January 2018

Missing Data in Complex Sample Surveys: Impact of Deletion and Imputation Treatments on Point and Interval Parameter Estimates

Anh Pham Kellermann
University of South Florida, napham@mail.usf.edu

Missing Data in Complex Sample Surveys: Impact of Deletion and Imputation Treatments on Point and Interval Parameter Estimates

by

Anh Pham Kellermann

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in Curriculum and Instruction with an emphasis in
Educational Measurement and Research
Department of Educational and Psychological Studies
College of Education
University of South Florida

Major Professor: John Ferron, Ph.D
Eun Sook Kim, Ph.D.
Yiping Lou, Ph.D.
Yi-hsin Chen, Ph.D.

Date of Approval
March 20 2018

Keywords: MCAR, MAR, MNAR, two-level data

Dedication

To the memory of my parents, who instilled in me the love and respect for knowledge, and who provided me with the foundation to reach far and beyond my dreams. Father and mother, I always want to make you proud.

To my beloved husband, who encouraged and supported me to pursue my dreams. Thank you, husband, for taking on my work and home responsibilities so that I could fulfill the heavy demands of my academic work. Thank you for believing in me and for building up my confidence to achieve my goals.

Acknowledgments

This dissertation would not have been possible without the guidance and professional support of many individuals from the University of South Florida and the immense physical and moral support of my extended family members.

I would like to express my very great appreciation to Dr. John Ferron, my major professor, for his critically valuable and constructive suggestions from the preparation throughout the completion of this dissertation. Dr. Ferron's food for thought and critical feedback have significantly enriched the final version of this dissertation. I also greatly appreciate Dr. Ferron for making my committee transition go smoothly and easing my anxiety during the writing/reviewing process with his very thoughtful and inspiring advice. I thank my committee members, Dr. Eun Sook Kim and Dr. Yiping Lou who have inspired me with intriguing and critical questions during the proposal and final defense process, helping improve the quality of the study and the clarity of the dissertation paper. I particularly wish to thank Dr. Yi-hsin Chen, my committee member, for his encouragement and valuable advice since my first year in the program and throughout multiple collaborative studies that I had the opportunity to work with him. I appreciate Dr. Chen's friendliness and especially his responsiveness to my quests for advice. I would like to express my deep gratitude to Dr. Jeffrey Kromrey, my retired major professor and my source of wisdom, knowledge, and inspiration during my years of doctoral studies. Forever I am indebted to him for his immense support and patient guidance leading me to achieving my academic goal.

I am very grateful for the extraordinary support given by the knowledgeable and exceptionally responsive USF Research Computing support staff, especially John DeSantis, who made an extra effort to quickly update the needed software version required for my simulation project, and who repeatedly and promptly trouble shot memory and storage issues during the simulation process.

I owe my deepest gratitude to my big sister and her husband, Diep and Frank Evener, who provided me unconditional support throughout my doctoral life; who took me to and picked me up from the airport on all of my conference trips over the years and provided me with emotional support, holding me up through the most difficult times for my family while I was in this doctoral program. My loving thanks are also extended to my caring younger sister, LeAnh, and my wonderful nephew, Alex, who have been my best travel companions to conferences during my years of study; who kept me company and looked after me while we were away from home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this simulation study was to evaluate the relative performance of five missing data treatments (MDTs) for handling missing data in complex sample surveys. The five missing data methods included in this study were listwise deletion (LW), single hot-deck imputation (HS), single regression imputation (RS), hot-deck-based multiple imputation (HM), and regression-based multiple imputation (RM). These MDTs were assessed in the context of regression weight estimates in multiple regression analysis in complex sample data with two data levels. In this study, the multiple regression equation had six regressors without missing data and two regressors with missing data. The four performance measures used in this study were statistical bias, RMSE, CI width, and coverage probability (i.e., 95%) of the confidence interval.

The five MDTs were evaluated separately for three types of missingness: MCAR, MAR, and MNAR. For each type of missingness, the studied MDTs were evaluated at four levels of missingness (10%, 30%, 50%, and 70%) along with complete sample conditions as a reference point for interpretation of results. In addition, ICC levels (.0, .25, .50) and high and low density population were also manipulated as studied factors.

The study's findings revealed that the performance of each individual MDT varied across missing data types, but their relative performance was quite similar for all missing data types except for LW's performance in MNAR. RS produced the most inaccurate estimates considering bias, RMSE, and coverage of confidence interval; RM and HM were the second poorest performers. LW as well as HS procedure outperformed the rest on the measures of accuracy and

precision in MCAR; however LW's measures of precision decreased in MAR and MNAR, and LW's CI width was the widest in MNAR data. In addition, in all three missing data types, those poor performers were less accurate and less precise on variables with missing data than they were on variables without missing data; and the degree of accuracy and precision of these poor performers depended mostly on the level of data ICC. The proportion of missing data only noticeably affected the performance of HM such that in higher missing data levels, HM yielded worse performance measures. Population density factor had negligible effects on most of the measures produced by all studied MDTs except for RMSE, CI width, and CI coverage produced by LW which were modestly influenced by population density.

CHAPTER ONE: INTRODUCTION

Background

Missing Data

Missing data are ubiquitous in quantitative social research. Just about every social analyst/researcher faces the problems of missing data especially in social survey research. Missing data occur at two levels: at the unit level, known as total or unit nonresponse, when no information is collected from a respondent or at the item level, known as item nonresponse, when respondents fail to respond to one or more of the survey items. Some sources of unit nonresponse are refusals to participate in the survey, language barrier, or illness. Some sources of item nonresponse are participants' refusal to answer a sensitive question or a question about behaviors perceived as socially undesirable; participants do not understand the question in a questionnaire or unintentionally skip a question or lose interest in a long questionnaire. Item nonresponse could also occur because of data collectors' activities during editing, recording, or storing processes.

Potential consequences of missing data include reducing the achieved sample size (hence decreasing statistical power) and reducing the precision and accuracy of estimates of parameters. Loss of accuracy leading to drawing wrong conclusions or biased inferences about outcomes and relationships of interest; and decreasing the precision of estimates results in degrading the performance of confidence intervals (e.g., wider confidence intervals on the estimates), and increasing standard errors.

To mitigate the inevitable problems of missing data, many methods have been developed to treat missing data. This dissertation studied the effectiveness of multiple widely-used methods that treat missing data resulting from item nonresponse in complex sample survey data. Followings are brief descriptions of (1) the features of complex sample surveys, (2) different sources leading to missing data (i.e., missing data mechanisms), (3) missing data treatments (MDT) based on three different strategies of deletion, single imputation, and multiple imputation, and (4) rationale for this dissertation study.

Complex Sample Survey

Data from complex sample surveys differ from those obtained via simple random sampling in several ways that impact how statistical analyses should be conducted. The key feature of a complex sample is that sample observations do not have equal probability of being selected because sampled units are sampled from clusters which are usually sampled within strata which often have different density levels; this leads to the need to incorporate sample weights. In addition, the multi-stage sampling feature also complicates the estimate of sampling error because variance among units within each cluster is smaller than the variance among units across clusters. This leads to the need to consider the homogeneity within clusters, measured by the intraclass correlation coefficient (ICC), in the analysis. These features are discussed in more detail in Chapter II.

Missing Data Mechanism

Missing data can happen to one or more measured variables that are used as predictors, covariates, and/or outcomes. Missing data in a sample survey can be unit non-response or item

non-response. One of the major contributions to missing-data literature is the differentiation of the underlying reasons for missing data which is referred to as missing data mechanisms. Rubin (1976) and colleague (Little & Rubin, 2002) identified three types of missing data mechanisms concerning the relationship between missingness and the values of variables in the dataset. According to Little and Rubin's theoretical framework, the three missing-data mechanisms that lead to missing data are (1) missing completely at random (MCAR), in which the cause of the missing data is unrelated to the values of any variables, whether missing or observed; (2) missing at random (MAR), in which the cause of the missing data is unrelated to the missing values, but may be related to the observed values of other variables; and (3) missing not at random (MNAR), in which the cause of the missing data is related to the missing values. Missing data mechanisms are applied to item non-response. Good understanding of these missing data mechanisms is essentially important in treating missing data as the performance of the missing data treatment methods depends on assumptions about the missing data mechanism.

Missing Data Methods

Two broad strategies to deal with missing data are elimination and imputation; elimination procedures eliminate observations with missing data from the analysis, and imputation procedures replace the missing values with estimates to create a complete dataset that can then be analyzed with traditional analysis methods. The traditional listwise deletion technique, which excludes observations with any missing variable values from the analysis, is the default setting in many statistical procedures. The benefit of listwise deletion is simplicity; however, the cost of this simplicity is information loss from those incomplete cases which are

left out from the analysis. To reduce the information loss from data elimination, researchers replace the values of the missing data with imputed values.

Two common approaches for the imputation strategy are hot-deck imputation and regression imputation. In hot-deck imputation, the value assigned for a missing item is taken from respondents in the current sample; the observation unit that contains the missing values is known as the recipient unit, and the observation unit that provides the value for imputation is known as the donor unit. Traditionally, the missing values are imputed one time resulting in one complete dataset for analysis.

In regression imputation, the value of the missing data variable is predicted using a regression model based on the relationship between that missing variable and other variables in the sample dataset. Regression imputation also fills missing values with predicted values generated from a regression equation one time resulting in one complete dataset for analysis. The regression imputation and hot-deck imputations as mentioned above are known as single imputation in which a missing value is substituted with a value; then the filled-in dataset is analyzed; (in this paper the regression imputation as mentioned here is referred to as regression imputation, and the hot-deck method is referred to as hot-deck imputation as opposed to regression multiple imputation and hot-deck multiple imputation, which will be discussed below). Major drawbacks of single imputation are that it does not reflect the uncertainty about the predictions or the imputed value of the unknown missing values, and the resulting estimated variances of the parameter estimates will be biased toward zero.

Rubin (1978) suggested a multiple imputation (MI) approach in which each missing value is replaced with a set of plausible values that represent the uncertainty about the right value to impute; the multiply imputed data sets are then analyzed by using standard procedures for

complete data; and the results from different imputed dataset are combined. Multiple imputation methods address the issue of uncertainty about the imputed value in single imputation. Any single imputation method can be used to obtain the imputed values in multiple imputation, but hot-deck imputation and regression imputation are commonly implemented in multiple imputation. In this paper, the hot-deck based multiple imputation is referred to as hot-deck multiple imputation and the regression-based multiple imputation as regression multiple imputation.

These missing data methods assume certain missing data mechanism, and their effects are influenced by these sources of missingness. The effects of missing data methods are also influenced by the complicated structure of complex sample survey data caused by differences in population density and by differences in variance among sampled unites within clusters compared to the one across clusters, and by the amount of missing data.

Rationale

The rationale of this study is derived from multiple sources. First, listwise deletion is widely used, and it is based on the assumption that the data are MCAR; the use of listwise deletion has been strongly discouraged (King et al., 1998, 2001; Wilkinson & Task Force on Statistical Inference, 1999; Cranmer & Gill, 2012) for causing substantial loss of information leading to insufficient power and biased estimates. However, none of the commonly-used methods is found to be clearly superior to listwise deletion (Allison, 2002), and various recent studies revealed that there are conditions where listwise deletion outperforms or at least performs as well as the more advanced imputation methods such as MI under different missing data mechanisms (Kellermann et al., 2016; Van Kuijk et al., 2016; White & Carlin, 2010). In the

modern age of big data, statistical power may not be a big concern for listwise deletion method and given its simplicity and high accessibility (e.g., default setting in most statistical software packages), applied researchers may be even more interested in findings about listwise deletion's efficiency compared to sophisticated alternatives in different theoretical scenarios, especially in the context of complex sample survey data under MAR or MNAR conditions.

Secondly, hot-deck imputation is the most commonly used imputation technique for survey data; but for situations where the data are not missing at random (i.e., MNAR), hot-deck imputation methods have not been well explored, and comparisons with alternative methods are limited (Andridge & Little, 2010), especially its comparisons with MI which is one of the most accepted methods among applied researchers to deal with missing data; also since MNAR data are not testable using sample data, it is of practical importance to provide applied researchers with information about the sensitivity of these missing data methods to this non-ignorable missingness.

Thirdly, although there are abundant studies comparing the performance of multiple imputation and other traditional imputation methods, existing missing data literature seems to lack empirical studies comparing the effect of multiple imputation based on different imputing approaches such as hot-deck based versus regression-based multiple imputation. A comparison of the sensitivity of these two methods under MNAR data and in complex sample survey data is much needed as the usage of complex sample survey data is becoming more common because of the usefulness of secondary data analysis using nationally representative surveys. This study seeks to fill this knowledge gap in literature and to provide practicing researchers and survey practitioners with practical utility of these missing data methods.

Summary

Problems

The accuracy and precision of the estimates of statistical parameters directly depend on the quality of underlying data; however, missing data is a common problem in quantitative social research studies, especially in social survey research, there are so many potential reasons for missing data. While the effect of a missing data method mainly depends on the cause of missing data, it is not possible to identify if missingness is MAR or MNAR. As such, empirical studies on the effect of missing data treatment methods for different types of missingness especially in complex sample survey data could help survey researchers to select a proper tool to handle their missing data problems posed in complex data structures. In addition, the multiple imputation methods based on both hot-deck imputation and model-based imputation are becoming popular among applied researchers, but information about application and effectiveness of these advanced methods relative to those traditional methods for different missing data mechanisms is still limited.

Purpose

The purpose of this study is to examine the effectiveness of five different MDTs for handling missing data in complex sample survey in the context of estimating points and intervals of parameters (i.e., regression coefficients) in multiple regression analysis. The five missing data methods include one deletion method: listwise deletion (LW), two single imputation methods: hot-deck imputation (HS) and regression imputation (RS), and two multiple imputation methods: hot-deck multiple imputation (HM) and regression multiple imputation (RM). This present study is a replication and extension of a previous study with a similar purpose conducted by Kellermann,

Trevathan, and Kromrey (2016). In this past study, the authors examined the effects of listwise deletion and regression-based multiple imputation methods on point and interval estimates in the context of complex sample survey analysis when data are MCAR and MAR. This present study expanded the scope of the earlier study by (1) examining three additional missing data methods: hot-deck imputation, regression imputation, and hot-deck based multiple imputation, and (2) examining the missing data methods under an additional type of missingness, MNAR. The findings from this present study will help equip researchers with information needed to handle missing data appropriately and adequately; also, based on these findings, areas for improvement can be identified and addressed by future researchers. The uniqueness of this study is that it investigated the relative performance of popular MDTs in treating missing data at level 1 of multiple level data for MNAR data.

Research Questions

1. For parameter estimates in a complex sample survey dataset, which missing data treatment method (LW, HS, HM, RS, or RM) produces more accurate and precise estimates in terms of bias and RMSE for point estimates, and CI width and CI coverage for interval estimates when data are MCAR, MAR, and MNAR?
2. For each missing data type (MCAR, MAR, MNAR) how do amount of missing data, Intraclass Correlation (ICC), and population density influence the accuracy and precision of parameter estimates produced by each missing data treatment method of LW, HS, HM, RS, and RM?

Limitations and Delimitations

This present study assessed the effects of the application of different MDTs on the estimation of regression weights using multiple regression on complex sample survey data. One limitation of this study is the simplicity of the missing data; because of the complexity of the simulation study, only two variables present missing data and they are always missing together. In field research, it is likely that data will be missing for more than two variables and the missing variables may not be missing together.

Also, other limits to the generalizability of this dissertation study are that the sample sizes are not generalized to simple random survey with small sample size; variables are limited to continuous data; and statistical analysis is limited to estimating regression weights in multiple regression analysis and focuses only on the point estimates of regression coefficients while researchers may be interested in mean, correlation, or variance. In addition, the results of this present study are limited to two-level data structures as two-level data are typically used in educational research; however, in real practices, there are various theoretical and practical reasons for combining more than two levels of data.

Regarding the methods of treating missing data, additional research may be needed to empirically compare the results obtained from the MDTs examined in this study to those obtained from the maximum likelihood methods which is one of the most current missing data treatment methods and the predictive-mean-matching-based MI, which was proposed to be used in MI (Little, 1998); predictive-mean-matching-based MI is an attractive way to do multiple imputation for missing data, especially for imputing quantitative variables that are not normally distributed (Allison, 2015).

Definitions of Frequently Used Terms

Definitions of terms frequently used throughout this paper are provided below in alphabetical order.

Complex Sample Survey Data: Data from complex sample surveys differ from those obtained via simple random sampling in several ways that impact how statistical analyses should be conducted. The main features of complex survey sample are stratification (e.g., population is divided into sub-populations, i.e., strata, and independent samples are then drawn from within each stratum), clustering sampling (e.g., sampling groups of individual units rather than the individual units themselves), and multistage sampling in which individual observations are sampled at the lowest stage, and at every stage above the lowest stage, clusters of observations are sampled.

Hot-deck Imputation: Hot-deck imputation is a method for handling missing data; it involves replacing missing values of one or more variables for a nonrespondent with observed values from a respondent that is similar to the non-respondent with respect to characteristics observed in both cases (Andridge & Little, 2010).

Hot-deck Multiple Imputation: In this present study, hot-deck multiple imputation is used to refer to a form of multiple imputation in which the imputation process is performed using a hot-deck technique. It was referred to as multiple hot-deck imputation in Cranmer and Gill's (2012) study.

Listwise Deletion: Listwise deletion is a method for handling missing data; it is accomplished by deleting from the sample any observations that have missing data on any variables in the model of interest and then applying conventional methods of analysis for complete data sets (Alisson, 2002). Listwise deletion is also known as complete case analysis.

Missing at Random (***MAR)***: When the data are missing at random, the cause of the missing data is unrelated to the missing values, but may be related to the observed values of other variables (Little & Rubin, 2002). For example, people with high education tend not to disclose their income.

Missing Data Mechanisms: The underlying reasons for missing data are referred to as missing data mechanisms which include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (see Rubin 1976; Little & Rubin, 2002).

Missing Completely at Random (MCAR): When the data are missing completely at random, there is no relationship between whether a data point is missing and any values in the data set, missing or observed (Little & Rubin, 2002); for example, surveyees accidentally skip one or more items in a questionnaire.

Missing not at Random (MNAR): When the data are missing not at random, there is a relationship between the probability of a value to be missing and its values (Little & Rubin, 2002). For example, people with high income tend not to reveal their income.

Multiple Imputation (MI): Multiple imputation is an advanced method to treat missing data; MI replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute; the multiply-imputed data sets are then analyzed by using standard procedures for complete data; and the results from these analyses are combined (Rubin 1978).

Regression Imputation: In regression imputation, the value of the missing data variable is predicted using a regression model. The regression model is built based on the relationship between the missing variable and other variables in the sample dataset; and the outcome of the

regression equation is the variable for which missing values are to be imputed. (Brick & Kalton, 1996).

Regression Multiple Imputation: Regression multiple imputation is a form of multiple imputation in which the imputation step is performed using regression technique.

Single Imputation: In single imputation, a missing value is imputed once resulting in one filled-in dataset as opposed to multiple imputation, where a missing value is replaced with a set of plausible values that represent the uncertainty about the right value to impute, resulting in multiply-imputed data sets. In this paper, regression imputation and hot-deck imputation are single imputation whereas regression multiple imputation and hot-deck multiple imputation are multiple imputation.

CHAPTER TWO: LITERATURE REVIEW

Complex Sample Surveys

The main features of complex survey sample are stratification, clustering sampling, and multistage sampling. In stratified sampling, the target population is classified into non-overlapping sub-populations or strata. The subpopulations are divided such that the characteristics of the population are homogenous within each stratum and heterogeneous between strata; independent samples are then drawn from within each stratum to ensure that each stratum is represented in the sample. For example, in sampling high school students, the country is divided by regions, and within each region, schools are randomly selected.

Cluster sampling involves sampling groups of individual units (rather than the units themselves) by first breaking the population into groups or clusters which are then randomly selected; then all individuals in the selected clusters are measured (e.g., high school students in a state are sampled by randomly selecting some schools in the state). In a complex sample design, clusters are usually sampled within strata. Sampled units from clusters sampled from within strata often have different probability of selection due to differences in strata's density levels; for example, sampled units from clusters sampled from heavy density strata will have lower probability of selection than the ones from light density strata.

In multistage sampling, individual observations are sampled at the lowest stage, and at every stage above the lowest stage, clusters of observations are sampled; for example, a survey of school children may be done by sampling regions, then schools within regions, and finally

children within schools. As a result of these features, data from complex sample surveys are different from data from simple random samples in several ways that need to be taken into consideration in analyzing complex sample data.

One of the most important considerations in analyzing data obtained from a complex sample survey is the sample weight. Sample weight is the inverse of the probability of selection; for example, if 100 of 1500 students of a school are selected to participate in a survey, the sampling fraction or the probability of being selected would be 1/15, and the raw weight would be 15; each student of the sample can be thought of as representing 15 students from the school. If the design was not stratified, the weights are exactly the same for all students in the sample. However, in complex sampling the sampled members do not have equal probability of being selected into the sample because they are being sampled within strata as mentioned above regarding cluster sampling. Units sampled from clusters in low density strata will have a higher probability of selection hence lower weights than units sampled from clusters in high density strata.

Variance is another important consideration in analyzing complex sample data. Multi-stage sampling yields clustered observations in which the variance among units within each cluster is smaller than the variance among units across clusters because observations from the same cluster will likely be more similar to each other than to those from a different cluster; and this complicates the estimation of sampling error. The homogeneity within clusters is measured by the intraclass correlation coefficient (ICC). A small ICC can increase the variance of estimates substantially if the cluster sizes are large (Green et al., 2012).

Stratification in the sampling design, while insuring appropriate sample representation on the stratification variable(s), yields negatively biased estimates of the population variance when

not considered in the analysis. The following section discusses the computation of sample weights in complex sample surveys.

Sample Weights

The weight most frequently used in survey data analysis is an expansion weight (Dargat & Hill, 1996; Lee et al., 1989). An expansion weight, known as a raw weight or base weight, is the inverse of the selection probability for the sampled unit. For simple random sampling

$$w_i = \frac{1}{p_i}$$

Where, w_i is the raw weight of unit *i* and p_i is the probability of selection for unit *i.* For multi-stage sampling, the raw weight is the inverse of the product of the probabilities of selection at each stage

$$w_i = \frac{1}{p_{1i} \times p_{2i}}$$

P_{1i} = probability of selection at stage 1
P_{2i} = probability of selection at stage 2

With raw weights, the sum of the weights equals the population size. That is

$$\sum_{k=1}^{s} \sum_{j=1}^{p_k} \sum_{i=1}^{n_{jk}} w_{ijk} = N$$

where w_{ijk} is the weight assigned to the i^{th} individual in the j^{th} primary sampling unit (PSU) of the k^{th} strata in a study with *s* strata where the k^{th} strata has p_k PSUs and N is the population size.

In the estimation process, the weight is used as a frequency count from the data item observed. A weight of 2 means that the case counts in the dataset as two identical cases; and a weight of 1 means that the case only counts as one case in the dataset. The sample mean is computed as:

$$\bar{X} = \frac{\sum_{k=1}^{s}\sum_{j=1}^{p_k}\sum_{i=1}^{n_{jk}} w_{ijk} X_{ijk}}{\sum_{k=1}^{s}\sum_{j=1}^{p_k}\sum_{i=1}^{n_{jk}} w_{ijk}}$$

where X_{ijk} is the value for the i^{th} individual in the j^{th} PSU of the k^{th} strata, and other symbols are defined as they were previously. Weights are used to compensate for unequal probability of selection to make statistics computed from the data more representative of the population. In order to generate unbiased population estimates and draw correct inferences, it is necessary to weight the sample data during the summarization process before the analysis. Findings are generally not representative of the larger population of interest if sample weights are not used. In this dissertation study, sample weights were computed during the data simulation process based on the probability of selection at stage 1 and the probability of selection at stage 2 which were obtained during Level-1 data simulation and Level-2 data simulation respectively. The computed sample weights were utilized when the simulated samples were analyzed.

Unequal probability of selection in complex sample survey leads to unequal weights which necessitate special techniques for variance estimation; this topic will be briefly discussed in the following section.

Variance Estimation

Information about the precision of point estimates (e.g., sample means or regression coefficients) of a sample survey is based on variance which is estimated from the same sample. Variance is used to construct confidence intervals around the point estimates to give some indication of the degree of precision associated with the estimates; variance is also used in hypothesis tests using the survey data. However, for complex sampling designs, variance

estimation is not straightforward. Multiple variance estimation techniques such as Taylor series linearization, or replication based methods like balanced repeated replications, jackknife, and bootstrap methods have been used to analyze data from complex sample surveys (Skinner, Holt, & Smith, 1989); and results of several studies comparing the Taylor series linearization, balanced repeated replications, and jackknife methods have shown that these methods produce variance estimates that are similar, and that no method for variance estimation is better than another in all situations (Rodgers-Farmer & Davis, 2001; Cohen, 1997; LaVange et al., 1996; Shah et al., 1997.)

Taylor series linearization is used in many statistical applications to obtain approximate values of functions that are difficult to calculate precisely. Because most statistical estimates from complex sample surveys are not simple linear functions of the observations, a Taylor series expansion may be used to obtain an approximation of the estimate based on the linear (first-order) part of the Taylor series. The variance of this approximation may then be used to estimate the variance of the original statistic of interest. The Taylor series approach tends to be computationally fast (in comparison with replication methods) but carries the limitation that a separate formula must be developed for each estimate of interest.

As an example, consider estimating the variance of the sample mean. Graubard and Korn (1996) show that Taylor linearization leads to

$$\hat{\sigma}^2_{\bar{X}} = \frac{\sum_{k=1}^{s} \frac{p_k}{p_k - 1} \sum_{j=1}^{p_k} \left[W_{jk}(\bar{X}_{jk} - \bar{X}) - \frac{1}{p_k} \sum_{t=1}^{p_k} W_{tk}(\bar{X}_{tk} - \bar{X}) \right]^2}{\left(\sum_{k=1}^{s} \sum_{j=1}^{p_k} W_{jk} \right)^2}$$

where W_{jk} is the sum of the weights in the jth PSU in the kth strata, $\bar{X}_{jk}$ is the mean for the jth PSU in the kth strata, and the other symbols are as previously defined.

Missing Data Mechanisms

In survey research, there is a variety of reasons data may be missing. McKnight (2007) suggested it could be related to the study participants, such as some participants were offended by certain questions on a survey; it could be related to the study design like a study required too much of the participants' time or the question is ambiguous or items are missing by design (e.g., branching surveys skip some items for some respondents based on previous responses); or it could be due to interaction of the participants and the study design, for example, participants who were the sickest were unable to complete the more burdensome aspects of the study. Missing data mechanisms describe relationships between measured variables and the probability of missing data, providing different explanations for why the data are missing.

The theoretical literature on missing data mechanisms was developed by Rubin (1976). The concept was later further elaborated by Little and Rubin (2002) by using terminology that is commonly used in the modern statistics literature on missing data. Missing data mechanism classifies three different reasons why the data are missing: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). This section covers the conceptual description, features of each mechanism, and a brief description of how social researchers diagnose missing data mechanisms.

MCAR

Define the complete data $Y= (y_{ij})$, where y_{ij} is the value of variable j for subject i, and the missing-data indicator matrix $M=(m_{ij})$, such that $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is present. The missing-data mechanism is characterized by the conditional distribution of M given Y. If missingness does not depend on the values of the data Y, either missing or observed, the data are called MCAR. In terms of probabilistic explanations, the missing data mechanism is called MCAR when the probability of missing data on a variable Y is not related to the value of Y itself or to any other variables in the data set; for example, survey respondents accidentally skipped some items or some items were answered with illegible handwriting. When the missing data are missing at random and the observed data are observed at random, ignoring the process that causes missing data when making sampling distribution inference about the data parameters is appropriate (Rubin, 1976). MCAR is said to be ignorable, meaning the mechanism does not need to be modeled, and it can be ‘‘ignored.’’

MCAR is a strict assumption and is rarely satisfied in practical applications, especially when the data on Y are missing because of uncontrolled events in the course of the data collection as these events are often associated with the study variables. The MCAR assumption is thought to be more plausible if the missing data are missing by design such as a design in which participants are randomly assigned to conditions in which they do not respond to all items, all measures, and/or all measurement occasions. MCAR is assumed in traditional missing data methods and modern methods which will be discussed in detail later in this chapter.

MAR

Again, let's define the complete data $Y = (y_{ij})$. Let Y_{obs} denote the observed components or entries of Y, and Y_{mis} the missing components. The missing-data mechanism is called MAR when missingness depends only on the components Y_{obs} of Y that are observed, and not on the components that are missing, Y_{mis}. Probabilistically stated, the missing data mechanism is called MAR when the probability of missing data on a variable Y does not depend on the value of Y itself but it depends on other observed variables in the study. For example, in depression surveys, male participants are more likely to refuse to fill out the survey, but it does not depend on the level of their depression; and after controlling for gender, the probability of missing data does not vary among male participants. MAR assumption is less restrictive than MCAR (Little & Rubin, 2002).

MAR is also considered an ignorable missing data mechanism since the mechanism that created the missing data is related to information that is known. With MAR, missingness is predictable from other variables in the observed data (e.g., if females tended to fill out the depression survey but males did not, then the missingness would be predicted by gender). MAR is more common in practice and most widely used assumption about missing data mechanisms, and many computational methods have been well developed for handling data under the MAR assumption. In general, there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents (Schafer & Graham, 2002).

MNAR

Finally, the mechanism is called MNAR if the distribution of M depends on the missing values in the data matrix Y. Probabilistically, the missing data mechanism is called MNAR when

the probability of missing data on a variable Y depends on the value of Y itself. MNAR is present when there is a systematic relation between the causes for data being missing and the missing data. For example a questionnaire item asking for personal information like income or salary is less likely to be completed by people with very low or high incomes as people do not want to share their income when it is too high or too low. Another example is asking smoking status among pregnant women. Women who still smoke during pregnancy are more likely to skip the question than women who do not smoke. It commonly occurs when people do not want to reveal something very personal or unpopular about themselves.

MNAR is considered non-ignorable (i.e., the missing mechanism should be modeled) as the reason for missing data is related to the missing values or related to the variable that is unobserved in the study. With MNAR, the missingness is predicted by the missing variables (e.g., people with very high income tended to skip the survey item asking about income; or pregnant women who are still smoking while being pregnant tend to skip the survey item asking about smoking status). Missing data mechanisms do not exist exclusively, but more than one mechanism could be present in a sample of data.

In practice, using knowledge during the data collection process and the data in general, most researchers have an idea about the reasons for the missing data to make assumptions about the missing data mechanism whether it is MCAR; for example if it appears as if most cases with missing values also score low (or high) on other variables, then these would be evidence indicating that the values are not missing completely at random. In addition, a statistical test may be used to determine if missingness is MCAR; and to obtain more valid conclusion whether the missingness is MCAR, it is recommended to combine these two approaches. For the statistical approach, the absence of MCAR mechanism can be reliably detected using Little's (1988)

proposed chi-square test which is the most common test for MCAR. If the p value for Little's MCAR test is not significant, then the data may be assumed to be MCAR. Prior to Little's method, data analysts used *t*-tests to assess whether missing data were, in fact, MCAR. Using *t*-test approach, MCAR can be confirmed by dividing respondents into those with and without missing data, then using *t*-tests of mean differences on key variables like income, age, gender to establish that the two groups do not differ significantly on any variable in the model, including the dependent variable.

Researchers accept MCAR as the missing data mechanism if they cannot rule it out. If MCAR is ruled out then either MAR or MNAR is accepted as the plausible missing data mechanism; however, it is not possible to test whether the missing data are MAR or MNAR because there is no information about the missing data themselves available. To distinguish between MAR and MNAR, researchers must establish whether the missing data mechanism is ignorable by relying on logic and a sound understanding of the study design and domain, using subject matter experts.

Missing Data Treatment Methods

In the missing data method literature, two broad strategies to deal with item nonresponse type of missing data are elimination and imputation; elimination procedures will eliminate observations with missing data from the analysis, and imputation procedures will replace the missing values with estimates to create a complete dataset that can then be analyzed with traditional analysis methods. Of different forms of deletion, LW is widely used and is the default setting in many analysis programs. Imputation strategies use multiple techniques to estimate the missing values; the objective of imputation is not to get the best possible predictions of the

missing values, but to replace them by plausible values in order to exploit the information in the recorded variables for the incomplete cases (Little & Rubin, 2002).The two popular imputation techniques studied in this dissertation are hot-deck imputation and regression imputation. In hot-deck imputation, the value assigned for a missing item is taken from respondents in the current sample. In regression imputation, regression is used to predict the value of the missing data variable based on the relationship between that variable and other variables in the sample dataset.

With imputation approaches, traditionally, a missing value is substituted with an estimate; then the filled-in dataset is analyzed. This method of imputation is called single imputation as opposed to multiple imputation, a more advanced imputation method; multiple imputation replaces each missing value with a set of plausible values that represent the uncertainty about the right value to impute. Commonly, a review of missing data treatment methods found in literature often focuses on the degree of implementation difficulty, the ability to properly reduce bias, and the ability to efficiently estimate standard errors. Using similar approach, this section will review listwise deletion, hot-deck imputation, regression imputation, and multiple imputation which includes hot-deck based multiple imputation and regression based multiple imputation.

Listwise Deletion

Listwise deletion also known as complete case analysis is among the oldest methods of treating missing data. With LW, all cases with missing values are removed from the analysis; that is the analysis use only cases with complete data on all variables used in the analysis. Listwise deletion is based on the assumption of MCAR (i.e., the set of complete cases is a random set of the whole sample); if this assumption holds, there will be no bias in analyses based

on complete cases (Allison, 2002). A major benefit of LW is simplicity: no special computational methods are required, and it is the default setting in many statistical software packages. However, LW can produce distorted parameter estimates when MCAR assumption does not hold or when the remaining cases are unrepresentative of the complete dataset.

Even if the MCAR assumption is plausible, eliminating data is wasteful especially when the discarded cases have data on a large number of variables. For example, when very high or low income data are missing, removing cases associated with these two groups also means that characteristics of interest related to these two groups will be removed from the data as well. Also dropping the incomplete cases can produce a dramatic reduction in the total sample size. According to King and colleagues (2001), on average, data analysis in political science research typically loses about a third of the cases due to LW of missing data; and it is strongly condemned (King et al., 1998); the APA Task Force on Statistical Inference also blamed that listwise and pairwise deletion are "among the worst methods available for practical application." (Wilkinson & Task Force on Statistical Inference, 1999, p. 598). In addition, recently, in comparing multiple imputation approaches for treating missing discrete data, Cranmer and Gill (2012) concluded that any imputation approach is an improvement over LW and LW is not a benign solution.

It is commonly known that LW distorts parameter estimates when the data are not MCAR and reduces power because of reduction in sample size. However, with the rise of "big data" in this advanced technology age, lacking of power due to insufficient data caused by LW tends to be not a major issue; and as regards to the bias issue, whether LW produces biased estimates depends not only on the source of missingness but also on the form of the analysis to be undertaken. Little (1992) observed that when missingness is a function of a predictor variable

and not the outcome variable, LW can produce unbiased estimates of regression slopes under all missing data mechanisms. Later, Little and Rubin (2002) also added that LW is valid when the focus of the study is on inference about the regression parameters and only the outcome variable is missing. This observation of robustness of LW to different missing data mechanisms is supported by multiple recent studies that LW can be unbiased under certain MAR or MNAR mechanisms if missingness is independent of outcome Y (Bartlett et al.,2015; White & Carlin, 2010).

An alternative deletion method is the pairwise deletion or available case analysis which attempts to mitigate the loss of data; with pairwise deletion, incomplete cases are deleted on an analysis-by-analysis basis. Consequently, any given case may contribute to some analyses but not to others. Using this approach the sample size will remain the same for some analyses and will be reduced for others. This inconsistency of the sample size can lead to problems in computing standard errors. Pairwise deletion is also based on the assumption of MCAR to produce unbiased estimates.

Another broad strategy to deal with missing data is imputation. Imputation strategies can be categorized based on whether the method is based on explicit or implicit modeling. For implicit modeling-based methods, the focus is on an algorithm, and their assumptions are implicit. An example of implicit modeling-based method is hot-deck imputation. The explicit modeling-based methods are based on a formal statistical model and the assumption is explicit (Little & Rubin, 2002). An example of explicit modeling-based method is regression imputation.

Hot-Deck Imputation

Hot-deck imputation is the best-known nonparametric approach to imputation (Allison, 2002). It involves replacing missing values of one or more variables for a nonrespondent with observed values from a respondent that is similar to the non-respondent with respect to characteristics observed in both cases (Andridge & Little, 2010); for example, missing income value for a nonrespondent is replaced with observed income value from a respondent who is similar to the nonrespondent in occupation, education, gender. It is commonly used for item nonresponse in survey practice due to its flexibility to handle various types of variables as well as the face validity of imputing values that have been observed in the same dataset on other individuals (Molenberghs, 2014). Various hot-deck methods have been proposed in the missing data literature. These methods range from a set of simple steps to a quite elaborate algorithm in attempts to find respondents as similar as possible to those with missing responses. Below are a few well-known hot-deck imputation methods.

Random Overall Imputation

The simplest forms of hot-deck imputation is *random overall imputation* in which, for each nonrespondent, a respondent is chosen at random from the total respondent sample, and the selected respondent's value is assigned to the nonrespondent for which this information is missing (Kalton & Kasprzyk, 1986). For example, assume that a sample survey has *n* targeted participants, with *r* individuals who provided responses and *n-r* individual who did not. With simple random hot-deck imputation, *n-r* values are randomly chosen from the *r* respondents to impute missing values to the *n-r* nonrespondents; the *r* respondents are called donors, and the *n-r*

nonrespondents are called recipients. With this simple hot-deck method, all respondents in a sample data have equal probability of being selected as a ‘‘donor’’ for a missing value.

Random Imputation within Class

In practice, the accuracy of imputation is improved by first dividing the sample into a set of classes called imputing classes or cells using control variables that are observed on all sample units, and then performing hot-deck imputation separately within each imputation class for each item with missing values. These imputation classes have responding and nonresponding units, and the matching donors with recipients process is carried out within classes. These imputation classes, also known as adjustment cells or donor pools, may be formed based on subject-matter expertise or on statistical analyses that determine which set of variables are predictive of nonrespondents (Brick & Kalton, 1996). Using a random imputation within classes method, a donor for each nonrespondent is randomly selected from within each classes; and this is what was used in this dissertation study to compare with other missing data method. If a recipient has multiple missing items, to preserve multivariate relationships, usually values from the same donor are used for all missing items of a recipient (Lorh, 2010).

Many methods have been proposed to select donors that are similar to recipients. Below are some popular donor selection methods.

