

Missing Data and Complex Sample Surveys Using SAS®: The Impact of Listwise Deletion vs. Multiple Imputation Methods on Point and Interval Estimates when Data are MCAR, MAR, and MNAR

Anh P. Kellermann and Jeffrey D. Kromrey, University of South Florida

ABSTRACT

Social scientists from many fields use secondary data analysis of complex sample surveys to answer research questions and test hypotheses. Despite great care taken to obtain the data needed, missing data are frequently found in such samples. Even though missing data is a ubiquitous problem, the methodological literature has provided little guidance to inform the appropriate treatment for such missingness. This Monte Carlo study used SAS® to investigate the impact of missing data treatment (hot deck-based multiple imputations versus regression-based multiple imputation versus listwise deletion) when data are MCAR, MAR, and MNAR. By using 10% to 70% of missing data (along with complete sample conditions as a reference point for interpretation of results), the research focused on the parameter estimates in multiple regression analysis in complex sample data. Results are presented in terms of statistical bias in the parameter estimates and both confidence interval coverage and width.

INTRODUCTION

COMPLEX SAMPLE SURVEY DATA

Secondary data analysis of nationally representative surveys is commonly conducted by researchers and can be extremely useful when investigating a variety of social and behavioral outcomes. By providing access to a vast array of variables on large numbers of people and their environments such as schools or neighborhoods, the nature of large-scale social science data is enticing to many researchers. The degree to which unbiased estimates and accurate inferences can be made from complex samples depends, however, on the care researchers take when analyzing the 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. For example, the probabilities of selection of the observations are often not equal leading to the need to incorporate sample weights. Further, multi-stage sampling yields clustered observations in which the variance among units within each cluster is less than the variance among units in general (i.e., intraclass correlation), which complicates the estimation of sampling error. In addition, stratification in the sampling design (e.g., geographical stratification) insures appropriate sample representation on the stratification variable(s), but yields negatively biased estimates of the population variance when not considered in the analysis. Finally, adjustments can be applied to the sample for unit nonresponse and other post-stratification to allow unbiased estimates of population characteristics (Brick & Kalton, 1996).

Sample Weights

Observations from complex sample surveys are typically weighted such that an observation's weight is based on the reciprocal of the observation's probability of being selected. That is, observations more likely to be selected (e.g., from oversampling) receive smaller weights than observations less likely to be selected. In data available from large-scale surveys, weights may be provided such that the sum of the weights equals the sample size (relative weights)

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

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

has a sample of n_{jk} individuals. Alternatively, weights may be provided such that the sum of the weights equals the population size (raw weights)

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

Sample weights are then applied in the computation of statistics from the sample observations. For example, 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} stratum, and other symbols are defined as they were previously. When researchers omit sample weights from the analysis of complex survey data, parameter estimates are typically biased and incorrect inferences can be drawn. Further, when sample weights are not used, findings are generally not representative of the larger population of interest.

Estimation of Variances

The estimation of sampling error is a critical component of survey analysis. Sampling error provides an index of the precision of point estimates (e.g., sample means or regression coefficients) and is used in the calculation of confidence intervals and hypothesis tests. For complex sample surveys (involving stratification, multi-stage sampling, and unequal probabilities of selection) the calculation of sampling error differs from the calculation used in simple random sampling. As an example of multi-stage sampling consider a context where schools are sampled within geographically defined strata, and then students are sampled within schools. In this sort of situation, the sampling variance of a statistic can be obtained by focusing on the between-cluster variance estimate of the statistic (Williams, 2000). Specific variance estimates can be obtained through Taylor series linearization, or through replication methods, such as balanced repeated replications, jackknife, and bootstrap methods (Skinner, Holt, & Smith, 1989).

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

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

where W_{jk} is the sum of the weights in the j^{th} PSU in the k^{th} stratum, $\bar{X}_{jk}$ is the mean for the j^{th} PSU in the k^{th} stratum, and the other symbols are as previously defined.

MISSING DATA

Missing data are a ubiquitous problem in social science research. As most common statistical procedures assume complete data, failing to address the presence of missing values can lead to a multitude of issues, ranging from decreased power to heightened bias and inaccurate Type I error control. Among the many methods developed to deal with missing data problems, multiple imputation (MI) has become one of the most accepted. Provided the assumptions are met, MI has shown excellent qualities (see, for example, reviews by Graham, Cumsille, & Elek-Fisk, 2003; Graham & Hoffer, 2000; Shafer & Olsen, 1998).

Missing data problems become even more complicated when we consider variation in the missing data mechanism. While most statistical analyses assume data are missing completely at random (MCAR) or missing at random (MAR), this is often not an assumption that can be tested using sample data; also, data that are missing not at random (MNAR) are not testable using sample data. In order to determine how robust our methodological choices are under different mechanisms for missingness, the missing data techniques are investigated using data that are MCAR, MAR, and MNAR.

Rubin's Missing Data Taxonomy

When applying a missing data treatment, e.g., MI, the process or mechanism that causes missingness might need to be incorporated into the statistical model because incorporating the missing data mechanism leads to efficient data modeling for fitting the data adequately. In other words, the missing data treatment should reflect the underlying uncertainty of the process or mechanism that best explains why data are missing (Rubin, 2005). Rubin's (1976) theoretical framework is the most widely accepted for missing data research.

Let $Y = (Y_1, Y_2, \dots, Y_p)$ be a random vector variable for p -dimensional multivariate data ($y \times p$ matrix), and θ and δ are the parameter of the data and the parameter of the conditional probability of the observed pattern of missing data, respectively. In the presence of missing data, ϕ , inferences about θ are conditional on δ , the mechanism that causes missingness. That is, given the data at hand, the purpose of the missing data process is to allow making valid inferences about θ but these inferences depend upon or are conditional on δ , the mechanism that generated missing data.

Missing Completely at Random (MCAR). When the probability of a missing value on a variable is independent of the observed or unobserved value for any variable, that is, the probability that y_i is missing is independent of the probability of missingness for any y_j , data are missing completely at random or MCAR.

Missing at Random (MAR). When the probability that y_i is missing for variable Y is dependent on the data of any other variables but not on the variable Y of interest, data are missing at random or MAR. In addition, MAR requires data on a variable to be missing randomly within subgroups (Roth, 1994).

Missing not at Random (MNAR). When the probability of a missing value depends on unobserved data or data that could have been observed, data are missing not at random or MNAR. That is, the missing value of y_i depends on the value of y_i . Under this scenario, why data are missing is not ignorable, thus requiring that the missing data mechanism is modeled to make valid inferences about the model parameter.

Treatment of Missing Data

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. One of the oldest elimination based methods is listwise deletion (LW), which excludes observations with any missing variable values from the analysis. Two of the most commonly used imputation-based missing data treatment methods 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. 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.

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 (MI), a more advanced imputation method; multiple imputation replaces each missing value with a set of m plausible values that represent the uncertainty about the right value to impute. MI creates m imputed datasets, then each imputed data set is analyzed independently and the results are combined to provide parameter estimates (point and interval estimates) that incorporate the uncertainty resulting from the missing data. Any single imputation method can be used to obtain the imputed values in multiple imputation. 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 multiple imputation (HM) versus regression-based multiple imputation (RM).

PURPOSE

The purpose of this study is to investigate the impact of the three missing data treatment methods (i.e., LW, HM, and RM), when the data are MCAR, MAR, and MNAR, in the context of multiple regression analysis of complex sample data using popular SAS procedures.

METHOD

In this Monte Carlo study, complex samples were generated from multivariate populations and each sample was analyzed using listwise deletion, hot deck-based MI, and regression-based MI. The sample simulation included both stratification and cluster sampling. Specifically, observations from ten strata were simulated in which each stratum differed in population means on all variables, with the maximum difference in stratum means being twice as large as the between PSU standard deviation. From each stratum, the sampling of PSUs and subsequently, observations within PSUs were simulated, controlling the relative variance between and within PSUs to produce target values of the intraclass correlation.

The Monte Carlo study included four factors in the design: population density, sample size, levels of intraclass correlation between cluster variance and total variance (ICC), and levels of missing data. The number of PSUs sampled from each of the ten strata was linked with the number of observations sampled from each PSU to provide 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). In addition to the number of PSUs and the sample size per PSU, the intraclass correlation was manipulated to investigate the effects of different degrees of observation clustering. Three levels of intraclass correlation were simulated (.00, .25, and .50) by controlling the ratio of the between PSU variance to the within PSU variance. Finally, in addition to complete samples, four levels of missing data were simulated: 10%, 30%, 50%, and 70%. Within each of these levels, 50% of the 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. Three missing data mechanisms were simulated to produce data that were either MCAR, MAR, or MNAR (Little & Rubin, 1987; Schafer & Graham, 2002).

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 would provide 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. The correlation matrix used in the simulations is provided in Table 1. Observations within each sample were weighted so

that the sample weight was proportional to the inverse probability of selection (taking into account the probability of PSU selection from the stratum and observation selection from the PSU) and the sample weights were incorporated in subsequent analyses.

	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

Table 1. Correlation Matrix Used as Template for Data Simulation

For all the MCAR, MAR, and MNAR design factors, the data generation was conducted using SAS/IML version 9.4, and each sample for all three missing data factors was analyzed separately, using the listwise deletion, PROC MI, and PROC SURVEYIMPUTE in SAS to treat missing data; for statistical analysis, the SURVEYREG procedure for complex sample survey analysis was used. Conditions for the study were run under Linux platform. Normally distributed random variables were 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 each condition investigated in this study, 5,000 samples were generated. The outcomes of interest in this simulation study included both point estimates (the bias and sampling error of the regression coefficients) and interval estimates (confidence interval coverage and width for the coefficients).

Generally, using multiple imputation approach involves three steps: imputing missing data multiple times, analyzing the individual, imputed datasets using standard statistical procedures, and pooling the results. In this study, these three steps were carried out using SAS procedures, specifically, PROC MI and PROC SURVEYIMPUTE were used to produce regression-based MI and hot-deck-based MI respectively; each simulated incomplete multivariate data were imputed 10 times resulting in 10 imputed datasets to be analyzed using PROC SURVEYREG; parameter estimates obtained from PROC SURVEYREG were pooled using PROC MIANALYZE.

