Paper 213-2013

Weighted Sequential Hot Deck Imputation: SAS Macro vs. SUDAAN's PROC HOTDECK.

David Izrael and Michael P. Battaglia,
Abt Associates Inc.

ABSTRACT

Item non-response is a challenge faced by virtually all surveys. Item non-response occurs when a respondent skips over a question, refuses to answer a question, or indicates that they do not know the answer to a question. Hot deck imputation is one of the primary item non-response imputation tools used by survey statisticians. Recently, new competitor in the field of Weighted Sequential Hotdeck Imputation has arrived: PROC HOTDECK of SUDAAN®, version 10. We compared the results of imputation using the new procedure with the results of the Hotdeck SAS® Macro with respect to: a) how close the post-imputation weighted distributions and standard errors of the estimates are to those of the item respondent data; b) whether there is a difference in the number of times donors contribute to the imputation.

INTRODUCTION

If a substantial percentage of respondents do not have a value on a survey question due to item non-response and these item non-respondents have a different distribution on the survey question, then that survey question will be subject to item non-response bias. Hot deck imputation fills in these missing question values using the following steps:

- Form imputation classes that do not have missing values and are correlated with the variables that have missing data, and
- Sort the respondents by the imputation classes and identify the respondents which will serve as donors within each imputation class.

Weighted Sequential Hot Deck Imputation (WSHD) differs from the hot deck in that the sample weights are used in selection of a donor. WSHD gives the resulting weighted distribution of imputed variables closer to the distribution of original variables of item responders which is a main measure of the quality of imputation. Besides that, the number of times the donor is selected for imputation is proportional to its weight. The topic of weighted sequential hot deck imputation received extensive consideration since introduction in 1980 [1]. An excellent review of the existing imputation tools can be found in [2].

The SAS macro implementing the WSHD was originated in 1980's by Research Triangle Institute (RTI) and since then has been developed, polished and expanded by many programmers including the authors of this paper. Recently, SUDAAN, version 10, presented the new PROC HOTDECK procedure based on Cox [1] and Iannacchione [3] and we asked ourselves how close would be the results of imputation by SAS Macro which we have been using for many years with the results of the new PROC HOTDECK.

TEST PROCESS

We compared the results of imputation using the mentioned SAS Macro and PROC HOTDECK using a national survey data set with 26,729 observations and variables of interest – Education and/or Marital status missing in 2,837 observations. We imputed the missing values using imputation cells created by the variables STATE, AGE GROUP, and SEX. We wanted both variables to be imputed simultaneously by the same donor.

The syntax of PROC HOTDECK looks like the following:

```
PROC HOTDECK data=mydata seed = 839276 filetype=sas;
weight final_weight;;
impby  STATE AGE_GRP SEX;                          /* IMPUTATION CELL */
```

```
impvar education marital_status;                    /* IMPUTED VARIABLES */
impid abtid;                                        /* UNIQUE ID         */
impcond  education >0 & marital_status>0;           /* CONDITIONS FOR DONORS *

impname education="i_education" marital_status =
 "i_marital_status";                  /* ASSIGN NAMES OF IMPUTED VARIABLES */
idvar   race2  children;              /* OTHER VARIABLES TO RETAIN */
output impid impby weight       /* VARIABLES RELATED TO IMPUTATION TO RETAIN */
idvar origval imputeval donorid
/ filename=hdout replace;
run;
```

The respective SAS macro call (actually, two macros are involved) looks like the following:

```
%IMP_VAR(EDUCATION,I_EDUCATION);
%IMP_VAR(MARITAL_STATUS,I_MARITAL_STATUS);

 %HOTDECK(abtid,
     EDUCATION > 0 and MARITAL_STATUS >0,
    &IMP_VARS, STATE AGE_GRP SEX, indset=mydata, weight=final_weight,
nrdset=NRDSET,sort=YES,seed=&seed);
```

