Paper SD-12

Multiple Imputation for Ordinal Variables: A Comparison of SUDAAN PROC IMPUTE and SAS® PROC MI

Kimberly Ault, RTI International, RTP, NC

ABSTRACT

This paper presents an outline for the process of performing multiple imputation for ordinal variables using two imputation methods – weighted sequential hot deck imputation (WSHD) and logistic regression when data can be described as having a monotone missing data pattern. Three examples are presented that describe the steps in performing multiple imputation in a sequential approach using the MI procedure in SAS and the IMPUTE procedure in SUDAAN. After multiple imputations are completed, the SURVEYMEANS procedure is used to compute estimates and variances for the ordinal variables. Finally, the MIANALYZE procedure is used to analyze the multiply imputed data.

INTRODUCTION

The IMPUTE procedure in SUDAAN version 11 and the MI procedure is SAS both produce multiply imputed data using different methods. The MI procedure in SAS performs multiple imputation for all types of variables. However, there are some restrictions based on the type of variables to be imputed. For ordinal variables, PROC MI performs logistic regression imputation for monotone missing data patterns. For data that does not have a monotone missing pattern, variables can be imputed sequentially where subsequent imputations are based on the imputed values of preceding variables. Similarly, PROC IMPUTE performs multiple imputation using weighted sequential hot deck imputation that is a non-model based method that can be used for all types of variables – binary, ordinal, nominal, and continuous- without imposing restrictions on the missing data patterns.

First, an example using PROC MI is presented that outlines the steps for performing multiple imputation for a set of ordinal variables. Following this example, two additional examples demonstrate how to perform multiple imputation on ordinal variables using PROC IMPUTE. Of the two examples using PROC IMPUTE, the first example demonstrates a univariate approach where prior imputed variables are used to impute subsequent variables in a monotone missing pattern. The second example shows a multivariate approach to demonstrate the ease of performing imputation without requiring the data to be monotone missing. After the imputations have been completed, means and standard deviations are computed from the multiply imputed using PROC SURVEYMEANS. Finally, a comparison between the point and variance estimates for each example is presented.

MULTIPLE IMPUTATION AND MISSING DATA PATTERNS

There are three general types of missing mechanisms as described by Little and Rubin (2002): Missing at Random (MAR), Missing Completely at Random (MCAR), and Not Missing at Random (NMAR). When the missing data does not depend on the unobserved data but depends only on observed data, then the missing data mechanism is known as MAR. Missing Completely at Random (MCAR) is defined when the missingness does not depend on observed data. Not Missing at Random (NMAR) is defined as when the missingness depends on both the observed data and missing data. When the missing mechanism is MCAR or MAR, we refer to this as ignorable nonresponse. When the mechanism is defined as NMAR, this is referred to as non-ignorable nonresponse.

As defined in the survey research perspective, imputation is the process of filling in missing survey responses to allow standard analysis techniques to be performed on the imputed data. Ideally, the analysis of imputed data should take into account that there is a greater degree of uncertainty than if the imputed values had actually been observed. Unfortunately, in common practice, this issue is often ignored. Imputation techniques are typically used to allow standard analysis techniques to be performed while, if assumptions hold true, reducing nonresponse bias in parameter estimates. Conventional single imputation methods, such as mean or regression imputation and some hot-deck methods, has been shown to underestimate standard errors, which affect confidence intervals and statistical tests (Little and Rubin,2002). Single imputation methods do not reflect the additional uncertainty due to imputing for missing data by treating the imputed values as if they were true values in variance estimation. Thus, multiple imputation (MI) methods have been suggested to help correct for the additional variance due to imputation (Rubin, 1987). The majority of imputation methods (implicitly) assume that the data are MAR. PROC IMPUTE and PROC MI both assume that the data are MAR.

Before performing any imputation method, it is important to examine the patterns of missing to help inform the approach for imputation. Missing data mostly occur in an arbitrary pattern that makes it difficult to perform imputation

using models that require complete cases. However, if data can be arranged such that order exists for variables $x_{[1]}, \ldots, x_{[k]}$ where at least one covariate $x_{[1]}$ has no missing and each $x_{[j]}$ has missing then $x_{[1]}, \ldots, x_{[j-1]}$ are not missing then this is called a monotone missing data pattern. Table 1 shows examples of monotone and arbitrary missing patterns.

