Paper SAS3520-2016

# Survey Data Imputation with PROC SURVEYIMPUTE

Pushpal K. Mukhopadhyay
SAS Institute Inc., Cary, NC

## ABSTRACT

Big data, small data—but what about when you have no data? Survey data commonly include missing values due to nonresponse. Adjusting for nonresponse in the analysis stage might lead different analysts to use different, and inconsistent, adjustment methods. To provide the same complete data to all the analysts, you can impute the missing values by replacing them with reasonable nonmissing values. Hot-deck imputation, the most commonly used survey imputation method, replaces the missing values of a nonrespondent unit by the observed values of a respondent unit.

In addition to performing traditional cell-based hot-deck imputation, the SURVEYIMPUTE procedure, new in SAS/STAT® 14.1, also performs more modern fully efficient fractional imputation (FEFI). FEFI is a variation of hot-deck imputation in which all potential donors in a cell contribute their values.

Filling in missing values is only a part of PROC SURVEYIMPUTE. The real trick is to perform analyses of the filled-in data that appropriately account for the imputation. PROC SURVEYIMPUTE also creates a set of replicate weights that are adjusted for FEFI. Thus, if you use the imputed data from PROC SURVEYIMPUTE along with the replicate methods in any of the survey analysis procedures—SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, or SURVEYPHREG—you can be confident that inferences account not only for the survey design, but also for the imputation.

This paper discusses different approaches for handling nonresponse in surveys, introduces PROC SURVEYIMPUTE, and demonstrates its use with real-world applications. It also discusses connections with other ways of handling missing values in SAS/STAT.

## INTRODUCTION

Nonresponse in survey data can compromise the quality of survey results. An observation unit (the unit in which a measurement is taken) that has no missing values is called a complete respondent, and an observation unit that contains missing values in some items is called an incomplete respondent. If the complete respondents differ from the incomplete respondents with regard to a survey effect or outcome, then survey estimates will not accurately represent the survey population.

You should plan to prevent nonresponse early in the data collection process. Methods to prevent nonresponse include designing a better questionnaire, changing the mode of the data collection, making several attempts to contact the nonrespondents, assigning a different interviewer, and providing incentives to complete the survey. For more information about preventing nonresponse, see Hidiroglou, Drew, and Gray (1993) and De Leeuw, Hox, and Huisman (2003). Prevention is the best approach for nonresponse.

No matter how hard you try to prevent it, you might still have nonresponse in your data. After data collection is complete, you can use imputation to replace missing values with acceptable values, and you can use sampling weight adjustments to compensate for nonresponse. This paper shows you how to perform imputation for survey data by using the SURVEYIMPUTE procedure, new in SAS/STAT 14.1. For reviews of imputation and weight adjustment methods that are commonly used in practice, see Kalton and Kasprzyk (1986) and Brick and Kalton (1996).

The primary objectives of imputation are to reduce nonresponse bias for the important survey variables and to publish a data set that does not contain any missing values. The nonresponse bias is the expected difference between the quantity that is computed from the set of respondents and the quantity that could be computed from the selected sample if all units in the selected sample could be observed, where the expectation is over all possible sets of respondents. Nonresponse bias generally does not decrease as the sample size increases. For a recent discussion on nonresponse bias, see Brick (2013).

In addition to reducing the nonresponse bias, imputation assures that the same data set is used by all analysts. If you publish your data with missing values, then different analysts might use different methods to adjust for the missing values and therefore obtain different estimated values for the same quantity. As an imputer, you have more information available about the data than the analysts do. If you use all the information available to you to impute the missing values and then publish the complete data without any missing values, all analysts will obtain the same efficient estimate for the same quantity.

The most commonly used imputation methods for survey data replace the missing values for the nonrespondent units by the observed values from one or multiple respondent units. These imputation techniques are known as *hot-deck imputation*. The recipient unit is defined as the observation unit that contains the missing values, and the donor unit is defined as the observation unit that provides the imputed values. For a recent review of hot-deck imputation, see Haziza (2009).

Imputation is only a part of the task. Users of the imputed data will use the data to produce estimates for various population quantities. Most commonly, imputation and analysis are two different tasks that are performed separately by different persons or by different organizations. This paper separates the analysis task from the imputation task and describes the analysis techniques in a separate section.

This paper introduces the SURVEYIMPUTE procedure and uses it to apply different hot-deck imputation methods to data from two national surveys. Examples show you how typical imputation projects can use different techniques that are available in the procedure, or call for running it in multiple steps. Sections of the paper describe the syntax for the new procedure, the imputation methods that are available, two imputation approaches when no donors are available, methods for analyzing imputed data sets, and some advantages and disadvantages for these imputation methods. Finally, the imputation techniques available in PROC SURVEYIMPUTE are summarized, and an appendix contains the SAS code for imputation in multiple steps.

Before the new procedure is introduced, let's review some features that all SAS/STAT survey analysis procedures include to handle missing values. First, you should not have any missing values in the STRATA, CLUSTER, WEIGHT, or REPWEIGHTS variables. Information about these variables are typically available before you begin your data collection. If you have missing values in any of these variables, you must check with your data provider. If you have missing values in the analysis variables, then, by default, SAS/STAT procedures do not use those observations in the analysis. In addition, the SAS/STAT survey procedures support the NOMCAR option in the PROC statement to treat the size of the set of respondents as random in the computation of the Taylor series linearized variance estimation. Thus, the NOMCAR option is equivalent to a domain analysis for the set of respondents. If you have missing values in the categorical variables (which include variables in the CLASS, CLUSTER, STRATA, and DOMAIN statements), then you can use the MISSING option to treat the missing values as a separate category.

In addition to these options, you can also use the MI procedure to impute missing values by using multiple imputation methods. For more information about PROC MI, see the chapter "The MI Procedure" in *SAS/STAT User's Guide*.

Nonresponse adjustments for survey data are extensively discussed in the literature. For example, see Särndal and Lundström (2005); Fuller (2009, Section 5.1 and 5.2); Lohr (2010, Chapter 8); Bethlehem, Cobben, and Schouten (2011); and Kim and Shao (2014). For more information about PROC SURVEYIMPUTE, see the chapter "The SURVEYIMPUTE Procedure" in *SAS/STAT User's Guide*.

## EXAMPLE DATA

This section introduces two public-use data sets from national surveys that are used to illustrate the imputation techniques throughout the paper.

The first data set, **Asthma**, contains 8,602 observation units from the Joint Canada/United States Survey of Health for the year 2003. The data set is similar to the data set used by Ghosh and Pahwa (2008); 2,003 observation units contain missing values in at least one variable. For more information, see `http://www.cdc.gov/nchs/nhis/jcush.htm`.

Table 1 shows the variables in the **Asthma** data set.

**Table 1** Variables in the Asthma Data Set

| Variable | Values | Units That Have Missing Values |
|---|---|---|
| **Age** | 1 if age is 18–34, 2 if age is 35–46, 3 if age is 47–61, 4 if age is 62–85 | 0 |
| **Asthma** | 1 if asthma is reported, 0 otherwise | 6 |
| **BMI** | 1 if underweight, 2 if normal weight, 3 if overweight, 4 if obese | 262 |
| **Birth** | 1 if Canada, 2 if United States, 3 if other | 228 |
| **Education** | 1 if less than high school, 2 if high school, 3 if more than high school | 274 |
| **FWGT** (full-sample weight) | Ranges from 1,113.41 to 235,012.00 | 0 |
| **Gender** | 1 if male, 2 if female | 0 |
| **Income** | 1 if low, 2 if low middle, 3 if high middle, 4 if high | 1,797 |
| **Race** | 1 if white, 2 if other | 278 |
| **Smoker** | 1 if current smoker, 2 if ex-smoker, 3 if never smoked | 39 |

In addition, the data set contains 1,000 bootstrap replicate weights, which are created by using the rescaling bootstrap method (Rao and Wu 1988). Variables **bsw1**–**bsw1000** contain the replicate weights.

The second data set, **ArthritisCell**, contains 16,954 observation units from the 2006 Health and Retirement Study. The **ArthritisCell** data set is similar to the first example data set used in Berglund and Heeringa (2014, Chapter 7); 127 observation units contain missing values in at least one variable. For more information about the survey, see `http://hrsonline.isr.umich.edu/`.

Table 2 shows the variables in the **ArthritisCell** data set.

**Table 2** Variables in the ArthritisCell Data Set

| Variable | Values | Units That Have Missing Values |
|---|---|---|
| **UnitID** (unit identification) | Ranges from 1 to 16,954 | 0 |
| **Stratum** (stratum identification) | Ranges from 1 to 56 | 0 |
| **SECU** (cluster identification) | Either 1 or 2 | 0 |
| **KWgtr** (full-sample weight) | Ranges from 930 to 16,532 | 0 |
| **Gender** | 1 if male, 2 if female | 0 |
| **KAge** (age in years) | Ranges from 36 to 105 | 0 |
| **SelfRHealth** (self-rated health) | 1 if excellent, 2 if very good, 3 if good, 4 if fair, 5 if poor | 24 |
| **Arthritis** | 1 if arthritis reported, 0 otherwise | 13 |
| **Diabetes** | 1 if diabetes is reported, 0 otherwise | 16 |
| **RaceCat** | 1 if Hispanic, 2 if white, 3 if black, 4 if other | 3 |
| **EdCat** | 1 if 0–11 years of education, 2 if 12 years of education, 3 if 13–15 years of education, 4 if 16 or more years of education | 79 |
| **ImpCell** (imputation cell) | Ranges from 1 to 6 | 0 |

## INTRODUCTION TO PROC SURVEYIMPUTE

The SURVEYIMPUTE procedure provides several imputation methods for replacing missing values in an item by observed values from the same item. These imputation methods can be used to impute missing values for survey data. The most useful imputation method available in PROC SURVEYIMPUTE is the fully efficient fractional imputation (FEFI) method, which uses information from all donors to impute missing values in each recipient unit. The primary literature for FEFI is from Kim and Fuller (2004) and Fuller and Kim (2005). When you use FEFI, PROC SURVEYIMPUTE also creates imputation-adjusted replicate weights that can be used in any SAS/STAT survey procedure to compute replication variance estimates. This section provides an overview of the syntax for PROC SURVEYIMPUTE. For more information about the syntax, see the section "Syntax: SURVEYIMPUTE Procedure" in *SAS/STAT User's Guide*.