Sequential Hot-Deck Method

This method begins by defining a set of groups or imputation classes. A starter value is assigned to each group. (This starter value may be taken either from a previous survey, class, or randomly from the variable’s domain; alternatively, it may also be the overall mean for the study

sample as a whole on the variable of interest for all cases for which the information is available or the group mean of the variable of interest for all cases for which the information is available within groups.) The survey records in the data file are ordered sequentially according to important variables which influence response. The first record within the group for which values are to be imputed is examined. If it is missing, then the starter value replaces the missing value. If the first case in the group does not have a missing value, this first real value replaces the starter value in this imputation class. The next record is examined; if the interest value is missing the new value is used to assign to this missing value. If it is not missing, it will be used as the current donor value for the next missing value. The process continues sequentially until all missing values are replaced by real values donated by the case preceding it within the same class or group. It is believed that if the data are arranged in some geographical order, adjacent units in the same class will tend to be more similar than randomly chosen units in the class (Lorh, 2010).

One of the disadvantages of sequential hot-deck method is that when two or more records with missing values occur in sequences in a given imputation class, these records receive values from the same donors, which was taken from the previous responding record. Using the same donors multiple times causes a loss in precision in the survey estimates (Kalton & Kasprzyk, 1986). Some of the suggested methods to deal with this problem are the random imputation within classes method as mentioned above and hierarchical hot-deck method.

Hierarchical Hot-Deck Method

In the hierarchical procedure, many imputation classes may be used initially. The survey records are grouped into a large number of subclasses based on detailed categorization of a large set of auxiliary variables. (Auxiliary variables are variables that are correlated to describe the

variables of interest; usually they are not part of the planned analysis but are collected to improve the accuracy of imputed values. In hot-deck imputation, auxiliary variables are used to track donors and they are used to predict the missing value in regression imputation; for example the variable education, occupation, gender, and neighborhood are used to predict income). The records within each subclass are separated into respondents and nonrespondents. Nonrespondents are then matched with respondents in these smallest classes first. If there are imputation classes with nonrespondents but without respondents or if the ratio of nonrespondents to respondents in some classes is too large, then some collapsing of classes may be employed. For example, the subclasses are categorized based on age by sex by race, these subclasses are then collapsed into a broad groupings by age only and the matching process is attempted at this level. The term 'hierarchical' is used to reflect this collapsing, with an initially detailed matching of respondents and nonrespondents, then collapsing the level of detail where necessary to ensure that donors are found for all nonrespondents.

Weighted Sequential Hot-Deck Imputation

The hot-deck imputation methods discussed above do not take into account the unequal probabilities of selection in the original sample. Ignoring sample weights may cause bias if the respondents have differing sampling weights because the distribution of responses within each imputation class of the imputation-revised data set may be distorted from that of the original distribution of responses (Andrige & Little, 2009). Cox (1980) proposed *weighted sequential hot-deck imputation,* which is a modified version of the sequential hot-deck imputation, to incorporate survey design weight into donor selection. Essentially, it replicates the weighted distribution of the available data in the imputed data by using the sample weights of item

respondents and nonrespondents. With *weighted sequential hot deck imputation,* means and proportions, estimated using the imputed data, will be equal in expectation to the weighted mean or proportion estimated using respondent data only. This is achieved by using the sampling weight from the original sample to specify the expected number of times a particular respondent's answer will be used as a donor. Each respondent record has a chance to be selected for use as a donor, and the number of times a respondent record can be used for imputation will be controlled (Andrige & Little, 2009, 2010). The weighted sequential hot-deck does not appear to have been widely implemented (Andrige & Little, 2010).

Weighted Sequential Hot-Deck Imputation is one form of weighted hot-deck imputation. Another form of weighted hot-deck imputation is *weighted random hot-decks.*

Weighted Random Hot-Decks

This imputation method is used to incorporate sampling weights when the donors are selected by simple random sampling from the donor pool. To incorporate sampling weights into donor selection, Platek and Gray (1983) suggested inflating the donated value by the ratio of the sampling weight of the donor to that of the recipient; while Rao and Shao (1992) and Rao (1996) suggested selecting donors via random draw with probability of selection proportional to the potential donor's sample weight. Assuming the response probability is constant within an adjustment cell, Rao's method yields an asymptotically unbiased estimator for Y (Andrige & Little, 2009). In general, applying weighting in imputation often reduces bias but increases variances of the estimates (Brick & Kalton, 1996).

One disadvantage of hot-deck imputation is that valid variance estimates cannot be obtained using standard formulae because the resulting variance will typically underestimate the

true variance of the sample statistic as the same individual's responses may be used repeatedly to supply missing information. Many special methods (Burns, 1990; Rao & Shao, 1992; Rao, 1996; Shao & Sitter, 1996; Chen & Shao, 1999) have been proposed to properly estimate variances when using hot-deck in single imputation. Rao and Shao's (1992) method was used to estimate hot-deck variance in this study. Another disadvantage is that hot-deck imputation requires good matches of donors to recipients, and finding good matches is more likely in large than in small samples. The positive side of hot-deck imputation is it is impossible for hot-deck imputation to impute values outside the observed range of the data because only observed values of the variable being imputed are candidates for imputation. Also with hot-deck imputation, nonsensical values cannot be imputed because within the observed range of the data, only values that have occurred for other cases can be imputed; hence it is more intuitively appealing than other imputation methods.

Regression Imputation

Regression methods involve a regression equation based on the nonmissing data to predict the values for the missing data; with this method, the missing values are the outcome variable, with one or more of the auxiliary variables in the data set serving as predictors. The simplest form of regression imputation is based on the simple linear regression with which the prediction of missing data is based on the regression of the variable with missing data on the single variable most highly correlated with it (Frane, 1976). However, simple regression imputation sacrifices other covariate information, and in the past it has received little attention (Affifi & Elashoff, 1969; Raymond, 1986; Raymond & Roberts, 1987). In order to make use of all available information, Buck (1960) proposed multiple regression imputation which is more

commonly-used regression imputation; and the relative performance of multiple regression imputation was assessed in this study.

In regression imputation, first a regression model is fitted with the variable with missing data as response variable and auxiliary variables as covariates. The coefficients are estimated and then missing values can be predicted by the fitted model.

Let *y* be the variable for which missing values are to be imputed and let $z= (z_1, z_2,\ldots,z_p)$ be the set of the auxiliary variables. Assume that *y* is a continuous variable and that there is no missing values for the z variables. Let y_{mi} and $\hat{y}_{mi}$ denote the actual and imputed values of *y* for record *i* for which the value of *y* is missing. Regression imputation methods can be represented by the regression equation

$$\hat{y}_{mi} = b_{r0} + \Sigma b_{rj} z_{mij} + \hat{e}_{mi}$$

Where b_{ro} is the intercept and b_{rj} are the estimated regression coefficients for the regression of y on *z* obtained from the records with *y* values observed. z_{mij} is the value of z_j for record *i* with a missing y value, and $\hat{e}_{mi}$ is residual term (Brick & Kalton, 1996).

In deterministic regression imputation, the value of $\hat{e}_{mi}$ is zero. The imputed values obtained from deterministic methods do not reflect the residual variability as all the imputed values are on the regression line or plane. As a result, the variance of the $\hat{y}_{mi}$ is attenuated as compared with the variance of the unobserved y_{mi}, and the shape of the distribution of the *y* values is distorted. With deterministic methods, estimates of means (e.g., mean income) are more precise; however, it underestimates the dispersion of Y and distorts the shape of the distribution of the variable Y (e.g., underestimating the proportion of the population with very high or low income) and distorts the correlation between variables, which are not used in the regression model. It might also artificially inflate the statistical association between Y and the auxiliary

variables (Kalton & Kasprzik, 1986; Durrant, 2005). To preserve the distributions in the imputed data and the associations between variables, a random residual is added to the estimate (which is also known as conditional mean).

In stochastic regression imputation, the non-zero residual, $\hat{e}_{mi,}$, is added to the conditional mean to reflect uncertainty in the predicted value. The residuals could be randomly chosen from a normal distribution with a mean of zero and a variance equal to the residual variance from the respondent regression. Another way of choosing residuals is to select a respondent at random and to take that respondent's residual, or select a residual from a respondent who has similar values on the auxiliary variables to the nonrespondent (Kalton & Kasprzik, 1986; Brick & Kalton, 1996). These stochastic values introduce variance into the imputed data that results in unbiased variance estimates while providing the same unbiased means as the deterministic regression imputation.

Regression imputation depends on several assumptions such as data must be missing at random, each missing variable must be highly correlated with one or more available variables, and the amount of missing data should not be excessive (Frane, 1976). Usually, linear regression is used for numeric variables, and logistic regression is used for categorical data. An advantage of regression imputation is that it can make use of many categorical and numeric variables. The method performs well for numeric data, especially if the variable of interest is strongly related to auxiliary variables. Regression imputation produces unbiased means under MCAR or MAR.

The main disadvantage of regression imputation is that the method depends on the construction of a suitable model and may be sensitive to misspecification of the regression model (Schenker & Taylor, 1996). If the regression model is not a good fit the predictive power of the model might be poor (Little & Rubin, 2002), and if a seriously misspecified model is used the

methods may generate poor, even impossible, imputed values (Kalton & Kasprzik, 1982). The regression imputation method increases correlation coefficients between outcome and predictors; and it distorts the shape of the correlation between variables which are not used in the regression model.

Compared to the imputation class/cell method in hot-deck imputation, regression imputation methods have an advantage in the level of detail of the auxiliary variables they can employ; for example, age can be taken as a continuous variable instead of being categorized into a few classes. Also, a regression mode allows more main effects to be included in the model, and it can also include polynomial terms and employ transformation. With careful modelling, regression models have the potential of providing better predictions for the imputed values (Kalton & Kasprzik, 1986).

A variant of regression imputation is called predictive mean matching (Little, 1988) (Little attributes the method to Rubin (1986)). With predictive mean matching, the nonrespondent is matched to the respondent with the closest predicted value, and then the respondent's observed value (not the predicted value) is assigned to the matched nonrespondent. Predictive mean matching was proposed to be used in multiple imputation, which will be described in the following section.

Multiple Imputation

The imputation methods mentioned above are called single imputation because each missing value is imputed only once. A disadvantage of single imputation is that imputing a single value treats that value as known, thus single imputation does not reflect the uncertainty about the predictions of the unknown missing values, and the resulting estimated variances of the

parameter estimates are biased toward zero (e.g., the underestimation of the true variance of the estimate increases as the proportion of missing values for an item increases). Instead of filling in a single value for each missing value, Rubin (1977, 1978) proposed the multiple imputation method in which the missing value is filled in with multiple imputed values, resulting in multiple imputed complete datasets from which the resultant statistics (e.g., regression coefficients) then can be averaged to get a pooled estimate across imputed datasets.

Specifically, multiple imputation inference involves three distinct phases: (1) the missing data are filled in *m* (m≥2) times to generate *m* complete data sets; (this step can be done with any existing single-imputation method such as hot-deck or regression method mentioned above to generate imputations); (2) the *m* complete data sets are analyzed independently by using standard procedures; and (3) the results from the *m* complete data sets are combined for the inference. Combining the resultant answers often only requires the calculation of the means and variances of the repeated complete-data statistics. The combined statistic is a simple average, but the variance of this estimate is computed using rules that combine within-imputation and between-imputation variability (Cranmer & Gill, 2012; Rubin, 1987).

Let $\hat{\theta}_m$, W_m, m=1,…, M be M complete-data estimates and their associated variances for an estimated parameter θ, calculated from M repeated imputations under one model. The combined estimate is

$$\bar{\theta}_M = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m$$

The variability associated with this estimate has two components: the average within-imputation variance,

$$\bar{W}_M = \frac{1}{M}\sum_{m=1}^{M} W_m$$

And the between-imputation component,

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{\theta}_m - \bar{\theta}_M)^2$$

The total variability associated with $\bar{\theta}_M$ is

$$T_M = \bar{W}_M + \frac{M+1}{M} B_M$$

Where (1 + 1 / M) is an adjustment for finite M, and the new degrees of freedom are

$$df = (M-1)(1 + \frac{1}{M+1}\frac{\bar{W}_M}{B_M})^2$$

These quantities are produced automatically for users by statistical software implementing multiple imputation.

As regard to the number of imputations, Rubin (1987) suggested a modest *m* between 2 and 10 being used for situations with a modest fraction (e.g., 10 - 30%) of information missing due to nonresponse. The author mentioned that when fractions of missing information are large, modest *m* multiple imputation is not fully satisfactory. However, Allison (2002) suggested that even with 50% missing data, five imputed datasets is sufficient. As commonly implemented, multiple imputation assumes a continuous metric for the missing data and therefore produces imputations that are continuous. Multiple imputation rectifies the problem of underrepresentation of uncertainty with single imputation. As used in practice, multiple imputation usually assumes missing data are ignorable (i.e., either MCAR or MAR). It is suggested that non-ignorable missing data (i.e., MNAR) biases model-base multiple imputation (e.g., regression-based

multiple imputation) since there is no observed information from which to build an imputation process for filling-in missing values (Cranner & Gill, 2012).

At the time of its development, multiple imputation raised concerns about more work being needed to produce multiple imputation than single imputation, more space is needed to store a multiply-imputed data set, and more work is needed to analyze a multiply-imputed data set than a singly-imputed data set. However, these disadvantages are not as a big concern in the current stage of advanced computing and storage capacity.

Obviously, the difference between regression multiple imputation and hot-deck multiple imputation, lies in the imputation steps. Once the multiply-imputed datasets are produced and analyzed, the results are combined into a single estimate using the same formula. Specifically, for the regression-based multiple imputation, a model is used to generate the imputations multiple times. The most popular model for multiple imputation is the multivariate normal model which implies that all variables have normal distributions; each variable can be represented as a linear function of all the other variables, together with a normal, homoscedastic error term (Alliston, 2002).

As a result of regression multiple imputation, each of the multiple imputed datasets contains, possibly different, imputed values for each nonrespondent; this addresses the uncertainty associated with missing data. However, for hot-deck multiple imputation, if the standard random hot-deck procedure is applied multiple times to create a set of imputed data sets, the imputations are not proper as they do not fully propagate variability across imputations, and the between imputation variance will be underestimated if no adjustment is made (Sullivan & Andridge, 2015; Rubin, 1987). To make hot-deck MI a work properly, Rubin and Schenker (1986) proposed a method called approximate Bayesian bootstrap (ABB) as a modification to the

usual simple random hot-deck. ABB involves a two stage sampling procedure which can be described as below.

For simplicity, Y is defined to be a single variable with missing values. Y_{obs} consists of the values of Y that are observed, and Y_{mis} consists of the values of Y that are missing. Let n_{obs} and n_{mis} be the number of cases associated with Y_{obs} and Y_{mis} respectively. For the purposes of illustration, assume there is a single pool of donors, that is all n_{obs} subjects with observed Y are eligible to donate to each of the n_{mis} subjects with missing Y. In a standard simple random hot-deck, n_{mis} values would be drawn with replacement and with equal probability from the n_{obs} donor values (Y_{obs}) and used as imputed values. The ABB method will first draw n_{obs} cases randomly with replacement from Y_{obs} to create a new set of Y^*_{obs}, where the * denotes a bootstrapped sample of $Y_{obs.}$ Then standard hot-deck imputation is applied, with n_{mis} observations selected randomly with replacement and with equal probability from Y^*_{obs} and used for imputation (Siddique & Belin, 2009; Sullivan & Andridge, 2015).

In practice, the sample is divided into a set of classes as mentioned previously. The approximate Bayesian bootstrap is performed separately within each imputation class. Application of the above procedure in each class the imputation process will include four steps: (1) from the set of n_{obs} cases with complete data, take a random sample (with replacement) of n_{obs} cases; (2) from this sample, take a random sample (with replacement) of n_{mis} cases; (3) assign the n_{mis} observed values of Y to the n_{mis} cases with missing data on Y; (4) repeat steps 1 to 3 for every group. When applied to all groups, these four steps produce one completed dataset. For multiple imputation, the whole process is repeated multiple times. Estimates of interest can then be combined using the standard rules as formulated by Rubin (1987). ABB creates the added variability in the donor pool; hence between-imputation variance will no longer be

underestimated. ABB was used in this present study to carry out the hot-deck multiple imputation method.

Under the approximate Bayesian bootstrap, the mean and variance estimator is asymptotically unbiased; however the variance estimator is known to be biased in small-to-moderate samples. Kim (2002) and Parzen and colleagues (2005) each proposed a modification of the method for reducing the bias of the variance estimator. However, evaluating the two modifications and the original approximate Bayesian bootstrap method, Demirtas and colleagues (2006) found that efficiency losses for the variance estimator outweighed bias improvements and suggested that the proposed modifications may be inferior to the original approximate Bayesian bootstrap depending on the primary focus of interest.

Also, in discussing the application of Approximate Bayesian Bootstrap technique in complex sample survey data, Rao and Shao (1992) commented that Approximate Bayesian Bootstrap technique works well if the imputation does not cross sample clusters (e.g., donor and recipients are in the same cluster); if the imputation does cross sample clusters, the Approximate Bayesian Bootstrap technique does not always lead to consistent estimators, even with many imputations; this led to the authors' proposal of a jackknife variance estimator for stratified multistage surveys. The jackknife variance estimator for stratified multistage surveys is shown to be consistent as the sample size increases, assuming a uniform response mechanism within each imputation class and a particular hot-deck imputation (Rao & Shao, 1992). Using the jackknife variance estimator, the missing value is imputed once and results in one complete dataset.

Past Studies

Relative Effects of Different Simple Data Missing Methods

Roth and Switzer III (1995) empirically compared the effects of eight simple missing data methods on the precision and bias of predictor-criterion correlation and regression weights. The study was conducted using a simulation approach in which 50 complete datasets were generated using a population correlation matrix. The generated data matrix includes three predictor variables and two criterion variables. Only one type of missingness (i.e., MCAR) was studied; controlled factors include levels of missing data (10%, 20% and 30%) and sample size (50,100, and 200).

Bias and RMSE were used to assess the accuracy and precision (i.e., the amount of dispersion around true scores) of the regression weight and correlation between predictor and criterion yielded by different MDTs. Generally, the results in most studied conditions showed that pairwise and listwise deletion methods are superior to regression imputation and hot-deck imputation. Specifically, for regression weight, LW yielded essentially no bias, regression imputation yielded relatively small level of bias; and hot-deck imputation yielded highest regression weight bias. With regards to the precision of the estimate of regression weight, LW was second after regression which produced the least dispersion, and hot-deck produced the most dispersion. In all conditions, hot-deck imputation was found inferior to LW and regression imputation.

Listwise deletion performance was also found superior to regression imputation on estimates of regression coefficient in MNAR data in another study in which Kromrey and Hines (1994) compared the performance of five missing data treatments in multiple regression analysis with MNAR data, using bootstrap samples (N=50, 100 & 200) drawn from actual field data with

10% to 60% missingness. The authors found that listwise and pairwise deletion generated accurate estimates of regression coefficients with up to 30% of missingness while regression imputation method produced biased estimates.

Listwise Deletion vs. Multiple Imputation

Literature does not seem to support the claim that all imputation methods are better than LW, even when LW was compared to MI (conventionally referring to parametric or regression-based MI). Unexpected findings about LW performance have been found in past and recent studies. In the 1994 study, Kromrey and Hines also found that RM consistently overestimated regression weight while LW yielded accurate estimates of regression weight with up to 30% of missingness. Gibson and Olejnik (2003) assessed the relative efficacy of listwise deletion, MI, and other missing data methods for treating data in level-2 variable in a two-level data, using a simulation study. In the study, the effects of these MDTs on estimation of regression weights and random effects in hierarchical linear statistical models were evaluated under MCAR condition; the study used an "intercept-and slopes-as outcomes" model (e.g., students nested in a school, schools nested in some classification factor) with only Level 2 presenting missing data, and Level 2 has two variables with only one variable presenting missing data. The hierarchical data were simulated with respect to four design factors: (a) sample size for the second-level units (30,160) (the sample size for Level 1 was 45), (b) the correlation between Level-1 intercepts and slopes (.2, .8), (c) the number of second-level variables, and (d) the percentage of missing data (10%, 40%). The results indicated that LW performed well in estimating the regression weight for the Level 2 variable that has missing values (LW produced regression weight estimates that

were not statistically different from those produced by complete data) while MI consistently underestimated the parameter for Level 2 variable model in almost every condition.

Recently, in a Monte Carlo simulation study to compare bias in regression coefficient estimates when MCAR, MAR, and MNAR data are treated by MI and LW, Kuijk and colleagues (2016) found rather surprising results that LW yielded unbiased regression parameter estimates whereas MI overestimated the regression coefficient and underestimate the intercept under MNAR.

The authors performed a stochastic simulation study in which data were drawn from a multivariate normal distribution with three variables. Specifically the data were simulated based on the linear model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Controlled factors included sample size (100, 500, 1000), error variance (0.5, 1, 2, 4, 8), correlation between the two independent variable (-.75, -.50, -.25, 0, .25, .50, .75), the size of the regression coefficients for X1 (.5, 1, 2), for X2 (.25, .5, .75), and the proportions of missing value (25%, 50% and 75%). The number of Monte Carlo iterations was 1,000.

Missing data were created for X1 variable: (1) for MCAR data, a randomly selected proportion of X1 was deleted; (2) for MAR data, the missing mechanism was divided into MAR conditional on the X_2 variable (i.e. the lower the X_2 value, the higher the probability that the corresponding X_1 value was missing), and MAR conditional on Y (i.e., the lower the Y value, the higher the probability that the corresponding *X1* value was missing); and (3) To simulate a MNAR mechanism, the probability for any value to be missing was associated with its own value (i.e., the lower the value, the higher the probability of being missing).

MI was performed using the default settings in the *mice* package in R; specifically, the imputations were generated using predictive mean matching method, and the number of multiple

imputations was 5. The imputed data were analyzed using linear regression analysis to estimate regression coefficients, and this linear model was of the same structure as the one used to simulate the data. The results show that under MCAR and MNAR, LW analysis yielded unbiased regression parameter estimates, but under MAR, regression coefficients from LW analysis were underestimated (especially when data were MAR conditional on Y, LW resulted in severely biased regression coefficients); whereas MI yielded unbiased parameter estimates under MCAR and MAR but overestimated the regression coefficient under MNAR, contrary to expectation. Also, in conditions where LW analysis led to biased results, the regression coefficients were consistently underestimated; conversely, in scenarios where MI analysis yielded biased results, the regression coefficients were in most cases overestimated.

Evidences of LW outperforming MI were also seen in Kellermann and colleagues' (2016) study in which LW consistently surpassed MI in performance by delivering virtually no bias under MCAR and very minimal negative bias under MAR while MI point estimates were considerably biased in both directions under MCAR and MAR. One thing in common to be seen about the results of the two studies is that LW tends to deliver negative bias under MAR condition though the level of bias is very minor in the latter study. Regarding to the performance difference for MI between two studies, it may be worth noting a few fundamental differences in designing and executing the two studies. Firstly, the two studies based on different data structure and number of covariates; the latter study based on two-level complex sample data structure with eight variables whereas the prior based on flat data with two variables; secondly, the MAR data in the latter study was conditional on the observed predictor variables instead of on the observed criterion variable as in the prior study; thirdly, in the latter study, the imputations were created

using regression imputation whereas in the prior study the imputations were created using the predictive mean matching method.

The results from the above studies are based on all continuous data. When the missing data are discrete, the relative performance of LW and regression-based multiple imputation can be different, and they both appeared to be inferior to hot-deck based multiple imputation; and this can be seen in the following study's results.

Listwise Deletion vs. Parametric Multiple Imputation vs. Nonparametric Multiple Imputation

Regression multiple imputation and hot-deck multiple imputation both overcome the uncertainty associated with single imputation; however, multiple hot-decking is a nonparametric approach to imputation, therefore it avoids the assumptions and problems associated with the parametric approach such as the normality assumption and model-fitted dependence (e.g., the imputation step depends on the construction of a suitable model and may be sensitive to misspecification of the regression model as mentioned previously); also, assumptions and algorithms of multiple hot-deck imputation are thought to be intuitive and easy to understand (Cranmer & Gill, 2012).

In addition, because multiple regression imputation produces continuous imputations, it can be problematic when used on discrete data because of problems associated with rounding. Using parametric multiple imputation with rounding on discrete data will necessarily result in some degree of bias; also coefficients from parametric multiple imputation and rounding on discrete data will have inappropriate standard errors as the distance between the imputation and the nearest integer is lost. To avoid this problem, Cranmer and Gill (2012) suggested utilizing

hot-deck technique to impute missing data in multiple imputation to impute discrete data in political science. The authors performed a Monte Carlo experiment to illustrate the relative performance of the proposed hot-deck MI, LW, and parametric MI in MCAR and MAR data using discrete data.

The hot-deck imputation technique used in this experiment is the *random overall imputation* (e.g., for each nonrespondent, a respondent is chosen at random from the total respondent sample) as described earlier in the hot-deck imputation section. To evaluate the performance of these three methods, the authors randomly generated datasets of 500 observations, each with five binary variables with fairly high correlations between them (r=.8). Binary values were created by rounding the draws to 0 or 1 depending on whether the draws were below or above 0 respectively. According to the authors, this data setup maximally advantages parametric MI and maximally handicaps the proposed hot-deck MI technique since it is based originally on a continuous parametric specification.

Missing values were created on two of the five variables in the datasets. The proportions of missing data were 20%, 50%, and 80%. To create MAR data, the probability that a value was missing from the two affected variables was conditional on the observed values of the other variables all being zero. The experiments were run for 10, 000 Monte Carlo iterations. The parametric MI method used in this experiment was implemented in R packages named Amelia (King et al., 2001; Honaker & King, 2010), which is one of the most commonly-used methods in political science. The authors used the sample mean, standard deviation of the mean, and regression weights of logistic regression to compare the performance of the three missing data method in the study, and they also used percent of values imputed correctly to compare the two

imputation methods. The study results showed that no method was biased under MCAR. Below are the results of the study under MAR data.

For sample mean, the results show that LW yielded bias and the bias increases in intensity as the proportion of missing values increases; parametric MI displays a smaller bias than LW, and the bias also increases as the proportion of missing values increases; standard deviations of the means yielded by LW and parametric MI also increase as the proportion of missing data increases. Superior to LW and parametric MI, hot-deck MI displays no bias across all three levels of missing data although there is a slight increase in the standard deviation of the means as the proportion of missing values increases, which indicates increasing uncertainty (a desirable result as the volume of missing values goes up).

Regarding the percent of imputed values that were imputed correctly for both imputation approaches, Both multiple hot-decking and parametric multiple imputation perform remarkably well as more missing values are added (e.g., percent of correctly imputed values does not change as the proportion of missing data increases) with hot-deck MI getting about 83% of the missing values correct and parametric MI getting about 75% correct.

For regression weights of logistic regressions in which one of the variables with missing values was the outcome variable, and two fully observed variables served as explanatory variables, the results show that LW results in consistently biased point estimates, and that this bias becomes more severe as proportion of missing data increases; parametric MI typically displays less bias than LW, however there were cases (i.e., β_2 across three missing data levels) in which parametric MI produces a more severe bias than LW. The authors noted that the bias of the point estimates produced by parametric MI is due to the rounding. As expected, hot-deck MI outperforms parametric MI and LW in all cases.

Summary

As more survey data are made available to researchers for secondary study, more applied researchers have been using hierarchical data in their studies; however, literature shows MDTs have been investigated extensively in the context of single-level data but not much in hierarchical structure data such as complex sample survey data, in which the multi-stage sampling feature complicates the estimate of sampling error. This present study investigated the efficacy of MDTs on treating hierarchical data, specifically on two-level, complex sample survey data. To examine the influences of the complex sample design factors on the efficacy of the studied MDTs, this present study not only examined the effect of different levels of ICC, a factor that affects the precision of parameter estimate obtained from complex sample survey data, but also the population density, a factor that may affect the cluster effects which in turn affects the variance of the estimates obtained from the complex sample survey data.

In addition, often, studies on the efficacy of MDTs empirically compared the MDTs under their assumed missing data mechanism (e.g., comparing single regression imputation and MI under MAR conditions), but few have been seen to explore the robustness of MDTs when they depart from their assumed missing data mechanisms, (e.g., study LW or single/multiple imputation under MNAR). This study compared the effectiveness of the five studied MDTs under three missing data mechanisms including MNAR data because one of the missing data issues that is important to practicing researchers is the feasibility of applying missing-data treatments that are recommended for randomly missing data to samples in which data are missing nonrandomly (Kromrey & Hines, 1994).

Also, to date, due to limitations on the type of MDTs that statistical software packages offer, MDT studies have been generally limited to comparing MDTs which are based on

different theoretical approaches (e.g., deletion vs. imputation) or comparing different deletion methods (e.g., listwise vs. pairwise deletion), or comparing different imputation approaches (e.g., single imputation vs. multiple imputation), but not comparing different statistics-based approaches used in multiple imputation (e.g., parametric-based MI vs. nonparametric-based MI). This present study investigated the relative performance of regression-based multiple imputation vs. hot-deck-based multiple imputation using continuous variables; this particular focus of MDT evaluation has not been seen in literature.

Moreover, with regards to the evaluation criteria, MDT studies in literature often limited to the use of bias and RMSE, or sometime a single criterion (i.e., bias) is used (e.g., Gibson & Olejnik, 2003). This present study focused on four quantitative evaluation criteria: bias, RMSE, CI width and CI coverage rate. The comprehensive approaches this dissertation study uses will provide broader perspectives on parameter estimation by each missing data treatment than previous studies with less evaluation criteria did.

In summary, in investigating the effects of MDTs, this study placed emphasis on four aspects that has not been explored or were not well explored in previous MDT studies: (1) the relative efficacy of the studied MDTs on treating missing data in complex sample survey data, (2) the robustness of the studied MDTs when departing from their missing data assumptions, (3) comparing effectiveness of parametric-based and nonparametric-based MI using continuous data, and (4) providing comprehensive evaluation criteria.

CHAPTER THREE: STUDY DESIGN

Objectives

This present study is designed to evaluate the performance of five MDTs for handling missing data in complex sample surveys in the context of multiple regression analysis of complex sample data. The five missing data methods include listwise deletion, hot-deck imputation, regression imputation, hot-deck multiple imputation and regression multiple imputation. Samples of complex sample survey data were simulated separately for each type of missingness (MCAR, MAR, MNAR) with respect to two survey design factors, population density and intraclass correlation (ICC). For each type of missingness, complete data (i.e., no missing data) and incomplete data at four different levels of missingness were created, and the five MDTs were assessed under different degrees of population density, ICC, and proportion of missing data. The following research questions were examined according to the study's objectives:

Given the various experimental conditions (population density, intraclass homogeneity (ICC), proportion of missing data, and missing data mechanisms),

1. For parameter estimates in a complex sample survey dataset, which missing data treatment method (LW, HS, HM, RS, or RM) produces more accurate and precise estimates in terms of bias and RMSE for point estimates, and CI width and CI coverage for interval estimates when data are MCAR, MAR, and MNAR?

2. For each missing data type (MCAR, MAR, MNAR) how do amount of missing data, Intraclass Correlation (ICC), and population density influence the accuracy and precision of parameter estimates produced by each of the studied MDTs?

Discussion of the research method has five sections. The first section discusses simulation conditions and number of replications; the second section describes data generation strategies; the third section provides information about the missing data treatment software package which were used in the study; the fourth section discusses the evaluation criteria used in assessing the performance of the MDTs; and the fifth section discusses data analysis.

Simulation Conditions and Number of Replications

For each type of missingness (MCAR, MAR, MNAR), the performance of the MDTs was evaluated based on the two complex survey design factors (ICC and population density) and the degrees of missing data. Data samples were generated separately for MCAR, MAR, and MNAR data; the generation of these samples is discussed in the Data Generation section. This section describes the three controlled factors (ICC, population density, and proportions of missing data) used in this study and discuss the number of replications.

Intraclass Correlation Coefficient (ICC)

In multistage sampling, the data structure in the population is hierarchical. As mentioned above, in multilevel models, with grouped data, observations from the same group are generally more similar to each other than the observations from different groups. ICC measures the extent to which individuals within the same group are more similar to each other than they are to

individuals in different groups. ICC ranges from 0 to 1 (or 0% to 100%): when the observations within groups are a lot more similar than observations from different groups (i.e., within-cluster variance is very small and much smaller than between-cluster variance), the ICC is close to one; on the other hand, when the observations within groups are not more similar than observations from different groups, the ICC is close to zero; in terms of variance, ICC is close to zero when the within-cluster variance is large and much greater than the between-cluster variance. The similarity among observations within classes violates the assumption of independence of all observations in regular regression. The effects of highly correlated observations within clusters can be expressed using ICC, and higher ICC values represent stronger clustering effects. Because different ICC levels have differential impact on parameter estimation (e.g., the smaller the ICC, the more accurate the parameter estimates), ICC is the desirable condition to vary.

The hierarchical data in this study has two levels: individual observations within primary sample units (PSU) and PSUs within strata. For this design, two variance components can be derived: the between-PSU variance and the within-PSU variance. The intraclass correlation is the proportion of cluster (i.e., PSU) mean variability over the total variability.

$$\text{ICC} = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

where σ_B^2 is the between-PSU variance, and σ_W^2 is the within-PSU variance.

For this study, the intraclass correlation is manipulated to investigate the effects of different degrees of cluster effect on parameter estimate. In general, the values of ICCs depend on specific cluster designs and the outcome variables; Carlin and Hocking (1999) found that ICCs between socio-economic variables tended to be greater than those between various health-related outcomes. They also found different ICC values for different outcome variables in the

same cluster design, for example, an estimated ICC value for the outcome values of "average income" for families living in the same area was found to be about 0.2, and the ICC value for the outcome variable of "currently married" within the same cluster design was found to be close to zero (Carlin & Hocking, 1999).

In a school-class cluster design, to estimate ICC for some common outcome variables used in a school-based smoking prevention study, Siddiqui and colleagues (1996) found that the ICC values were between 0.04 and 0.09. However, data from national samples (e.g., NELS, TIMMS, and PIRLS) tended to have larger cluster effects than those from smaller scale samples. Examining the US data from Trends in Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), Zopluoglu (2012) found that the average two-level ICC estimates across multiple test cycles (1995, 1999, 2003, 2007) was .37 for mathematics and science achievement domains at the 8^{th} grade level, .29 for mathematics, .33 for science, and .25 for reading achievement domains at the 4^{th} grade level. In addition, a meta-analysis (Stockford, 2009) of 176 ICCs from 63 studies (which were authored after 1985 with data collected in or after 1970, including data from national samples as well as smaller scale samples) employing multilevel models to analyze K-12 student achievement in the United States showed that an average ICCs from two-level models using either classroom or school as the cluster unit were about .22 with the ICCs ranging from .01 to .60. Guided by these findings, this dissertation study controled three levels of ICC: .00, .25, and .50. These target ICC values were controlled by controlling the ratio of the between-PSU variance to the within-PSU variance. Table 1 below describes the target ICCs and its associated population variance between and variance within.

Table 1

Controlled ICC Levels and Associated Population Variance

Controlled ICC	Variance Within	Variance between
ICC = 0.00	100	0
ICC = 0.25	75	25
ICC = 0.50	50	50

Population Density

Density refers to the extent of clustering in the samples. A sparse arrangement has many clusters, but each cluster has only a few observations. A dense arrangement has few clusters but the clusters are large**.** In this present study, the complex survey design has ten strata. The number of PSUs sampled from each of the ten strata were linked with the number of observations sampled from each PSU to provide different density levels: low density (100 PSUs per stratum with 10 – 30 observations per PSU) and high density (20 PSUs per stratum with 50 – 150 observations per PSU) samples. To obtain realistic samples in the Monte Carlo study, the number of observations per PSU was a random factor in the simulations. This combination of the number of PSUs with the average sample size per PSU provided consistent overall sample sizes across these two factors (i.e., a mean of 2000 observations per stratum for an average total sample size of 20,000 for each complex sample).

Population density is expected to affect the impact of clustering on the standard errors of parameter estimates. With few, large clusters (high density) the impact is greater than with many, small clusters (low density). The two levels, high and low, population densities were simulated by controlling the numbers of PSUs per stratum and number of observations per PSU.

Proportion of Missing Data

Five levels of missing data were simulated: 0%, 10%, 30%, 50%, and 70%. Within each of these levels (except the 0% missing level), 50% of the data missing data were selected at the observation level and 50% at the PSU level. Through this process, not only were entire PSUs completely removed from the simulated samples when listwise deletion was used, but the structure of the remaining PSUs was also altered. For example, some of the remaining PSUs lost some, but not all, observations, thus resulting in a reduced clustering effect, while some PSUs retained their original structure and number of observations.

Number of Replications

It is observed that the numbers of replications used in past simulation-based studies on MDTs range from as few as 50 replications (e.g., Roth & Switzer III, 1995) to 10,000 irterations (e.g., Cramner & Gill, 2012; Demirtas et al., 2006); and the numbers of 1,000 replications are commonly seen in past and current MDT studies (e.g., Kromrey & Hines, 1994; Collins et al., 2001; Zhu, 2014; Dohoo et al., 2016; Kuijk et al., 2016). Based on Robey's and Barcikowski's (1992) logic, a replication number of 5,000 would provide adequate precision of the estimate, e.g., providing maximum 95% confidence interval of ±.014 around an observed proportion. In addition, Mooney (1997) suggested that evaluating Type I error rate would require more replications than would evaluating a statistic's bias because the former deals with the tails of a distribution and the latter works on the central part; also, according to the SAS guide for Monte Carlo Studies (Fan et al., 2002), a number of 5,000 replications can provide "reasonable accuracy" in estimating Type I error in an ANOVA analysis. Drawing from these remarks, a

number of 5000 replications would be a sufficient number for evaluating bias. For this reason, the number of replications used in this present study was 5,000 iterations, which are also seen in previous studies on MDTs (e.g., Mukaka et al., 2016).

Data Generation

One of the practical concerns evident in applied data analysis that research on missing data treatment should address is the data matrices investigated should reflect realistic data encountered in actual field research (Kromrey & Hines, 1994). To increase the usefulness of the results of this dissertation study to applied researchers, the data in this dissertation study were simulated from actual data. This section describes the process of generating two-level data and creating missing data.

Two-Level Data Simulation

In this simulation study, the sample simulation included both stratification and cluster sampling. The sampling of PSUs was simulated from each of ten strata of which each stratum had different size and mean; then within PSUs, observations were simulated, controlling the ratio of the between PSU variance to the within PSU variance to produce desired intraclass correlation level. Depending on the population density, sample size drawn from each PSU were randomly determined resulting in approximately 20,000 observations per replication. Table 2 below describes the size and mean of ten strata from which observations were generated.

