WEIGHTED SEQUENTIAL HOT DECK IMPUTATION MACROS

Vincent G. Iannacchione, Research Triangle Institute

1. Introduction

Item nonresponse occurs when questions from an otherwise completed survey questionnaire are not answered. Since the population estimates formed by ignoring missing data are often biased, nonresponse adjustment procedures to reduce this bias should be considered.

Hot deck imputation is a commonly used nonresponse adjustment procedure that replaces missing data with available survey data. Conventional hot deck methods do not, however, consider the weighted distribution of the data in the imputation process. Weighted sequential hot deck imputation (Cox, 1980) replicates the weighted distribution of the available data in the imputed data by using the sample weights, or inverse selection probabilities, of item respondents and nonrespondents.

2. Description of the Procedure

The procedure takes account of the unequal probabilities of selection in the original sample by using the sample weights to specify the expected number of times that a particular respondent's answer will be used to replace a missing item. These expected selection frequencies are specified so that, over repeated applications of the algorithm, the expected value of the weighted distribution of substitute, or imputed, values will equal the weighted distribution of respondent answers.

This imputation strategy is based on the presumption that, over repeated initial samples, the weighted answer distributions for respondents and nonrespondents have the same expectation. This assumption is more plausible when the respondents and nonrespondents with similiar known characteristics are partitioned into subclasses refered to as imputation poststrata. The algorithm is then applied separately within each poststratum.

The sample of respondent answers are sequentially drawn and matched to nonrespondents within a poststratum. Therefore, additional control may be gained by purposively sorting respondents and nonrespondents in the same order. Williams and Chromy (1980) discuss a hierarchic serpentine sorting method that places observations with many characteristics in common close to each other.

3. Execution of the Macros

Although the imputation algorithm contains rather formidable mathematical formulas, it is easily programmed in SAS by creating two SAS data sets, one containing respondents and the other nonrespondents, and then interleaving them. This interleave positions each nonrespondent between one or more respondents with the ordering in the combined listing determined by the cumulative sum of the scaled sampling weights. When two or more respondents precede a nonrespondent, a weighted sample selection algorithm selects the respondent to be used for imputation. That respondent's data is then attached to the nonrespondent's record and placed on an output dataset of nonrespondents. The user defines macros that specify stratification, identification, and weight variables and that distinguish respondents from nonrespondents.

The execution of the algorithm is best described by way of an example. For simplicity, an input SAS data set with one imputation poststratum is used. The data set has two variables: ID, the identifier of the survey member, and WGT, the sampling weight.

This data set is first split into a respondent data set RESPS and a nonrespondent data set NONRESPS as in Step 1 of Figure 1. During the same step, the ratio of the sum of respondents' weights to the sum of nonrespondents' weights (80/40 in the example) is computed for each poststratum. In Step 2 this ratio is used to scale each nonrespondent's weight so that their sum equals that of the respondents. These scaled nonrespondent weights define selection zones from which a respondent is selected for imputation.

These zones are formed by interleaving the respondent and nonrespondent data sets by SUMWGTS, the cumulative sum of their scaled weights. During this interleave a nonrespondent's zone consists of all respondents after the previous nonrespondent along with part of the next respondent. Since the algorithm is executed sequentially, the zone length must be known at the beginning of each zone. Therefore, in Step 2, the variable NEXTZONE is created by lagging the WGT variable. Then, when the interleave brings in a nonrespondent the zone length of the next nonrespondent will be known. If the first observation of the interleave is a respondent instead of a nonrespondent, the first zone length will not be known. Therefore, an extra observation having the first zone length is added to the beginning of the nonrespondent data set. As shown in Step 2, the SUMWGTS variable is set to zero so that it is always interleaved first.

An example of the interleave process is shown in Step 3 of Figure 1. The first zone length is 12. Therefore, respondents R1 and R2 and part of R3 belong to this selection zone. The variable _EXPN is the expected number of times a respondent can be selected for imputation. Since R1 and R2 belong to zone 1, their _EXPN is 4/12 and 6/12 respectively. Portions of R3 are in zones 1, 2, and 3, so its _EXPN is 2/12 + 8/8 + 10/14. Notice the influence of the nonrespondents' weights on the values of _EXPN. Even though R2 and R8 have the same weight, R2

is more likely to be selected since it is in a smaller zone than R8.

A weighted sequential selection procedure (Chromy, 1979) determines _N, the actual number of times a respondent is used for imputation. The possible values of _N are the values of _EXPN when rounded up or down to the next integer, i.e. INT(_EXPN) or INT(_EXPN)+1. In the example R3 may be imputed to either NR1, NR2, or NR3. With certainty it will be imputed to NR2 since it completely covers that zone. However, since its _EXPN is 1.88 it can be imputed to either NR1 or NR3, but not both. This control over the imputation process restricts the number of times a particular respondent's data is used and thus insures that the imputed data reflects the weighted distribution of the respondents' data.

