Missing Data and Complex Samples: The Impact of Listwise Deletion vs. Subpopulation Analysis on Statistical Bias and Hypothesis Test Results when Data are MCAR and MAR

Bethany A. Bell[1], Jeffrey D. Kromrey[2], John M. Ferron[2]
[1]Educational Psychology, Research, and Foundations, University of South Carolina, College of Education, Wardlaw 133, Columbia, SC 29208
[2]Educational Measurement and Research, University of South Florida, College of Education, 4202 East Fowler Avenue, EDU 162, Tampa, FL 33620

Abstract

Secondary data analysis of complex sample survey results is common among social scientists. Yet, the degree to which unbiased estimates and accurate inferences can be made from complex samples depends on the care researchers take when analyzing the data, including strategies for the treatment of missing data. Several studies have illustrated that the results of subpopulation analysis may diverge from those obtained through listwise deletion. However, given the paucity of simulation work in this area, it is not clear how frequently discernable discrepancies will arise. This Monte Carlo study focuses on the impact of listwise deletion versus a subpopulation analysis, when the data are MCAR and MAR, in the context of multiple regression analysis of complex sample data. Results are presented in terms of statistical bias in parameter estimates and both confidence interval width and coverage.

Key Words: survey research, complex samples, missing data, variance estimation

1. 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).

1.1 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)

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

where w_{ijk} is the weight assigned to the ith individual in the jth primary sampling unit (PSU) of the kth strata in a study with *s* strata where the kth strata has p_k PSUs and the jth PSU within the kth strata 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)

$$\sum_{k=1}^{s}\sum_{j=1}^{p_k}\sum_{i=1}^{n_{jk}} w_{ijk} = N \tag{2}$$

Sample weights are then applied in the computation of statistics from the sample observations. For example, the sample mean is computed as

$$\bar{X} = \frac{\sum_{k=1}^{s}\sum_{j=1}^{p_k}\sum_{i=1}^{n_{jk}} w_{ijk} X_{ijk}}{\sum_{k=1}^{s}\sum_{j=1}^{p_k}\sum_{i=1}^{n_{jk}} w_{ijk}} \tag{3}$$

where X_{ijk} is the value for the ith individual in the jth PSU of the kth strata, 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.

1.2 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

$$\hat{\sigma}^2_{\bar{X}} = \frac{\sum_{k=1}^{s} \frac{p_k}{p_k - 1} \sum_{j=1}^{p_k} \left[W_{jk}(\bar{X}_{jk} - \bar{X}) - \frac{1}{p_k} \sum_{t=1}^{p_k} W_{tk}(\bar{X}_{tk} - \bar{X}) \right]^2}{\left(\sum_{k=1}^{s} \sum_{j=1}^{p_k} W_{jk} \right)^2} \tag{4}$$

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

1.2.1 Estimation of variances for subpopulations

In many contexts, researchers are interested in obtaining variance estimates for sample statistics of a subpopulation. For example, a researcher may be examining predictors of academic achievement among English Language Learners. The researcher may do this by examination of data available from a complex national sample where schools were the primary sampling unit (PSU) and students were then sampled within each school. Not all sampled students would be in the subgroup of interest (English Language Learners), and thus the analysis is focused on a subgroup of the population that was originally sampled.

One common approach for analyzing a subgroup is to listwise delete all participants that are not part of the subgroup. For our example this approach would lead to deleting all students who were not English Language Learners. The analyses would then be carried out using this reduced data set. An alternative, sometimes referred to as subpopulation analysis, is to make the sample weights zero for all participants outside the subgroup of interest. For our example, the weights of all students that were not English Language Learners would be changed to zero, and then the analysis would then be conducted using the modified, but complete, data set.

If we consider Equation 3 for the sample mean, we see that these two approaches will lead to the same estimate for the sample mean. If we consider Equation 4 for the variance of the sample mean, we see that the two approaches lead to the same result as long as each PSU has multiple members of the subpopulation. There are instances, however, where this will not be the case. In our example, it is easy to imagine a sample of students from a school, which happens to contain no English Language Learners. Under these conditions, the variance estimates from these two approaches tend to differ.