Table 2

Size and Mean of Strata Used Simulating Observations

Strata	1	2	3	4	5	6	7	8	9	10
Size	30,000	30,000	50,000	50,000	50,000	50,000	100,000	100,000	100,000	100,000
Mean	0	0.497	0.993	1.49	1.988	2.485	2.982	3.479	3.967	4.47

Within each PSU, multivariate normal data were generated using a correlation matrix derived from an actual matrix obtained from the NELS-88 survey (National Center for Educational Statistics [NCES], 2007). Specifically, the intercorrelations between eight predictor variables were taken directly from the NELS-88 results. Zero-order correlations with a hypothetical criterion variable were calculated so that the predictors provided a range of effect sizes in the eight predictor regression equation. Two predictors were generated to provide small (X7 and X8), medium (X1 and X2), and large (X3 and X4) effect sizes, respectively, as well as two predictors (X5 and X6) which were approximately null (i.e., regression coefficients were generated to be practically zero) in the multiple regression equation. In addition, the hypothetical criterion was generated so that there were differences in cluster mean based on the desired ICC of 0, .25, and .50; and also, the data were generated such that each regressor (X1, X2, X3, X4, X5, X6, X7, X8) has a different cluster mean. The correlation matrix used in the simulations is provided in Table 3, and the details of the data generation are provided in appendix A.

Observations within each sample were weighted so that the sample weight were proportional to the inverse probability of selection, taking into account the probability of PSU selection from the stratum and observation selection from the PSU (i.e., sample weight = 1 / (PSU_prob * observation_prob where PSU_prob is the probability of PSU selection from a stratum, and observation_prob is the probability of observation selection from this PSU) and the

sample weights were incorporated in subsequent analyses (e.g., in estimating population parameters).

Table 3

Correlation Matrix Used as Template for Data Simulation

	Y	X1	X2	X3	X4	X5	X6	X7	X8
Y	1.00000								
X1	0.29354	1.00000							
X2	0.28902	0.03716	1.00000						
X3	0.33003	-0.02342	-0.08097	1.00000					
X4	0.42926	0.02039	0.05139	-0.15033	1.00000				
X5	0.17179	0.04689	0.07601	-0.14001	0.40799	1.00000			
X6	0.05367	0.07268	0.11877	-0.21079	0.16350	0.25853	1.00000		
X7	0.10842	0.09224	-0.06382	0.11601	-0.05750	-0.10975	-0.20160	1.00000	
X8	0.15151	0.05810	0.21698	-0.13668	0.10849	0.17502	0.34115	-0.21985	1.00000

Missing Data Creation

Using the process described above, samples were simulated separately for each type of missingness (MCAR, MAR, and MNAR). Each observation in the simulated dataset has four flags which are indicators for four missing data levels (10%, 30%, 50%, and 70%). For example, if an observation has the flag variable miss1=1, the observation is being selected for missing at 10% missing level; similarly if an observation has the flag variable miss2 =1, the observation is being selected for missing at 30% missing level and so on. A complete dataset (i.e., 0% missing data) for each type of missingness was obtained by disregarding those missing flags.

Observations in the simulated samples were flagged as missing in two steps: missing at the cluster level (Level-2) step and missing at the observation level (Level-1) step. In the Level-2 step, entire observations of a missing PSU were flagged, and the total number of missing PSUs in a sample is the product of half of the missing proportion and total number of PSUs in the

sample (i.e., half of the missing data are missing at cluster level). Which PSU of which the entire observations were flagged depends on the type of missingness for which the sample was simulated. For example: (1) for Level-2 missing for MCAR data, a missing PSU was randomly selected to have missing data imposed; (2) for Level-2 missing for MAR data, a missing PSU was randomly selected to have missing data imposed, conditioned on the cluster mean value for variable X1 in the PSU; and (3) for Level-2 missing for MNAR data, a missing PSU was randomly selected to have missing data imposed, conditioned on the mean value of cluster means for variable X7 and X8 in the PSU (X7 and X8 are designed to have missing data in the incomplete samples). As regarding to the conditions for MAR missing, missing PSUs are more likely to occur with large cluster mean values for the regressor X1 in the PSU and less likely to occur with its small cluster mean values (e.g., 80% of the missing PSUs were randomly selected from the upper half of the distribution of cluster mean for variable X1, and 20% of the missing PSUs were randomly selected from the lower half of this distribution). Similarly, for the MNAR missing, missing PSUs are more likely to occur with large mean values of the cluster mean for regressor X7 and X8 and less likely to occur with their small mean values (e.g., 80% of the missing PSUs were randomly selected from the upper half of the distribution of the mean of the cluster means for variable X7 and X8, and 20% of the missing PSUs were randomly selected from the lower half of this distribution).

In the Level-1 step, the total number of observations being flagged is the product of half of the missing proportion and total number of observations in the current sample (i.e., half of the missing data are missing at observation level). In this step, observations that had not been flagged for missing at Level-2 missing were randomly selected for being missing in MCAR samples and randomly selected, conditioned on the values of X1 or conditioned on the mean

values of X7 and X8 in MAR or MNAR samples respectively. Similarly to the approach described in selecting missing PSUs above, in Level-1 missing for MAR data, missing observations are more likely to occur with large values of the regressor X1 and less likely to occur with its small values; and for MNAR data, missing observations are more likely to occur with large mean values of regressor X7 and X8 and less likely to occur with their small mean values. SAS programs to generate complete samples and impose according missing flags for MCAR, MAR, and MNAR data can be seen in appendix A which has three sections: MCAR Data, MAR Data, and MNAR Data.

Using the missing flags imposed during the data simulation process, incomplete samples were extracted from the simulated data in accordance to each study condition. Also, of eight independent variables (x1- x8) in the simulated dataset, two variables (x7 and x8) of the flagged observations have missing data; the resulted missing datasets were imputed using each of the MDTs mentioned above. A SAS program to extract incomplete samples from the simulated data can be found in appendix B.

Software Applications and Computing Systems Used

SAS software package was used to generate data, impute missing data as well as statistically analyze the data. A complete sample complex survey data as well as the samples of three different types of missing data (MCAR, MAR, and MNAR) were created using SAS/IML version 9.4. Each sample for the three missing data types was handled separately, using the five missing data treatment approaches mentioned in the Objective section. In order to evaluate the performance of each of these methods, multiple regression analyses were performed for each experimental condition: points and interval parameter estimates were obtained from the sample

data using the SURVEYREG procedure in SAS. To apply the Listwise deletion method, all observations with missing data were removed from the incomplete sample data before doing statistical analysis. To apply the hot-deck single imputation and hot-deck multiple imputation to impute the missing data, the SURVEYIMPUTE procedure (SAS/STAT® 14.1) was used; and to apply the regression single imputation and regression multiple imputation methods, the MI procedure (SAS/STAT® 14.1) was used to impute the missing data. Regarding the number of imputations in multiple imputation, in the past the default values implemented in the SAS® PROC MI was five, but in SAS/STAT 14.1, the default imputation number is 25. The imputation number of ten was used in this present study. This number was selected based on analyzing a simulated sample with the highest level of missing data (e.g., 70%) using 5 imputations, 10 imputations, and 25 imputations. It was found that the difference in the resulting point and interval estimates between 5 imputations and 25 imputations was noticeable, but the difference in those estimates between 10 imputations and 25 imputations was negligible.

Using the above approach for the multiple imputation method, an incomplete multivariate dataset was imputed ten times leading to ten imputed datasets to be used in statistical analysis; and regardless whether the missing data are imputed using hot-deck or regression imputation approach, analyzing multiply-imputed data followed two steps: within-imputation analysis and between-imputation analysis. In this present study, the SURVEYREG procedure was used to obtain the within-imputation estimates and the MIANALYZE procedure will be used to obtain the between-imputation estimates (i.e., pooling the estimates obtained from the within-imputation analysis). A SAS program to impute missing data, analyze the imputed data, and pool the results using the regression-based multiple imputation can be found in Appendix B. Below is a brief description for each of the mentioned software procedures.

SURVEYIMPUTE Procedure

SAS's SURVEYIMPUTE procedure assumes that data are MCAR and MAR (SAS/STAT® 14.1 User's Guide); it can be used for single-stage or multistage designs, with or without stratification, and with or without unequal weighting; it implements single and multiple hot-deck imputation and other variation of hot-deck imputation. For hot-deck imputation, the SURVEYIMPUTE procedure provides multiple donor selection techniques of which two methods were used in this present study: (1) the approximate Bayesian bootstrap selection (Runbin & Schenker, 1986) was used in hot-deck multiple imputation and (2) the probability proportional to weights selection (Rao & Shao, 1992) was used in single hot-deck imputation.

MI Procedure

The MI procedure is a multiple imputation procedure that creates multiply imputed data sets for incomplete p-dimensional multivariate data; it can be used to produce regression-based single imputation and multiple imputation. The MI procedure assumes that the data are from a multivariate distribution and contain missing values that can occur for any of the variables; the MI procedure in SAS/STAT® 14.1 can be used to impute MCAR, MAR and MNAR data. It is worth mentioning that PROC MI is a single-level multiple imputation procedure. To incorporate the complex sample design in the multiple imputation, a variable that represents the unique identifier for each PSU was included in the class statement of the procedure; this identifier is a composite variable representing the stratum number and PSU (see Berglund & Heeringa, 2014).

SURVEYREG Procedure

The SURVEYREG procedure was used to do regression analysis on the complete sample data as well as on the imputed sample data; the SURVEYREG procedure can handle complex survey sample designs with stratification, clustering, and unequal weighting; (the sample weights which were generated in the data generation process were utilized by this SURVEYREG procedure). It computes regression coefficients and provides significance tests for any specified estimable parameters. It used the Taylor Series approximation to estimate the sampling variances (Kiecolt & Nathan, 1985).

MIANALYZE Procedure

The MIANALYZE procedure combines the results of the analyses of imputations and generates valid statistical inferences; the procedure is designed to read results from different types of analysis (e.g., results from different types of regression analysis, correlation analysis). Specifically, in this study, the MIANALYZE procedure reads parameter estimates and associated standard errors and the covariance matrix that is computed by the PROC SURVEYREG for each of the ten imputed datasets in each studied condition then derives valid univariate inference for these parameters. The diagram in appendix C describes the general steps in doing multiple using SAS procedures. The two branches in the diagram are only different in the first step, which is the step of imputing missing data.

Conditions for the study were run under Linux platforms. Normally distributed random variables will be generated using the RANNOR random number generator in SAS. A different seed value for the random number generator was used in each execution of the program. The program code was verified by hand-checking results from benchmark datasets; for example, the

properness of the simulated missingness was verified using the correlations between the missing variables and the variables conditioning for the missingness; i.e., MAR data were verified based on the correlations between X1 and the missing of X7 and X8, and MNAR data were verified based on the correlations between the mean of X7 and X8 and the missing of X7 and X8. Table 4 summarized the obtained correlations for MCAR, MAR, and MNAR data: (1) in MCAR there was no correlation between X1 and the missing of X7 and X8; neither was there correlation between the mean of X7 and X8 and the missing of X7 and X8; (2) in MAR, the correlation between X1 and the missing of X7 and X8 was seen to be different for different levels of ICC (it should be mentioned that these correlations were not very strong because x7 and X8 were selected to be missing at both ends of the distribution of X1; (3) in MNAR the correlation between the mean of X7 and X8 and the missing of X7 and X8 was higher in higher ICC data, and these correlations were not very strong because x7 and X8 were selected to be missing at both ends of the distribution of mean of X7 and X8.

Table 4

Correlation between Missing Variable and the Variable Conditioning for the Missingness at 70% Missing Level

	MCAR		**MAR**		**MNAR**	
	ICC=0	ICC=.5	ICC=0	ICC=.5	ICC=0	ICC=.5
Correlation between X1 and the missing of X7 & X8	0.00	0.00	**.38**	**.47**	.06	.02
Correlation between mean of X7&X8 and the missing of X7 & X8	0.00	0.00	.05	.01	**.39**	**.49**

Evaluation Criteria

The performance of the MDTs was assessed by assessing the quality of the estimated multivariate regression coefficients resulted from analyzing the treated data obtained by each of

the MDTs. Four evaluation measures were used: statistical bias, root mean square error (RMSE), confidence interval coverage rate and confidence interval width. Results were presented using these four measures which will be discussed below.

Statistical Bias

Bias is the average difference between a single sample value and the true parameter. In this study, the smaller the bias is the better the performance of the data treatment method.

$$bias = \frac{\sum_1^n (\hat{\theta}_i - \theta\ )}{n}$$

where $\hat{\theta}_i$ is the parameter estimate obtained from i^{th} replication within each cell; θ is the corresponding true population parameter value; and *n* is the number of replications within each cell. The population regression parameters that were used to estimate bias in this study were obtained by computing the average of ten 1,000,000-observation samples for each ICC level (.00, .25, .50). Below are those obtained populations parameters that are expressed in form of regression equations (respectively, the parameters in equation (1), equation (2), and equation (3) were obtained using ICC level 00, .25, and .50).

Y = 0.2609645X1 + 0.2649556X2 + 0.4174972X3 + 0.461104X4 - 0.005698X5
- 0.008731X6 + 0.0887275X7 + 0.0993191X8 (1)

Y = 0.245915X1 + 0.248279X2 + 0.3884303X3 + 0.4452925X4 - 0.017297X5
- 0.026503X6 + 0.062703X7 + 0.0834575X8 (2)

Y = 0.236224X1 + 0.2375396X2 + 0.3697126X3 +0.4351108X4 - 0.024766X5
- 0.037948X6 + .0459445X7 + 0.0732439X8 (3)

RMSE

Root mean square error is the square root of the average squared difference between the estimate and the true parameter value; combining the concepts of bias and variability of estimation, RMSE is a measure of overall accuracy and precision of the estimate (i.e., small RMSE means small prediction error). It is desirable that an estimate is more accurate (i.e., less biased) and has minimum variance, so the smaller the RMSE the better the performance of the data treatment method.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta)^2}{n-1}}$$

where $\hat{\theta}_i$ is the parameter estimate obtained from i^{th} replication within each cell; θ is the corresponding true population parameter value; and n is the number of replications within each cell.

Confidence Interval Coverage

Generally, confidence interval coverage is computed as the percentage of replications within a design cell in which the confidence interval around a parameter estimate contained the true population parameter. In this study, the confidence interval coverage for the estimated regression coefficient parameter is computed as the percentage of 5,000 replications in which the 95% confidence intervals around the regression coefficient contain the true regression coefficient parameter. The data were generated with 95% confidence interval, hence the probability that the true regression coefficient falling in the confidence interval is expected to be 95%; so the closer the *coverage* is to 95% the better the data missing method treatment. In other words, a better

MDT should yield 95% confidence intervals that actually contain the true parameter about 95% of the time.

Confidence Interval Width

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, which is the true regression coefficient in this present study. CI width is computed as the distance between upper limit and lower limit (CI width = upper limit – lower limit). The narrower the CI width is the more precise the population estimate; so a better MDT is expected to yield a narrower CI width.

Data Analysis

Technically, this present study has three designs (one for MCAR, MAR, and MNAR). For each design there are three between-subject factors [ICC (0, .25, .50); population density (low, high); and proportion of missing data (0, .10, .30, .50, .70)] and one within subject factor [MDT (LW, HS, RS, HM, RM)]. Therefore the study is composed of three 3 x 2 x 5 x 5 designs. The sample estimate of the multiple regression coefficients was analyzed. The data were analyzed by computing the four aforementioned measures that are used to assess these sample estimates and that are obtained from the MDT conditions relative to the complete sample condition (e.g., no missing data). The analysis of the imputed data was done using multivariate regression model to estimate regression coefficients. The model used in analyzing the imputed data is of the same structure as the model used to impute the missing data (e.g., both have the same eight regressors with the same associated correlations between the regressors and the criteria variable).

All analysis were conducted and reported separately for each type of missingness, MCAR, MAR, and MNAR. For each type of missingness, the four evaluation criteria (bias, RMSE, CI coverage, and CI width) for the estimated regression coefficient across all conditions (missing levels, ICC degrees, population density degrees) were computed when the missing data are treated by each of the studied MDTs; the computations of these four evaluation criteria were also be performed on the complete data (i.e., no missing data – NM or missing level = 0%) which is used as the base for comparison.

To answer the first research question, which MDT produces more accurate and precise estimates when data are MCAR, MAR, and MNAR, box plots were used to illustrate the distributions of the four evaluation criteria yielded by each of the studied MDTs. Specifically, for each type of missingness (MCAR, MAR, MNAR), and for each evaluation criteria (bias, RMSE, CI coverage, or CI width), a box plot was used to describe the distributions of the evaluation criteria yielded by each of studied MDTs (LW, HS, HM, RS, RM) along with the distribution of each of the four evaluation criteria obtained from the complete data (i.e., NM).

To answer the 2nd research question of how the amount of missing data, Intraclass Correlation (ICC), and population density influence the accuracy and precision of parameter estimates produced by each MDT on each type of missingness, eta-squared analyses were used to identify the major influencing factors on the evaluation criteria (i.e., bias, RMSE, CI coverage, CI width), and Excel line charts were used to describes the results of the eta-squared analysis.

CHAPTER FOUR: RESULTS

The present study was designed to evaluate the performance of five MDTs for handling missing data in complex sample surveys. The five missing data methods were listwise deletion (LW), single hot-deck imputation (HS), single regression imputation (RS), hot-deck-based multiple imputation (HM), and regression-based multiple imputation (RM). These MDTs were assessed in the context of parameter estimates in multiple regression analysis in complex sample data with two data levels. Specifically, the evaluations were based on evaluating parameter point and interval estimates obtained from each of the studied MDTs; the four performance measures used in this study were statistical bias, RMSE, CI width, and coverage probability (i.e., 95%) of the confidence interval obtained from each of the MDTs. It should be noted that the missing data treatments do not account for the entire obtained bias, RMSE, and confidence interval coverage, but at least part of each of these measures are accounted by the sampling variability. The five MDTs were evaluated separately for three types of missingness: MCAR, MAR, and MNAR. For each type of missingness, the studied MDTs were evaluated at four levels of missingness (10%, 30%, 50%, and 70%) along with complete sample conditions as a reference point for interpretation of results. The performance measures were also obtained in three ICC levels (.0, .25, .50) and in high and low density population. This chapter reports the results of the performance measures obtained from each type of missingness: MCAR, MAR, and MNAR.

MCAR

Bias

Figure 1 shows the distribution of bias estimates for the five studied MDTs along with the bias estimates obtained from the no-missing data samples (NM) in MCAR. Across conditions in MCAR data, the estimated bias yielded by LW and the estimated bias obtained from the samples with no missing data (NM) were nearly identical. In addition, bias estimates yielded by HS were slightly larger than LW bias estimates whereas bias estimates produced by RM, HM and RS appeared to vary more in both directions with their negative bias showing greater variability. Also, among the two multiple imputation methods, RM appeared to produce more negative outliers than HM; and among the five MDTs, RS appeared to produce the largest and most varied bias.

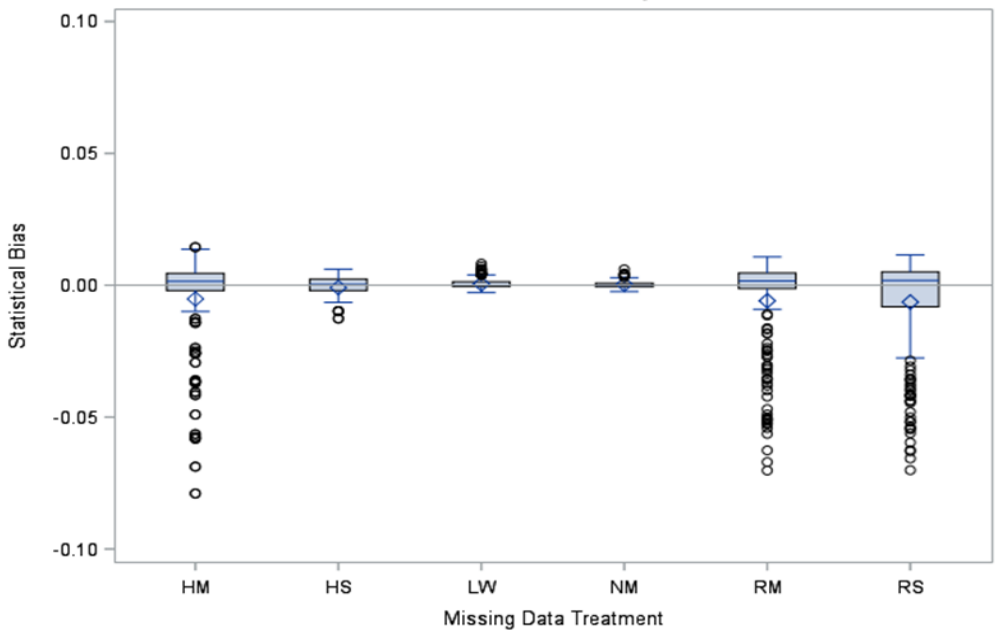

Figure 1. Distributions of Bias Estimates by Missing Data Treatment in MCAR

An eta squared analysis revealed that (1) the percent of missing data, by itself, had either negligible or no influence on the point estimates bias produced by most of the studied MDTs in MCAR data except for the HM method. Specifically, percent of missing data was found accounting for 3% of variability in bias estimates produced by HM, 0% of variability in bias produced by HS, 0.3% of variability in bias produced by LW, 0.4% of variability in bias produced by RM, and 0.2% of variability in bias produced by RS. (2) The major influencing factor on the bias produced by all the studied MDTs except LW was the parameter (e.g., the differences among the parameter estimates for variables with missing data vs. variables without missing data); parameter accounted for most of the variability in bias yielded by HM ($\eta 2 = 67\%$), SH ($\eta 2 = 72\%$), RM ($\eta 2 = 76\%$), and RS ($\eta 2 = 83\%$). (3) Interaction between ICC and parameter was the most influencing factor on bias produced by LW ($\eta 2 = 34\%$), and it was also the second most influencing factor on bias produced by HS ($\eta 2 = 20\%$), by RM ($\eta 2 = 18\%$), and by RS ($\eta 2 = 12\%$).

Figures 2, 3, and 4 describe the impacts of the top influencing factor on bias estimates obtained from RS, RM, and HM which are three of the MDTs producing the highest bias measures. In all three figures, bias estimates associated with missing-data regressors (X7 and X8) were larger and varied more compared to those associated with non-missing-data regressors; minimal bias was also seen with X4 and X5 in all three graphs.

Also as mentioned above, the interaction effect between percent missing and parameter was the 2nd most influencing factor on the bias estimates yielded by HM, and the interaction between ICC and parameter was the 2^{nd} most influencing factor on the bias estimates yielded by RM, and RS. Figure 5 describes the interaction effect of percent missing and parameter on bias

estimates yielded by HM: the negative bias associated with X7 and X8 increased as the percentage of missing increased.

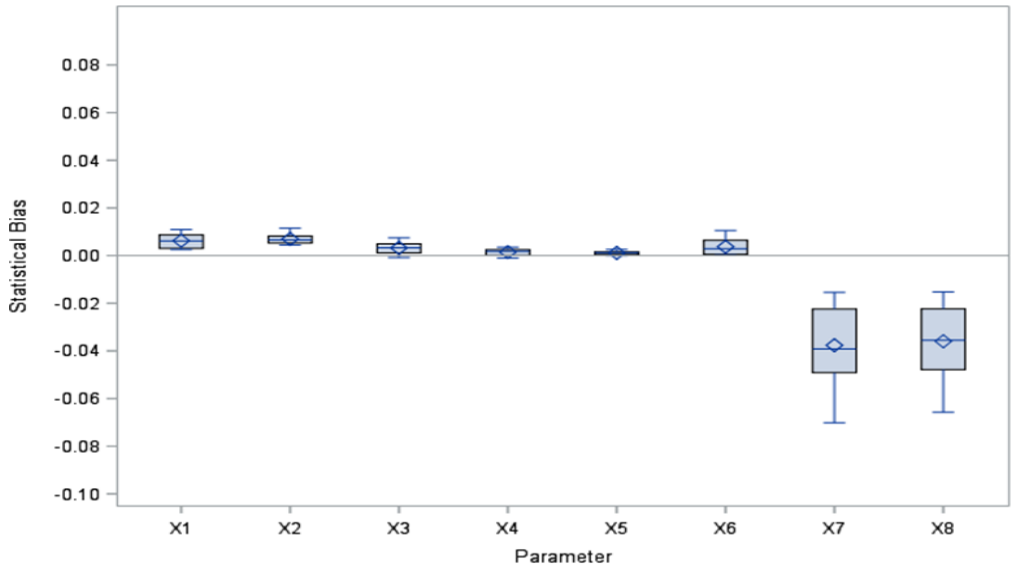

Figure 2. Distributions of Bias Estimates by Parameter for RS in MCAR

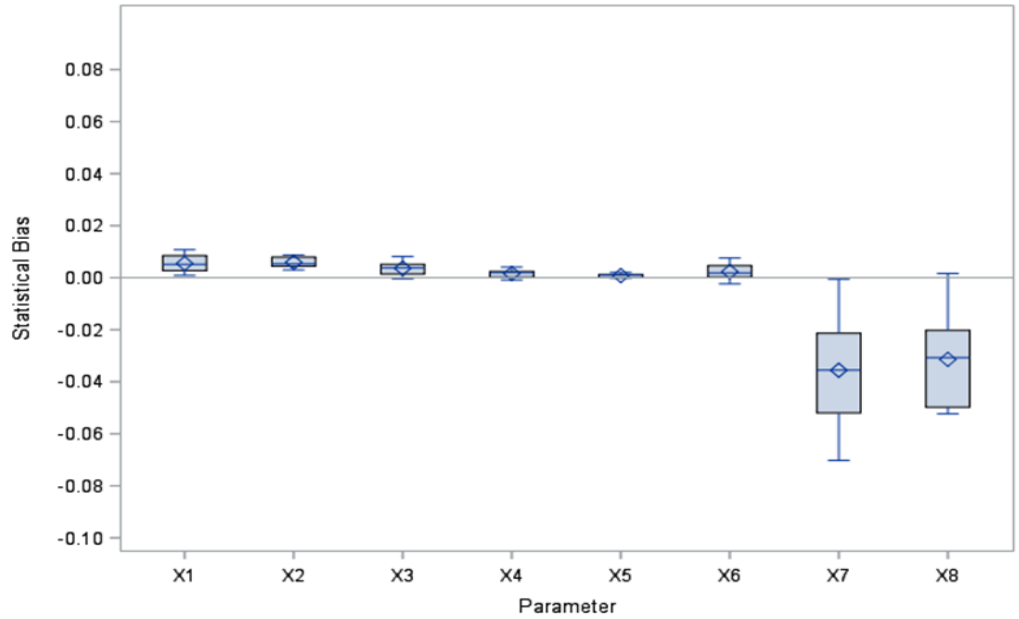

Figure 3. Distributions of Bias Estimates by Parameter for RM in MCAR

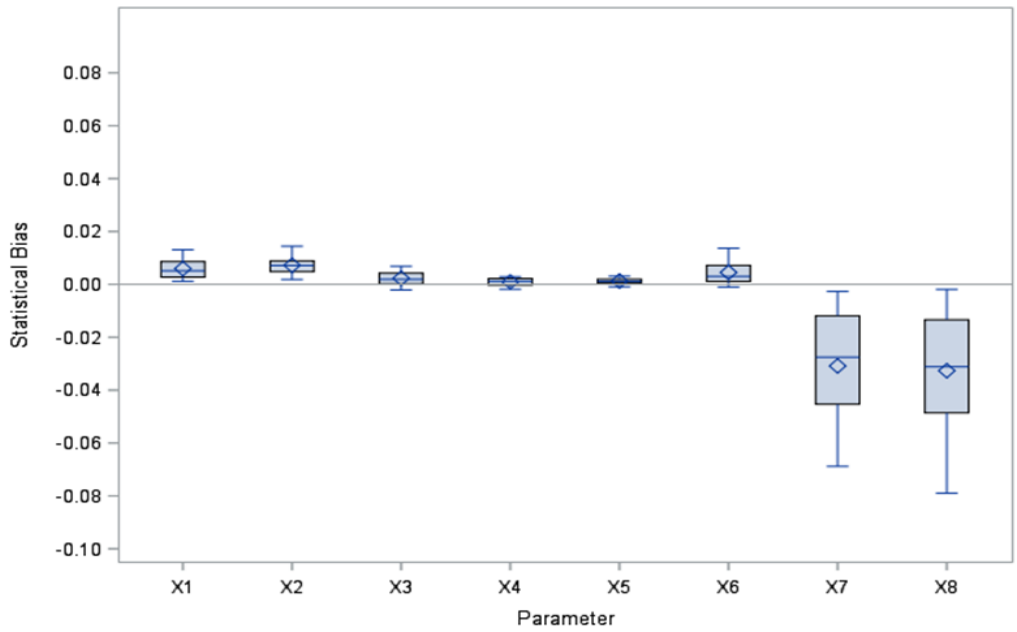

Figure 4. Distributions of Bias Estimates by Parameter for HM in MCAR

Figure 6 describes the interaction effect of ICC and parameter on bias estimates yielded by RS: negative bias associated with X7 and X8 decreased as the ICC level increased. This interaction effect also influenced the bias yielded by RM in a similar manner (i.e., negative bias associated with X7 and X8 decreased as the ICC level increased).

RMSE

RMSE estimates associated with HS and LW appeared more comparable to RMSE obtained from the complete samples (NM); and RMSE for HS appeared to be smallest and vary least among the five MDTs, whereas RMSE for RM and RS appeared to be the largest and varied most (Figure 7).

Results from an eta squared analysis on RMSE were similar to those obtained from the eta squared analysis on bias on at least two aspects: (1) the percent of missing data, by itself, had

either weak or negligible influence on the RMSE produced by the studied MDTs in MCAR. Specifically, for those MDTs associated with larger RMSE such as HM, RM and RS, the percent of missing data, respectively, accounted for 6%, 0.02% ,and .8% of variability in RMSE obtained from these MDTs. (2) The major influencing factor on RMSE produced by most studied MDTs (except LW and HS) was parameter; the parameter accounted for most of the variability in RMSE yielded by RM ($\eta 2 = 63\%$), by RS ($\eta 2 = 60\%$) and HM ($\eta 2 = 24\%$). However, RMSE produced by HS and LW were most influenced by ICC, with SH ($\eta 2 = 61\%$) and LW ($\eta 2 = 64\%$).

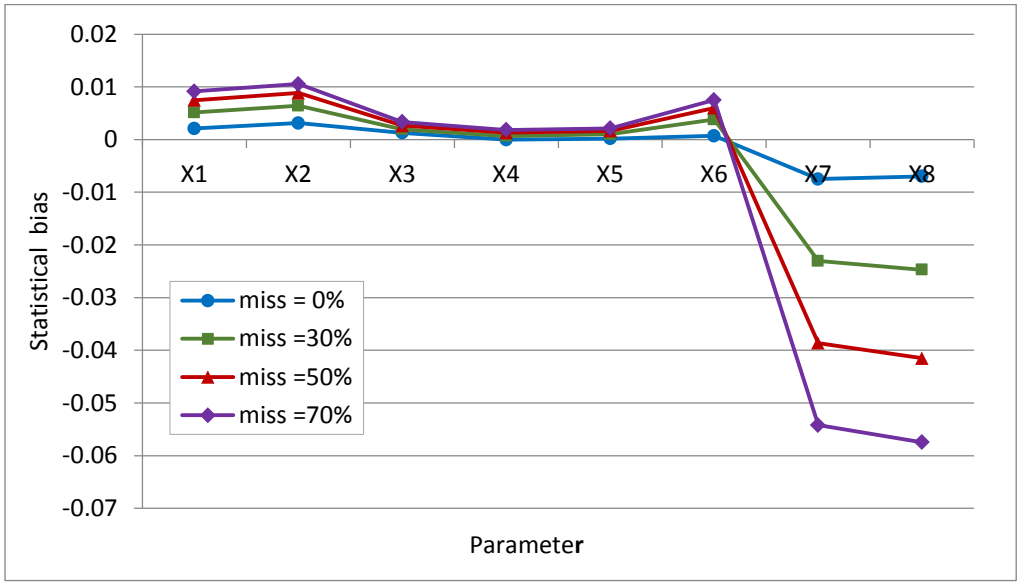

Figure 5. Mean Estimated Bias by Percent Missing Data and Parameter for HM in MCAR

Figure 8, 9, and 10 describe the distribution of RMSE by parameter for HM, RM, and RS respectively. In all three cases the RMSE estimates associated with X7 and X8 were larger compared to those associated with the other regressors.

RMSE estimates associated with the above three MDTs were also modestly influenced by the interaction effect between ICC and parameter; Figure 11, 12, and 13 describe the impact of

this interaction effect on RMSE produced by HM, RM, and RS respectively: generally, RMSE associated with parameters of non-missing-data variables increased as ICC level increased; however, RMSE associated with parameter of missing-data variables decreased as ICC levels increased.