Monotone Missing Pattern					Arbitrary Missing Pattern				
Pattern	**X_1**	**X_2**	**X_3**	**X_4**	**Pattern**	**X_1**	**X_2**	**X_3**	**X_4**
1	*x*	*x*	*x*	*x*	***1***	*x*	*x*	*x*	*x*
2	*x*	*x*	*x*	.	***2***	*x*	.	*x*	.
3	*x*	*x*	.	.	***3***	.	*x*	.	*x*
4	*x*	.	.	.	***4***	*x*	.	.	*x*

Table 1. Missing Data Patterns

A key advantage of imputing data in a monotone missing data patterns is computational efficiency. When the missing data pattern is monotone, performing model-based imputation where each subsequent model uses additional variables to help improve the final imputed values is easily done with the MI procedure in SAS. However, if the variables selected for imputation can't be rearranged to produce a monotone missing pattern, then using a non-model based method, such as WSHD, can be implemented easily and quickly using the IMPUTE procedure in SUDAAN. Depending on the amount of missing data, a simple method may be appropriate.

DATA SOURCE

This paper uses data from the 2004 National Health Interview Survey (NHIS) public use data file (NCHS, 2005) which provides data on various health conditions and health care access for the civilian, noninstitutionalized United States population. Data from the NHIS are weighted to account for different selection probabilities and nonresponse and post stratified to census control totals to provide estimates for the U.S. civilian, noninstitutionalized population. Further description of the NHIS design and methodology are available on the Center for Disease Control website.

The examples presented in this paper use six variables from the 2004 NHIS data file including: (1) age in years (6-level) ; (2) gender; (3) race/ethnicity; (4) body mass index; (5) summed chronic health condition indicator (heart disease, stroke, cancer, lung disease, diabetes, and hypertension) and (6) a five-category measure for serious psychological distress.

The variable names and missing patterns for the six variables are as follows:

- AGEGRP: Age at Interview has no missing
- SEX: Gender has no missing
- RACEHISP: Race/Hispanicity has no missing
- BMICAT: Body Mass Index has missing
- CHRONIC: Number of Chronic Health Conditions has missing
- SPD5: Serious Psychological Distress has missing

The three variables with missing values (BMICAT, CHRONIC, and SPD5) are all ordinal variables that are defined as categorical variables that assume *r* different values where order is implied in the levels of the variable. The three variables requiring imputation will be imputed by modeling the relationships between the three variables without missing data and each of the three variables with missing values.

MULTIPLE IMPUTATION WITH PROC MI - EXAMPLE 1

Example 1 provides the steps to use when performing multiple imputation for ordinal variables using PROC MI. In order to use a correct imputation method for categorical variables, PROC MI requires a monotone missing data pattern. For ordinal variables, the PROC MI uses a cumulative logistic regression model. For data with monotone missing patterns, the variables with missing values can be imputed sequentially with variables constructed from their corresponding sets of preceding variables.

Step1: Determine the pattern of missing data by using the NIMPUTE=0 option.

```
proc mi data=in nimpute=0;
   var agegrp sex racehisp bmicat chronic spd5;
run;
```

Missing Data Patterns

Group	AGEGRP	SEX	RACEHISP	BMICAT	CHRONIC	SPD5	Freq	Percent
1	x	x	x	x	x	x	27657	94.74
2	x	x	x	x	x	.	324	1.11
3	x	x	x	x	.	x	175	0.60
4	x	x	x	x	.	.	31	0.11
5	x	x	x	.	x	x	901	3.09
6	x	x	x	.	x	.	77	0.26
7	x	x	x	.	.	x	17	0.06
8	x	x	x	.	.	.	11	0.04

Output 1. Missing Data Patterns from PROC MI

Output 1 shows the missing data pattern for the six variables where "x" denotes a nonmissing value and SAS missing (.) denotes missing values. Three variables have missing data – BMICAT (1006=901+77+17+11 missing values), CHRONIC (234=175+31+17+11 missing values), and SPD5 (443=324+31+77+11). The missing data pattern is not monotone and even if the variables were re-ordered, it would not be possible to produce a monotone missing data pattern. Thus, a sequential approach to imputing the variables is needed to produce a monotone missing data pattern.

Step 2: Impute in sequential process. Let $j = 1, \ldots, k$ denote the columns of the variables in the dataset and then predict $x_{[j]}$ from $x_{[1]}, \ldots, x_{[j-1]}$. The sequential approach creates monotone patterns where the following assumptions are made:

- Age, race, and gender predicts body mass index
- Body mass index, age, race, and gender predicts number of chronic health conditions
- Number of chronic health conditions, body mass index, age, race, and gender predicts serious psychological distress

The following code shows the steps for performing the sequential imputation:

Step 2a: Model relation between AGEGRP, SEX, RACEHISP, and BMICAT and then impute BMICAT.

