Paper 246-2009

Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2

Patricia A. Berglund, Institute For Social Research-University of Michigan, Ann Arbor, Michigan

ABSTRACT

This paper presents practical guidance on three common survey data analysis techniques: repeated replication methods for variance estimation, subpopulation analyses, and techniques for handling missing data. A number of new features in the SAS® 9.2 survey procedures are demonstrated: Jackknife and Balanced Repeated Replication methods for variance estimation, subpopulation analysis with use of the DOMAIN option and the subsetting approach and the use of the NOMCAR (not missing completely at random) and multiple imputation options for handling missing data. Descriptive statistics as well as logistic regression are demonstrated. The analytic techniques presented can be used on any operating system and are intended for an intermediate level audience.

INTRODUCTION

Three important features of the SAS® v9.2 SURVEY analysis procedures are demonstrated: Repeated Replication methods, use of the DOMAIN statement for subpopulation analysis, and use of the NOMCAR statement for handling missing data. The Jackknife Repeated Replication and Balanced Repeated Replication methods are compared to the default Taylor Series linearization method for variance estimation. Use of the DOMAIN statement for correct subgroup analysis of survey data is compared to a "BY" group or sub-setting approach. Finally, the use of the NOMCAR option for including missing data as a domain subgroup in survey data analysis is demonstrated and subsequently compared to a multiple imputation approach with PROC MI and PROC MIANALYZE.

BACKGROUND INFORMATION ON COMPLEX SAMPLE SURVEYS AND DATA SETS

Complex surveys are comprised of data derived from sample designs that adjust for non-response and differing probabilities of selection. Complex samples differ from simple random samples (SRS) in that SRS designs assume independence of observations while complex samples do not. Most SAS® procedures assume a simple random sample and under-estimate variances when analyzing data from complex samples. Therefore, analysis of data from complex surveys should include methods of variance estimation that account for these sample design features (Kish, 1965 and Rust, 1985).

The analyses in this paper use data from the National Comorbidity Survey Replication, a nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Kessler et al, 2004 and Heeringa, 1996) and the NHANES 1999-2000 data set, with a focus on the medical examination data. In the NHANES data set, replicate weights constructed by the project staff are provided instead of strata and cluster variables. For each data set, weights that adjust for non-response and differing probabilities of selection are routinely used in analyses. Both data sets also include variables that allow analysts to incorporate the complex survey design into variance estimation computations: the stratum and SECU (Sampling Error Computing Unit) for the NCS-R data and replicate weights in the NHANES data.

Many projects elect to publish replicate weights rather than stratum and/or cluster variables. The rationale is often due to concerns about protecting respondent confidentiality. For the data analyst, however, providing only replicate weights places a limitation in the choice of appropriate variance estimation methods in that a replication approach must be employed. Although some users had programmed a JRR or BRR approach via SAS® MACRO language (Berglund, 2001), with the advent of Jackknife Repeated Replication and Balanced Repeated Replication options in SAS® v9.2, the analyst can easily perform both any of the linearization or repeated replication methods directly in the SURVEY procedures of SAS® v9.2.

VARIANCE ESTIMATION METHODS AND COMPLEX SAMPLE DESIGN VARIABLES

THE TAYLOR SERIES APPROACH

The Taylor Series Linearization approach (Rust, 1985) is based on a method that derives a linear approximation of variance estimates that are in turn used to develop corrected standard errors and confidence intervals for statistics of interest. A major advantage of the Taylor Series method is that it is very efficient computationally and is available for use with most important analyses of survey data including descriptive estimation of population statistics as well as linear and logistic regression. The SAS® SURVEYMEANS, SURVEYFREQ, SURVEYREG and SURVEYLOGISTIC procedures all use the Taylor Series method as the default.

RESAMPLING APPROACHES

Resampling approaches follow a specified method of selecting observations defined as probability sub-samples or replicates from which variance estimates are derived. Because the formulation of the probability samples is based upon the complex design, unbiased, design-corrected variance estimates can be derived.

