4 MAR on Y are similar to the curves when X_2 is MAR on Y so these curves are shown in Figures A.1-A.3 in Appendix A.

Table 4.3: Missing Data Mechanism 1: Bias of $\hat{\beta}_0$ and percent relative bias of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ of the predictive model created by the training data for the full data, complete case, and imputed data for each missing variable and missing rate investigated.

Missing Variable	Scenario	**Bias**	**% Relative Bias**			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\mathbf{X_2}$	**10% Missing**					
	Full Data	0.016	2.83	2.75	2.30	0.10
	Complete Case	-0.050	2.83	2.85	2.33	-0.10
	Imputed	0.023	2.23	1.95	1.53	-0.80
	30% Missing					
	Full Data	0.003	1.70	2.15	2.45	0.60
	Complete Case	-0.166	2.20	2.60	3.10	2.00
	Imputed	0.019	0.00	-0.65	0.20	-1.00
	50% Missing					
	Full Data	-0.003	2.50	2.30	1.73	4.10
	Complete Case	-0.352	4.30	4.45	3.50	9.40
	Imputed	0.016	-0.20	-2.35	-1.40	1.70
$\mathbf{X_4}$	**10% Missing**					
	Full Data	-0.009	1.90	1.85	2.10	4.70
	Complete Case	-0.076	2.40	2.30	2.48	4.60
	Imputed	-0.006	1.87	1.80	2.05	3.50
	30% Missing					
	Full Data	-0.008	1.67	1.95	1.75	3.80
	Complete Case	-0.184	3.10	2.75	2.85	4.90
	Imputed	-0.003	1.63	1.90	1.70	1.80
	50% Missing					
	Full Data	-0.003	2.00	2.15	2.08	2.10
	Complete Case	-0.357	4.17	4.10	4.58	5.00
	Imputed	0.001	1.97	2.05	2.03	-2.20
$\mathbf{X_2, X4}$	**10% Missing**					
	Full Data	-0.002	2.27	2.20	2.08	1.40
	Complete Case	-0.137	2.67	2.10	2.48	0.80
	Imputed	0.009	1.57	1.10	1.25	-1.00
	30% Missing					
	Full Data	0.007	2.20	1.80	1.58	3.80
	Complete Case	-0.335	3.53	3.45	2.55	3.90
	Imputed	0.031	0.63	-0.15	-0.43	-1.10
	50% Missing					
	Full Data	-0.002	1.70	1.75	1.90	1.60
	Complete Case	-0.697	8.47	7.65	7.15	1.90
	Imputed	0.039	-0.70	-2.45	-1.43	-8.90

Table 4.4: Missing Data Mechanism 1: Estimated sensitivity, specificity, error rate, and AUC results of the testing data for the full data, complete case, and imputed data for each missing variable and missing rate investigated.

Missing Variable	Scenario	Sensitivity	Specificity	Error Rate	AUC
$\mathbf{X_2}$	**10% Missing**				
	Full Data	0.891	0.758	0.152	0.955
	Complete Case	0.884	0.759	0.156	0.955
	Imputed	0.891	0.758	0.153	0.954
	30% Missing				
	Full Data	0.891	0.746	0.155	0.955
	Complete Case	0.879	0.771	0.159	0.955
	Imputed	0.891	0.746	0.156	0.953
	50% Missing				
	Full Data	0.891	0.758	0.152	0.956
	Complete Case	0.869	0.789	0.160	0.955
	Imputed	0.885	0.742	0.157	0.953
$\mathbf{X_4}$	**10% Missing**				
	Full Data	0.891	0.758	0.152	0.955
	Complete Case	0.885	0.759	0.154	0.955
	Imputed	0.891	0.754	0.153	0.955
	30% Missing				
	Full Data	0.891	0.758	0.153	0.956
	Complete Case	0.879	0.776	0.158	0.956
	Imputed	0.891	0.758	0.153	0.956
	50% Missing				
	Full Data	0.891	0.758	0.154	0.955
	Complete Case	0.869	0.789	0.161	0.955
	Imputed	0.891	0.758	0.152	0.955
$\mathbf{X_2, X4}$	**10% Missing**				
	Full Data	0.891	0.758	0.152	0.956
	Complete Case	0.879	0.764	0.156	0.955
	Imputed	0.891	0.758	0.153	0.955
	30% Missing				
	Full Data	0.884	0.758	0.154	0.955
	Complete Case	0.867	0.789	0.166	0.954
	Imputed	0.891	0.742	0.155	0.953
	50% Missing				
	Full Data	0.891	0.758	0.153	0.956
	Complete Case	0.852	0.833	0.162	0.953
	Imputed	0.884	0.746	0.157	0.953

Table 4.5: Missing Data Mechanism 1: Percent relative bias of sensitivity, specificity, error, and AUC Results of the testing data for the full data, complete case, and imputed data for each missing variable and missing rates investigated.

Missing Variable	Scenario	% Relative Bias			
		Sensitivity	Specificity	Error Rate	AUC
$\mathbf{X_2}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	-0.79	0.13	2.63	0.00
	Imputed	0.00	0.00	0.66	-0.10
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	-1.35	3.35	2.58	0.00
	Imputed	0.00	0.00	0.65	-0.21
	50% Missing				
	Full Data	-	-	-	-
	Complete Case	-2.47	4.09	5.26	-0.10
	Imputed	-0.67	-2.11	3.29	-0.31
$\mathbf{X4}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	-0.67	0.13	1.32	0.00
	Imputed	0.00	-0.53	0.66	0.00
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	-1.35	2.37	3.27	0.00
	Imputed	0.00	0.00	0.00	0.00
	50% Missing				
	Full Data	-	-	-	-
	Complete Case	-2.47	4.09	4.55	0.00
	Imputed	0.00	0.00	-1.30	0.00
$\mathbf{X_2, X4}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	-1.35	0.79	2.63	-0.10
	Imputed	0.00	0.00	0.66	-0.10
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	-1.92	4.09	7.79	-0.10
	Imputed	0.79	-2.11	0.65	-0.21
	50% Missing				
	Full Data	-	-	-	-
	Complete Case	-4.38	9.89	5.88	-0.10
	Imputed	-0.79	-1.58	2.61	-0.31

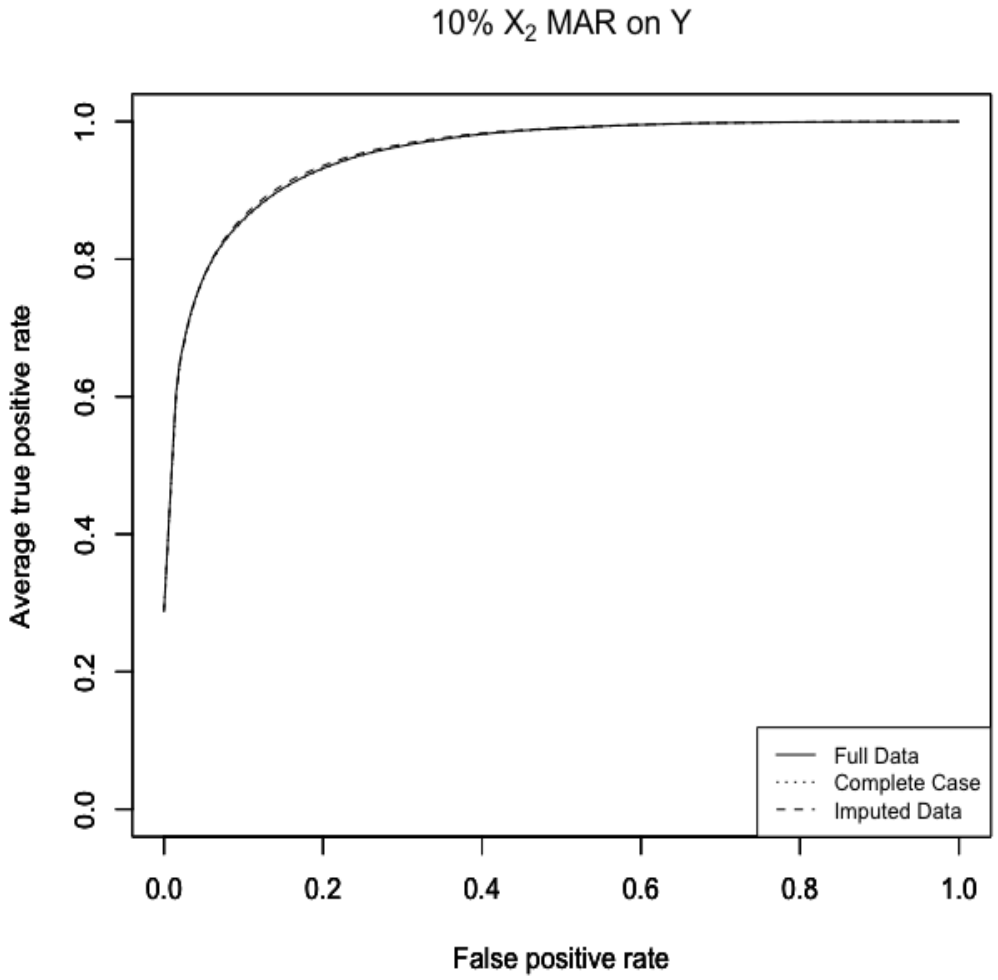

Figure 4.2: ROC Curves for 10% X_2 MAR on Y - full data, complete case, and imputed data.

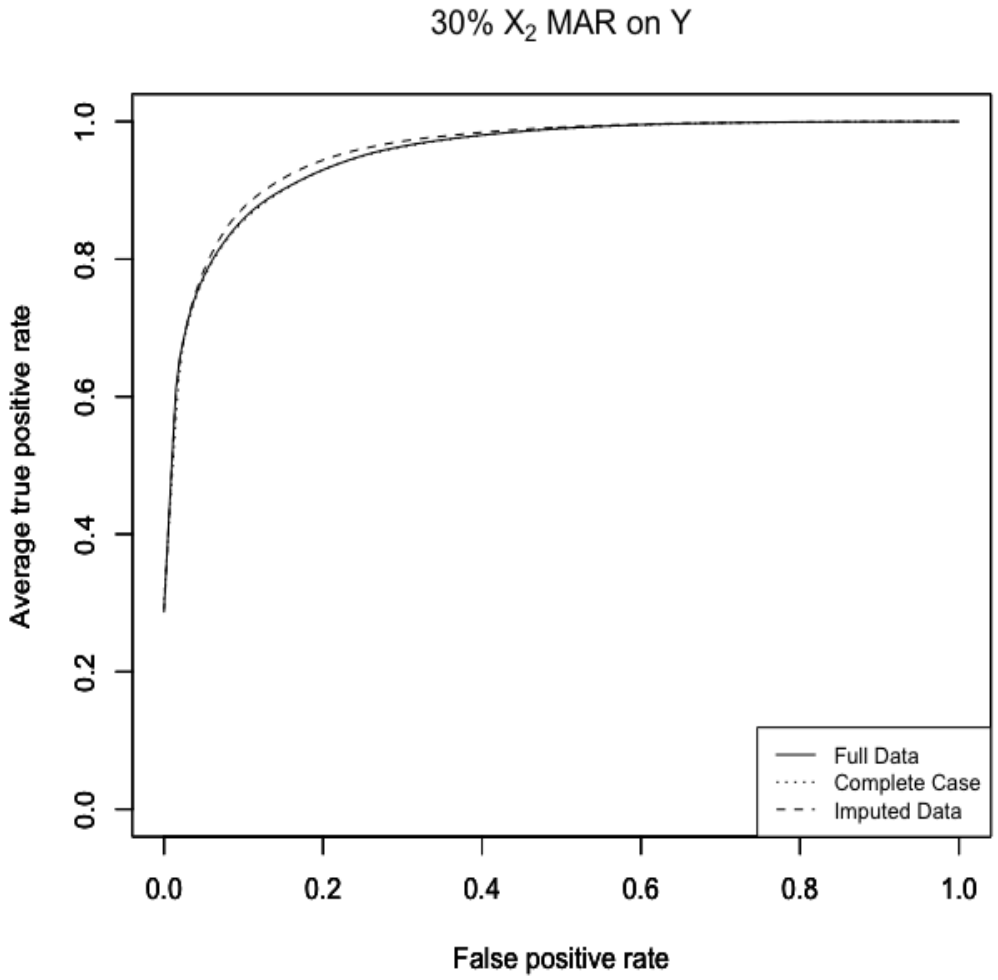

Figure 4.3: ROC Curves for 30% X_2 MAR on Y - full data, complete case, and imputed data.