A closer examination of Equation 4 shows that the differences arise from several changes in the numerator. Most obviously, the fractions $\frac{p_k}{p_k - 1}$ and $\frac{1}{p_k}$ when using listwise deletion will be based on the number of PSUs in the kth strata that happen to have multiple members of the subpopulation, but for the subpopulation analysis these fractions will be based on the number of PSUs in the kth strata of the complex sampling plan. Furthermore, it can be seen that with listwise deletion empty PSUs provide no contribution to the sum-of-squared deviations, whereas subpopulation analysis gets a contribution from each PSU from the complex sample. More specifically, all empty PSUs (i.e., those with no members from the subpopulation) add to the sum-of-squared deviations an amount equal to

$$\left[\; 0 \;-\frac{1}{p_k}\sum_{t=1}^{p_k} W_{tk}\,(\bar{X}_{tk}-\bar{X})\right]^2 \tag{5}$$

For simplicity, we have used the sample mean and its variance to illustrate differences between the two approaches for analyzing subgroups. Similar results could also be shown for regression coefficients (Graubard & Korn, 1996).

It has been suggested that when PSUs exist with no members of a subpopulation, the subpopulation analysis is the more appropriate analysis because it is based on the full complex sampling design (Chantala, 2006; Graubard & Korn, 1996; West, Berglund, & Heeringa, 2007). Although a PSU may contain no members of the subpopulation in the realized sample, the subgroup theoretically could be represented in a different sample from that PSU. Thus, the subgroup sample size is conceptually a random variable and all PSUs should be represented in the variance estimates.

Subpopulation analysis has also been suggested as an appropriate method for handling missing data (Chantala, 2006), although there is some question about how to best accommodate such analyses using current software packages (West, Berglund, & Heeringa, 2007). In the context of missing data, those with complete data can be considered the subpopulation of interest. Those with missing data, which are assumed to be missing at random, could then have their sample weights turned to zero. Doing so allows the full complex sampling design to be taken into account when conducting analyses on a subset of the sample that has complete data.

Several studies have shown for a specific analysis of a particular data set that subpopulation analysis leads to results that diverge from those obtained through listwise deletion (Chantala, 2006; Graubard & Korn, 1996; West, Berglund, & Heeringa, 2007; West, Berglund, & Heeringa, 2008). It is not clear how frequently discernable discrepancies will arise, or when there are differences that the theoretical advantages of subpopulation analysis are being realized. Consequently, it is difficult for researchers to know how much caution to use when reading literature based on the listwise deletion approach to subgroup analysis, or how critical it is to obtain access to software that accommodates subpopulation analyses. The purpose of this study was to investigate the impact of using the listwise deletion method versus a subpopulation analysis, when the data are missing completely at random (MCAR) and missing at random (MAR), in the context of multiple regression analysis of complex sample data (Little & Rubin, 1987; Rubin, 1987; Collins, Schafer, & Kam, 2001; Schafer & Graham, 2002). Data that are MCAR represent missingness that is related to neither the variable presenting missing data nor other variables in the analysis. In contrast, MAR represents missing data that are unrelated to value of the variable presenting missingness, but are related to other measured variables.

2. Method

For this Monte Carlo study, complex samples were generated from multivariate populations and each sample was analyzed using both listwise deletion and subpopulation approaches. 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 was simulated, controlling the relative variance between and within PSUs to produce target values of intraclass correlation.

The Monte Carlo study included four factors in the design. 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 ($\rho_I =$ .00, .05, and .10) by controlling the ratio of the between PSU variance to the within PSU variance. Finally, 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, 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. Two missing data mechanisms were simulated to produce data that were either missing completely at random (MCAR) or missing at random (MAR; 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.