Weighted distribution for item respondents is shown in Table 1:

Table 1. Weighted distribution for item respondents

Table of education by marital_status				
education	marital_status			
Frequency Percent Row Pct Col Pct	Married	Never married, member unmarried couple	Divorced, Widowed, Separated	Total
Less than HS	136013 4.73 51.75 7.18	39923.7 1.39 15.19 10.34	86908.8 3.02 33.06 14.60	262845 9.14
HS Grad	505596 17.59 61.82 26.70	114917 4.00 14.05 29.76	197282 6.86 24.12 33.14	817796 28.45
Some College	481818 16.76 62.17 25.45	113220 3.94 14.61 29.32	180013 6.26 23.23 30.24	775051 26.96
College Grad	769880 26.78 75.54 40.66	118103 4.11 11.59 30.58	131146 4.56 12.87 22.03	1019129 35.45
Total	1893307 65.86	386164 13.43	595350 20.71	2874821 100.00

The distribution after using PROC HOTDECK and the SAS Hotdeck macro to impute the 2,837 item non-respondents are shown in Table 2 and Table 3 respectively.

Table 2. Weighted post-imputation distribution after PROC HOTDECK

Table of i_education by i_marital_status				
i_education(Imputation #1 for Imputed variable #1)	i_marital_status(Imputation #1 for Imputed variable #2)			
Frequency Percent Row Pct Col Pct	Married	Never married, member unmarried couple	Divorced, Widowed, Separated	Total
Less than HS	159518 4.98 53.79 7.55	43262.8 1.35 14.59 10.12	93781.9 2.93 31.62 14.12	296562 9.25
HS Grad	566071 17.66 62.06 26.78	128694 4.02 14.11 30.11	217436 6.78 23.84 32.73	912201 28.46
Some College	538829 16.81 62.54 25.50	123769 3.86 14.37 28.96	198944 6.21 23.09 29.95	861542 26.88
College Grad	849023 26.49 74.82 40.17	131638 4.11 11.60 30.80	154135 4.81 13.58 23.20	1134795 35.41
Total	2113440 65.94	427364 13.33	664297 20.73	3205101 100.00

Table 3. Weighted post-imputation distribution after SAS Hot deck macro

I_EDUCATION(Post-Imputation value of EDUCATION)	I_MARITAL_STATUS(Post-Imputation value of MARITAL_STATUS)			
Frequency Percent Row Pct Col Pct	Married	Never married, member unmarried couple	Divorced, Widowed, Separated	Total
Less than HS	155492 4.85 52.57 7.37	44719.1 1.40 15.12 10.36	95571.7 2.98 32.31 14.42	295783 9.23
HS Grad	570408 17.80 62.18 27.02	127342 3.97 13.88 29.51	219580 6.85 23.94 33.14	917330 28.62
Some College	531324 16.58 62.23 25.17	127924 3.99 14.98 29.64	194601 6.07 22.79 29.37	853849 26.64
College Grad	853718 26.64 75.01 40.44	131551 4.10 11.56 30.48	152869 4.77 13.43 23.07	1138139 35.51
Total	2110943 65.86	431536 13.46	662621 20.67	3205101 100.00

The reader can easily observe the closeness of both distributions to the distribution of original variables for responders. Figure 1 and Figure 2 displays almost identical total percent for Marital status and Education respectively for item responders, SAS Hotdeck macro, and SUDAAN's PROC HOTDECK.