Commonly used methods of resampling include Balanced Repeated Replication (BRR) and Jackknife Repeated Replication (JRR), (Wolter, 1985). Balanced Repeated Replication is a method that reduces the number of sub-samples needed by dividing each stratum into halves. Once the statistic of interest is derived from the half samples, the design corrected variance estimations can be developed through use of the usual formulas for variance. The Jackknife Repeated Replication method is similar to the BRR in that it performs replicate calculations of interest after developing replicates by deletion of a small and different portion of the total sample for each of the sample subsets. The repeated replication approaches are preferred for some non-linear statistics and in situations where the coefficient of variation exceeds 0.2 from the Taylor Series linearization method (Kish, 1965). Additional advantages of repeated replication methods are the simplicity of the process and applicability to a wide range of statistics.

COMPARISON OF LINEARIZATION AND REPEATED REPLICATION APPROACHES WITH STRATA AND CLUSTER VARIABLES

This section compares the means and standard errors for lifetime major depressive episode derived from three variance estimation methods: linearization (Taylor Series), JRR and BRR. The data set used is the NCS-R and weights, strata, and SECU (Sampling Error Computing Unit) variables are employed to account for the complex sample design. Use of the VARMETHOD option in the procedure statement allows specification of the variance estimation method desired.

SAS® CODE AND RESULTS

The following code illustrates how to use the VARMETHOD option in the procedure statement of PROC SURVEYMEANS. The default is the Taylor Series linearization method with the JRR and BRR as additional available options. The first set of statements perform a default Taylor Series variance estimation method and use the strata and cluster variables called “str” and “secu” along with the weight (“finalp1wt”) to account for the complex sample design. The use of the optional “cv” statistic requests the coefficient of variation from the Taylor Series method.

```
proc surveymeans data=ncsr mean stderr cv;
strata str;
cluster secu;
weight finalp1wt;
var dsm_mde;
run;
```

The next set of statements use the “VARMETHOD=JACKKNIFE” option to specify the JRR method. Replicate weights are automatically generated with the number of replicates equal to the number of PSU’s or 84 in this data set.

```
proc surveymeans data=ncsr varmethod=jackknife;
strata str;
cluster secu;
weight finalp1wt;
var dsm_mde;
run ;
```

The following code uses the "VARMETHOD=BRR" for the Balanced Repeated Replication method and also requests an optional printout of the Hadamard matrix generated for the BRR.

```
proc surveymeans data=ncsr varmethod=brr (printh);
strata str;
cluster secu;
weight finalp1wt;
var dsm_mde;
run;
```

Table 1.1 Statistics

Variable	Label	Mean	Std Error of Mean	Coeff of Variation
DSM_MDE	DSM-IV Major Depressive Episode(Lifetime) (Method=Taylor Series)	0.191711	0.004877	0.025438
	DSM-IV Major Depressive Episode(Lifetime) (Method= JRR)	0.191711	0.004879	NA
	DSM-IV Major Depressive Episode(Lifetime) (Method= BRR)	0.191711	0.004952	NA

Source: National Comorbidity Survey Replication, coefficient of variation for the Taylor Series = 0.025438.

Table 1.2 Variance Estimation

Method	
Number of Replicates	
BRR	44
Number of Replicates JRR	84

Analysis of Table 1.1 shows very little difference in the SE's between the three variance estimation methods and as expected, the exact same weighted point estimates of the mean of MDE (19.2%). The coefficient of variation (CV) for the Taylor Series method is well below .2, implying that any of the three methods are appropriate for a means analysis of MDE. Table 1.2 details the number of replicates produced by the JRR method with 84 replicates (42*2=84) and BRR with 44 replicates (42 +2 =44, BRR requires a multiple of 4 for the Hadamard matrix). See SAS®/STAT documentation on details of Balanced Repeated Replication and Hadamard matrices.