The syntax in PROC SURVEYIMPUTE for design and replicate weights is similar to that syntax in other SAS/STAT survey procedures. You use the WEIGHT, STRATA, and CLUSTER statements to specify your survey design. If you have a set of replicate weights, then you use the REPWEIGHTS statement to specify the replicate weights. However, PROC SURVEYIMPUTE does not perform any data analysis or variance estimation. PROC SURVEYIMPUTE uses design information or the replicate weights to create a set of imputation-adjusted weights and imputation-adjusted replicate weights.

PROC SURVEYIMPUTE also supports the BY, CLASS, ID, and OUTPUT statements. The syntax for these statements is similar to that in any other SAS/STAT procedure. You use the BY statement to perform independent imputations within the BY groups, the CLASS statement to specify the categorical variables in your data, the ID statement to specify the variable that contains the observation IDs, and the OUTPUT statement to store the imputed data. In addition, PROC SURVEYIMPUTE supports a VAR statement that you use to name the variables to be imputed.

PROC SURVEYIMPUTE supports the CELLS and IMPJOINT statements, which are unique to this procedure. The CELLS statement specifies the variable that identifies the imputation cells. The IMPJOINT statement specifies the variables that are to be imputed jointly. By default, all variables in the VAR statement are imputed jointly. The IMPJOINT statement is applicable only for the fully efficient fractional imputation (FEFI) method.

Important options in the PROC SURVEYIMPUTE statement include the DATA=, METHOD=, NDONORS=, and REPWEIGHTSTYPE= options. You can use the DATA= option to name the input data set that contain observations with missing values, the REPWEIGHTSTYPE= option to name a replication method (either delete-1 jackknife or balanced repeated replication), and the NDONORS= option to specify the number of donors to use for a recipient unit. Use the METHOD= option to specify the imputation method. You can specify either METHOD=FEFI (fully efficient fractional imputation) or METHOD=HOTDECK (hot-deck imputation). When you specify METHOD=HOTDECK, you can also request a donor selection method by using the SELECTION= option. The donor selection methods include simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), weighted sampling (WEIGHTED), and the approximate Bayesian bootstrap (ABB).

Use the OUT= option in the OUTPUT statement to store the imputed data. When you use FEFI or when you use multiple donor units to impute one observation unit, then the OUTPUT OUT= imputed data set contains more rows than the DATA= input data set contains. The observation units in the output data are identified by the values of the variable in the ID statement. If you do not specify an ID statement, PROC SURVEYIMPUTE creates a new variable named **UnitID**. In addition, the output data set includes a new variable named **RecipientIndex**, which contains the recipient index.

## IMPUTATION METHODS IN PROC SURVEYIMPUTE

There are many ways to impute missing values. Every method essentially involves some model assumptions for what the missing values would have been if had they been observed; in principle these assumptions can bias the inferences you make from the imputed data. The methods implemented in PROC SURVEYIMPUTE strive to make minimal model assumptions; moreover, particular attention has been paid to how your final analysis can best account for the fact that some values are imputed. This section defines and discusses the three general imputation techniques available in PROC SURVEYIMPUTE: hot-deck imputation, FEFI, and the ABB method. Examples of each are shown using the **Asthma** and **ArthritisCell** data sets.

## HOT-DECK IMPUTATION

### Definition and Discussion

In the early days of modern computing, data were read from a deck of punch cards. The term "hot-deck" was used by the US Census Bureau to refer to an imputation procedure where donors were close to the recipients in the deck of cards. In contrast, "cold-deck" imputation used imputed values from an external source.

Hot-deck imputation is arguably the most commonly used imputation method for survey data. The observed data are first partitioned into imputation cells such that observations in the same cell are similar in some sense (Brick and Kalton 1996). The imputation is performed independently within each cell. Observations that contain no missing values are used as donors. For every recipient, one or more donors are selected randomly from the same imputation cell. The observed values of the donors are used as the imputed values for the missing items of the recipient.

In practice, different randomization techniques are used to select donor units. The properties of the estimators that are constructed from the imputed data depend on the randomization technique that is used for donor selection. For example, a selection with replacement increases the variability compared to a selection without replacement (Kalton and Kish 1984, p. 1925). If you want to achieve design unbiasedness, then a weighted selection is preferred over an unweighted selection (Rao and Shao 1992, p. 816).

Although the hot-deck imputation method is straightforward to implement, constructing a variance estimator that appropriately accounts for the imputation is challenging. Treating the imputed data as observed ignores the imputation variability and might underestimate the variance. Often, selecting multiple donor units for a recipient unit reduces the imputation variability and thus reduces the underestimation of the variance (Kalton and Kish 1984). For replication methods that account for hot-deck imputation, see Rao and Shao (1992), Fay (1993) and Shao and Tu (1995, Section 6.5).

If the observation weights are unequal, then it is reasonable to use a weighted selection of donors instead of an unweighted selection. You can use the SELECTION=WEIGHTED selection option in the METHOD=HOTDECK option in the PROC SURVEYIMPUTE statement to request a probability proportional to respondent weights with replacement selection of donors.

### Example 1: Hot-Deck Imputation for Asthma Data

Suppose you want to impute the missing values in the **Asthma** data by applying the hot-deck imputation method. You use the METHOD=HOTDECK option in the PROC SURVEYIMPUTE statement to perform hot-deck imputation. The following statements request a cell-based hot-deck imputation for the **Asthma** data (which are described in the section "EXAMPLE DATA" on page 2):

```
proc surveyimpute data=Asthma method=hotdeck(selection=srswor) seed=3242 ndonors=20;
   var   Asthma BMI Birth Education Income Race Smoker;
   class Asthma BMI Birth Education Income Race Smoker;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   cells Gender Age;
   output out=AsthmaHD donorid;
run;
```

The SELECTION=SRSWOR selection option in the METHOD=HOTDECK option requests the SRSWOR method for donor selection. The SEED= option specifies the seed for random number generation. The NDONORS= option specifies the number of donor units to select for every recipient unit.

The VAR statement specifies the variables that contain the missing values, and the CLASS statement specifies the classification variables. The ID statement names the variable that contains the observation IDs. The WEIGHT statement names the weight variable that contains the sampling weights, and the REPWEIGHTS statement names the variables that contain the bootstrap weights.

You can use the CELLS statement to identify the variables that define the imputation cells. In this example, imputation cells are created by using **Gender** and **Age**. All observations in the input data set contain nonmissing values for both **Gender** and **Age**. Variables in the CELLS statement should not contain any missing values.

The OUT= option in the OUTPUT statement names a data set, **AsthmaHD**, to store the imputed data. The **AsthmaHD** data set contains all observations from the **Asthma** input data set and replaces the missing values in the **Asthma**, **BMI**,

**Birth**, **Education**, **Income**, **Race**, and **Smoker** variables by observed values. In addition, PROC SURVEYIMPUTE creates a new variable, **RecipientIndex**, to store the recipient index. The DONORID option in the OUTPUT statement stores the unit IDs for the donors.

The summary information displayed in Figure 1 indicates that 2,003 observation units have missing values in the input data set. All of them are imputed. The **AsthmaHD** output data set contains 46,659 rows, which include 6,599 rows that are fully observed and 40,060 = 20 × 2003 rows that have imputed values. Each of the 2,003 units are imputed by using 20 rows.

**Figure 1** Imputation Summary for Hot-Deck Imputation

**The SURVEYIMPUTE Procedure**

Imputation Summary

| Observation Status | Number of Observations | Sum of Weights |
|---|---|---|
| **Nonmissing** | 6599 | 169945054 |
| **Missing** | 2003 | 58051020.7 |
| **Missing, Imputed** | 2003 | 58051020.7 |
| **Missing, Not Imputed** | 0 | 0 |

Figure 2 displays the first three imputed values for two observation units from the **AsthmaHD** data set. Both observations are in the same imputation cell, which is defined by **Gender**=Female and **Age** between 35 and 46. Observation unit 20 (**UnitID**=20) contains a missing value in **Income**, and observation unit 41 contains a missing value in **Smoker**. The **RecipientIndex** variable contains the recipient index, which ranges from 1 to 20. All nonmissing items are unchanged. The **DonorID** variable contains the **UnitID** of the donor units. For example, the missing value in **Income** for **UnitID** 20 and **RecipientIndex** 1 is replaced by the observed value of **Income** (=3) from **UnitID** 66. **UnitID** 66 is the donor unit for **UnitID** 20 and **RecipientIndex** 1.

Donors are selected independently in every **RecipientIndex**. The independent selection of donors in every **RecipientIndex** is equivalent to independent repetitions of the imputation. Thus, it is possible that the same donor unit is selected multiple times for different recipient indexes. However, if you use without-replacement sampling (as in the current example), then the same donor unit will not be used to impute multiple recipient units within the same recipient index.

**Figure 2** Selected Observations from Hot-Deck Imputation for Income and Smoker Variables

| Obs | UnitID | RecipientIndex | FWGT | Age | Gender | Asthma | BMI | Birth | Education | Income | Race | Smoker | DonorID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **96** | 20 | 1 | 14782.87 | 2 | 1 | 2 | 2 | 1 | 2 | 3 | 2 | 2 | 66 |
| **97** | 20 | 2 | 14782.87 | 2 | 1 | 2 | 2 | 1 | 2 | 3 | 2 | 2 | 30 |
| **98** | 20 | 3 | 14782.87 | 2 | 1 | 2 | 2 | 1 | 2 | 4 | 2 | 2 | 17 |
| **174** | 41 | 1 | 9579.09 | 2 | 1 | 2 | 3 | 1 | 2 | 2 | 1 | 2 | 74 |
| **175** | 41 | 2 | 9579.09 | 2 | 1 | 2 | 3 | 1 | 2 | 2 | 1 | 3 | 130 |
| **176** | 41 | 3 | 9579.09 | 2 | 1 | 2 | 3 | 1 | 2 | 2 | 1 | 3 | 92 |

Because there are 20 donors for each recipient (NDONORS=20), there are 20 observation rows in the **AsthmaHD** output data set that correspond to each observation unit in the input data that contains a missing value. You must change the observation weights, or create 20 data sets, before using the **AsthmaHD** data set for analysis. Do not use the unadjusted **AsthmaHD** data set for analysis. To analyze the imputed data, see the sections "ANALYSIS FOR IMPUTED DATA" on page 14 and "ANALYSIS FOR HOT-DECK IMPUTED DATA" on page 15.