Figure 1. Marital Status

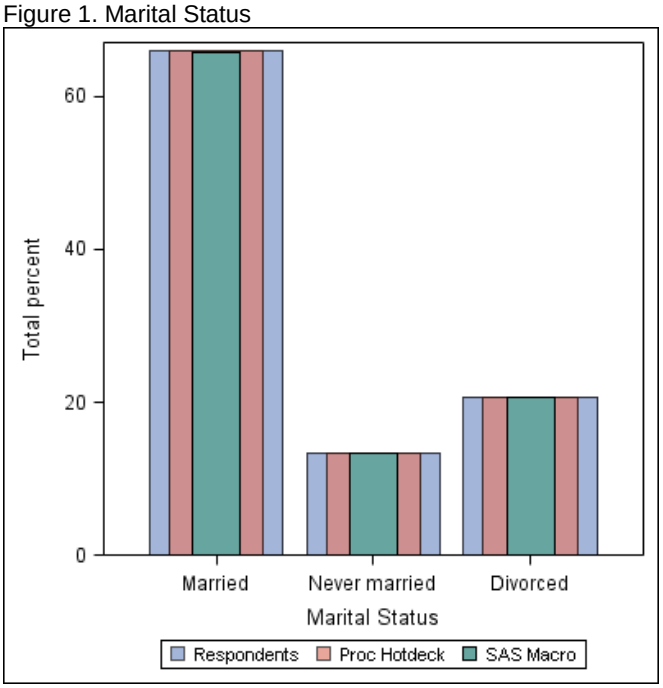
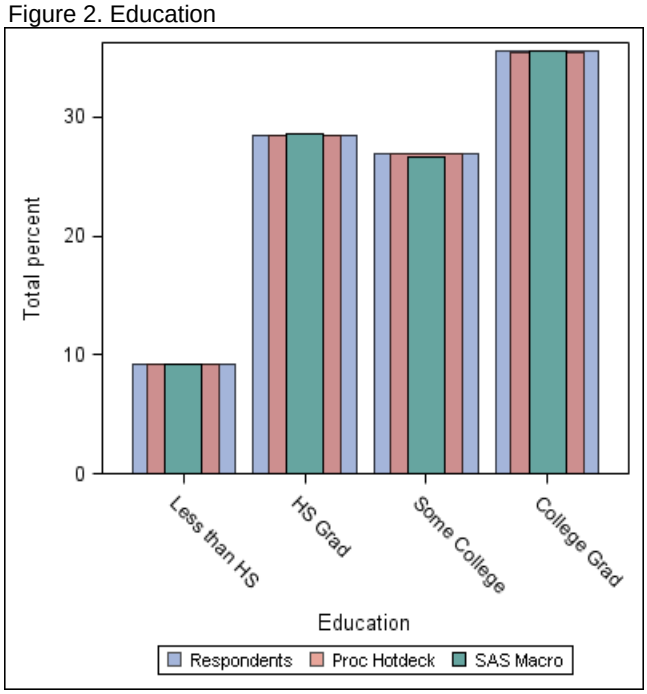


Figure 2. Education



To calculate standard error of percent for Education, Marital Status, and combination of those two variables we used PROC SURVEYFREQ. Table 4, Table 5, and Table 6 show standard errors for item respondents, imputed PROC HOTDECK, and SAS macro HOTDECK respectively:

Table 4. Standard Error of Percent for Education

education	Std Error of Percent		
	Respondents	PROC HOTDECK	SAS Macro
Less than HS	0.2943	0.2844	0.2915
HS Grad	0.4547	0.4429	0.4434
Some College	0.4506	0.4384	0.4387
College Grad	0.4812	0.4671	0.4705

Table 5. Standard Error of Percent for Marital Status

marital_status	Std Error of Percent		
	Respondents	PROC HOTDECK	SAS Macro
Married	0.4640	0.4507	0.4540
Never married, member unmarried couple	0.3553	0.3422	0.3508
Divorced, Widowed, Separated	0.3676	0.3598	0.3583