The following statements invoked the MI procedure and imputed missing values for the “miss” data set. The SAS variable ‘decode’ is a categorical variable that represents the stratum number and PSU as a single composite variable (see Berglund & Heeringa, 2014). The FCS statement requests multivariate imputation by fully conditional specification methods. The 10 imputed datasets were concatenated and stored in the dataset named “mi_out” with the variable –Imputation– indicating the imputation numbers.

```
proc mi data = miss out = mi_out nimpute=10;
  by replication;
  class decode;
  FCS;
  var y x1 - x8 decode new_wt;
run;
```

The following statements invoked the SURVEYIMPUTE procedure and imputed missing values for variables x7 and x8 in the “miss” data set. For every missing value, 10 donors were selected randomly from the same stratum. The imputation was independently repeated 10 times (i.e., the missing values of x7 and x8 were imputed 10 times) and the output data were stored in the dataset named “hd_out” which contains the data of 10 individual, imputed datasets. The dataset, hd_out, was then restructured such that each individual, imputed dataset is identified by _IMPUTATION_ variable; (sample code to restructure the imputed data can be found in The SURVEYIMPUTE Procedure User Guide).

```
proc surveyimpute data=miss method=hotdeck(selection=ABB) ndonors=10;
var x7 x8;
weight new_wt;
cells stratum;
output out=hd_out;
run;
```

In the 2nd step, the SURVEYREG procedure was used to do regression analysis on the imputed sample data. The PROC SURVEYREG can handle complex survey sample designs with stratification, clustering, and unequal weighting. It computes regression coefficients and provides significance tests for any specified estimable parameters. The WEIGHT statement was used to correctly weigh the population estimates; if an observation had a missing value or a nonpositive value for the WEIGHT variable, then PROC SURVEYREG excluded that observation from the analysis. An observation was also excluded if it had a missing value for any STRATA variable, CLUSTER variable, dependent variable, or any variable used in the independent effects. PROC SURVEYREG used the Output Delivery System (ODS) to create output data sets. In the below code snippet, the BY _IMPUTATION_ statement was used to request separate analyses of each MI _IMPUTATION_ repetition.

```
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 _imputation_ var_within var_schools kids_min kids_max number_schools
pctmiss MDT;
ods output parameterestimates = mi_out2;
run;
```

The MIANALYZE procedure combined the results of analyses of imputations and generates statistical inferences. The MIANALYZE procedure read parameter estimates and associated standard errors or covariance matrix that were computed by the PROC SURVEYREG for each imputed data set. The MIANALYZE procedure then derived valid univariate inference for these parameters. In the following code snippet, The EDF option was used to ensure a correct determination of the complete data degrees of freedom used in the construction of MI confidence intervals for the population statistics of interests (df = #PSUs - #Strata).

```
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= mianalyze_out;
run;
```

RESULTS

MCAR

Across the MCAR conditions, the estimated bias yielded by LW and the bias obtained from the complete samples (NM) were nearly identical; whereas, the estimated bias produced by RM and HM appeared to vary more in both directions with the negative bias showing greater variability; and RM appeared to produce more outliers than HM (Figure 1).

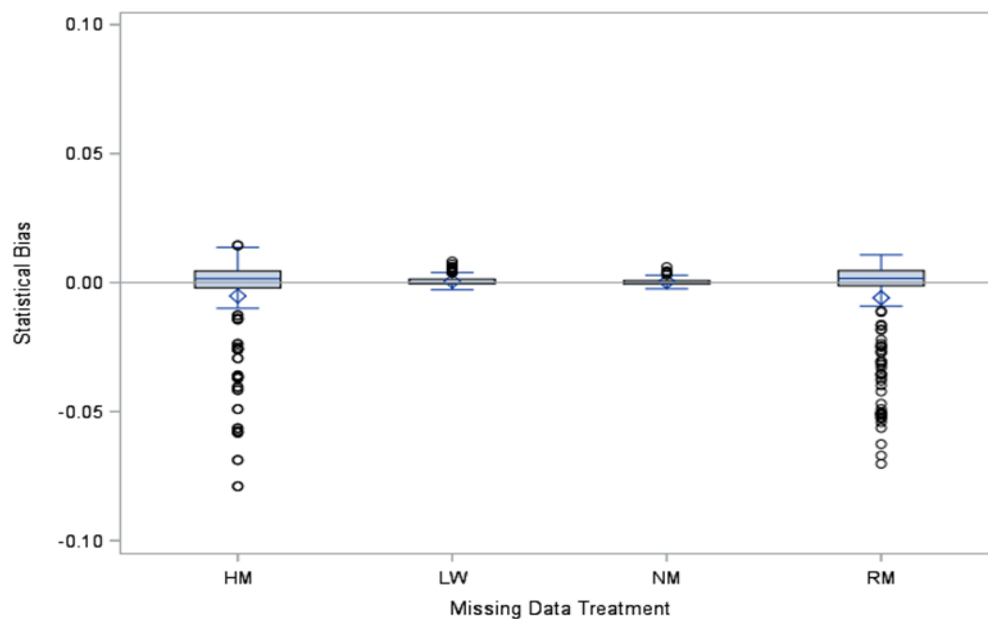


Figure 1. Distributions of Bias by Missing Data Treatment in MCAR

In addition, RMSE estimates associated with LW method were close to those obtained from the complete data samples (NM) in both magnitude and variation, and they appeared smaller compared to those associated with HM and RM. RMSE estimates associated with RM were larger and varied more compared to those associated with HM (Figure 2).

An eta squared analysis revealed that the percent of missing data, by itself, had either very weak or negligible influence on the point estimates bias and RMSE produced by the three studied missing data treatment methods (MDTs) in MCAR. Specifically, percent of missing data was found accounting for 3% of variability in bias and 6% of variability in RMSE estimates produced by HM and .3% of variability in bias and .002% of variability in RMSE produced by RM.

The main influencing factor on the bias and RMSE produced by the three studied MDTs was found to be the parameter factor (i.e., the differences among the parameter estimates for the eight regressors); parameter factor accounted for most of the variability in bias yielded by HM ($\eta^2 = 62\%$) and by RM ($\eta^2 = 76\%$); parameter was also the most influencing factor accounting for variability in RMSE yielded by RM ($\eta^2 = 63\%$), by LW ($\eta^2 = 64\%$) and by HM ($\eta^2 = 24\%$).

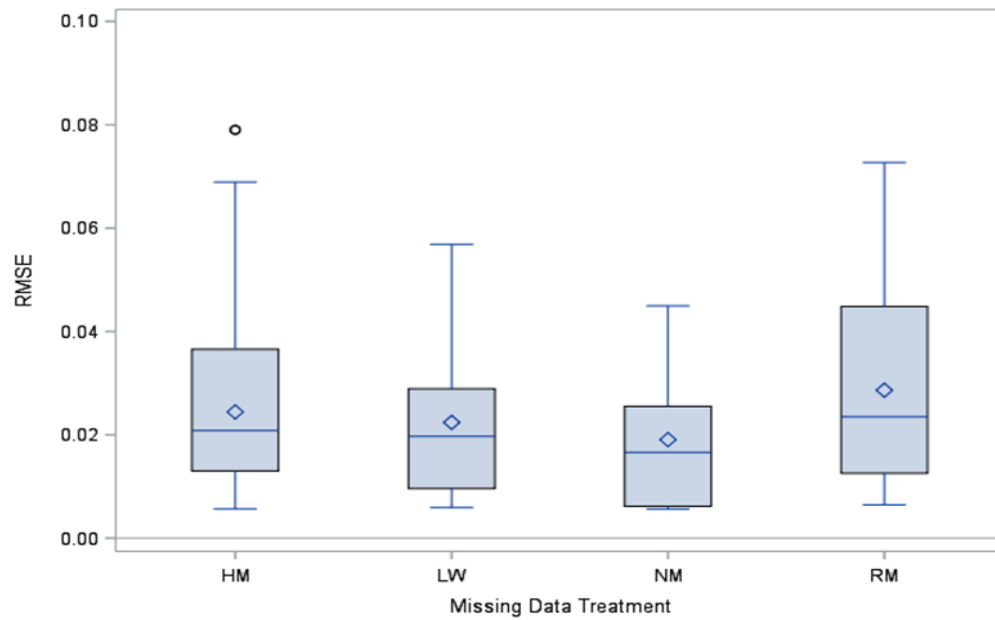


Figure 2. Distributions of RMSE by Missing Data Treatment in MCAR

Figures 3 and 4 describe the distribution of bias estimates by parameter for HM and RM respectively. In both figures, bias estimates associated with X7 and X8 were larger and varied more compared to those associated with other regressors; it's worth mentioning that X7 and X8 were the two variables with missing data.

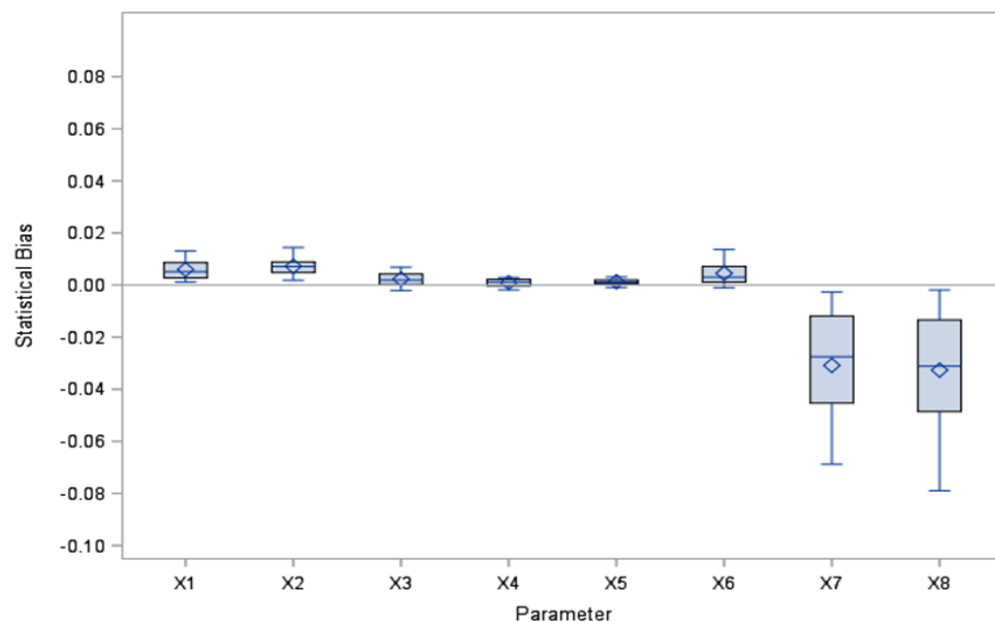


Figure 3. Distributions of Bias Estimates for HM by Parameter in MCAR

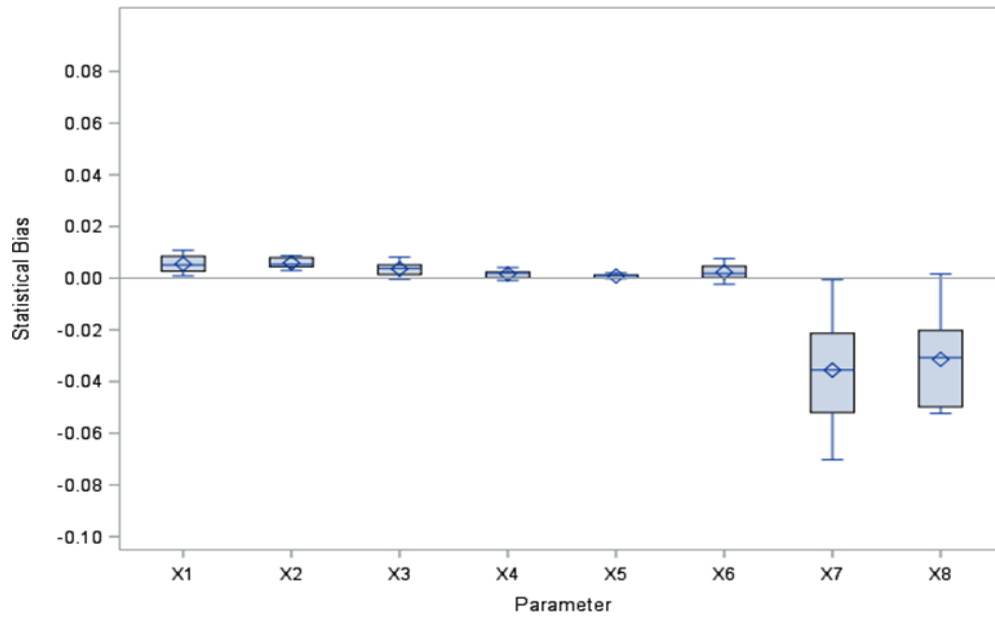


Figure 4. Distributions of Bias Estimates for RM by Parameter in MCAR

Figures 5 and 6 describe the distribution of RMSE by parameter for HM and RM method respectively. RMSE estimates associated with X7 and X8 were larger than those associated with other regressors. The figures also show that RMSE associated with X7 and X8 yielded by RM were higher and varied less compared to those yielded by HM. Besides being moderately influenced by parameter, RMSE produced by HM was also modestly influenced by the interaction between ICC and parameter ($\eta^2 = 16\%$), ICC ($\eta^2 = 15\%$), and interaction between percent missing and parameter ($\eta^2 = 13\%$).