REPEATED REPLICATION WITH REPLICATE WEIGHTS

This section demonstrates the use of replicate weights for a means analysis of diastolic blood pressure. The data is from the interview and medical examination portions of the NHANES 1999-2000 survey. Only one of the three variance estimation methods, JRR is demonstrated here as the NHANES replicate weights are specifically labeled as Jackknife Repeated Replication weights. The replicate weights are the only design related variables published in this particular data set and as explained previously, repeated replication methods are the appropriate choice for a data set with solely replicate weights. The NHANES documentation explains more on how these variables were constructed and how they should be used in analysis (http://www.cdc.gov/nchs).

SAS® CODE AND RESULTS

The following code illustrates how to use the REPWEIGHTS statement in PROC SURVEYMEANS along with the VARMETHOD=JACKKNIFE option in the procedure statement.

```
proc surveymeans data=nhanes9900 varmethod=jackknife;
repweights wtmrep01-wtmrep52;
weight wtmec2yr;
```

```
var bpxdi1;
run;
```

Table 2.1 Variance Estimation	
Method	Jackknife
Number of Replicates	52
Replicate Weights Data Set	DEMO_BPX_NHANES9900

Table 2.2 Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
BPXDI1	Diastolic: Blood pres (1st rdg) mm Hg	6457	70.233553	0.518031	69.1940482	71.2730571

Source: NHANES 1999-2000 data

Tables 2.1 and 2.2 present results for the JRR analysis using the 52 replicate weights, (wtmrep01-wtmrep52). The mean for diastolic blood pressure is 70.23 with a standard error of .518. The 95% confidence limits are calculated using the corrected standard error with the usual formula for confidence limits around a mean (see SAS® documentation for details and formulae).

SUBPOPULATION ANALYSIS AND THE DOMAIN STATEMENT

Analysis of subpopulations is a common analytic practice. Confusion about how to correctly analyze subpopulations from survey data persists however, but with the addition of a DOMAIN statement in PROC SURVEYREG and PROC SURVEYLOGISTIC (new in SAS® 9.2), the analyst can correctly perform DOMAIN analyses in each of the main survey analysis SAS® procedures: SURVEYMEANS, SURVEYFREQ, SURVEYREG, and SURVEYLOGISTIC.

The temptation to merely subset the data to the subpopulation of interest through use of a “BY or WHERE” statement and then analyze the subgroup of interest is seemingly sensible yet statistically incorrect. One mistake made with this approach can be potential problems with strata with “singleton” clusters due to the sub-setting step rather than including the entire sample in the analysis. A second mistake is that a “BY” group or conditional approach analyzes only a part of the original sample and ignores the variability of the domain sample sizes across the strata of the sample design (Cochran, 1977). The result of these issues is generally under-estimated variances.

In summary, a subpopulation analysis should use the entire sample in the analysis and also take the sample size of the created domain into account. This is important because created domains are often not related to the original sample design. Use of the “domain” statement in the SAS® Survey procedures ensures that the correct use of the entire data set occurs and separate analyses per domain are performed while accounting for the random variability introduced by domain sample sizes unrelated to sample design.

SAS® CODE AND RESULTS

The following example compares a subpopulation analysis that uses a DOMAIN statement versus a WHERE or subset approach. As previously stated, the use of the DOMAIN statement produces statistically correct standard errors for analyses for subpopulations. Consider a domain defined as those people with each of the following disorders: General Anxiety Disorder, Major Depressive Episode, and Drug Abuse. This domain would then be analyzed and compared to all other people in the data set. The selected domain group is defined by having each of three disorders and the disorder variables of interest are not included in the original sample design of the data set, (Kessler et al, 2004). The analysis is first run with a DOMAIN statement to perform a subpopulation analysis in each domain while including the entire data set in the analysis and secondly, with a “where” statement that effectively subsets the data prior to analysis.

