An Introduction to Multiple Imputation of Complex Sample Data using SAS® v9.2

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

ABSTRACT

This paper presents practical guidance on the proper use of multiple imputation tools in SAS® 9.2 and the subsequent analysis of multiple imputed data sets from a complex sample survey data set. Use of the MI and MIANALYZE procedures and SAS survey procedures for typical descriptive and inferential analyses is demonstrated. The analytic techniques presented can be used on any operating system and are intended for an intermediate level audience.

INTRODUCTION

This paper presents an outline of the process of multiple imputation and application of the three step process of imputation using PROC MI, analysis of imputed data sets using SAS analysis procedures including Survey procedures for complex survey data and use of PROC MIANALYZE for analysis of imputed data sets and output from general analytic procedures. Procedures used during the 2nd step of this process are PROC SURVEYMEANS, PROC SURVEYREG, and PROC SURVEYLOGISTIC.

A brief overview of imputation strategies and recommended methods is provided along with detailed examples using a public use data set, the National Comorbidity Survey Replication, a nationally representative complex sample data set focused on mental health in the US. The examples cover imputation of both continuous and categorical variables as well as analysis of imputed data sets using adjustments for survey data assumptions.

MISSING DATA AND MULTIPLE IMPUTATION

Missing data is a pervasive and persistent problem in many data sets. Common reasons for missing data include survey structure that deliberately results in missing data (questions asked only of women), refusal to answer (sensitive questions), insufficient knowledge (month of first words spoken), and attrition due to death or loss of contact with respondents in longitudinal surveys. Missing data can be categorized as unit non-response (entire survey is missing) or item non-response (some questions are missing within a survey).

Analysts often make assumptions about the nature of missing data including categorizations such as: Missing at Random (MAR), Missing Completely at Random (MCAR) or Not Missing at Random (NMAR). MAR means that the existing missingness depends only on the observed variables, MCAR means that missingness does not depend on observed variables and NMAR is used to describe missingness that depends on both observed and non-observed variables. PROC MI and PROC MIANALYZE both use the MAR assumption for all analyses.

Imputation methods can be defined as simple or multiple. Though simple imputation is attractive and often used to impute missing data, the focus of this paper is use of multiple imputation methods in SAS. This is due to the ability of the multiple imputation process to incorporate statistically sophisticated techniques and draw from distributions of "plausible" values while accounting for the variability introduced by the process of selecting a value for the missing data point (Rubin, 1987). Simple imputation methods such as inserting a mean value or a value selected from a similar type of respondent are attractive due to ease of concept and implementation but do not account for the variability introduced by the imputation process. They also tend to distort the variable distribution once imputation is complete. Given these limitations, multiple imputation is generally considered a preferred method for dealing with missing data.

The analyses in this paper use data from the National Comorbidity Survey Replication, a public release, nationally representative sample based on a stratified, multi-stage area probability sample of the United States population (Kessler et al, 2004 and Heeringa, 1996). The NCS-R data set includes variables that allow analysts to incorporate the complex survey design into variance estimation computations through the use of the "SESTRAT" (strata) and "SECLUSTR" (Sampling Error Computing Unit or cluster) variables in addition to probability weights (NCSRWTSH, NCSRWTLG). All analyses should account for the complex sample and be correctly weighted in order to produce statistically correct variance estimates. These variables will be used in the 2nd step of the imputation process (analysis of imputed data sets) with use of the SAS Survey procedures. Combined with PROC MIANALYZE, the variance estimates will be fully corrected due to variability introduced by multiple imputation plus the variance adjustments required to account for the complex sample design. For more information on complex sample data analysis see the SAS Survey procedures documentation, Kish (1965), and Rust, (1985).