4. How to Use The Macros

4.1 User Defined Macros

The user specifies stratification, weight, and identification variable names, input and output data set names, and the random number seed by defining the following required and optional macros.

Required Macros

_ID, a SAS variable that uniquely identifies each observation on the input data set, e.g. MACRO _ID PID %. The variable _IMPID is the imputed respondent's identifier.

_RESPOND, a SAS expression that defines the set of respondents, e.g. MACRO _RESPOND RESP = 1 %, or MACRO _RESPOND _N_ LE 76 %. The complement of this expression must be the set of nonrespondents.

_VARS, a list of SAS variables to be used for imputation, e.g. MACRO _VARS INCOME AGE %. At least one variable must be specified.

_IMPVARS, a SAS variable of the form _IMPVARs where s is the number of imputation variables listed in the _VARS macro, e.g. MACRO _IMPVARS _IMPVAR2 %.

Optional Macros

_INDSET, a SAS data set name for an input data set, sorted by imputation poststrata, that contains both respondent and nonrespondent observations, e.g. MACRO _INDSET IN.PIDFILE %. If omitted, the most recently created SAS dataset is used.

_STRATA, a list of SAS variables (numeric variables only) defining the imputation poststrata, e.g. MACRO _STRATA STATE COUNTY %. If omitted, no stratification is used.

_WEIGHT, a SAS variable for the sampling weight assigned to both respondents and nonrespondents, e.g. MACRO _WEIGHT PIDWGT %. If omitted, equal weighting is used.

_SEED, a 5, 6, or 7 digit odd integer used as the seed for the random number generator, e.g. MACRO _SEED 12345 %. If omitted, the computer's clock provides the seed which means that the results cannot be reproduced on a subsequent run.

_OTHER, a list of other SAS variables to be taken from the input data set and placed on the output data set(s), e.g. MACRO _OTHER PSU SEGMENT %.

_NRDSET, a SAS data set name for an output data set of nonrespondents, e.g. MACRO _NRDSET OUT.NONRESP %. The data set will contain the following variables:

1) the variable specified in the macro _ID
2) the variable(s) specified in the macro _STRATA
3) the variable specified in the macro _WEIGHT
4) the variable(s) specified in the macro _OTHER (if any)
5) the variable(s) specified in the macro _VARS
6) _IMPVAR1 - _IMPVARs
7) _IMPID

If omitted, the name of the dataset is NRDSET.

_RDSET, a SAS data set name for an optional output data set of respondents, e.g. MACRO _RDSET OUT.RESP %. The data set contains the following variables:

1) the variable specified in the macro _ID
2) the variable specified in the macro _WEIGHT
3) the variable(s) specified in the macro _STRATA
4) the variables(s) specified in the macro _OTHER, if any
5) _N, the actual number of time the respondent was used for imputation
6) _EXPN, the expected number of times the response is used for imputation

If omitted, the dataset will not be created.

4.2 The Macro _HOTDECK

After the above macros have been defined, the macro _HOTDECK is called to perform the actual imputations. _HOTDECK creates one temporary respondent data set, RESP1, and two temporary nonrespondent datasets, NONRESP1 and NONRESP2, as shown in Steps 1 and 2 of Figure 1. The interleave of NONRESP2 and RESP1 by SUMWGTS then defines the selection zones and determines the _EXPN as shown in Step 3. At this point the sequential selection procedure is called to choose a respondent from each zone. Since a respondent may be linked to many nonrespondents, the imputed data are attached to the data set NONRESP1 during the interleave as follows.

```
DATA _NRDSET(KEEP=ID WGT _IMPID)
     _RDSET (KEEP=ID WGT _EXPN _N);
SET NONRESP2 (IN=INNR)
    RESP1 (IN=INR);
BY SUMWGTS;
*For each nonrespondent;
IF INNR THEN DO; bring in next zone
                 begin computing _EXPN
             END;
*For each respondent;
IF INR;
    finish computing _EXPN
    determine _N
    OUTPUT _RDSET;
    _IMPID=ID;
    Do I=1 to _N;
       SET NONRESP1;
       OUTPUT _NRDSET;
    END;
```

Notice that if a respondent is not selected, i.e. _N = 0, the data set NONRESP1 is not read. If _N > 0 the values of ID and WGT from NONRESP1 replace the values of ID and WGT from RESP1.

The program provided in Figure 2 illustrates this process for the data set given in Figure 1. The variable AGE is imputed to the five nonrespondents.