Table 1: Correlation Matrix Used as Template for Data Simulation

	Y	X1	X2	X3	X3	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					
X3	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

For both the MCAR and MAR design factors, the data generation was conducted using SAS/IML version 9.1, and each sample for both missing data factors was analyzed separately, using the listwise deletion approach in SAS (SAS Institute, 2004) and the subpopulation strategy in SUDAAN SAS-Callable (Research Triangle Institute, 2004). The available packaged procedures for complex sample survey analysis (i.e., PROC SURVEYREG in SAS and PROC REGRESS in SUDAAN) used the Taylor Series approximation to estimate the sampling variances (Kiecolt & Nathan, 1985). Conditions for the study were run under Windows XP. 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, 1,000 samples were generated. The use of 1,000 estimates provides adequate precision for the investigation of the sampling behavior of point and interval estimates of the regression coefficients. For example, 1,000 samples provide a maximum 95% confidence interval width around an observed proportion that is ± .03 (Robey & Barcikowski, 1992).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).

3. Results

Because the individual design factors included in the current study explained minimal amounts of variability in the outcomes of interest, results are presented across conditions for both MCAR and MAR data structures. Across MCAR conditions, SAS (i.e., listwise deletion) and SUDAAN (i.e., subpopulation analysis) point estimates were identical, very low levels of statistical bias were evident (Figure 1), and differences between standard errors were trivial (Figure 2). As shown in Figure 2, a standard error ratio equal to one represented identical standard errors from the SAS and SUDAAN analyses; a standard error ratio greater than one occurred when a SAS standard error was larger than a SUDAAN standard error; and a standard error ratio less than one occurred when a SAS standard error was less than a SUDAAN standard error.

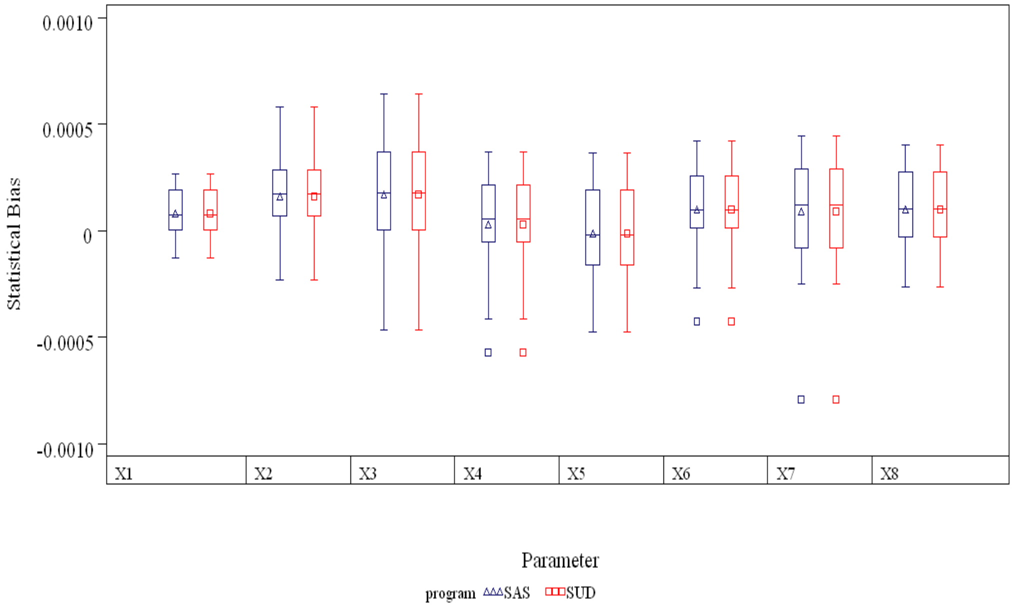

Figure 1: Box-and-Whisker Plots for Statistical Bias for SAS and SUDAAN (MCAR)