The strategy used in PROC MI for imputing monotone missing data for ordinal variables include estimating the probability for each category using an cumulative logistic regression model and then imputes each category for missing values based on estimated probabilities. With the MONOTONE statement, the missing values of a variable are imputed sequentially in the order specified in the VAR statement. For example, the following PROC MI statements use the cumulative logit model to impute variable BMICAT from effects AGEGRP, SEX, and RACEHISP (since BMICAT is the only variable that has missing data). The CLASS statement specifies the categorical variables in the VAR statement.

```
proc mi data=in nimpute=5 seed=51343673 out=mi1;
   class agegrp sex racehisp bmicat;
   var agegrp sex racehisp bmicat;
   monotone logistic (bmicat = agegrp sex racehisp);
run;
```

Step 2b: Model relation between AGEGRP, SEX, RACEHISP, BMICAT, and CHRONIC and then impute CHRONIC.

```
proc mi data=mi1 nimpute=1 seed=51343672 out=mi2;
   class agegrp sex racehisp bmicat chronic;
   var agegrp sex racehisp bmicat chronic;
   monotone logistic (chronic=agegrp sex racehisp bmicat);
run;
```

Step 2b demonstrates how to execute the second imputation using the data set produced in the first imputation, "MI1" and output the imputed data set called "MI2". Because Step2a produced m=5 imputations to produce monotone missing data, only one imputation is done in Step 2b to complete the full sequence. This results in five imputed data sets for the next step.

Step 2c: Model relation between AGEGRP, SEX, RACEHISP, BMICAT, CHRONIC and SPD5 and then impute SPD5.

```
proc mi data=mi2 nimpute=1 seed=51343671 out=mi3;
   class agegrp sex racehisp bmicat chronic spd5;
   var agegrp sex racehisp bmicat chronic spd5;
   monotone logistic (spd5=agegrp sex racehisp bmicat chronic);
run;
```

Output 2 shows the results from PROC MI for Step 2c. The "Monotone Model Specification" table shows that logistic regression method was used to impute SPD5 from the five other variables. The "Missing Data Patterns" table lists distinct missing data patterns with corresponding frequencies and percentages. The number of cases with missing data for SPD5 is 5 times as large as the original missing rate (443*5= 2215, see Output 1) since the input dataset contains five set of imputations based on Steps 2a and 2b. The total number of observations in the final “MI3”dataset is 145,965 = 29193 * 5 and contains all of the imputed data and a variable called _IMPUTATION_ that denotes the imputation set. This data will be used in a later section to produce means and standard deviations based on the multiply imputed data.

The MI Procedure

Monotone Model Specification

Method	Imputed Variables
Logistic Regression	SPD5

Missing Data Patterns

Group	AGEGRP	SEX	RACEHISP	BMICAT	CHRONIC	SPD5	Freq	Percent
1	x	x	x	x	x	x	143750	98.48
2	x	x	x	x	x	.	2215	1.52

Output 2. Output from PROC MI Statement for Step 2c

PERFORMING MULTIPLE IMPUTATION WITH PROC IMPUTE

In version 10 of SUDAAN, PROC HOTDECK performs weighted sequential hot deck (WSHD) imputation for item nonresponse as described in Cox (1980) and Iannacchione (1982). In version 11 of SUDAAN, PROC HOTDECK is replaced with PROC IMPUTE[1], which is an enhanced version of the PROC HOTDECK. PROC IMPUTE includes the following imputation methods: WSHD, cell mean imputation, linear regression for continuous variables, and logistic regression for binary variables. PROC IMPUTE performs both multivariate imputations (i.e., multiple variables imputed at the same time) as well as multiple imputations (i.e., multiple imputed versions of the same variable) for the WSHD method. However, the multiple imputation option is currently only available for the WSHD method. Multiple imputation could be performed with the other methods by repeatedly apply the chosen method. The imputation methodology used in IMPUTE assumes that, for a given variable with missing data, the missing-data mechanism within each imputation class is MAR.