Although SELECTION=SRSWOR is used in this example, it is also reasonable to use SELECTION=WEIGHTED to account for the unequal sampling weights in the **Asthma** data.

## FULLY EFFICIENT FRACTIONAL IMPUTATION

### Definition and Discussion

The fully efficient fractional imputation (FEFI) method uses all observed values from an item to replace the missing values in that item. A fractional weight is assigned to each imputed value; the fractional weight represents the proportion of that imputed value in the observed data. Imputed values that are observed more frequently in the input data set are assigned higher fractional weights. The sum of the fractional weights over all imputed values for every recipient unit is 1. The imputation-adjusted weights for a recipient are computed by multiplying the fractional weights by the full-sample weight of the recipient. The imputation-adjusted weight for observations that do not require imputation is unchanged from that observation's full-sample weight.

FEFI does not introduce additional variability due to the selection of donor units, and hence it is called "fully efficient" (Kim and Fuller 2004, p. 563). For more information about fractional imputation, see Kalton and Kish (1984), Fay (1996), Kim and Fuller (2004), Fuller and Kim (2005), Fuller (2009), and Kim and Shao (2014).

### Example 2: Fully Efficient Fractional Imputation

To illustrate FEFI, consider the 14 observation units in Figure 3. The variable **ID** contains the observation ID, the variable **Job** takes two values (Doctor and Teacher), and the variable **Income** takes three values (High, Medium, and Low). The first 10 observation units contain no missing values. **ID**=11 earns high income but has a missing value in **Job**. **ID**=12 is a doctor and **ID**=13 is a teacher, but both units have missing values in **Income**. **ID**=14 has missing values in both **Job** and **Income**. Each unit has a weight of 100.

**Figure 3** Income Data

| ID | Weight | Job | Income |
|---|---|---|---|
| 1 | 100 | Doctor | High |
| 2 | 100 | Doctor | High |
| 3 | 100 | Doctor | Medium |
| 4 | 100 | Doctor | Low |
| 5 | 100 | Doctor | Medium |
| 6 | 100 | Doctor | High |
| 7 | 100 | Teacher | Low |
| 8 | 100 | Teacher | Medium |
| 9 | 100 | Teacher | High |
| 10 | 100 | Teacher | Low |
| 11 | 100 | | High |
| 12 | 100 | Doctor | |
| 13 | 100 | Teacher | |
| 14 | 100 | | |

The following statements request FEFI. Variables that contain the missing values are specified in both CLASS and VAR statements. The WEIGHT statement names the weight variable, and the ID statement names the variable that contains the observation IDs. The OUTPUT statement names a data set to store the imputed data.

```
proc surveyimpute data=Income method=FEFI;
   class Job Income;
   var Job Income;
   weight Weight;
   id ID;
   output out=IncomeImputed;
run;
```

FEFI uses four steps: initialization, the E-step, the M-step, and repetition. After computing the imputed values and initializing the fractional weights, FEFI iterates between the so-called M-step (which computes joint probabilities) and the E-step (which recomputes fractional weights) until convergence. The expectation-maximization (EM) steps in FEFI are similar to the EM-by-weighting algorithm described in Ibrahim (1990).

Imputed values and the initial imputation-adjusted weights are computed only from the nonmissing observations.

Thus, four observations have **Income**=High, as shown in the following table.

| ID | Weight | Job | Income |
|---|---|---|---|
| 1 | 100 | Doctor | High |
| 2 | 100 | Doctor | High |
| 6 | 100 | Doctor | High |
| 9 | 100 | Teacher | High |

Because **Job** needs to be imputed for **ID**=11, these four units with **Income**=High are the set of possible donor units. They can be further categorized into two "donor cells": (Doctor, High) with total weight 300 and (Teacher, High) with total weight 100. It is these donor cells for which fractional weights and imputation-adjusted weights are computed. The imputation-adjusted weights for **ID**=11, with **Income**=High but **Job** missing, are thus initialized to 300/4 = 75 for **Job**=Doctor and 100/4 = 25 for **Job**=Teacher.

Similar calculations yield the initial imputation-adjusted weights (**ImpWt**) for all units as shown in Figure 4.

The number of rows after the initial computation is 24, which includes 10 rows for the first 10 observation units that do not contain any missing values, 2 imputed rows for **ID**=11, 3 imputed rows for **ID**=12 and **ID**=13, and 6 imputed rows for **ID**=14.

**Figure 4** Data after Initial FEFI

| Obs | ID | RecipientIndex | ImpWt | Job | Income |
|---|---|---|---|---|---|
| **1** | 1 | 0 | 100.000 | Doctor | High |
| **2** | 2 | 0 | 100.000 | Doctor | High |
| **3** | 3 | 0 | 100.000 | Doctor | Medium |
| **4** | 4 | 0 | 100.000 | Doctor | Low |
| **5** | 5 | 0 | 100.000 | Doctor | Medium |
| **6** | 6 | 0 | 100.000 | Doctor | High |
| **7** | 7 | 0 | 100.000 | Teacher | Low |
| **8** | 8 | 0 | 100.000 | Teacher | Medium |
| **9** | 9 | 0 | 100.000 | Teacher | High |
| **10** | 10 | 0 | 100.000 | Teacher | Low |
| **11** | 11 | 1 | 75.000 | Doctor | High |
| **12** | 11 | 2 | 25.000 | Teacher | High |
| **13** | 12 | 1 | 50.000 | Doctor | High |
| **14** | 12 | 2 | 16.667 | Doctor | Low |
| **15** | 12 | 3 | 33.333 | Doctor | Medium |
| **16** | 13 | 1 | 25.000 | Teacher | High |
| **17** | 13 | 2 | 50.000 | Teacher | Low |
| **18** | 13 | 3 | 25.000 | Teacher | Medium |
| **19** | 14 | 1 | 30.000 | Doctor | High |
| **20** | 14 | 2 | 10.000 | Doctor | Low |
| **21** | 14 | 3 | 20.000 | Doctor | Medium |
| **22** | 14 | 4 | 10.000 | Teacher | High |
| **23** | 14 | 5 | 20.000 | Teacher | Low |
| **24** | 14 | 6 | 10.000 | Teacher | Medium |

The M-step uses all 24 rows from Figure 4 and recomputes the weighted percentages for **Job** and **Income** by using the imputation-adjusted weights, **ImpWt**. The adjusted precentages from the first iteration are shown in the following table.

| Job | Income=High | | Income | Job=Doctor |
|---|---|---|---|---|
| Doctor | 73.98 | | Low | 15.17 |
| Teacher | 26.02 | | Medium | 30.34 |
| | | | High | 54.49 |

| Job | Income | | |
|---|---|---|---|
| | Low | Medium | High |
| Doctor | 9.05 | 18.10 | 32.50 |
| Teacher | 19.29 | 9.64 | 11.43 |

The E-step then updates the imputation-adjusted weights again for the imputed rows by using the weighted proportions from the previous M-step, and the whole process is iterated to convergence.

PROC SURVEYIMPUTE also computes replicate weights that you should use to estimate the standard errors for the estimators that use the FEFI data. By default, the procedure creates the delete-one jackknife replicate weights. Because you have 14 observation units, PROC SURVEYIMPUTE creates 14 replicate weights. Each set of replicate weights is further adjusted for imputation by applying the EM steps independently in each replicate.

For more information about the implementation of FEFI in PROC SURVEYIMPUTE, see the section “Fully Efficient Fractional Imputation” in the chapter “The SURVEYIMPUTE Procedure” in *SAS/STAT User’s Guide*; for more information about the replication methods in SAS/STAT, see Mukhopadhyay et al. (2008).

### APPROXIMATE BAYESIAN BOOTSTRAP IMPUTATION

#### Definition and Discussion

ABB is a donor selection method for hot-deck imputation that is recommended for use with multiple imputation (Rubin 1987, p. 124). ABB first creates a donor set by selecting a sample of size $r$ from $r$ donor units by using a SRSWR sample. ABB then selects $m$ donor units from the donor set by using another SRSWR sample, where $r$ and $m$ are the number of donor units and the number of recipient units, respectively.

PROC SURVEYIMPUTE implements a cell-based ABB, where you specify the imputation cells. If your survey design has stratification, then it is recommended that you define the imputation cells that are nested within the strata (Little and Rubin 2002, p. 89). You must specify the STRATA variables in the CELLS statement to define the imputation cells that are nested within the strata.

A different version of ABB is available in the MI procedure. PROC MI uses estimated response propensity for a variable in order to partition the data into several groups such that the observations within a group have similar response propensity. The ABB is then applied within each group. If you have missing values in multiple variables, then the process is repeated sequentially for each variable. For more information, see Yuan (2000).

For more information about ABB, see Rubin and Schenker (1986), Rubin (1987), Little and Rubin (2002), and Kim (2002).

#### Example 3: ABB Imputation for Asthma Data

Use the SELECTION=ABB selection option in the METHOD=HOTDECK option in the PROC SURVEYIMPUTE statement to request the ABB imputation method. The SEED= option specifies the seed for random number generation, and the NDONORS= option specifies the number of donor units to select for every recipient unit.

The following statements request ABB for the **Asthma**, **BMI**, **Birth**, **Education**, **Income**, **Race**, and **Smoker** variables in the **Asthma** data. Because you specify NDONORS=20, the imputation is independently repeated 20 times. All other statements are similar to the example in the section “HOT-DECK IMPUTATION” on page 5.

```
proc surveyimpute data=Asthma method=hotdeck(selection=abb) seed=3242 ndonors=20;
   var   Asthma BMI Birth Education Income Race Smoker;
   class Asthma BMI Birth Education Income Race Smoker;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   cells Gender Age;
```

```
      output out=AsthmaABB;
   run;
```

Figure 5 shows that all 2,003 observation units with missing values are imputed.