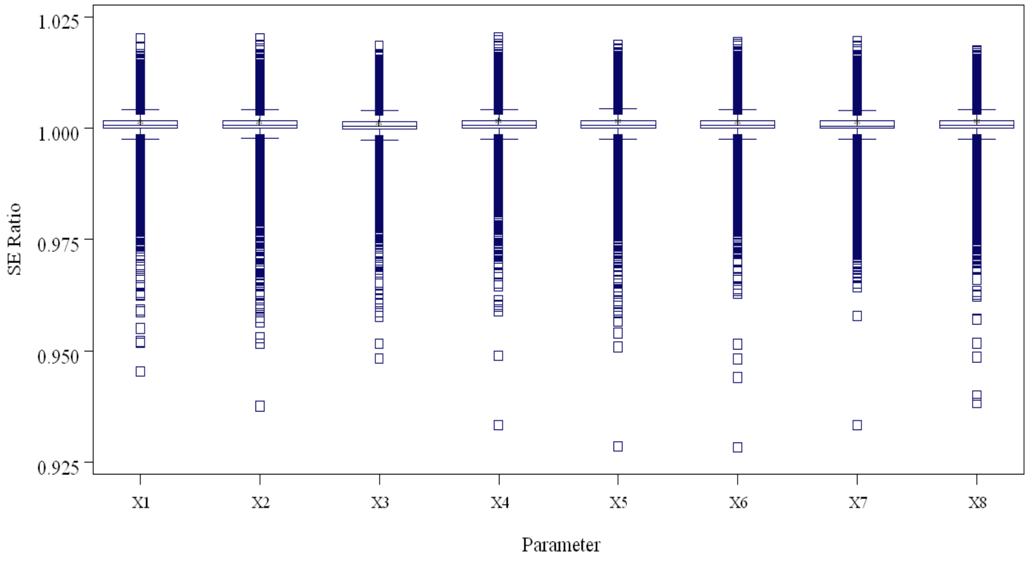

Figure 2: Box-and-Whisker Plots for Standard Error Ratios [SAS/SUDAAN] (MCAR)

Similarly, as shown in Figures 3 and 4, although confidence interval coverage and width varied across the eight parameters included in the model, differences between SAS and SUDAAN were trivial. Thus, when one approach over or undercovered on a particular parameter, so did the other. Also, except for the two parameters that were approximately null in the OLS model, average statistical power for each parameter was nearly perfect ($\bar{x}$ = .997 for both SAS and SUDAAN) with miniscule differences between the two analytic approaches (i.e., minimum value for SAS was .789 and the minimum value for

SUDAAN was .797). Furthermore, even with the two nearly null effects, the average statistical power was still above the desired .80 for both programs.