MISSING DATA PATTERNS AND TYPES OF VARIABLES

Missing data patterns provide important information about the amount and structure of missing data. Through examination of the missing data pattern, the missingness can be characterized as arbitrary or a more specialized pattern of missing data such as monotone missing data. Arbitrary missing data is used to describe a missing data pattern that has missingness interspersed among full data values while monotone missing data is a pattern in which the missing data exists at the end (reading from left to right) of the data record with no gaps between full and missing data. In other words, once a variable has missing data, all variables to the right of the missing data variable in a rectangular data array are also missing. This is an important distinction due to the manner in which missing data is imputed, moving from left to right across the rectangular data array of columns and rows. (See Table 1.0 for a graphic of common missing data patterns). The implication for the imputation step and selection of imputation method is that a monotone missing data pattern allows the analyst more flexibility in selecting an appropriate imputation technique. Analysis of existing missing data patterns is a critical first step in planning the overall imputation and is done by PROC MI (see subsequent examples for details).
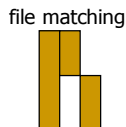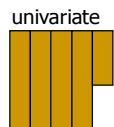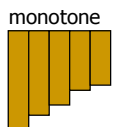
Table 1.0 Patterns of Missing Data

Another important consideration in planning an imputation is the type of variables (numeric or character) that either require imputation or will contribute to the imputation process. Careful attention to the variable type will help ensure that the imputation is done correctly. For example, categorical variables (binary, nominal or ordinal) and continuous variables can be imputed using PROC MI but require different methods for the imputation step. Knowledge of the variables with missing data as well as variables used during the imputation will allow the analyst to make correct decisions about how to set up the imputation. SAS V9.2 allows the use of the CLASS statement for categorical variables in both PROC MI and PROC MIANALYZE.

IMPUTATION OF MISSING DATA – PROC MI

Numerous methods for the imputation step are available in PROC MI and are fully detailed in the PROC MI documentation (see summary Table 54.3 of the documentation for a nice overview). However, most imputation methods require a monotone missing data pattern if any categorical variables are included in the PROC MI step. The default method of MCMC (Markov Chain Monte Carlo, Schafer, 1997) is appropriate for continuous arbitrary missing data while for categorical variables, a monotone missing data pattern offers the widest range of imputation options.

Once the imputation method is determined, the question of how many imputations arises. This number (called m) is a balance between the amount of missing data and the relative efficiency desired. For example, to achieve relative efficiency (RE) of 1.0 would require an infinite number of imputations but for most problems, a few imputations, 3-10 are all that is needed for a RE of .90 or higher. (See the Details section of the PROC MI documentation for the relative efficiency formula and Table 54.5 for more information on RE.)

ANALYSIS OF IMPUTED DATA SETS

Once the imputation step is complete, one data set with m concatenated imputed data sets will be created by PROC MI along with an automatic variable called _IMPUTATION_. The concatenated data set can, in turn, be used with SAS Survey procedures and analyzed using typical analytic techniques. The _IMPUTATION_ variable serves as an identifier for analysis of each imputed data set with values of m=1 to number of imputations performed. The recommended approach is to use the desired analytic technique using the _imputation_ variable as a DOMAIN variable and save the results in a data set appropriate for further analysis using PROC MIANALYZE. The point of this step is to analyze the imputed data sets using the analytic

technique originally selected, save the estimates and standard errors from that procedure and finally use MIANALYZE for additional analysis of the variability introduced by the process of multiple imputation.

In the examples presented in this paper, the 2nd step requires the use of the SAS Survey procedures due to the complex sample design of the NCS-R data set. This type of design generally results in increased variance due to clustering and other design features (Kish, 1965) and use of the Survey procedures correctly adjusts the standard errors to account for the complex sample design. A key reason for this approach is that the needed output from the 2nd step analyses consists of a point estimate (i.e. a mean or parameter estimate) along with a complex sample corrected standard error needed for further use in PROC MIANALYZE.

SYNTHESIS OF IMPUTATION AND ANALYSIS RESULTS - PROC MIANALYZE
Once the imputation step and analysis of imputed data sets using the selected procedure are complete, the final step in the process is analysis of the combined data sets using PROC MIANALYZE. This procedure synthesizes the results by producing means of the point estimates of interest (means, parameter estimates, etc.) across the imputed data sets along with adjusted variance and standard errors taking the uncertainty introduced by the imputation into account. Without use of PROC MIANALYZE, the analyst would be underestimating the variability due to the process of multiple imputation.