Sequential hot-deck imputation is a common method used for item nonresponse in survey research. This method uses the respondent survey data (donors) to provide imputed values for records with missing values by defining imputation classes, which generally consist of a cross-classification of covariates, and then replacing missing values sequentially from a single pass through the survey data within the imputation classes. When sequential hot-deck imputation is performed using the sampling weights of the item respondents and nonrespondents, the method is called *weighted* sequential hot-deck imputation (WSHD). This method takes into account the unequal probabilities of selection in the original sample by using the sampling weight to specify the expected number of times a particular respondent's answer is used to replace a missing item. Selection frequencies are specified so that, over repeated applications of the algorithm, the expected value of the weighted distribution of the imputed values will equal the weighted distribution of the reported answers within imputation class. An advantage of WSHD imputation is that it controls the number of times a respondent record can be used for imputation and gives each respondent record a chance to be selected for use as a hot-deck donor.

EXAMPLE 2 - PERFORMING MULTIPLE IMPUTATION WITH PROC IMPUTE FOR MONOTONE MISSING DATA PATTERNS

The SUDAAN code shown below performs multiple imputation for the three ordinal variables (BMICAT, CHRONIC, and SPD5) using a sequential approach where prior imputed variables were used to impute subsequent variables in a monotone missing pattern.

[1] Programs using the key words “PROC HOTDECK” created in SUDAAN 10 will run without change in SUDAAN 11.

Step 1: Perform multiple imputation for BMICAT using WSHD imputation and imputation classes AGEGRP, SEX, and RACEHISP.

```
/* PROC IMPUTE - WSHD - Sequential */
proc impute data=in filetype=sas method=wshd notsorted;
   weight sa_wgt_new;
   impby agegrp sex racehisp;
   impvar bmicat/ multimp=5;
   impname bmicat="ibmicat";
   impid numpublicid;
   output / impute=default filename=wshd_bmi;
```

The option METHOD=WSHD indicates that weighted sequential hot deck imputation will be performed. Because the input data is not sorted by the imputation classes, the NOTSORTED option was used on the procedure statement. When the NOTSORTED is specified, then SUDAAN will perform the sorting internally prior to completing the imputation process. The WEIGHT statement specifies the sample weight (SA_WGT_NEW) to be used in the WSHD algorithm.

The IMPBY statement identifies the variables that make up the imputation classes (AGEGRP, SEX, and RACEHISP) and the variables are treated as categorical. The actual imputation classes are the cross-classification of the variables listed. The IMPVAR statement requests the variable to be imputed – BMICAT. Within each imputation class, item respondents (potential donor records) are identified. For this example, donor records will be those records that have a valid response for BMICAT. Additionally, the IMPVAR statement shows that a multiple imputation will be performed as noted by the MULTIMP=5 option indicating that WSHD imputation will be performed 5 times. The output dataset will include the results of all five of the multiple imputations. In the case of multiple imputation, an item nonrespondent record may receive values from different donors for each round of multiple imputation.

The IMPNAME statement defines the names for variables that will hold the imputed values in the output dataset after imputation. The output dataset will contain five imputed versions of the BMICAT variable. The imputed versions of the BMICAT variable will be labeled as follows: IBMICAT_1, IBMICAT_2, IBMICAT_3, IBMICAT_4, and IBMICAT_5. The IMPID statement identifies the variable that holds the record identifier on the input dataset. The OUTPUT statement tells IMPUTE create an output file labeled “WSHD_BMI” and the IMPUTE=DEFAULT option requests that all variables from the input dataset be included on the output dataset.

Output 3 below shows the summary statistics that are generated from PROC IMPUTE. It states that the procedure “completed successfully” and that the WSHD method was used. The number of respondent records (28,187) and the number of nonrespondent records (1,006) are presented. It shows that five imputations were performed as noted by the “Number of Donor Records for Imputation #1 to #5”.

```
                              S U D A A N
            Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute         August 2012
                                Release 11.0.0

The Impute Process has completed successfully.
Method: WSHD

Total Records Read from File: 29193
Random Number Seed: 56237485

Total Respondent Records: 28187

Total Nonrespondent Records: 1006
   Total Donor Records Imputation #1: 1006
   Total Donor Records Imputation #2: 1006
   Total Donor Records Imputation #3: 1006
   Total Donor Records Imputation #4: 1006
   Total Donor Records Imputation #5: 1006

   Total Records Imputed: 1006
```

Output 3. Default Summary Statistics from PROC IMPUTE for BMICAT

Step 2: Perform multiple imputation for CHRONIC using WSHD imputation and imputation classes AGEGRP, SEX, and RACEHISP. This sequential approach ensures that the data is in a monotone missing data pattern. The output dataset WSHD_BMI will be used as the input dataset for the second step. The SUDAAN code is similar to the first step with the exception of the variable being imputed.

```
proc impute data=wshd_bmi filetype=sas method=wshd notsorted;
   weight sa_wgt_new;
   impby agegrp sex racehisp;
   impvar chronic/ multimp=5;
   impname chronic="ichronic";
   impid numpublicid;
   output / impute=default filename=wshd_chronic;
```