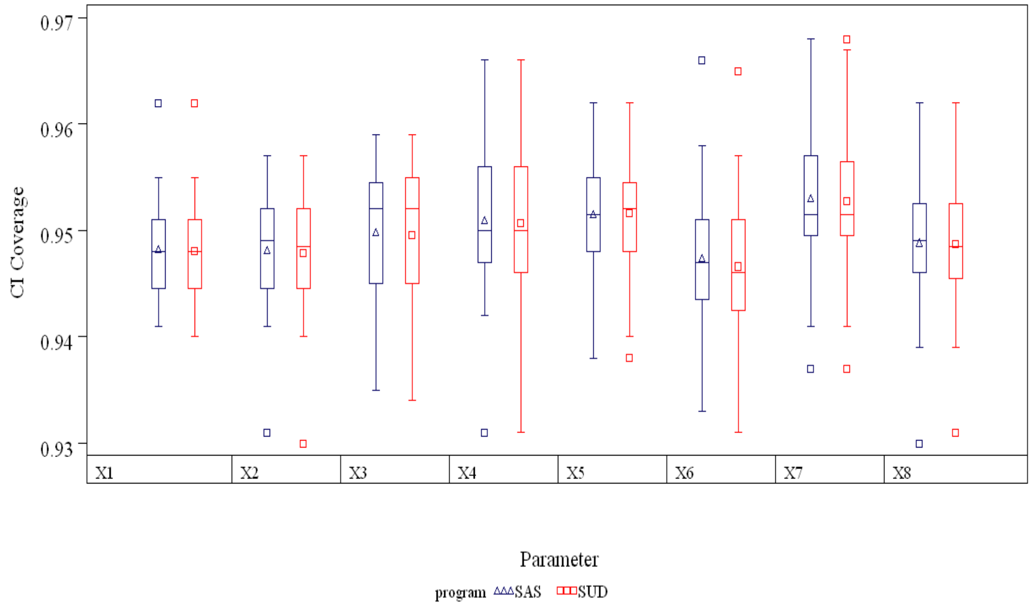

Figure 3: Box-and-Whisker Plots for 95% CI Coverage for SAS and SUDAAN (MCAR)

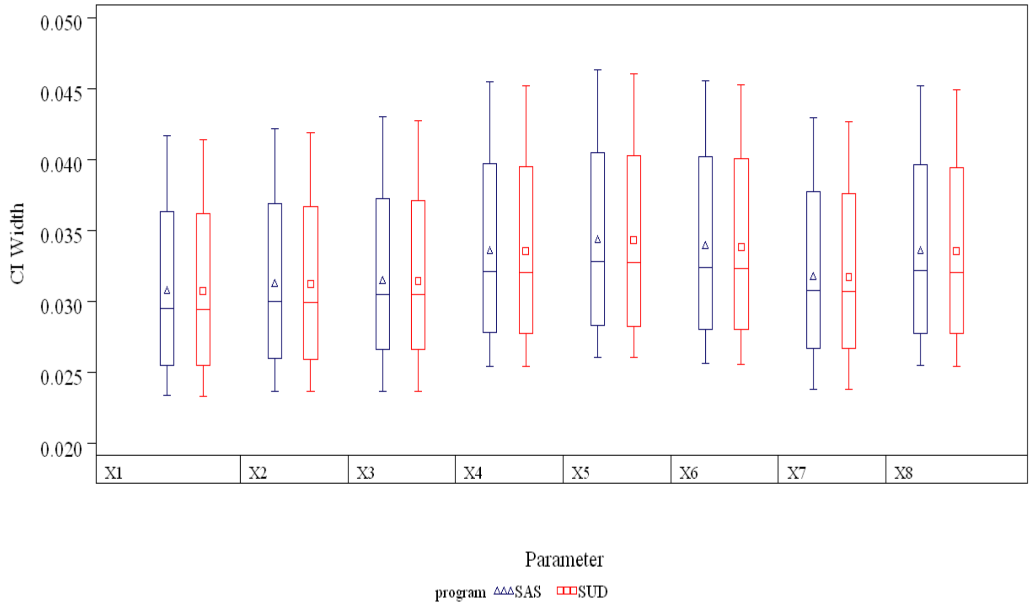

Figure 4: Box-and-Whisker Plots for 95% CI Width for SAS and SUDAAN (MCAR)

Similar results were evident when data were MAR. SAS and SUDAAN point estimates were identical, very little statistical bias was evident (Figure 5), and differences between standard errors were trivial (Figure 6), with the most variability in X1, the variable that was systematically missing.