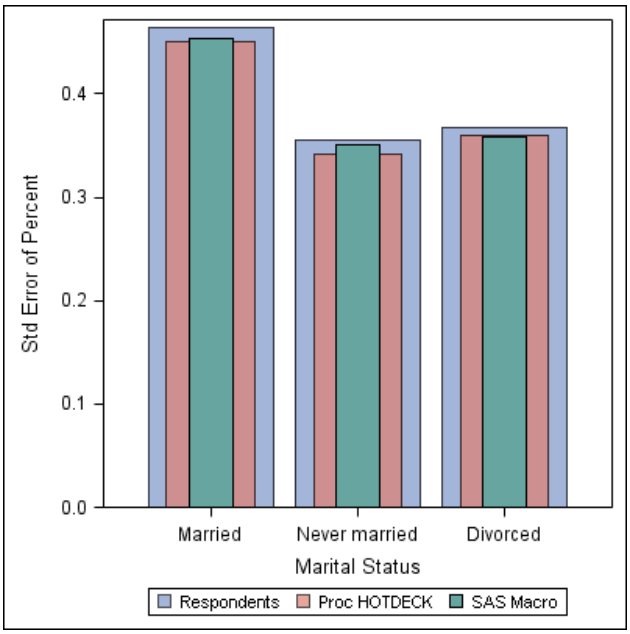
Table 6. Standard Error for Percent of combined distribution

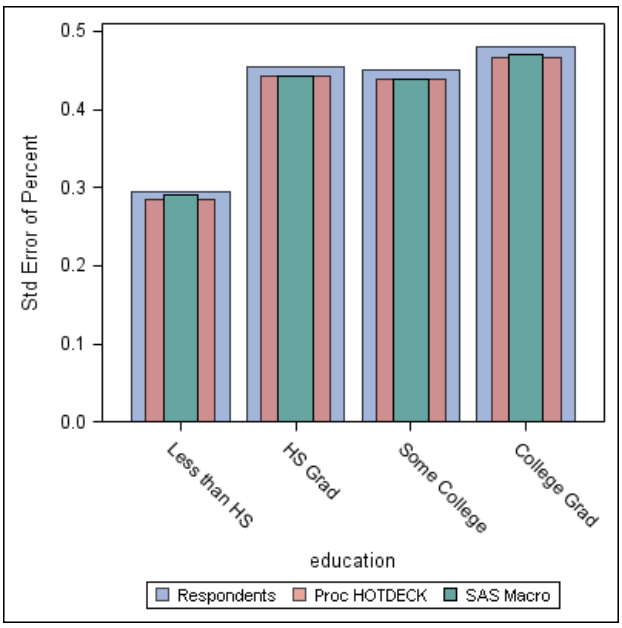
Education	marital_status	Std Err of Percent		
		Respondents	PROC HOTDECK	SAS macro
Less than HS	Married	0.2370	0.2311	0.2313
	Never married, member unmarried couple	0.1231	0.1145	0.1165
	Divorced, Widowed, Separated	0.1657	0.1556	0.1546
	Total	0.3037	0.2911	0.2915
HS Grad	Married	0.4112	0.3867	0.3872
	Never married, member unmarried couple	0.2100	0.1994	0.1997
	Divorced, Widowed, Separated	0.2255	0.2120	0.2133
	Total	0.4700	0.4427	0.4434
Some College	Married	0.4002	0.3802	0.3761
	Never married, member unmarried couple	0.2233	0.2063	0.2111
	Divorced, Widowed, Separated	0.2241	0.2109	0.2076
	Total	0.4669	0.4407	0.4387
College Grad	Married	0.4690	0.4397	0.4411
	Never married, member unmarried couple	0.2145	0.2019	0.2015
	Divorced, Widowed, Separated	0.1783	0.1740	0.1746
	Total	0.4990	0.4697	0.4705
Total	Married	0.4811	0.4533	0.4540
	Never married, member unmarried couple	0.3712	0.3482	0.3508
	Divorced, Widowed, Separated	0.3796	0.3591	0.3583

The standard errors are slightly lower after imputing the missing values because the sample size is now larger. Figure 3 and 4 confirm this. But the lower standard error does not reflect the uncertainty in the imputed values. To get valid standard errors one would need to use multiple imputations. PROC HOTDECK in SUDAAN has an option of multiple imputations.