Output 4 shows the results of the IMPUTE run. There were 234 missing values for CHRONIC and all missing values were imputed as noted by the “Total Records Imputed: 234” statement. Similar to Step 1, five imputations were performed.

```
                              S U D A A N
            Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute         August 2012
                                 Release 11.0.0

The Impute Process has completed successfully.
Method: WSHD

Total Records Read from File: 29193
Random Number Seed: 56237485

Total Respondent Records: 28959

Total Nonrespondent Records: 234
   Total Donor Records Imputation #1: 234
   Total Donor Records Imputation #2: 234
   Total Donor Records Imputation #3: 234
   Total Donor Records Imputation #4: 234
   Total Donor Records Imputation #5: 234

   Total Records Imputed: 234
```

Output 4. Default Summary Statistics from PROC IMPUTE for CHRONIC

Step 3. Perform multiple imputation for SPD5 using WSHD imputation and imputation classes AGEGRP, SEX, and RACEHISP. The input dataset for this step contains the five different versions of the imputed variables for BMICAT (IBMICAT_1 – IBMICAT_5) and CHRONIC (ICHRONIC_1 – ICHRONIC_5). This last step completes the multiple imputation for the three variables.

```
proc impute data=wshd_chronic filetype=sas method=wshd notsorted;
   weight sa_wgt_new;
   impby agegrp sex racehisp;
   impvar spd5/ multimp=5;
   impname spd5="ispd5";
   impid numpublicid;
   print / donorstat=default;
   output / impute=default filename=wshd_spd;
run;
```

Output 5 shows the default summary statistics for the last step in the sequential approach for performing multiple imputation for monotone missing data patterns. There were 443 missing values for SPD5 and all missing values were imputed within the 14 imputation classes based on AGEGRP, SEX, and RACEHISP.

```
                              S U D A A N
          Software for the Statistical Analysis of Correlated Data
        Copyright      Research Triangle Institute         August 2012
                              Release 11.0.0

The Impute Process has completed successfully.
Method: WSHD

Total Records Read from File: 29193
Random Number Seed: 56237485

Total Respondent Records: 28750

Total Nonrespondent Records: 443
   Total Donor Records Imputation #1: 443
   Total Donor Records Imputation #2: 443
   Total Donor Records Imputation #3: 443
   Total Donor Records Imputation #4: 443
   Total Donor Records Imputation #5: 443

   Total Records Imputed: 443
```

Output 5. Default Summary Statistics from PROC IMPUTE for SPD5

The next example shows more details of the output provided by PROC IMPUTE and demonstrates how to use IMPUTE to perform multivariate imputations (imputing more than one variable at time) and ease of performing this method as compared to the sequential approach that accommodate monotone missing data.

EXAMPLE 3 - PERFORMING MULTIVARIATE WSHD IMPUTATION WITH PROC IMPUTE

The SUDAAN code shown below performs multivariate WSHD imputation for the three variables with missing data: BMICAT, CHRONIC, and SPD5. The variables are imputed in a multivariate fashion where all three variables receive imputed values from the same donor record. The donor record contributes only values for missing values and will not replace any valid values for the variables being imputed. In other words, variables that are not missing on the item nonrespondent record are not replaced by the donor's response.

```
proc impute data=in filetype=sas method=wshd notsorted;
   weight sa_wgt_new;
    class bmicat chronic spd5;
   impby agegrp sex racehisp;
   impvar bmicat chronic spd5/ multimp=5;
   impname bmicat="ibmicat" chronic="ichronic" spd5="ispd5";
   impid numpublicid;
   print / donorstat=default percents=all;
   output / impute=default filename=wshd;
run;
```

The sample code is similar to the code provided in Example 2 with the exception being that all three variables requiring imputation are specified on the IMPVAR statement. For multivariate imputations (i.e. multiple variables on IMPVAR statement), the default set of records considered "respondents" are those in which each IMPVAR variable is non-missing; and "non-respondents" are those in which at least one IMPVAR variable is missing.