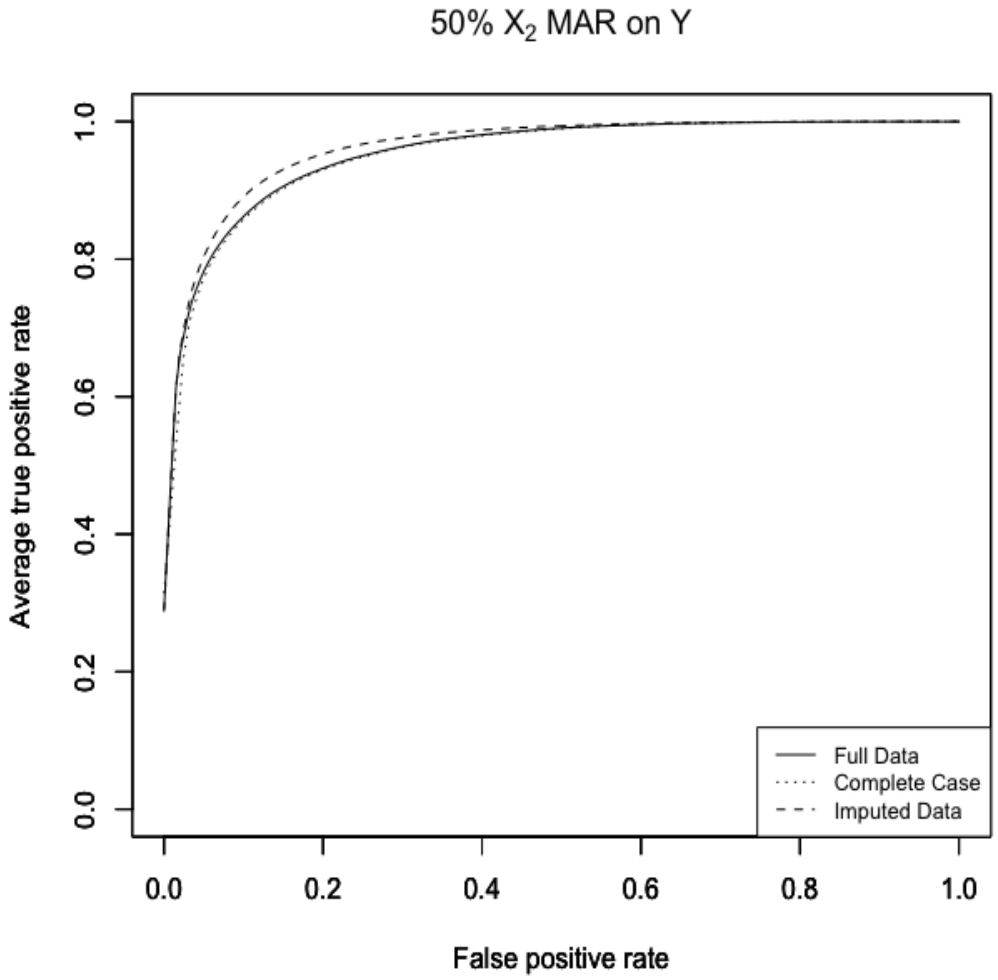

Figure 4.4: ROC Curves for 50% X_2 MAR on Y - full data, complete case, and imputed data.

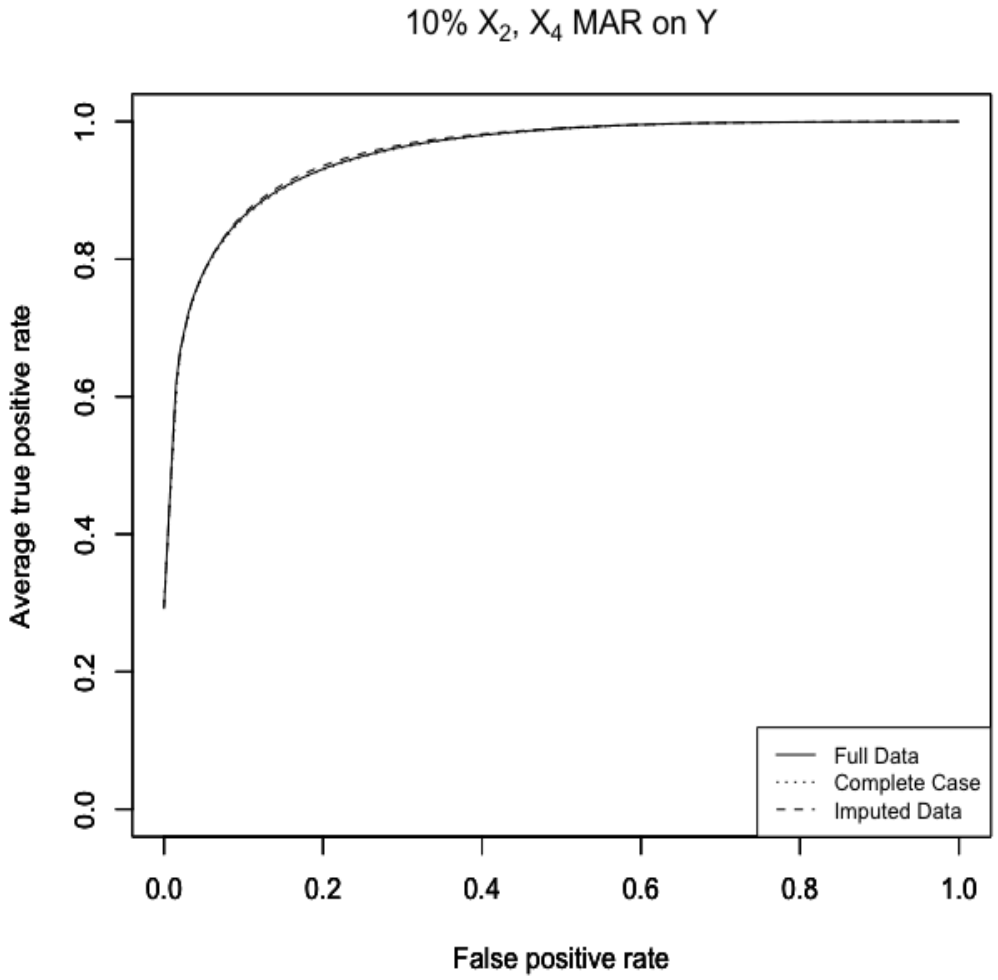

Figure 4.5: ROC Curves for 10% X_2, X_4 MAR on Y - full data, complete case, and imputed data.

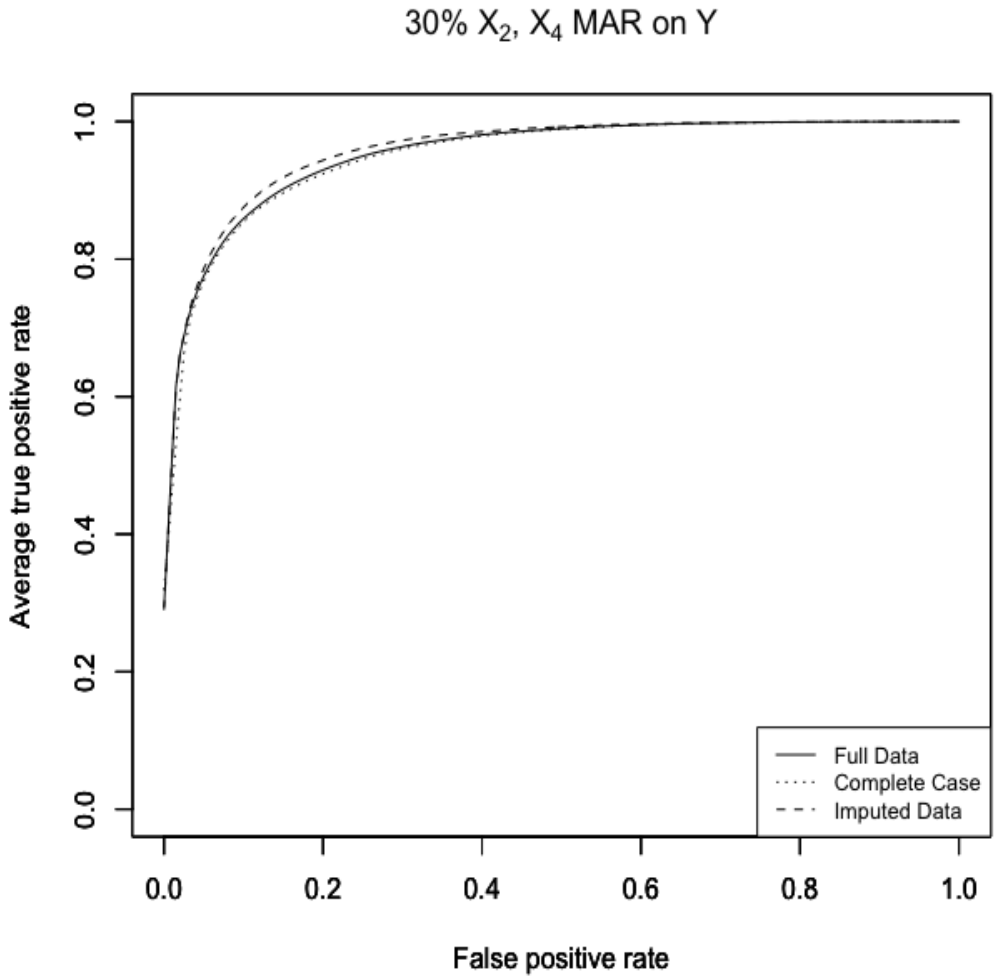

Figure 4.6: ROC Curves for 30% X_2, X_4 MAR on Y - full data, complete case, and imputed data.

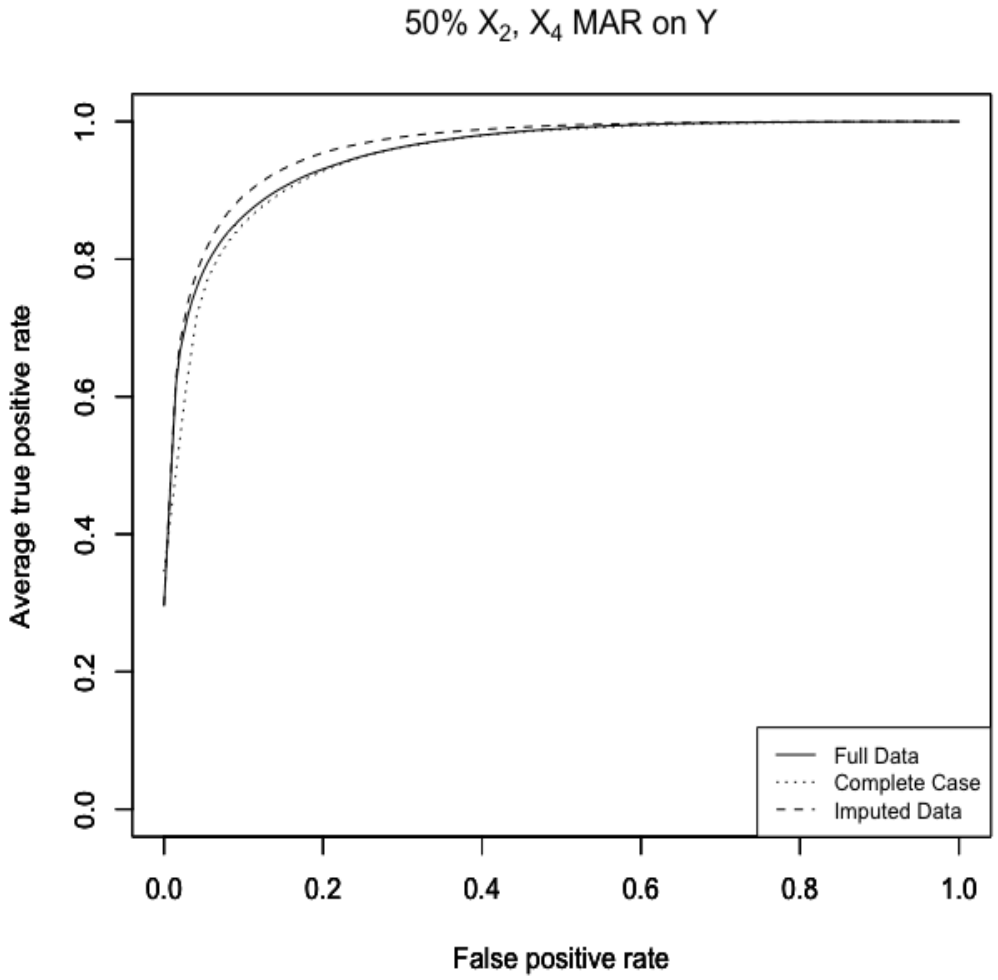

Figure 4.7: ROC Curves for 50% X_2, X_4 MAR on Y - full data, complete case, and imputed data.

4.3.2 Missing Data Mechanism 2

Due to 'perfect prediction' no results were obtained for the 50% X_2, X_4 missing scenario. When both covariates have 50% missingness, the total cases with missing values is approximately 65%. Therefore, since there are only 70 cases available to test with compared to 200 for the full dataset, there were instances where the model completely separated observations into the correct categories.

The results for the remaining scenarios of the second missing data mechanism with covariates MAR on Y, X_1, and X_3 are summarized in Tables 4.6-4.8. The estimated coefficients of the predictive model for each missing variable and missing rate scenario investigated can be found in Table A.3 in Appendix A. Table 4.6 displays the bias of $\hat{\beta}_0$ and the percent relative bias of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ for each of these scenarios relative to the true values. It is shown that $\hat{\beta}_0$ of a complete case analysis is biased for each missing variable scenario. Further, the bias is shown to increase as the missing rate increases. For all missing scenarios investigated, $\hat{\beta}_3$ has the greatest percent relative bias and it also has the greatest increase in percent relative bias as the missing rate increases. The percent relative bias of $\hat{\beta}_2$ and $\hat{\beta}_4$ also increase as the missing rate increases. However, the percent relative bias of $\hat{\beta}_1$ increases in magnitude as the missing rate increases from 10% to 30% but then decreases in magnitude from 30% to 50%.