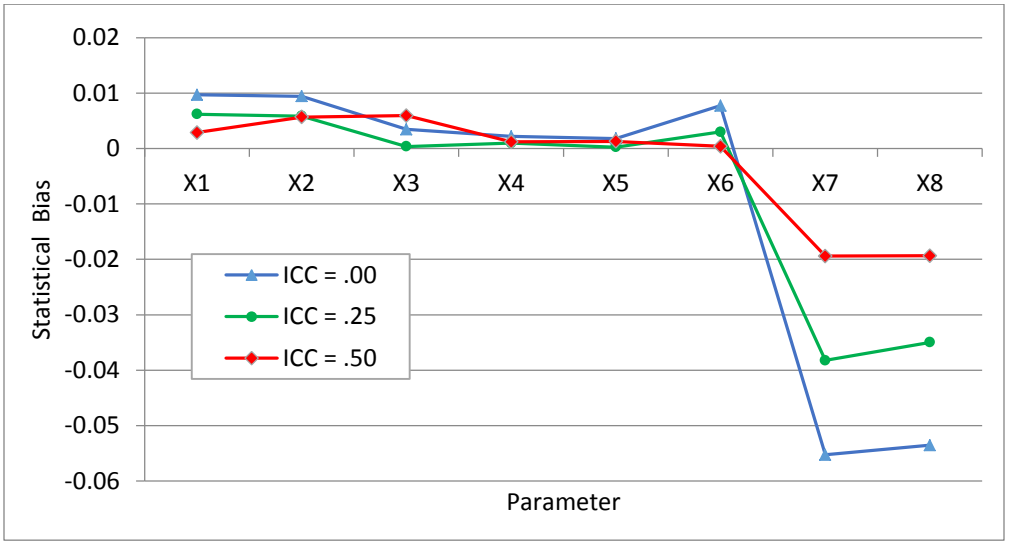

Figure 6. Mean Estimated Bias by ICC and Parameter for RS in MCAR

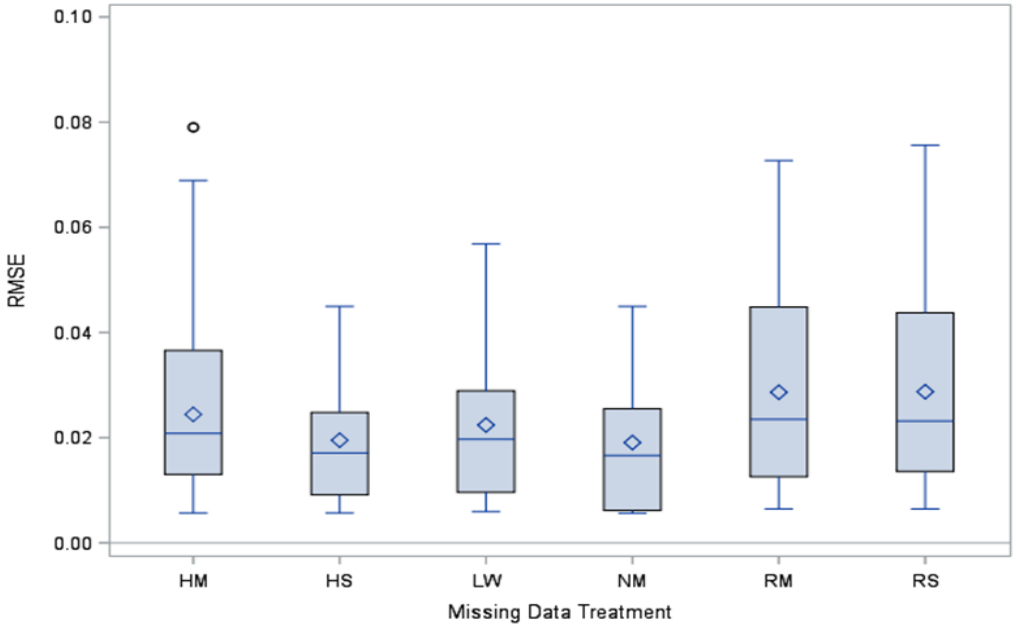

Figure 7. Distributions of RMSE by Missing Data Treatment in MCAR

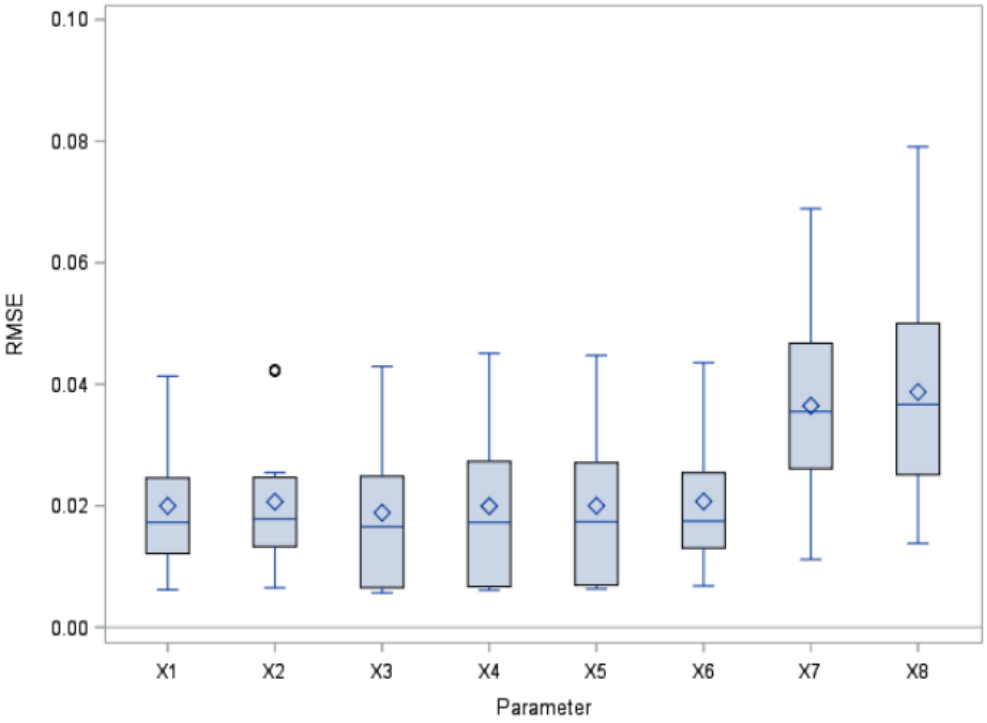

Figure 8. Distributions of RMSE Estimates by Parameter for HM in MCAR

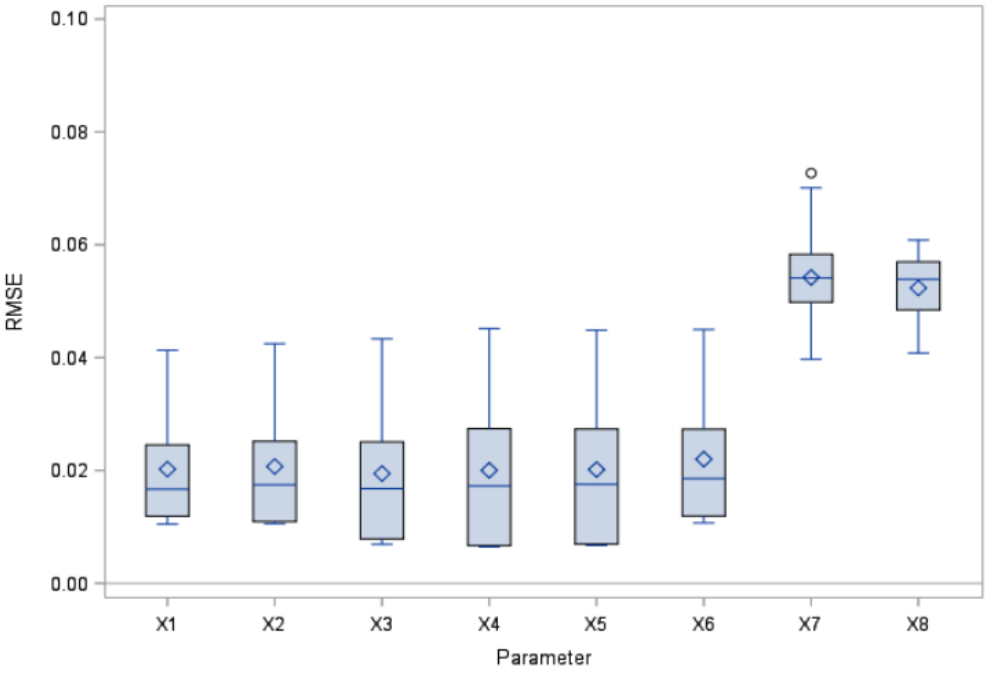

Figure 9. Distributions of RMSE Estimates by Parameter for RM in MCAR

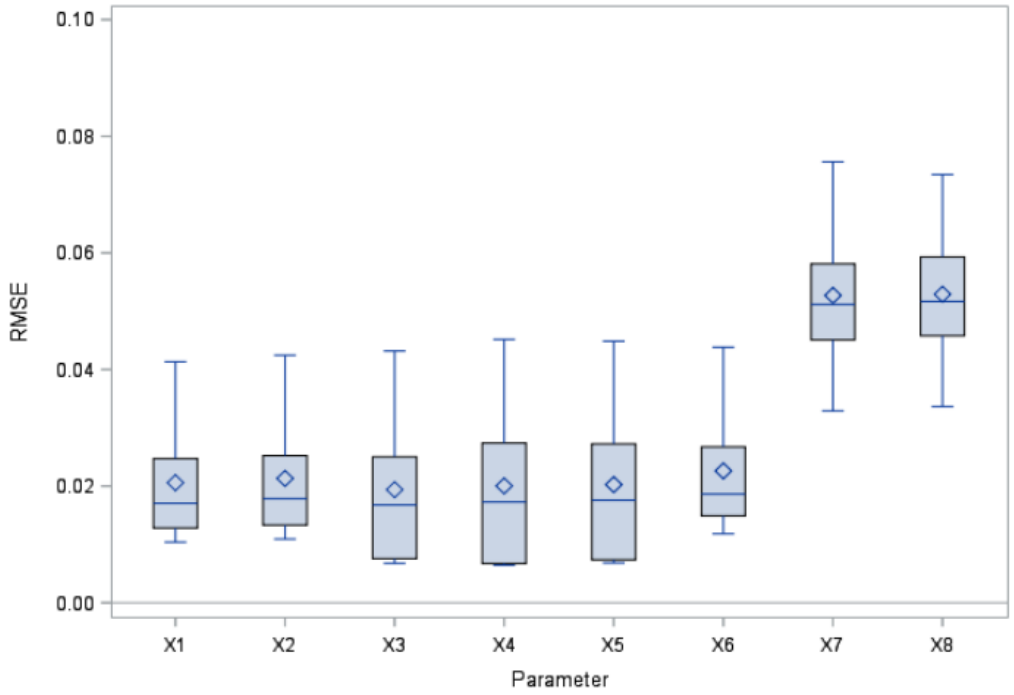

Figure 10. Distributions of RMSE Estimates by Parameter for RS in MCAR

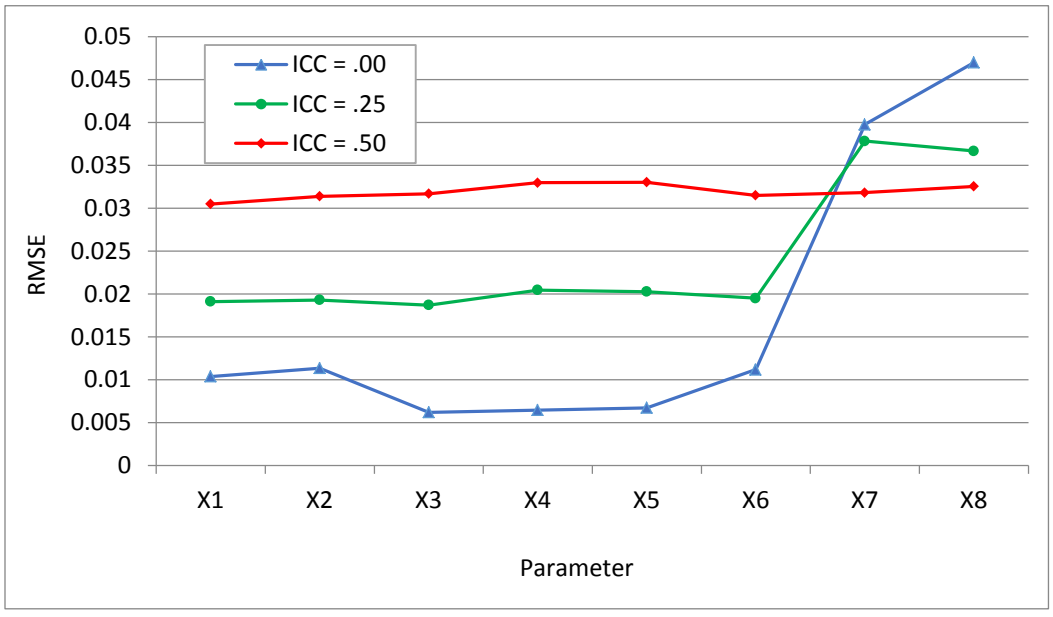

Figure 11. Mean Estimated RMSE by ICC and Parameter for HM in MCAR

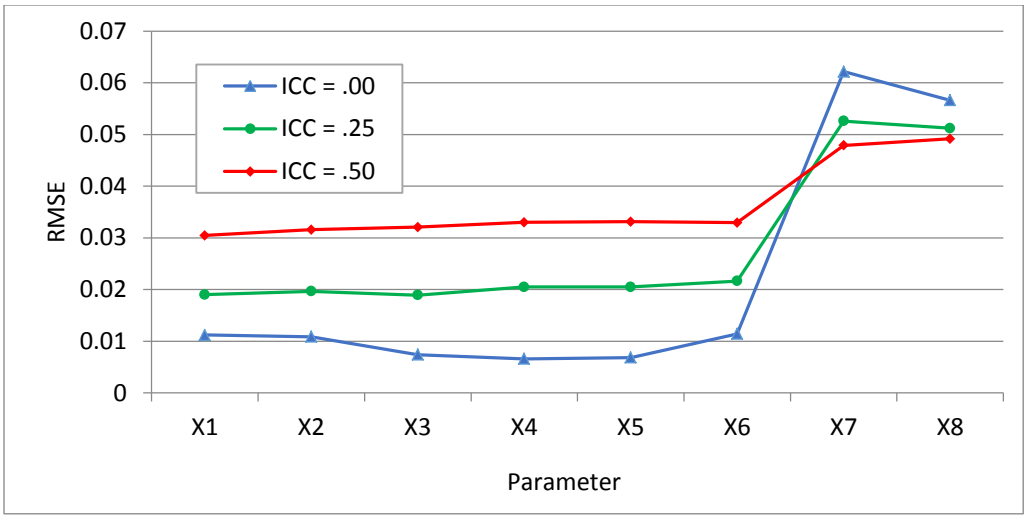

Figure 12. Mean Estimated RMSE by ICC and Parameter for RM in MCAR

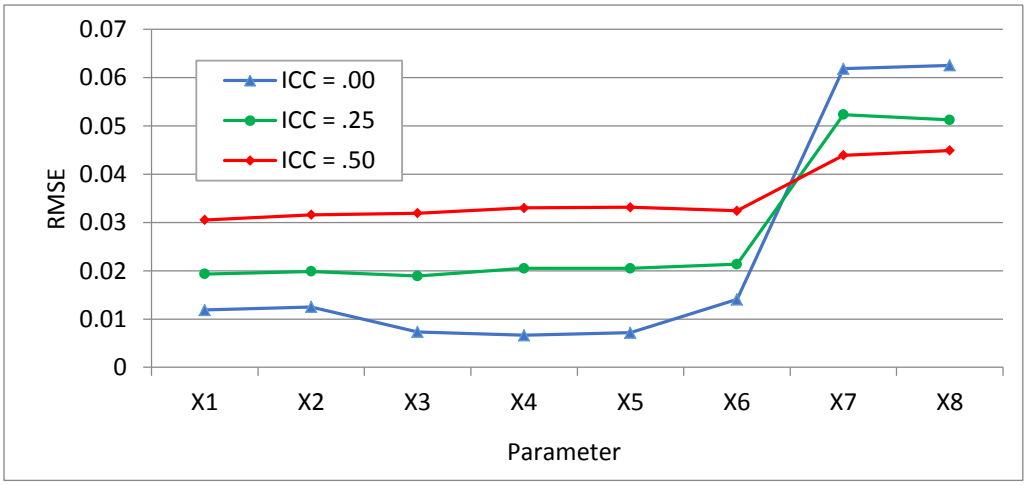

Figure 13. Mean Estimated RMSE by ICC and Parameter for RS in MCAR

CI Width

Regarding the precision measure estimates produced by the five MDTs, CI width estimates produced by HM, HS, and RS appeared comparable, and they were narrower than those produced by LW and RM, and CI width estimated by RM was the widest (Figure 14).

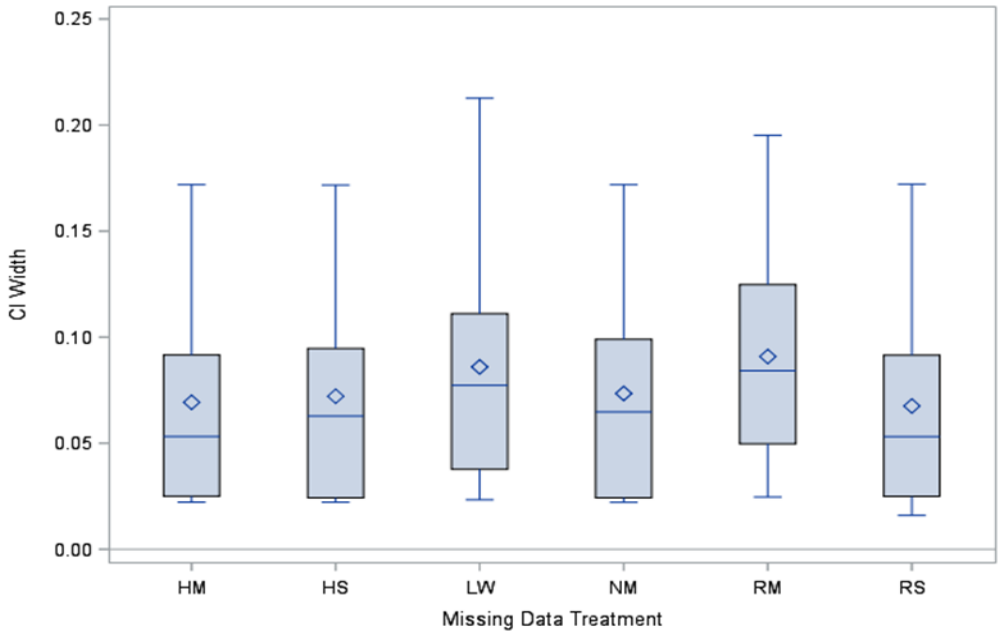

Figure 14. Distributions of Confidence Interval Width by Missing Data Treatment in MCAR

Via an eta squared analysis, it was found that ICC and population density, respectively, were the most and 2nd most influencing factors on CI width estimates produced by all MDTs in MCAR except RM. Each of these two factors influenced the CI width estimates produced by each MDT methods at similar degrees; for example, ICC accounted for 61% of variability in CI width estimates produced by HM, 66% of those produced by HS, 65% of those produced by LW, and 60% of those produced by RS; and population density accounted for 19% of variability in CI width estimates produced by HM, 20% of those produced by HS, 20% of those produced by LW, and 20% of those produced by RS. In addition, CI width estimate produced by RM was also comparably impacted by three factors: the interaction effect between ICC and parameter ($\eta 2$ = 25%), parameter ($\eta 2$ = 24%), and ICC ($\eta 2$ = 22%).

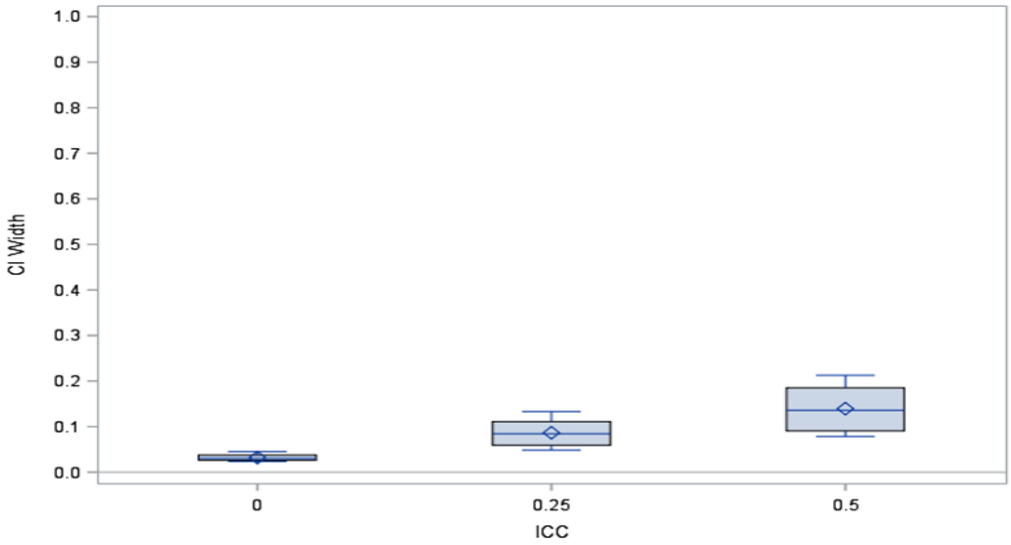

Figure 15. Distributions of Confidence Interval Width by ICC for LW in MCAR

As seen in Figure 14, RM and LW produced larger CI width estimates than other MDTs in MCAR data; for the CI width estimates produced by LW, it was found that the estimates increased as the ICC level increased (Figure 15); for the CI width estimates produced by RM, the estimates associated with non-missing data regressors were higher in higher ICC level and lower in lower ICC level; but the estimates associated with missing-data regressors (X7 and X8) were higher in zero-ICC data than those in non-zero ICC data (Figure 16).

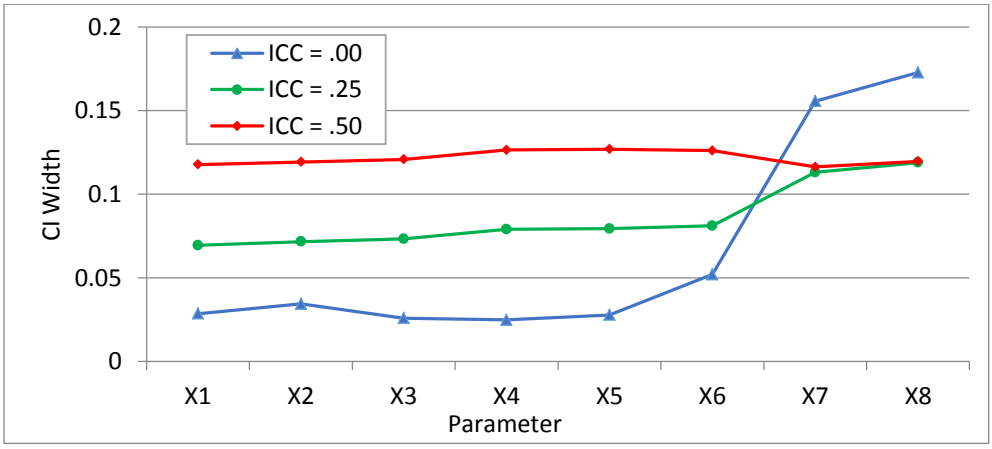

Figure 16. Mean Estimated CI Width by ICC and Parameter for RM in MCAR

CI Coverage

As for the confidence interval coverage measure, LW produced the highest coverage estimate which was about 95% (similar to those estimated from the complete samples, NM), and except for a few outliers, coverage estimates produced by HS were over 90% whereas RS produced the lowest and most varied coverage estimates. Of the two multiple imputation methods, RM produced higher and less varied coverage estimates than HM, but the coverages associated with these both MI methods varied below the nominal level by a considerable amount. These features can be seen in Figure 17 below.

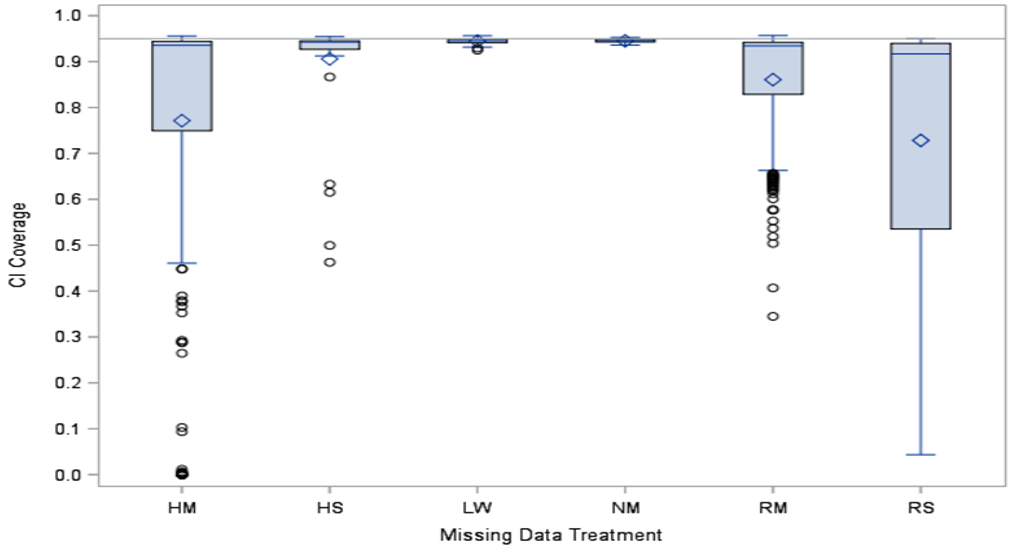

Figure 17. Distributions of Coverage Probability Estimates by Missing Data Treatment in MCAR

An eta squared analysis showed that the top two influencing factors on coverage estimated by RS was parameter ($\eta 2$=72%) and ICC ($\eta 2$=14%). Figure 18 describes the distribution of coverage estimates by parameter for RS method: coverage estimates associated with X7 and X7 varied widely at substantially low coverage levels; coverage estimates

associated with X1, X2, and X6 also varied widely but at higher coverage levels; and coverage estimates associated with X4 appeared consistent at about 95%.

Parameter was also the most influencing factor on coverage estimates produce by RM ($\eta 2$=74%), but the manner of the influence of parameter on the coverage estimates produced by the two regression imputation methods, RS and RM, were quite different. Figure 19 describes the distribution of coverage estimates by parameter for RM method. Compared to Figure 18, Figure 19 shows that coverage estimates associated with each regressor were a lot higher and varied less than those shown in Figure 18.

From Figure 17, it was also seen that HM yielded the 2nd lowest coverage estimates in MCAR data. Parameter, ICC, and the interaction effect of missing level and parameter, respectively, were the most ($\eta 2$=45%), 2nd most ($\eta 2$=13%), and 3rd most ($\eta 2$=12%) influencing factors on coverage estimated by HM. Figure 20 describes the distribution of mean coverage estimates by parameter and levels of missing data for HM in MCAR data: except for X3, X4, and X5, the coverage estimates decreased as missing data level increased.

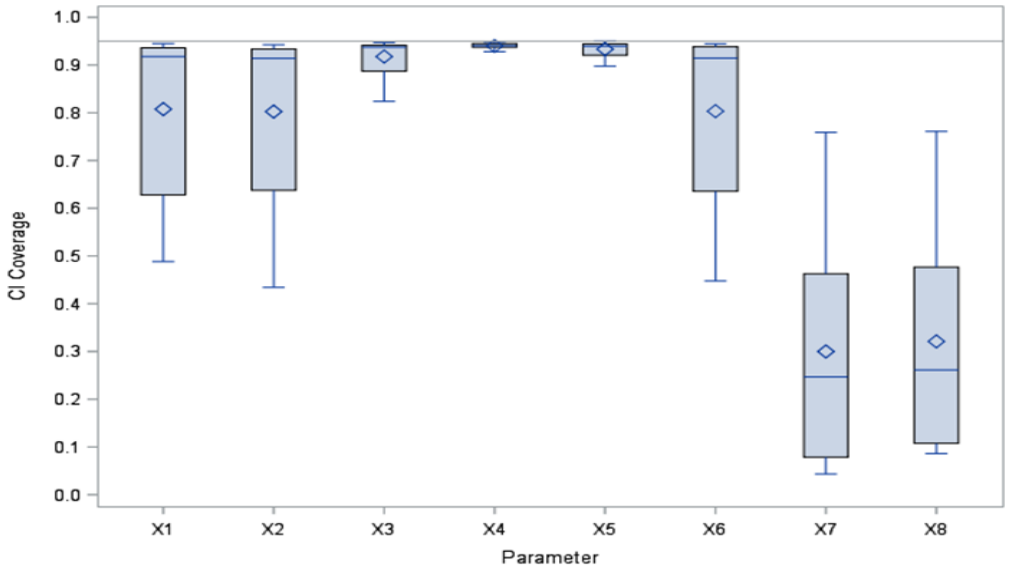

Figure 18. Distributions of Coverage Estimates by Parameter for RS in MCAR

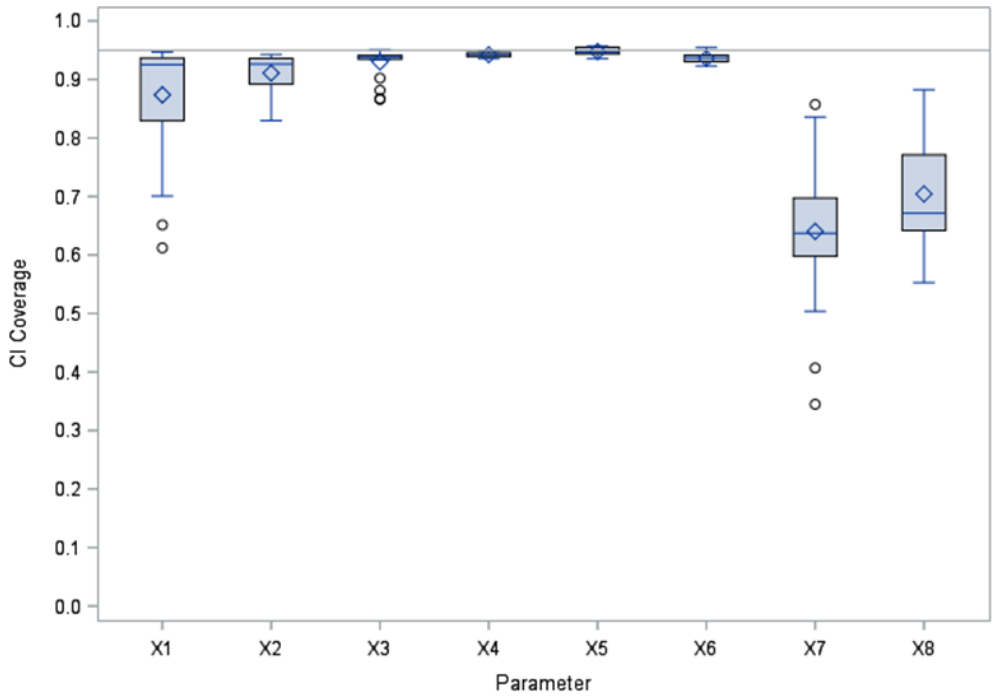

Figure 19. Distributions of Coverage Estimates by Parameter for RM in MCAR

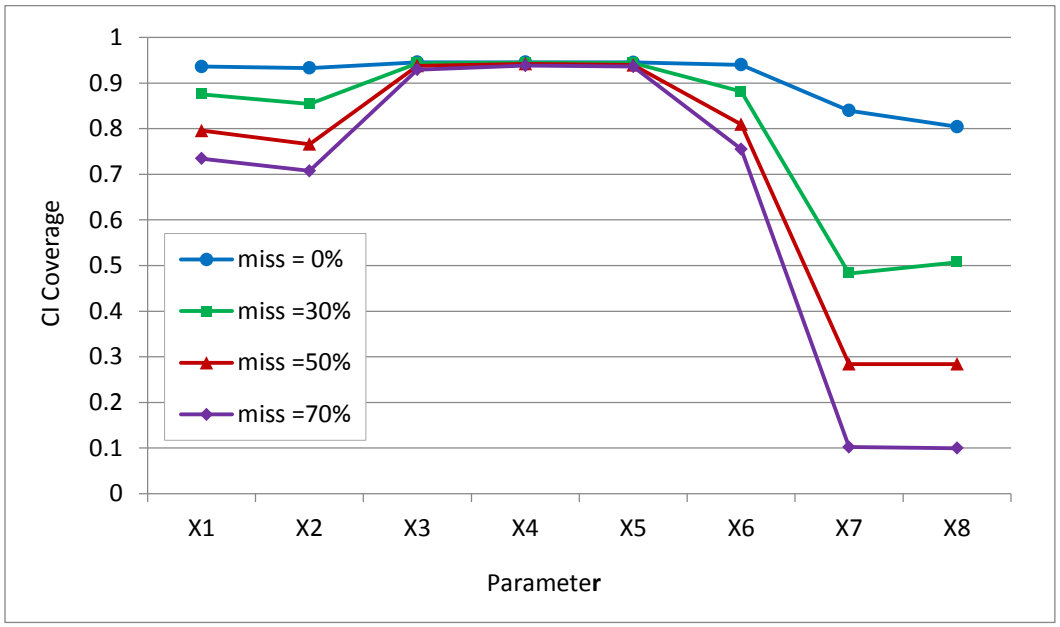

Figure 20. Mean Estimated Coverage by Percent Missing and Parameter for HM in MCAR

MAR

Bias

Across conditions in MAR data, the estimated bias yielded by LW and HS, except for few outliers, were close to bias estimates obtained from the complete samples (NM). However, compared to HS, LW produced outliers in both directions (which was not seen in MCAR). Also, in MAR, RS produced the largest and most varied bias estimates, and RM and HM produced comparable bias estimate in terms of magnitude and variability (Figure 21). In general, compared to the results obtained in MCAR, LW produced a few more outliers in both directions in MAR while HS appeared to produce consistent bias measures in MCAR and MAR; and RS produced largest and most varied bias estimates in both data types.

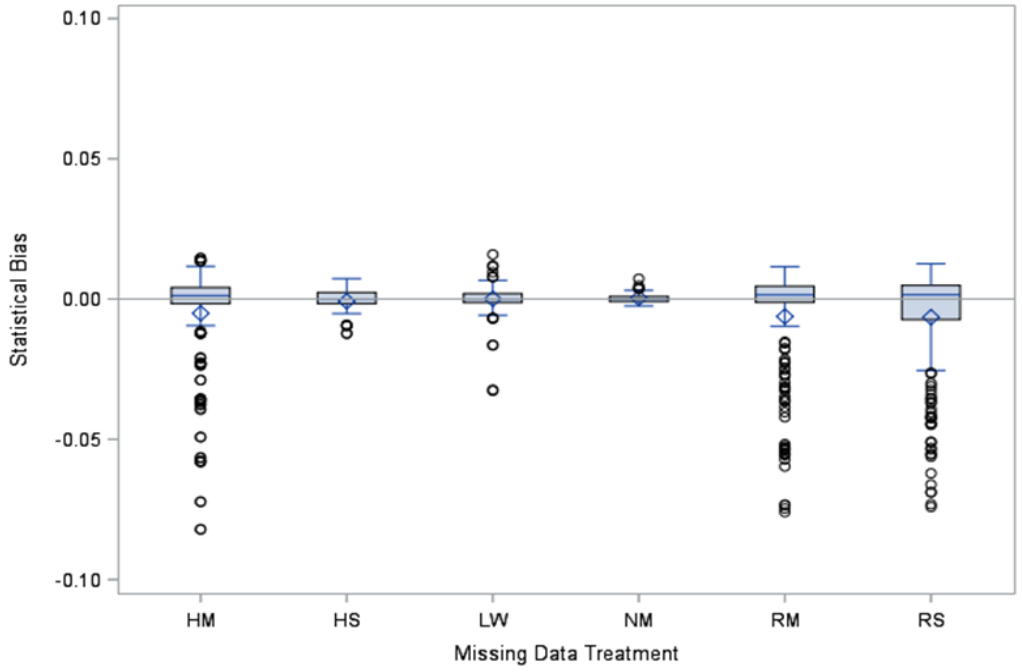

Figure 21. Distributions of Bias by Missing Data Treatment in MAR

As for the major factors influencing the bias estimates in MAR, parameter remained the most influencing factor on bias estimates produced by HM ($\eta 2$=64%), HS ($\eta 2$=66%), RM ($\eta 2$=71%), and by RS ($\eta 2$=80%). The general pattern of the distribution of bias estimates by parameter for RS, RM, and HM in MAR was similar to what seen in MCAR, specifically for all three methods, bias associated with X7 and X8 was negative and larger in magnitude and variability than those associated with non-missing-data regressors of which bias were positive. Figure 22, 23, 24 describes the distributions of bias by parameters for RS, RM, and HM respectively.

In MAR, the interaction effect between ICC and parameter also remained the most influencing factor on bias estimates produced by LW ($\eta 2$=31%) as previously seen in MCAR, but the bias estimates occurred in both directions with a few more outliers than what found in MCAR data. Figure 25 describes the mean estimated bias by ICC and Parameter for LW in MAR: for most parameters, negative bias was associated with mid ICC level (ICC=.25), and positive bias was associated with high ICC level (.50). At zero level ICC, there was no or minimal bias for all parameters except for X1. In fact, there was no specific pattern of association between ICC level and bias estimates for each particular regressor (e.g., the impact of the interaction effect of the two factors seems to be random).

As found in MCAR, the interaction effect between percent miss and parameter also had a modest impact ($\eta 2$=22%) on bias estimates produced by HM. Figure 26 describes the mean bias estimates for HM by missing levels and parameter: bias associated with missing-data regressors (X7 and X8) increased downward as percent of missing increased.

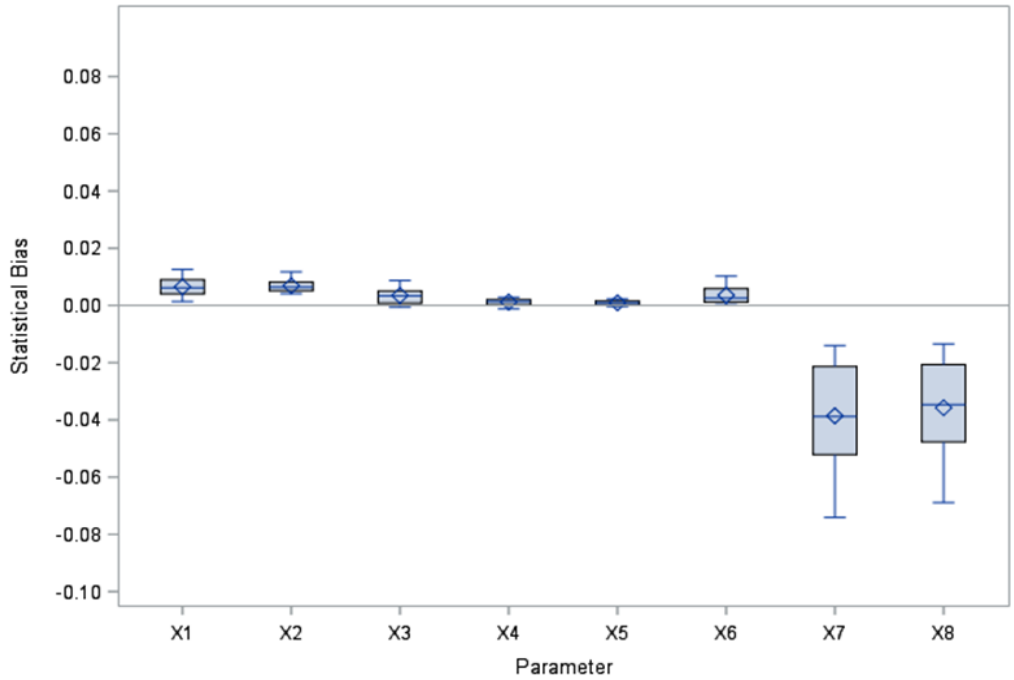

Figure 22. Distributions of Bias Estimates by Parameter for RS in MAR

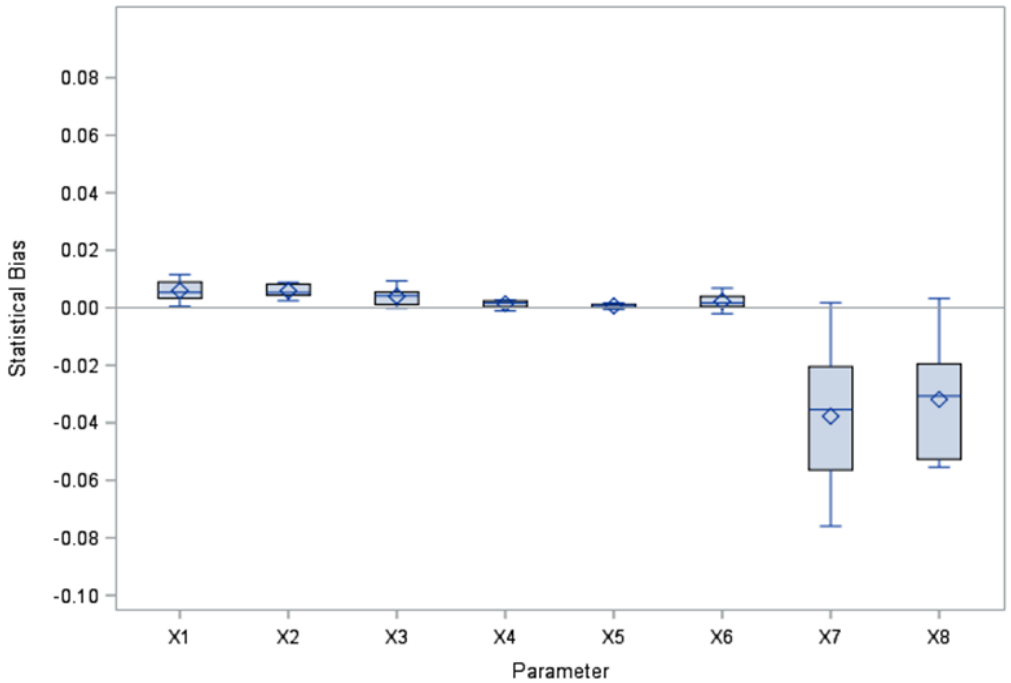

Figure 23. Distributions of Bias Estimates by Parameter for RM in MAR

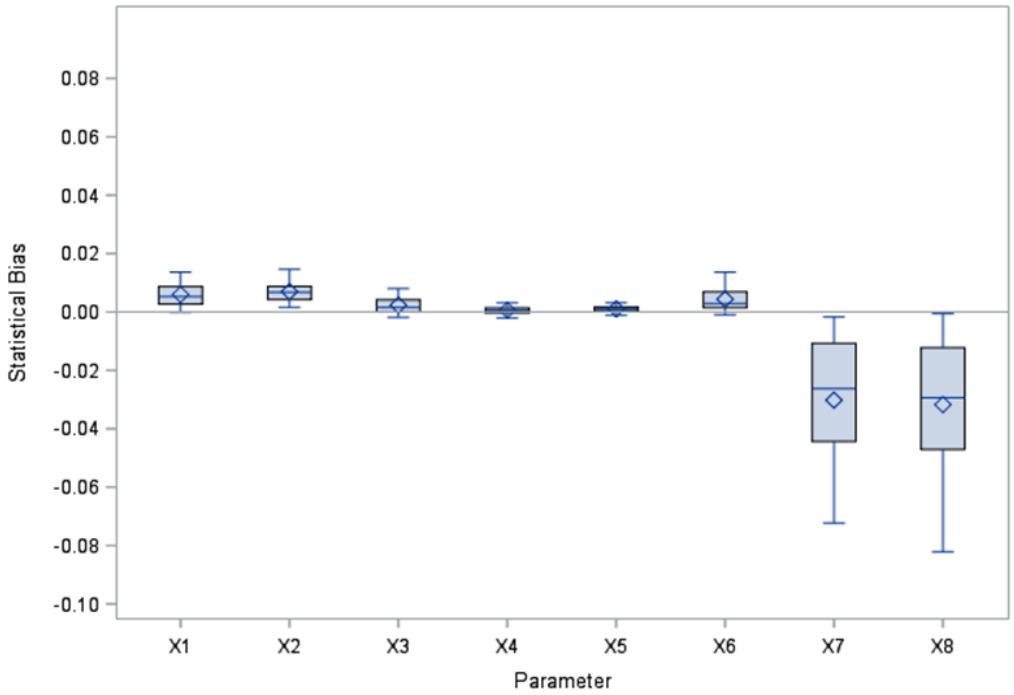

Figure 24. Distributions of Bias Estimates by Parameter for HM in MAR

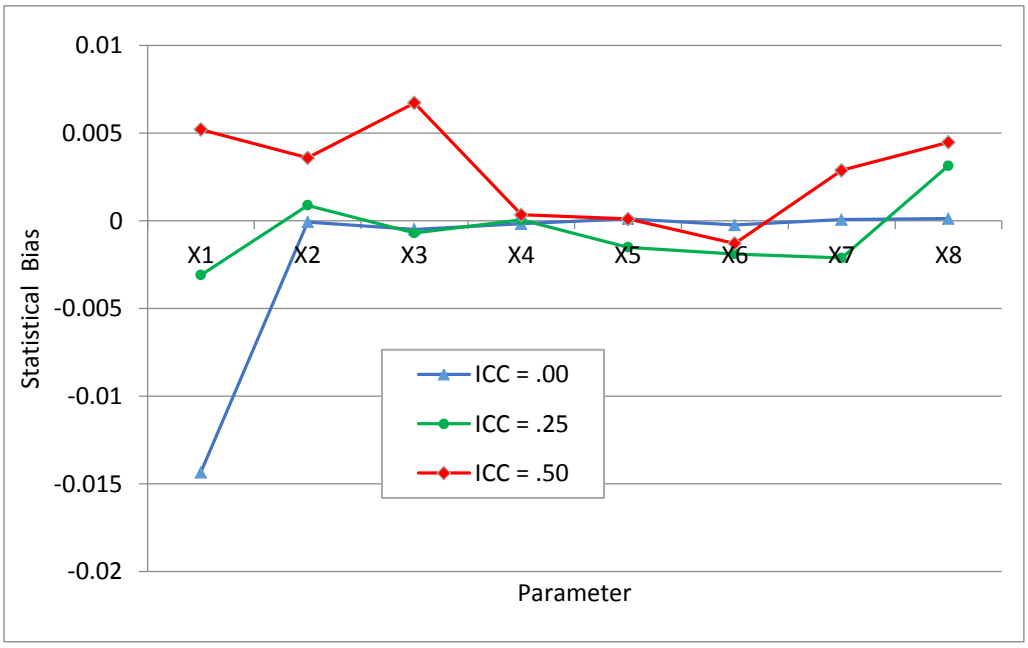

Figure 25. Mean Estimated Bias by ICC and Parameter for LW in MAR

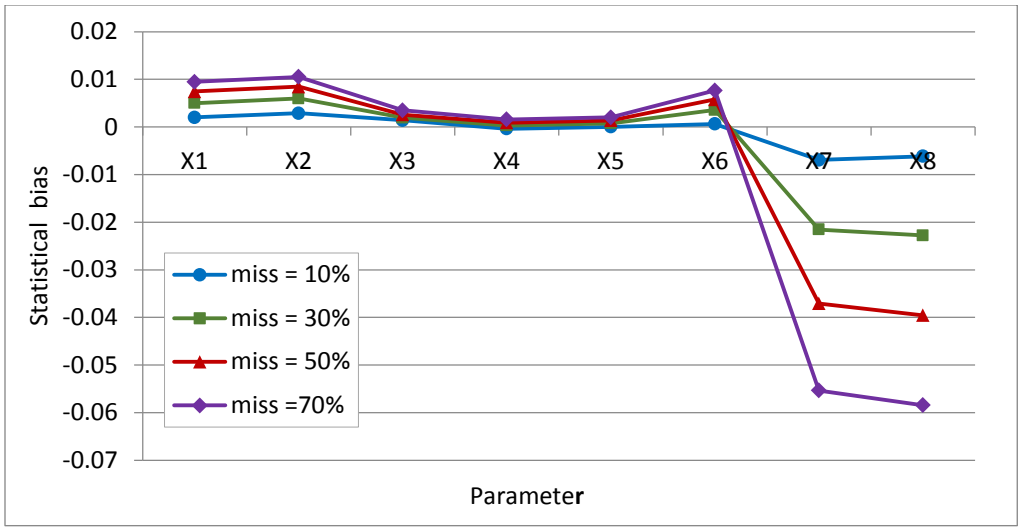

Figure 26. Mean Estimated Bias for HM by Percent Missing Data and Parameter in MAR

RMSE

Figure 27 shows the distribution of RMSE estimated by the studied MDTs along with the estimates obtained from the complete samples (NM): RMSE estimates obtained from HS was comparable to those obtained from the complete samples, and it appeared to be smallest and least varied compared to those obtained from the other MDTs; whereas RMSE estimates yielded by RM and RS appeared to be the largest and varied most.

