NORCSuite_Impute: A Two-Way Search Hot-Deck Imputation Macro using SAS IML

Fang Wang, Steven Pedlow, Yongyi Wang

NORC at the University of Chicago

Abstract

Traditional hot-deck imputation macros utilize SAS arrays and data steps. They keep only one nearest donor above the recipient in the sort order in the memory without limiting the number of times that a donor can be used. The new NORCSuite Impute macro using SAS IML has the following new functionalities: 1) a two-way search (above and below) algorithm that selects the global nearest donor without violating the user-specified donor use limit to control imputation bias; and 2) the number of donors kept in memory is raised automatically to avoid violating the donor use limit, and to minimize the underestimation of data variance inherent in hot-deck. The new macro also supports serpentine sorting using multiple sort variables with different variable types, imputing a set of variables together, and applying unequal donor use limits to different subsets of the data. In this paper, we will compare our new and previous hot deck imputation macros in terms of mean, variance (the amount of underestimation) and imputation bias.

Key words: hot-deck imputation, two-way search, donor use limit, number of donors

1. Background

Hot-deck imputation is a well-respected method to deal with missing values in surveys due to its many advantages. It preserves the distribution of data and is efficient and reliable for surveys with a large number of data records. In hot-deck imputation, a non-missing value is donated to a record with a missing data item from a donor record thought to be similar to the recipient. All records are categorized into homogeneous classes, using variables strongly associated with the variables being imputed. Within classes, records are sorted by variables related to the imputed variables, so that data records ending up adjacent to one another are as similar as possible.

One popular hot-deck imputation method is unweighted sequential nearest neighbor hot-deck imputation. "Unweighted" means the case weight (a sampling weight, for example, in survey research) is ignored in determining which donor is used, while "sequential nearest neighbor" means the last encountered non-missing observation from within the same imputation class is used to impute the currently encountered missing value. The class variables divide the data into mutually exclusive blocks. Donors and recipients must agree on the class variables. But within these blocks, donors and recipient can be different on the sort variables. To ensure that adjacent records are as similar as possible, serpentine sorting is preferred.

A SAS macro to carry out the hot-deck imputation procedure described above was presented by Barbara Lepidus Carlson, Brenda Cox and Linda S. Bandeh at the Twentieth Annual SAS User's Group International Conference on April 2-5, 1995, in "SAS® Macros Useful in Imputing Missing Survey Data." This macro has been inherited

and improved by NORC and used in many large NORC surveys such as the Survey of Doctorate Recipients, the National Immunization Survey, and the Residential Energy Consumption Survey. There are several limitations for the 1995 macro:

1. There is no way to limit the number of times a donor can be used. Only one non-missing record can be held as donor in the imputation for a missing record. Using the same record as donor too many times will further increase the underestimation of variance that already occurs in hot-deck imputation, and could skew the data distribution.

2. With multiple sort variables specified, the donor can only come from "previous" records (non-missing record above the missing record in the dataset). No "next" records (below) can be used as donors. This restriction can lead the imputation procedure to be a biased process.

3. Only one sort variable can be used in a two way search. However, hot-deck imputation relies on the relationship of the missing variable to the sort variables. Only having one sort variable is insufficient to be used in practice.

2. NORCSuite_Impute

In this paper, we describe the SAS macro NORCSuite_Impute, a two-way search unweighted hot-deck imputation macro with a user-specified donor use limit and a flexible number of donors applied to avoid violating the donor use limit. This procedure first divides the dataset into mutually-exclusive blocks by user-specified class variables and performs serpentine sorting within each block by user-specified sort variables. Next, each block is imputed separately using SAS IML. Finally, quality control examination of the imputed variables is conducted and imputed values are merged back into the original dataset.

Since this macro performs serpentine sorting using PROC SURVEYSELECT, there is no requirement on the type of sort and class variables (numeric or character). Mixed variable types are allowed. After the sorted dataset is divided into mutually-exclusive blocks by class variables, the variable(s) to be imputed in a block is read into a matrix with SAS IML. The donor can come from either above or below a missing record in the imputation matrix.