In general, the magnitude of bias and percent relative bias for each coefficient in each imputed data scenario is less than that of the values for the corresponding complete case dataset. There is one instance where this is not the case: $\hat{\beta}_1$ when 10% X_4 is missing. However, the percent relative bias of the imputed dataset is similar in magnitude and direction to the value for the full data. Thus, while the magnitude of the percent relative bias for the imputed dataset relative to the full data is greater than that for the complete case dataset, the bias for the imputed data relative to full data is less than the bias of the complete case data relative to the full data.

Table 4.7 summarizes the sensitivity, specificity, error rate, and AUC when the test data is applied to the predictive model. For the values of the confusion matrices used to calculate the sensitivity and specificity, see Table A.4

in Appendix A. Table 4.8 displays the percent relative bias of these measures for the complete case dataset and imputed dataset relative to the full dataset. Regardless of which variable is missing, the sensitivity from a complete case analysis is augmented relative to the full dataset. This may be due to the relative proportion of missing data for $Y = 0$ and $Y = 1$ being opposite that compared to study 1. The magnitude of this effect increases as the missing rate increases such that the more missing values there are, the greater the sensitivity is overestimated.

The specificity from a complete case analysis is attenuated relative to the full dataset and the magnitude of the effect increases with the missing rate such that the specificity is underestimated more as there are more missing values present in the dataset. The classification error rate from a complete case analysis is always lower than the error rate of the full dataset. As the missing rate increases, the error decreases. Finally, the AUC is unchanged relative to the full dataset when a complete case analysis is performed.

The full data, complete case, and imputed data ROC curves for 10%, 30%, and 50% X_2 MAR on Y, X_1, and X_3 are shown by Figures 4.8-4.10. Figures 4.11-4.13 display the full data, complete case, and imputed data ROC curves for 10%, 30%, and 50% X_2, X_4 MAR on Y, X_1, and X_3. It is seen that as the missing rate increases, the full data and complete case curves separate farther apart. Further, when both X_2 and X_4 are missing, the distance between the curves is greater than when only one variable is missing. The ROC curves for X_4 MAR on Y, X_1, and X_3 are similar to the curves when X_2 is MAR on Y, X_1, and X_3 so these curves are shown in Figures A.4-A.6 in Appendix A.

Table 4.6: Missing Data Mechanism 2: Bias of $\hat{\beta}_0$ and percent relative bias of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ of the predictive model created by the training data for the full data, complete case, and imputed data for each missing variable and missing rate investigated.

Missing Variable	Scenario	**Bias**	**% Relative Bias**			
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\mathbf{X_2}$	**10% Missing**					
	Full Data	-0.002	1.77	2.25	2.28	3.60
	Complete Case	0.113	-1.50	2.50	9.00	3.50
	Imputed	-0.001	1.13	1.10	1.23	2.80
	30% Missing					
	Full Data	-0.004	2.13	2.05	2.18	1.20
	Complete Case	0.489	-3.63	3.80	16.83	3.90
	Imputed	-0.009	0.30	-0.85	-0.150	-0.20
	50% Missing					
	Full Data	-0.009	2.07	1.95	1.45	1.00
	Complete Case	0.935	-1.90	5.80	23.85	8.50
	Imputed	-0.037	-1.13	-4.90	-2.60	-1.30
$\mathbf{X_4}$	**10% Missing**					
	Full Data	0.002	1.97	2.45	2.05	2.10
	Complete Case	0.115	-1.30	2.95	8.53	2.60
	Imputed	0.005	1.90	2.40	2.03	1.20
	30% Missing					
	Full Data	-0.001	2.10	2.15	1.60	0.50
	Complete Case	0.487	-2.97	4.30	17.20	4.10
	Imputed	0.007	2.13	2.20	1.58	-1.20
	50% Missing					
	Full Data	0.004	1.60	2.30	1.88	2.60
	Complete Case	0.961	-2.20	6.50	25.00	9.20
	Imputed	0.032	1.60	2.20	1.60	-4.10
$\mathbf{X_2}, \mathbf{X_4}$	**10% Missing**					
	Full Data	0.002	2.17	1.95	2.03	1.70
	Complete Case	0.227	-3.73	2.50	14.98	3.10
	Imputed	0.004	1.77	1.20	1.13	0.90
	30% Missing					
	Full Data	0.002	2.20	2.60	2.08	1.10
	Complete Case	0.985	-6.50	6.30	77.80	8.70
	Imputed	0.000	0.60	-0.80	-0.35	-2.40
	50% Missing					
	Full Data	NA	NA	NA	NA	NA
	Complete Case	NA	NA	NA	NA	NA
	Imputed	NA	NA	NA	NA	NA

Table 4.7: Missing Data Mechanism 2: Estimated sensitivity, specificity, error, and AUC Results of the testing data for the full data, complete case, and imputed data for each missing variable and missing rates investigated.

Missing Variable	Scenario	Sensitivity	Specificity	Error Rate	AUC
$\mathbf{X_2}$	**10% Missing**				
	Full Data	0.891	0.754	0.153	0.955
	Complete Case	0.903	0.723	0.142	0.955
	Imputed	0.891	0.754	0.154	0.954
	30% Missing				
	Full Data	0.891	0.758	0.151	0.956
	Complete Case	0.922	0.680	0.120	0.956
	Imputed	0.884	0.742	0.156	0.954
	50% Missing				
	Full Data	0.891	0.758	0.153	0.954
	Complete Case	0.946	0.615	0.098	0.953
	Imputed	0.884	0.742	0.161	0.951
$\mathbf{X_4}$	**10% Missing**				
	Full Data	0.891	0.754	0.153	0.955
	Complete Case	0.902	0.723	0.144	0.955
	Imputed	0.891	0.758	0.152	0.956
	30% Missing				
	Full Data	0.891	0.754	0.153	0.955
	Complete Case	0.922	0.680	0.120	0.954
	Imputed	0.891	0.758	0.152	0.955
	50% Missing				
	Full Data	0.891	0.742	0.153	0.955
	Complete Case	0.946	0.615	0.099	0.952
	Imputed	0.891	0.754	0.153	0.955
$\mathbf{X_2}, \mathbf{X_4}$	**10% Missing**				
	Full Data	0.891	0.758	0.152	0.954
	Complete Case	0.914	0.711	0.133	0.955
	Imputed	0.891	0.746	0.154	0.954
	30% Missing				
	Full Data	0.891	0.758	0.151	0.956
	Complete Case	0.949	0.583	0.091	0.952
	Imputed	0.891	0.754	0.156	0.954
	50% Missing				
	Full Data	NA	NA	NA	NA
	Complete Case	NA	NA	NA	NA
	Imputed	NA	NA	NA	NA

Table 4.8: Missing Data Mechanism 2: Percent relative bias of sensitivity, specificity, error, and AUC Results of the testing data for the full data, complete case, and imputed data for each missing variable and missing rates investigated.

Missing Variable	Scenario	% Relative Bias			
		Sensitivity	Specificity	Error Rate	AUC
$\mathbf{X_2}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	1.35	-4.11	-7.19	0.00
	Imputed	0.00	0.00	0.65	-0.08
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	3.48	-10.29	-20.53	0.00
	Imputed	-0.79	-2.11	3.31	-0.24
	50% Missing				
	Full Data	-	-	-	-
	Complete Case	6.17	-18.87	-35.95	-0.10
	Imputed	-0.79	-2.11	5.23	-0.36
$\mathbf{X_4}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	1.23	-4.11	-5.88	0.00
	Imputed	0.00	0.53	-0.65	0.02
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	3.48	-9.81	-21.57	-0.10
	Imputed	0.00	0.53	-0.65	0.00
	50% Missing				
	Full Data	-	-	-	-
	Complete Case	6.17	-17.12	-35.29	-0.31
	Imputed	0.00	1.62	0.00	-0.01
$\mathbf{X_2, X_4}$	**10% Missing**				
	Full Data	-	-	-	-
	Complete Case	2.58	-6.20	-12.50	0.10
	Imputed	0.00	-1.58	1.32	-0.01
	30% Missing				
	Full Data	-	-	-	-
	Complete Case	6.51	-23.09	-39.73	-0.42
	Imputed	-0.79	-0.53	3.31	-0.22
	50% Missing				
	Full Data	NA	NA	NA	NA
	Complete Case	NA	NA	NA	NA
	Imputed	NA	NA	NA	NA

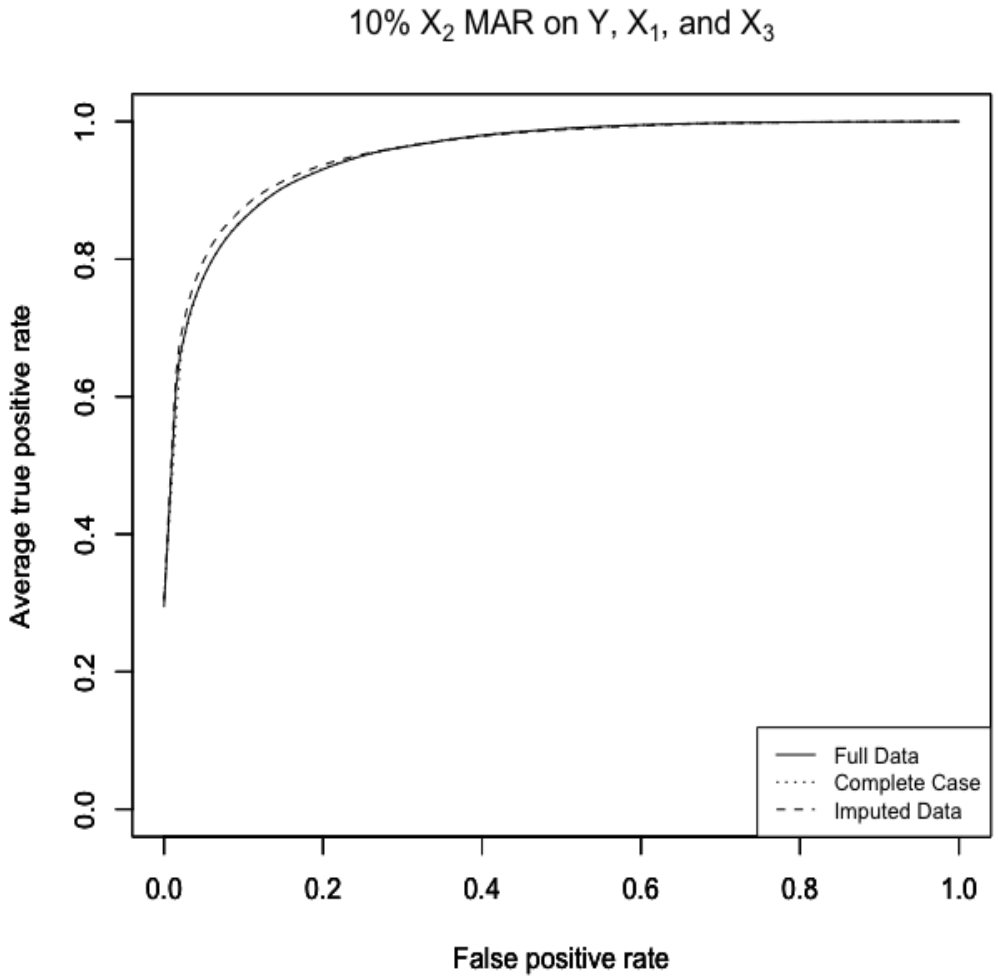

Figure 4.8: ROC Curves for 10% X_2 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