5. Concluding Remarks

The macros described above have been used extensively for missing data imputation in two large national health surveys. The types of data imputed include income, employment, disability days, health insurance premium, and medical visit charges. The data sets involved were as large as 200,000 records.

At present, the variance attributable to this type of imputation has not been derived. However, software is available to perform up to five independent replicated imputations from which imputation variance estimates conditional on the original sample can be obtained.

6. References

Cox, Brenda G. (1980). The Weighted Sequential Hot Deck Imputation Procedure. Proceedings of the American Statistical Association, Survey Research Methods Section.

Chromy, James R. (1979). Sequential Sample Selection Methods. Proceedings of the American Statistical Assocation, Survey Methodology Section.

Williams, Rick L. and James R. Chromy (1980). SAS Sample Selection Macros. Proceedings of the Fifth Annual SAS Users Group International Conference.

Figure 1. Example of the Formation of Selection Zones Within an Imputation Poststratum

STEP 1. Split respondents and nonrespondents into two SAS data sets, and sum weights.

RESP1		
PID	WGT	SUMWGTS
R1	4	4
R2	6	10
R3	20	30
R4	3	33
R5	4	37
R6	8	45
R7	9	54
R8	6	60
R9	7	67
R10	13	80

NONRESP1		
PID	WGT	SUMWGTS
NR1	6	6
NR2	4	10
NR3	7	17
NR4	13	30
NR5	10	40

STEP 2. Scale nonrespondents' weights; create variable NEXTZONE by lagging WGT; add an observation so that the first zone length is known at beginning of interleave.

NONRESP2			
PID	WGT	NEXTZONE	SUMWGTS
.	.	12	0
NR1	12	8	12
NR2	8	14	20
NR3	14	26	34
NR4	26	20	60
NR5	20	.	80

STEP 3. Interleave NONRESP2 and RESP1 by their cumulative sum of weights to form selection zones, i.e.

SET NONRESP2 RESP1; BY SUMWGTS;

PID	WGT	NEXTZONE	_EXPN	SUMWGTS	(possible) IMPID
.	.	12		0	
R1	4		.33	4	
R2	6		.50	10	
NR1	12	8		12	R1,R2, or R3
NR2	8	14		20	R3
R3	20		1.88	30	
R4	3		.21	33	
NR3	14	26		34	R3,R4, or R5
R5	4		.19	37	
R6	8		.31	45	
R7	9		.35	54	
NR4	26	20		60	R5,R6,R7, or R8
R8	6		.23	60	
R9	7		.35	67	
NR5	20	.		80	R8,R9, or R10
R10	13		.65	80	

Figure 2. An Example Program Using the Data Set Shown in Figure 1

```
//Jobcard
// EXEC SAS
//SYSIN DD DSN=HOTDECK.MACROS
//      DD *
DATA EXAMPLE;
     INPUT ID $ WGT AGE @@;
CARDS;
R1 4 14 R2 6 27 R3 20 35 R4 3 7 R5 4 49
R6 8 15 R7 9 29 R8  6 21 R9 7 40 R10 13 11
NR1 6 . NR2 4 . NR3 7 . NR4 13 . NR5 10 .
;
MACRO _INDSET EXAMPLE %
MACRO _NRDSET NONRESPS %
MACRO _RDSET RESPS %
MACRO _ID ID %
MACRO _RESPOND AGE NE . %
MACRO _VARS AGE %
MACRO _IMPVARS _IMPVAR1 %
MACRO _WEIGHT WGT %
MACRO _SEED 13589 %
_HOTDECK
PROC PRINT DATA=RESPS;
     ID ID;
     VAR WGT _EXPN _N AGE;
     SUM WGT _EXPN _N;
     TITLE RESPONDENT OUTPUT DATA SET;
PROC PRINT DATA=NONRESPS;
     ID ID;
     VAR WGT AGE _IMPID _IMPVAR1;
     SUM WGT;
     TITLE NONRESPONDENT OUTPUT DATA SET;
/*
```

RESPONDENT OUTPUT DATA SET

ID	WGT	_EXPN	_N	AGE
R1	4	0.33333	0	14
R2	6	0.50000	1	27
R3	20	1.88095	2	35
R4	3	0.21429	0	7
R5	4	0.18681	0	49
R6	8	0.30769	0	15
R7	9	0.34615	0	29
R8	6	0.23077	1	21
R9	7	0.35000	0	40
R10	13	0.65000	1	11
	80	5.00000	5	

NONRESPONDENT OUTPUT DATA SET

ID	WGT	AGE	_IMPID	_IMPVAR1
NR1	6	.	R2	27
NR2	4	.	R3	35
NR3	7	.	R3	35
NR4	13	.	R8	21
NR5	10	.	R10	11
	40			