The following code performs a means analysis of lifetime alcohol abuse and dependence with a DOMAIN statement for subpopulation analysis of those with lifetime General Anxiety Disorder, Major Depressive Episode, and Drug Abuse (n=88). An examination of a cross-tabulation (not shown) of the strata*cluster variables indicates that only 33 strata with 46 clusters exist in the subpopulation yet due to the use of the DOMAIN statement, all 42 strata and 84 clusters are still used in the analysis.

```
proc surveymeans data=ncsr;
strata str;
cluster secu;
weight finalp2wt;
domain gad_mde_dra;
var dsm_ala dsm_ald;
run;
```

Table 3.1 Data Summary	
Number of Strata	42
Number of Clusters	84
Number of Observations	9836
Number of Observations Used	5692
Number of Obs with Nonpositive Weights	4144
Sum of Weights	5692.00049

Table 3.2 Domain Analysis				
gad_mde_dra	Variable	Label	Mean	Std Error of Mean
0	DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.125227	0.005485
	DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.049645	0.003283
1	DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.836008	0.039781
	DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.513451	0.051268

Table 3.1 details the number of observations, strata and clusters used in the analysis. Since this is a DOMAIN analysis, the full strata and cluster array is used. Table 3.2 shows that 83.6% (4.0%) of those with GAD, MDE, and Drug Abuse have Alcohol Abuse and 51.3% (5.1%) are diagnosed with Alcohol Dependence.

In the second demonstration of a subpopulation analysis, the data set is subset through use of a “where” statement . This approach reduces the number of records included in the analysis to n=88 with only 33 stratum and 46 clusters present in the subpopulation analyzed.

```
proc surveymeans data=ncsr mean stderr;
strata str;
cluster secu;
weight finalp2wt;
where gad_mde_dra=1;
var dsm_ala dsm_ald;
run;
```

Table 3.3 Data Summary	
Number of Strata	**33**
Number of Clusters	**46**
Number of Observations	**88**
Sum of Weights	54.2420583

Table 3.4 Statistics			
Variable	Label	Mean	Std Error of Mean
DSM_ALA	DSM-IV Alcohol Abuse(Lifetime)	0.836008	0.032639
DSM_ALD	DSM-IV Alcohol Dependence(Lifetime)	0.513451	0.036225

Analysis of the output from the two approaches (Tables 3.1-3.4) shows a difference in the number of strata and clusters used: 42 strata and 84 clusters in the DOMAIN approach but only 33 strata and 46 clusters with the "where" approach. This is an issue because the second approach excludes those stratum/clusters that do not happen to exist ("Sampling Zeros") in the subset of 88 people with GAD, MDE and Drug Abuse (DSM-IV definitions), though they theoretically could.

Note that this exclusion also has an impact on the size of the standard errors for the second approach in that the SE's are smaller than those from the DOMAIN analysis. For example, the SE for Alcohol Abuse is .0397 for the DOMAIN method and .0326 for the "where" group method and for Alcohol Dependence the DOMAIN method SE is .0513 and .0362 for the "BY" group method. This illustrates how use of a subset rather than the full data set with a DOMAIN analysis incorrectly underestimates the variance through the exclusion of strata/clusters.

The concepts presented in this section would also apply to other types of analysis procedures such as frequency tables, quantiles, and regression. See Appendix A for an example of a domain analysis using PROC SURVEYLOGISTIC. In summary, for subgroup analyses use of the DOMAIN statement provides a statistically appropriate approach and is available in each of the SAS® SURVEY analysis procedures in version 9.2.

MISSING DATA, PROC MI/MIANALYZE AND THE NOMCAR OPTION

Analysts have various options for handling missing data in survey data sets. The first and default in nearly all SAS® standard and survey procedures is to exclude missing data from the analysis, yet this is clearly incomplete and can lead to erroneous conclusions.