For this example, "respondents" will be those cases with all three variables (BMICAT, CHRONIC, and SPD5) not missing. As in the prior example, five imputations will be performed as noted by the MULTIMP=5 option. The output dataset will contain five imputed versions of each of the three variables being imputed. The imputed versions of the BMICAT variable will be labeled as follows: IBMICAT_1, IBMICAT_2, IBMICAT_3, IBMICAT_4, IBMICAT_5. Similarly, ICHRONIC_1 – ICHRONIC_5 and ISPD5_1 – ISPD5_5 will be created.

The PRINT statement includes the PERCENTS=ALL option which will print summary statistics for categorical imputed variables as defined by a CLASS statement. The PRINT statement requires that the categorical variables be defined by a CLASS statement.

Output 6 displays the number of respondent records (27,657) and the number of nonrespondent records (1,536). The total number of nonrespondent records is for the multivariate missing data patterns is less than the total number of nonrespondents for the three monotone missing data pattern steps (1,006 + 234 +443=1,683) presented in Example 2. This is due to some records having missing values for two or more of the variables.

```
                                S U D A A N
            Software for the Statistical Analysis of Correlated Data
          Copyright      Research Triangle Institute         August 2012
                                 Release 11.0.0

The Impute Process has completed successfully.
Method: WSHD

Total Records Read from File: 29193
Random Number Seed: 56237485

Total Respondent Records: 27657

Total Nonrespondent Records: 1536
   Total Donor Records Imputation #1: 1536
   Total Donor Records Imputation #2: 1536
   Total Donor Records Imputation #3: 1536
   Total Donor Records Imputation #4: 1536
   Total Donor Records Imputation #5: 1536

   Total Records Imputed: 1536
```

Output 6. Results from PROC IMPUTE for multivariate imputation for BMICAT, CHRONIC, and SPD5

Output 7 shows a portion of the IMPUTE output from the DONORSTAT print group. For Imputation #1, the donor statistics shown for the imputation class for AGEGRP=1. Each row lists the number of records considered respondents and nonrespondents within the imputation class, the number of donors contributing to imputation (Donor Count), and the number of records with missing data post-imputation. For the imputation class where GENDER=1 and RACEHISP=1, there are 890 respondent records, 51 nonrespondent records (missing either BMICAT, CHRONIC, or SPD5), 51 donors contributing data, and no records with missing BMICAT, CHRONIC, or SPD5 after imputation.

```
Hot- Deck form Imputing BMICAT, CHRONIC, and SPD5
Imputation #1
by: AGEGRP, SEX, RACEHISP
for: AGEGRP=1
                                Item            Item
                           Respondent      Non-Respondent      Donor       Missing Data -
                              Count           Count          Count       Post Imputation
SEX = 1
RACEHISP = 1                    890              51             51              0
RACEHISP = 2                   2098              57             57              0
RACEHISP = 3                    470              20             20              0
RACEHISP = 4                    174               8              8              0

SEX = 2
RACEHISP = 1                   1099              81             81              0
RACEHISP = 2                   2327             100            100              0
RACEHISP = 3                    745              26             26              0
RACEHISP = 4                    166               8              8              0
```

Output 7. IMPUTE Results for WSHD Imputation: DONORSTAT PRINT Group

Output 8 shows the results generated from the PRINT statement used the PERCENTS=ALL option. For each imputation class, PROC IMPUTE computes the weighted percentage in each level before and after imputation for each variable that was imputed, as well as the absolute difference of the pre-imputation and post-imputation percentages and the relative percent difference for each variable imputed. For example, for the AGEGRP=1, SEX=1 and RACEHISP=1 imputation class, the weighted percentage for BMICAT=1 before imputation is 3.47% and after imputation is 3.42%. The absolute difference is 0.05% and the relative difference before and after imputation is the 1.5%. When the PERCENTS=ALL options is used on the PRINT statement, PROC IMPUTE prints these statistics for all imputation classes and variables being imputed.

Method = WSHD
BMI – Categorical
by: AGEGRP, SEX, RACEHISP, BMICAT
for: AGEGRP = 1, SEX = 1

	% Pre-Imputation	% Post Imputation	Absolute Difference	Relative Difference (%)
RACEHISP = 1				
BMICAT = 1	3.47	3.42	0.05	-1.50
BMICAT = 2	31.68	31.54	0.14	-0.43
BMICAT = 3	44.04	43.70	0.34	-0.78
BMICAT = 4	20.81	21.34	0.53	2.54
RACEHISP = 2				
BMICAT = 1	5.95	6.04	0.09	1.57
BMICAT = 2	37.28	37.23	0.05	-0.13
BMICAT = 3	36.89	36.86	0.03	-0.08
BMICAT = 4	19.88	19.87	0.02	-0.08
RACEHISP = 3				
BMICAT = 1	5.33	5.31	0.03	-0.50
BMICAT = 2	33.54	32.42	1.12	-3.33
BMICAT = 3	34.36	34.77	0.41	1.2
BMICAT = 4	26.77	27.51	0.73	2.73
RACEHISP = 4				
BMICAT = 1	9.77	9.68	0.09	-0.93
BMICAT = 2	49.78	49.31	0.46	-0.93
BMICAT = 3	32.51	33.14	0.63	1.94
BMICAT = 4	7.94	7.87	0.07	-0.93