**Figure 5** Imputation Summary for ABB Imputation

**The SURVEYIMPUTE Procedure**

| Imputation Summary | | |
|---|---|---|
| Observation Status | Number of Observations | Sum of Weights |
| Nonmissing | 6599 | 169945054 |
| Missing | 2003 | 58051020.7 |
| Missing, Imputed | 2003 | 58051020.7 |
| Missing, Not Imputed | 0 | 0 |

All observation units that contain missing values in the input data are imputed 20 times in the output data set, **AsthmaABB**. You must change the observation weights (**ImpWt**) or create 20 data sets before using the **AsthmaHD** for analyses. Do not use the unadjusted **AsthmaHD** data set for analyses. To analyze the ABB imputed data, see the section "ANALYSIS FOR ABB DATA" on page 18.

## ADVANCED FEFI EXAMPLES WHEN NO DONORS ARE AVAILABLE

The basic principle of hot-deck imputation is to impute values for nonrespondents from donors who are as similar to them as possible. You can control which observations are judged to be similar by defining imputation cells.

As with any hot-deck imputation method, missing values in a recipient unit are not imputed if there are no donors available for that recipient unit. The most common reason for not finding a donor unit is the size of the imputation cell. For this reason, you often need to merge similar imputation cells into one cell so that some donor units are available for every recipient.

In addition to imputation cells, FEFI further screens donors for missing values in an item based on the observed values for the other items in that observation unit. Thus, with these two ways of screening donors, FEFI runs into the no-donor problem more frequently than other hot-deck methods.

The following two examples describe two approaches to implementing FEFI when no donors are available for some recipients.

### EXAMPLE 4: FEFI IN MULTIPLE STEPS FOR ASTHMA DATA

This example uses four steps to impute missing values in the analysis variables **Asthma**, **BMI**, **Birth**, **Education**, **Income**, **Race** and **Smoker** in the **Asthma** data. Assume that it is reasonable to create eight imputation cells by dividing **Age** into four age groups and by using the two levels of **Gender**.

The following statements request fully efficient fractional imputation for **Asthma**, **BMI**, **Birth**, **Education**, **Income**, **Race**, and **Smoker**:

```
   proc surveyimpute data=Asthma method=fefi;
      var   Asthma BMI Birth Education Income Race Smoker;
      class Asthma BMI Birth Education Income Race Smoker;
      cells Age Gender;
      weight fwgt;
      repweights bsw:;
      output out=Asthma1FEFI;
   run;
```

The DATA= option in the PROC SURVEYIMPUTE statement names the input data set, and the METHOD=FEFI option requests the FEFI method. The CELLS statement specifies the imputation cells. The WEIGHT statement specifies full-sample weights for the observation units. The REPWEIGHTS statement specifies the bootstrap replicate weights that are available in the **Asthma** data. The OUTPUT OUT= option names a data set to store the imputed data.

Figure 6 displays the imputation summary.

**Figure 6** Imputation Summary for FEFI

**The SURVEYIMPUTE Procedure**

| | |
|---|---|
| **Number of Observations Read** | 8602 |
| **Number of Observations Used** | 8602 |
| **Sum of Weights Read** | 227996075 |
| **Sum of Weights Used** | 227996075 |

**Imputation Summary**

| Observation Status | Number of Observations | Sum of Weights |
|---|---|---|
| **Nonmissing** | 6599 | 169945054 |
| **Missing** | 2003 | 58051020.7 |
| **Missing, Imputed** | 1845 | 53293940.6 |
| **Missing, Not Imputed** | 158 | 4757080.07 |
| **Missing, Partially Imputed** | 0 | 0 |

The 8,602 observation units in the sample represent 228 million individuals in the population for the year 2003. There are 2,003 observation units that have missing values in at least one variable. The 2,003 observation units represent 58 million individuals in the population. The number of donor cells range from 1 to 295, but only 25 observation units require more than 50 donor cells. The output data set, **Asthma1FEFI**, has 20,890 rows.

If all units with missing values are imputed, then you finish the imputation task after this step. However, there are 158 observation units that have no donor cells. Missing items in these units are not imputed. In addition, 207 observation units are imputed by using only one donor cell. Using only one donor cell to impute a missing observation provides appropriate variance estimation when that donor cell contains more than one primary sampling unit (PSU). However, if the donor cell contains only one observation unit (or all observation units are in the same PSU), then the replicate sample where that unit (or the PSU) is deleted does not properly account for the imputation. Therefore, it is reasonable to relax some of the imputation criteria for these observation units.

It is reasonable to use subsequent FEFI steps to impute the missing values in the problematic units by using different imputation cells or by imputing different variables. When performing FEFI with multiple steps, your objective is to collapse similar imputation cells and drop variables from the VAR statement until you find donor cells for every recipient unit. Drop the variables that you believe are not related to the variables you are imputing. You must retain the variables that require imputation in the VAR statement. A key point is that you must use all observation units from an imputation cell at each step to compute the fractional weights appropriately in that imputation cell.

This example uses four FEFI steps to impute missing values in all variables, as summarized in the following table:

| Step | CELL Statement Variables | VAR Statement Variables | OUTPUT OUT= |
|---|---|---|---|
| 1 | **Age Gender** | **Asthma BMI Birth Education Income Race Smoker** | **Asthma1FEFI** |
| 2 | **Gender** | **Asthma BMI Birth Education Income Race Smoker** | **Asthma2FEFI** |
| 3 | | **Asthma BMI Education Income Race** | **Asthma3FEFI** |
| 4 | **Gender** | **Education Income Race** | **Asthma4FEFI** |

Finally, you combine appropriate observation units from the four output data sets **Asthma1FEFI**–**Asthma4FEFI** to create a new data set **AsthmaFEFI** that contains imputed values, fractional weights, and imputation-adjusted replicate weights for all 8,602 observation units.

The **AsthamFEFI** data set contains rows from the various data sets as follows:

1. All respondent units from **Asthma1FEFI**
2. All imputed units that have at least two donor cells from **Asthma1FEFI**
3. All imputed units that have at least two donor cells from **Asthma2FEFI** but are not in 2
4. All imputed units that have at least two donor cells from **Asthma3FEFI** but are not in 2 or 3

5. All imputed units from **Asthma4FEFI** that are not in 2, 3, or 4

The imputed data set **AsthmaFEFI** contains 21,638 rows. The maximum number of donor cells is 295, but only 22 recipient units require more than 50 donor cells. The four-step approach imputes 1,638 observation units by using all nine variables together, 255 observation units by using a reduced set of eight variables, 103 observation units by using a reduced set of five variables, and 7 observation units by using a further reduced set of four variables.

A different number of FEFI steps might be necessary for some missing data. You should always use the entire data in every step. Otherwise, the imputation-adjusted weights and the imputation-adjusted replicate weights will be incorrect.

It is a good practice to verify the following in the final imputed data after a multi-step FEFI:

- All units should be included in the imputed data.
- All units that have missing values should have imputed values.
- The sum of the weights before the imputation should match the sum of the imputation-adjusted weights after the imputation for both the entire data set and for every observation unit.
- The sum of the replicate weights before the imputation should match the sum of the imputation-adjusted replicate weights after the imputation.

The SAS code for all four steps and for combining the final imputed data are shown in the Appendix.

Here's a simple trick to speed up the computation time in PROC SURVEYIMPUTE. If you are interested only in determining the number of donor cells for a recipient unit, then you don't need to create the imputation-adjusted weights and the imputation-adjusted replicate weights. These can be computationally expensive operations, so PROC SURVEYIMPUTE gives you a way to avoid them—the MAXEMITER=1 and REPWEIGHTSTYPE=NONE options in the PROC SURVEYIMPUTE statement. However, don't use these options once you've settled on an imputation scheme and you are ready to create the final imputed data.

### EXAMPLE 5: FEFI USING DIFFERENT VARIABLES SEPARATELY FOR ARTHRITISCELL DATA

You can use PROC SURVEYIMPUTE to perform FEFI on two groups of variables separately. This method is useful when the two groups of variables are not related to each other. This example uses two IMPJOINT statements to impute missing values in the **ArthritisCell** data in two separate groups.

The following statements request FEFI for variables **SelfRHealth**, **Arthritis**, **Diabetes**, **RaceCat**, and **EdCat**. These statements are similar to the statements in Example 4.

Because you don't have the replication weights available in the **ArthritisCell** data, you can't use the REPWEIGHTS statement. Instead, you know the strata and cluster identifications. Therefore, you use the STRATA, CLUSTER, and WEIGHT statements to specify the stratum, cluster, and full-sample weight variables to compute the imputation-adjusted weights and the imputation-adjusted replicate weights. When you specify METHOD=FEFI in the PROC SURVEYIMPUTE statement, PROC SURVEYIMPUTE creates a set of imputation-adjusted delete-1 jackknife replicate weights by default. But the **ArthritisCell** data set uses a design with two PSUs per stratum. Therefore, it is reasonable to use the REPWEIGHTSTYPE=BRR option in the PROC SURVEYIMPUTE statement to create a set of imputation-adjusted balanced repeated replication (BRR) weights.

The CELLS statement specifies the imputation cells. In this example, the imputation cells are created by using a clustering algorithm that is available in the CLUSTER procedure. Variables **Gender** and **KAge** do not contain missing values in any observation units and are used to create the imputation cells.

```
proc surveyimpute data=ArthritisCell method=fefi repweightstype=brr;
   class selfrhealth arthritis diabetes racecat edcat;
   var   selfrhealth arthritis diabetes racecat edcat;
   cells ImpCell;
   weight kwgtr;
   strata stratum;
   cluster secu;
   id UnitID;
   output out=ArthFEFI;
run;
```

There are 127 observation units with missing values in the **ArthritisCell** data—126 observation units are imputed, but no donor cells are available for 1 observation unit. The observation unit that is not imputed in the previous step is displayed in Figure 7. There are no observation units in imputation cell 5 that have **SelfRHealth** equal to 2 and **Arthritis**, **Diabetes**, and **RaceCat** equal to 1. Therefore, the missing value in **EdCat** for **UnitID**=16409 is not imputed.