A number of important statistics are provided by PROC MIANALYZE. The combined point estimate is the mean of the point estimates over m imputations: (Formulae and additional information available from SAS v9.2 PROC MI/MIANALYZE documentation)

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i$$

The within imputation variance is the average variance within the imputed data sets:

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i$$

The between imputation variance is the variance across the m imputed data sets:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \bar{Q}\right)^2$$

And, total variance is the sum of the within and between variances:

$$T = \bar{U} + \left(1 + \tfrac{1}{m}\right)B$$

The within, between and total variance estimates provide information about the variability introduced due to imputation with the standard errors for the statistic of interest are adjusted to account for the process. Concrete examples will be provided in the application of the three step process in the next sections of this paper.

APPLICATION OF THE THREE STEP PROCESS
This paper focuses on three practical examples with monotone or arbitrary missing data and continuous and categorical variables to be imputed and/or used in the imputation. The examples include 1. use of the default MCMC method for imputation of a continuous variable using all continuous covariates, 2. use of the CLASS statement for analyses with categorical variables used to impute a continuous variable with a monotone missing data pattern, and 3. use of the MCMC method to impute enough data to achieve a monotone missing data pattern followed by use of logistic regression to complete the imputation step. Other key imputation options such as round, min, max, and seed values are included in these examples.

All examples include application of the three step process by presenting 1. imputation using PROC MI, 2. analytic procedures for proper analysis of imputed complex survey data sets (PROC SURVEYMEANS, PROC SURVEYLOGISTIC, and PROC SURVEYREG) and 3. use of PROC MIANALYZE for analysis of the output from the analysis step including accounting for the variability introduced during the imputation step.

SAS EXAMPLES-CODE AND RESULTS

EXAMPLE 1: MCMC IMPUTATION FOR CONTINUOUS VARIABLES

Step 1: The code below illustrates how to use PROC MI with the **nimpute=0** option to examine the missing data pattern. This is followed by use of the default MCMC method to impute a continuous variable using all continuous covariates. This is a basic approach to introduce the use of PROC MI for a simple imputation with continuous variables.

```
proc mi nimpute=0 ;
   var hhinc age weight ;
run ;
```

Table 1.1 Missing Data Patterns								
						Group Means		
Group	HHINC	AGE	WEIGHT	Freq	Percent	HHINC	AGE	WEIGHT
1	X	X	X	5590	98.21	59443	43.345796	174.656530
2	X	X	.	102	1.79	60467	45.156863	.

Table 1.1 shows that 1.79% of the n of 5692 have missing data on the WEIGHT variable, indicated by a '.'. This is a monotone missing data pattern because the missing data exists only on WEIGHT with full data on HHINC and AGE.

The code below uses PROC MI to impute the missing data on WEIGHT with the default method of MCMC, **nimpute=5** option to produce 5 imputed data sets, and a seed value for later replication of results. Because this is a monotone missing data pattern with continuous variables, the default imputation method is suitable for this problem. Note the use of the **seed=** option to ensure the ability to replicate these results at a later time.

```
proc mi data=one nimpute=5 seed=454 out=outimputedex1;
   var hhinc age weight ;
run ;
```

Table 1.2 Variance Information							
	Variance				Relative Increase	Fraction Missing	Relative
Variable	Between	Within	Total	DF	in Variance	Information	Efficiency
WEIGHT	0.003612	0.312064	0.316399	4441.7	0.013890	0.013792	0.997249

Table 1.3 Parameter Estimates									
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
WEIGHT	174.666852	0.562493	173.5641	175.7696	4441.7	174.580348	174.735665	0	310.52	<.0001

Table 1.2 provides information about the between, within, and total variance from the imputation step in PROC MI. Other useful statistics are the relative increase in variance due to non-response, fraction of missing information, and relative efficiency. See previous section and PROC MI documentation for details on these formulae and further interpretation.

Table 1.3 provides the mean of WEIGHT (174.67) across the m=5 imputed data sets along with the standard error, confidence limits, degrees of freedom and other univariate statistics. The t test for the null hypothesis of mean=0 can be changed to suit other analytic tests.

The following PROC MEANS code uses the concatenated data set called "outimutedex1" with a BY statement to perform a means analysis for each imputed data set (indicated by the _imputation_ variable automatically produced by PROC MI).