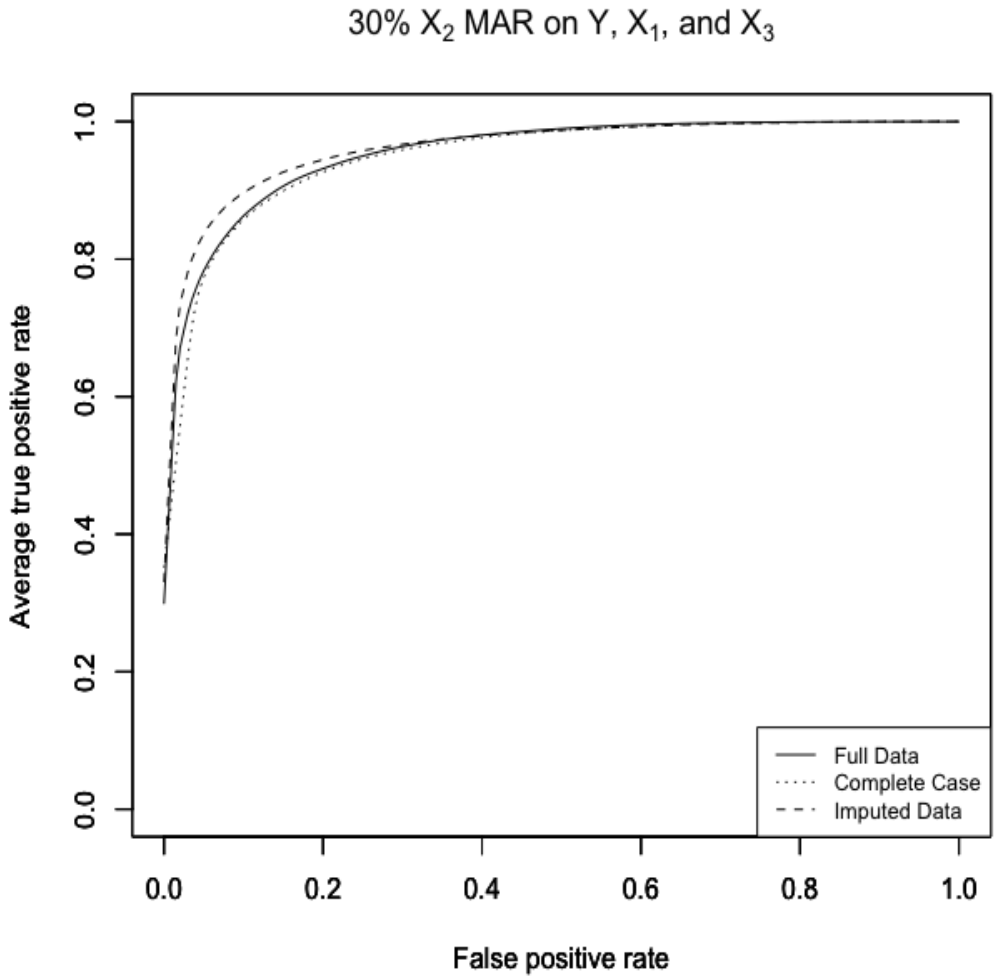

Figure 4.9: ROC Curves for 30% X_2 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

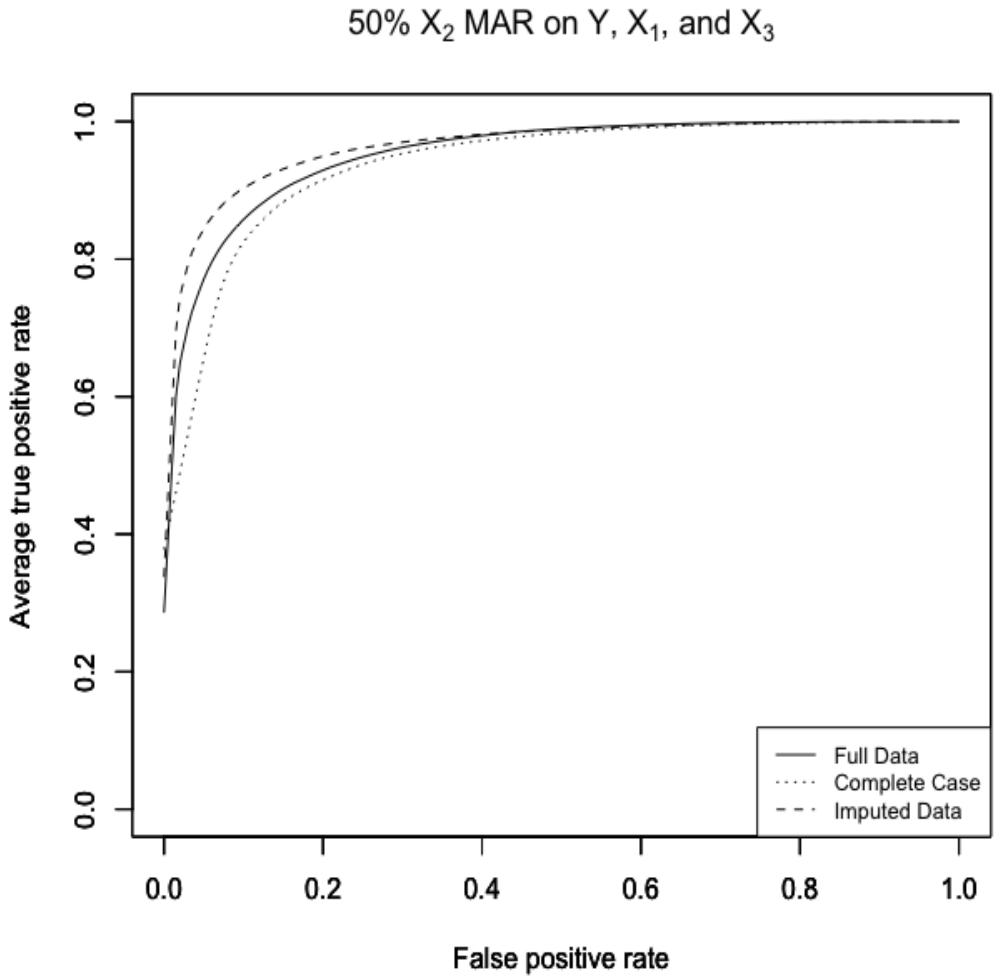

Figure 4.10: ROC Curves for 50% X_2 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

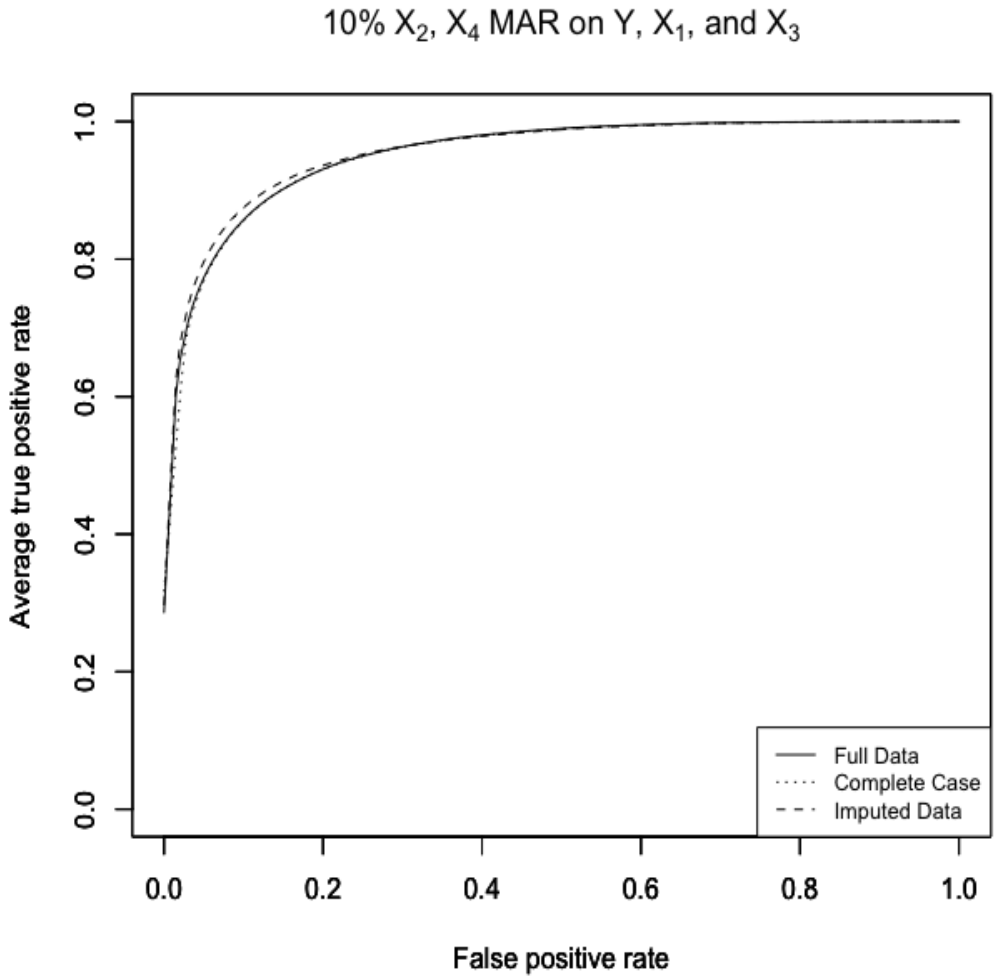

Figure 4.11: ROC Curves for 10% X_2, X_4 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

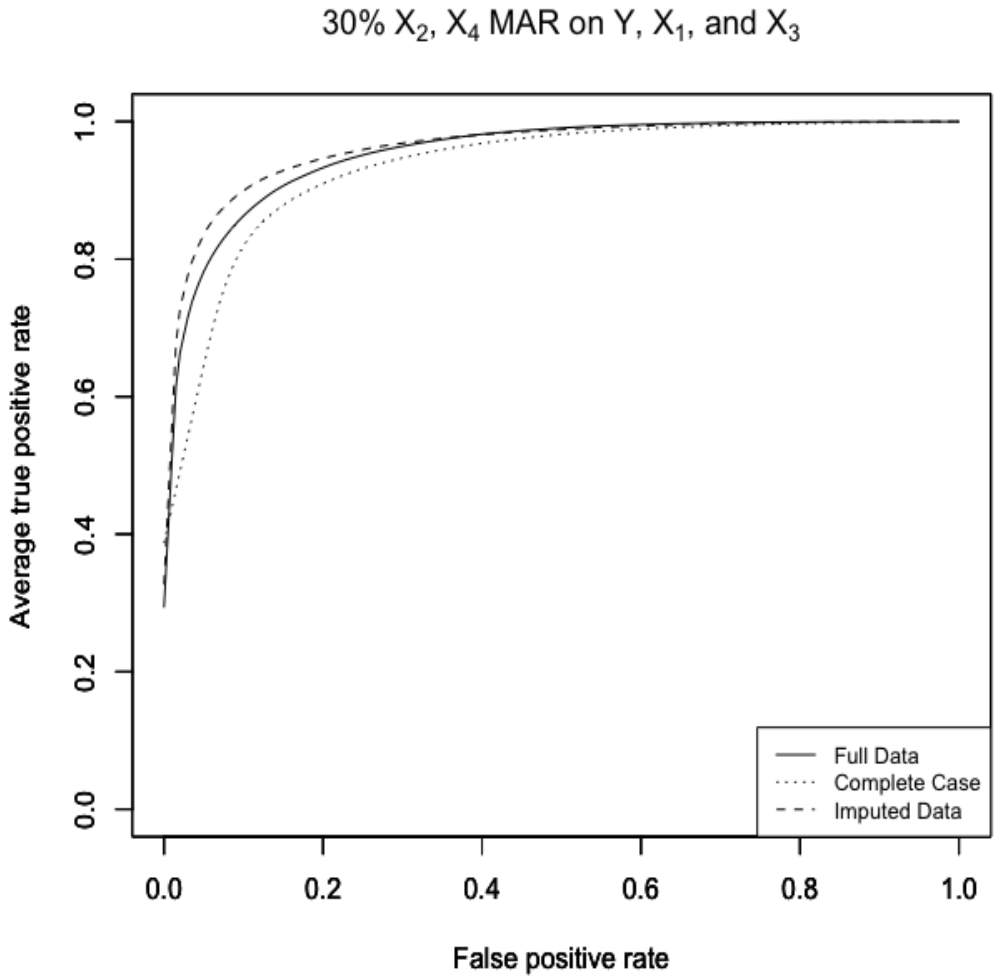

Figure 4.12: ROC Curves for 30% X_2, X_4 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

4.4 Discussion

4.4.1 Parameter Estimates

As expected, the intercept coefficient β_0 is biased when a complete case analysis is performed because both missing data mechanisms depend on the response Y. It is also shown that the bias of β_0 increases as the missingness increases regardless of the variable missing or the complexity of the missing data mechanism. The magnitude of bias of β_0 from a complete case analysis is greater for missing data mechanism 2 than missing data mechanism 1 for each scenario. This suggests that a more complex missing data mechanism that includes other covariates will lead to more biased parameter estimates than a simple mechanism that includes the response only when a complete case analysis is performed.

The direction of bias is opposite between mechanisms as mechanism 1 underestimates β_0 and mechanism 2 overestimates β_0. This is likely due to the design of the mechanisms. As shown in Table 4.1, the probability of missing when $Y = 1$ is two times greater than when $Y = 0$ for mechanism 1. However, mechanism 2 is designed such that the probability of missing when $Y = 0$ is greater than when $Y = 1$ (see Table 4.2). For both mechanisms, multiple imputation decreases the bias of β_0 for all scenarios.

Further, the bias of β_3 from a complete case analysis is greater for missing data mechanism 2 than mechanism 1. This bias is present because mechanism 2 includes X_3 but mechanism 1 does not. However, there is not a noticeable difference in bias for X_1 between the mechanisms. For most scenarios in both mechanisms, multiple imputation decreases the bias of $\beta_1, \beta_2, \beta_3$, and β_4.

4.4.2 Sensitivity, Specificity, and Error Estimates

For both missing data mechanisms, a complete case analysis results in biased estimates of sensitivity and specificity. It is further seen that the magnitude of bias for both measures increases as the missing rate increases for both mechanisms. The direction of the bias for sensitivity is always opposite that of the bias for specificity. However, the direction of bias is not constant between the missing data mechanisms. For mechanism 1, the sensitivity is underestimated

and the specificity is overestimated; however, for mechanism 2, the sensitivity is overestimated and the specificity is underestimated.