**Figure 7** Observation Unit Not Imputed

| UnitID | KWGTR | ImpCell | SELFRHEALTH | ARTHRITIS | DIABETES | RACECAT | EDCAT |
|---|---|---|---|---|---|---|---|
| 16409 | 7372 | 5 | 2 | 1 | 1 | 1 | . |

You can merge two imputation cells as in Example 4 to impute **EdCat** in **UnitID**=16409. However, strictly for illustration purposes, this example demonstrates an alternative approach.

Assume that the health variables **SelfRHealth**, **Arthritis**, and **Diabetes** are related to each other and the demographic variables **RaceCat** and **EdCat** are related to each other but the two groups of variables are not related.

You use two IMPJOINT statements in PROC SURVEYIMPUTE to request FEFI independently on two groups of variables. The health variables are specified in one IMPJOINT statement, and the demographic variables are specified in another IMPJOINT statement.

The following statements request FEFI on two groups of variables separately:

```
proc surveyimpute data=ArthritisCell method=FEFI repweightstype=brr;
   class selfrhealth arthritis diabetes racecat edcat;
   var   selfrhealth arthritis diabetes racecat edcat;
   impjoint selfrhealth arthritis diabetes;
   impjoint racecat edcat;
   cells ImpCell;
   strata stratum;
   cluster secu;
   weight kwgtr;
   id UnitID;
   output out=ArthFEFI2;
run;
```

The OUT= option in the OUTPUT statement names a data set to store the output data, which contain both the observed values and the imputed values.

The imputation summary in Figure 8 shows that all observation units that have missing values are imputed. The missing data pattern table in Figure 9 shows that 99.25% observation units in the data set contain no missing values, 0.47% observation units contain missing values in **EdCat**, and 0.12% observation units contain missing values in **SelfRHealth**. Columns Sum of Weights and Weighted Percent represent the estimated number of observation units and the estimated percentages of the observation units in the population, respectively. Columns Number of Observations and Unweighted Percent represent the number of observation units and the percentages in the sample, respectively.

**Figure 8** Imputation Summary for FEFI

**The SURVEYIMPUTE Procedure**

| Imputation Summary | | |
|---|---|---|
| **Observation Status** | **Number of Observations** | **Sum of Weights** |
| **Nonmissing** | 16827 | 75915519 |
| **Missing** | 127 | 625148 |
| **Missing, Imputed** | 127 | 625148 |
| **Missing, Not Imputed** | 0 | 0 |
| **Missing, Partially Imputed** | 0 | 0 |

**Figure 9** Missing Data Pattern for **ArthritisCell**

| Group | SELFRHEALTH | ARTHRITIS | DIABETES | RACECAT | EDCAT | Number of Observations | Sum of Weights | Unweighted Percent | Weighted Percent |
|---|---|---|---|---|---|---|---|---|---|
| 1 | X | X | X | X | X | 16827 | 75915519 | 99.25 | 99.18 |
| 2 | X | X | X | X | . | 79 | 405286 | 0.47 | 0.53 |
| 3 | X | X | X | . | X | 3 | 16584 | 0.02 | 0.02 |
| 4 | X | X | . | X | X | 11 | 67166 | 0.06 | 0.09 |
| 5 | X | . | X | X | X | 8 | 30940 | 0.05 | 0.04 |
| 6 | X | . | . | X | X | 2 | 9446 | 0.01 | 0.01 |
| 7 | . | X | X | X | X | 20 | 81825 | 0.12 | 0.11 |
| 8 | . | X | . | X | X | 1 | 4023 | 0.01 | 0.01 |
| 9 | . | . | X | X | X | 1 | 4396 | 0.01 | 0.01 |
| 10 | . | . | . | X | X | 2 | 5482 | 0.01 | 0.01 |

Using two IMPJOINT statements in one PROC SURVEYIMPUTE invocation is similar to performing FEFI on two groups of variables by using two PROC SURVEYIMPUTE invocations. The two groups of variables are imputed independently as described in the following list:

- If an observation unit contains a missing value in a variable from the first group but no missing values in the variables in the second group, then the observed levels of the other variables in only the first group are used in the imputation. The observed levels for the second group of variables are not used when imputing the first group of variables.
- If an observation unit contains a missing value in a variable from the second group but no missing values in the variables in the first group, then the observed levels of the other variables in only the second group are used in the imputation. The observed levels for the first group of variables are not used when imputing the second group of variables.
- If an observation unit contains missing values in both groups, then FEFI is first used to impute the missing values in the first group. Then for each imputed level from the first group, FEFI is used to impute the missing values in the second group.

You can use multiple IMPJOINT statements in one PROC SURVEYIMPUTE invocation to divide the variables into multiple groups.

Imputing different variables separately can significantly increase the number of rows in the imputed data set. In addition, the relation between the two groups of variables in the observed data might be distorted in the imputed data.

## ANALYSIS FOR IMPUTED DATA

Before you analyze imputed data sets, you should know how the missing values in the data were imputed. Both point estimation and variance estimation depend on the imputation method. For example, if you use multiple imputation, then you should analyze each imputed data set separately to compute the within-imputation estimates and then combine all within-imputation estimates to compute the final results. On the other hand, if you use fractional imputation, then you should use the imputation-adjusted fractional weights and the imputation-adjusted replicate weights in your analyses.

If you publish your imputed data set, then you should identify the units and the items that contain the imputed values and explain how the imputation was performed.

If you use a public-use data set that contains imputed values, then you should understand how the imputation was performed before you use your favorite SAS/STAT procedures for analyses.

This paper uses PROC SURVEYFREQ and PROC SURVEYLOGISTIC to analyze imputed data, but you can use the analysis techniques described in this section with any SAS/STAT survey data analysis procedure.

## ANALYSIS FOR HOT-DECK IMPUTED DATA

### Definition and Discussion

If the variance added by the imputation process is small compared to the total variance, then you can ignore the imputation variance in the analysis step. If you use a without-replacement hot-deck imputation procedure and you use multiple donors to impute missing items in each recipient unit, the variance added by the imputation might be small for many statistics. For more information about the increase in variance due to hot-deck imputation, see Kalton and Kish (1984). The following example analyzes hot-deck imputed data by ignoring the imputation variability when multiple donor units are used to impute missing values in each recipient unit.

Consider the hot-deck imputed data from the section "HOT-DECK IMPUTATION" on page 5. Recall that 20 donor units were used to replace the missing values in every recipient unit. The following statements create full-sample weights and replicate weights such that the weights are divided by 20 for the units that require imputation. The weights are unchanged for units that do not require imputation. The new weights are **ImpWt** for the full-sample weight and **ImpRepWt_1** to **ImpRepWt_1000** for 1,000 replicate samples.

```
data AsthmaHD; set AsthmaHD;
   array ImpRepWt_{1000};
   array bsw      {1000};
   if RecipientIndex = 0 then mul = 1;
   else  mul = 1/20;
   /*---Assign fractional weights                                  ---*/
   do i = 1 to 1000;
      ImpRepWt_{i} = mul*bsw{i};
   end;
   ImpWt = mul*fwgt;
run;
```

### Example 1 Follow-Up: Frequency Analysis for Hot-Deck Imputed Data

The following statements request a one-way frequency table for **BMI** and a domain analysis for **BMI** and **Education** for domains that are defined by **Race**. The original replicate weights are created by using the rescaling bootstrap method (Rao and Wu 1988). The replicate coefficient for this bootstrap is $1/R$ where $R$ is the number of bootstrap replicates. The VARMETHOD=BRR option in the PROC SURVEYFREQ statement specifies $1/R$ as the replicate coefficient.

```
proc surveyfreq data=AsthmaHD varmethod=brr;
   tables BMI Race*BMI*Education;
   repweights ImpRepWt_:;
   weight ImpWt;
run;
```

The TABLES statement names the variables that you want to tabulate. The REPWEIGHTS statement names the variables that contain the adjusted replicate weights, and the WEIGHT statement names the variable that contain the adjusted full-sample weights.

Figure 10 displays the weighted frequencies, estimated percentages, and their standard errors for **BMI**. The Frequency column represents the frequencies for the number of rows in the **AsthmaHD** data rather than the frequencies for the observation units. You can use the NOFREQ option in the TABLES statement to supress the Frequency column from the output. The weighted frequency represents the estimated number of observation units in the population. The table shows that 2.35% (0.21) of the individuals in the population from 2003 are underweight, which represents an estimated 5,364,515 (472,909) individuals in the population. The standard errors for the estimates are given in parenthesis.

**Figure 10** Frequency Table for **BMI** Using Hot-Deck Imputed Data

**The SURVEYFREQ Procedure**

| Table of BMI | | | | | |
|---|---|---|---|---|---|
| BMI | Frequency | Weighted Frequency | Std Err of Wgt Freq | Percent | Std Err of Percent |
| 1 | 1566 | 5364515 | 472909 | 2.3529 | 0.2071 |
| 2 | 21761 | 99907463 | 1531391 | 43.8198 | 0.6648 |
| 3 | 14960 | 76885739 | 1486426 | 33.7224 | 0.6532 |
| 4 | 8372 | 45838357 | 1266117 | 20.1049 | 0.5554 |
| Total | 46659 | 227996075 | 334711 | 100.000 | |

The results of the domain analysis are not shown.

Although 20 donor units were used in this example to replace missing values in every recipient unit, five donor units are often sufficient to reduce the imputation variance unless you have a massive amount of missing data.

## ANALYSIS FOR FEFI DATA

### Definition and Discussion

You should use imputation-adjusted weights and imputation-adjusted replicate weights to analyze FEFI data. The number of rows in the imputed data does not represent the number of observation units. Therefore, you must be careful to interpret some results from SAS/STAT procedures. For example "Number of Observations" from the "Data Summary" table in PROC SURVEYMEANS does not represent the number of observation units; instead it represents the number of rows in the input data. With FEFI data, you should always use the statistics that are constructed by using the imputation-adjusted weights and imputation-adjusted replicate weights.