Figure 3. Marital Status

Figure 4. Education





OBSERVED PECULIARITIES

In our imputation work, we require that the number of donors in a donor's cell exceed 5. As a preliminary step, we review the donor's cell distribution and, if the mentioned condition is not met, we collapse cells correspondingly. Frequently we have a situation when there are no donors in one or more cells at all. In this case, the SAS Macro terminates and the log displays the values of the variables that create the empty imputation cell. In contrast, PROC HOTDECK ignores the situation with an empty donor's cell not giving any warning and the respective recipients are just assigned missing values; that is, imputation in those empty donor cells is not performed. Therefore, it is important to review the combined distribution of variables that create the donor's cell and, if needed, to collapse them appropriately.

It was also of interest to look into how many times a donor contributed into imputation. Let's take for comparison one cell: State: Alabama, Age Group: 26-30 years old, Sex: Female. This imputation cell had 72 donors and 12 recipients – missing either Education or Marital status. Table 7 and 8 demonstrate how many times and which donors contribute to imputation. The donor's weight is shown as well;

Table 7. SAS Hot deck Macro

Donor ID	Times contributed	Donor's weight
12009031968	1	583.671
12009033023	1	84.996
12009033049	1	127.493
12009034511	1	230.958
12009034544	1	346.437
12009035339	1	287.511
12009038197	1	96.865
12009038236	1	193.729
12009038331	1	145.151
12009041403	1	136.635
12009044811	1	102.161
12009053256	1	48.960

Table 8. PROC HOTDECK

Donor ID	Times contributed	Donor's weight
12009031968	1	583.671
12009031999	2	778.228
12009033023	2	84.996
12009034353	1	115.479
12009034544	1	346.437
12009039124	1	68.412
12009039187	1	273.646
12009043956	1	32.831
12009050034	1	173.219
12009057966	1	136.635

Of course, assignment of the donor is a complicated process involving the weights of a current recipient and the donors, as well as how the assignment went before the current iteration, plus a random factor. However, it is intuitively clear that 72 donors are quite enough to cover 12 recipients using each donor just once. In summary, out of total 587 imputation cells, PROC HOTDECK yields 178 cells where donors contributed more than once, versus

38 cells yielded by the SAS Hotdeck macro. In this sense the SAS Hotdeck macro seems to do a better job not violating the main principle: keeping the distribution close to item respondents.

CONCLUSIONS

The SAS Hotdeck macro and SUDAAN's SAS-callable PROC HOTDECK demonstrate very similar imputation results in terms of distribution of post-imputed variables and the standard error of estimates (percentages). The SAS Macro syntax, in our opinion, is somewhat simpler to execute and the built in warning on empty donor's cells appears to be a very helpful feature. The SAS Hotdeck macro implements a single imputation. PROC HOTDECK, on the other hand, includes an option that allows the user to specify one imputation or multiple imputations. If a user does not have SUDAAN the SAS macro presents a valuable WSHD imputation tool which we would highly recommend.

.

REFERENCES

1. Brenda G. Cox, Research Triangle Institute THE WEIGHTED SEQUENTIAL HOT DECK IMPUTATION PROCEDURE www.amstat.org/sections/srms/proceedings/papers/1980_152.pdf

2. Rebecca R. Andridge, Roderick J. A. Little A Review of Hot Deck Imputation for Survey Non-response International Statistical Review, Volume 78, Issue 1 Pages 1–159

http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2010.00103.x/full

3. Vincent G. Iannacchione, "Weighted Sequential Hot Deck Imputation Macros", Seventh Annual SAS User's Group International Conference, San Francisco CA, February 1982

ACKNOWLEDGEMENTS

CONTACT INFORMATION

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

Author Name: David Izrael
Company: Abt Associates Inc.
Work Phone: 617.349.2434
Email: david_izrael@abtassoc.com