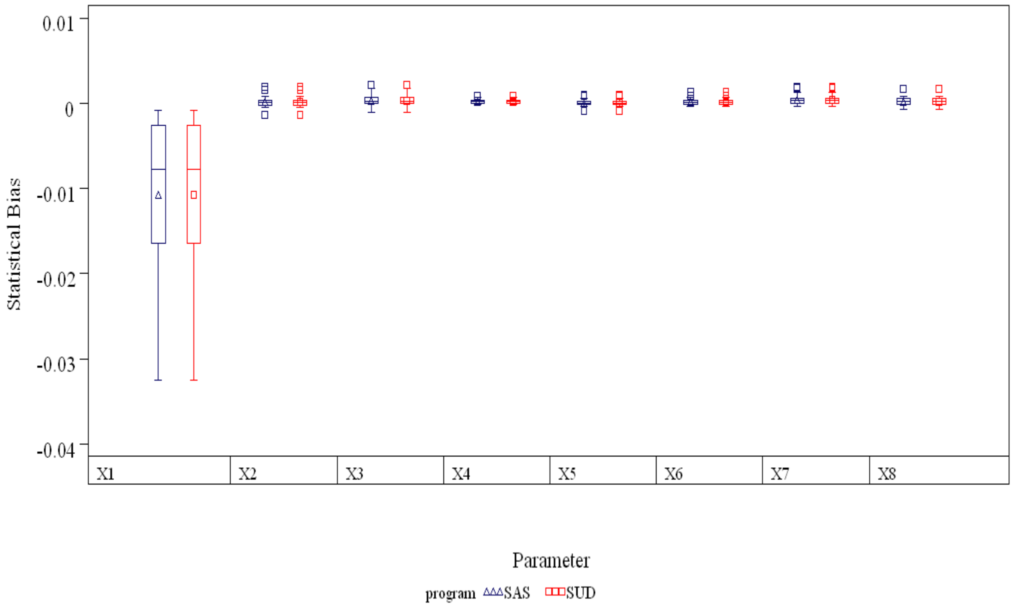

Figure 5: Box-and-Whisker Plots for Statistical Bias for SAS and SUDAAN (MAR)

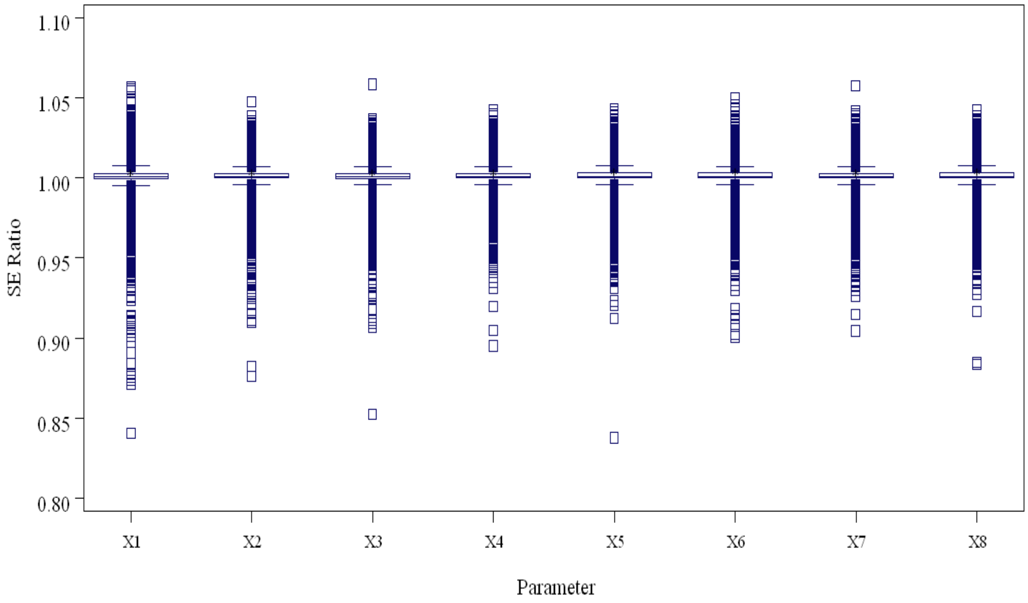

Figure 6: Box-and-Whisker Plots for Standard Error Ratios [SAS/SUDAAN] (MAR)

Moreover, although the nature of the confidence interval coverage and widths varied across parameters and, X1 had noticeably worst coverage and larger widths than the other parameters, differences between SAS and SUDAAN estimates were trivial (Figures 7 and 8). Also, except for the two parameters that were approximately null in the OLS model, average statistical power for each parameter was nearly perfect ($\bar{x}$ = .997 for both SAS and SUDAAN) with miniscule differences between the two analytic approaches (i.e., minimum value for SAS was .806 and the minimum value for SUDAAN was .818). Furthermore, even with the two nearly null effects, the average statistical power was still above the desired .80 for both programs. Moreover, unlike the other outcomes presented above, even though X1 was the variable that was systematically missing, the statistical power for X1 did not vary more than the other parameters.