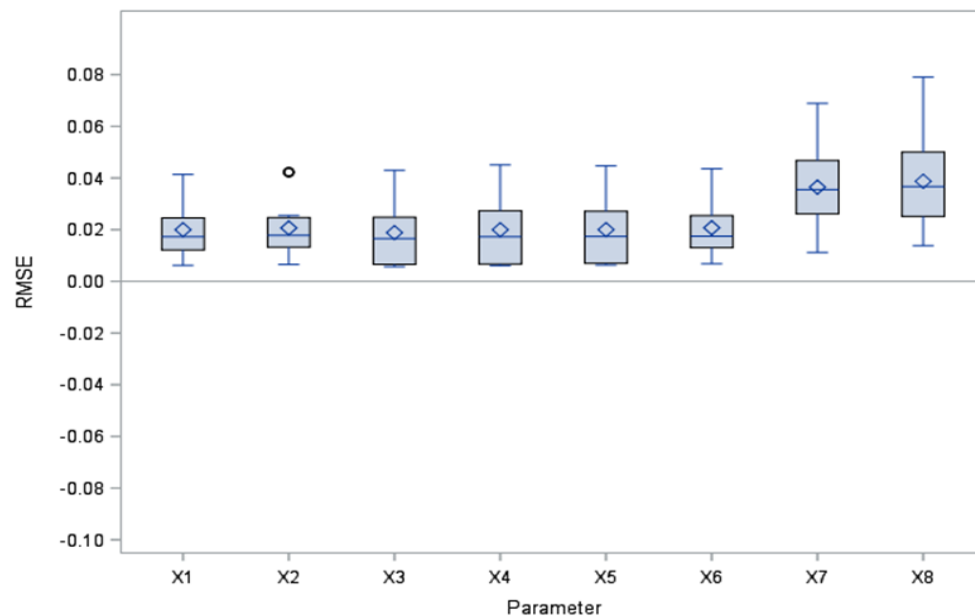


Figure 5. Distributions of RMSE Estimates for HM by Parameter in MCAR

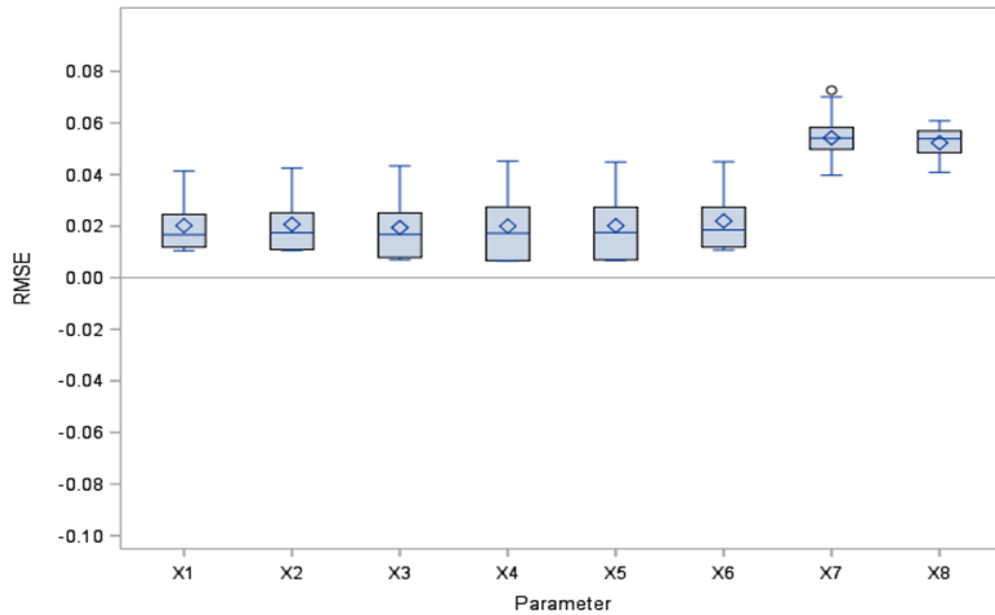


Figure 6. Distributions of RMSE Estimates for RM by Parameter in MCAR

While RMSE yielded by HM and RM were mostly affected by parameter, RMSE yielded by LW was mostly affected by ICC ($\eta^2 = 64\%$); and as expected (due to the design effect), ICC also strongly influenced RMSE obtained from complete samples ($\eta^2 = 66\%$). Figure 7 describes the distribution of RMSE estimated by LW by ICC: RMSE got larger and varied more as ICC increased.

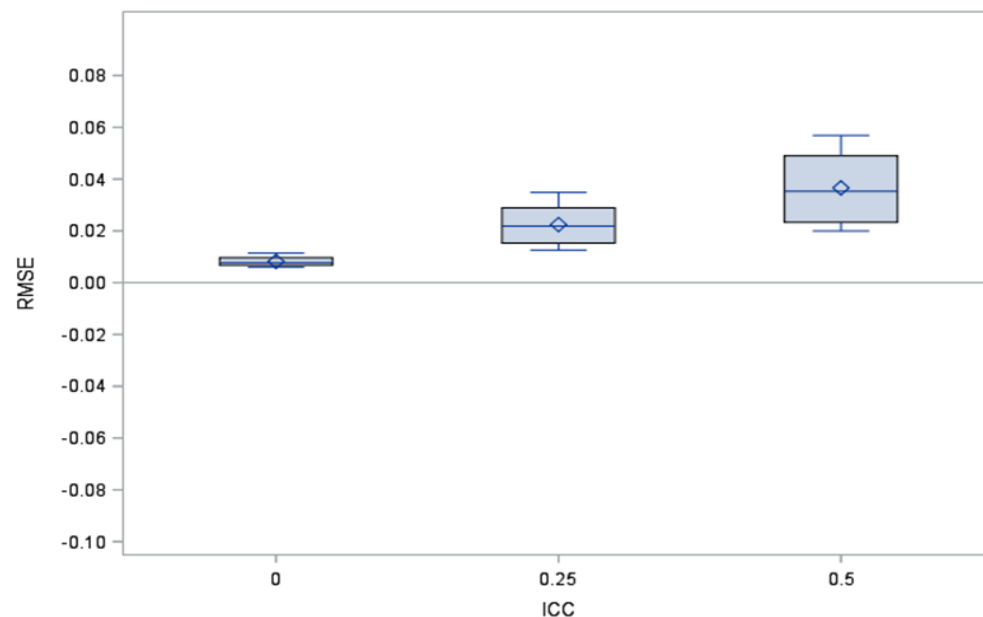


Figure 7. Distributions of RMSE Estimates for LW by ICC in MCAR

The 2nd most influencing factor on the bias estimates yielded by HM was the interaction effect between percent missing and parameter factors ($\eta^2 = 21\%$). Figure 8 describes the interaction effect of percent missing and parameter on bias estimates yielded by HM: bias associated with X7 and X8 increased as

the percentage of missing increased. This interaction also influenced the bias yielded by RM in similar manner but with a lesser degree ($\eta^2 = 18\%$).

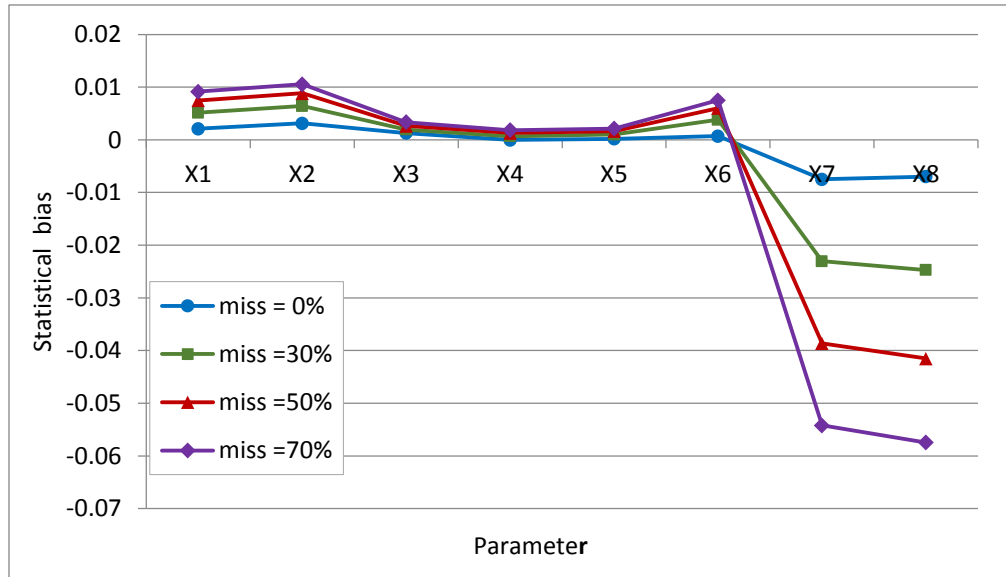


Figure 8. Mean Estimated Bias for HM by Percent Missing and Parameter in MCAR

Regarding confidence interval width estimates, HM produced the narrowest width, and RM produced the widest width (Figure 9). As for confidence interval coverage, LW produced the highest coverage which were around 95% (similar to those estimated from the complete samples), and the coverages associated with HM and RM were below the nominal level by a considerable amount, and they varied substantially across simulation conditions. Also compared to the coverage produced by HM, those produced by RM appeared to vary less. These features can be seen in Figure 10 below.

ICC and density respectively were the most and 2nd most influencing factors on CI width produced by HM and LW (HM: $\eta^2 = 61\%$ for ICC and $\eta^2 = 19\%$ for density; LW: $\eta^2 = 65\%$ for ICC and $\eta^2 = 20\%$ for density); and the interaction of ICC and parameter was the most influencing factor on CI width produced by RM, which produced the widest CI width in MCAR (Figure 9). Figure 11 describes the distribution of CI width for RM by parameter and by ICC levels in MCAR data: except for X7 and X8 which were the two regressors with missing data, wider CI-width estimates were associated with larger ICC data.

As for the most influencing factors on CI coverage estimates produced by the three MDTs, parameter estimates were the factor accounting for most of the variability of CI coverage yielded by RM ($\eta^2 = 74\%$) and by HM ($\eta^2 = 45\%$); other factors modestly influencing the IC coverage yielded by HM were ICC ($\eta^2 = 13\%$), the interaction of percent miss and parameter ($\eta^2 = 11\%$), and percent of missing data ($\eta^2 = 10\%$).

In observing how parameter affected CI coverage yielded by HM and by RM, Figures 12 and 13, respectively, describe the distribution of CI coverage estimated by HM and by RM by parameter: in both cases, coverage estimates associated with X7 and X8 were substantially lower than those associated with other regressors, but the influence of parameters on HM's coverage was significantly stronger than its influence on RM's coverage (e.g., coverage associated with X7 and X8 in HM varied from 0% to 95%).