For the major factors influencing RMSE estimates produced in MAR, as found in MCAR data, parameter was the most influencing factor on RMSE estimates produced by RM ($\eta 2$=64%), HM ($\eta 2$=24%), and RS ($\eta 2$=60%); and ICC was the most influencing factor on RMSE estimates produced by SH ($\eta 2$=61%) and LW ($\eta 2$=57%). Figure 28, 29 shows the distribution of RMSE by parameter for RM and RS respectively. In both cases, RMSE associated with missing-data regressors (X7 and X8) were larger than those associated with non-missing-data regressors.

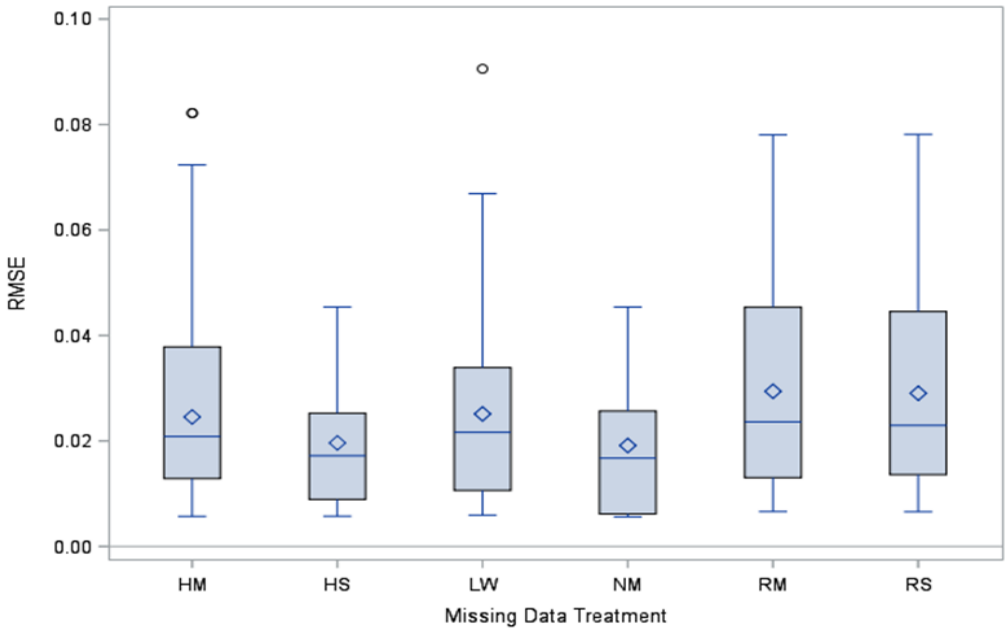

Figure 27. Distributions of RMSE by Missing Data Treatment in MAR

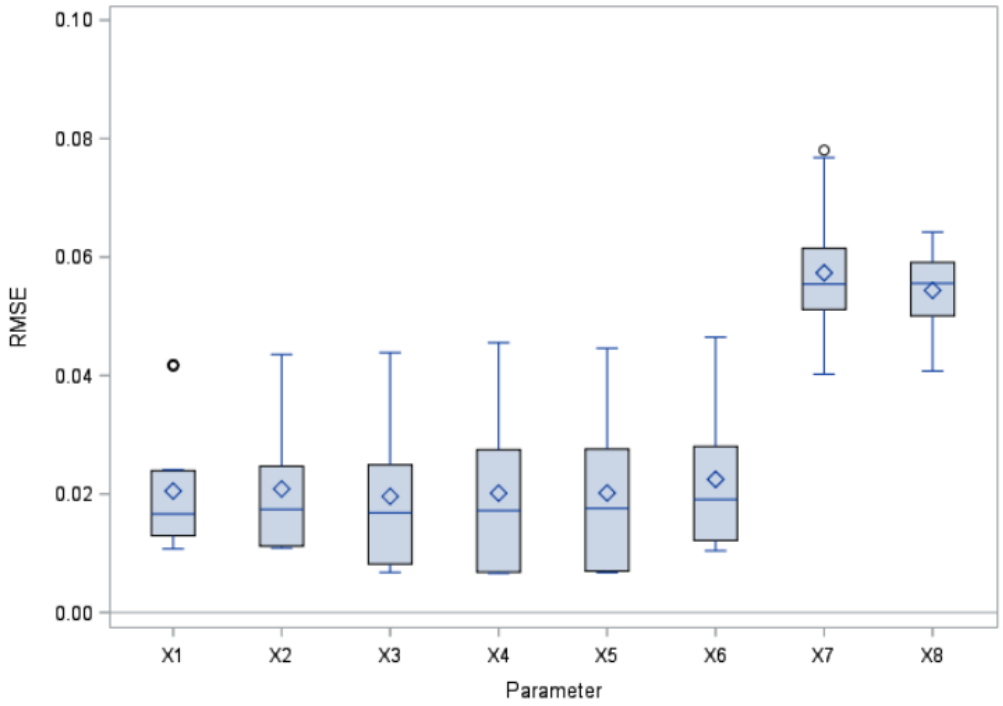

Figure 28. Distributions of RMSE Estimates for RM by Parameter in MAR

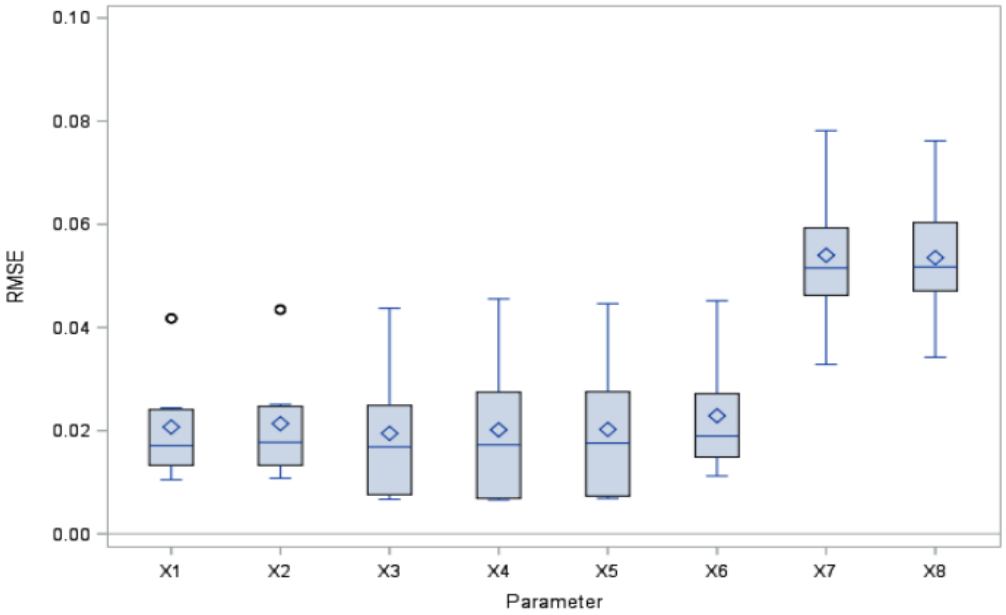

Figure 29. Distributions of RMSE Estimates for RS by Parameter in MAR

Figure 30 describes the distributions of RMSE estimates by ICC for LW: larger and more varied RMSE estimates were associated with larger ICC.

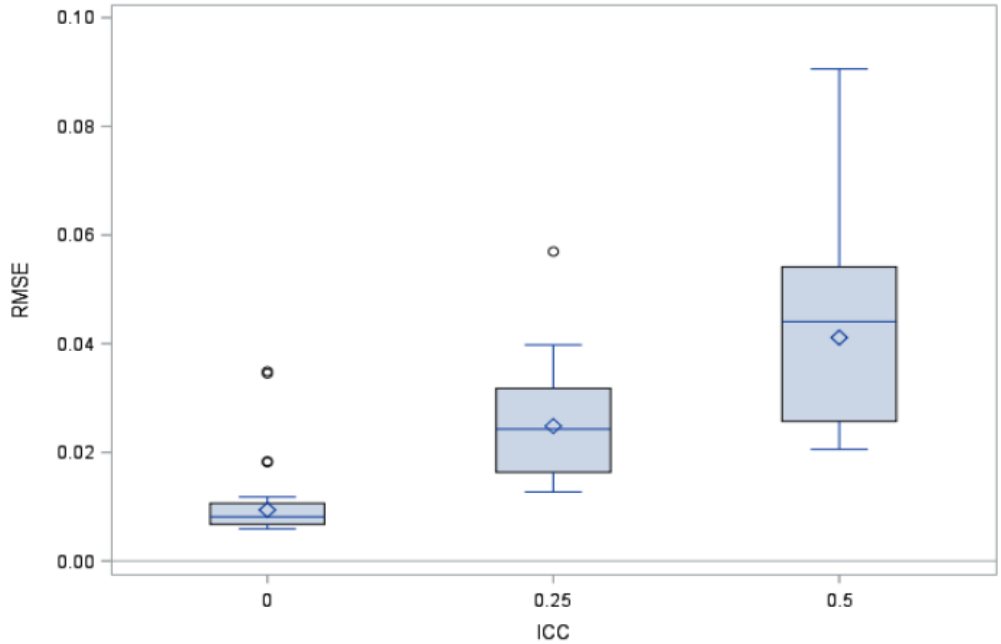

Figure 30. Distributions of RMSE Estimates for LW by ICC in MAR

CI Width

In MAR data, CI width estimated by LW was the widest and varied most while CI width estimates obtained from HM, HS, and RS were narrower and comparable to those obtained from the complete sample, NM (Figure 31).

Considering the major influencing factors on CI width estimates obtained in MAR data, ICC was the top influencing factor on CI width estimated by all of the studied MDTs: HM ($\eta 2$=63%), HS ($\eta 2$=66%), LW ($\eta 2$=62%), RM ($\eta 2$=26%), and RS ($\eta 2$=61%). For LW which produced the largest CI width estimates, the estimate increased in magnitude and variation as ICC level increased (Figure 32). Population density and the interaction effect between ICC and population density were also the 2nd most and 3rd most influencing factors on CI width estimated by LW.

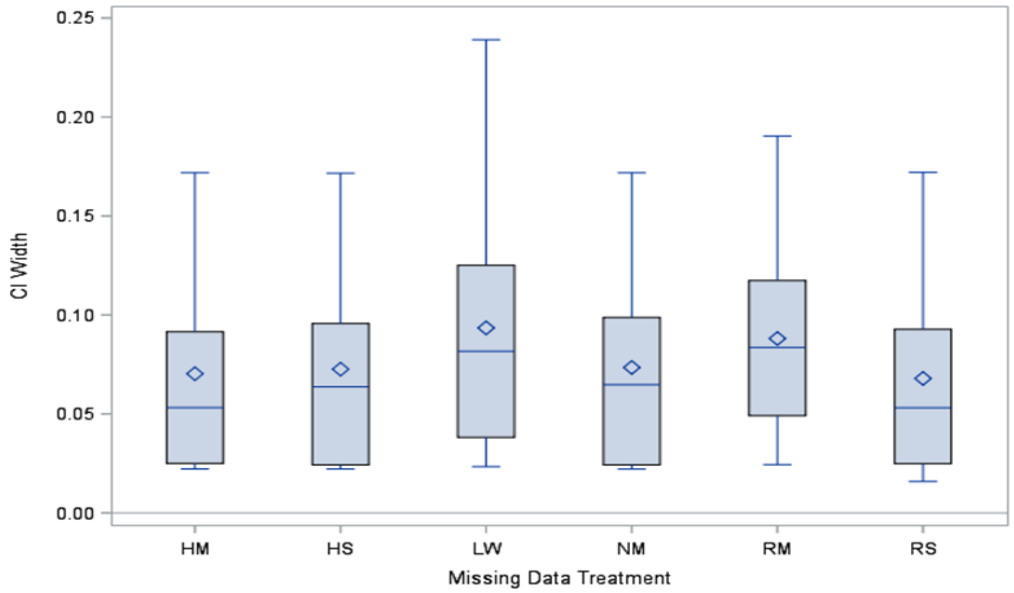

Figure 31. Distributions of Confidence Interval Width by Missing Data Treatment in MAR

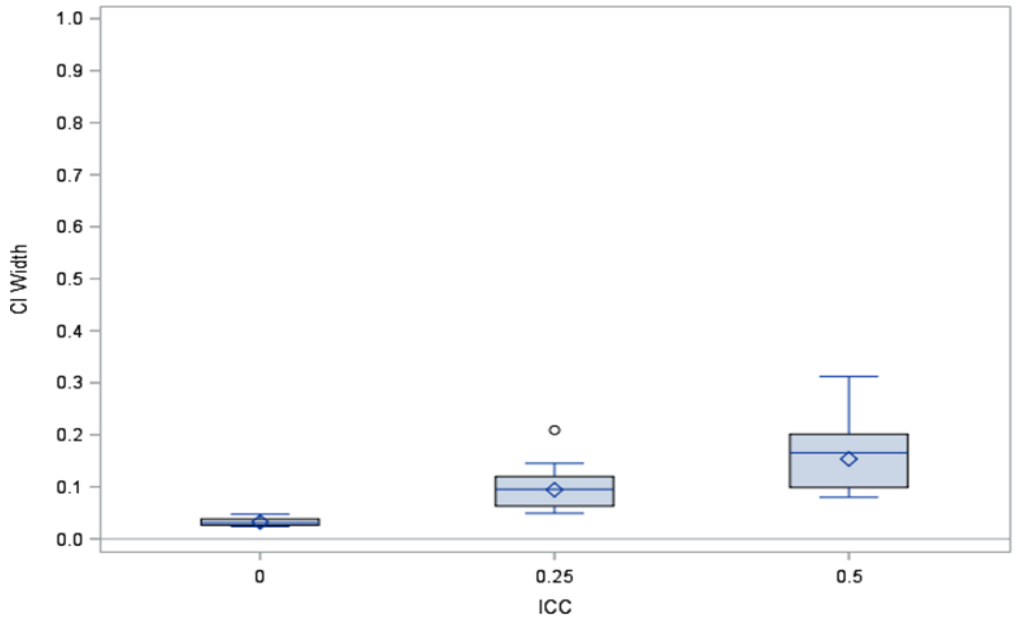

Figure 32. Distributions of Confidence Interval Width by ICC for LW in MAR

Figure 33 describes the mean estimated CI width for LW by population density and ICC in MAR: CI width estimates for high and low population density increased as ICC increased; at zero ICC level, there was no difference in CI width for high and low density population; however at higher ICC levels, CI width for high population density was higher than those for low population density.

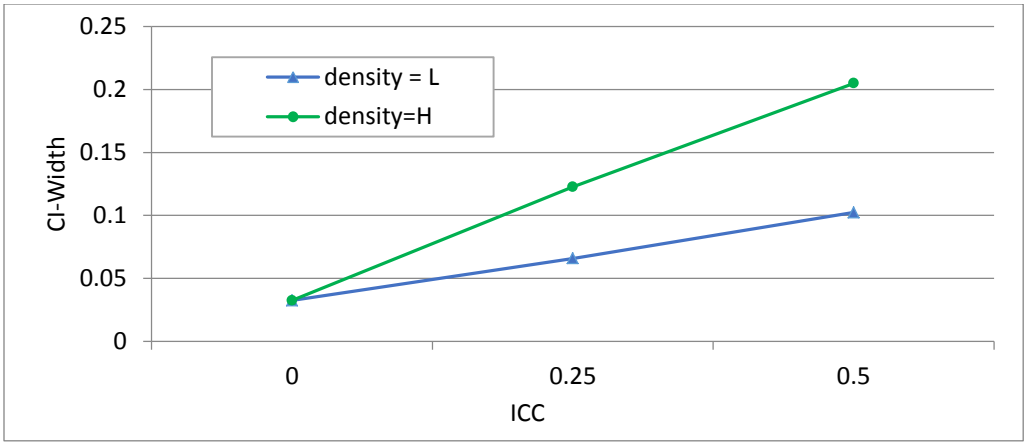

Figure 33. Mean Estimated CI Width by Population Density and ICC for LW in MAR

CI Coverage

As for the coverage probability measures estimated in MAR, except for a few outliers, LW remained producing the highest coverage estimate (about 95%, which was similar to those estimated from the complete samples, NM); and coverage estimates produced by HS, except for a few outliers, were also over 90%, whereas RS produced the lowest and most varied coverage estimates (similar to what found in MCAR data). In addition, coverage estimates produced by the two MI methods HM and RM were comparable and varied well below the nominal level, but HM produced more outliers than RM. These features can be seen in Figure 34 below.

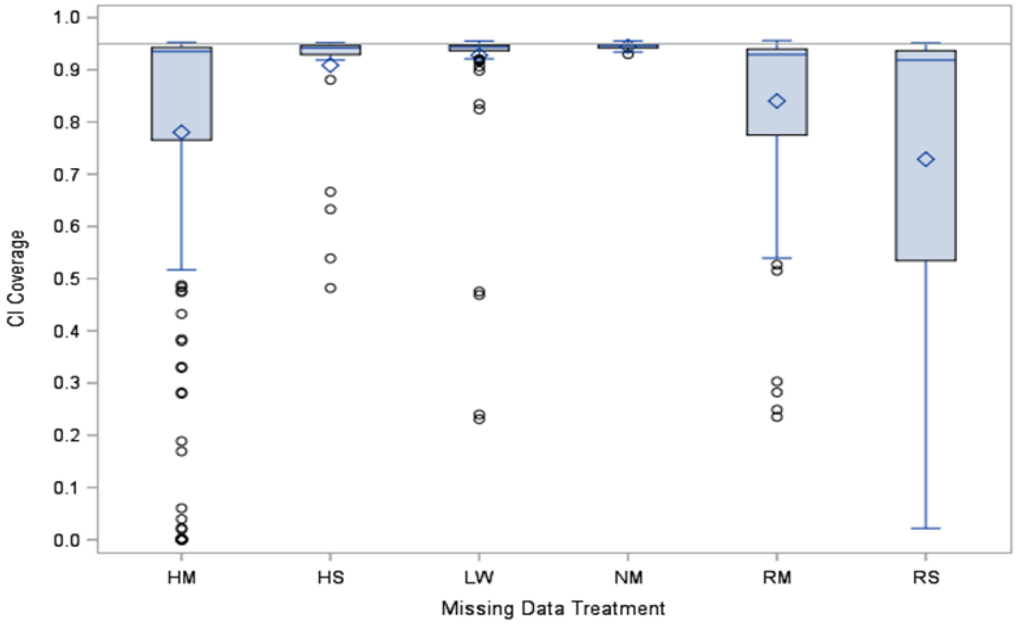

Figure 34. Distributions of Coverage Probability Estimates by Missing Data Treatment in MAR

As mentioned above, coverage estimates for RS were the lowest; the top three influencing factors on coverage estimates produced by RS were parameter ($\eta 2$=69%), ICC ($\eta 2$=16%), and the interaction between parameter and ICC ($\eta 2$=8%). Figure 35 describes the mean coverage estimates by parameter and ICC for RS in MAR data: In zero ICC data, coverage

estimates associated with all parameter were lower than those in non-zero ICC data, and coverage estimates associated with X7 and X8 were the lowest. In non-zero ICC data, coverages estimates associated with non-missing regressors appeared comparably high regardless of ICC levels; but coverage associated with missing-data regressors were low in low ICC level and high in high ICC level.

Overall, in MAR data the patterns of the influence of the simulation factors on the evaluation measures yielded by each of the five MDTs are, in general, similar to those found in MCAR data except that the factors influencing the coverage probability yielded by LW were different from those found in MCAR: (1) in MAR, ICC had a weak influence (η2=3%) on coverage yielded by LW whereas this effect was moderate (η2=31%) in MCAR; (2) parameter had a modest effect (η2=18%) on this measure in MAR but it had a negligible effect (η2=0.9%) in MCAR; and (3) density had almost no effect (η2=0.2%) on this measure in MAR, but it had a modest effect (η2=21%) in MCAR.

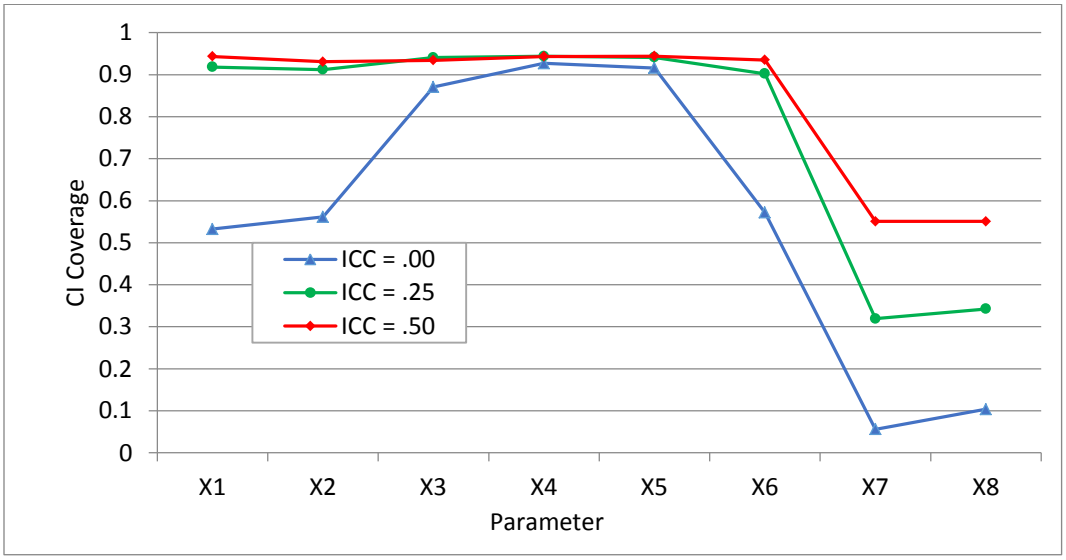

Figure 35. Mean Estimated Coverage by Parameter and ICC for RS in MAR

MNAR

Bias

Across the conditions in MNAR data, virtually the bias estimates yielded by each of the five MDTs were similar to those found in MAR data except that LW appeared to produce less positive outliers. Besides that, bias estimates yielded by LW and by HS were still close to those obtained from the complete samples (NM) with LW producing few more outliers than HS in both directions; and bias estimates yielded by RS were the largest and varied most compared to those obtained from the other MDTs. Figure 36 describes the distributions of bias estimates by missing data treatment methods in MNAR.

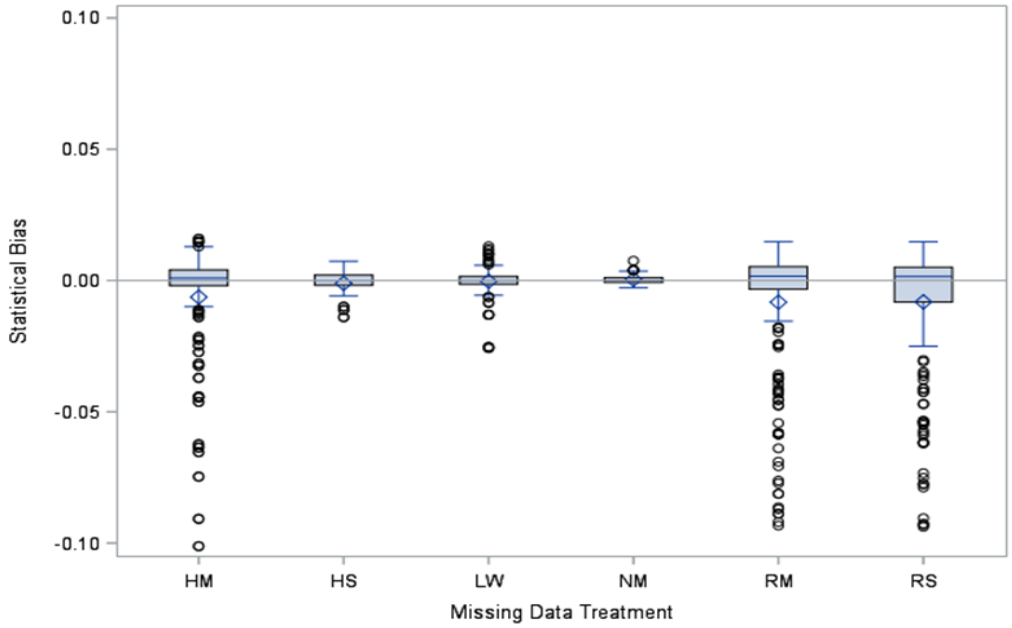

Figure 36. Distributions of Bias by Missing Data Treatment in MNAR

The major influencing factors on bias estimates by each MDTs in MNAR were also similar to those found in MAR data; for example, the top three influencing factors on bias estimates produced by RS, in order, were parameter, interaction effect between ICC and

parameter, and ICC; and the top three influencing factors on bias estimates produced by HM, in order, were parameter, interaction effect between percent miss and parameter, and interaction effect between ICC and parameter. However, the proportions of variance explained by each factor found in MNAR were different than what found in MAR; for example, in MNAR data, parameter accounted for 60%, 73%, and 76% of variation in bias estimates produced by HM, RM, and RS respectively while these numbers in MAR were 64%, 71%, and 80%.

In addition, the general pattern of the distribution of bias estimates by parameter for RS, RM, and HM in MNAR was also similar to what seen in MAR and MCAR, specifically for these three methods, bias associated with X7 and X8 was negative and larger in magnitude and variability than those associated with other regressors of which the bias estimates were positive. Figure 37 describes the distributions of bias by parameters for RS in MNAR.

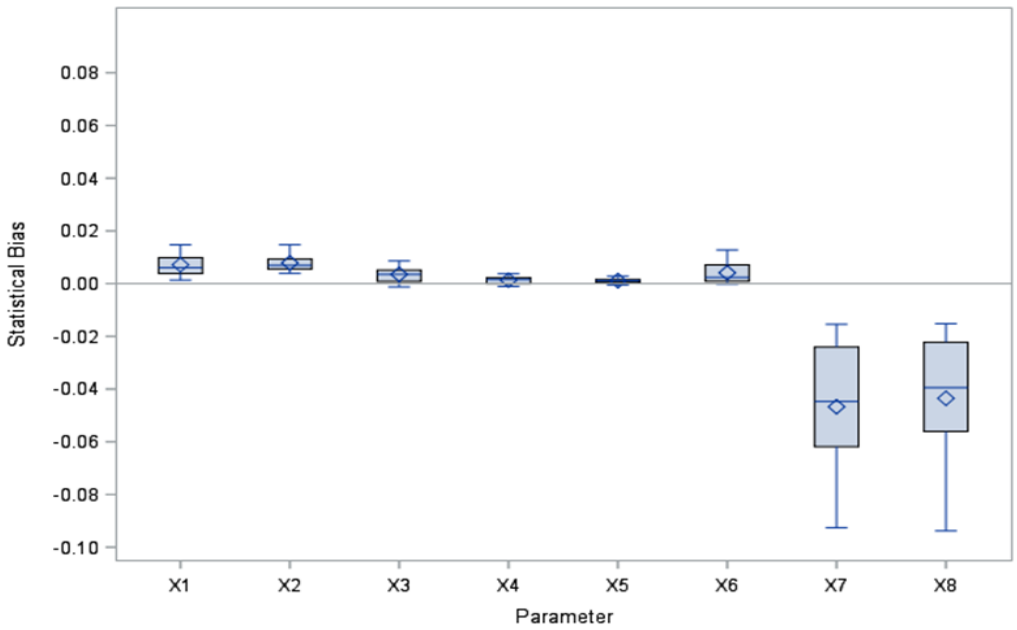

Figure 37. Distributions of Bias Estimates by Parameter for RS in MNAR

In MNAR data, though LW still produced relatively smaller bias estimates compared to other MDT methods, it produced some distance outliers which were not seen in MCAR. An eta squared analysis showed that the interaction effect between ICC and parameter was the most influencing factor on bias estimates produced by LW as previously seen in MCAR and MAR. Figure 38 describes the mean estimated bias by ICC and parameter for LW in MNAR: in zero-level ICC, there was no bias seen for non-missing-data regressors, and bias estimates for missing-data regressor were negative; but in non-zero-level ICC, there was no specific pattern of association between ICC level and bias estimates for each particular regressor (e.g., the impact of the influencing factor appears to be random).

The interaction effect between ICC and parameter also modestly influenced ($\eta 2$=17%) the bias estimates produced by RS in MNAR. The pattern of the influence of this interaction effect on the bias estimates in RS condition was more specific than what seen in LW condition. Figure 39 shows that for missing-data regressors (X7 and X8), bias estimates were negative and substantially larger than those associated with non-missing-data regressors in a way that smaller bias were seen in larger ICC level; but for non-missing-data regressors, bias estimates for different ICC levels did not show a lot of difference. Similar pattern of influence was also seen in bias estimated by RM (Figure40) and by HM (Figure41).

As found in MCAR and MAR, missing data level had very little impact on bias estimates by all five studied MDTs in MNAR (especially it had no impact on bias estimated by HS); however, the interaction effect between missing data level and parameter had a modest effect ($\eta 2$=11%) on bias estimates yielded by HM. The pattern of the influence of this interaction effect on bias estimates yielded by HM was clear, i.e., for all parameters, bias estimates in low missing

data level were smaller than those estimated in higher missing data level, and the bias estimates were negative and larger for X7 and X8 (Figure 42).