Because this is an exploratory analysis, use of PROC MEANS rather than PROC SURVEYMEANS is an acceptable way to examine the imputed data.

```
proc means data=outimputedex1 ;
  by _imputation_ ;
  var weight ;
run ;
```

Table 1.4 Means by Imputation (Partial Output) Analysis Variable : WEIGHT Weight in pounds or kgs					
Imputation Number	N	Mean	Std Dev	Minimum	Maximum
1	5692	174.5803478	42.1174393	75.5132350	300.0000000
2	5692	174.6426673	42.1503365	84.0550331	300.9452342

Table 1.4 shows partial output of the PROC MEANS results and illustrates how the means and standard deviations are slightly different across two imputed data sets. This is expected due to different values being imputed during the imputation step.

Step 2: The second step of the process consists of analysis of the 5 imputed data sets using PROC SURVEYMEANS. Use of the SURVEYMEANS procedure is required to correctly estimate variances due to the complex sample design of the NCS-R data set (see previous section for details and references). Use of the DOMAIN statement rather than a BY statement is recommended for an unconditional analysis (see SURVEYMEANS documentation for more on this topic).

```
proc surveymeans data=outimputedex1 ;
  strata sestrat ;  cluster seclustr ;  weight ncsrwtlg ;
  var weight ; domain _imputation_ ;
  ods output domain = outex1 ;
run ;

proc print data=outex1 ;
run ;
```

Table 1.5 Data Summary	
Number of Strata	42
Number of Clusters	84
Number of Observations	28460
Sum of Weights	28460.0024

Table 1.6 Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
WEIGHT	Weight in pounds or kgs	28460	174.786703	0.842607	173.086254	176.487152

Table 1.7 Domain Analysis: Imputation Number							
Imputation Number	Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
1	WEIGHT	Weight in pounds or kgs	5692	174.755614	0.876056	172.987661	176.523567
2	WEIGHT	Weight in pounds or kgs	5692	174.760515	0.836552	173.072285	176.448745
3	WEIGHT	Weight in pounds or kgs	5692	174.803738	0.851814	173.084708	176.522768
4	WEIGHT	Weight in pounds or kgs	5692	174.690528	0.816075	173.043621	176.337435
5	WEIGHT	Weight in pounds or kgs	5692	174.923120	0.850072	173.207604	176.638635

Table 1.5 includes important information about the sample design such as number of strata and clusters (42 and 84 respectively) along with number of observations used 5*5692=28460 for the 5 data sets produced by PROC MI. Table 1.6 includes the overall mean estimate of body weight (WEIGHT) of 174.79 with a standard error of .842. Examination of the Domain analysis in Table 1.7 illustrates how the means and SE's change slightly across the five imputed data sets.

Step 3: Use of PROC MIANALYZE for analysis of the imputed data sets and output from the SURVEYMEANS analysis completes the three step process. Here, the output data set from PROC SURVEYMEANS is used as input for the PROC MIANALYZE analysis and therefore includes complex sample corrected standard errors and accounts for the variability introduced by multiple imputation via use of the MIANALYZE procedure. Note again that the overall n used by PROC MIANALYZE is 5*5692 or 28460.

```
proc mianalyze data=outex1 ;
  modeleffects mean ;
  stderr stderr ;
run ;
```

Table 1.8 Model Information	
Data Set	WORK.OUTEX1
Number of Imputations	5

Table 1.9 Variance Information							
	Variance				**Relative**	**Fraction**	
					Increase	**Missing**	**Relative**
Parameter	**Between**	**Within**	**Total**	**DF**	**in Variance**	**Information**	**Efficiency**
mean	0.007450	0.716297	0.725237	26321	0.012482	0.012403	0.997526

Table 1.10 Parameter Estimates										
									t for H0:	
Parameter	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**Minimum**	**Maximum**	**Theta0**	**Parameter=Theta0**	**Pr > |t|**
Mean	174.786703	0.851608	173.1175	176.4559	26321	174.690528	174.923120	0	205.24	<.0001

Table 1.8 provides information about the data set read into PROC MIANALYZE, "work.outex1", as well as the number of imputed data sets included in the analysis (5). Table 1.9 includes the variance (between, within and total) and the relative increase in variance and fraction missing information which both provide information about the increase in variability due to missing data. Table 1.10 provides the overall estimate for body weight of 174.79 (.852) along with the standard error and other descriptive information.