The macro parameter "donor use limit" is specified by the user. A record is considered an eligible donor as long as it has not exceeded the donor use limit. A variable "count" is created and updated throughout imputation to record the number of times a non-missing record has been used as a donor. For a record at row i, initially for any i count[i]=0; when a record is used as donor, then

$$count[i] = count[i] + 1$$

The distance from a donor to a recipient is computed as

$$distance = abs(\text{row number of the donor} - \text{row number of the recipient})$$

Both the "distance" and "eligibility" of each donor are considered in selecting the "global-nearest eligible" donor, which is the donor from either above or below the recipient with the smallest distance without exceeding the donor use limit. As the imputation process begins, the record with the smallest distance (whether it is above or below the recipient) is first considered as a possible donor. If the eligibility check suggests that it has already exceeded the donor limit, the second nearest donor is considered. If the distances from two donors are the same and both have not exceeded the donor use limit, one donor is randomly selected from the two. If both have exceeded the donor use limit, the number of donors to be considered is automatically raised by one, which means one more possible donor in each direction is searched until an eligible donor is found.

Although NORCSuite_Impute can search for as many donors as needed to find one that has not exceeded the donor use limit, an upper bound of "number of donors considered" is specified by the user to make sure the distance between the donor and the recipient is within a reasonable range. If this upper bound has been reached but the donor use limit is still exceeded, the donor use limit will be raised automatically. Note that the "number of donors considered " can be set to be a number larger than the sample size, which means no upper bound is imposed; similarly, the "donor use limit" can be set to be a number larger than the sample size, which means no donor use limit is in effect. Since each block is imputed separately, even if the donor use limit has to be raised in one block, other blocks can maintain the desired pre-set donor use limit. This is also true for the number of donors to be considered.

It is important to note that NORCSuite_Impute handles simultaneous imputation for multiple variables without wiping out the reported values. If a case is missing on some variables but not on the others, the missing variables would be imputed while the non-missing variables would keep the reported values.

To summarize, the imputation process includes the following steps:

1. Divide the data into blocks by the combination of class variables and serpentine sort the dataset within each block by the sort variables to impute each block separately.
2. Read the imputation variables of a block into a matrix.
3. Select one nearest donor above and one nearest donor below. The parameter "ndonor" indicates how many donors are used in each direction; initially, ndonor=1 (two in total).
4. Check the donor to recipient distances and donor use counts to choose the donor with the smallest distance without violating the donor use limit.
5. If both the nearest donors above and below exceed the donor use limit, use the second nearest donors above and below; if the donor use limit is still violated, repeat this step: raise the number of donors by adding one more donor from each direction continuously until the block is imputed without exceeding the donor use limit.
6. If the number of donors used reaches the upper bound before meeting the donor use limit, keep the number of donors used at the upper bound and raise the donor use limit by one. This step is repeated until the entire block can be imputed without violating the donor use limit.
7. Perform quality control checks, combine the imputed blocks back together, and merge the post-imputation variables back into the pre-imputation dataset.

3. Compare NORCSuite_Impute with the 1995 hot-deck imputation macro

To prove the advantages of this new hot-deck imputation procedure, a simulation study was conducted to compare the imputation results from the new NORCSuite_Impute and the 1995 published hot deck imputation macro. The dataset utilized for this study is from the 2010 National Immunization Survey – Teen (NIS-Teen) Public Use File. The NIS-Teen is the largest survey ever to assess vaccination levels of adolescents 13-17 years of age in the U.S., and is conducted for the Centers for Disease Control and Prevention by NORC. The NIS-Teen collects data by interviewing households randomly selected from 56 areas: all 50 States, the District of Columbia, and 5 areas designated for oversampling. The 2010 NIS-Teen public use file and corresponding codebook can be downloaded at http://www.cdc.gov/nchs/nis/data_files_teen.htm.

The variable to be imputed, class variables and sort variables for the hot-deck imputation are listed below. For detailed coding rules, please check the 2010 NIS-Teen PUF codebook.

Variable to be imputed:

- INCQ298A Family income category

Class Variables

- EDUC1 Education level of mother with 4 categories

Sort Variables (sorted by sort order in the left column)

1. LANGUAGE Language of interview (English, Spanish, Other)
2. MARITAL Marital status of mother
3. RACE_K Race of teen with multi-race category
4. CEN_REG Census region based on state of residence
5. AGEGRP_M_I Mother's age category
6. I_HISP_K Is teen Hispanic or Latino?
7. CHILDNM Number of children under 18 years of age in household
8. NUM_PROVR Number of valid and unique health providers identified by respondent

The class variable EDUC1 divided the entire dataset into 4 blocks with mother's education from lowest to highest. The 8 sort variables are listed from top to bottom in the sort order.