To understand what is causing these measures to be biased, I examine the true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) for each scenario in both missing data mechanisms (see Tables A.2 and A.4 in Appendix A). The number of false positives and false negatives for missing data mechanism 1 are always slightly greater than the number for the corresponding scenario in mechanism 2. There is a minimum difference of 1 and maximum difference of 3 for these values between the two mechanisms. Considering there is only a slight difference, these measures do not appear to be the reason the sensitivity (affected by FN) and specificity (affected by FP) are so biased.

However, the number of true positives for mechanism 1 are lower than the number for the corresponding scenario in mechanism 2. Likewise, the number of true negatives for mechanism 1 are higher than the number for the corresponding scenario in mechanism 2. For example, when 30% X_2 and X_4 are missing, mechanism 1 only has 52 TP but mechanism 2 has 93 TP. In the same scenario, mechanism 1 has 30 TN but mechanism 2 only has 7.

It appears that the probability of missing data for both levels of the response is what is causing the direction of bias in sensitivity and specificity. Missing data mechanism 1 underestimates sensitivity because the probability of missing "positives" ($Y = 1$) is greater than the probability of missing "negatives" ($Y = 0$). Therefore, there will not be as many positives to classify into the true positive category. Likewise, mechanism 2 underestimates specificity because there will not be as many negatives to classify since the probability of missing "negatives" ($Y = 0$) is greater than the probability of missing "positives" ($Y = 1$).

When multiple imputation is used to correct the missing data, the magnitude of bias for sensitivity and specificity decrease for both missing data mechanisms. This suggests multiple imputation is a useful method for correcting the bias of these estimates across different MAR mechanisms, different types of missing covariates, and varying missing rates.

Lastly, the error rate is also biased when performing a complete case analysis. The error rate is always biased relative to the full dataset. However, the error rate increases for missing data mechanism 1 but decreases for mechanism 2. For missing data mechanism 1, the error rate is higher because the value in the denominator decreases more than the values in numerator. This means the true positives and true negatives are affected by the mechanism more than the false positives and false negatives.

For mechanism 2, the error rate decreases when a complete case analysis is performed. This implies that the false positives and false negatives are most affected by this mechanism. Comparing mechanism 1 to mechanism 2, it has already been stated that mechanism 1 always has a higher total number of false positives and false negatives than the corresponding scenario in mechanism 2. However, while it appears that a decrease in the error rate would be ideal, we must realize that other measures are severely biased.

I return to further explore the sensitivity and specificity estimates of the two missing data mechanisms. Simply stated, missing data mechanism 2 is overestimating sensitivity and severely underestimating specificity. Since there are more positive responses in the complete case dataset based on the design of missing data mechanism 2, small increases in sensitivity decrease the error rate more than large decreases in specificity would increase the error rate.

We see that while the sensitivity and specificity are both biased for mechanism 1, the magnitude of bias for the two measures are similar and always lower than 10% relative to the full data. This means that the the decrease in the number of true positives correctly being identified is similar to the increase in the number of true negatives correctly being identified.

However, for missing data mechanism 2 the magnitude of bias for specificity is greater than 10% for 4 of the 9 scenarios investigated and goes as high as 23% biased relative to the full data. This means that the predictive model is classifying less negative responses as negative and more positive responses as positive. Therefore, since there are more positive responses in the dataset and more of them are correctly being classified as positive, the error rate will decrease. This comes at a huge cost to the specificity of the predictive model

since there are less negative responses in the dataset and they are also being identified as negative less frequently.

Once again, multiple imputation is useful for decreasing the bias of the complete case error rate relative to the full data error rate. While this corresponds to an increase in error rate for missing data mechanism 2, we have already seen that the lower error rate came at a cost of the specificity.

4.4.3 ROC Curves and AUC

The vertical averaged ROC curves for the full data and complete case data tend to spread further apart as the missing rate increases. However, the difference is not as evident in missing data mechanism 1 as it is in missing data mechanism 2. For both missing data mechanisms, the ROC curve of the imputed data is always slightly above the full data ROC curve. Overall, there appears to be regions in the ROC curves where the full data and complete case data curves are spread apart but the significance of the difference is unknown.

The AUC values reported in this thesis are the average of the AUC values from each ROC curve created over the simulations, not of the vertically averaged ROC curves. The AUC is shown to be negligibly affected by missing data regardless of the complexity of the missing data mechanism, type of covariate missing, or the missing rate. Therefore, the AUC may not be the best measure for determining the accuracy of prediction when there is missing data because we know that other values such as sensitivity and specificity are affected. It is possible that the AUC was not affected due to setting the beta coefficients to have such a strong signal but this would need to be examined in another study.

Chapter 5

Conclusion

An important observation from this thesis is that the AUC was not found to be biased when performing a complete case analysis, even when the missing rate was as high as 50%. Oftentimes researchers report the AUC to describe the performance of their predictive model, but the results of this thesis show that the AUC is unaffected even if sensitivity and specificity are biased. Thus, caution must be taken when reporting results and the AUC should only be reported alongside other measures such as sensitivity and specificity.

It was found that performing a complete case analysis logistic regression when data are missing at random on the response leads to biased coefficient, sensitivity, specificity, and error rate estimates with increasing bias as the missingness increases. Further, a more complex missing data mechanism that includes covariates leads to estimates of coefficients, sensitivity, specificity, and error rates that are more biased than a simple missing data mechanism that only includes the response.

Multiple imputation was found to be effective in reducing the bias of coefficient, sensitivity, specificity, and error rate estimates. Since many statistical programs include multiple imputation methods, it is suggested that multiple imputation be used as a correction when data are assumed to be missing at random. The methods and results from this thesis can be helpful in making better predictions if used appropriately. For example, if a researcher is exploring EMR data to generate a hypothesis for conducting a new research study and the outcome is known for all subjects, then multiple imputation can be used to

correct for missing values in the covariates. This allows the researcher to use more available information than a complete case analysis and the predictive model will result in less biased estimates for sensitivity, specificity, and error rate.

Once a predictive model has been created with these data, the researcher then conducts a study that requires values for all independent variables of interest to be collected (eg., the system used for collecting medical test results requires a value to be entered). The newly collected data can then be applied to the predictive model to make predictions for each person's probability of having an outcome (or event) in the future. The results of this thesis show that using multiple imputation to correct for missing data can improve the utility of a predictive model. Therefore, the predictions for the newly collected data will be classified more accurately than if multiple imputation was not used when building the model.

5.1 Future Work

There are a number of routes that can be explored for future research. First, it would be worthwhile to apply these methods to a real dataset. An EMR dataset with similar covariate structure would be the best initial choice for comparing results. Other types of datasets could then be investigated to see if the results are more generalizable. Further, a natural extension of these methods is for multilayer neural networks.

A number of other scenarios could also be investigated. The magnitude of the beta coefficients could be decreased to determine if the measures are still biased when the signal is not as strong as in this thesis. The amount of correlation between the covariates could be increased or the sample size could vary to assess finite sample properties. The multiple imputation parameters of n_{burn}, $n_{between}$, and K datasets can also be changed. This multiple imputation step requires the most computational time so if n_{burn} can be decreased and achieve similar results this would serve a practical purpose. A 5-fold cross validation could also be investigated instead of the 80/20 split into training and test datasets as used in this thesis.

The entire simulation study can also be redesigned to further break down the sources of bias. Rather than generating a new full dataset for each iteration, one full dataset can be generated in the beginning and then missing values can be introduced to create multiple imperfect datasets from the same full dataset. This would help partition the sources of error.

For evaluating results, the AUC of the vertically averaged ROC curves could be calculated. The vertically averaged ROC curves appear to show that the AUC of these would be different than simulation study average AUC reported; this should be investigated further. Threshold averaged ROC curves can also be examined to see if this is a useful method for combining ROC curves in this setting.

This thesis does not present new information for handling missing values in the test dataset, but handling missing values in a real life test set needs to be explored. If the value for the outcome is unknown and the missing data mechanism is dependent upon the outcome, multiple imputation can not be used because this is now an MNAR mechanism, hence the missing data mechanism would be inconsistent across datasets.

Appendix A

Additional Results

The $\hat{\beta}$ coefficients of the full data, complete case, and imputed data for missing data mechanism 1 are shown in Table A.1. The true positives, false positives, false negatives, and true negatives used to calculate sensitivity and specificity for missing data mechanism 1 are shown in Table A.2. Figures A.1-A.3 display the 10%, 30%, and 50% X_4 MAR on Y ROC curves.

The $\hat{\beta}$ coefficients of the full data, complete case, and imputed data for missing data mechanism 2 are shown in Table A.3. The true positives, false positives, false negatives, and true negatives used to calculate sensitivity and specificity for missing data mechanism 1 are shown in Table A.4. Figures A.4-A.6 display the 10%, 30%, and 50% X_4 MAR on $Y, X_1,$ andX_3 ROC curves.

Table A.1: Missing Data Mechanism 1: $\hat{\beta}$ coefficients of the predictive model created by the training data for the full data, complete case, and imputed data for each missing variable and missing rates investigated.

Missing Variable	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
	Truth	0	3	2	4	1
$\mathbf{X_2}$	**10% Missing**					
	Full Data	0.016	3.085	2.055	4.092	1.001
	Complete Case	-0.050	3.085	2.057	4.093	0.999
	Imputed	0.023	3.067	2.039	4.061	0.992
	30% Missing					
	Full Data	0.003	3.051	2.043	4.098	1.006
	Complete Case	-0.166	3.066	2.052	4.124	1.020
	Imputed	0.019	3.000	1.987	4.008	0.990
	50% Missing					
	Full Data	-0.003	3.075	2.046	4.069	1.041
	Complete Case	-0.352	3.129	2.089	4.140	1.094
	Imputed	0.016	2.994	1.953	3.944	1.017
$\mathbf{X_4}$	**10% Missing**					
	Full Data	-0.009	3.057	2.037	4.084	1.047
	Complete Case	-0.076	3.072	2.046	4.099	1.046
	Imputed	-0.006	3.056	2.036	4.082	1.035
	30% Missing					
	Full Data	-0.008	3.050	2.039	4.070	1.038
	Complete Case	-0.184	3.093	2.055	4.114	1.049
	Imputed	-0.003	3.049	2.038	4.068	1.018
	50% Missing					
	Full Data	-0.003	3.060	2.043	4.083	1.021
	Complete Case	-0.357	3.125	2.082	4.183	1.050
	Imputed	0.011	3.059	2.041	4.081	0.978
$\mathbf{X_2}, \mathbf{X_4}$	**10% Missing**					
	Full Data	-0.002	3.068	2.044	4.083	1.014
	Complete Case	-0.137	3.080	2.042	4.099	1.008
	Imputed	0.009	3.047	2.022	4.050	0.990
	30% Missing					
	Full Data	0.007	3.066	2.036	4.063	1.038
	Complete Case	-0.335	3.106	2.069	4.102	1.039
	Imputed	0.031	3.019	1.997	3.983	0.989
	50% Missing					
	Full Data	-0.002	3.051	2.035	4.076	1.016
	Complete Case	-0.697	3.254	2.153	4.286	1.019
	Imputed	0.039	2.979	1.951	3.943	0.911

Table A.2: Missing Data Mechanism 1: Confusion matrices displaying the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) used for calculating the sensitivity and specificity.

Missing Variable	Scenario	TP	FP	FN	TN	Sensitivity	Specificity
$\mathbf{X_2}$	**10% Missing**						
	Complete	123	15	15	47	0.891	0.758
	Missing	107	14	14	44	0.884	0.759
	Imputed	123	15	15	47	0.891	0.758
	30% Missing						
	Complete	122	16	15	47	0.891	0.746
	Missing	80	11	11	37	0.879	0.771
	Imputed	122	16	15	47	0.891	0.746
	50% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	53	8	8	30	0.869	0.789
	Imputed	123	16	16	46	0.885	0.742
$\mathbf{X_4}$	**10% Missing**						
	Complete	123	15	15	47	0.891	0.758
	Missing	108	14	14	44	0.885	0.759
	Imputed	123	15	15	46	0.891	0.754
	30% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	80	11	11	38	0.879	0.776
	Imputed	122	15	15	47	0.891	0.758
	50% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	53	8	8	30	0.869	0.789
	Imputed	123	15	15	47	0.891	0.758
$\mathbf{X_2}, \mathbf{X_4}$	**10% Missing**						
	Complete	123	15	15	47	0.891	0.758
	Missing	94	13	13	42	0.879	0.764
	Imputed	123	15	15	47	0.891	0.758
	30% Missing						
	Complete	122	15	16	47	0.884	0.758
	Missing	52	8	8	30	0.867	0.789
	Imputed	122	16	15	46	0.891	0.742
	50% Missing						
	Complete	122	15	15	47	0.891	0.758
	Missing	23	4	4	20	0.852	0.833
	Imputed	122	16	16	47	0.884	0.746