An examination of the change in the standard errors across the three steps shows a steady increase as the complex sample and then the combination of the complex sample and the variability introduced by the imputation are accounted for: Step 1 SE=.562, Step 2: SE=.843 and Step 3 SE=.852. As expected, the mean of the parameter of interest for body weight remains 174.79 (weighted) for Steps 2 and 3 with an unweighted mean of 174.67 from Step 1.

EXAMPLE 2: MONOTONE REGRESSION IMPUTATION WITH MISSING DATA ON CONTINUOUS VARIABLE AND CATEGORICAL COVARIATES

Step 1: The second example builds upon the first by adding additional variables to contribute to the imputation and also introduces the complexity of using categorical variables such as gender, education, and region. The code below illustrates the use of PROC MI with a NIMPUTE=0 option to initially examine the existing missing data pattern. Table 2.1 shows that the variable called body weight (WEIGHT) has missing data and the pattern of missing data is monotone.

```
proc mi data=one nimpute=0 ;
   var hhinc age sex ed4cat region weight ;
run ;
```

Table 2.1 Missing Data Patterns												
								Group Means				
Group	HHINC	AGE	SEX	ED4CAT	REGION	WEIGHT	Freq	Percent	HHINC	AGE	SEX	ED4CAT	REGION	WEIGHT
1	X	X	X	X	X	X	5590	98.21	59443	43.345796	1.575134	2.650805	2.577818	174.656530
2	X	X	X	X	X	.	102	1.79	60467	45.156863	1.931373	2.647059	2.421569	.

The next code segment uses PROC MI to produce 5 imputed data sets combined into a data set called "outimputex2" and includes a CLASS statement for the categorical variables REGION, ED4CAT, and SEX. The MONOTONE REGRESSION statement requests the use of the regression method for imputation of the continuous variable WEIGHT. This is a different method than was used in the first example where the default of MCMC was used. The regression method is recommended for imputation of a continuous variable with a monotone missing data pattern and categorical covariates contributing to the imputation step. Note that the CLASS statement can be used only with a monotone missing data pattern.

```
proc mi data=one nimpute=5 out=outimputex2 seed=20102 ;
   class region ed4cat sex ;
   monotone regression ;
   var hhinc age region sex ed4cat weight ;
run ;
```

Table 2.2 Monotone Model Specification	
Method	Imputed Variables
Regression	WEIGHT

Table 2.3 Variance Information							
	Variance				Relative Increase	Fraction Missing	Relative
Variable	Between	Within	Total	DF	in Variance	Information	Efficiency
WEIGHT	0.000945	0.311813	0.312947	5564.8	0.003637	0.003630	0.999274

The selected output in Tables 2.2 and 2.3 comes from the PROC MI imputation. The data is now fully imputed and *m*=5 imputed data sets are created using the regression imputation method. Table 2.3 provides details for the imputed variable WEIGHT. Note a very high relative efficiency with 5 imputations, indicating even fewer imputations might have been performed.

Step 2: Step 2 uses PROC SURVEYREG to perform linear regression with the 5 imputed data sets stored in the file called "outimputex2". The results from the SURVEYREG analysis are then saved for later use in PROC MIANALYZE. The following syntax uses a DOMAIN statement to analyze each data set separately. Because the output data set includes records where the variable _IMPUTATION_ is set to missing (a default when using a DOMAIN statement), a WHERE statement is used to exclude those records in the output data set.

```
proc surveyreg data=outimputex2 ;
  strata sestrat ;  cluster seclustr ;
  weight ncsrwtlg ; class region sex ed4cat ;
  domain _imputation_ ;
  model weight=hhinc age region sex ed4cat / solution ;
  ods output ParameterEstimates=outregex2 (where=(_imputation_ ne . ));
run ;
```

	Table 2.4 Partial SURVEYREG Output Imputation =1 Estimated Regression Coefficients			
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	149.562453	3.28476205	45.53	<.0001
HHINC	-0.000021	0.00001242	-1.71	0.0951
AGE	0.076099	0.04337713	1.75	0.0867
REGION 1	4.546014	2.57144796	1.77	0.0843
REGION 2	5.068807	1.66139387	3.05	0.0039
REGION 3	7.925830	1.85368849	4.28	0.0001
REGION 4	0.000000	0.00000000	.	.
SEX 1	33.666299	1.61972248	20.79	<.0001
SEX 2	0.000000	0.00000000	.	.
ED4CAT 1	-2.037483	2.21794693	-0.92	0.3635
ED4CAT 2	3.536673	1.75089985	2.02	0.0498
ED4CAT 3	4.642264	1.75254386	2.65	0.0113
ED4CAT 4	0.000000	0.00000000	.	.