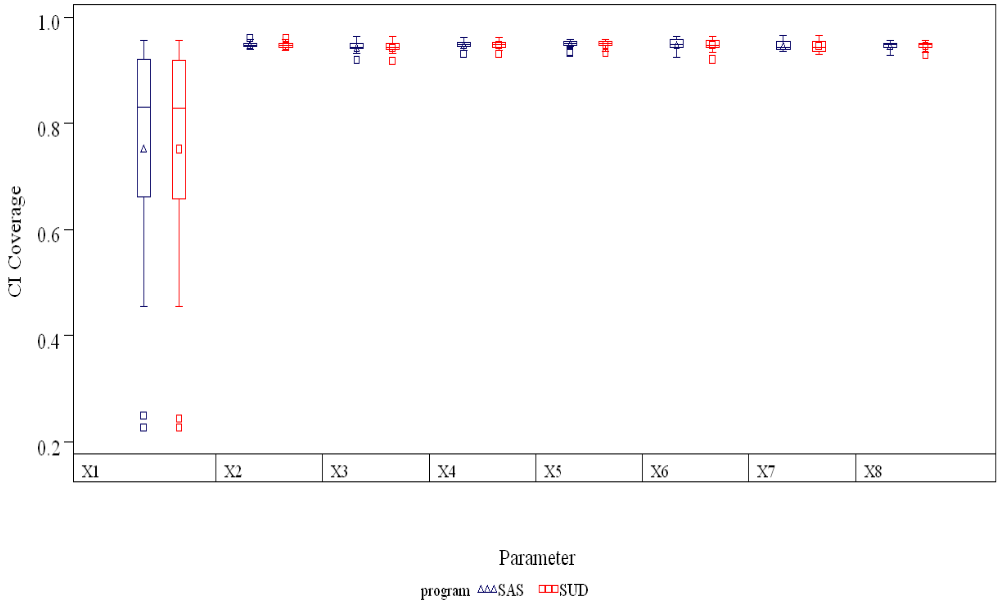

Figure 7: Box-and-Whisker Plots for 95% CI Coverage for SAS and SUDAAN (MAR)

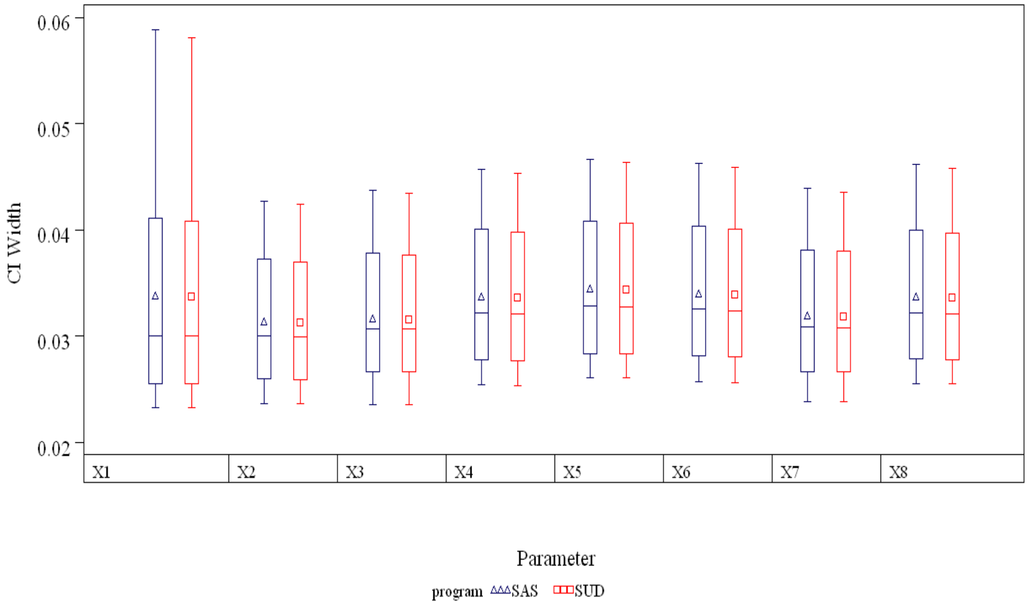

Figure 8: Box-and-Whisker Plots for 95% CI Width for SAS and SUDAAN (MAR)