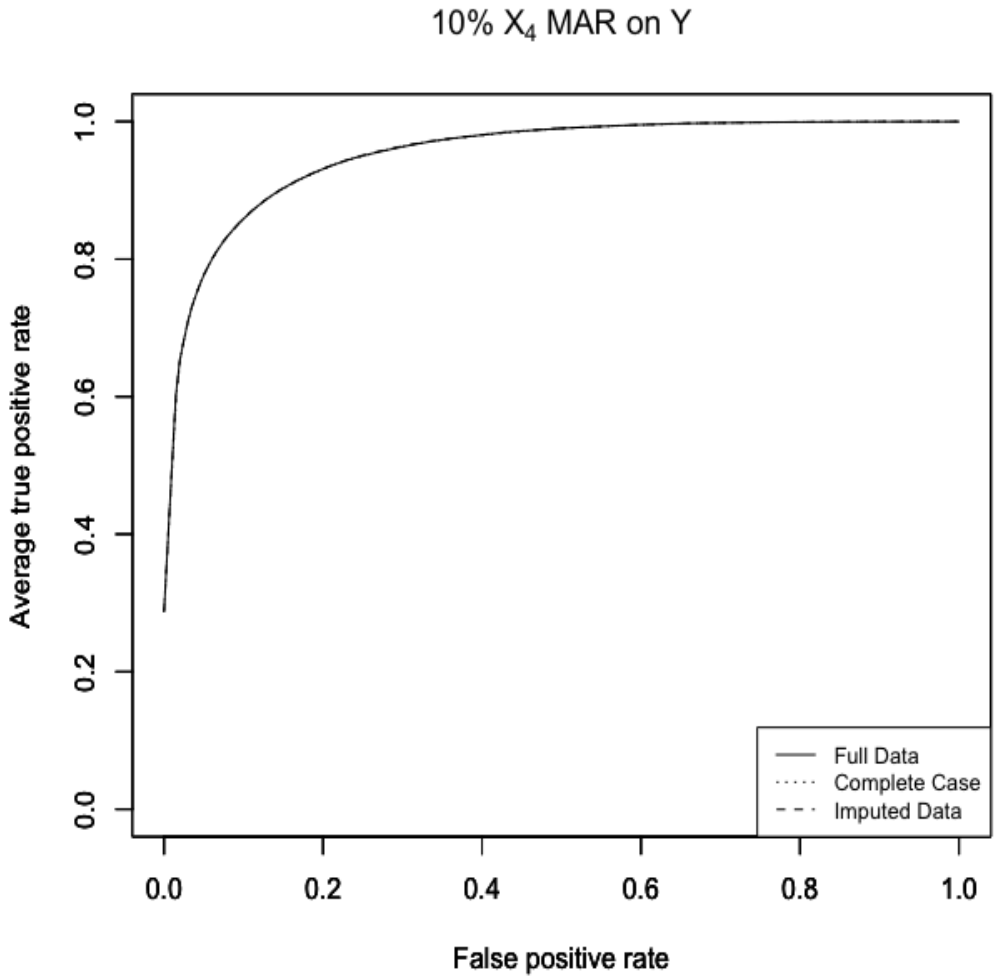

Figure A.1: ROC Curves for 10% X_4 MAR on Y - full data, complete case, and imputed data.

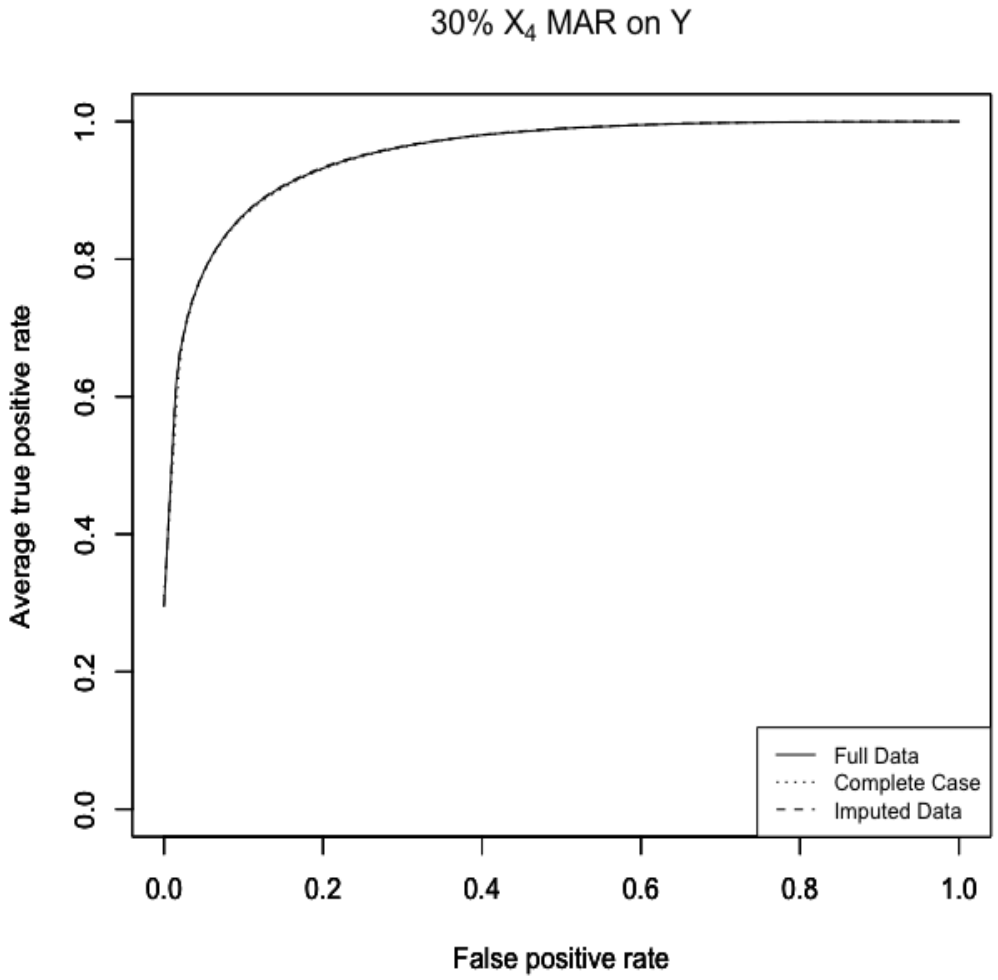

Figure A.2: ROC Curves for 30% X_4 MAR on Y - full data, complete case, and imputed data.

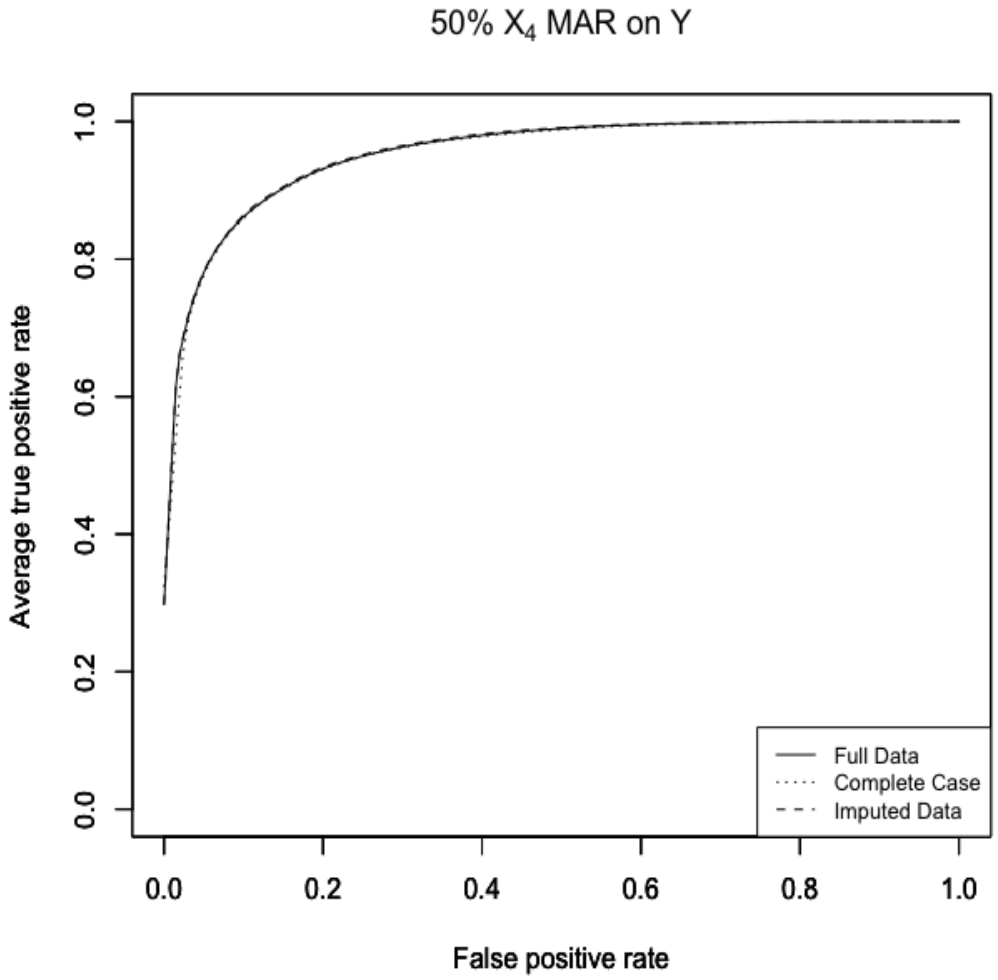

Figure A.3: ROC Curves for 50% X_4 MAR on Y - full data, complete case, and imputed data.

Table A.3: Missing Data Mechanism 2: $\hat{\beta}$ coefficients of the predictive model created by the training data for the full data, complete case, and imputed data for each missing variable and missing rates investigated.

Missing Variable	Scenario	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
	Truth	0	3	2	4	1
$\mathbf{X_2}$	**10% Missing**					
	Complete	-0.002	3.053	2.045	4.091	1.036
	Missing	0.113	2.955	2.050	4.360	1.035
	Imputed	-0.001	3.034	2.022	4.049	1.028
	30% Missing					
	Complete	-0.004	3.064	2.041	4.087	1.012
	Missing	0.489	2.891	2.076	4.673	1.039
	Imputed	-0.009	3.009	1.983	3.994	0.998
	50% Missing					
	Complete	-0.009	3.062	2.039	4.058	1.010
	Missing	0.935	2.943	2.116	4.954	1.085
	Imputed	-0.037	2.966	1.902	3.896	0.987
$\mathbf{X_4}$	**10% Missing**					
	Complete	0.002	3.059	2.049	4.082	1.021
	Missing	0.115	2.961	2.059	4.341	1.026
	Imputed	0.005	3.057	2.048	4.081	1.012
	30% Missing					
	Complete	-0.001	3.063	2.043	4.064	1.005
	Missing	0.487	2.911	2.086	4.688	1.041
	Imputed	0.007	3.064	2.044	4.063	0.988
	50% Missing					
	Complete	0.004	3.048	2.046	4.075	1.026
	Missing	0.961	2.934	2.130	5.000	1.092
	Imputed	0.032	3.048	2.044	4.064	0.959
$\mathbf{X_2}, \mathbf{X_4}$	**10% Missing**					
	Complete	0.002	3.065	2.039	4.081	1.017
	Missing	0.227	2.888	2.050	4.599	1.031
	Imputed	0.004	3.053	2.024	4.045	1.009
	30% Missing					
	Complete	0.002	3.066	2.052	4.083	1.011
	Missing	0.985	2.805	2.126	7.112	1.087
	Imputed	0.000	3.018	1.984	3.986	0.976
	50% Missing					
	Complete	NA	NA	NA	NA	NA
	Missing	NA	NA	NA	NA	NA
	Imputed	NA	NA	NA	NA	NA

Table A.4: Missing Data Mechanism 2: Confusion matrices displaying the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) used for calculating the sensitivity and specificity.