A second and new feature included in SAS® v9.2 is the use of the NOMCAR (Not Missing Completely at Random) option. The NOMCAR feature includes and analyzes missing data cases as a separate domain rather than simply excluding these cases from the analysis. This assumption is built upon the notion that often there is reason to believe that the missing data is not completely random. Or, that the missing data differs from the full data on the statistic of interest and inclusion of the missing data as a separate domain will provide valuable information that would otherwise be excluded. The advantage of this approach is that the entire data set is still analyzed and the complex sample is fully accounted for. By including the missing data as a separate domain and not assuming the data are missing completely at random, the analyst can examine the impact of the missing data as a group and the effect on the variance estimates.

A third option is to use multiple imputation to impute missing data prior to analysis, analyze imputed data sets using standard or survey SAS® procedures, and then use tools for correctly analyzing imputed data sets. See Rubin, (1987) or Shafer, (1999) for more information. SAS® offers PROC MI for multiple imputation and PROC MIANALYZE for subsequent analysis of the multiply imputed data sets (see SAS® documentation for details on these procedures). PROC MIANALYZE is fully capable of analyzing point estimates and standard errors produced by SAS® survey procedures and will therefore account for both the variability introduced by multiple imputation as well as the complex sample design features.

The next section of the paper demonstrates and compares three approaches for handling missing data in the NCS-R dataset via a means analysis for respondent income. The approaches considered are 1. exclusion of missing data (default in PROC MEANS and PROC SURVEYMEANS), 2. use of the NOMCAR (not missing completely at random) where missing data is analyzed as a separate domain in PROC SURVEYMEANS, and finally 3. multiple imputation of missing data using PROC MI and subsequent analysis of imputed data sets using PROC SURVEYMEANS and PROC MIANALYZE. The final approach combines multiple imputation of missing data, analysis of the imputed data using PROC SURVEYMEANS to account for the complex sample of the NCS-R data, and use of PROC MIANALYZE to analyze the imputed and complex sample adjusted variance. A comparison of the variances among the three approaches will illustrate how the way missing data is handled can affect variances.

SAS® CODE AND RESULTS
EXAMPLE 1A - SIMPLE RANDOM SAMPLE MEANS ANALYSIS WITH MISSING DATA EXCLUDED

This code uses a simple random sample variance estimation approach but is weighted using the "finalp2wt" in the data set. The personal income variable "inc_rsp" is analyzed and use of the nmiss statistic shows that there are 857 cases with missing data on the respondent income variable. This is from an overall n of 5692 (from part 2 of the survey) and the missing data is limited to stratum 35-42 and SECU=1. The missing data in this example was simulated to demonstrate analysis of data not missing completely at random (NOMCAR).

```
proc means data=ncsr nmiss sum mean std stderr min max;
var inc_rsp;
weight finalp2wt;
run;
```

Table 4.1
Analysis Variable : Respondent Income

N Miss	Sum	Mean	Std Dev	Std Error	Minimum	Maximum
857	114812102	**24837.02**	26886.18	**395.4440844**	0	125000.00

Table 4.1 shows the mean for respondent income of $24837.02 with a standard error of 395.44. This standard error is based on a simple random sample assumption but is weighted. Given that this data is from a survey data set, the standard error is under-estimated. Also, the nmiss column shows 857 cases are excluded from the analysis, due to missing data on respondent income.

EXAMPLE 1B - PROC SURVEYMEANS ANALYSIS WITH MISSING DATA EXCLUDED

Example 1b builds on Example 1a in that it uses a correct approach for a means analysis of survey data but still excludes missing data on the respondent income variable (default setting in PROC SURVEYMEANS). Use of the weight and complex sample design variables ensures correct standard errors but still excludes missing data from the analysis. If the missing data is not completely missing at random, there is reason to believe that if the missing data is included in the analysis, it will affect the variance estimates.

```
proc surveymeans nmiss mean stderr data=ncsr;
strata str;
cluster secu;
weight finalp2wt;
var inc_rsp;
run;
```

Table 4.2 Data Summary

Number of Strata	42
Number of Clusters	84
Number of Observations	9836
Number of Observations Used	5692
Number of Obs with Nonpositive Weights	4144
Sum of Weights	5692.00