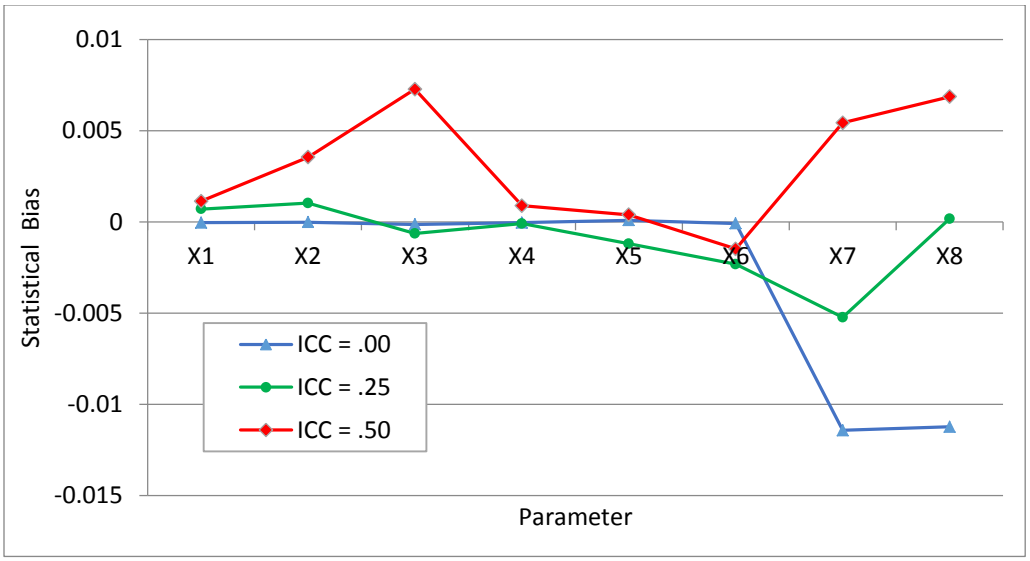

Figure 38. Mean Estimated Bias by ICC and Parameter for LW in MNAR

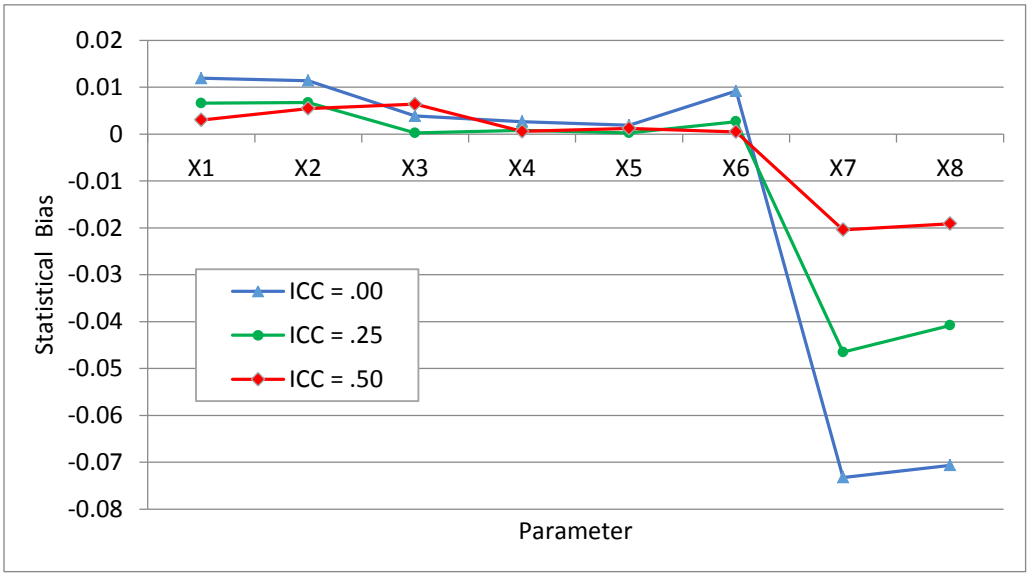

Figure 39. Mean Estimated Bias by ICC and Parameter for RS in MNAR

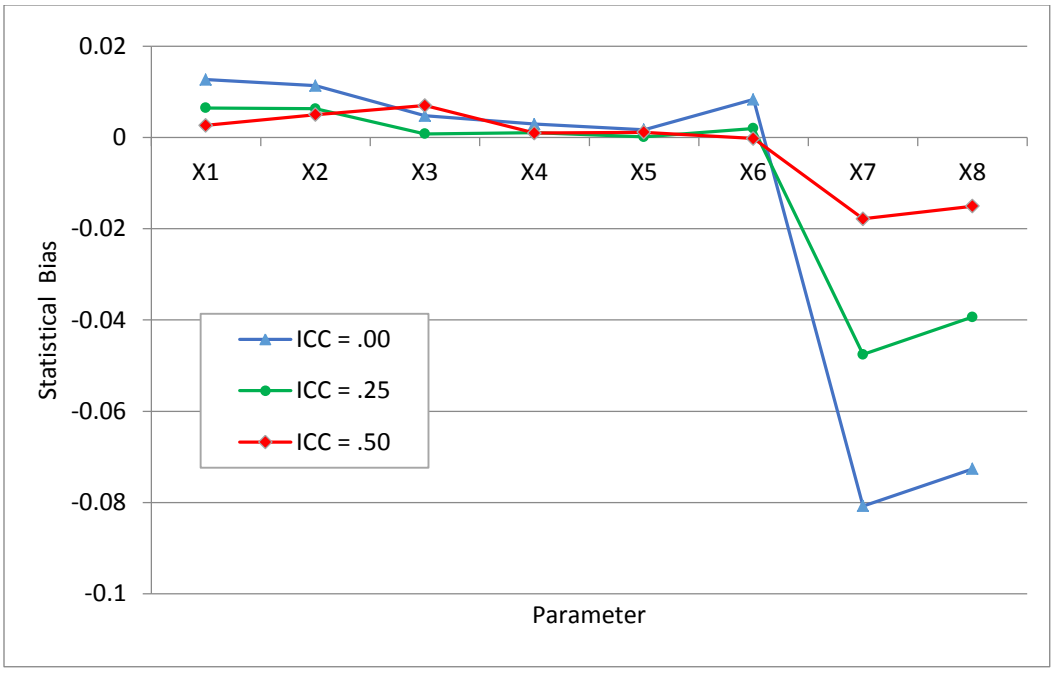

Figure 40. Mean Estimated Bias by ICC and Parameter for RM in MNAR

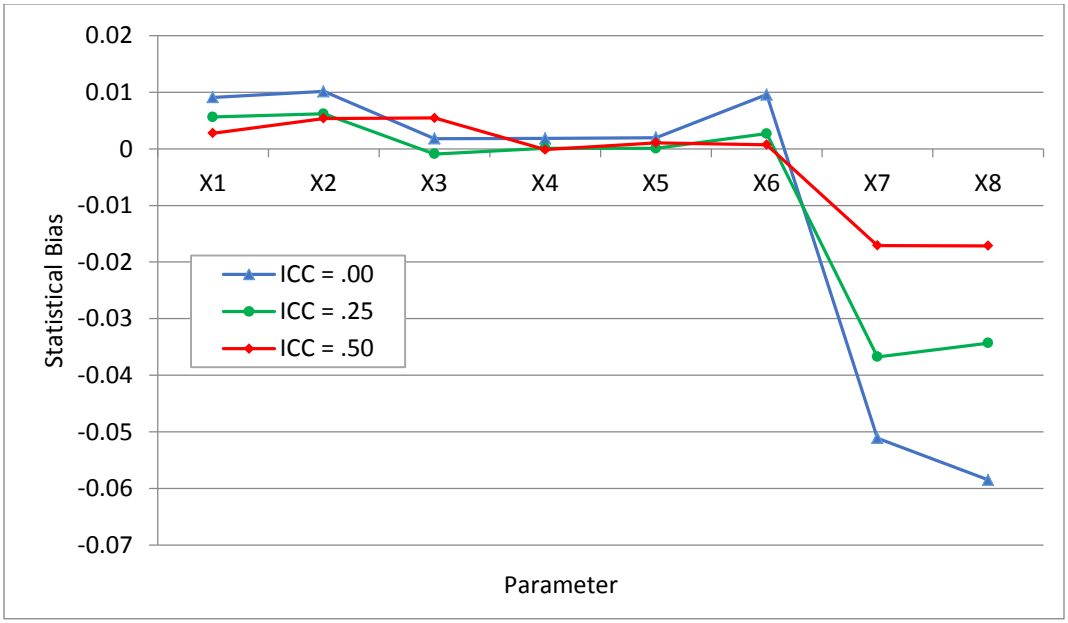

Figure 41. Mean Estimated Bias for HM by ICC and Parameter in MNAR

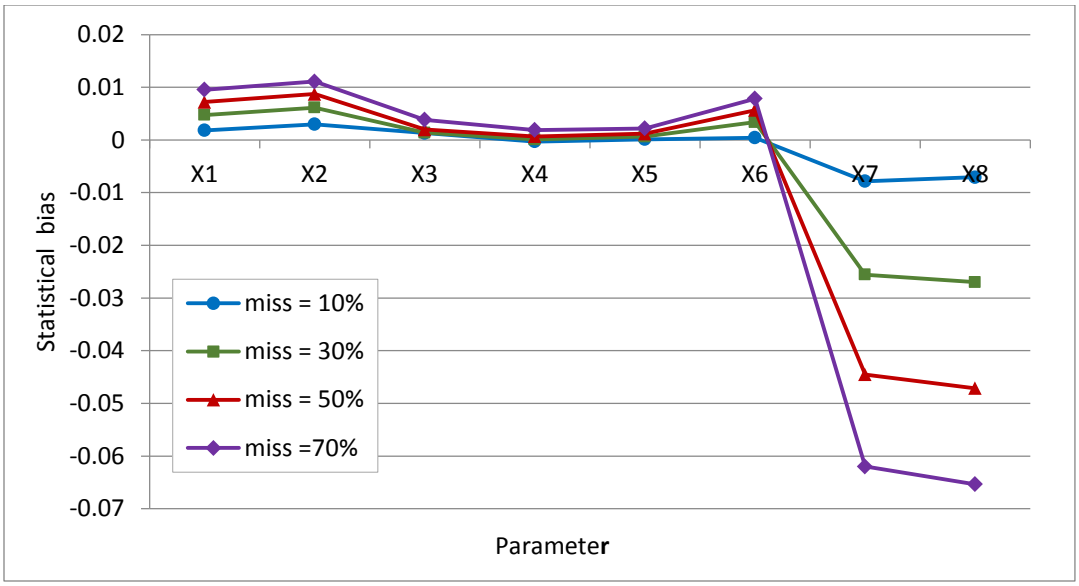

Figure 42. Mean Estimated Bias for HM by Percent Missing Data and Parameter in MNAR

RMSE

Figure 43 shows the distribution of RMSE estimates for the five studied MDTs along with the RMSE obtained from the complete samples (NM): RMSE estimates obtained from HS were more comparable to those obtained from the complete samples in terms of average magnitude and variation; and regarding these two aspects, RMSE estimates for LW were second to those for HS, whereas RMSE estimates yielded by RM and RS appear to be the largest and most varied.

Parameter was the most influencing factor on RMSE estimates produced by RS ($\eta 2$=63%) and by RM ($\eta 2$=68%) in MNAR. Figure 44, 45 describes the distribution of RMSE estimates by parameter for RS and RM in MNAR, respectively. In both graphs, RMSE estimates associated with X7 and X8 were larger compared to those associated with other regressors.

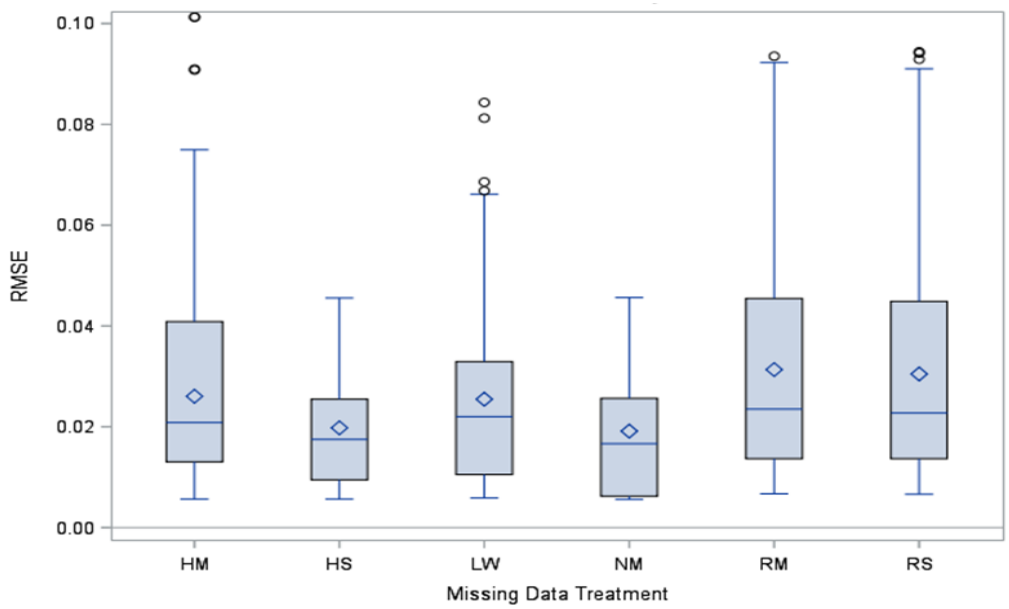

Figure 43. Distributions of RMSE estimates by Missing Data Treatment in MNAR

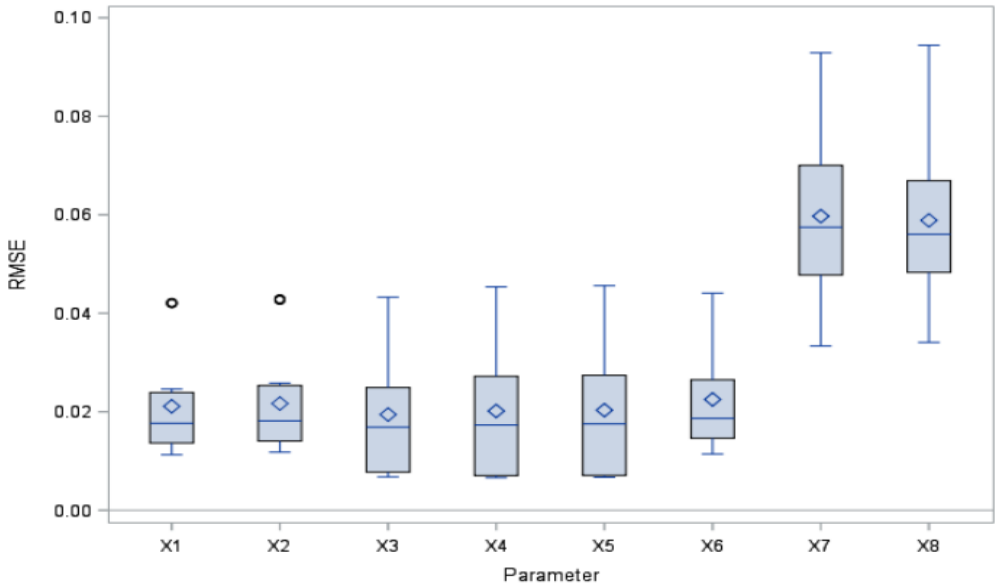

Figure 44. Distributions of RMSE Estimates by Parameter for RS in MNAR

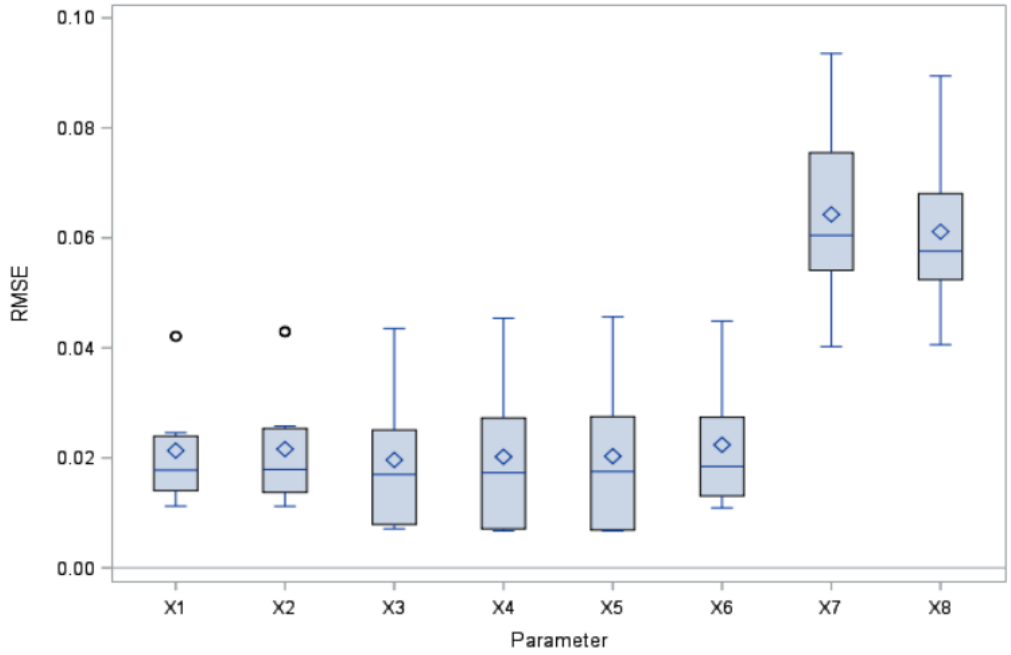

Figure 45. Distributions of RMSE Estimates by Parameter for RM in MNAR

CI Width

In MNAR data, CI width estimated by LW varied most compared to those estimated by other MDTs, and it was also the largest; whereas CI width estimated by HM, HS, and RS appeared comparable to those obtained from the complete sample (NM) (Figure 46). As found in MCAR, ICC was the top influencing factor on CI width produced by all five MDTs in MNAR: HM ($\eta 2$=65%), HS ($\eta 2$=66%), LW ($\eta 2$=62%), RM ($\eta 2$=46%), RS ($\eta 2$=63%).

For LW (which produced the largest and most varied CI width in MNAR), population density and the interaction effect between ICC and population density also had a modest effect (18% and 11% respectively) on the estimated CI width. The pattern of this influence was similar to what seen in MCAR and MAR data, i.e., the CI width estimates were larger and varied more as ICC levels increased (Figure 47); and also in zero ICC data, there was no difference in CI

width estimates between low and high population density levels, but in non-zero ICC data, CI width estimate for high density population was larger than CI width estimate for low density population (Figure 48); it is also noted that the mean CI width estimates in MAR was larger than in MNAR (see Figure 48 vs. Figure 33).

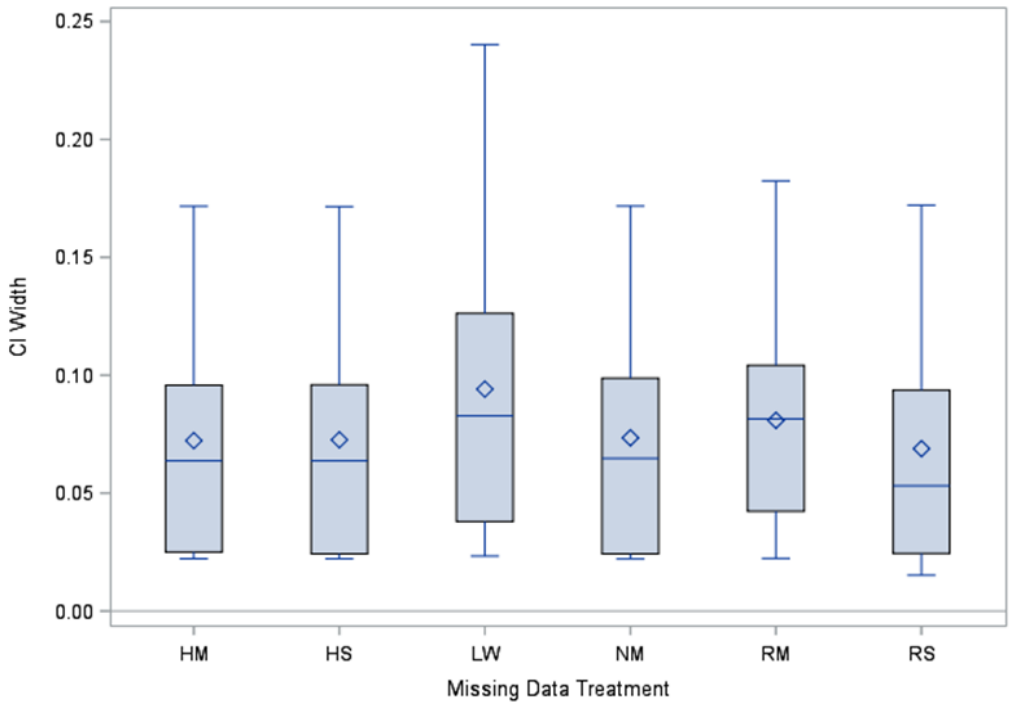

Figure 46. Distributions of CI Width Estimate by Missing Data Treatment in MNAR

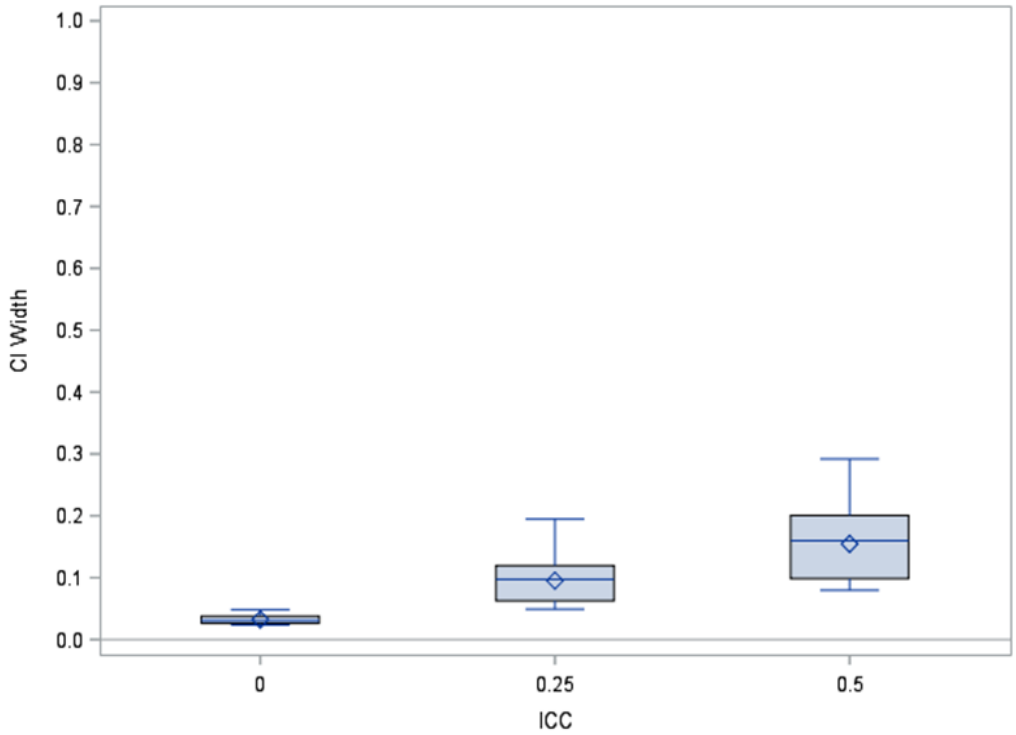

Figure 47. Distributions of Confidence Interval Width by ICC for LW in MNAR

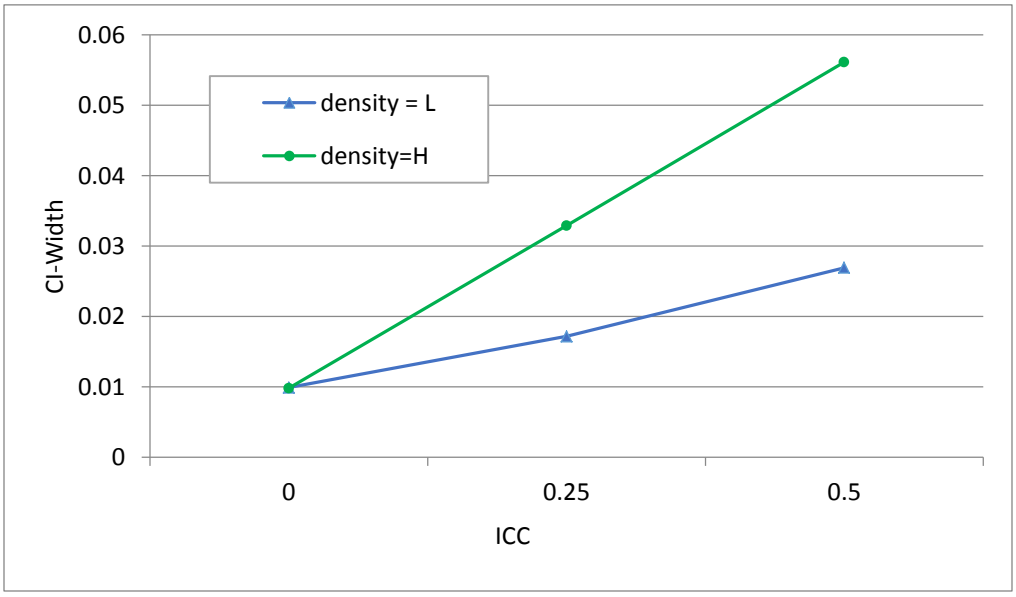

Figure 48. Mean Estimated CI Width by Population Density and ICC for LW in MNAR

CI Coverage

As found in MCAR and MAR data, RS produced the lowest and most varied coverage estimates, and LW and HS produced comparable coverage estimates which were closer to that obtained from the complete data. Of the two multiple imputation methods RM and HM, estimates associated with RM were generally lower and varied more (Figure 49).

Concerning the top factors influencing CI coverage produced by the five MDTs, parameter was the most influencing factor on the estimates produced by HM ($\eta 2$=36%), RM ($\eta 2$=61%), and RS ($\eta 2$=63%); and the interaction effect between ICC and parameter was the most influencing factor on the estimates produced by HS ($\eta 2$=50%) and LW ($\eta 2$=26%).

Also, in examining major factors influencing the lowest coverage estimates produced in MNAR, besides parameter factor which had a strong effect, ICC and the interaction effect between parameter and ICC had a modest effect (20% and 12%, respectively) on this estimate (produced by RS). The pattern of the influence of the interaction effect of parameter and ICC on the coverage estimated by RS was similar to what previously seen in MCAR and MNAR data, and it is described in Figure 50: In zero ICC data, coverage estimates associated with all parameter were lower than those in non-zero ICC data, and coverage estimates associated with X7 and X8 were the lowest. In non-zero ICC data, coverages estimates associated with non-missing regressors appeared comparably high regardless of ICC levels; but coverage associated with missing-data regressors were low in low ICC level and high in high ICC level.

In general, of the design factors, parameter, ICC, and the interaction effect between ICC and parameter were the top most influencing factors on the performance measures yielded by the studied MDTs for all three types of missingness. For example, in all three missing data types,

parameter was the top influencing factor on bias produced by all studied MDTs except LW, of which the bias was mostly influenced by the interaction effect between ICC and Parameter; also ICC was the top influencing factor on CI width estimated by all the studied MDTs. It is also noted that in different forms of effect, ICC was the top factor influencing all performance measures produced by LW, while parameter was the most influencing factor on most measures produced by HM, RM, and RS (see Table 4).

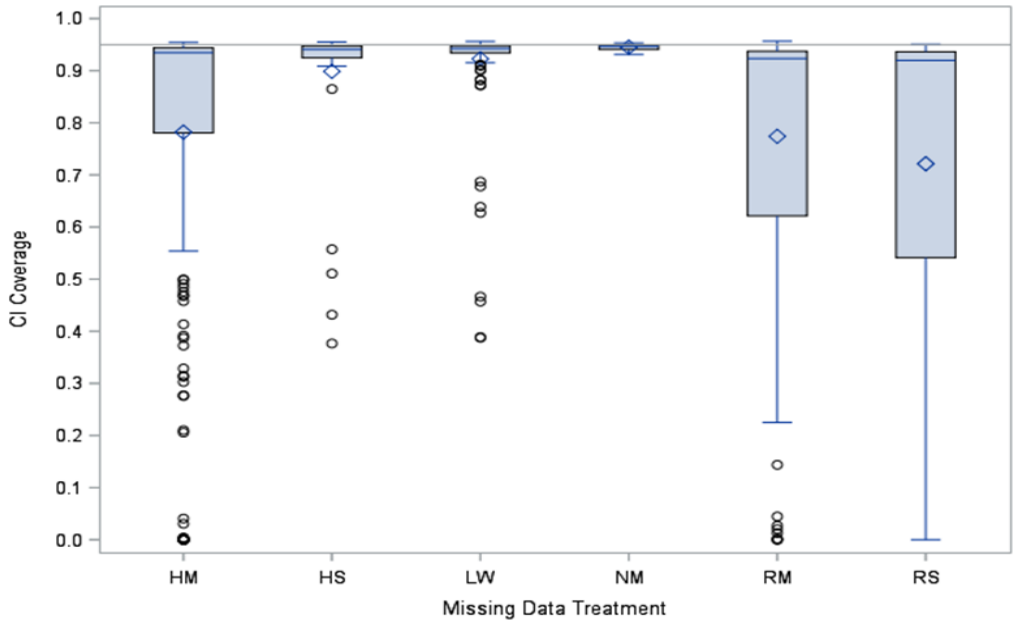

Figure 49. Distributions of Coverage Estimate by Missing Data Treatment in MNAR

In summary, it was observed that (1) in all three types of missingness (MCAR, MAR, and MNAR), the pattern of what simulation factor had a major influence on a performance measure (bias, RMSE, CI width, or CI coverage) produced by a certain MDT was consistent, but the degrees of influence of the factor on the measure could be different or the same among the three data types. For example, parameter was the most influencing factor on bias estimates produced by HM in all three data types, but it accounted for 67% of variation in bias in MCAR, 64% in

MAR, and only 59% in MNAR; ICC was the most influencing factor on CI width produced by LW in all three data types, and it accounted for 65% of variation in CI width estimates in MCAR and approximately 62% of variation in CI width estimates in MAR and MNAR. (2) In the three types of missingness, the top-most-influencing factor on the four evaluation measures was either parameter, ICC, or interaction effect between ICC and parameter where parameter was always the most-influencing factor on measures estimated by a multiple imputation method (i.e., HM and RM); and either ICC or the interaction effect of ICC and parameter was the most-influencing factor on measures estimated by LW. (3) Level of missing data, by itself, had no or very weak effects on those performance measures yielded by the five studied MDTs, particularly it had some effects on HM but no effect on HS. (4) Population density factor had negligible effects on most of the measures produced by all studied MDTs except for RMSE, CI width, and CI coverage produced by LW which were modestly influenced by population density.

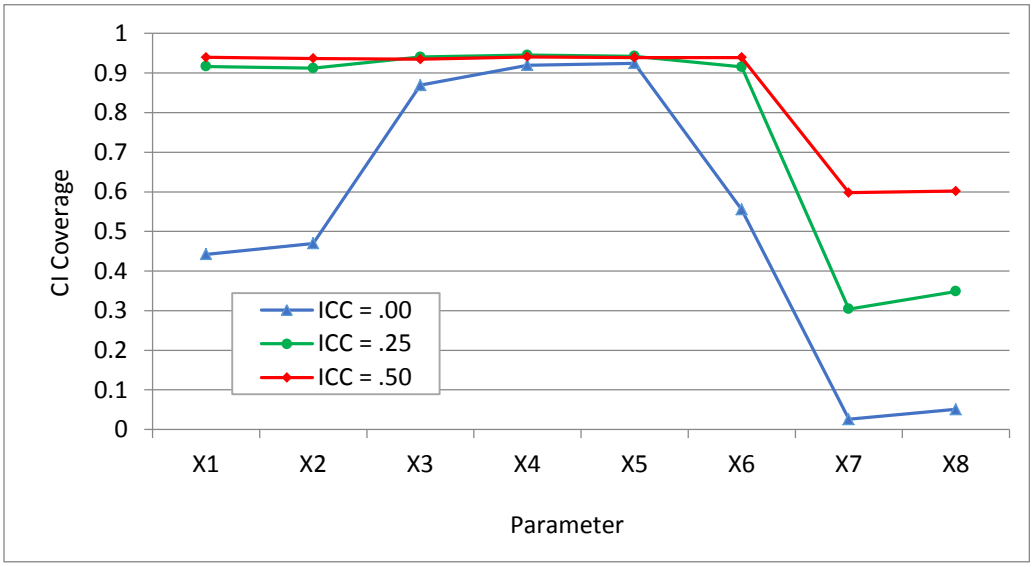

Figure 50. Mean Estimated Coverage by Parameter and ICC for RS in MNAR

Table 5

The Most Influencing Design Factor on Performance Measures in MCAR, MAR, and MNAR

MDTs / Measures		HM	HS	LW	RM	RS
MCAR	Bias	parameter	parameter	ICC* parameter	parameter	parameter
	RMSE	parameter	ICC	ICC	parameter	parameter
	CI width	ICC	ICC	ICC	ICC	ICC
	Coverage	parameter	ICC*parameter	ICC*parameter	parameter	parameter
MAR	Bias	parameter	parameter	ICC* parameter	parameter	parameter
	RMSE	parameter	ICC	ICC	parameter	parameter
	CI width	ICC	ICC	ICC	ICC	ICC
	Coverage	parameter	ICC*parameter	ICC* parameter	parameter	parameter
MNAR	Bias	parameter	parameter	ICC* parameter	parameter	parameter
	RMSE	parameter	ICC	ICC	parameter	parameter
	CI width	ICC	ICC	ICC	ICC	ICC
	Coverage	parameter	ICC*parameter	ICC* parameter	parameter	parameter

CHAPTER FIVE: DISCUSSION

The present study was designed to evaluate the performance of five different MDTs (listwise deletion-LW, single hot-deck imputation-HS, single regression imputation-RS, hot-deck-based multiple imputation-HM, and regression-based multiple imputation-RM) for handling missing data in MCAR, MAR, and MNAR data. These MDTs were evaluated in the context of parameter estimates in multiple regression analysis in complex sample data with two data levels. Specifically, the evaluations were based on evaluating parameter point and interval estimates yielded by each of the studied MDTs; and the four performance indicators used in this study were statistical bias, RMSE, CI width, and coverage rate (i.e., 95%) of the confidence interval. The design factors directly controlled in this study were levels of missing data, ICC, population density (high, low) and types of missingness (MCAR, MAR, and MNAR).

This chapter summarizes the major findings about the performance of the MDTs and discusses the potential implications of the findings for practice and suggestions for future research. Generally, this chapter contains four sections: (1) Summary of the relative performance of the five studied MDTs and discussion of the impact of the design factors on their performance; (2) discussion of reflection and contradiction of the present findings with results from the previous studies reviewed in Chapter 2; (3) recommendations for practice; and (4) limitations of the present study and suggestions for future studies.

Result Summaries: Relative Performance and Influencing Factors

Essentially, this section provides the answers for the research questions, and the discussion will be grouped in three subsections in accordance with the three types of missingness: MCAR, MAR, and MNAR. In each subsection, the answer for the first question, which MDT is more accurate and precise, and the second question, how the controlled factors affect the procedures' performance, will be discussed in order. In general, the accuracy of the studied MDTs will be assessed based on the combination of bias, CI coverage probability, and RMSE; and the precision will be assessed based on the combination of RMSE and CI width produced by each of the MDTs.

MCAR

Across conditions in MCAR, RS procedure performed poorest considering accuracy (e.g., poorest bias, RMSE, and coverage of confidence interval); RM produced the most imprecise estimates (e.g., poorest RMSE and CI width); HM's performance was not the poorest nor the best considering the combination of all performance measures; and LW as well as HS procedure outperformed the rest on all performance measures of accuracy and precision (with one exception that CI width yielded by LW was not the narrowest). Between the two top performers LW did slightly better than HS regarding bias and coverage of estimates whereas HS did slightly better than LW regarding RMSE and CI width.

Regarding the influence of the design factors on the performance of MDTs, the factor of missing-data variables vs. non-missing-data variables had the most impact followed by ICC level; and the influences of these factors and their interaction effects were evident on those

relatively-poor performers (HM, RM, and RS). Specifically, the slopes of variables without missing data were slightly overestimated while the slopes of variables with missing data were largely underestimated. As for the influences of the ICC factor, on variables with missing data, RS and RM were less biased in higher ICC data; but on variables without missing data, there was no specific pattern of the influence of ICC. Besides these two major influencing factors, the proportion of missing data greatly impacted the accuracy of HM on variables with missing data such that the bias of the estimated slopes increased as the proportion of missing data increased.

Examining the influence of the design factors on the precision of estimates, it is observed that RM, which yielded the least precise estimates, produced larger RMSE and CI width for variables with missing data than for variables without missing data, especially in low ICC data; while LW's CI width, which was the second largest, was mainly influenced by ICC levels such that the higher the ICC, the wider the CI width estimates. Of the three relatively poor performers, RS yielded the poorest coverage probability followed by HM and RM in order; and the estimated coverage measures were lower for variables with missing data and higher for variables without missing data. In addition, particularly for HM the higher the proportion of missing data the lower the coverage probability; and for RS the lower the ICC level the lower the coverage probability.

In summary, the three relatively-poor-performance MDTs (RS, RM, and HM) were less accurate and less precise in estimating the slopes of variables with missing data than of variables without missing data; and the slopes of variables without missing data were slightly overestimated whereas the slopes of variables with missing data were largely underestimated. The inaccuracy and imprecision also depended on data ICC level; and for HM, the performance measures also depended on the proportion of missing data (e.g., higher missing data level, less accurate and lower coverage).

MAR

The relative performance of these MDTs in MAR data was mostly similar to that found in MCAR data with some variations. Specifically RS produced the most inaccurate estimates (e.g., its bias, RMSE, and CI coverage were the poorest - though its CI width was one of the smallest); RM and HM were the second poorest performers (RM and HM yielded equally poor bias, but just like RS, HM produced one of the narrowest CI width estimates); and HS and LW remained the top two performers. LW slightly trailed HS in RMSE, plus LW yielded the largest CI width estimate. The differences in performance of LW in MAR compared to its performance in MCAR were that it produced the largest CI width, and it yielded slightly more bias outliers in MAR than in MCAR while HS bias estimates appeared consistent between two data types.

Regarding the influence of the design factors on the relatively poor performance MDTs (RS, RM, and HM) in MAR data, these poor performers were less accurate and less precise on variables with missing data than they were on variables without missing data; plus, they largely underestimated the parameters of variables with missing data while slightly overestimated the parameters of variables without missing data.

Besides the design factor of missing variables vs. non-missing variables, ICC was another factor strongly influencing the performance of these three MDTs in MAR: in higher ICC data, HM, RM, and RS performed better on variables with missing data than they did in lower ICC data (e.g., they produced smaller bias/RMSE and larger coverage for variables with missing data in higher ICC data than in lower ICC data). ICC levels also affected the poor CI width estimates produced by LW; the higher the ICC level, the larger the CI width estimates produced.

In addition, as seen in MCAR data, the proportion of missing data only noticeably affected HM's accuracy and precision such that the higher the missing data proportion the less

accurate and precise the estimates (e.g., larger bias, larger RMSE, and lower coverage estimate obtained from higher missing data level data).

MNAR

Despite the performance of each individual MDT varied between MAR and MNAR data, their relative performance in MNAR data was similar to that in MAR data. In MNAR data, RS remained the poorest performer (it still produced one of the smallest CI width estimates as it did in MCAR and MAR data); HM and RM also remained the second poorest performers; and HS as well as LW were the top performers with HS' RMSE and CI width being smaller than that of LW's, but LW's coverage being higher than HS'.

As for the influence of the design factors on those poor performers in MNAR data, missing data variables vs. non-missing data variables was still the most influencing factor and the impact of this factor depended on other factors like ICC level and proportion of missing data as described above in the MAR data section. Particularly, those poor performers were less accurate on variables with missing data than they were on variables without missing data; furthermore, they largely underestimated the parameters of variables with missing data while slightly overestimated the parameters of variables without missing data. They also produced more precise estimates on variables without missing data than on variables with missing data. Their performance measures on variables with missing data vs. variables without missing data depended on the data ICC such that for variables with missing data, smaller bias/RMSE and higher coverage estimates were obtained from higher ICC data samples; whereas there was no specific pattern of performance found across levels of ICC in variables without missing data.

In addition, as seen in MCAR and MAR, the proportion of missing data only noticeably affected the performance of HM such that in higher missing data levels, HM yielded worse performance measures (e.g., higher bias/RMSE and lower coverage); however, unlike in MAR and MCAR data, in MNAR, the impact of proportion of missing data on the coverage yielded by HM was minimal.

Reflection and Contradiction of Present and Previous Findings

Those findings discussed above reflect what have been found about the relative performance of MDTs from previous studies, and they also identify some differences from what was previously found. Published studies about the relative performance of MDTs found in the literature were mostly done with MCAR data; some were done with MAR data; and very few studies found were done with MNAR data. This section will separately discuss the present findings that reflect and that differ from previous findings for each types of missingness. In addition, since each study employed different performance measures to compare its studied MDTs, the comparative analysis below will be based on the common measures between the present and previous studies.

MCAR

In general, most of the present results with MCAR support previous findings; yet, there also are some contradictions. The superior performance of LW regarding regression weight bias over the RS in MCAR data found in the present study supports the results found in Roth and Switzer's 1995 study. However, the superior performance of HS over RS found in this present study is not consistent to Roth and Switzer's findings, in which HS produced higher regression

bias and more dispersion in regression weight estimates compared to RS, whereas the present findings show that HS produced smaller bias and RMSE than RS did. There are major differences in study designs between the two studies; and these design differences could possibly contribute to the difference in findings. Particularly in the former study, imputation cells for hot deck imputation were categorized by the range (e.g., low, medium, high) of auxiliary variables in the sample of the single level data leading to a much more finely categorized group which could afford small percentages (10-30%) of missing data while in the current study, imputation cells were based on a stratum which was a level 2 variable leading to a much larger pool of donors to afford high percentage (10-70%) of missing data; moreover, the relatively weak covariates in the population correlation matrix used in the present study might not benefit the performance of RS.

No similar research into MDTs at Level 1 of multilevel data has been found in literature, but Gibson and Olejnik (2003) investigated the performance of LW, RM, and other MDTs on missing data at the second level of a two-level hierarchical data structure with MCAR data. In their study, the authors found RM consistently underestimated regression weights for Level 2 while LW produced non-significant bias. The present findings support the former findings regarding the relative performance the two procedures, but in the present study RM was also seen to slightly overestimate the parameters for variables without missing data.

The relative small bias in the regression weights estimated by LW found in this present study is also consistent with the findings in Kuijk and colleagues' 2016 study in which the controlled missing data level was similar to that of the present study but with smaller sample size and higher correlation between independent variables; however, Kuijk and colleagues also found that RM produced unbiased regression weights, which contradicts to the present study's finding. Yet, the RM procedure used in Kuijk and colleagues' study employed the predictive mean

matching mean version of the regression imputation method (i.e., the nonrespondent was matched to the respondent with the closest predicted value, and then the respondent's observed value - not the predicted value - was assigned to the matched nonrespondent).