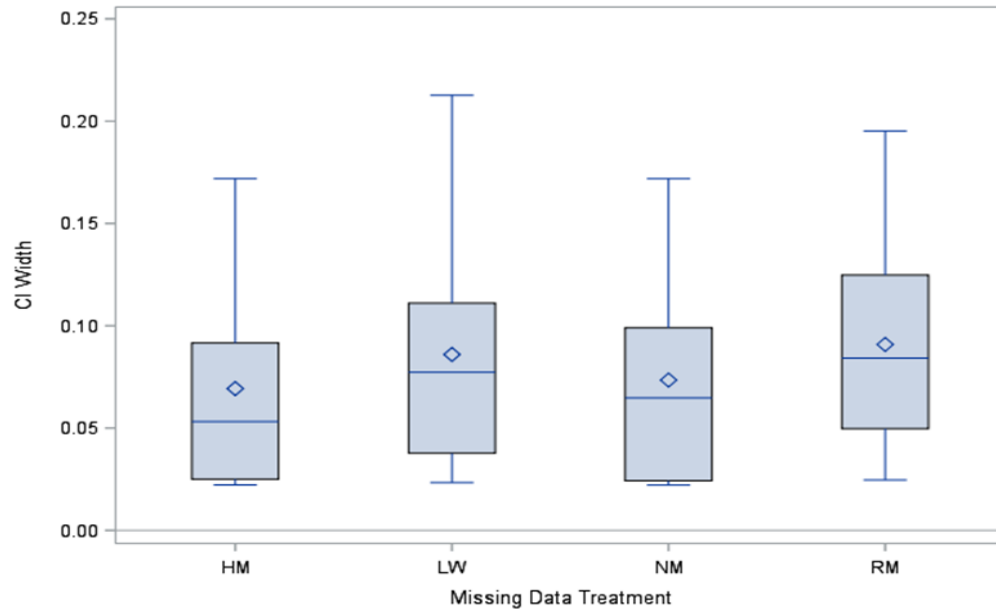


Figure 9. Distributions of Confidence Interval Width by Missing Data Treatment in MCAR

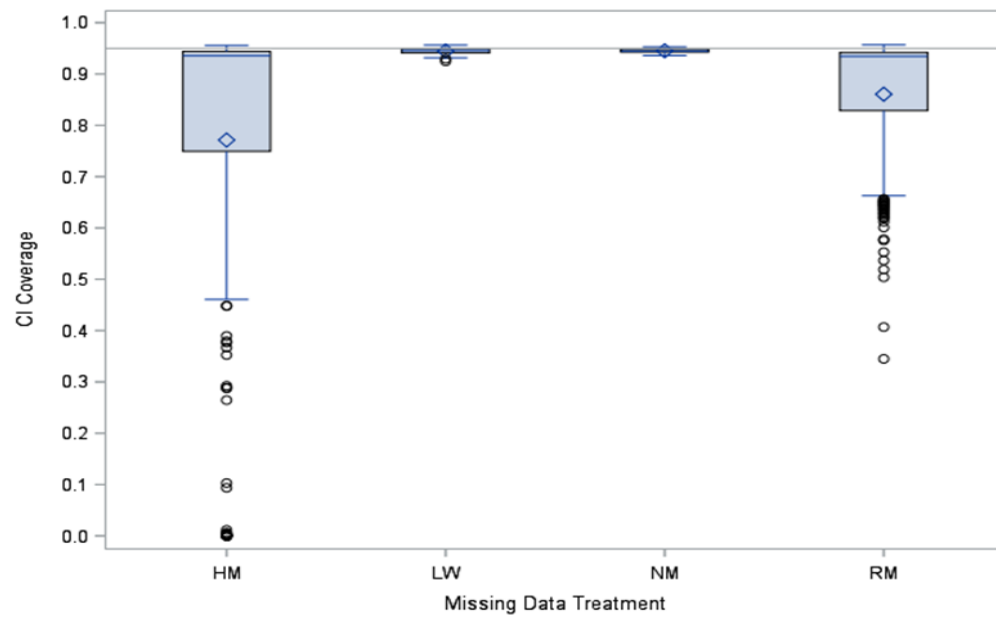


Figure 10. Distributions of Coverage Probability Estimates by Missing Data Treatment in MCAR

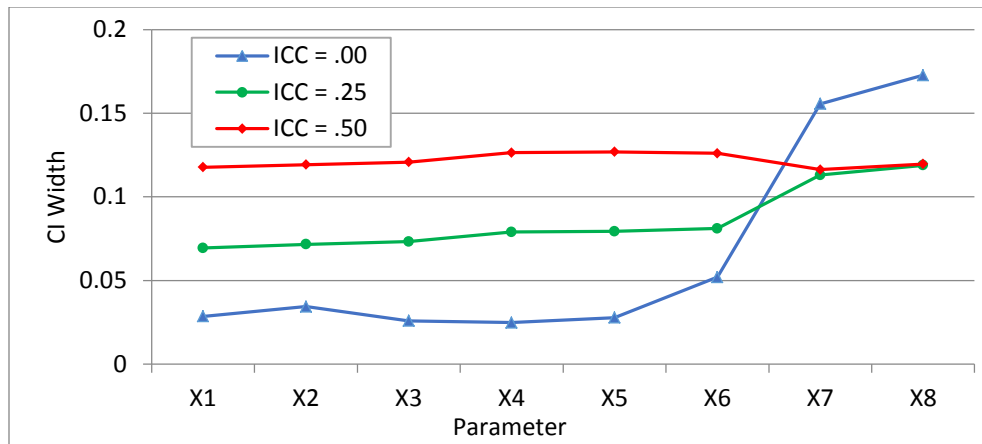


Figure 11. Mean Estimated CI Width for RM by ICC and Parameter in MCAR

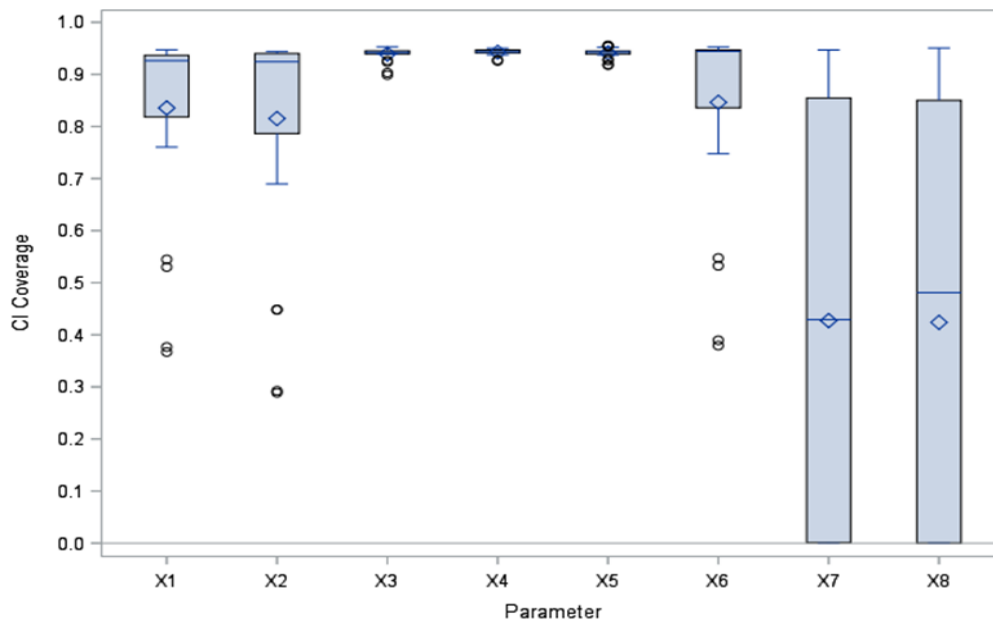


Figure 12. Distributions of CI Coverage Estimates for HM by Parameter in MCAR

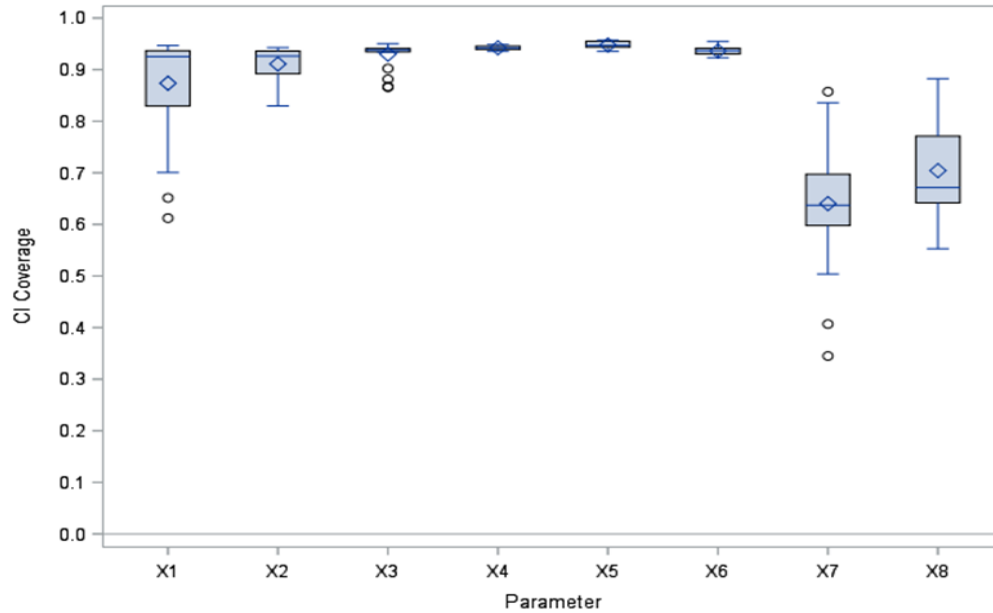


Figure 13. Distributions of CI Coverage Estimates for RM by Parameter in MCAR

MAR

In MAR data, LW's bias varied slightly more than it did in MCAR data, but the degree of variation was clearly less than those obtained from RM and HM. Also, there was no well defined difference in bias yielded by RM and HM regarding variation and degree of bias (though RM appeared to produce somewhat more outliers than HM). Figure 14 describes the distribution of bias estimates by MDTs and bias obtained from the complete samples (NM). Figure 15 shows the distribution of RMSE estimates yielded by the three MDTs: RMSE estimated by RM were averagely larger and varied more compared to those estimated by LW and HM as seen in MCAR data; but in MAR, LW's RMSE was just slightly smaller than HM's RMSE while this difference was more obvious in MCAR (Figure 2).

As for the major factors influencing the bias and RMSE produced in MAR, parameter was also the most influencing factor on bias estimates produced by HM ($\eta^2=64\%$) and by RM ($\eta^2=71\%$); and the interaction between ICC and parameter was also the most influencing factor on bias estimates produced by LW ($\eta^2=31\%$) as previously seen in MCAR. Figures 16 and 17 describe the distribution of bias by parameter for HM and for RM respectively: as seen in MCAR, in both conditions, bias associated with X7 and X8 was negative and larger in magnitude and variability than those associated with other regressors of which bias were positive.