Missing Variable	Scenario	TP	FP	FN	TN	Sensitivity	Specificity
$\mathbf{X_2}$	**10% Missing**						
	Complete	123	15	15	46	0.891	0.754
	Missing	121	13	13	34	0.903	0.723
	Imputed	123	15	15	46	0.891	0.754
	30% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	107	8	9	17	0.922	0.680
	Imputed	122	16	16	46	0.884	0.742
	50% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	87	5	5	8	0.946	0.615
	Imputed	122	16	16	46	0.884	0.742
$\mathbf{X_4}$	**10% Missing**						
	Complete	123	15	15	46	0.891	0.754
	Missing	120	13	13	34	0.902	0.723
	Imputed	123	15	15	47	0.891	0.758
	30% Missing						
	Complete	123	15	15	46	0.891	0.754
	Missing	107	8	9	17	0.922	0.680
	Imputed	123	15	15	47	0.891	0.758
	50% Missing						
	Complete	123	16	15	46	0.891	0.742
	Missing	87	5	5	8	0.946	0.615
	Imputed	123	15	15	46	0.891	0.754
$\mathbf{X_2, X_4}$	**10% Missing**						
	Complete	123	15	15	47	0.891	0.758
	Missing	117	11	11	27	0.914	0.711
	Imputed	123	16	15	47	0.891	0.746
	30% Missing						
	Complete	123	15	15	47	0.891	0.758
	Missing	93	5	5	7	0.949	0.583
	Imputed	122	15	16	46	0.884	0.754
	50% Missing						
	Complete	NA	NA	NA	NA	NA	NA
	Missing	NA	NA	NA	NA	NA	NA
	Imputed	NA	NA	NA	NA	NA	NA

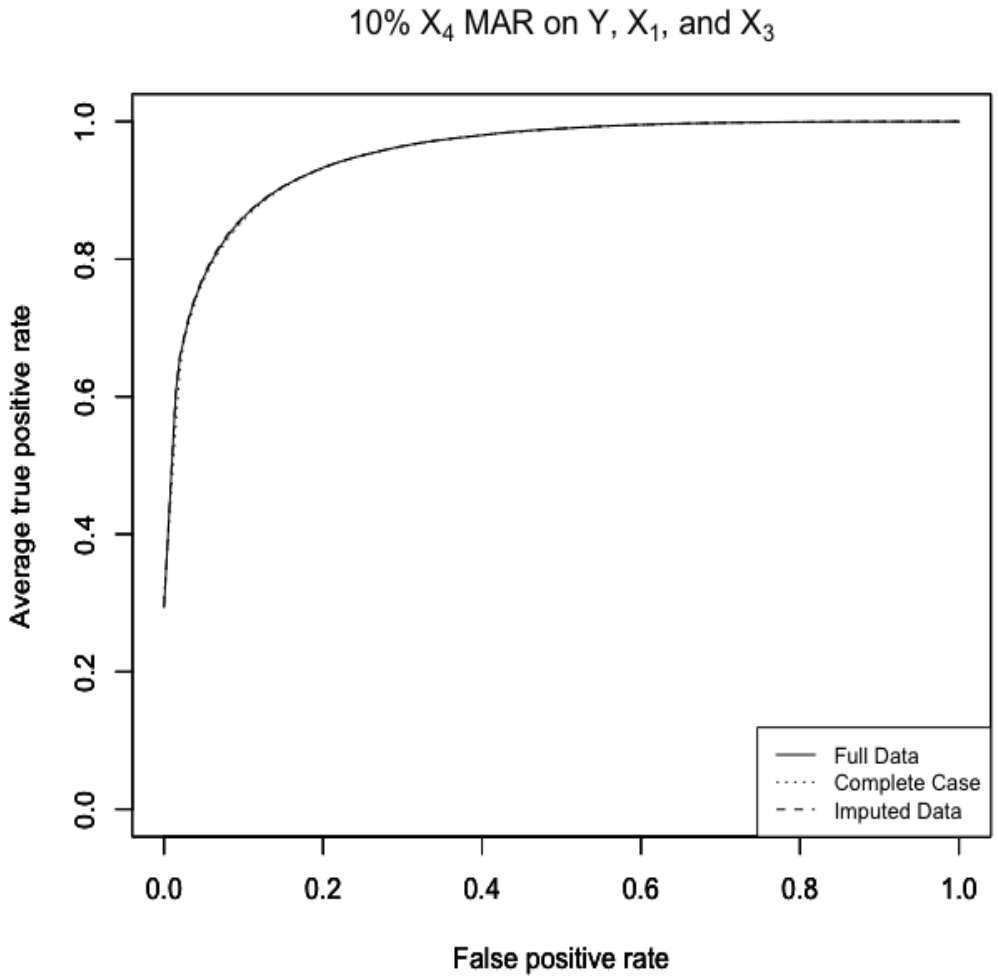

Figure A.4: ROC Curves for 10% X_4 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

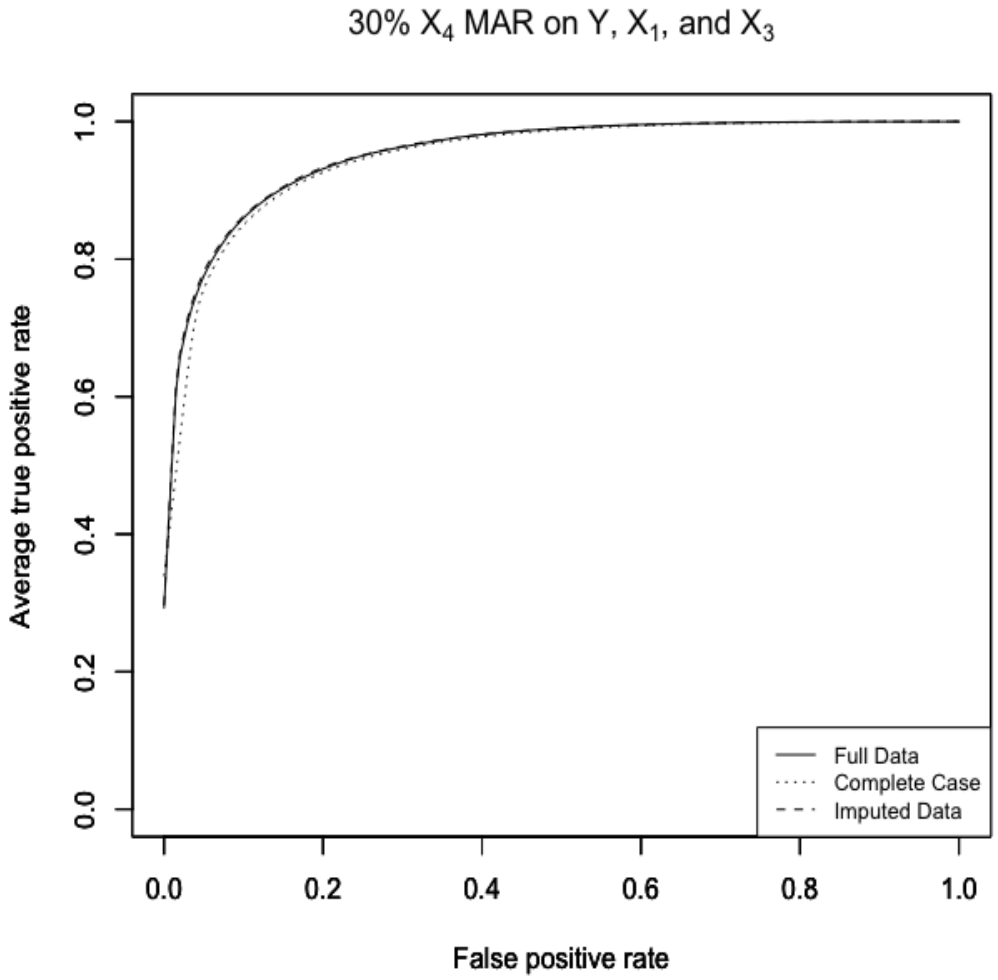

Figure A.5: ROC Curves for 30% X_4 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

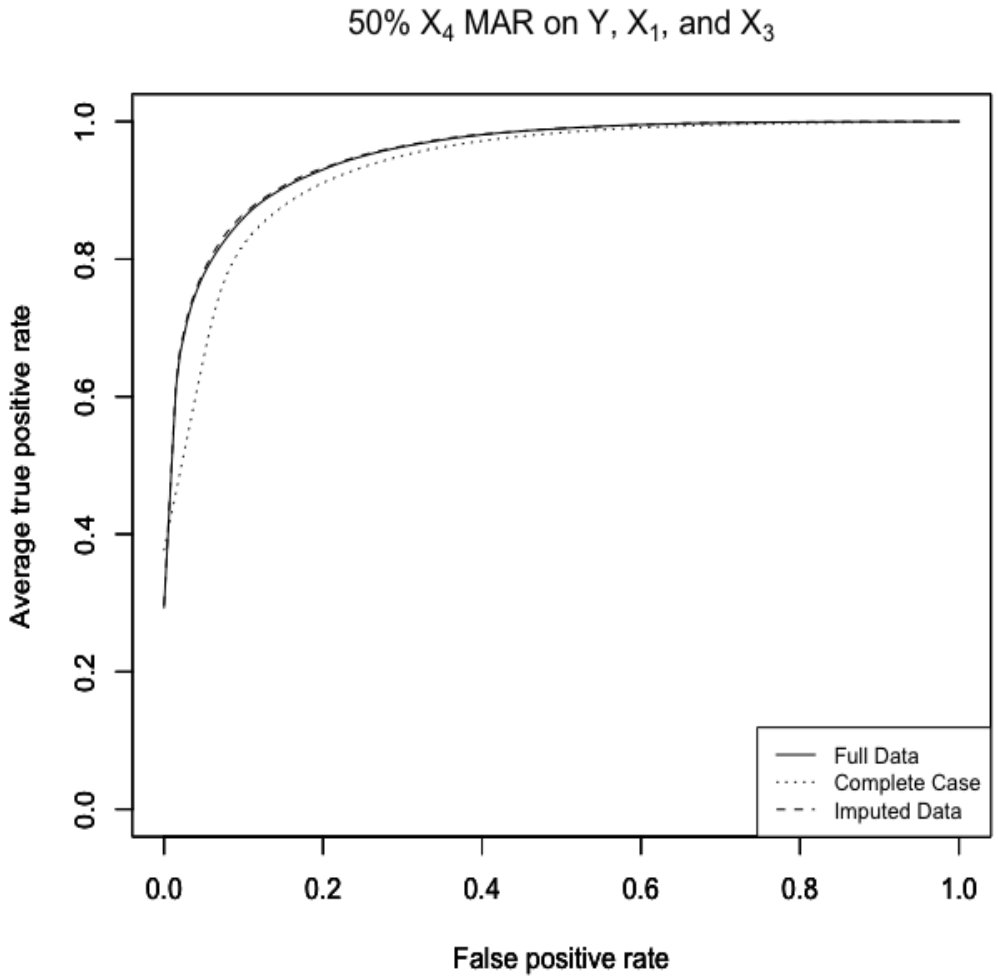

Figure A.6: ROC Curves for 50% X_4 MAR on Y, X_1, and X_3 - full data, complete case, and imputed data.

Appendix B

Additional Theory

B.1 provides detailed steps of the Gibbs Sampler used in continuous and binary variable imputation. B.2 explains the method of estimating coefficients in generalized linear models.

B.1 Gibbs Sampler

B.1.1 Continuous Variable Imputation

The following steps explain how to draw $\mathbf{Y}_M^r \sim f(\mathbf{Y}_M|\boldsymbol{\beta}^r, \boldsymbol{\Omega}^r, \mathbf{Y}_O)$:

1. For each unit $i = 1, ..., n$, re-order the variables such that $Y_{i,1}, ..., Y_{i,p_1}$ are observed and $Y_{i,p_1+1}, ..., Y_p$ are missing.

2. Re-order $\boldsymbol{\beta}$ and partition such that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ where $\boldsymbol{\beta}_1^T = (\beta_1, ..., \beta_{p_1})$ and $\boldsymbol{\beta}_2^T = (\beta_{p_1+1}, ..., \beta_p)$

3. Re-order $\boldsymbol{\Omega}$ and partition such that $\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{1,1} & \boldsymbol{\Omega}_{1,2} \\ \boldsymbol{\Omega}_{2,1} & \boldsymbol{\Omega}_{2,2} \end{pmatrix}$

4. Draw $\mathbf{Y}_{i,M} \sim \text{N}\{\boldsymbol{\beta}_2 + (\mathbf{Y}_{i,O} - \boldsymbol{\beta}_1)^T \boldsymbol{\Omega}_{1,1}^{-1} \boldsymbol{\Omega}_{1,2}, \boldsymbol{\Omega}_{2,2} - \boldsymbol{\Omega}_{2,1} \boldsymbol{\Omega}_{1,1}^{-1} \boldsymbol{\Omega}_{1,2}\}$

B.1.2 Binary Variable Imputation

Using a latent normal variable approach designed such that

$$\begin{aligned} \text{Pr}(Y_i = 1) = \text{Pr}(Z_i > 0) &= \text{Pr}(Z_i - \beta > -\beta) \\ &= 1 - \Phi(-\beta) \\ &= \Phi(\beta). \end{aligned} \tag{B.1}$$

is equivalent to regressing a binary variable $\mathbf{Y}$ on a constant in a probit model [5],

$$\Phi^{-1}\{\text{Pr}(Y_i = 1)\} = \beta, \quad i \ \epsilon \ (1, ..., n). \tag{B.2}$$

The fitted probabilities from probit and logit models are only slightly different so a probit model can be used to impute missing values and a logit model can still be fit. Therefore, the multivariate normal model can be used to model both the latent normal variables and continuous variables, with the restriction that the variance of the latent normal variable must be 1 [5].

Suppose there is a continuous variable Y_1 and binary variable Y_2. The joint model using a latent normal approach is

$$\begin{aligned} Y_{i,1} &= \beta_{0,1} + e_{i,1} \\ Z_{i,2} &= \beta_{0,2} + e_{i,2} \\ \text{where} \begin{pmatrix} e_{i,1} \\ e_{i,2} \end{pmatrix} &\sim N_2 \left[\mathbf{0}, \mathbf{\Omega} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & 1 \end{pmatrix} \right] \end{aligned} \tag{B.3}$$

and $Z_{i,2}$ is the latent variable associated with Y_2. Following the general outline of the Gibbs sampler in section 2.4.1, use the complete data to set initial values for $\beta_{0,1}^0, \beta_{0,2}^0$, and $\mathbf{\Omega}^0$. At iteration r of the Gibbs sampler [5],

1. For $i = 1, ..., n$, draw $\tilde{Z}_{i,2}$ from the conditional normal given $Y_{i,1}$,

$$\tilde{Z}_{i,2} \sim N\{\beta_{0,2} + (Y_{i,1} - \beta_{0,1})\mathbf{\Omega}_{1,1}^{-1}\mathbf{\Omega}_{1,2}, 1 - \mathbf{\Omega}_{2,1}(\mathbf{\Omega}_{1,1})^{-1}\mathbf{\Omega}_{1,2}\} \tag{B.4}$$

   If $Y_{i,2} = 1$ and $\tilde{Z}_{i,2} > 0$ or $Y_{i,2} = 0$ and $\tilde{Z}_{i,2} \leq 0$ then accept and set $Z_{i,2}^r = \tilde{Z}_{i,2}$. Otherwise draw a new $\tilde{Z}_{i,2}$.