To simulate the non-response indicators, we built a propensity model to estimate the individual probability of item non-response using all the records in the dataset. The propensity model is:

$$Logit(INCQ298A\_FLAG) = PDAT + EDUC1 + LANGUAGE + MARITAL + RACE\_K + CEN\_REG + AGEGRP\_M\_I + I\_HISP\_K + CHILDNM + NUM\_PROVR$$

*PDAT: ADEQUATE PROVIDER DATA FLAG (only used in the model)

For the logistic regression model, the dependent variable is the non-response indicator of the INCQ298A in the original public use file (if INCQ298A is missing in the original data, INCQ298A = 1, otherwise 0); 10 independent variables are included in the model: the class variable EDUC1, eight sort variables and one extra variable PDAT.

A non-response indicator for each subject was drawn independently from a Bernoulli distribution using the item non-response probability estimated in the propensity model. Then we induced missingness into the observed non-missing cases according to this simulated non-response indicator. The data with records and pseudo non-records was used as our input file for imputation. Cases with real missing values in the original data were dropped from the imputation input file because we can study the differences between observed and imputed data only. This process of creating non-records was repeated 100 times; thus, 100 different input files were generated and evaluated to account for the variance caused by various missing patterns of the input file.

While using NORCSuite_Impute, researchers can specify a donor use limit (N) and a maximum number of donors considered in either direction (M) as needed. For this simulation study, we specified N = 4 and M = 2 (each donor can be used up to 4 times; two donors are considered in each direction - 4 in total). By comparison, the 1995 macro does not impose a donor use limit and only the nearest non-missing record from above the missing case can be used as a donor.

First, we compared the average statistics across the 100 imputed datasets and listed the results in Table 1. The following variables are analyzed:

1. INCQ298A0: the complete variable without missing values;
2. INCQ298A: the incomplete, un-imputed variable with non-missing records and simulated missing records (the variable to be imputed);
3. Imp_new: the imputed INCQ298A using NORCSuite_Impute;
4. Imp_95: the imputed INCQ298A using the 1995 hot-deck macro.

Table 1: Average Mean, Variance and Correlation from 100 Simulations

	INCQ298A0	INCQ298A	Imp_new	Imp_95
Average Mean	11.3430	11.3487	11.3440	11.3404
Average Variance	11.1908	11.1654	11.1924	11.2043
Average Correlation	N/A	N/A	0.945620	0.944410

$Mean_i$ is the mean value of the variable in the i-th simulation (i = 1 …100)

$$Average\ Mean = \frac{1}{100}\sum_{1}^{100} Mean_i$$

Var_i is the variance of the variable in the i-th simulation (i = 1 …100)

$$Average\ Variance = \frac{1}{100}\sum_{1}^{100} \mathrm{Var_i}$$

$Corr_i$ is the Pearson correlation between the imputed variable and the complete control variable INCQ298A0 in the i-th simulation (i = 1 …100).

$$Average\ Correlation = \frac{1}{100}\sum_{1}^{100} \mathrm{Corr_i}$$

By comparing the three columns INCQ298A, Imp_new and Imp_95 with the column INCQ298A0 which is the complete variable, we observe that:

1. Imp_new is closest to the true variable INCQ298A0 in terms of average mean and variance; this is followed by Imp_95 and then the un-imputed variable INCQ298A.
2. The average correlation between Imp_new and INCQ298A0 is larger than the correlation between Imp_95 and INCQ298A0.

Table 1 shows that NORCSuite_Impute provides better imputation results than the 1995 hot-deck macro in terms of mean, variance and correlation, and the imputed variable created by both macros are closer to the full data mean (INCQ298A0) than the unimputed variable (INCQ298A).