Also as found in MCAR, the second most influencing factor on the bias estimates produced by HM was the interaction between percent miss and parameter ($\eta^2=22\%$). Figure 18 describes the mean bias estimates for HM by missing levels and parameter: bias associated with X7 and X8 increased downward as percent of missing increased. Also as seen in MCAR data, parameter was the most influencing factor on RMSE estimates produced by RM ($\eta^2=64\%$) and by HM ($\eta^2=24\%$), and ICC was the most influencing factor on RMSE estimates produced by LW ($\eta^2=57\%$). Figure 19 shows the distribution of RM's RMSE by parameter: RMSE associated with X7 and X8 were larger than those associated with other regressors. Figure 20 describes the distributions of LW's RMSE estimates by ICC: larger and more varied RMSE associated with larger ICC.

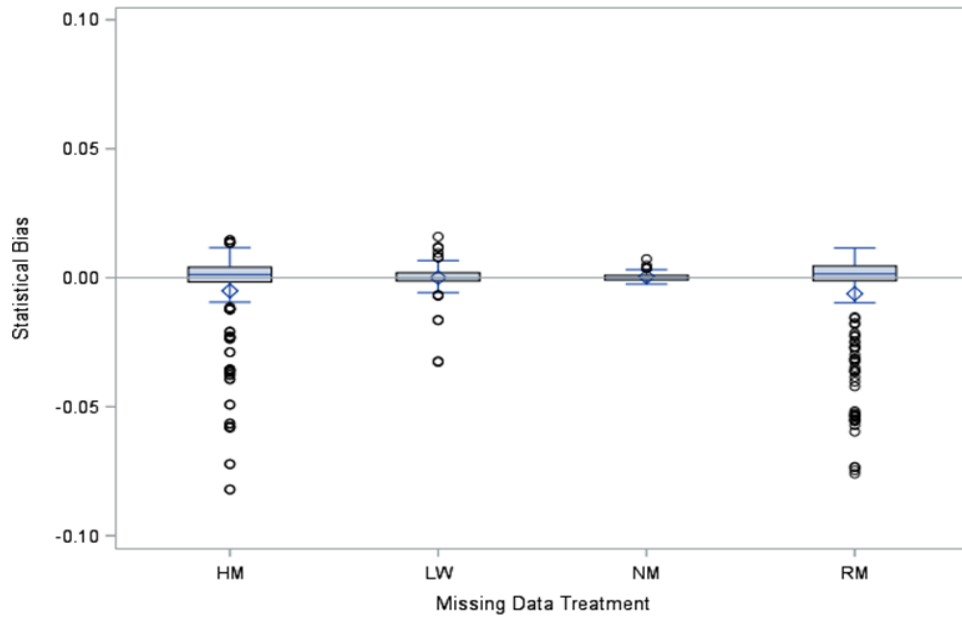


Figure 14. Distributions of Bias by Missing Data Treatment in MAR

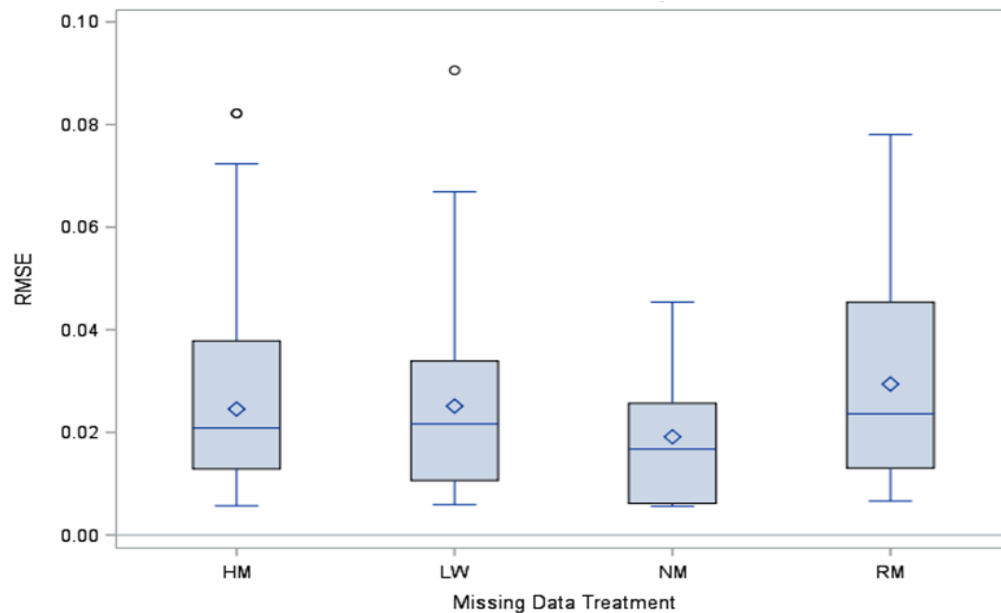


Figure 15. Distributions of RMSE by Missing Data Treatment in MAR

Observations of CI-width estimates in MAR data showed that CI width produced by LW was virtually larger and varied more than those produced by RM and HM; and CI width produced by RM was larger than those produced by HM. These features can be viewed in Figure 21. Though LW produced larger CI width than HM and RM did, it produced very high CI coverage compared to RM and HM. Figure 22 shows that except for some outliers, CI coverage produced by LW was as high as that obtained from the complete samples (NM); whereas CI coverage yielded by RM and HM varied substantially with some CI coverage for HM were as low as 0%.

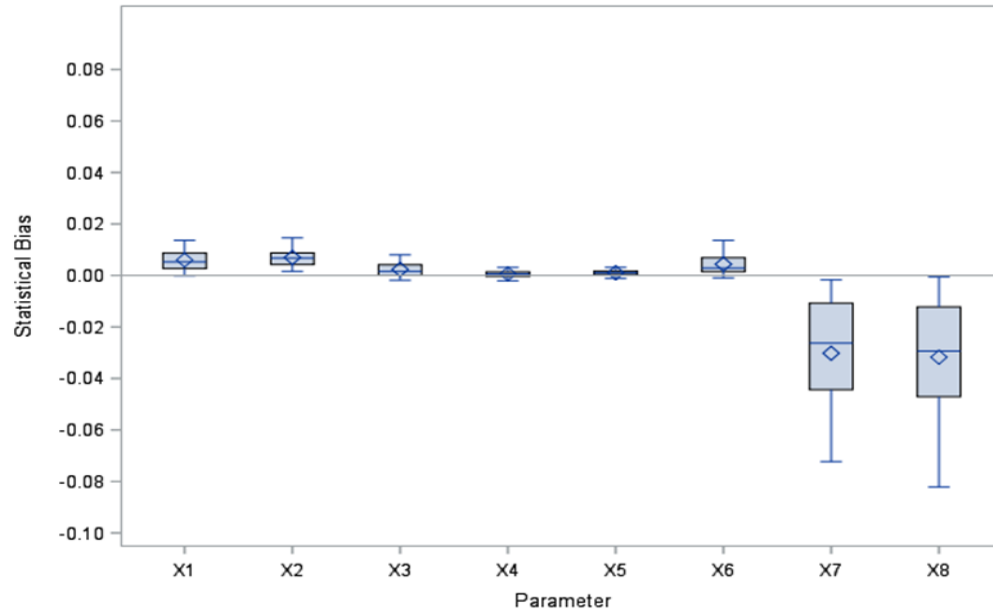


Figure 16. Distributions of Bias Estimates for HM by Parameter in MAR

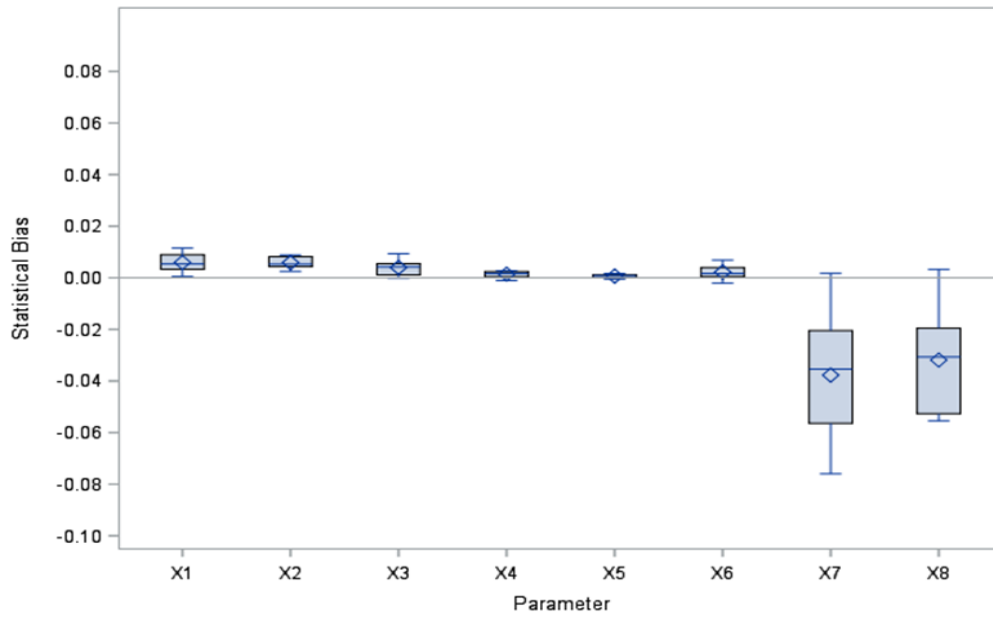


Figure 17. Distributions of Bias Estimates for RM by Parameter in MAR

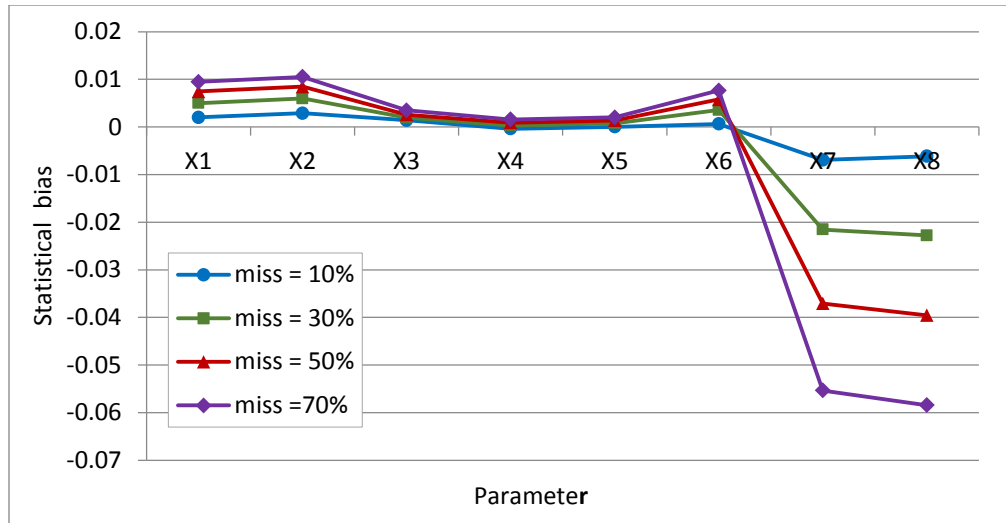


Figure 18. Mean Estimated Bias for HM by Percent Missing Data and Parameter in MAR