The relative performance of LW vs. RM found in MCAR data from this present study is also consistent with that found in a similar study by Kellermann and colleagues (2016), in which LW consistently surpassed MI by delivering virtually no bias, less dispersion in regression weight estimates, narrower CI width, and higher coverage rate, while RM point estimates were considerably biased in both directions. The only difference in design between the two studies was the controlled ICC levels: the ICC levels in the former study were much smaller than that in the present study which showed more effect of sample variance on bias and dispersion measures yielded by the MDTs.

MAR

The superior performance of LW over RM found in the present study by producing less bias, less dispersion, and higher coverage in MAR data is also consistent with what was found in Kellermann and colleagues' 2016 study. However, it is not consistent with what was found for MAR data in Kuijk and colleagues' 2016 study, from which RM produced unbiased while LW yielded underestimated regression weights. In addition, for MAR data, the present findings also completely contradict to the results found in Cranmer and Gill's 2012 study in which HM outperformed LW and RM in all studied conditions, and RM often produced less biased regression weights than LW. Considering the differences in study design and donor selection method, at least some of the contradicted results could be explained. Regarding the superior performance by HM in the former study, the donors were selected based on the random overall

imputation, e.g., for each nonrespondent, a respondent was chosen at random from the total respondent sample; whereas in the present study, for each nonrespondent, a respondent was chosen at random from the imputation cell (e.g., in the same stratum), hence in the latter study, the numbers of donors were more limited; and when the missing level increased, the number of recipient units was greater than the number of donor units in the imputation cell level. Regarding the finding that LW produced bias estimates more often than RM in the former study, the sample size in the former study was only a small portion (2.5%) of the sample size in the latter study, hence removing cases with missing data in the small samples severely affect the performance of LW especially when missing data level increased. In addition, the consistently high correlation between variables in the former study (vs. the mostly low-to-modest covariates in the present study) might advantage the performance of regression based imputation in the former study.

MNAR

There are not many published studies about the performance of MDTs in MNAR data; Kromrey and Hines' 1994 study is one of the few studies found that investigated the relative performance of LW, RS, RM and some other MDTs in MNAR data. From their study, the authors found that the LW procedure appeared to provide the best performance overall in the estimation of the regression weights. Considering the relative performance of LW vs. RS and RM, the present findings support the former findings; however, in the present study RM bias was seen in both directions while in the former study RM consistently overestimated the regression weight. This general difference in RM bias direction between the two studies could be due to the difference in design factors especially the control of different ICC levels which was applied in

the present study only (in the present study, ICC was seen to strongly influence the negative bias of parameter estimates for variables with missing data).

LW surpassing RM in MNAR data found in the present study also supports the findings in Kuijk and colleagues' 2016 study; but in the former study, RM overestimated the regression coefficient of variables with missing data and produced no bias on the regression coefficient of variables without missing data whereas in the present study RM underestimated the regression coefficient of variables with missing data and overestimated the regression coefficient of variables without missing data. A major difference in study design such as use of a single predictor variable with and without missing data in the former study vs. multiple predictor variables with and without missing data plus the control of ICC level in the present study could be the reasons for the difference in bias direction between the two studies. (In the present study, RM was seen to produce no bias on the two moderately-correlated predictor variables that have no missing data; and lower ICC levels exacerbated the negative bias of parameter estimates for variables with missing data).

On a different aspect, the present study also extended Kuijk and colleagues' 2016 study by using a different approach to regression based multiple imputation. Kuijk and colleagues used the mean matching regression while the present study used the traditional regression in doing RM. The results from these studies suggest that LW surpassed each of the RM approaches studied in MNAR data.

Implications and Recommendations

The results suggest at least three implications. First, depending on the type of missingness and the researchers' goal, researchers may wish to use certain MDTs and avoid

others. For a complex sample survey with large sample size and with data ICC similar to the data in this study, researchers should consider using LW when they are concerned about bias. LW is accurate (though LW bias slightly changed from MCAR to MAR and MNAR, it is still one of the most accurate compared to others in these studied conditions) and is the easiest to use in a computational sense. If researchers are concerned more about precision of estimates, LW could be still the technique of choice in MCAR, but in MAR and MNAR, researchers should note that LW precision decreases especially in high ICC and high population density data; so if the data are not MCAR plus the ICC level is high and the precision of estimate is more important, HS could be the method of choice. HS is accurate and precise with a high coverage rate, but HS requires extra effort regarding computation, whereas LW is automatic in most standard analysis procedures. When data have a low ICC level and the correlations between predictors are low as in this study, researchers should generally avoid RS.

Second, the most advanced MDT is not always the best, but rather the effectiveness of an MDT depends on the features of the data and the researcher's goal in using the data; and thus the key to choosing an appropriate MDT procedure is knowing the data. Specifically researchers should investigate the data for the correlations between variables, proportion of missing data, data ICC level, and data structure (e.g., multilevel vs. single level); if these data features are similar to that in this study, and if multiple imputation methods should be considered, researchers should note that RM produced larger and more varied RMSE than HM in all three data types, and the CI width yielded by RM was also slightly wider than that yielded by HM; plus, in MNAR, coverage probability for RM varied more than that for HM. However, HM was found to be most susceptible to the amount of missing data.

Finally, and also equally important is studying the user documentation for the software being used for imputation to be aware of options and default settings and alter the options and settings as appropriate for the context of the study. Imputation default settings such as the imputation approach and the number of imputations for multiple imputation procedures might be different in different statistical software packages and they might be even different in different versions of the same software package. With respect to computing resources, multiple imputation could utilize large amounts of memory and storage depending on the number of imputations, the percent of missing data, and number of variables and records; hence if researchers need to impute a multiple-variable, large dataset with high percent of missing data but have limited computing resources, it can be helpful to investigate the data to determine the smallest number of imputations needed.

Limitations and Recommendation for Future Studies

There are some limitations in this study that may affect the generalization of the study findings. First of all, multicollinearity is one concern in multivariate analysis. Multicollinearity occurs when one or more of the predictor variables highly correlates with the other predictor variables in a regression equation, and one of the consequences of multicollinearity is less precise estimates (e.g., wider confidence interval width). In ordinary least squares (OLS), the effect of multicollinearity is well studied in literature; and the variance inflation factor (VIF) is often used to diagnose multicollinearity. VIF shows how much the variance of the coefficient estimate is, compared to what it would be if the variable were uncorrelated with any other variable in the model; and the square root of the VIF shows how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other variables in

the equation. Cohen and colleagues (2003) suggested that parameter estimates and standard errors may noticeably change at VIF values around 5.

The effect of multicollinearity in multilevel models is not widely discussed in literature; however, there are some studies that examined and evidenced the general influence of multicollinearity among level-1 predictors on multilevel model parameter estimates and their standard errors (Kubitschek and Hallinan,1999; Shieh and Fouladi, 2003). Shieh and Fouladi's study suggested that bias can be introduced at a level of correlation around .5. The data in this present study were simulated based on a correlation matrix with mostly weak correlations (the average correlation of variates was approximately .17) and a quite low level of VIF as seen in Table 5, so the analysis results may not clearly reflect the effects of MDTs in contexts where there is multicollinearity and thus the conclusions from this dissertation may not generalize to real-world contexts that have data with a high level of multicollinearity.

Table 6

VIF Associated with Each Regressor

Regressors	X1	X2	X3	X4	X5	X6	X7	X8
VIF	1.02222	1.05543	1.07449	1.21588	1.27095	1.24015	1.09328	1.21718

In addition, in real practices, missing data mechanisms may not exist exclusively, but more than one mechanism could be present in a data sample. Therefore the findings about the performance of the MDTs in this study may not represent their performance on samples with mixed missing data mechanisms. Another limitation of this study is the simplicity of the missing data; because of the complexity of the simulation study, only two variables present missing data

and they are always missing together. In field research, it is likely that data will be missing for more than two variables and the missing variables may not be missing together.

Furthermore, the samples used in this study are not comparable to those obtained from a simple random sample survey with a small sample size; variables were limited to continuous data; and statistical analysis was limited to estimating regression weights in multiple regression analyses and focuses only on the point estimates of regression coefficients, whereas researchers may be interested in means, correlations, or variances. Also, the results of this present study are limited to two-level data structures as two-level data are typically used in educational research; however, in real practices, there are various theoretical and practical reasons for combining more than two levels of data.

Future studies can be considered to explore the impact of different levels of correlation between variates to investigate the impact of multicollinearity on the effectiveness of MDTs (particularly on the effectiveness of hot-deck based imputation and regression based imputation) on multilevel data. Computationally, the levels of correlation between variates influence multiple regression hence regression imputation; and theoretically, the levels of correlation between variates also influences the effects of hot-deck imputation; it would be interesting to see how susceptible to correlation between covariates the regression imputation is compared to the hot-deck imputation.

In practice, the types of missingness whether MCAR, MAR, MNAR, or mixed missing data mechanisms can be different for different fields. Research can be conducted to study the sources of missing data that are likely to occur in the field's real data to identify the pattern of mixed missing data mechanisms in the data and study the performance of MDTs under the identified mixed missing data mechanisms. Other research could also be conducted to examine

performance of MDTs for a MAR condition where the probability of missing data on a regressor depends on the dependent variable instead of on an independent variable (studies with MAR data conditioning on dependent and independent variables have been seen in flat data but have not been seen in multilevel data).

Furthermore, since the performance of the traditional regression imputation and its multiple imputation version were poor compared to other MDTs in the present study, it would be interesting to replicate this study to investigating the performance of the predictive mean matching imputation which is a variant of the regression imputation and was proposed to be used in MI by Little (1998). According to Allison (2015), predictive-mean-matching-based MI is an attractive way to do multiple imputation for missing data.

In addition, different software package may have different level of reliability, and studies of relative reliability of different statistical software packages have been seen in literature of numerical analysis software. The analysis in this study was conducted using SAS/STAT®, the standard for data analysis; and the analysis results in this study have not been double checked using other software package. Interested researchers could duplicate the study using a different statistical software package.

REFERENCES

Allison, P. (2002). *Missing Data. .* Thousand Oaks, CA: Sage.

Allison, P. (March 5, 2015). Imputation by predictive mean matching: Promise & peril. Retrieved from http://statisticalhorizons.com/predictive-mean-matching

Andridge, R. R. and Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review* 78, 40–64.

Bartlett, J. W., Harel, O., & Carpenter, J. R. (2015). Asymptotically unbiased estimation of exposure odds Ratios in complete records logistic regression. *American Journal of Epidemiology, 182*(8), 730–736. http://doi.org/10.1093/aje/kwv114

Berglund, P., & Heeringa, S. G. (2014). Multiple imputation of missing data using SAS. SAS Institute.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research, 5*(3), 215-238.

Buck, S.F. (1960). A method of estimation of missing values of multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society Series* , 22 (3), 302-306.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences,* 3rd edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Carlin, J. B., & Hocking, J. (1999). Design of cross-sectional surveys using cluster sampling: an overview with Australian case studies. *Australian and New Zealand Journal of Public Health*, 23(5), 546-551

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 721–726.

Cranmer, S. J., & Gill, J. (2013). We have to be discrete about this: A Non-parametric Imputation Technique for Missing Categorical Data. *British Journal of Political Science*, 43(02), 425-449.

Dargatz, D. A., & Hill, G. W. (1996). Analysis of survey data. *Preventive Veterinary Medicine, 28*(4), 225-237.

David, M., Roderick J. A. Little, Samuhel, M., & Triest, R. (1986). Alternative methods for CPS income imputation. *Journal of the American Statistical Association, 81*(393), 29-41. doi:10.2307/2287965

Frane, J. W. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika, 41*(3), 409-415.

Gibson, N. M. & Olejnik, S. (2003). Treatment of missing data at the second level of hierarchical Linear models. *Educational and Psychological Measurement*, 63, 204-238.

Graubard, B. I., & Korn, E. L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology*, 144, 102-106.

Green, J. L., Camilli, G., & Elmore, P. B. (2012). *Handbook of Complementary Methods in Education Research*. Routledge.

Harrell Jr FE. (2001). *Regression Modeling Strategies*. New York: Springer.

Kalton, G. & Kasprzyk, D. (1986), The treatment of missing survey data, *Survey Methodology*, 12, 1-16.

Kellermann, P. A.,Travathan, D., & Kromrey, J. (2016). *Missing Data and Complex Sample Surveys Using SAS®: The Impact of Listwise Deletion vs. Multiple Imputation on Point and Interval Estimates when Data are MCAR and MAR*. SouthEast SAS Users Group 2016 Proceedings. Cary, NC: SAS Institute Inc.

King, G., Honaker, J., Joseph, A., & Scheve, K. (1998, July). List-wise deletion is evil: what to do about missing data in political science. In *Annual Meeting of the American Political Science Association, Boston*.

King, G., Honaker, J., Joseph, A., & Scheve, K. (2001, March). Analyzing incomplete political Science data: An alternative algorithm for multiple imputation. In *American Political Science Association* (Vol. 95, No. 01, pp. 49-69). Cambridge University Press.

Kim, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika*, *89*(2), 470-477.

Kromrey, J. D. & Hines, C. V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Education and Psychological Measurement, 54*, 573-593.

Kubitschek, W. N. & Hallinan, M. T. (1999). Collinearity, bias, and effect size: Modeling "the" effect of track on achievement. *Social Science Research, 28*, 380-402.

Little, R. J., & Rubin, D. B. (2002). *Statistical Analysis with Missing data* (2nd ed.). New York, NY: John Wiley & Sons.

Lohr, S. (2009). *Sampling: Design and Analysis*. Nelson Education.

Mander, A., & Clayton, D. (2003). Weighted hotdeck imputation. *Statistical Software Components, Boston College Department of Economics, http://fmwww. bc.edu/repec/bocode/w/whotdeck. pdf*.

McKnight, P. E. (2007). *Missing Data: a Gentle Introduction*. New York : Guilford Press, c2007.

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A., & Verbeke, G. (Eds.). (2014). *Handbook of Missing Data Methodology*. CRC Press.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 581-592.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association* (Vol. 1, pp. 20-34). American Statistical Association.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Van Kuijk, S. M., Viechtbauer, W., Peeters, L. L., & Smits, L. (2016). Bias in regression Coefficient estimates when assumptions for handling missing data are violated: a simulation study. *Epidemiology, Biostatistics and Public Health*, *13*(1).

Rodgers-Farmer, A. Y., & Davis, D. (2001). Analyzing complex survey data. *Social Work Research*, *25*(3), 185.

Parzen, M., Lipsitz, S. R., & Fitzmaurice, G. M. (2005). A note on reducing the bias of the approximate Bayesian bootstrap imputation variance estimator. *Biometrika*, 971-974.

Rao, J. N., & Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, *79*(4), 811-822.

Rockwell, Richard C. 1975. An investigation of imputation and differential quality of data in the 1970 census. *Journal of the American Statistical Association* 70(349):39–42.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147-177.

Siddiqui, O., Hedeker, D., Flay, B. R., & Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. *American Journal of Epidemiology*, 144(4), 425-433.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys: Wiley Series in Probability and Mathematical Statistics.* New York: John Wiley & Sons.

Shieh, Y-Y. & Fouladi, R. T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement, 63*(6), 951-985.

Stockford, S. M. (2009). Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement (Doctoral dissertation, Arizona State University, 1990).

Sullivan, D., & Andridge, R. (2015). A hot deck imputation procedure for multiply imputing nonignorable missing data: The proxy pattern-mixture hot deck. *Computational Statistics & Data Analysis*, 82, 173-185.

Westreich, D. (2012). Berkson's bias, selection bias, and missing data. *Epidemiology (Cambridge, Mass.), 23*(1), 159.

White, I. R., & Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28), 2920-2931.

Zopluoglu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3(1), 242-278.

APPENDICES

Appendix A: SAS Program for Data Generation

MCAR Data

```
options ls = 250 ps = 500;
proc iml;
ICC = .00; * ICC conditions: 0, .25, and .50;
density = 1; * density conditions: 1, 2;
n_samples = 5000;
seed1=round(1000000*ranuni(0));
number_strata = 10;
strata_size = {30000,30000,50000,50000,50000,50000,
100000,100000,100000,100000}; *the # of schools in each stratum;
strata_means = {0,0.497,0.993,1.49,1.988,
2.485,2.982,3.479,3.976,4.47};
if ICC = 0 then do;
var_within = 100;
var_schools=0;
end;
if ICC = .25 then do; *ICC = var_schools/(var_schools + var_within);
var_within = 75;
var_schools=25;
end;
if ICC = .50 then do;
var_within = 50;
var_schools=50;
end;
if density = 1 then do; * low density;
number_schools = 100;
kids_min = 10; * min # of kids to sample from each school;
kids_max = 30; * max # of kids to sample from each school;
end;
if density = 2 then do; * high density;
number_schools = 20;
kids_min = 50;
kids_max = 150;
end;
full_R_matrix =
{1.00000 0.2935422 0.2890171 0.3300259 0.4292551 0.1717928 0.0536683 0.1084152 0.1515053,
0.2935422 1.00000 0.03716 -0.02342 0.02039 0.04689 0.07268 0.09224 0.05810,
0.2890171 0.03716 1.00000 -0.08097 0.05139 0.07601 0.11877 -0.06382 0.21698,
0.3300259 -0.02342 -0.08097 1.00000 -0.15033 -0.14001 -0.21079 0.11601 -0.13668,
0.4292551 0.02039 0.05139 -0.15033 1.00000 0.40799 0.16350 -0.05750 0.10849,
```

```
0.1717928 0.04689 0.07601 -0.14001 0.40799 1.00000 0.25853 -0.10975 0.17502,
0.0536683 0.07268 0.11877 -0.21079 0.16350 0.25853 1.00000 -0.20160 0.34115,
0.1084152 0.09224 -0.06382 0.11601 -0.05750 -0.10975 -0.20160 1.00000 -0.21985,
0.1515053 0.05810 0.21698 -0.13668 0.10849 0.17502 0.34115 -0.21985 1.00000};
* +-----------------------------------------------------------------------------------------+
Subroutine to generate a random sample.
User specifies the population means and standard deviations, as well as the correlation
matrix. For population shapes,Fleishman constants are used.
Inputs to the subroutine are
NN - desired sample size
mu - row vector of population means
variance - row vector of population variances
bb,cc,dd - Fleishman constants
r_matrix - population correlation matrix
Outputs are
Rawdata - matrix of NN observations
from the specified population
+-----------------------------------------------------------------------------------------+;
start gendata(NN,seed1,variance,bb,cc,dd,mu,r_matrix,rawdata);
COLS = NCOL(r_matrix);
G = ROOT(r_matrix);
rawdata=rannor(repeat(seed1,nn,COLS));
rawdata = rawdata*G;
do r = 1 to NN;
do c = 1 to COLS;
rawdata[r,c] = (-1*cc) + (bb*rawdata[r,c]) + (cc*rawdata[r,c]##2) + (dd*rawdata[r,c]##3);
rawdata[r,c] = (rawdata[r,c] * SQRT(variance[1,c])) + mu[1,c];
end;
end;
finish;
do replication = 1 to N_samples;
* +-----------------------------------------------------------------------------+
Randomly missing data indicators: Half the missingness occurs at the school level
+-----------------------------------------------------------------------------+;
nmiss1 = 0; need1 = .5#.1#number_Schools#number_strata;
nmiss2 = 0; need2 = .5#.3#number_Schools#number_strata;
nmiss3 = 0; need3 = .5#.5#number_Schools#number_strata;
nmiss4 = 0; need4 = .5#.7#number_Schools#number_strata;
total = number_schools#number_strata;
total_miss1 = 0;
total_miss2 = 0;
total_miss3 = 0;
total_miss4 = 0;
do stratum = 1 to nrow(strata_size);
school_prob = number_schools/strata_size[stratum,1];
do school = 1 to number_schools;
miss1 = 0;
miss2 = 0;
miss3 = 0;
miss4 = 0;
* Randomly pick a sample size (between max and min # of kids) to draw from each school;
N_students = (kids_min - 1) + round(uniform(seed1)*(kids_max - kids_min + 1) +.49999999);
school_mean=strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
school_mean_vec = school_mean;
```

```
do i=1 to 8;
school_meanx = strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
school_mean_vec = school_mean_vec || school_meanx;
end;
school_var_vec=REPEAT(var_within,9)`;
* +-----------------------------------------------------+
Randomly missing data indicators for entire schools
+-----------------------------------------------------+;
if ranuni(0) < need1/total then do;
nmiss1 = nmiss1 + 1;
need1 = need1 - 1;
miss1 = 1;
total_miss1 = total_miss1 + N_students;
end;
if ranuni(0) < need2/total then do;
nmiss2 = nmiss2 + 1;
need2 = need2 - 1;
miss2 = 1;
total_miss2 = total_miss2 + N_students;
end;
if ranuni(0) < need3/total then do;
nmiss3 = nmiss3 + 1;
need3 = need3 - 1;
miss3 = 1;
total_miss3 = total_miss3 + N_students;
end;
if ranuni(0) < need4/total then do;
nmiss4 = nmiss4 + 1;
need4 = need4 - 1;
miss4 = 1;
total_miss4 = total_miss4 + N_students;
end;
total = total - 1;
*generate sample of observations for one PSU ;
run GENDATA(N_students,seed1,school_var_vec,1,0,0,school_mean_vec,full_R_matrix,
student_sample);
ID = replication||stratum||school||school_prob||miss1||miss2||miss3||miss4||var_within||
var_schools||kids_min||kids_max||number_schools;
* Arbitrary school size = 800;
Kid_prob = N_students / 800;
ID_info = REPEAT(ID||Kid_prob,N_students);
final_sample = final_sample//(ID_info||student_sample);
end;
end;
* +-----------------------------------------------------------+
Randomly missing data indicators for individual students
Half of the total missingness occurs at the student level
+-----------------------------------------------------------+;
nmiss1 = 0; need1 = .5#.1#NROW(final_sample);
nmiss2 = 0; need2 = .5#.3#NROW(final_sample);
nmiss3 = 0; need3 = .5#.5#NROW(final_sample);
nmiss4 = 0; need4 = .5#.7#NROW(final_sample);
total1 = NROW(final_sample)-total_miss1;
total2 = NROW(final_sample)-total_miss2;
```

```
total3 = NROW(final_sample)-total_miss3;
total4 = NROW(final_sample)-total_miss4;
do i = 1 to NROW(final_sample);
t1 = 0; t2 = 0; t3 = 0; t4 = 0;
if final_sample[i,5] = 0 then t1 = 1;
if final_sample[i,6] = 0 then t2 = 1;
if final_sample[i,7] = 0 then t3 = 1;
if final_sample[i,8] = 0 then t4 = 1;
if (need1 > 0 & total1 > 0) then do;
if (final_sample[i,5] = 0 & ranuni(0) < need1/total1) then do;
nmiss1 = nmiss1 + 1;
need1 = need1 - 1;
final_sample[i,5] = 1; * set the record to be missing;
end;
end;
if (need2 > 0 & total2 > 0) then do;
if (final_sample[i,6] = 0 & ranuni(0) < need2/total2) then do;
nmiss2 = nmiss2 + 1;
need2 = need2 - 1;
final_sample[i,6] = 1; * set the record to be missing;
end;
end;
if (need3 > 0 & total3 > 0) then do;
if (final_sample[i,7] = 0 & ranuni(0) < need3/total3) then do;
nmiss3 = nmiss3 + 1;
need3 = need3 - 1;
final_sample[i,7] = 1; * set the record to be missing;
end;
end;
if (need4 > 0 & total4 > 0) then do;
if (final_sample[i,8] = 0 & ranuni(0) < need4/total4) then do;
nmiss4 = nmiss4 + 1;
need4 = need4 - 1;
final_sample[i,8] = 1;
end;
end;
total1 = total1 - t1;
total2 = total2 - t2;
total3 = total3 - t3;
total4 = total4 - t4;
end;
* Set the header for the dataset;
if replication = 1 then do;
cname = {"Replication" "Stratum" "School_ID" "School_Prob" "miss1" "miss2" "miss3" "miss4"
"Var_Within" "Var_Schools" "Kids_Min" "Kids_Max" "Number_Schools" "Kid_Prob"
"Y" "X1" "X2" "X3" "X4" "X5" "X6" "X7" "X8"};
create TEST_FILE from final_sample [ colname=cname ];
append from final_sample;
free final_sample;
end;
if replication > 1 then do;
* +-------------------------------------------------------------------------=+
Send simulated samples to regular SAS for analysis.
(the data set test_file contains the generated samples with no missing data.
```

```
the missing data conditions will be created by using the miss1 – miss4 variables)
+-----------------------------------------------------------------------------+;
setout TEST_FILE;
append from final_sample;
free final_sample;
end;
end; *end the replication loop;
quit; *end PROC IML;
*creating useable weight for each observation;
data test_file;
set test_file;
wt = 1/(school_prob*kid_prob);
proc means noprint data = test_file;
var wt;
by replication;
output out = q sum = sum_wt n = howmany;
*Normalizing the weights*;
data test_file;
merge test_file q;
by replication;
new_wt = wt*(howmany/sum_wt);
pctmiss = 0;
MDT = 'NM'; * No missing data;
ods listing close;
run;
```

MAR Data

```
options ls = 250 ps = 500;
proc iml;
ICC = .00; * ICC conditions: 0, .25, and .50;
density = 1; * density conditions: 1, 2;
n_samples = 5000;
seed1=round(1000000*ranuni(0));
number_strata = 10;
strata_size = {30000,30000,50000,50000,50000,50000,
100000,100000,100000,100000}; *# of schools in each stratum;
strata_means = {0,0.497,0.993,1.49,1.988,
2.485,2.982,3.479,3.976,4.47};
if ICC = 0 then do;
var_within = 100;
var_schools=0;
end;
if ICC = .25 then do; *ICC = var_schools/(var_schools + var_within);
var_within = 75;
var_schools=25;
end;
if ICC = .50 then do;
var_within = 50;
var_schools=50;
end;
if density = 1 then do; * low density;
```

```
number_schools = 100;
kids_min = 10; * min # of kids to sample from each school;
kids_max = 30; * max # of kids to sample from each school;
end;
if density = 2 then do; * high density;
number_schools = 20;
kids_min = 50;
kids_max = 150;
end;
full_R_matrix =
{1.00000 0.2935422 0.2890171 0.3300259 0.4292551 0.1717928 0.0536683 0.1084152 0.1515053,
0.2935422 1.00000 0.03716 -0.02342 0.02039 0.04689 0.07268 0.09224 0.05810,
0.2890171 0.03716 1.00000 -0.08097 0.05139 0.07601 0.11877 -0.06382 0.21698,
0.3300259 -0.02342 -0.08097 1.00000 -0.15033 -0.14001 -0.21079 0.11601 -0.13668,
0.4292551 0.02039 0.05139 -0.15033 1.00000 0.40799 0.16350 -0.05750 0.10849,
0.1717928 0.04689 0.07601 -0.14001 0.40799 1.00000 0.25853 -0.10975 0.17502,
0.0536683 0.07268 0.11877 -0.21079 0.16350 0.25853 1.00000 -0.20160 0.34115,
0.1084152 0.09224 -0.06382 0.11601 -0.05750 -0.10975 -0.20160 1.00000 -0.21985,
0.1515053 0.05810 0.21698 -0.13668 0.10849 0.17502 0.34115 -0.21985 1.00000};
* +-----------------------------------------------------------------------------------+
Subroutine to generate a random sample.
User specifies the population means and standard deviations, as well as the correlation
matrix. For population shapes,Fleishman constants are used.
Inputs to the subroutine are
NN - desired sample size
mu - row vector of population means
variance - row vector of population variances
bb,cc,dd - Fleishman constants
r_matrix - population correlation matrix
Outputs are
Rawdata - matrix of NN observations
from the specified population
+-----------------------------------------------------------------------------------+;
start gendata(NN,seed1,variance,bb,cc,dd,mu,r_matrix,rawdata);
COLS = NCOL(r_matrix);
G = ROOT(r_matrix);
rawdata=rannor(repeat(seed1,nn,COLS));
rawdata = rawdata*G;
do r = 1 to NN;
do c = 1 to COLS;
rawdata[r,c] = (-1*cc) + (bb*rawdata[r,c]) + (cc*rawdata[r,c]##2) + (dd*rawdata[r,c]##3);
rawdata[r,c] = (rawdata[r,c] * SQRT(variance[1,c])) + mu[1,c];
end;
end;
finish;
do replication = 1 to N_samples;
* +-----------------------------------------------------------------------------+
MAR Randomly missing data indicators: Half the missingness occurs at the school level
+-----------------------------------------------------------------------------+;
nmiss1a = 0; need1a = .8#.5#.1#number_Schools#number_strata;
nmiss1b = 0; need1b = .2#.5#.1#number_Schools#number_strata;
nmiss2a = 0; need2a = .8#.5#.3#number_Schools#number_strata;
nmiss2b = 0; need2b = .2#.5#.3#number_Schools#number_strata;
nmiss3a = 0; need3a = .8#.5#.5#number_Schools#number_strata;
```

```
nmiss3b = 0; need3b = .2#.5#.5#number_Schools#number_strata;
nmiss4a = 0; need4a = .8#.5#.7#number_Schools#number_strata;
nmiss4b = 0; need4b = .2#.5#.7#number_Schools#number_strata;
totala = .5#number_schools#number_strata;
totalb = .5#number_schools#number_strata;
total_miss1 = 0;
total_miss2 = 0;
total_miss3 = 0;
total_miss4 = 0;
sch_mean = j(totala+totalb,3,0);
do stratum = 1 to nrow(strata_size);
school_prob = number_schools/strata_size[stratum,1];
do school = 1 to number_schools;
miss1 = 0;
miss2 = 0;
miss3 = 0;
miss4 = 0;
* Randomly select a sample size to draw from each school;
N_students = (kids_min - 1) + round(uniform(seed1)*(kids_max - kids_min + 1) +.49999999);
school_mean=strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
*each variable will have different cluster mean ;
school_mean_vec = school_mean;
do i=1 to 8;
school_meanx = strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
school_mean_vec = school_mean_vec || school_meanx;
end;
sch_mean [(stratum-1)#number_schools+school,1] = school_mean_vec[1,2];
sch_mean [(stratum-1)#number_schools+school,2] = (N_students);
if (stratum-1)#number_schools+school = 1 then
sch_mean [(stratum-1)#number_schools+school,3]=(N_students);
if (stratum-1)#number_schools+school > 1 then
sch_mean [(stratum-1)#number_schools+school,3]=
(N_students+sch_mean [(stratum-1)#number_schools+school-1,3]);
school_var_vec=REPEAT(var_within,9)`;
run GENDATA(N_students,seed1,school_var_vec,1,0,0,school_mean_vec,full_R_matrix,
student_sample);
ID = replication||stratum||school||school_prob||miss1||miss2||miss3||miss4||var_within||
var_schools||kids_min||kids_max||number_schools;
* probability of kids selection, given that school was sampled. Arbitrary school size = 800;
Kid_prob = N_students / 800;
ID_info = REPEAT(ID||Kid_prob,N_students);
final_sample = final_sample//(ID_info||student_sample);
end;
end; * end replication loop; *at this stage, the sample simulated hasn't had missing assinged yet
*rank the school mean from 1 to 200 which is the # of school in high density strata;
Rank_sch_mean = rank(sch_mean[,1]);
in_loop = J(1,2,0);
*compute the value of a half of # of schools ;
half = (totala + totalb)/2;
* +-----------------------------------------------------+
MAR indicators for entire schools in the upper half of the distribution
+-----------------------------------------------------+;
do i = 1 to (totala+totalb);
if Rank_sch_mean[i] > half then do;
```

```
in_loop[1,1] = in_loop[1,1]+1;
if ranuni(0) < need1a/totala then do;
nmiss1a = nmiss1a + 1;
need1a = need1a - 1;
total_miss1 = total_miss1 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],5]=1;
if i = 1 then final_sample [1:sch_mean[i,3],5]=1;
end;
if ranuni(0) < need2a/totala then do;
nmiss2a = nmiss2a + 1;
need2a = need2a - 1;
total_miss2 = total_miss2 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],6]=1;
if i = 1 then final_sample [1:sch_mean[i,3],6]=1;
end;
if ranuni(0) < need3a/totala then do;
nmiss3a = nmiss3a + 1;
need3a = need3a - 1;
total_miss3 = total_miss3 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],7]=1;
if i = 1 then final_sample [1:sch_mean[i,3],7]=1;
end;
if ranuni(0) < need4a/totala then do;
nmiss4a = nmiss4a + 1;
need4a = need4a - 1;
total_miss4 = total_miss4 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],8]=1;
if i = 1 then final_sample [1:sch_mean[i,3],8]=1;
end;
totala = totala - 1;
end; *end the upper half rank loop;
* +-----------------------------------------------------+
MAR indicators for entire schools in the lower half of the distribution
+-----------------------------------------------------+;
if Rank_sch_mean[i] <= half then do;
in_loop[1,2] = in_loop[1,2]+1;
if ranuni(0) < need1b/totalb then do;
nmiss1b = nmiss1b + 1;
need1b = need1b - 1;
total_miss1 = total_miss1 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],5]=1;
if i = 1 then final_sample [1:sch_mean[i,3],5]=1;
end;
if ranuni(0) < need2b/totalb then do;
nmiss2b = nmiss2b + 1;
need2b = need2b - 1;
total_miss2 = total_miss2 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],6]=1;
if i = 1 then final_sample [1:sch_mean[i,3],6]=1;
end;
if ranuni(0) < need3b/totalb then do;
nmiss3b = nmiss3b + 1;
need3b = need3b - 1;
total_miss3 = total_miss3 + sch_mean[i,2];
```

```
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],7]=1;
if i = 1 then final_sample [1:sch_mean[i,3],7]=1;
end;
if ranuni(0) < need4b/totalb then do;
nmiss4b = nmiss4b + 1;
need4b = need4b - 1;
total_miss4 = total_miss4 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],8]=1;
if i = 1 then final_sample [1:sch_mean[i,3],8]=1;
end;
totalb = totalb - 1;
end; *end the LOWER half rank loop;
end; *end the sample loop ;
* +-----------------------------------------------------------------------------+
MAR Randomly missing data indicators: Half the missingness occurs at the student level
+-----------------------------------------------------------------------------+;
nmiss1a = 0; need1a = .8#.5#.1#NROW(final_sample);
nmiss1b = 0; need1b = .2#.5#.1#NROW(final_sample);
nmiss2a = 0; need2a = .8#.5#.3#NROW(final_sample);
nmiss2b = 0; need2b = .2#.5#.3#NROW(final_sample);
nmiss3a = 0; need3a = .8#.5#.5#NROW(final_sample);
nmiss3b = 0; need3b = .2#.5#.5#NROW(final_sample);
nmiss4a = 0; need4a = .8#.5#.7#NROW(final_sample);
nmiss4b = 0; need4b = .2#.5#.7#NROW(final_sample);
total1 = NROW(final_sample)-total_miss1;
total2 = NROW(final_sample)-total_miss2;
total3 = NROW(final_sample)-total_miss3;
total4 = NROW(final_sample)-total_miss4;
new_vector1 = j(total1,1,0);
new_vector2 = j(total2,1,0);
new_vector3 = j(total3,1,0);
new_vector4 = j(total4,1,0);
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,5] = 0 then do;
new_vector1[t,1] = final_sample[i,16];
t=t+1;
end;
end;
rank_vector1 = rank(new_vector1);
do i = 1 to NROW(rank_vector1);
if rank_vector1[i] = round(total1/2) then do;
median1 = new_vector1[i];
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,6] = 0 then do;
new_vector2[t,1] = final_sample[i,16];
t=t+1;
end;
end;
rank_vector2 = rank(new_vector2);
do i = 1 to NROW(rank_vector2);
```

```
if rank_vector2[i] = round(total2/2) then do;
median2 = new_vector2[i];
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,7] = 0 then do;
new_vector3[t,1] = final_sample[i,16];
t=t+1;
end;
end;
rank_vector3 = rank(new_vector3);
do i = 1 to NROW(rank_vector3);
if rank_vector3[i] = round(total3/2) then do;
median3 = new_vector3[i];
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,8] = 0 then do;
new_vector4[t,1] = final_sample[i,16];
t=t+1;
end;
end;
rank_vector4 = rank(new_vector4);
do i = 1 to NROW(rank_vector4);
if rank_vector4[i] = round(total4/2) then do;
median4 = new_vector4[i];
end;
end;
in_loop = J(4,2,0);
total1 = 0.5#(NROW(final_sample)-total_miss1);
total2 = 0.5#(NROW(final_sample)-total_miss2);
total3 = 0.5#(NROW(final_sample)-total_miss3);
total4 = 0.5#(NROW(final_sample)-total_miss4);
* +-----------------------------------------------------------+
MAR indicators for individual students in the upper half
of the distribution
+-----------------------------------------------------------+;
do i = 1 to NROW(final_sample);
t1 = 0; t2 = 0; t3 = 0; t4 = 0;
if final_sample[i,5] = 0 then t1 = 1;
if final_sample[i,6] = 0 then t2 = 1;
if final_sample[i,7] = 0 then t3 = 1;
if final_sample[i,8] = 0 then t4 = 1;
if total1 <= 0 then total1 = 1;
if total2 <= 0 then total2 = 1;
if total3 <= 0 then total3 = 1;
if total4 <= 0 then total4 = 1;
if (final_sample[i,16] > median1 & need1a > 0 & total1 > 0) then do;
in_loop[1,1] = in_loop[1,1] + 1;
if (final_sample[i,5] = 0 & ranuni(0) < need1a/total1) then do;
nmiss1a = nmiss1a + 1;
need1a = need1a - 1;
```

```
final_sample[i,5] = 1;
end;
end;
if (final_sample[i,16] > median2 & need2a > 0 & total2 > 0) then do;
in_loop[2,1] = in_loop[2,1] + 1;
if (final_sample[i,6] = 0 & ranuni(0) < need2a/total2) then do;
nmiss2a = nmiss2a + 1;
need2a = need2a - 1;
final_sample[i,6] = 1;
end;
end;
if (final_sample[i,16] > median3 & need3a > 0 & total3 > 0) then do;
in_loop[3,1] = in_loop[3,1] + 1;
if (final_sample[i,7] = 0 & ranuni(0) < need3a/total3) then do;
nmiss3a = nmiss3a + 1;
need3a = need3a - 1;
final_sample[i,7] = 1;
end;
end;
if (final_sample[i,16] > median4 & need4a > 0 & total4 > 0) then do;
in_loop[4,1] = in_loop[4,1] + 1;
if (final_sample[i,8] = 0 & ranuni(0) < need4a/total4) then do;
nmiss4a = nmiss4a + 1;
need4a = need4a - 1;
final_sample[i,8] = 1;
end;
end;
total1 = total1 - t1;
total2 = total2 - t2;
total3 = total3 - t3;
total4 = total4 - t4;
end;
total1 = 0.5#(NROW(final_sample)-total_miss1); * Total of records that are not missing (yet);
total2 = 0.5#(NROW(final_sample)-total_miss2);
total3 = 0.5#(NROW(final_sample)-total_miss3);
total4 = 0.5#(NROW(final_sample)-total_miss4);
do i = 1 to NROW(final_sample);
t1 = 0; t2 = 0; t3 = 0; t4 = 0;
if final_sample[i,5] = 0 then t1 = 1;
if final_sample[i,6] = 0 then t2 = 1;
if final_sample[i,7] = 0 then t3 = 1;
if final_sample[i,8] = 0 then t4 = 1;
if total1 <= 0 then total1 = 1;
if total2 <= 0 then total2 = 1;
if total3 <= 0 then total3 = 1;
if total4 <= 0 then total4 = 1;
* +------------------------------------------------------------+
MAR indicators for individual students in the lower half
of the distribution
+------------------------------------------------------------+;
if (final_sample[i,16] <= median1 & need1b > 0 & total1 > 0) then do;
in_loop[1,2] = in_loop[1,2] + 1;
if (final_sample[i,5] = 0 & ranuni(0) < need1b/total1) then do;
nmiss1b = nmiss1b + 1;
```

```
need1b = need1b - 1;
final_sample[i,5] = 1;
end;
end;
if (final_sample[i,16] <= median2 & need2b > 0 & total2 > 0) then do;
in_loop[2,2] = in_loop[2,2] + 1;
if (final_sample[i,6] = 0 & ranuni(0) < need2b/total2) then do;
nmiss2b = nmiss2b + 1;
need2b = need2b - 1;
final_sample[i,6] = 1;
end;
end;
if (final_sample[i,16] <= median3 & need3b > 0 & total3 > 0) then do;
in_loop[3,2] = in_loop[3,2] + 1;
if (final_sample[i,7] = 0 & ranuni(0) < need3b/total3) then do;
nmiss3b = nmiss3b + 1;
need3b = need3b - 1;
final_sample[i,7] = 1;
end;
end;
if (final_sample[i,16] <= median4 & need4b > 0 & total4 > 0) then do;
in_loop[4,2] = in_loop[4,2] + 1;
if (final_sample[i,8] = 0 & ranuni(0) < need4b/total4) then do;
nmiss4b = nmiss4b + 1;
need4b = need4b - 1;
final_sample[i,8] = 1;
end;
end;
total1 = total1 - t1;
total2 = total2 - t2;
total3 = total3 - t3;
total4 = total4 - t4;
end;
missing_students1 = nmiss1a + nmiss1b;
missing_students2 = nmiss2a + nmiss2b;
missing_students3 = nmiss3a + nmiss3b;
missing_students4 = nmiss4a + nmiss4b;
if replication = 1 then do;
* +-----------------------------------------------------+
Send simulated samples to regular SAS for analysis
+-----------------------------------------------------+;
cname = {"Replication" "Stratum" "School_ID" "School_Prob" "miss1" "miss2" "miss3" "miss4"
"Var_Within" "Var_Schools" "Kids_Min" "Kids_Max" "Number_Schools"
"Kid_Prob" "Y" "X1" "X2" "X3" "X4" "X5" "X6" "X7" "X8"};
create TEST_FILE from final_sample [ colname=cname ];
append from final_sample;
free final_sample;
end;
if replication > 1 then do;
* +-----------------------------------------------------+
Send simulated samples to regular SAS for analysis
+-----------------------------------------------------+;
setout TEST_FILE;
append from final_sample;
```

```
free final_sample;
end;
end; * end the big replication loop;
quit; *end PROC IML;
**creating useable weight for each observation**;
data test_file;
set test_file;
wt = 1/(school_prob*kid_prob);
proc means noprint data = test_file;
var wt;
by replication;
output out = q sum = sum_wt n = howmany;
**Normalizing the weights**;
data test_file;
merge test_file q;
by replication;
new_wt = wt*(howmany/sum_wt);
pctmiss = 0;
MDT = 'NM'; * No missing data;
ods listing close;
run;
```