Table 4.3 Statistics

Variable	N Miss	Mean	Std Error of Mean
inc_rsp	**857**	**24837**	**783.534079**

Tables 4.2 and 4.3 present the results of the PROC SURVEYMEANS analysis with missing data again excluded. Table 4.2 details the strata and clusters in the sample as well as the number of observations and sum of weights. Table 4.3

shows the Statistics from the means analysis of respondent income. Note that the mean for respondent income remains the same (compared to Table 4.1) but the standard error now includes the effect of weighting and the design stratification and clustering and thus, is much higher than the SRS standard error.

EXAMPLE 2 - PROC SURVEYMEANS ANALYSIS WITH THE NOMCAR OPTION

The second example extends Examples 1a and 1b through use of the NOMCAR option. This option includes the missing data as a separate domain and analyzes the missing data along with the non-missing data. The implication of this approach is that there is a meaningful difference in respondent income between the full and missing data cases and that the missing is not completely at random. In this situation, the missing data is restricted to stratum 35-42 and SECU=1 so that the **missing at random** assumption is violated.

The code is essentially the same as Example 1b with the exception of the NOMCAR option on the procedure statement. This specifies that the missing data be analyzed as a separate domain.

```
proc surveymeans data=ncsr nomcar nmiss mean stderr;
strata str;
cluster secu;
weight finalp2wt;
var inc_rsp;
run;
```

Table 4.4 Data Summary	
Number of Strata	42
Number of Clusters	84
Number of Observations	9836
Number of Observations Used	5692
Number of Obs with Nonpositive Weights	4144
Sum of Weights	5692.00

Table 4.5 Variance Estimation	
Method	Taylor Series
Missing Values	**Included (NOMCAR)**

Table 4.6 Statistics			
Variable	N Miss	Mean	Std Error of Mean
inc_rsp	857	**24837**	**910.410784**

Tables 4.4-4.6 show that the inclusion of the missing data as a separate domain increases the standard errors substantially from 395.44 (SRS) to 783.53 (Taylor Series with missing data excluded) to 910.41 (Taylor Series with the NOMCAR option). This implies that the exclusion of the missing data would incorrectly reduce variance by excluding the missing strata and SECU's with missing data on respondent income. Note the message displayed in Table 4.5 "Included(NOMCAR)"; indicating the use of the NOMCAR option of PROC SURVEYMEANS.

EXAMPLE 3 - MULTIPLE IMPUTATION AND ANALYSIS WITH PROC MI, PROC SURVEYMEANS, AND PROC MIANALYZE

The final example compares the previous results (Examples 1a,1b, and 2) with output from multiple imputation and subsequent analysis of imputed data sets using PROC SURVEYMEANS and PROC MIANALYZE. The advantage of this approach is that missing data is imputed using a regression based approach prior to analysis in PROC

SURVEYMEANS and PROC MIANALYZE. Imputation eliminates the need to include missing data as a separate domain and addresses imputation through use of a statistically valid approach (Rubin, 1987).

Multiple imputation and analysis of imputed data generally follow three distinct steps when implemented in SAS®: 1. imputation of missing data using PROC MI, 2. analysis of imputed data sets using standard SAS procedures, and 3. analysis of the point estimates and standard errors from step 2 using PROC MIANALYZE.

This demonstration illustrates how to code and implement these steps: the use of PROC MI for the imputation step, analysis of the 5 imputed data sets with PROC SURVEYMEANS, and finally, use of PROC MIANALYZE for analysis of the SURVEYMEANS point estimates and standard errors. The output from last step provides a comparison to the other methods of handling missing data presented in examples 1a, 1b, and 2. Given three differing methods of handling missing data, users can examine the impact on the variance from each approach. For users unfamiliar with multiple imputation, see SAS® documentation on PROC MI/MIANALYZE and publications such as Rubin, 1987 and Schafer, 1999.