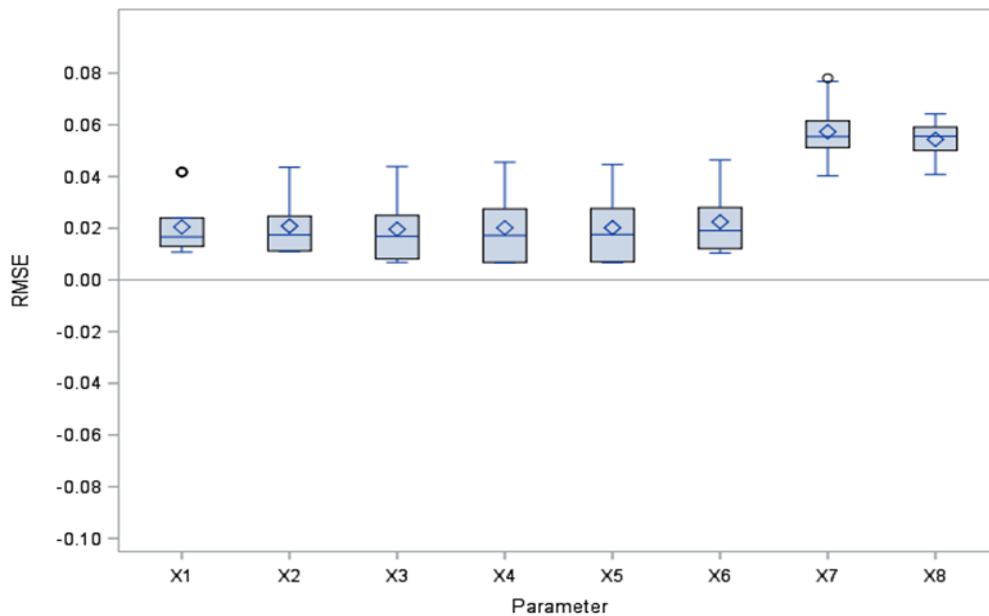


Figure 19. Distributions of RMSE Estimates for RM by Parameter in MAR

Considering the major influencing factors on CI width yielded by the three MDTs in MAR data, ICC was the most influencing factor on CI width produced by HM ($\eta^2=63\%$), by LW ($\eta^2=62\%$), and by RM ($\eta^2=26\%$). In MAR, compared to HM and RM, LW produced the widest and most varied CI width which appeared to be larger in larger ICC data; also so when ICC was not zero, LW's CI width was also higher in high density than in lower density (Figure 23).

Regarding CI coverage, as seen in Figure 22, HM produced the poorest coverage; the most influencing factor on the CI coverage produced by HM was parameter ($\eta^2=40\%$); coverage for HM was also modestly affected by ICC ($\eta^2=15\%$), the interaction between ICC and parameter ($\eta^2=11\%$), the interaction between percent miss and parameter ($\eta^2=11\%$), and percent miss ($\eta^2=10\%$). Parameter was also the most influencing factor on CI coverage produced by RM ($\eta^2=73\%$); and the most influencing factor on CI

coverage produced by LW was the interaction between ICC and parameter ($\eta^2=35\%$). Figures 24 describe the distribution of CI coverage by parameter for HM: parameter estimates associated with X7 and X8 were the driving influence on CI coverage produced by HM as seen in MCAR.

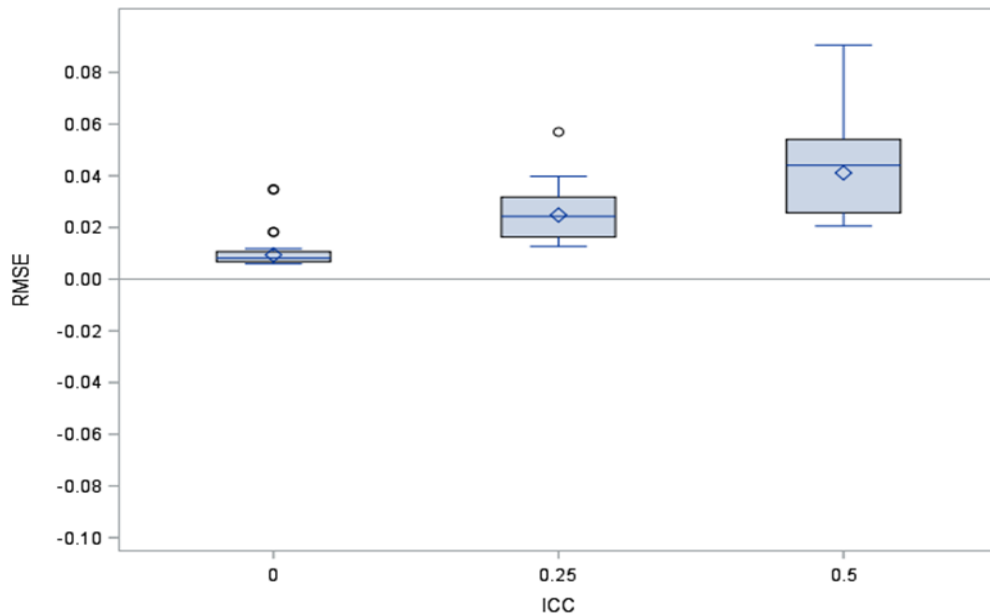


Figure 20. Distributions of RMSE Estimates for LW by ICC in MAR

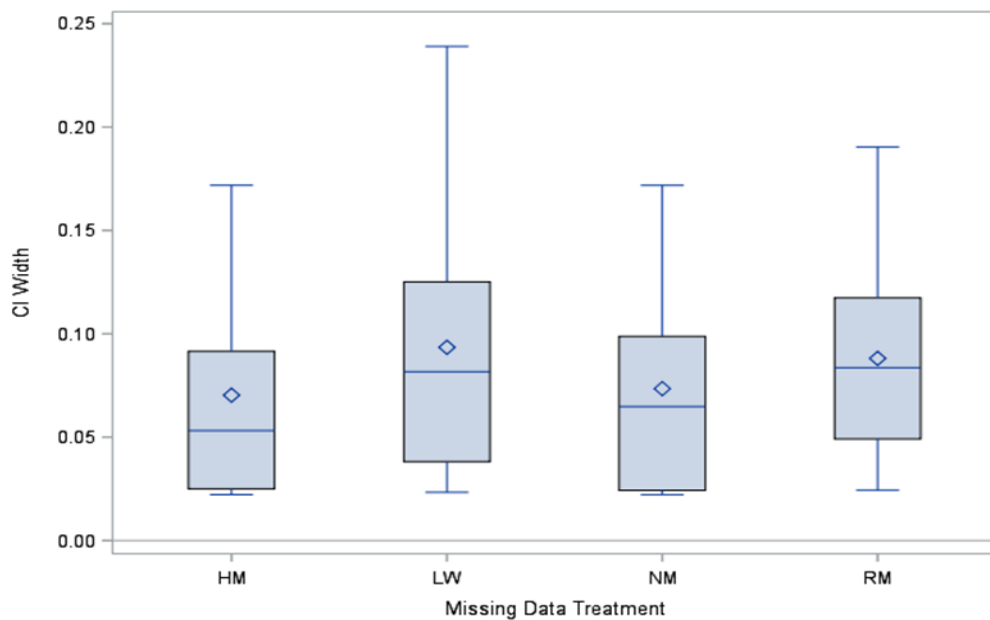


Figure 21. Distributions of CI-Width Estimates by Missing Data Treatment in MAR

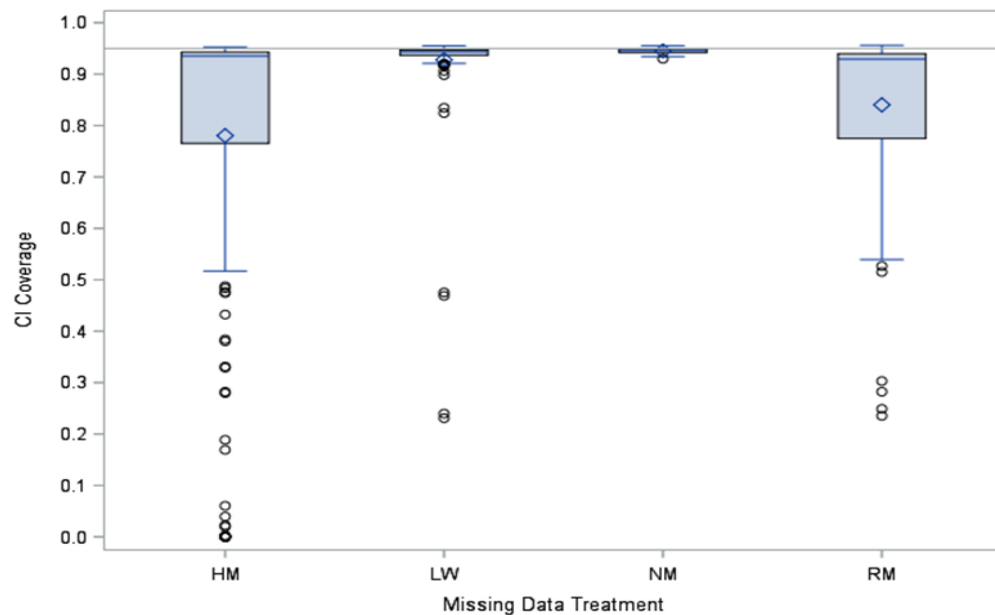


Figure 22. Distributions of CI-Coverage Estimates by Missing Data Treatment in MAR

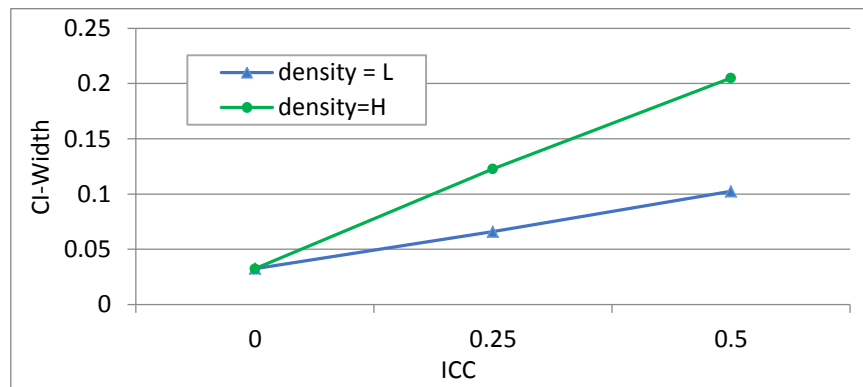


Figure 23. Mean Estimated CI Width for LW by Population Density and ICC in MAR

Overall, in MAR data the patterns of the influence of the simulation factors on the evaluation measures yielded by the three 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 negligible influence ($\eta^2=3\%$) on coverage yielded by LW whereas this effect was moderate ($\eta^2=31\%$) in MCAR; (2) parameter had a modest effect ($\eta^2=18\%$) on this measure in MAR but it had a negligible effect ($\eta^2=0.9\%$) in MCAR; and (3) density had almost no effect ($\eta^2=0.2\%$) on this measure in MAR, but it had a modest effect ($\eta^2=21\%$) in MCAR.