MNAR Data

```
options ls = 250 ps = 500;
proc iml;
ICC = .00; * ICC conditions: 0, .25, and .50;
density = 1; * density conditions: 1, 2;
n_samples = 5000;
seed1=round(1000000*ranuni(0));
number_strata = 10;
strata_size = {30000,30000,50000,50000,50000,50000,
100000,100000,100000,100000}; *# of clusters in each stratum;
strata_means = {0,0.497,0.993,1.49,1.988,
2.485,2.982,3.479,3.976,4.47};
if ICC = 0 then do;
var_within = 100;
var_schools=0;
end;
if ICC = .25 then do; *ICC = var_schools/(var_schools + var_within);
var_within = 75;
var_schools=25;
end;
if ICC = .50 then do;
var_within = 50;
var_schools=50;
end;
if density = 1 then do; * low density;
number_schools = 100;
kids_min = 10; * min # of kids to sample from each school;
kids_max = 30; * max # of kids to sample from each school;
end;
if density = 2 then do;
```

```
number_schools = 20;
kids_min = 50;
kids_max = 150;
end;
full_R_matrix =
{1.00000 0.2935422 0.2890171 0.3300259 0.4292551 0.1717928 0.0536683 0.1084152 0.1515053,
0.2935422 1.00000 0.03716 -0.02342 0.02039 0.04689 0.07268 0.09224 0.05810,
0.2890171 0.03716 1.00000 -0.08097 0.05139 0.07601 0.11877 -0.06382 0.21698,
0.3300259 -0.02342 -0.08097 1.00000 -0.15033 -0.14001 -0.21079 0.11601 -0.13668,
0.4292551 0.02039 0.05139 -0.15033 1.00000 0.40799 0.16350 -0.05750 0.10849,
0.1717928 0.04689 0.07601 -0.14001 0.40799 1.00000 0.25853 -0.10975 0.17502,
0.0536683 0.07268 0.11877 -0.21079 0.16350 0.25853 1.00000 -0.20160 0.34115,
0.1084152 0.09224 -0.06382 0.11601 -0.05750 -0.10975 -0.20160 1.00000 -0.21985,
0.1515053 0.05810 0.21698 -0.13668 0.10849 0.17502 0.34115 -0.21985 1.00000};
* +-----------------------------------------------------------------------------------+
Subroutine to generate a random sample.
User specifies the population means and standard deviations, as well as the correlation
matrix. For population shapes,Fleishman constants are used.
Inputs to the subroutine are
NN - desired sample size
mu - row vector of population means
variance - row vector of population variances
bb,cc,dd - Fleishman constants
r_matrix - population correlation matrix
Outputs are
Rawdata - matrix of NN observations
from the specified population
+-----------------------------------------------------------------------------------+;
start gendata(NN,seed1,variance,bb,cc,dd,mu,r_matrix,rawdata);
COLS = NCOL(r_matrix);
G = ROOT(r_matrix);
rawdata=rannor(repeat(seed1,nn,COLS));
rawdata = rawdata*G;
do r = 1 to NN;
do c = 1 to COLS;
rawdata[r,c] = (-1*cc) + (bb*rawdata[r,c]) + (cc*rawdata[r,c]##2) + (dd*rawdata[r,c]##3);
rawdata[r,c] = (rawdata[r,c] * SQRT(variance[1,c])) + mu[1,c];
end;
end;
finish;
do replication = 1 to N_samples;
* +-----------------------------------------------------------------------------+
MNAR Randomly missing data indicators: Half the missingness occurs at the school level
+-----------------------------------------------------------------------------+;
nmiss1a = 0; need1a = .8#.5#.1#number_Schools#number_strata;
nmiss1b = 0; need1b = .2#.5#.1#number_Schools#number_strata;
nmiss2a = 0; need2a = .8#.5#.3#number_Schools#number_strata;
nmiss2b = 0; need2b = .2#.5#.3#number_Schools#number_strata;
nmiss3a = 0; need3a = .8#.5#.5#number_Schools#number_strata;
nmiss3b = 0; need3b = .2#.5#.5#number_Schools#number_strata;
nmiss4a = 0; need4a = .8#.5#.7#number_Schools#number_strata;
nmiss4b = 0; need4b = .2#.5#.7#number_Schools#number_strata;
totala = .5#number_schools#number_strata;
totalb = .5#number_schools#number_strata;
```

```
total_miss1 = 0;
total_miss2 = 0;
total_miss3 = 0;
total_miss4 = 0;
sch_mean = j(totala+totalb,3,0);
do stratum = 1 to nrow(strata_size);
school_prob = number_schools/strata_size[stratum,1];
do school = 1 to number_schools;
miss1 = 0;
miss2 = 0;
miss3 = 0;
miss4 = 0;
* Randomly select a sample size to draw from each school;
N_students = (kids_min - 1) + round(uniform(seed1)*(kids_max - kids_min + 1) +.49999999);
school_mean=strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
*each variable will have different cluster mean ;
school_mean_vec = school_mean;
do i=1 to 8;
school_meanx = strata_means[stratum,1]+(rannor(seed1)* SQRT(var_schools));
school_mean_vec = school_mean_vec || school_meanx;
end;
*compute mean of X7&X8 ;
x7x8mean = (school_mean_vec[1,8] + school_mean_vec[1,9])/2;
sch_mean [(stratum-1)#number_schools+school,1] = x7x8mean;
sch_mean [(stratum-1)#number_schools+school,2] = (N_students);
if (stratum-1)#number_schools+school = 1 then
sch_mean [(stratum-1)#number_schools+school,3]=(N_students);
if (stratum-1)#number_schools+school > 1 then
sch_mean [(stratum-1)#number_schools+school,3]=
(N_students+sch_mean [(stratum-1)#number_schools+school-1,3]);
school_var_vec=REPEAT(var_within,9)`;
run GENDATA(N_students,seed1,school_var_vec,1,0,0,school_mean_vec,full_R_matrix,
student_sample);
ID = replication||stratum||school||school_prob||miss1||miss2||miss3||miss4||var_within||
var_schools||kids_min||kids_max||number_schools;
* probability of kids selection, given that school was sampled. Arbitrary school size = 800;
Kid_prob = N_students / 800;
ID_info = REPEAT(ID||Kid_prob,N_students);
final_sample = final_sample//(ID_info||student_sample);
end; * end school loop;
end; *end replication loop; *at this stage, the sample simulated hasn't had missing assinged yet;
*rank the school mean from 1 to 200 which is the # of school in high density strata;
Rank_sch_mean = rank(sch_mean[,1]);
in_loop = J(1,2,0); *create a matrix of zero with 1 row and 2 columns;
*compute the value of a half of # of schools ;
half = (totala + totalb)/2;
* +-----------------------------------------------------+
MNAR indicators for entire schools in the upper half of the distribution
+-----------------------------------------------------+;
do i = 1 to (totala+totalb);
if Rank_sch_mean[i] > half then do;
in_loop[1,1] = in_loop[1,1]+1;
if ranuni(0) < need1a/totala then do;
nmiss1a = nmiss1a + 1;
```

```
need1a = need1a - 1;
total_miss1 = total_miss1 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],5]=1;
if i = 1 then final_sample [1:sch_mean[i,3],5]=1;
end;
if ranuni(0) < need2a/totala then do;
nmiss2a = nmiss2a + 1;
need2a = need2a - 1;
total_miss2 = total_miss2 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],6]=1;
if i = 1 then final_sample [1:sch_mean[i,3],6]=1;
end;
if ranuni(0) < need3a/totala then do;
nmiss3a = nmiss3a + 1;
need3a = need3a - 1;
total_miss3 = total_miss3 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],7]=1;
if i = 1 then final_sample [1:sch_mean[i,3],7]=1;
end;
if ranuni(0) < need4a/totala then do;
nmiss4a = nmiss4a + 1;
need4a = need4a - 1;
total_miss4 = total_miss4 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],8]=1;
if i = 1 then final_sample [1:sch_mean[i,3],8]=1;
end;
totala = totala - 1;
end; *end the upper half rank loop;
* +-----------------------------------------------------+
MNAR indicators for entire schools in the lower half of the distribution
+-----------------------------------------------------+;
if Rank_sch_mean[i] <= half then do;
in_loop[1,2] = in_loop[1,2]+1;
if ranuni(0) < need1b/totalb then do;
nmiss1b = nmiss1b + 1;
need1b = need1b - 1;
total_miss1 = total_miss1 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],5]=1;
if i = 1 then final_sample [1:sch_mean[i,3],5]=1;
end;
if ranuni(0) < need2b/totalb then do;
nmiss2b = nmiss2b + 1;
need2b = need2b - 1;
total_miss2 = total_miss2 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],6]=1;
if i = 1 then final_sample [1:sch_mean[i,3],6]=1;
end;
if ranuni(0) < need3b/totalb then do;
nmiss3b = nmiss3b + 1;
need3b = need3b - 1;
total_miss3 = total_miss3 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],7]=1;
if i = 1 then final_sample [1:sch_mean[i,3],7]=1;
end;
```

```
if ranuni(0) < need4b/totalb then do;
nmiss4b = nmiss4b + 1;
need4b = need4b - 1;
total_miss4 = total_miss4 + sch_mean[i,2];
if i > 1 then final_sample [sch_mean[i-1,3]+1:sch_mean[i,3],8]=1;
if i = 1 then final_sample [1:sch_mean[i,3],8]=1;
end;
totalb = totalb - 1;
end; *end the LOWER half rank loop;
end; *end the sample loop ;
* +-----------------------------------------------------------------------------+
MNAR Randomly missing data indicators: Half the missingness occurs at the student level
+-----------------------------------------------------------------------------+;
nmiss1a = 0; need1a = .8#.5#.1#NROW(final_sample);
nmiss1b = 0; need1b = .2#.5#.1#NROW(final_sample);
nmiss2a = 0; need2a = .8#.5#.3#NROW(final_sample);
nmiss2b = 0; need2b = .2#.5#.3#NROW(final_sample);
nmiss3a = 0; need3a = .8#.5#.5#NROW(final_sample);
nmiss3b = 0; need3b = .2#.5#.5#NROW(final_sample);
nmiss4a = 0; need4a = .8#.5#.7#NROW(final_sample);
nmiss4b = 0; need4b = .2#.5#.7#NROW(final_sample);
total1 = NROW(final_sample)-total_miss1;
total2 = NROW(final_sample)-total_miss2;
total3 = NROW(final_sample)-total_miss3;
total4 = NROW(final_sample)-total_miss4;
new_vector1 = j(total1,1,0);
new_vector2 = j(total2,1,0);
new_vector3 = j(total3,1,0);
new_vector4 = j(total4,1,0);
t=1;
*for each level of missing%, obtain the median value for mean of X7 and X8;
do i = 1 to NROW(final_sample);
if final_sample[i,5] = 0 then do;
new_vector1[t,1] = (final_sample[i,22] + final_sample[i,23])/2;
t=t+1;
end;
end;
rank_vector1 = rank(new_vector1);
do i = 1 to NROW(rank_vector1);
if rank_vector1[i] = round(total1/2) then do;
median1 = new_vector1[i];
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,6] = 0 then do;
new_vector2[t,1] = (final_sample[i,22] + final_sample[i,23])/2;
t=t+1;
end;
end;
rank_vector2 = rank(new_vector2);
do i = 1 to NROW(rank_vector2);
if rank_vector2[i] = round(total2/2) then do;
median2 = new_vector2[i];
```

```
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,7] = 0 then do;
new_vector3[t,1] = (final_sample[i,22] + final_sample[i,23])/2;
t=t+1;
end;
end;
rank_vector3 = rank(new_vector3);
do i = 1 to NROW(rank_vector3);
if rank_vector3[i] = round(total3/2) then do;
median3 = new_vector3[i];
end;
end;
t=1;
do i = 1 to NROW(final_sample);
if final_sample[i,8] = 0 then do;
new_vector4[t,1] = (final_sample[i,22] + final_sample[i,23])/2;
t=t+1;
end;
end;
rank_vector4 = rank(new_vector4);
do i = 1 to NROW(rank_vector4);
if rank_vector4[i] = round(total4/2) then do;
median4 = new_vector4[i];
end;
end;
* END obtain median value;
in_loop = J(4,2,0);
total1 = 0.5#(NROW(final_sample)-total_miss1);
total2 = 0.5#(NROW(final_sample)-total_miss2);
total3 = 0.5#(NROW(final_sample)-total_miss3);
total4 = 0.5#(NROW(final_sample)-total_miss4);
* +----------------------------------------------------------+
MNAR indicators for individual students in the upper half
of the distribution
+----------------------------------------------------------+;
do i = 1 to NROW(final_sample);
t1 = 0; t2 = 0; t3 = 0; t4 = 0;
if final_sample[i,5] = 0 then t1 = 1;
if final_sample[i,6] = 0 then t2 = 1;
if final_sample[i,7] = 0 then t3 = 1;
if final_sample[i,8] = 0 then t4 = 1;
if total1 <= 0 then total1 = 1;
if total2 <= 0 then total2 = 1;
if total3 <= 0 then total3 = 1;
if total4 <= 0 then total4 = 1;
meanx7x8 = (final_sample[i,22] + final_sample[i,23])/2;
if (meanx7x8 > median1 & need1a > 0 & total1 > 0) then do;
in_loop[1,1] = in_loop[1,1] + 1;
if (final_sample[i,5] = 0 & ranuni(0) < need1a/total1) then do;
nmiss1a = nmiss1a + 1;
need1a = need1a - 1;
```

```
final_sample[i,5] = 1;
end;
end;
if (meanx7x8 > median2 & need2a > 0 & total2 > 0) then do;
in_loop[2,1] = in_loop[2,1] + 1;
if (final_sample[i,6] = 0 & ranuni(0) < need2a/total2) then do;
nmiss2a = nmiss2a + 1;
need2a = need2a - 1;
final_sample[i,6] = 1;
end;
end;
if (meanx7x8 > median3 & need3a > 0 & total3 > 0) then do;
in_loop[3,1] = in_loop[3,1] + 1;
if (final_sample[i,7] = 0 & ranuni(0) < need3a/total3) then do;
nmiss3a = nmiss3a + 1;
need3a = need3a - 1;
final_sample[i,7] = 1;
end;
end;
if (meanx7x8 > median4 & need4a > 0 & total4 > 0) then do;
in_loop[4,1] = in_loop[4,1] + 1;
if (final_sample[i,8] = 0 & ranuni(0) < need4a/total4) then do;
nmiss4a = nmiss4a + 1;
need4a = need4a - 1;
final_sample[i,8] = 1;
end;
end;
total1 = total1 - t1;
total2 = total2 - t2;
total3 = total3 - t3;
total4 = total4 - t4;
end;
total1 = 0.5#(NROW(final_sample)-total_miss1);
total2 = 0.5#(NROW(final_sample)-total_miss2);
total3 = 0.5#(NROW(final_sample)-total_miss3);
total4 = 0.5#(NROW(final_sample)-total_miss4);
do i = 1 to NROW(final_sample);
t1 = 0; t2 = 0; t3 = 0; t4 = 0;
if final_sample[i,5] = 0 then t1 = 1;
if final_sample[i,6] = 0 then t2 = 1;
if final_sample[i,7] = 0 then t3 = 1;
if final_sample[i,8] = 0 then t4 = 1;
if total1 <= 0 then total1 = 1;
if total2 <= 0 then total2 = 1;
if total3 <= 0 then total3 = 1;
if total4 <= 0 then total4 = 1;
* +------------------------------------------------------------+
MNAR indicators for individual students in the lower half
of the distribution
+------------------------------------------------------------+;
meanx7x8 = (final_sample[i,22] + final_sample[i,23])/2;
if (meanx7x8 <= median1 & need1b > 0 & total1 > 0) then do;
in_loop[1,2] = in_loop[1,2] + 1;
if (final_sample[i,5] = 0 & ranuni(0) < need1b/total1) then do;
```

```
nmiss1b = nmiss1b + 1;
need1b = need1b - 1;
final_sample[i,5] = 1;
end;
end;
if (meanx7x8 <= median2 & need2b > 0 & total2 > 0) then do;
in_loop[2,2] = in_loop[2,2] + 1;
if (final_sample[i,6] = 0 & ranuni(0) < need2b/total2) then do;
nmiss2b = nmiss2b + 1;
need2b = need2b - 1;
final_sample[i,6] = 1;
end;
end;
if (meanx7x8 <= median3 & need3b > 0 & total3 > 0) then do;
in_loop[3,2] = in_loop[3,2] + 1;
if (final_sample[i,7] = 0 & ranuni(0) < need3b/total3) then do;
nmiss3b = nmiss3b + 1;
need3b = need3b - 1;
final_sample[i,7] = 1;
end;
end;
if (meanx7x8 <= median4 & need4b > 0 & total4 > 0) then do;
in_loop[4,2] = in_loop[4,2] + 1;
if (final_sample[i,8] = 0 & ranuni(0) < need4b/total4) then do;
nmiss4b = nmiss4b + 1;
need4b = need4b - 1;
final_sample[i,8] = 1;
end;
end;
total1 = total1 - t1;
total2 = total2 - t2;
total3 = total3 - t3;
total4 = total4 - t4;
end;
missing_students1 = nmiss1a + nmiss1b;
missing_students2 = nmiss2a + nmiss2b;
missing_students3 = nmiss3a + nmiss3b;
missing_students4 = nmiss4a + nmiss4b;
if replication = 1 then do;
* +-----------------------------------------------------+
Send simulated samples to regular SAS for analysis
+-----------------------------------------------------+;
cname = {"Replication" "Stratum" "School_ID" "School_Prob" "miss1" "miss2" "miss3"
"miss4" "Var_Within" "Var_Schools" "Kids_Min" "Kids_Max"
"Number_Schools" "Kid_Prob" "Y" "X1" "X2" "X3" "X4" "X5" "X6" "X7" "X8"};
create TEST_FILE from final_sample [ colname=cname ];
append from final_sample;
free final_sample;
end;
if replication > 1 then do;
* +-----------------------------------------------------+
Send simulated samples to regular SAS for analysis
+-----------------------------------------------------+;
setout TEST_FILE;
```

```
append from final_sample;
free final_sample;
end;
end; * end the big replication loop;
quit; *end PROC IML;
**creating useable weight for each observation**;
data test_file;
set test_file;
wt = 1/(school_prob*kid_prob);
proc means noprint data = test_file;
var wt;
by replication;
output out = q sum = sum_wt n = howmany;
**Normalizing the weights**;
data test_file;
merge test_file q;
by replication;
new_wt = wt*(howmany/sum_wt);
pctmiss = 0;
MDT = 'NM'; * No missing data;
run;
```

Appendix B: SAS Code for Extracting and Imputing Missing Data Using Regression-Based Multiple Imputation, and Analyzing Imputed Data

```
options ls = 250 ps = 500 ;
**------------------------------------------------------------------------------;
**Extract the 10% missing dataset from the simulated dataset**;
**------------------------------------------------------------------------------;
ods listing close;
data Miss;
set mcar_d2_icc00;
pctmiss=.1;
if miss1 = 1 then do;
x7 = .;
x8 = .;
end;
MDT = 'RM';
descode = 1000*stratum + school_id;
error_df1 = number_schools*10 - 10; *EDF needed in the proc MIanalysis;
call symput('error_df',trim(left(put(error_df1,8.))));
run;
*do regression-based MULTIPLE IMPUTATION on 10% missing data;
proc mi data = miss nimpute=10 out = mi_out;
by replication;
class descode;
FCS;
var y x1 - x8 descode new_wt;
run;
*Analyze each of every individual imputed dataset ;
```

```
proc surveyreg data = mi_out;
cluster school_id;
strata stratum;
weight new_wt;
model y = x1 x2 x3 x4 x5 x6 x7 x8/ clparm deff;
by replication _imputation_ var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
ods output parameterestimates = mi_out2;
run;
proc sort data = mi_out2;
by replication parameter _imputation_ var_within var_schools kids_min kids_max
number_schools pctmiss MDT;
run;
*combine results obtained from indiviual imputed dataset;
proc mianalyze data = mi_out2 EDF = &error_df;
by replication parameter var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
modeleffects estimate;
stderr stderr;
ods output parameterestimates=mi_out3;
run;
data mi_out3;
set mi_out3;
* Evaluation of results;
if probt < .001 then reject = 1; else reject = 0;
if probt = . then reject = .;
CI_Width = UCLMean - LCLMean;
if var_within = 100 then ICC = 0;
if var_within = 75 then ICC = .25;
if var_within = 50 then ICC = .50;
if kids_min = 10 then density = 'Low ';
if kids_min = 50 then density = 'High';
inband = 0;
if icc = 0 then do;
if Parameter = 'X1' then truth_B = 0.2610322;
if Parameter = 'X2' then truth_B = 0.2649848;
if Parameter = 'X3' then truth_B = 0.4177301;
if Parameter = 'X4' then truth_B = 0.4612977;
if Parameter = 'X5' then truth_B = -0.0058010;
if Parameter = 'X6' then truth_B = -0.0085900;
if Parameter = 'X7' then truth_B = 0.0886790;
if Parameter = 'X8' then truth_B = 0.0992116;
end;
if icc = .25 then do;
if Parameter = 'X1' then truth_B = 0.2009905;
if Parameter = 'X2' then truth_B = 0.2017216;
if Parameter = 'X3' then truth_B = 0.3038786;
if Parameter = 'X4' then truth_B = 0.3297423;
if Parameter = 'X5' then truth_B = 0.0349549;
if Parameter = 'X6' then truth_B = 0.0014293;
if Parameter = 'X7' then truth_B = 0.0793123;
if Parameter = 'X8' then truth_B = 0.0778457;
end;
if icc = .50 then do;
```

```
if Parameter = 'X1' then truth_B = 0.1415543;
if Parameter = 'X2' then truth_B = 0.1380404;
if Parameter = 'X3' then truth_B = 0.1923219;
if Parameter = 'X4' then truth_B = 0.2174791;
if Parameter = 'X5' then truth_B = 0.0457584;
if Parameter = 'X6' then truth_B = 0.0098770;
if Parameter = 'X7' then truth_B = 0.0570907;
if Parameter = 'X8' then truth_B = 0.0581175;
end;
if LCLMean < truth_B and UCLMean > truth_B then inband = 1;
bias = estimate - truth_B;
relbias = bias/truth_B;
MSE = (estimate - truth_B)**2;
if Parameter = 'Intercept' then delete;
run;
proc print data = mi_out3;
title1 'SAS SurveyReg RM Treatment';
run;
proc sort data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
run;
proc means noprint data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
var reject ci_width inband bias relbias mse stderr;
output out = mi_out4 mean = ;
run;
data mi_out4;
set mi_out4;
RMSE = SQRT(MSE);
run;
ods listing;
proc print data = mi_out4 heading = horizontal;
run;
proc datasets;
delete miss mi_out mi_out2 mi_out3 mi_out4;
run; quit;
**------------------------------------------------------------;
**Extract the 30% missing dataset from the simulated dataset**;
**------------------------------------------------------------;
ods listing close;
data Miss;
set mcar_d2_icc00;
pctmiss=.3;
if miss2 = 1 then do;
x7 = .;
x8 = .;
end;
MDT = 'RM';
descode = 1000*stratum + school_id;
error_df1 = number_schools*10 - 10;
call symput('error_df',trim(left(put(error_df1,8.))));
run;
*do regression-based MULTIPLE IMPUTATION on 30% missing data;
proc mi data = miss nimpute=10 out = mi_out;
```

```
by replication;
class descode;
FCS;
var y x1 - x8 descode new_wt;
run;
proc surveyreg data = mi_out;
cluster school_id;
strata stratum;
weight new_wt;
model y = x1 x2 x3 x4 x5 x6 x7 x8/ clparm deff;
by replication _imputation_ var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
ods output parameterestimates = mi_out2;
run;
proc sort data = mi_out2;
by replication parameter _imputation_ var_within var_schools kids_min kids_max
number_schools pctmiss MDT;
run;
proc mianalyze data = mi_out2 EDF = &error_df;
by replication parameter var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
modeleffects estimate;
stderr stderr;
ods output parameterestimates=mi_out3;
run;
data mi_out3;
set mi_out3;
* Evaluation of results;
if probt < .001 then reject = 1; else reject = 0;
if probt = . then reject = .;
CI_Width = UCLMean - LCLMean;
if var_within = 100 then ICC = 0;
if var_within = 75 then ICC = .25;
if var_within = 50 then ICC = .50;
if kids_min = 10 then density = 'Low ';
if kids_min = 50 then density = 'High';
inband = 0;
if icc = 0 then do;
if Parameter = 'X1' then truth_B = 0.2610322;
if Parameter = 'X2' then truth_B = 0.2649848;
if Parameter = 'X3' then truth_B = 0.4177301;
if Parameter = 'X4' then truth_B = 0.4612977;
if Parameter = 'X5' then truth_B = -0.0058010;
if Parameter = 'X6' then truth_B = -0.0085900;
if Parameter = 'X7' then truth_B = 0.0886790;
if Parameter = 'X8' then truth_B = 0.0992116;
end;
if icc = .25 then do;
if Parameter = 'X1' then truth_B = 0.2009905;
if Parameter = 'X2' then truth_B = 0.2017216;
if Parameter = 'X3' then truth_B = 0.3038786;
if Parameter = 'X4' then truth_B = 0.3297423;
if Parameter = 'X5' then truth_B = 0.0349549;
if Parameter = 'X6' then truth_B = 0.0014293;
```

```
if Parameter = 'X7' then truth_B = 0.0793123;
if Parameter = 'X8' then truth_B = 0.0778457;
end;
if icc = .50 then do;
if Parameter = 'X1' then truth_B = 0.1415543;
if Parameter = 'X2' then truth_B = 0.1380404;
if Parameter = 'X3' then truth_B = 0.1923219;
if Parameter = 'X4' then truth_B = 0.2174791;
if Parameter = 'X5' then truth_B = 0.0457584;
if Parameter = 'X6' then truth_B = 0.0098770;
if Parameter = 'X7' then truth_B = 0.0570907;
if Parameter = 'X8' then truth_B = 0.0581175;
end;
if LCLMean < truth_B and UCLMean > truth_B then inband = 1;
bias = estimate - truth_B;
relbias = bias/truth_B;
MSE = (estimate - truth_B)**2;
if Parameter = 'Intercept' then delete;
run;
proc print data = mi_out3;
title1 'SAS SurveyReg RM Treatment';
run;
proc sort data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
run;
proc means noprint data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
var reject ci_width inband bias relbias mse stderr;
output out = mi_out4 mean = ;
run;
data mi_out4;
set mi_out4;
RMSE = SQRT(MSE);
run;
ods listing;
proc print data = mi_out4 heading = horizontal;
run;
proc datasets;
delete miss mi_out mi_out2 mi_out3 mi_out4;
run; quit;
**------------------------------------------------------------;
**Extract the 50% missing dataset from the raw dataset**;
**------------------------------------------------------------;
ods listing close;
data Miss;
set mcar_d2_icc00;
pctmiss=.5;
if miss3 = 1 then do;
x7 = .;
x8 = .;
end;
MDT = 'RM';
descode = 1000*stratum + school_id;
error_df1 = number_schools*10 - 10;
```

```
call symput('error_df',trim(left(put(error_df1,8.))));
run;
*do regression-based MULTIPLE IMPUTATION on 50% missing data;
proc mi data = miss nimpute=10 out = mi_out;
by replication;
class descode;
FCS;
var y x1 - x8 descode new_wt;
run;
proc surveyreg data = mi_out;
cluster school_id;
strata stratum;
weight new_wt;
model y = x1 x2 x3 x4 x5 x6 x7 x8/ clparm deff;
by replication _imputation_ var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
ods output parameterestimates = mi_out2;
run;
proc sort data = mi_out2;
by replication parameter _imputation_ var_within var_schools kids_min kids_max
number_schools pctmiss MDT;
run;
proc mianalyze data = mi_out2 EDF = &error_df;
by replication parameter var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
modeleffects estimate;
stderr stderr;
ods output parameterestimates=mi_out3;
run;
data mi_out3;
set mi_out3;
* Evaluation of results;
if probt < .001 then reject = 1; else reject = 0;
if probt = . then reject = .;
CI_Width = UCLMean - LCLMean;
if var_within = 100 then ICC = 0;
if var_within = 75 then ICC = .25;
if var_within = 50 then ICC = .50;
if kids_min = 10 then density = 'Low ';
if kids_min = 50 then density = 'High';
inband = 0;
if icc = 0 then do;
if Parameter = 'X1' then truth_B = 0.2610322;
if Parameter = 'X2' then truth_B = 0.2649848;
if Parameter = 'X3' then truth_B = 0.4177301;
if Parameter = 'X4' then truth_B = 0.4612977;
if Parameter = 'X5' then truth_B = -0.0058010;
if Parameter = 'X6' then truth_B = -0.0085900;
if Parameter = 'X7' then truth_B = 0.0886790;
if Parameter = 'X8' then truth_B = 0.0992116;
end;
if icc = .25 then do;
if Parameter = 'X1' then truth_B = 0.2009905;
if Parameter = 'X2' then truth_B = 0.2017216;
```

```
if Parameter = 'X3' then truth_B = 0.3038786;
if Parameter = 'X4' then truth_B = 0.3297423;
if Parameter = 'X5' then truth_B = 0.0349549;
if Parameter = 'X6' then truth_B = 0.0014293;
if Parameter = 'X7' then truth_B = 0.0793123;
if Parameter = 'X8' then truth_B = 0.0778457;
end;
if icc = .50 then do;
if Parameter = 'X1' then truth_B = 0.1415543;
if Parameter = 'X2' then truth_B = 0.1380404;
if Parameter = 'X3' then truth_B = 0.1923219;
if Parameter = 'X4' then truth_B = 0.2174791;
if Parameter = 'X5' then truth_B = 0.0457584;
if Parameter = 'X6' then truth_B = 0.0098770;
if Parameter = 'X7' then truth_B = 0.0570907;
if Parameter = 'X8' then truth_B = 0.0581175;
end;
if LCLMean < truth_B and UCLMean > truth_B then inband = 1;
bias = estimate - truth_B;
relbias = bias/truth_B;
MSE = (estimate - truth_B)**2;
if Parameter = 'Intercept' then delete;
run;
proc print data = mi_out3;
title1 'SAS SurveyReg RM Treatment';
run;
proc sort data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
run;
proc means noprint data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
var reject ci_width inband bias relbias mse stderr;
output out = mi_out4 mean = ;
run;
data mi_out4;
set mi_out4;
RMSE = SQRT(MSE);
run;
ods listing;
proc print data = mi_out4 heading = horizontal;
run;
proc datasets;
delete miss mi_out mi_out2 mi_out3 mi_out4;
run; quit;
**------------------------------------------------------------;
**Extract the 70% missing dataset from the raw dataset**;
**------------------------------------------------------------;
ods listing close;
data Miss;
set mcar_d2_icc00;
pctmiss=.7;
if miss4 = 1 then do;
x7 = .;
x8 = .;
```

```
end;
MDT = 'RM';
descode = 1000*stratum + school_id;
error_df1 = number_schools*10 - 10;
call symput('error_df',trim(left(put(error_df1,8.))));
run;
*do regression-based MULTIPLE IMPUTATION on 70% missing data;
proc mi data = miss nimpute=10 out = mi_out;
by replication;
class descode;
FCS;
var y x1 - x8 descode new_wt;
run;
proc surveyreg data = mi_out;
cluster school_id;
strata stratum;
weight new_wt;
model y = x1 x2 x3 x4 x5 x6 x7 x8/ clparm deff;
by replication _imputation_ var_within var_schools kids_min kids_max
number_schools pctmiss MDT;
ods output parameterestimates = mi_out2;
run;
proc sort data = mi_out2;
by replication parameter _imputation_ var_within var_schools kids_min kids_max
number_schools pctmiss MDT;
run;
proc mianalyze data = mi_out2 EDF = &error_df;
by replication parameter var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
modeleffects estimate;
stderr stderr;
ods output parameterestimates=mi_out3;
run;
data mi_out3;
set mi_out3;
* Evaluation of results;
if probt < .001 then reject = 1; else reject = 0;
if probt = . then reject = .;
CI_Width = UCLMean - LCLMean;
if var_within = 100 then ICC = 0;
if var_within = 75 then ICC = .25;
if var_within = 50 then ICC = .50;
if kids_min = 10 then density = 'Low ';
if kids_min = 50 then density = 'High';
inband = 0;
if icc = 0 then do;
if Parameter = 'X1' then truth_B = 0.2610322;
if Parameter = 'X2' then truth_B = 0.2649848;
if Parameter = 'X3' then truth_B = 0.4177301;
if Parameter = 'X4' then truth_B = 0.4612977;
if Parameter = 'X5' then truth_B = -0.0058010;
if Parameter = 'X6' then truth_B = -0.0085900;
if Parameter = 'X7' then truth_B = 0.0886790;
if Parameter = 'X8' then truth_B = 0.0992116;
```

```
end;
if icc = .25 then do;
if Parameter = 'X1' then truth_B = 0.2009905;
if Parameter = 'X2' then truth_B = 0.2017216;
if Parameter = 'X3' then truth_B = 0.3038786;
if Parameter = 'X4' then truth_B = 0.3297423;
if Parameter = 'X5' then truth_B = 0.0349549;
if Parameter = 'X6' then truth_B = 0.0014293;
if Parameter = 'X7' then truth_B = 0.0793123;
if Parameter = 'X8' then truth_B = 0.0778457;
end;
if icc = .50 then do;
if Parameter = 'X1' then truth_B = 0.1415543;
if Parameter = 'X2' then truth_B = 0.1380404;
if Parameter = 'X3' then truth_B = 0.1923219;
if Parameter = 'X4' then truth_B = 0.2174791;
if Parameter = 'X5' then truth_B = 0.0457584;
if Parameter = 'X6' then truth_B = 0.0098770;
if Parameter = 'X7' then truth_B = 0.0570907;
if Parameter = 'X8' then truth_B = 0.0581175;
end;
if LCLMean < truth_B and UCLMean > truth_B then inband = 1;
bias = estimate - truth_B;
relbias = bias/truth_B;
MSE = (estimate - truth_B)**2;
if Parameter = 'Intercept' then delete;
run;
proc print data = mi_out3;
title1 'SAS SurveyReg RM Treatment';
run;
proc sort data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
run;
proc means noprint data = mi_out3;
by parameter var_within var_schools kids_min kids_max number_schools pctmiss MDT;
var reject ci_width inband bias relbias mse stderr;
output out = mi_out4 mean = ;
run;
data mi_out4;
set mi_out4;
RMSE = SQRT(MSE);
run;
ods listing;
proc print data = mi_out4 heading = horizontal;
run;
ods listing close;
proc datasets;
delete miss mi_out mi_out2 mi_out3 mi_out4;
run; quit;
```

Appendix C: Multiple Imputation Using SAS Package