Output 8. IMPUTE Results: PERCENTS PRINT Group for WSHD Imputation

COMPARING MULTIPLE IMPUTATION RESULTS

This section describes how to analyze the multiple imputed data and compares the results from the three examples. PROC SURVEYMEANS provides estimates based on survey data. PROC MIANALYZE combines the estimates obtained from the SURVEYMEANS for multiply imputed data to produce valid statistical inferences. For the examples presented in this paper, the mean and standard errors are computed for the multiply imputed data.

ANALYZING SURVEY DATA

After completing the multiple imputations, the means and standard deviations are computed using PROC SURVEYMEANS since the data is based on a sample survey. The sample code below computes the means and standard errors for the multiply imputed data from PROC MI for Example 1. For the NHIS data, a stratified sample survey design is used and this procedure will produce standard errors of the sample means that reflect the complex sample design. The STRATA statement is used to indicate the stratification variable (STRATUM) and the CLUSTER statement is used to identify the first-stage primary sampling units (PSU). The WEIGHT statement indicates the variable containing the sampling weight (SA_WGT_NEW). The CLASS and VAR statements indicate that the means will be computed for categorical variables by each level of the variable. The BY statement requests that the analyses will be performed for each imputation set (_IMPUTATION_). Finally, ODS OUTPUT statement saves the means and standard errors into an output dataset (as noted by MEAN and STD on the PROC statement) that will be used to perform the multiple imputation analysis.

```
proc surveymeans data=mi3 mean std;
   strata stratum;
   cluster psu;
   weight sa_wgt_new;
   class bmicat chronic spd5;
   var bmicat chronic spd5;
   by _imputation_;
   ods output statistics=out;
run;
```

Output 9 shows the partial results from PROC SURVEYMEANS. The means and standard errors for the multiply imputed data from Example 1 are computed for each set of imputations. The output displays the means and standard errors for each level of BMICAT.

IMPUTATION	Variable	Level	Mean	Std Error of Mean
1	BMICAT	1	0.061981	0.001676
		2	0.342066	0.003121
		3	0.352002	0.002802
		4	0.243951	0.002979
2	BMICAT	1	0.062218	0.001688
		2	0.341767	0.003142
		3	0.352360	0.002871
		4	0.243655	0.002987

Output 9. By Imputation Analysis from SURVEYMEANS

For each of the three variables imputed, the mean and standard deviation for each variable level were computed and stored in an output dataset. The output dataset "OUT" is used in the next step of the process – to combine the results from each of the multiply imputed datasets.

ANALYZING MULTIPLY IMPUTED DATA

PROC MIANALYZE combines the results of the analyses of imputations and generates valid statistical inferences. In this example, PROC MIANALYZE reads in the means and standard errors from the five imputed datasets and then creates a total variance estimate that can be used to make confidence statements about the mean estimates.

```
proc mianalyze data=out notsorted;
   by varname varlevel;
   modeleffects mean;
   stderr stderr;
   ods output varianceinfo=mi_var ParameterEstimates=mi_parm;
run;
```

The BY statement is used to obtain separate analyses by the groups (VARNAME and VARLEVEL).The NOTSORTED option is used to let SAS know that the input dataset is not sorted by the variables on the BY statement. The MODELEFFECTS statement lists the estimates (MEAN) to be analyzed. The STDERR statement indicates the standard errors associated with the estimators indicated on the MODELEFFECTS statement. The ODS OUTPUT statement saves the variance information table (VARIANCEINFO) and the parameter estimates (PARAMETERESTIMATES) as output datasets.

Output 10 shows the results from PROC MIANALYZE for the first level of the BMICAT variable. The variance information and parameter estimate tables are presented. The variance information includes the following items: between-imputation variance, within-imputation variance, total variance, degrees of freedom for the total variance, relative increase in variance due to imputation, the fraction of missing data, and the relative efficiency for each imputed variable. The output shows that the five imputed data sets were used in PROC MIANALYZE and presents the variance information. Note a very high relative efficiency with five imputations, indicating even fewer imputations might have been performed. The parameter estimates table shows the mean value for the first level of the BMICAT variable (0.061981), the standard error (0.001750), the 95% confidence interval, and a t-test with the associated p-value for the hypothesis that the parameter is equal to zero.