4. Conclusions

As the mathematical formulas suggest, listwise deletion and subpopulation analysis techniques yielded different standard errors. However, the differences were so slight that hypothesis test results did not substantially vary between the two approaches. These findings were consistent for both the MCAR and MAR data structures. Thus, even though the standard errors were not always identical across the two methods, when data were MCAR and MAR, listwise deletion for conducting subpopulation analyses on observations with complete data did not lead to incorrect inferences.

Given the common use (among many social science researchers) of listwise deletion for handing missing data, these findings are encouraging when analyzing the data through OLS regression. First, based on the findings from the current study, it does not seem imperative for researchers who use complex sample data to obtain software that accommodates subpopulation analyses. Second, researchers do not need to be overly skeptical of complex sample survey findings in the literature based on the listwise deletion approach to subgroup analysis. Third, researchers who have conducted secondary analyses of complex sample survey data using a listwise deletion approach should feel no guilt about employing such a strategy. Furthermore, although the findings from the current study contradict the literature that suggests that listwise deletion methods for complex sample data are inappropriate, it is important to note that the results from this study are based on a controlled Monte Carlo simulation study of the two analytic methods, whereas much of the literature in this area is based on case studies of the two methods (e.g., comparisons between the listwise deletion and subpopulation analysis have been conducted on various national data sources such as the National Longitudinal Survey of Adolescent Health, National Health and Nutrition Examination Survey, and National Hospital Ambulatory Medical Care Survey; Chantala, 2006; West et al., 2007; West et al., 2008). Thus, the results of this study help fill an important gap in the methodological literature related to complex sample survey data analysis.

Because education, public health, and other social policies are often informed by analyses based on nationally representative complex sample data, it is important to understand how various analytic methods (i.e., listwise deletion vs. subpopulation analysis) impact resulting point estimates, standard errors, and hypothesis testing. Without this knowledge, inaccurate conclusions could be drawn from biased results. Methodological research on analysis options with complex sample data provides much-needed guidance for applied researchers who use such data to address substantive issues in the social and behavioral sciences. However, it is also important to note that the findings presented here are limited to complex data structures that have at least 20 PSUs per strata. Differences between listwise deletion and subpopulation analysis would likely be different for data structures that only include a few PSUs per strata.

References

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5,* 215 – 238.

Chantala, K. (2006). *Guidelines for analyzing Add Health data*. Retrieved June 6, 2007, from http://www.cpc.unc.edu/projects/addhealth/files/wt_guidelines.pdf

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.

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

Kiecolt, K.J. and Nathan, L.E. 1985. *Secondary Data Analysis of Survey Data*. Beverly Hills, CA: Sage.

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/

Research Triangle Institute. (2004). *SUDAAN Example Manual, Release 9.0***.** Research Triangle Park, NC: Research Triangle Institute.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, *45*, 283-288.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons, Inc.

SAS Institute. (2004). *SAS/STAT® 9.1 User’s Guide.* Cary, NC: SAS Institute Inc.

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.

West, B. T., Berglund, P., Heeringa, S.G. (2007). Alternative approaches to subclass analysis of complex sample survey data. *Proceedings of the 2007 JSM Proceedings, Survey Research Methods Section*.

West, B. T., Berglund, P., Heeringa, S.G. (2008). A closer examination of subpopulation analysis of complex-sample survey data. *The Stata Journal, 8*, 520-531.

Williams, R. L. (2000). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56*, 645-646.