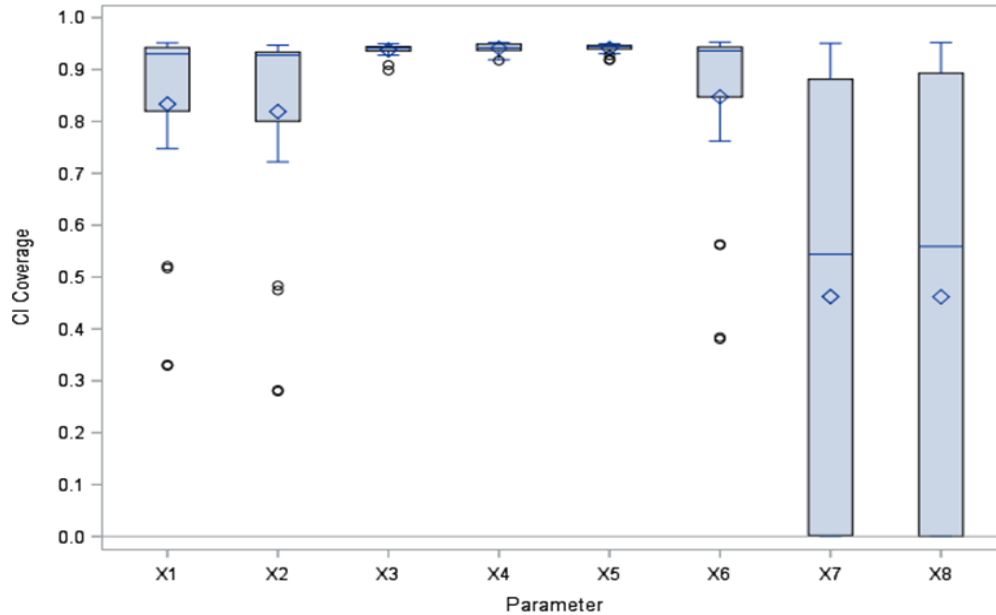


Figure 24. Distributions of CI-Coverage Estimates for HM by Parameter in MAR

MNAR

In MNAR, as previously seen in MCAR and MAR data, bias yielded by LW varied less than those yielded by HM and RM, and RM's bias appeared to vary slightly more than HM's bias. Figure 25 describes the distribution of bias estimated by the three MDTs and bias obtained from the complete samples (NM).

Also, as previously seen in MCAR and MAR, of the three MDTs, RMSE estimated by RM appeared to be the largest and varied most, and RMSE estimated by LW was smallest and varied least (Figure 26). In addition, as seen in MAR, CI width estimated by LW appeared to be wider than those estimated by the two multiple imputation methods, RM and HM; and CI width estimates produced by RM was slightly wider than those produced by HM (Figure 27).

Regarding CI coverage, as previously seen in MCAR and MAR, coverage estimates produced by LW in MNAR were more consistent and close to the nominal level of 95% compared to those produced by HM and RM which varied substantially; and being different from those observed in MCAR and MAR, RM's coverage varied much more than HM's coverage (Figure 28).

With regard to the major influencing factors on these estimated measures in MNAR, parameter estimates were the most influencing factor on bias produced by HM ($\eta^2=60\%$) and by RM ($\eta^2=73\%$) whereas the interaction between ICC and parameter estimate was the most influencing factor on bias produced by LW ($\eta^2=33\%$). Figures 29, 30 show the distribution of bias by parameter for HM and RM respectively. In both conditions, similarly to the pattern of influence previously seen in MCAR and MAR data, parameter estimates associated with X7 and X8 were the driving influence on the bias estimates in MNAR, but the bias estimated in MNAR varied more than those estimated in MCAR and MAR data.

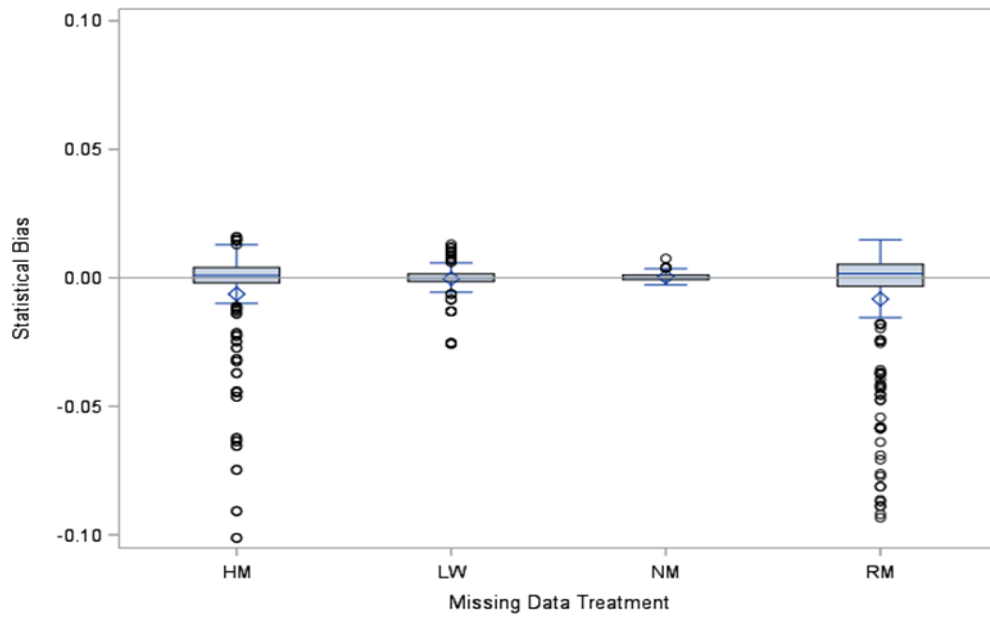


Figure 25. Distributions of Bias by Missing Data Treatment in MNAR

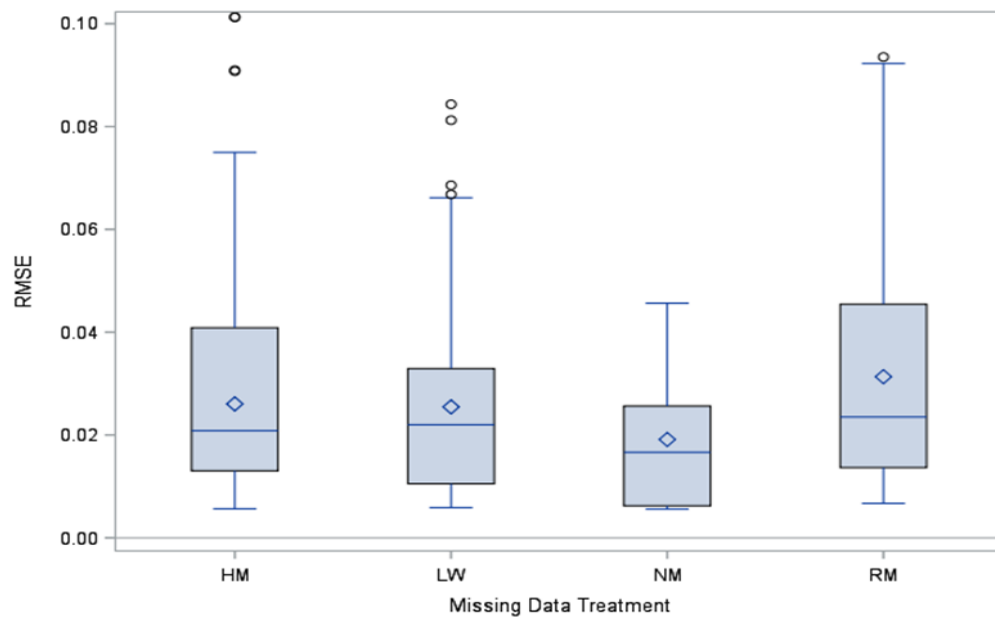


Figure 26. Distributions of RMSE by Missing Data Treatment in MNAR

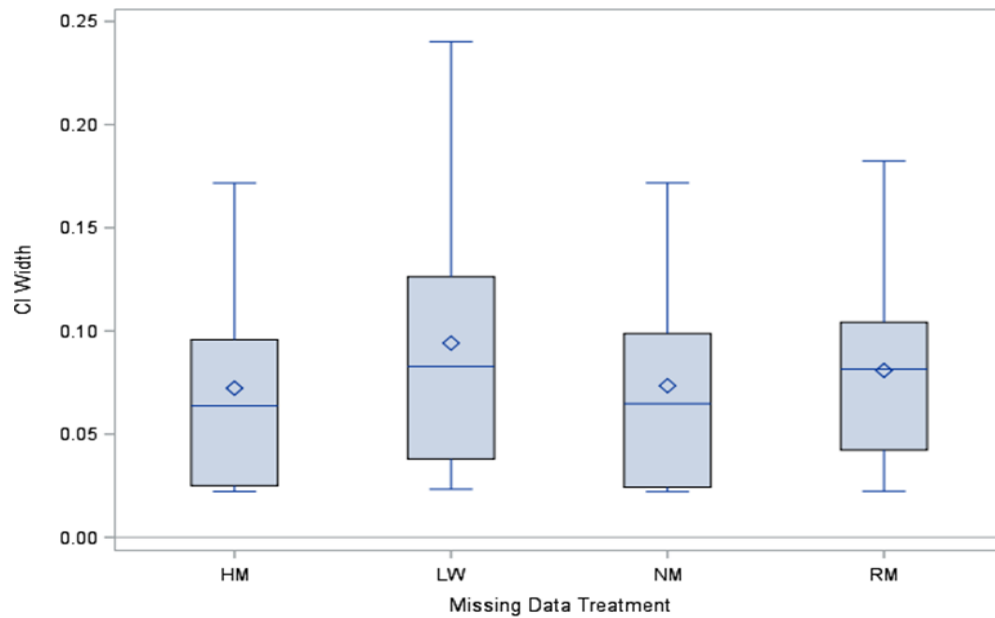


Figure 27. Distributions of CI-Width Estimates by Missing Data Treatment in MNAR

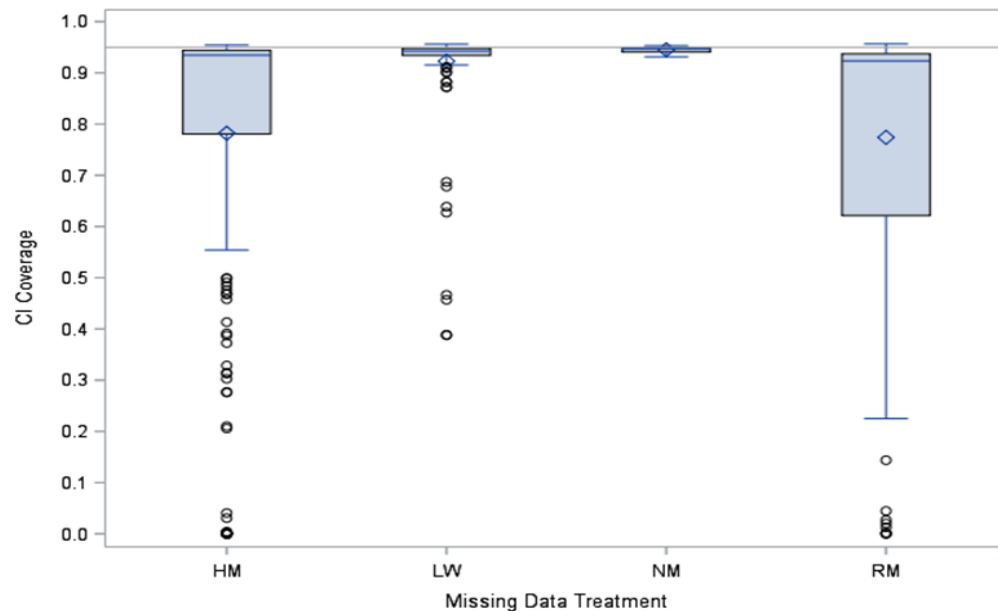


Figure 28. Distributions of CI Coverage by Missing Data Treatment in MNAR

Parameter factor was also the most influencing factor on RMSE estimates produced by HM ($\eta^2=30\%$) and by RM ($\eta^2=67\%$) in MNAR. In both cases, the RMSE estimates associated with X7 and X8 were larger compared to those associated with other regressors. Figure 31 describes the distribution of RMSE for RM by parameter estimates in MNAR.