The MIANALYZE Procedure
Variable Name=BMICAT Variable Level=1

Model Information	
Data Set	WORK.OUT
Number of Imputations	5

Variance Information

Parameter	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
	Between	Within	Total				
Mean	0.000000150	0.000002882	0.000003062	1163	0.062300	0.060261	0.988091

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter =Theta0	Pr > |t|
Mean	0.061981	0.001750	0.058548	0.065414	1163	0.061474	0.062487	0	35.42	<.0001

Output 10. Variance Information and Parameter Estimates from MIANALYZE for BMICAT=1

COMPARING THE RESULTS

Table 2 below shows the results from PROC MIANALYZE of the three examples presented in this paper. The means and standard errors for each level of the three categorical variables after imputation are shown. There are small differences in the estimates among each example. The simplest method for performing multiple imputation is the multivariate imputation using PROC IMPUTE since it does not require as many steps as the sequential approaches presented in both PROC IMPUTE and MI. For this example, the ease of performing this procedure does not comprise the results of the imputed data.

Variable	Level	Example 1 - PROC MI (Sequential)		Example 2 - PROC IMPUTE (Sequential)		Example 3 - PROC IMPUTE (Multivariate)	
		Mean	Standard Error	Mean	Standard Error	Mean	Standard Error
BMICAT	1	0.0620	0.0018	0.0623	0.0017	0.0623	0.0018
BMICAT	2	0.3421	0.0032	0.3428	0.0032	0.3427	0.0032
BMICAT	3	0.3520	0.0029	0.3507	0.0028	0.3508	0.0028
BMICAT	4	0.2440	0.0031	0.2442	0.0031	0.2442	0.0031
CHRONIC	0	0.5934	0.0040	0.5936	0.0040	0.5937	0.004
CHRONIC	1	0.2416	0.0032	0.2415	0.0032	0.2415	0.0032
CHRONIC	2	0.1649	0.0027	0.1650	0.0027	0.1648	0.0027
SPD5	0	0.4891	0.0051	0.4893	0.0051	0.4887	0.0051
SPD5	1	0.2058	0.0031	0.2060	0.0031	0.2061	0.0031
SPD5	2	0.1539	0.0026	0.1537	0.0027	0.1539	0.0027
SPD5	3	0.1001	0.0023	0.0999	0.0023	0.1003	0.0023
SPD5	4	0.0511	0.0016	0.0511	0.0017	0.0510	0.0016

Table 2. Contents of the Parameter Estimates Table from PROC MIANALYZE

CONCLUSION

This paper provides practical guidance for implementing multiple imputation for monotone missing data patterns and offers an alternative approach when data is not monotone. Three examples are presented with detailed steps on how to perform multiple imputation using the SAS MI procedure and the SUDAAN IMPUTE procedure. The examples illustrate how to use analyze the results to accommodate for the complex survey design of the data using the SURVEYMEANS procedure and how to analyze the multiply imputed datasets using the MIANALYZE procedure. These simple examples can be generalized to more complex missing data patterns and different variable types.

REFERENCES

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the 1980 American Statistical Association, Survey Research Methods Section, Houston, TX* (pp. 721-726). Washington, DC: American Statistical Association.

Iannacchione, V. (1982). Weighted sequential hot deck imputation macros. In *Proceedings of the Seventh Annual SAS Users Group International Conference* (pp. 759-763). Cary, NC: SAS Corporation.

Little, R. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Second ed. Hoboken, NJ: John Wiley & Sons, Inc.

National Center for Health Statistics (2005). *Data File Documentation, National Health Interview Survey, 2004 (machine-readable data file and documentation)*. National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, Maryland. (http://www.cdc.gov/nchs/data/nhis/srvydesc.pdf)

RTI International. (2012). *SUDAAN® language manual, Release 11.0*. Research Triangle Park, NC: RTI International.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: John Wiley & Sons, Inc.

SAS Institute Inc. 2011. SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

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

Name: Kimberly Ault
Enterprise: RTI International
Address: 3040 Cornwallis Rd, Cox 230
City, State ZIP: RTP, NC 27709
Work Phone: 919-541-7455
E-mail: ault@rti.org