### Example 4 Follow-Up: Frequency Analysis for FEFI Data

Consider the data set **AsthmaFEFI** from "EXAMPLE 4: FEFI IN MULTIPLE STEPS FOR **ASTHMA** DATA" on page 10, which uses the FEFI method to impute missing values in **Asthma**, **BMI**, **Birth**, **Education**, **Income**, **Race**, and **Smoker**.

The following statements request a one-way frequency table for **BMI** and a domain analysis for **BMI** and **Education** for domains that are defined by **Race**. In "EXAMPLE 4: FEFI IN MULTIPLE STEPS FOR **ASTHMA** DATA" on page 10, you provided a set of bootstrap replicate weights, and PROC SURVEYIMPUTE adjusted the full-sample weights and the bootstrap weights for FEFI. The imputation-adjusted full-sample weights are specified by using the WEIGHT statement, and the imputation-adjusted replicate weights are specified by using the REPWEIGHTS statement. These statements are similar to the example in the section "ANALYSIS FOR HOT-DECK IMPUTED DATA" on page 15.

```
proc surveyfreq data=AsthmaFEFI varmethod=brr;
   tables BMI Race*BMI*Education;
   repweights ImpRepWt_:;
   weight ImpWt;
run;
```

Figure 11 displays the estimated percentages and standard error for **BMI**. The Frequency column displays the frequencies for the number of rows in the **AsthamFEFI** data. The Weighted Frequency column displays the estimated number of observation units in the population for the year 2003 for each level of **BMI**, and the Std Err of Wgt Freq column represents the standard error of the estimated weighted frequencies. The Percent and the Std Err of Percent columns display the estimated percentage of the observation units that fall in each category of BMI in the population and the standard error of the estimated percentage, respectively. For example, 2.31% (0.21) of the individuals are underweight in the 2003 population.

**Figure 11** Frequency Table for BMI Using FEFI Data

**The SURVEYFREQ Procedure**

| Table of BMI | | | | | |
|---|---|---|---|---|---|
| **BMI** | **Frequency** | **Weighted Frequency** | **Std Err of Wgt Freq** | **Percent** | **Std Err of Percent** |
| **1** | 674 | 5270866 | 487693 | 2.3118 | 0.2136 |
| **2** | 10099 | 99710166 | 1573961 | 43.7333 | 0.6838 |
| **3** | 7079 | 77036064 | 1524007 | 33.7883 | 0.6699 |
| **4** | 3786 | 45978980 | 1306400 | 20.1666 | 0.5728 |
| **Total** | 21638 | 227996075 | 334711 | 100.000 | |

The frequency table for **BMI** and **Education** for **Race**=1 subpopulation is displayed in Figure 12.The total from the Weighted Frequency column shows that an estimated 167,866,610 (954,562) observation units are white in the population for the year 2003. Among the whites, 1.87% (0.21) of the individuals are underweight, 46.07% (0.77) of the individuals are normal weight, 33.36% (0.72) of the individuals are overweight, and 18.70% (0.60) of the individuals are obese in the population. In the white subpopulation, 5.93% (0.36) of the individuals are obese and have education level above high school, whereas 19.55% (0.61) of the individuals are normal weight and have education level above high school.

**Figure 12** Frequency Table for BMI and Education for the White Subpopulation Using FEFI Data

| Table of BMI by Education<br>Controlling for Race=1 | | | | | | |
|---|---|---|---|---|---|---|
| **BMI** | **Education** | **Frequency** | **Weighted Frequency** | **Std Err of Wgt Freq** | **Percent** | **Std Err of Percent** |
| **1** | **1** | 84 | 316545 | 82813 | 0.1886 | 0.0493 |
| | **2** | 218 | 1820783 | 278244 | 1.0847 | 0.1652 |
| | **3** | 158 | 1001608 | 197602 | 0.5967 | 0.1173 |
| | **Total** | 460 | 3138935 | 347857 | 1.8699 | 0.2061 |
| **2** | **1** | 1174 | 6508506 | 511146 | 3.8772 | 0.3035 |
| | **2** | 3762 | 38009703 | 1117345 | 22.6428 | 0.6463 |
| | **3** | 2742 | 32811896 | 1047637 | 19.5464 | 0.6139 |
| | **Total** | 7678 | 77330105 | 1382441 | 46.0664 | 0.7707 |
| **3** | **1** | 940 | 5343495 | 406659 | 3.1832 | 0.2417 |
| | **2** | 2584 | 28816416 | 1064697 | 17.1663 | 0.6285 |
| | **3** | 1744 | 21847194 | 903752 | 13.0146 | 0.5338 |
| | **Total** | 5268 | 56007105 | 1245364 | 33.3641 | 0.7240 |
| **4** | **1** | 542 | 3304388 | 327413 | 1.9685 | 0.1942 |
| | **2** | 1438 | 18129898 | 833529 | 10.8002 | 0.4917 |
| | **3** | 829 | 9956180 | 609137 | 5.9310 | 0.3625 |
| | **Total** | 2809 | 31390466 | 1014507 | 18.6996 | 0.5959 |
| **Total** | **1** | 2740 | 15472933 | 697610 | 9.2174 | 0.4128 |
| | **2** | 8002 | 86776800 | 1463386 | 51.6939 | 0.8103 |
| | **3** | 5473 | 65616877 | 1378445 | 39.0887 | 0.7950 |
| | **Total** | 16215 | 167866610 | 954562 | 100.000 | |

### Example 5 Follow-Up: Logistic Regression Analysis for FEFI Data

To analyze FEFI data, you don't need to know whether FEFI was performed by using one step, multiple steps, or by using different variables separately. You should always use the imputation-adjusted weights and the imputation-adjusted replicate weights in your analysis.

Consider the data set **ArthFEFI2** from the section "EXAMPLE 5: FEFI USING DIFFERENT VARIABLES SEPARATELY FOR **ARTHRITISCELL** DATA" on page 12, which uses FEFI to impute missing values in **SelfRHealth**, **Arthritis**, **Diabetes**, **RaceCat**, and **EdCat**.

Suppose you want to estimate the odds of reporting arthritis for diabetic individuals after adjusting for age. The following statements fit a logistic regression model for **Arthritis** on **Diabetes** and **KAge**:

```
proc surveylogistic data=ArthFEFI2 varmethod=brr;
   class Diabetes;
   model Arthritis = Diabetes KAge;
   weight ImpWt;
   repweights ImpRepWt_:;
run;
```

Recall that PROC SURVEYIMPUTE created a set of imputation-adjusted BRR weights in the **ArthFEFI2** data set. The VARMETHOD=BRR option in the PROC SURVEYLOGISTIC statement requests the BRR variance estimation method. The WEIGHT and REPWEIGHTS statements specify the imputation-adjusted full-sample weights and the imputation-adjusted replicate weights, respectively.

The estimated odds of reporting arthritis for an individual who has diabetes is 1.7 times the estimated odds of reporting arthritis for an individual who does not have diabetes after adjusting for the age of the individual in the 2006 population (Figure 13). The degrees of freedom to compute the confidence intervals for the odds ratios is 60, which is the number of BRR weights created by PROC SURVEYIMPUTE in the section "EXAMPLE 5: FEFI USING DIFFERENT VARIABLES SEPARATELY FOR **ARTHRITISCELL** DATA" on page 12.

**Figure 13** Estimated Odds Ratios for Arthritis by Using the FEFI Data

**The SURVEYLOGISTIC Procedure**

| Odds Ratio Estimates | | | |
|---|---|---|---|
| Effect | Point Estimate | 95% Confidence Limits | |
| DIABETES 0 vs 1 | 1.711 | 1.533 | 1.908 |
| KAGE | 0.950 | 0.946 | 0.954 |

**NOTE: The degrees of freedom in computing the confidence limits is 60.**

## ANALYSIS FOR ABB DATA

### Definition and Discussion

ABB relies on the principle of multiple imputation. Therefore, you should use the same technique to analyze ABB data that you use to analyze any multiply imputed data. The analysis is divided into two parts: the within-imputation part and the between-imputation part. For survey data, use the survey procedures along with the complete design information for the within-imputation analysis. Use PROC MIANALYZE for the between-imputation analysis for both survey and non-survey data.

Consider the data set **AsthmaABB** from the section "APPROXIMATE BAYESIAN BOOTSTRAP IMPUTATION" on page 9, which uses the ABB method to impute missing values in **Asthma**, **BMI**, **Birth**, **Education**, **Income**, **Race**, and **Smoker**. Twenty donor units are used to impute missing values in each recipient unit.

To facilitate the analysis, the following statements create 20 imputed data sets. Each data set contains all the observed units and exactly one donor unit for every recipient unit.

```
data AsthmaABBMI;
   set AsthmaABB;
   if (Recipient = 0) then do;    /* Include complete respondents */
      do _Imputation_=1 to 20;    /* in all imputations.          */
         output;
      end;
```

```
      end;
      else do;                        /* Put incomplete respondents   */
         _Imputation_ = Recipient;  /* in separate imputations.     */
         output;
      end;
   proc sort data=AsthmaABBMI;
      by _Imputation_ UnitID;
   run;
```

### Example 3 Follow-Up: Frequency Analysis for ABB Imputed Data

Suppose you want to compute a one-way frequency table for **BMI** and a domain analysis for a two-way frequency table for **BMI** and **Education** for the **Race**=White domain.

The following statements request separate frequency tables within every imputed data set. The ODS SELECT statement suppresses the output from individual analysis. The TABLES statement names the variables that you want to tabulate.

```
   ods select none;
   proc surveyfreq data=AsthmaABBMI varmethod=brr;
      tables BMI Race*BMI*Education;
      repweights bsw:;
      weight fwgt;
      by _imputation_;
      ods output table1.oneway=BMIFreqWI
                 table1of2.crosstabs=BMIEduWhiteWI;
   run;
   ods select all;
```

Use the complete design information (including strata, cluster, and weights) to compute the within-imputation estimates. If you perform domain analysis, then you must use the domain identification within each imputed data set. In the **AsthmaABB** data set, you do not have the strata or the cluster information, but a set of bootstrap weights is available to you. Specify the bootstrap weights in the REPWEIGHTS statement. As in the previous examples, the VARMETHOD=BRR option in the PROC SURVEYFREQ statement specifies $1/R$ as the replication coefficient.