To compare the overall statistics, we combined the 100 imputed datasets and analyzed it as one single file with 100*29394 records (Table 2). In addition to INCQ298A0, INCQ298A, Imp_new and Imp_95, 4 more variables were generated:

1. INCQ298A_bnew: absolute imputation bias between Imp_new and INCQ298A0 (|Imp_new - INCQ298A0|**)**. For non-missing records, INCQ298A_bnew = 0; for imputed records, INCQ298A_bnew ≥ 0.
2. INCQ298A_b95: absolute imputation bias between Imp_95 and INCQ298A0 (|Imp_95 - INCQ298A0|).
3. True_new: Indicator of whether the Imp_new value is within plus or minus 1 of the value in the true variable INCQ298A0.
   For imputed values:
   True_new = 1 if INCQ298A0 - 1 ≤ Imp_new ≤ INCQ298A0 + 1, 0 otherwise;
   For un-imputed records, True_new is missing.
4. True_95: Indicator of whether the Imp_95 value is within plus or minus 1 of the value in the true variable INCQ298A0.
   For imputed values:
   True_95 = 1 if INCQ298A0 - 1 ≤ Imp_95 ≤ INCQ298A0 + 1, 0 otherwise;
   For un-imputed records, True_95 is missing.

Table 2 Overall Statistics of the Combined Dataset

Variable	Nobs	N Miss	Sum	Mean	Variance
INCQ298A0	2,939,400	0	33,341,700	11.3430	11.1905
INCQ298A	2,637,469	301931	29,931,736	11.3487	11.1650
Imp_new	2,939,400	0	33,344,530	11.3440	11.1921
Imp_95	2,939,400	0	33,334,108	11.3404	11.2040
INCQ298A_bnew	2,939,400	0	693,500	0.2359	1.1615
INCQ298A_b95	2,939,400	0	703,184	0.2392	1.1879
True_new	301,931	2637469	153,136	0.5072	0.249949
True_95	301,931	2637469	152,072	0.5037	0.249987

There are **2,637,469** non-missing records and **301,931** missing records in the variable INCQ298A. Comparing Imp_new and Imp_95 in Table 2, the combined NORCSuite imputed variables has:

1. smaller sum of bias (the INCQ298A_bnew and INCQ298A_b95 rows);
2. larger accuracy rate (the True_new and True_95 rows); and
3. closest overall mean and variance to the complete variable (the first four rows).

Table 1 and Table 2 both indicate that NORCSuite_Impute provides better imputed results: a more accurate mean and variance, smaller imputation bias, larger correlation with the complete variable and a higher imputation accuracy rate.

4. Conclusion:

NORCSuite_Impute allows a two-way search to select the global nearest donor with a user-specified number of donors searched in each direction and a maximum number of times a donor can be used. It automatically increases the number of donors searched in each direction in order to avoid exceeding the donor use limit; the class and sort variables can be a mix of numeric and character variables; a series of variables to be imputed together.

NORCSuite_Impute provides better imputed results according to the mean, variance, sum of absolute bias, correlation to the complete variable and accuracy rate compared with the 1995 hot deck imputation macro. But both NORCSuite_Impute and MPR 95 imputed variables provide better estimates than the un-imputed variable.

5. Acknowledgement

The authors would like to offer their sincere appreciation to Kirk Wolter and Dan Kasprzyk for their helpful support and research advice.

Reference

SAS OnlineDoc® 9.1.3, Copyright © 2002-2005, SAS Institute Inc., Cary, NC, USA; All rights reserved. Produced in the United States of America.

Barbara Lepidus Carlson, Brenda Cox and Linda S. Bandeh. April 2-5, 1995: "SAS® Macros Useful in Imputing Missing Survey Data," at 20th SAS User's Group International Conference on.

Contact Information

Fang Wang, Survey Statistician I
NORC at the University of Chicago
55 East Monroe Street, 20th Floor, Chicago IL 60603
Work Phone: (312) 325-2558
Email: wang-fang@norc.org

Steven Pedlow, Sr. Survey Statistician II
NORC at the University of Chicago
55 East Monroe Street, 20th Floor, Chicago IL 60603
Work Phone: (312) 759-4084
Email: pedlow-steven@norc.org

Yongyi Wang, Sr. Survey Statistician I
NORC at the University of Chicago
55 East Monroe Street, 20th Floor, Chicago IL 60603
Work Phone: (312) 759-4067
Email: wang-yongyi@norc.org