2. Update elements of $\mathbf{\Omega}$ to obtain $\mathbf{\Omega}^r$ using a Metropolis Hastings algorithm.

a) Draw $\tilde{\mathbf{\Omega}}_{k,l}$ from a symmetric proposal distribution.

b) Ensure $\mathbf{\Omega}$ is positive definite when $\mathbf{\Omega}_{k,l}$ is updated with $\tilde{\mathbf{\Omega}}_{k,l}$. If not, draw a new $\tilde{\mathbf{\Omega}}_{k,l}$.

c) Accept $\tilde{\mathbf{\Omega}}_{k,l}$ with probability

$$\min\left(1, \frac{L(\boldsymbol{\beta}, \tilde{\mathbf{\Omega}}_{k,l}, \mathbf{\Omega}_{-k,l})p(\tilde{\mathbf{\Omega}}_{k,l}, \mathbf{\Omega}_{-k,l})}{L(\boldsymbol{\beta}, \mathbf{\Omega})p(\Omega)}\right) \tag{B.5}$$

where L is the bivariate normal likelihood,

$$L(\boldsymbol{\beta}, \mathbf{\Omega}) \propto |\mathbf{\Omega}|^{-n/2} \times \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Y_{i,1} - \beta_{0,1}, Y_{i,2} - \beta_{0,2})\mathbf{\Omega}^{-1}(Y_{i,1} - \beta_{0,1}, Y_{i,2} - \beta_{0,2})\right\}, \tag{B.6}$$

$\mathbf{\Omega}_{-k,l}$ refers to elements of $\mathbf{\Omega}$ excluding the $(k,l)^{th}$, and p(.) is the prior distribution for $\mathbf{\Omega}$. See Appendix B in *Multiple Imputation and Its Application* for more information [5].

d) If $\tilde{\mathbf{\Omega}}_{k,l}$ is accepted, set $\mathbf{\Omega}^r_{k,l} = \tilde{\mathbf{\Omega}}_{k,l}$. Otherwise, keep $\mathbf{\Omega}^r_{k,l} = \mathbf{\Omega}^r_{k,l}$.

3. Draw $(\beta^r_{0,1}, \beta^2_{0,2})$ from $N_2[(\bar{\mathbf{Y}}_1, \bar{\mathbf{Z}}^r_2), n^{-1}\mathbf{\Omega}^r]$

If data are missing, draw a value for each missing variable by sampling with replacement from the observed values of the corresponding variable. Step 1 of the Gibbs sampler is modified as follows [5]:

1. If binary $Y_{i,2}$ is missing, draw $Z^r_{i,2}$ from (2.24). If $Z^r_{i,2} > 0$ then set $Y^r_{i,2}$ = 1; otherwise, $Y^r_{i,2} = 0$.

2. If continuous $Y_{i,1}$ is missing, draw $Z^r_{i,2}$ from the marginal normal $N(\beta^{r-1}_0, 1)$. Follow the steps for accepting or rejecting this proposal. $Y^r_{i,1}$ is then drawn from the conditional normal given $Z^r_{i,2}$,

$$N\{\beta^{r-1}_{0,1} + (Z^r_{i,2} - \beta^{r-1}_{0,2})(\mathbf{\Omega}^{r-1}_{2,2})^{-}1\mathbf{\Omega}^{r-1}_{2,1}, \mathbf{\Omega}^{r-1}_{1,1} - \mathbf{\Omega}^{r-1}_{1,2}(\mathbf{\Omega}^{r-1}_{2,2})^{-1}\mathbf{\Omega}^{r-1}_{2,1}\} \tag{B.7}$$

B.2 Generalized Linear Model

B.2.1 Iteratively Reweighted Least Squares

For the iteratively reweighted least squares procedure, the dependent variable is a linearized form of the link function applied to y, denoted as z. To linearize $g(y)$, let $\eta = g(\mu)$ and $\mu =$E(Y) and perform a Taylor expansion such that [22]

$$\begin{aligned} g(y) &\approx g(\mu) + (y-\mu)g'(\mu) \\ &= \eta + (y-\mu)\frac{d\eta}{d\mu} \\ &\equiv z. \end{aligned} \tag{B.8}$$

The regression of z on $\mathbf{X}$ uses weights w_i that are inversely proportional to Var$(g(y))$

$$\widehat{\text{Var}}(z) = \left(\frac{d\eta}{d\mu}\right)^2 V(\mu) = \frac{1}{w} \tag{B.9}$$

in order to estimate the new estimates for $\boldsymbol{\beta}$ [22]. The process is iterative because the adjusted dependent variable z and the weights w_i are functions of the fitted values $\hat{\mu}$. The iteratively reweighted least squares procedures is then as follows [22]:

1. Set initial estimates for $\hat{\eta}_0$ and $\hat{\mu}_0$.
2. Form the adjusted dependent variable $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}$.
3. Form the weights $w_0^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 \Big|_{\hat{\eta}_0} V(\hat{\mu}_0)$.
4. Re-estimate β to get $\hat{\eta}_1$.
5. Iterate steps 2-4 until convergence.

The above steps represent how these parameters are estimated in statistical software packages. See *Generalized Linear Models* by McCullagh and Nelder for more details [22].

B.2.2 Logistic Regression Maximum Likelihood

The coefficients $\beta_0, \beta_1, ..., \beta_p$ are estimated such that the predicted probability π_i for each individual is as close as possible to their true outcome status. This is accomplished by maximizing the likelihood function [17], written as

$$L(\beta_0, \beta_1, ..., \beta_p) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{B.10}$$

This is transformed by taking the log of both sides to create the log-likelihood,

$$\begin{aligned} l(\beta_0, \beta_1, ..., \beta_p) &= \sum_{i=1}^{n} y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i) \\ &= \sum_{i=1}^{n} y_i \log(\pi_i) + \log(1-\pi_i) - y_i \log(1-\pi_i) \\ &= \sum_{i=1}^{n} \log(1-\pi_i) + \sum_{i=1}^{n} y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) \end{aligned} \tag{B.11}$$

Typically maximum likelihood problems are solved by differentiating the log-likelihood with respect to each β_{p+1} separately, setting it equal to 0, and then solving for the respective β. However, logistic regression does not have a closed form. Thus, it is best to follow the iteratively reweighted least squares method highlighted in the previous section.

Bibliography

[1] Baneshi M.R., T.A.R. (2012), "Does the missing data imputation method affect the composition and performance of prognostic models?" *Iranian Red Crescent Medical Journal*, 14.

[2] Bartlett, J.W., Harel, O. and Carpenter, J.R. (2015), "Asymptotically Unbiased Estimation of Exposure Odds Ratios in Complete Records Logistic Regression," *American Journal of Epidemiology*, 182, 730-736.

[3] Bernhard J., Cella D.F., Coates A.S., Fallowfield L., Ganz P.A., Moinpour C.M., Mosconi P., Osoba D., Simes J., Hürny C. (1998), "Missing quality of life data in cancer clinical trials: serious problems and challenges." *Statistics in medicine*, 17, 15.

[4] Bounthavong, M., Watanabe, J.H. and Sullivan, K.M. (2015), "Approach to Addressing Missing Data for Electronic Medical Records and Pharmacy Claims Data Research," *Pharmacotherapy*, 35, 380-387.

[5] Carpenter, J.R., Kenward, M.G. (2013), *Multiple imputation and its application*.

[6] Choi, Y., Nam, C. and Kwak, M. (2004), "Multiple imputation technique applied to appropriateness ratings in cataract surgery," *Yonsei medical journal*, 45, 829-837.

[7] Enders, C. K. (2010), *Applied missing data analysis*.

[8] Engels, J.M., Diehr, P. (2003), "Imputation of missing longitudinal data: a comparison of methods," *Journal of Clinical Epidemiology*, 56, 968-976.

[9] Fawcett, T. (2006), "An introduction to ROC analysis," *PATREC Pattern Recognition Letters*, 27, 861-874.

[10] Greenland, S. and Finkle, W. (1995), "A critical look at methods for handling missing covariates in epidemiologic regression analyses," *American Journal of Epidemiology*, 142, 1255-1264.

[11] Hallgren, K. A., Witkiewitz, K., Kranzler, H. R., Falk, D. E., Litten, R. Z., O'Malley, S. S., Anton, R. F. and Conjunction Alcohol Clinical Trial. (2016), "Missing Data in Alcohol Clinical Trials with Binary Outcomes," *Alcoholism-Clinical and Experimental Research*, 40, 1548-1557.

[12] Herring, A.H., Ibrahim, J.G., Lipsitz, S.R. (2004), "Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial," *Journal of the Royal Statistical Society.Series C, Applied Statistics*, 53, 293-310.

[13] Horton, N.J, Kleinman, K.P. (2007), "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, 61, 79-90.

[14] Horton, N. J., Lipsitz, S.R. (2001), "Multiple Imputation in Practice: Comparison of Software Pack- ages for Regression Models With Missing Variables," *The American Statistician*, 55, 244–254.

[15] Ibrahim, J.G., Chen, M., Lipsitz, S.R., Herring, A.H.. (2005), "Missing-Data Methods for Generalized Linear Models: A Comparative Review," *Journal of the American Statistical Association*, 100, 332-346.

[16] Ibrahim, J., Lipsitz, S. (1999), "Missing covariates in generalized linear models when the missing data mechanism is non-ignorable," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 61, 173-190.

[17] James, G. (2013), *An introduction to statistical learning : with applications in R*.

[18] Lipsitz, S., Parzen, M., Natarajan, S., Ibrahim, J. and Fitzmaurice, G. (2004), "Generalized linear models with a coarsened covariate," *Journal of the Royal Statistical Society Series C-Applied Statistics*, 53, 279-292.

[19] Little, R.J.A. (1988), "A Test of Missing Completely at Random for Multivariate Data with Missing Values," *Journal of the American Statistical Association*, 83, 1198-1202.

[20] Little, R.J.A, Rubin, D.B. (2002), *Statistical analysis with missing data*.

[21] Masconi, K.L., Matsha, T.E., Erasmus, R.T., Kengne, A.P. (2015), "Effects of Different Missing Data Imputation Techniques on the Performance of Undiagnosed Diabetes Risk Prediction Models in a Mixed-Ancestry Population of South Africa." *PloS one*, 10.

[22] McCullagh, P., Nelder, J.A. (1983), *Generalized linear models*, London; New York: Chapman and Hall.

[23] Novo, A.A (Ported to R). Original by Schafer, J.L. (2013). norm: Analysis of multivariate normal datasets with missing values. R package version 1.0-9.5.

[24] Peng, L., Lei, L., and Naijun, W. (2005). "A quantitative study of the effect of missing data in classifiers," *Fifth International Conference on Computer and Information Technology*, 28–33.

[25] Plaia, A., Bondì, A.L. (2006), "Single imputation method of missing values in environmental pollution data sets," *AEA Atmospheric Environment*, 40, 7316-7330.

[26] Quartagno, M., Carpenter, J. (2016), jomo: A package for Multilevel Joint Modelling Multiple Imputation.

[27] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[28] Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-592.

[29] Rubin, D. B. (1987), *Multiple imputation for nonresponse in surveys*, New York: Wiley.

[30] Rubin, D. (1996), "Multiple imputation after 18+ years," *Journal of the American Statistical Association*, 91, 473-489.

[31] Saar-Tsechansky, M.,Provost, F. (2008), "Handling Missing Values when Applying Classification Models," *Journal of machine learning research : JMLR.*, 8, 1623-1658.

[32] SAS Institute Inc. 2011. *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

[33] Schafer JL. (1999), "Multiple imputation: a primer." *Statistical methods in medical research*, 8, 3-15.

[34] Schafer JL, Graham, J.W. (2002), "Missing data: our view of the state of the art." *Psychological methods*, 7, 147-77.

[35] Stoltzfus, J.C. (2011), "Logistic Regression: A Brief Primer," *Academic Emergency Medicine*, 18.

[36] van Buuren, S., Groothuis-Oudshoorn, K. (2011), mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67.

[37] van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T. and Moons, K. G. M. (2006), "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example," *Journal of clinical epidemiology*, 59, 1102-1109.

[38] Williams, D., Liao, X., Xue, Y. and Carin, L. (2005), "Incomplete-Data Classification Using Logistic Regression," In *Proceedings of the 22nd International Conference on Machine Learning (ICML)*.

[39] Zhong, H. (2010), "The impact of missing data in the estimation of concentration index: a potential source of bias," *The European Journal of Health Economics : HEPAC*, 11, 255-66.

[40] Zhou, X., Zhou, C., Liu, D. ,Ding, X. (2014), *Applied missing data analysis in the health sciences*.