The BY statement requests independent analysis for each imputed data set. The ODS OUTPUT statement stores the frequency tables.

The following statements request the between-imputation analysis. The MODELEFFECTS statement specifies the variable that contains the point estimates, and the STDERR statement specifies the variable that contains the standard errors from the within-imputation analysis.

```
   proc sort data=BMIFreqWI;
      by BMI _Imputation_;
   run;
   data BMIFreqWI; set BMIFreqWI;
      where BMI ne .;
   run;
   proc mianalyze data=BMIFreqWI edf=1000;
      by BMI;
      modeleffects percent;
      stderr stderr;
      ods output ParameterEstimates=BMIFreq;
   run;
```

The degrees of freedom for survey data are often much less than the number of observation units. The degrees of freedom for survey data are computed by using the number of clusters and the number of strata, or by using the number of replicate weights. You need to adjust the degrees of freedom when using multiply imputed data sets for survey analyses. You should specify the survey degrees of freedom by using the EDF= option in PROC MIANALYZE. For more information about how the degrees of freedoms are adjusted, see the EDF=*number* option in the chapter "The MIANALYZE Procedure" in *SAS/STAT User's Guide*; also see Barnard and Rubin (1999). However, in this example the design degrees of freedom is large and therefore specifying EDF=1000 has no significant effect.

The **BMIFreq** data set contains the parameter estimates and the standard errors after combining results from the within-imputation and between-imputation data sets. Figure 14 displays the estimated percentages and their standard errors. The table shows that an estimated 2.32% (0.22) of the individuals are underweight.

**Figure 14** Estimated Percentages and Their Standard Errors for BMI Using ABB Imputed Data

| BMI | Estimate | StdErr |
|---|---|---|
| 1 | 2.317474 | 0.217195 |
| 2 | 43.833642 | 0.699094 |
| 3 | 33.795569 | 0.680678 |
| 4 | 20.053315 | 0.573428 |

Use the estimated percentages and their standard errors from **BMIEduWhiteWI** in another invocation of PROC MIANALYZE to combine the within-imputation results for the domain analysis for the white subpopulation.

#### Example 3 Follow-Up: Logistic Regression for ABB Imputed Data

Suppose you want to know whether more smokers than nonsmokers reported asthma in the 2003 population, after adjusting for **BMI**, **Age**, and **Gender**. You can perform a logistic regression for **Asthma** on **Gender**, **Age**, **BMI**, and **Smoker** by using the ABB imputed data.

The following PROC SURVEYLOGISTIC invocation computes the within-imputation estimates. The WEIGHT, REPWEIGHTS, and BY statements are exactly the same as in the previous example, which used PROC SURVEYFREQ. The MODEL statement specifies the regression model, and the COVB option in the MODEL statement requests the estimated covariance matrix. The ODS OUTPUT statement stores the parameter estimates and the estimated covariance matrix from every imputed data set.

```
ods select none;
proc surveylogistic data=AsthmaABBMI varmethod=brr;
   class BMI Smoker Gender;
   model Asthma=BMI Smoker Age Gender / covb;
   repweights bsw:;
   weight fwgt;
   by _imputation_;
   ods output parameterestimates=Estimates covb=Covariances;
run;
ods select all;
```

It is a good practice to verify that the maximum likelihood estimates converge in the full sample and in every replicate sample for all BY groups.

Use PROC MIANALYZE to combine the within-imputation estimates and the covariances. The MODELEFFECTS statement specifies the regression effects, and the CLASS statement specifies the CLASS variables that are used in the previous PROC SURVEYLOGISTIC analysis. The results are not shown.

```
proc mianalyze parms(classvar=classval)=Estimates
               covb(effectvar=stacking)=Covariances
               edf=1000;
   class BMI Smoker Gender;
   modeleffects Intercept BMI Smoker Age Gender;
   ods output parameterestimates=ABBLogisticAnalysis;
run;
```

## SAME ANALYSIS BUT DIFFERENT RESULTS

When you use imputed data, the estimated values for your parameters of interest and their standard errors depend on how the imputation was performed. It is not surprising that the estimated percentages and their standard errors in Figure 10, Figure 11, and Figure 14 are different. Although the three analyses use the same **Asthma** data set before imputation, they use different imputation methods and different variance estimation methods. This is why it is important to mention what imputation method is used in your data and what variance estimator is used for statistical inference.

To choose an imputation method, you must consider the advantages and disadvantages for that method. Any imputation method replaces missing values with plausible values. All methods work well if the underlying assumptions are satisfied. But if the assumptions are not satisfied, the performance will be sacrificed. This section describes some advantages and disadvantages for the three imputation methods available in PROC SURVEYIMPUTE.

With hot-deck imputation, the imputed values are always observed in the data set. Therefore, unreasonable imputed values (such as 1.5 children) are not possible. In addition, if the same donor unit is used to impute all missing items in a recipient unit, then unreasonable combinations of items (such as pregnant males) are not possible. However, constructing a variance estimator that appropriately accounts for the imputation for complex estimators is not easy.

Some advantages of FEFI include the following:

- Is applicable to complex surveys with multi-stage designs
- Does not add any extra variability due to the selection of donors
- Does not use any explicit models
- Uses only observed values as the imputed values
- Preserves the relationship between multiple survey items

The primary disadvantage of FEFI is that it could significantly increase the number of rows in the imputed data if the missing items have many levels in the observed data. For example, if you impute a variable that has 100 observed levels, then the output data might contain 100 new rows for every observation unit that has a missing value in that variable.

Adding many imputed rows might not be a major disadvantage in the modern age of "big data," but you might also encounter situations where no donors are available for a recipient. FEFI is a form of hot-deck imputation, and as in any hot-deck imputation, auxiliary information is incorporated by using only the imputation cells. If the imputation cells are small and you impute many variables, then you might not find donors for some recipient units. Therefore, you need to merge some imputation cells, and thus you might compromise the quality of your imputed data.

A large number of donor units per recipient unit also helps you to create jackknife replicates that have better properties. The jackknife variance estimator of the variance for FEFI has a positive bias because of the finite number of donor units per recipient. Recall that the $(n-1)/n$ adjustment factor is used in the standard jackknife for a simple random sample of size $n$ to reduce the bias. The same type of bias occurs when you create replicates for the set of donor units for a recipient. However, this bias is specific to the response variable, and thus no adjustments are made in creating the jackknife replicates for FEFI. The bias is small when the number of donor units per recipient unit is large. Because the bias is positive, the inference from the FEFI data will be conservative.

The analysis techniques that you can apply to FEFI data are also restricted. Imputed data from FEFI should be used only with survey procedures that support replication methods. You should not use a non-survey procedure, such as PROC GLIMMIX, to analyze FEFI data. In addition, you should not use the Taylor series linearized variance estimator that is available in the survey analysis procedures to analyze FEFI data. You should use the REPWEIGHTS statement in the survey analysis procedures to analyze FEFI data. To analyze data from complex surveys, you should always use the SAS/STAT survey procedures, so this restriction is not significant.

The ABB method relies on the principle of multiple imputation. You must ensure that the imputation is proper (Rubin 1987, Section 4.2) in order to construct randomization-valid confidence intervals from your imputed survey data. However, it is challenging to design an imputation technique that is proper for many variables that you typically observe in a complex survey, which also involves stratification, clustering, and unequal weights (Binder and Sun 1996, p. 286). Multiple imputation that is proper for one variable might not be proper for another variable. Moreover, multiple imputation that is proper for a variable in one analysis might not be proper for the same variable in another analysis. For example, an imputation that is proper for a variable in the overall analysis might not be proper for the same variable in a domain analysis (Fay 1993; Kim et al. 2006; Little and Rubin 2002, Example 10.6). For a good discussion on multiple imputation and other alternatives for survey data, see Rubin (1996), Fay (1996), and Rao (1996) in addition to comments by Binder, Eltinge, and Judkins, and rejoinders by Rao, Fay, and Rubin on pages 507–520 in the same journal issue.

## SUMMARY

Imputation is a very challenging task. The imputer should have sound understanding about the survey selection process and the nonresponse process, and have sufficient auxiliary information and background information about the variables to be imputed. It might be better not to do any imputation than to perform poor imputation. After you choose an imputation method, you can use PROC SURVEYIMPUTE to impute the missing values.

PROC SURVEYIMPUTE supports FEFI and four other hot-deck imputation methods. The METHOD=FEFI option requests the FEFI method, and the METHOD=HOTDECK option along with its SELECTION= selection option specifies the other hot-deck imputation methods. The available donor selection methods are SRSWR, SRSWOR, WEIGHTED, and ABB.

You can use any SAS/STAT survey procedure to analyze survey data that contain imputed values; but before analyzing the imputed data, you must identify the imputation method. The choice of your analysis technique depends on the imputation method used during the data preparation stage. The following list summarizes the imputation methods and the corresponding analysis techniques that are described in this paper:

- **FEFI:** This is the preferred imputation method for complex surveys. To analyze FEFI data, you should use the imputation-adjusted weights and the imputation-adjusted replicate weights in the survey procedures.
- **Hot-Deck:** This is arguably the most commonly used imputation method for survey data. If you use it, you should use an efficient donor selection method that minimizes the imputation variability. When it comes to analysis, the added variability due to the imputation is often ignored in practice.
- **ABB:** This imputation method relies on the principle of multiple imputation. Accordingly, although you use the survey procedures to compute the within-imputation analysis, you need to use the MIANALYZE procedure to compute the between-imputation analysis.

## APPENDIX

This appendix provides the complete SAS® code that is used to perform FEFI in multiple steps in the section "EXAMPLE 4: FEFI IN MULTIPLE STEPS FOR **ASTHMA** DATA" on page 10. First, four PROC SURVEYIMPUTE invocations create four FEFI data sets. Then DATA statements are used to combine the four imputed data sets. To simplify the code, a SAS macro, %OneOrZeroDonor, is used to identify observation units that might require further imputation after each step. The definition of this macro is given at the end of this section.