The output of Table 2.4 is typical of what PROC SURVEYREG produces for each value of the variable _IMPUTATION_. The parameter estimates and standard errors are the usual linear regression output with the exception of the standard errors being adjusted to account for the complex sample of the NCS-R data set. In general, they will be larger than simple random sample assumption standard errors due to the features of complex samples.

The data set produced by the ODS OUTPUT statement of PROC SURVEYREG requires the use of the COMPRESS option to format the data such that PROC MIANALYZE can correctly process the parameter estimates, (Agnelli, SAS Tech Support). The syntax below illustrates how to correctly remove the blanks in the variable called PARAMETER in the output data set "outregex2":

```
data outregex2;
  set outregex2;
  parameter=compress(parameter);
run;
```

Step 3: Once the data set generated by PROC SURVEYREG is correctly structured, use of PROC MIANALYZE concludes the process. An additional programmatic detail is the correct manner to reference the values of the class variables region, sex and education in the PROC MIANALYZE step: refer to the variables as region1, region2, region3 and region4 in the MODELEFFECTS statement. A similar naming strategy is required for the categorical variables SEX and ED4CAT. Note that the omitted categories of region4, sex2, and ed4cat4 are excluded in the modeleffects statement.

```
proc mianalyze parms=outregex2;
  modeleffects intercept hhinc age region1 region2 region3 sex1 ed4cat1 ed4cat2   ed4cat3 ;
  run;
```

Table 2.5 Parameter Estimates							
Parameter	**Estimate**	**Std Error**	**95% Confidence Limits**		**DF**	**Minimum**	**Maximum**
Intercept	149.803377	3.272402	143.3895	156.2173	67949	149.562453	150.179837
Hhinc	-0.000019838	0.000012928	-0.0000	0.0000	4528.7	-0.000022684	-0.000017908
Age	0.072143	0.044948	-0.0160	0.1603	3434.4	0.060867	0.080630
region1	4.475203	2.693912	-0.8054	9.7558	10868	3.914326	4.843456
region2	4.774186	1.624644	1.5887	7.9597	3090.3	4.449094	5.068807
region3	7.673220	1.831686	4.0826	11.2639	7040	7.390544	7.968183
region4	0	0	.	.	.	0	0
sex1	33.532870	1.636445	30.3254	36.7403	37188	33.280993	33.666299
sex2	0	0	.	.	.	0	0
ed4cat1	-2.048017	2.281712	-6.5208	2.4247	8062.1	-2.397360	-1.567889
ed4cat2	3.864954	1.823668	0.2854	7.4445	827.05	3.340223	4.446401
ed4cat3	4.834270	1.825331	1.2523	8.4162	987.75	4.392407	5.510266
ed4cat4	0	0	.	.	.	0	0

Table 2.5 includes the synthesized output for the 5 imputed data sets with the complex sample adjusted standard errors (Step 2 using PROC SURVEYREG) and the variance further adjusted by PROC MIANALYZE. The parameter estimates and other statistics are mean values from the 5 imputed values and are now adjusted correctly to account for the multiple imputation and the complex survey features.

EXAMPLE 3: TWO STEP PROCESS - IMPUTE SUFFICIENT MISSING DATA TO PRODUCE MONOTONE MISSING DATA AND USE OF LOGISTIC REGRESSION TO IMPUTE REMAINING MISSING DATA

The final example uses a two step process to impute enough missing data to produce a monotone missing data pattern and then employs a subsequent imputation using logistic regression for imputation of remaining missing data (with categorical variables). The two step process is needed because current imputation methods in PROC MI for imputing categorical variables require monotone missing data. See the PROC MI documentation for additional details.