As mentioned previously, LW produced the widest CI-width estimates compared to HM and RM in MNAR. The top three influencing factors on this measure were ICC ($\eta^2=62\%$), density ($\eta^2=18\%$), and the interaction of these two factors ($\eta^2=11\%$). The pattern of the influence was similar to what seen in MAR

data: CI width was larger for larger ICC and when ICC was not zero, CI width from high density data was higher than CI width from low density data (Figure 32).

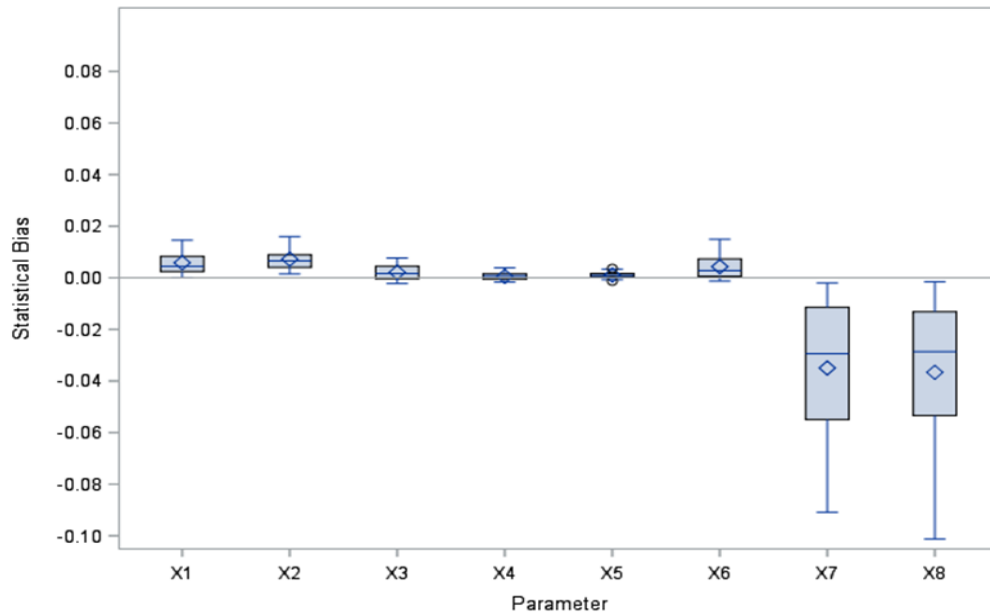


Figure 29. Distributions of Bias Estimates for HM by Parameter in MNAR

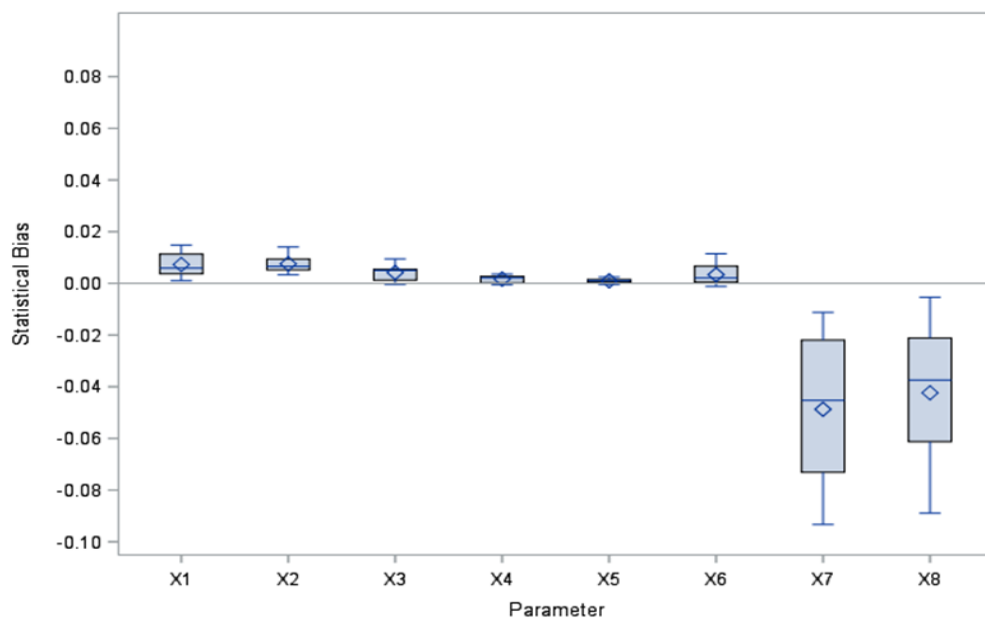


Figure 30. Distributions of Bias Estimates for RM by Parameter in MNAR

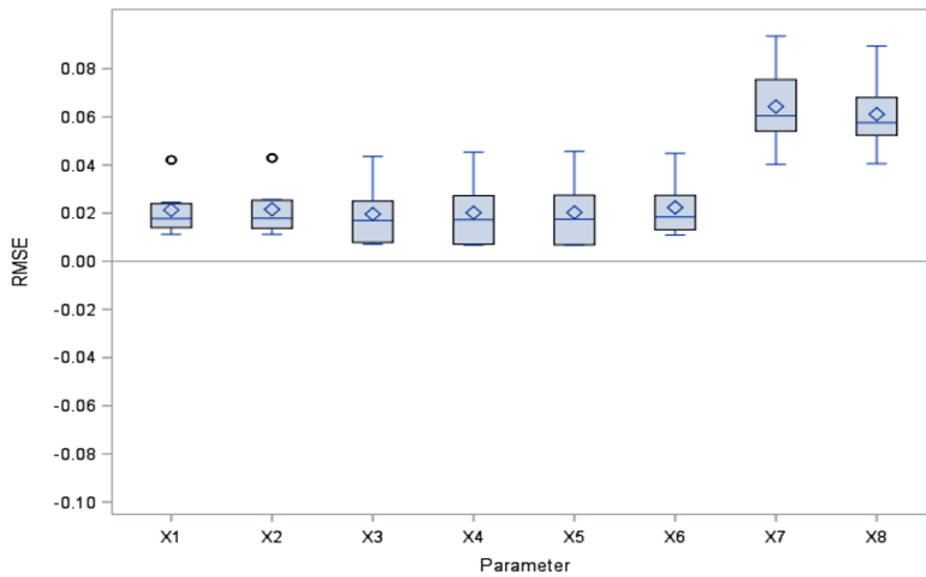


Figure 31. Distributions of RMSE Estimates for RM by Parameter in MNAR

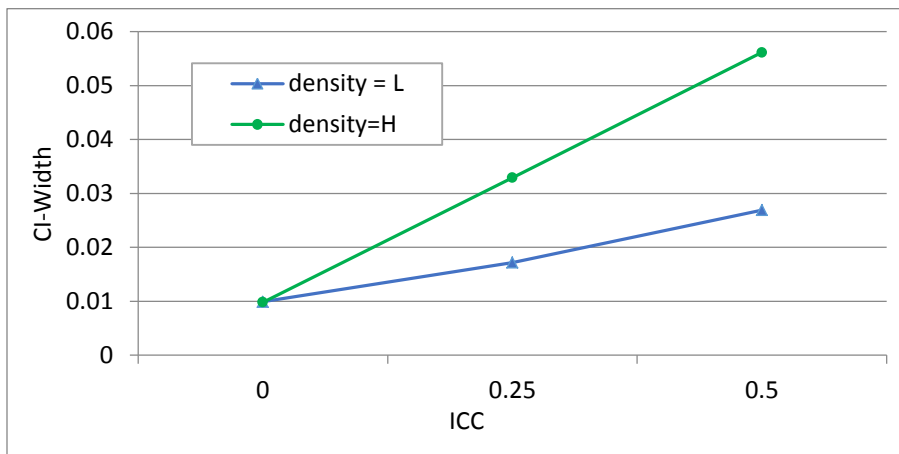


Figure 32. Mean Estimated CI Width Estimates for LW by Population Density and ICC in MNAR

Concerning the factors influencing CI coverage produced by the three MDTs, parameter was the most influencing factor on CI coverage produced by HM ($\eta^2=36\%$) and by RM ($\eta^2=61\%$). The pattern of influence was similar to what previously seen in MCAR and MNAR data; i.e., the parameter estimates associated with X7 and X8 were the driving influence especially for the influence of parameter on CI coverage yielded by HM.

In summary, it was observed that (1) in each types of missingness (MCAR, MAR, and MNAR), the pattern of what simulation factor had a major influence on a measure (bias, RMSE, CI width, or CI coverage) produced by a certain MDT was consistent, but the degrees the 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 in

MCAR and approximately 62% of variation in CI width in MAR and MNAR. (2) In the three types of missingness, the most-influencing factor on the four evaluation measures was either parameter, ICC, or interaction of 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 interaction of ICC and parameter was the most-influencing factor on measures estimated by LW. (3) Percent of missing data, by itself, had negligible effects on those performance measures yielded by the three studied MDTs. (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.

CONCLUSION

In all three MCAR, MAR, and MNAR conditions, the performance indicators obtained from the three missing data treatment methods showed clear evidence that listwise deletion outperformed the two multiple imputation methods, hot-deck-based and regression-based multiple imputation, in regarding to bias, RMSE, and coverage probability. Specifically, in all three missing data mechanisms, listwise deletion was found to be less biased and to produce smaller RMSE and higher and less varied coverage probability compared to each of the two multiple imputation methods. The findings contradict the literature that suggests that listwise deletion methods for complex sample data are inappropriate; so for complex sample survey data, if you are a “listwise deleter,” no need to feel guilty unless CI width is your most concern and your data have high density and large ICC.

If multiple imputation methods should be considered, it should be noted that RM produced larger and more varied RMSE than HM in all three data types; and in MNAR, coverage probability for RM varied more and CI width for RM was also slightly wider than those for HM. With respect to computing resources, depending on the number of imputations and the percent of missing data, HM method (i.e., PROC SURVEYIMPUTE and the process of restructuring the imputed data) utilized a substantial amount of work space; whereas RM (i.e., PROC MI) utilized a very large amount memory.

REFERENCES

- Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215 – 238.
- Berglund, P., & Heeringa, S. G. (2014). *Multiple imputation of missing data using SAS*. SAS Institute.
- Collins, L. M., Schafer, J. L., & Kam, C. M (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330-351.
- Graham, J. W., Cumsille, P. E. & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Research Methods in Psychology* (pp. 87-114). New York: Wiley.
- Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schanabel, & J. Baumert (Eds.). *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201-218). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Graubard, B. I., & Korn, E. L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology*, 144, 102-106.
- Little, R. J., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: John Wiley & Sons.
- National Center for Education Statistics. n.d. “Surveys and Programs”. Retrieved January 15, 2007, from <http://nces.ed.gov/surveys/nels88/>
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537-560.

- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469), 322-331.
- SAS Institute. (2004). *SAS/STAT® 9.4 User's Guide*. Cary, NC: SAS Institute Inc.
- Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4), 545-571.
- Schafer, J. L. & Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.
- 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.
- Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, 56, 645-646.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Anh Kellermann
University of South Florida
4202 East Fowler Ave., EDU 105
Tampa, FL 33620
Email: napham@mail.usf.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.