The following statements use four PROC SURVEYIMPUTE invocations to create four imputed data sets:

```
/*---FEFI using the full set of variables (FSV)                ---*/
proc surveyimpute data=Asthma method=fefi;
   var   Asthma BMI Birth Education Income Race Smoker;
   class Asthma BMI Birth Education Income Race Smoker;
   cells Age Gender;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   output out=Asthma1FEFI;
run;
/*---FEFI using the first reduced set of variables (RSV1)      ---*/
proc surveyimpute data=Asthma method=fefi;
   var   Asthma BMI Birth Education Income Race Smoker;
   class Asthma BMI Birth Education Income Race Smoker;
   cells Gender;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   output out=Asthma2FEFI;
run;
/*---FEFI using the second reduced set of variables (RSV2)     ---*/
proc surveyimpute data=Asthma method=fefi;
   var   Asthma BMI Education Income Race;
```

```
   class Asthma BMI Education Income Race;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   output out=Asthma3FEFI;
run;
/*---FEFI using the third reduced set of variables (RSV3)          ---*/
proc surveyimpute data=Asthma method=fefi;
   var   Education Income Race;
   class Education Income Race;
   cells Gender;
   id UnitID;
   weight fwgt;
   repweights bsw:;
   output out=Asthma4FEFI;
run;
```

The following statements combine the four imputed data sets by taking appropriate observation units from each imputed data set:

```
/*---Units that are imputed by using FSV                            ---*/
%OneOrZeroDonor(FEFIData=Asthma1FEFI,OneDonorData=OneDonorFSV);
data AsthmaFEFI;
   merge Asthma1FEFI(in=_ina) OneDonorFSV(in=_inb);
   by UnitID;
   if _inb then delete;
run;
/*---Units that are imputed by using RSV1                           ---*/
%OneOrZeroDonor(FEFIData=Asthma2FEFI,OneDonorData=OneDonorRSV1);
data Asthma21FEFI;
   merge Asthma2FEFI(in=_ina) OneDonorRSV1(in=_inb);
   by UnitID;
   if _inb then delete;
run;
data A1FEFIID;
   set AsthmaFEFI;
   keep UnitID;
   by UnitID;
   if first.UnitID;
run;
data Asthma22FEFI;
   merge Asthma21FEFI(in=_ina) A1FEFIID(in=_inb);
   by UnitID;
   if _inb then delete;
run;
proc append base=AsthmaFEFI data=Asthma22FEFI;
run;
/*---Add UnitID 5350, which is imputed by using only one donor cell---*/
/*---but the donor cell contains two donor units --- 8317 and 8537 ---*/
proc append base=AsthmaFEFI data=Asthma2FEFI(where=(UnitID=5350));
run;
proc sort data=AsthmaFEFI;
   by UnitID;
run;
/*---Units that are imputed by using RSV2                           ---*/
%OneOrZeroDonor(FEFIData=Asthma3FEFI,OneDonorData=OneDonorRSV2);
data Asthma31FEFI;
   merge Asthma3FEFI(in=_ina) OneDonorRSV2(in=_inb);
   by UnitID;
   if _inb=1 then delete;
run;
data A2FEFIID;
   set AsthmaFEFI;
```

```
   keep UnitID;
   by UnitID;
   if first.UnitID;
run;
data Asthma32FEFI;
   merge Asthma31FEFI(in=_ina) A2FEFIID(in=_inb);
   by UnitID;
   if _inb then delete;
run;
proc append base=AsthmaFEFI data=Asthma32FEFI;
run;
proc sort data=AsthmaFEFI;
   by UnitID;
run;
/*---Units that are imputed by using RSV3                          ---*/
data A3FEFIID;
   set AsthmaFEFI;
   keep UnitID;
   by UnitID;
   if first.UnitID;
run;
data Asthma41FEFI;
   merge Asthma4FEFI(in=_ina) A3FEFIID(in=_inb);
   by UnitID;
   if _inb then delete;
run;
proc append base=AsthmaFEFI data=Asthma41FEFI;
run;
proc sort data=AsthmaFEFI;
   by UnitID;
run;
```

Finally, the following SAS macro creates a data set to contain the observation units that were either not imputed or imputed by using only one donor cell. You use the FEFIData= argument to input the imputed data set, and the OneDonorData= argument to provide a data set name to contain the output.

```
%macro OneOrZeroDonor(FEFIData=,OneDonorData=);
/*---Create a data set containing only one or zero donor cells     ---*/
/*---UnitID must be present in both data sets                      ---*/
options nonotes;
data &OneDonorData; set &FEFIData;
run;
data &OneDonorData;
   merge &OneDonorData
         &OneDonorData(firstobs=2 rename=(UnitId=NextUnit) keep=UnitID);
run;
data &OneDonorData; set &OneDonorData;
   if NextUnit=UnitID then delete;
   if Recipient ne 1 then delete;
   drop NextUnit;
run;
options notes;
data &OneDonorData; set &OneDonorData;
   keep UnitID;
run;
%mend OneOrZeroDonor;
```

## REFERENCES


Barnard, J., and Rubin, D. B. (1999). "Small-Sample Degrees of Freedom with Multiple Imputation." *Biometrika* 86:948–955.

Berglund, P., and Heeringa, S. (2014). *Multiple Imputation of Missing Data Using SAS*. Cary, NC: SAS Institute Inc.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: John Wiley & Sons.

Binder, D. A., and Sun, W. (1996). "Frequency Valid Multiple Imputation for Surveys with a Complex Design." In *Proceedings of the Survey Research Methods Section*, 281–286. Alexandria, VA: American Statistical Association.

Brick, J. M. (2013). "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29:329–353.

Brick, J. M., and Kalton, G. (1996). "Handling Missing Data in Survey Research." *Statistical Methods in Medical Research* 5:215–238.

De Leeuw, E. D., Hox, J., and Huisman, M. (2003). "Prevention and Treatment of Item Nonresponse." *Journal of Official Statistics* 19:153–176.

Fay, R. E. (1993). "Valid Inferences from Imputed Survey Data." In *Proceedings of the Survey Research Methods Section*, 41–48. Alexandria, VA: American Statistical Association.

Fay, R. E. (1996). "Alternative Paradigms for the Analysis of Imputed Survey Data." *Journal of the American Statistical Association* 91:490–498.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

Fuller, W. A., and Kim, J. K. (2005). "Hot Deck Imputation for the Response Model." *Survey Methodology* 31:139–149.

Ghosh, S., and Pahwa, P. (2008). "Assessing Bias Associated with Missing Data from Joint Canada/U.S. Survey of Health: An Application." In *Proceedings of the Biometrics Section*, 3394–3401. Alexandria, VA: American Statistical Association.

Haziza, D. (2009). "Imputation and Inference in the Presence of Missing Data." In *Sample Surveys: Design, Methods, and Applications*, edited by D. Pfeffermann, and C. R. Rao, 215–246. Vol. 29A of Handbook of Statistics. Amsterdam: North-Holland.

Hidiroglou, M. A., Drew, J. D., and Gray, G. B. (1993). "A Framework for Measuring and Reducing Nonresponse in Surveys." *Survey Methodology* 19:81–94.

Ibrahim, J. G. (1990). "Incomplete Data in Generalized Linear Models." *Journal of the American Statistical Association* 85:765–769.

Kalton, G., and Kasprzyk, D. (1986). "The Treatment of Missing Survey Data." *Survey Methodology* 12:1–16.

Kalton, G., and Kish, L. (1984). "Some Efficient Random Imputation Methods." *Communications in Statistics—Theory and Methods* 13:1919–1939.

Kim, J. K. (2002). "A Note on Approximate Bayesian Bootstrap Imputation." *Biometrika* 89:470–477.

Kim, J. K., Brick, J. M., Fuller, W. A., and Kalton, G. (2006). "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Soceity, Series B* 68:509–521.

Kim, J. K., and Fuller, W. A. (2004). "Fractional Hot Deck Imputation." *Biometrika* 91:559–578.

Kim, J. K., and Shao, J. (2014). *Statistical Methods for Handling Incomplete Data*. Boca Raton, FL: CRC Press.

Little, R. J. A., and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Lohr, S. L. (2010). *Sampling: Design and Analysis*. 2nd ed. Boston: Brooks/Cole.

Mukhopadhyay, P. K., An, A. B., Tobias, R. D., and Watts, D. L. (2008). "Try, Try Again: Replication-Based Variance Estimation Methods for Survey Data Analysis in SAS 9.2." In *Proceedings of the SAS Global Forum 2008 Conference*. Cary, NC: SAS Institute Inc. `http://www2.sas.com/proceedings/forum2008/367-2008.pdf`.

Rao, J. N. K. (1996). "On Variance Estimation with Imputed Survey Data." *Journal of the American Statistical Association* 91:499–506.

Rao, J. N. K., and Shao, J. (1992). "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation." *Biometrika* 79:811–822.

Rao, J. N. K., and Wu, C. F. J. (1988). "Resampling Inference with Complex Survey Data." *Journal of the American Statistical Association* 83:231–241.

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

Rubin, D. B. (1996). "Multiple Imputation after 18+ Years." *Journal of the American Statistical Association* 91:473–489.

Rubin, D. B., and Schenker, N. (1986). "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse." *Journal of the American Statistical Association* 81:366–374.

Särndal, C. E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Chichester, UK: John Wiley & Sons.

Shao, J., and Tu, D. (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.

Yuan, Y. (2000). "Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0)." In *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*. Cary, NC: SAS Institute Inc. Revised 2009. http://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf.

## ACKNOWLEDGMENTS

I am grateful to Randall D. Tobias, Donna L. Watts, Anthony B. An, and Anne H. Baxter of SAS Institute Inc. for their valuable assistance in the prepartion of this paper. I would also like to thank Wayne A. Fuller and Jae-kwang Kim of Iowa State University for their continuous support in developing the FEFI method.

## CONTACT INFORMATION

Pushpal K. Mukhopadhyay
SAS Institute Inc.
R4110 SAS Campus Drive
Cary, NC, 27513
919-531-2123
pushpal.mukhopadhyay@sas.com