Step1a: Initial examination of the missing data pattern (Table 3.1) illustrates how even with re-ordering variables, it would not be possible to produce a monotone missing data pattern. This pattern is needed due to the categorical nature of the variables with missing data (OBESE6CAT is a 6 category obesity status variable and WKSTAT3C is a 3 category variable representing work status). In order to use a correct imputation method for categorical variables (logistic for ordinal or binary variable or discriminant for nominal variables) a monotone missing data pattern is required. This example will treat WKSTAT3C and OBESE6CA as ordinal variables and use a logistic regression method once the monotone missing data pattern is achieved.

```
proc mi nimpute=0 data=one ;
  var age hhinc sex region ed4cat wkstat3c obese6ca ; run ;
```

Table 3.1 Missing Data Patterns								
Group	**AGE**	**HHINC**	**SEX**	**REGION**	**ED4CAT**	**WKSTAT3C**	**OBESE6CA**	**Freq**	**Percent**
1	X	X	X	X	X	X	X	5581	98.05
2	X	X	X	X	X	X	.	98	1.72
3	X	X	X	X	X	.	X	13	0.23

The syntax below imputes 5 data sets with enough data values to produce a monotone missing data pattern (**mcmc impute=monotone;**). Note that imputed values are rounded and bounded during imputation so that imputed values match the format of observed values as well as existing ranges. The values in the round, min, and max statements correspond to the variables listed in the VAR statement.

```
proc mi nimpute=5 data=one out=outimputex3 seed=33
  round=1  1      1 1 1 1 1
  min=  18 0      1 1 1 1 1
  max=  98 2000000 2 4 4 3 6 ;
  mcmc impute=monotone;
  var age hhinc sex region ed4cat wkstat3c obese6ca ;
run ;
```

Table 3.2 Missing Data Patterns

Group	AGE	HHINC	SEX	REGION	ED4CAT	WKSTAT3C	OBESE6CA	Freq	Percent
1	X	X	X	X	X	X	X	5581	98.05
2	X	X	X	X	X	X	O	98	1.72
3	X	X	X	X	X	.	X	13	0.23

Table 3.2 uses the usual "X" to show full data, a "O" to show missing that will not be imputed during the initial imputation and the usual "." to indicate cases that will be imputed in the first step.

Step1b: Step1b demonstrates how to execute the second imputation using the data set produced in the first imputation, "outimputex" and output the final imputed data set called "outimputex3full". Because Step1a produced m=5 imputations to produce monotone missing data, only 1 imputation is done in Step1b to complete the full imputation. This results in 5 imputed data sets for subsequent analysis.

Use of the MONOTONE LOGISTIC statements requests that PROC MI perform two separate logistic regressions to first impute WKSTAT3C and then use WKSTAT3C in the imputation of OBESE6CA.

```
proc mi nimpute=1 data=outimputex3 seed=333 out=outimputex3full ;
  class sex region ed4cat wkstat3c obese6ca ;
  var age hhinc sex region ed4cat wkstat3c obese6ca ;
  monotone logistic (wkstat3c=age hhinc sex region ed4cat) ;
  monotone logistic (obese6ca=age hhinc sex region ed4cat wkstat3c) ;
run ;
```

Table 3.3 Monotone Model Specification

Method	Imputed Variables
Logistic Regression	WKSTAT3C OBESE6CA

The output in Table 3.3 indicates that two variables were imputed, WKSTAT3C and OBESE6CA.

Step 2: Logistic Regression Analysis using PROC SURVEYLOGISTIC
The next section of code uses PROC SURVEYLOGISTIC with continuous and class variables to predict a binary outcome of being obese (created from the imputed OBESE6CA variable from Steps 1a and 1b) predicted by age, sex, region and education. The use of the domain statement and the _IMPUTATION_ variable performs a logistic regression for each of the imputed data sets and then saves the output in "ex3est", produced by the ods output statement.

```
data outimputex3full1 ;
  set outimputex3full ;
  if obese6ca > 3 then obese=1 ; else obese=0 ;
run ;
```

```
ods output parameterestimates=ex3est (where=(_Imputation_ ne .))  ;
  proc surveylogistic data=outimputex3full1 ;
  strata sestrat ;  cluster seclustr ; weight ncsrwtlg ;
  class sex region ed4cat / param=ref ;
  model obese (event='1')  = age sex region ed4cat ;
  domain _imputation_ ;
  format sex sexfor. region regionf. ed4cat ed4catf.  ;
run ;

proc print data=ex3est ;
run ;
```