SAS® CODE AND RESULTS

Step1: Use PROC MI with number of imputations=0 (nimpute=0) to examine missing data patterns. The type of missing data pattern is important for understanding where missing data occurs in the data set as well as informed specification of an appropriate imputation method. PROC MI output is shown in Table 4.7 and indicates 857 missing data on respondent income and a monotone missing data pattern.

```
proc mi data=ncsr nimpute=0;
run;
```

Table 4.7 Missing Data Patterns

Group	SEXF	Age	inc_rsp	Freq	Percent	Group Means	
						Age	inc_rsp
1	X	X	X	4835	84.94	43.174354	25470
2	X	X	.	857	15.06	44.527421	.

Step 2: Use PROC MI to multiply impute missing data (monotone pattern) on respondent income using a regression method for continuous variables (monotone method=reg;). Note that the imputation is unweighted and “sexf” is declared a class variable (“class sexf;”) and sexf and age are full data and therefore, donor variables that contribute to the imputation of respondent income. Also, the order of the variables in the var statement is important since variables are imputed from left to right. The statement “Out=outmi1” creates an output data set called “outmi1” and this data set contains 5 imputed data sets of n=5692 each “(nimpute=5)”. A variable called _imputation_ is generated and retained for identification of each imputed data set. In this case, _imputation_ has values from 1 to 5.
Output is shown in Tables 4.8 and 4.9. Note that the mean respondent income is $25,464 with a standard error of $375.60 (based on analysis of the 5 imputed data sets). This standard error accounts for the variability introduced by imputation but not for the complex sample design. It is also unweighted.

```
proc mi nimpute=5 data=ncsr out=outmi;
class sexf;
monotone method=reg;
var sexf age inc_rsp;
run;
```

Table 4.8 Model Information

Data Set	WORK.FOUR
Method	Monotone
Number of Imputations	5
Seed for random number generator	659887001

Table 4.9 Parameter Estimates										
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
inc_rsp	25464	375.602876	24725.46	26201.70	458.29	25327	25607	0	67.79	<.0001

Step 3: Use PROC SURVEYMEANS to analyze the "outmi1" data set by the variable _imputation_ (automatically generated SAS® variable used to define each imputed file). A means analysis of the imputed "inc_rsp" variable is performed while taking the complex sample into account. The means and standard errors for each domain (or imputed data set) are stored for further analysis in PROC MIANALYZE via "ods output domain=outsummary" statement. (Output shown in Table 4.10). Note that this standard error is now corrected taking the complex sample into account and the results are also weighted, hence producing a slightly different mean value for "inc_rsp" ($25,039.00).

```
ods output domain=outsummary;
proc surveymeans data=outmi1;
domain _imputation_;
var inc_rsp;
strata str;
cluster secu;
weight finalp2wt;
run;
ods output close;
```

Table 4.10 Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
inc_rsp	28460	25039	786.893470	23450.5317	26626.5623

Step 4: Use PROC MIANALYZE to analyze the output from Step 3 and account for the variability introduced by the imputation step. Because the standard errors used in Step 4 are derived from PROC SURVEYMEANS, the complex sample design variance adjustment is included in the variance produced here. The data set (data=outsummary) from step 3 is used with the "modeleffects mean" and "stderr stderr " statements, which refer to the mean and standard errors from the PROC SURVEYMEANS analysis. Note that this data set contains 5*5692 or 28460 records due to the 5 imputed data sets being analyzed together. Output is shown in Tables 4.11 and 4.12.

```
proc mianalyze data=outsummary;
modeleffects mean;
stderr stderr;
run;
```

Table 4.11 Variance Information							
Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Mean	23731	660362	688840	2340.4	0.043124	0.042160	0.991639

Table 4.12 Parameter Estimates							
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum
Mean	25039	829.963570	23411.01	26666.09	2340.4	24896	25276