				Table 3.4 (Partial Output)					
Obs	Variable	ClassVal0	DF	Estimate	StdErr	WaldChiSq	ProbChiSq	_Imputation_	Domain
1	Intercept		1	-1.7484	0.1894	85.1965	<.0001	1	Imputation Number=1
2	AGE		1	0.00337	0.00257	1.7149	0.1904	1	Imputation Number=1
3	SEX	Female	1	-0.0220	0.1092	0.0407	0.8401	1	Imputation Number=1
4	REGION	MW	1	0.2558	0.1208	4.4839	0.0342	1	Imputation Number=1
5	REGION	NE	1	0.2160	0.1584	1.8585	0.1728	1	Imputation Number=1
6	REGION	S	1	0.4204	0.1334	9.9396	0.0016	1	Imputation Number=1
7	ED4CAT	0-11	1	0.3348	0.1432	5.4683	0.0194	1	Imputation Number=1
8	ED4CAT	12	1	0.4107	0.0968	18.0209	<.0001	1	Imputation Number=1
9	ED4CAT	13-15	1	0.3393	0.1050	10.4493	0.0012	1	Imputation Number=1

Table 3.4 illustrates the output from the "ex3est" data set for one of the five imputed data sets. This output is then used in PROC MIANALYZE to account for the imputation variability.

Step 3: PROC MIANALYZE
The final step utilizes PROC MIANALYZE to synthesize the results from Steps 1 and 2 to fully incorporate the variance adjustments from PROC SURVEYLOGISTIC and PROC MI. Use of the "parms(classvar=classval) option specifies use of the values of the class variables included in the output data set from step 2.

```
proc mianalyze parms(classvar=classval)=ex3est  ;
  class sex region ed4cat ;
  modeleffects intercept age sex ed4cat region ;
run  ;
```

				Table 3.5 Parameter Estimates								
Parameter	sex	region	ed4cat	Estimate	Std Error	95% Confidence Limits		Pr > |t|	DF	Minimum	Maximum
Intercept				-1.786459	0.193196	-2.16521	-1.40791	<.0001	4701.9	-1.815581	-1.748364
Age				0.003660	0.002612	-0.00146	0.00878	0.1609	15006	0.003370	0.004081
Sex	Female			-0.014200	0.110491	-0.23081	0.20229	0.8977	4810.7	-0.030979	0.011859
ed4cat		0-11		0.351483	0.140605	0.07590	0.62691	0.0124	73209	0.334786	0.363908
ed4cat		12		0.431743	0.094373	0.24667	0.61671	<.0001	2221.5	0.410726	0.455701
ed4cat		13-15		0.350312	0.104771	0.14487	0.55564	0.0008	2614.5	0.328333	0.373181
Region			MW	0.279029	0.122934	0.03806	0.51986	0.0232	13513	0.255843	0.294675
Region			NE	0.225320	0.157928	-0.08421	0.53468	0.1534	958609	0.215963	0.232987
Region			S	0.422088	0.134431	0.15860	0.68543	0.0017	28870	0.401129	0.437654

Table 3.5 shows significant and positive estimates for education levels 0-11 yrs., 12 yrs., and 13-15 yrs., compared to the referent category of 16+ yrs. in predicting being obese. Those in the Midwest and South regions have significantly positive

estimates predicting being obese, as compared to those in the West region. Age and gender are both non-significant in predicting being obese.

CONCLUSION
The focus of this paper is to provide the data analyst with practical guidance on use of a variety of features in the SAS® v9.2 multiple imputation and Survey procedures. Typical examples of how to use PROC MI and PROC MIANALYZE combined with SAS Survey procedures for imputation and analysis of imputed complex survey data are demonstrated and discussed.   The examples utilize both continuous and categorical variables and demonstrate use of common options such as seed values, min and max and round. These simple examples can be generalized and expanded to perform more complex imputations and analysis of complex sample survey data sets.

REFERENCES

Berglund, P. (2008) "Getting the Most out of the SAS® Survey Procedures: Repeated Replication Methods, Subpopulation Analysis, and Missing Data Options in SAS® v9.2", SAS Global Forum 2008.

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

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

Rubin, D.B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91: 473-489.

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

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Yuan, Y.C. "Multiple Imputation for Missing Data: Concepts and New Development." *Proceedings of the SAS Users Group International Conference*. Paper 267-25.

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