Table 4.13 Summary of Four Approaches (PROC MEANS, PROC SURVEYMEANS WITHOUT NOMCAR, PROC SURVEYMEANS WITH NOMCAR, AND MULTIPLE IMPUTATION APPROACH)			
Variable (Method)	N Miss	Mean	Std Error of Mean
Inc_rsp (Proc Means with Missing Data Excluded)	857	24837	395.444
Inc_rsp (Proc SurveyMeans without NOMCAR)	857	24837	783.534
Inc_rsp (Proc SurveyMeans with NOMCAR)	857	24837	910.411
Inc_rsp (Imputed and Analyzed with Proc MI and MIANALYZE)	None	25039	829.964

A side-by-side comparison of the means and standard errors on respondent income (Table 4.13) indicates that the use of the NOMCAR and multiple imputation approaches produce the highest and likely the most accurate standard errors along with very similar means. It should be noted that the mean values are slightly different due to the production of 5 multiply imputed data sets that are then analyzed together, thus producing a mean of $25039 from the imputed files with n=28460. The standard errors for the two methods that either recognize the missing data (NOMCAR) in the estimation of standard errors or impute the missing data (PROC MI) are similar and represent the square root of the variance under complex sample and in the case of the imputed data analysis, also the variability introduced by the imputation step.

CONCLUSION

The focus of this paper is to provide the survey data analyst with practical guidance on use of a variety of features in the SAS® v9.2 survey procedures. Repeated Replication methods, subpopulation analysis, and techniques for handling missing data were discussed and demonstrated using public-use data sets such as NCS-R and NHANES.

APPENDIX A

```
proc surveylogistic data=ncsr;
strata str;
cluster secu;
weight finalp2wt;
domain gad_mde_dra;
model dsm_ala (event='1')= sexm;
run;

proc surveylogistic data=ncsr;
weight finalp2wt;
where gad_mde_dra=1;
model dsm_ala (event='1')= sexm;
run;
```

These two PROC SURVEYLOGISTIC code samples illustrate how to use a DOMAIN statement within PROC SURVEYLOGISTIC along with design variables “str” and “secu” and the final weight “finalp2wt”. This is contrasted with a “conditional” subpopulation analysis using a WHERE statement to subset the data. As previously explained, the DOMAIN approach provides statistically correct subpopulation analyses.

Appendix Table 1.1
Analysis of Maximum Likelihood Estimates from Domain Analysis

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5785	0.3690	2.4580	0.1169
SEXM	1	**3.1678**	**0.9020**	12.3338	0.0004

Appendix Table 1.2
Analysis of Maximum Likelihood Estimates from WHERE Analysis

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.5785	0.3741	2.3916	0.1220
SEXM	1	**3.1678**	**0.8257**	14.7190	0.0001

REFERENCES

Berglund, P. (2001) “Analysis of Complex Sample Survey Data Using the SURVEYMEANS and SURVEYREG Procedures and Macro Coding” SUGI 27.

Cochran, W.C. (1977) Sampling Techniques, Third Edition. John Wiley and Sons, NY

Heeringa, S. (1996) “National Comorbidity Survey (NCS): Procedures for Sampling Error Estimation”.

Kalton, G. (1977). “Practical Methods for Estimating Survey Sampling Errors,” Bulletin of the International Statistical Institute, Vol 47, 3, pp. 495-514.

Kessler, R.C., Berglund, P., Chiu, W.T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B-E., Walters, E.E., Zaslavsky, A., Zheng, H. (2004). The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *The International Journal of Methods in Psychiatric Research*, 13(2), 69-92.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

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

Rust, K. (1985). Variance Estimation for Complex Estimation in Sample Surveys. Journal of Official Statistics, Vol 1, 381-397. (CP)

Schafer, J. L. (1999), "Multiple Imputation: A Primer," *Statistical Methods in Medical Research*, 8, 3–15).

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag

CONTACT INFORMATION

Patricia Berglund
Institute for Social Research
University of Michigan
426 Thompson St.
Ann Arbor, MI 48106
pberg@umich.edu