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 Hot Deck 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 1980s 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>
<thead>
<tr>
<th>Table 1. Weighted distribution for item respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of education by marital_status</td>
</tr>
<tr>
<td>education</td>
</tr>
<tr>
<td>Frequency</td>
</tr>
<tr>
<td>Percent</td>
</tr>
<tr>
<td>Row Pct</td>
</tr>
<tr>
<td>Col Pct</td>
</tr>
<tr>
<td>Married</td>
</tr>
<tr>
<td>136013</td>
</tr>
<tr>
<td>4.73</td>
</tr>
<tr>
<td>51.75</td>
</tr>
<tr>
<td>7.18</td>
</tr>
<tr>
<td>Never married, member unmarried couple</td>
</tr>
<tr>
<td>39923.7</td>
</tr>
<tr>
<td>1.39</td>
</tr>
<tr>
<td>15.19</td>
</tr>
<tr>
<td>10.34</td>
</tr>
<tr>
<td>Divorced, Widowed, Separated</td>
</tr>
<tr>
<td>86908.8</td>
</tr>
<tr>
<td>3.02</td>
</tr>
<tr>
<td>33.06</td>
</tr>
<tr>
<td>14.60</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>262845</td>
</tr>
<tr>
<td>9.14</td>
</tr>
<tr>
<td>Less than HS</td>
</tr>
<tr>
<td>HS Grad</td>
</tr>
<tr>
<td>Some College</td>
</tr>
<tr>
<td>College Grad</td>
</tr>
<tr>
<td>Total</td>
</tr>
<tr>
<td>1893307</td>
</tr>
<tr>
<td>386164</td>
</tr>
<tr>
<td>595350</td>
</tr>
<tr>
<td>2874821</td>
</tr>
<tr>
<td>26.86</td>
</tr>
<tr>
<td>6.26</td>
</tr>
<tr>
<td>23.23</td>
</tr>
<tr>
<td>30.24</td>
</tr>
<tr>
<td>1019129</td>
</tr>
<tr>
<td>35.45</td>
</tr>
<tr>
<td>100.00</td>
</tr>
<tr>
<td>2874821</td>
</tr>
<tr>
<td>35.45</td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>i_education</th>
<th>i_marital_status</th>
<th>Frequency</th>
<th>Percent</th>
<th>Row Pct</th>
<th>Col Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Married</td>
<td>159518</td>
<td>4.98</td>
<td>53.79</td>
<td>7.55</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>43262.8</td>
<td>1.35</td>
<td>14.59</td>
<td>10.12</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>93781.9</td>
<td>2.93</td>
<td>31.62</td>
<td>14.12</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>296562</td>
<td>9.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than HS</td>
<td>Married</td>
<td>566071</td>
<td>17.66</td>
<td>62.06</td>
<td>26.78</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>128694</td>
<td>4.02</td>
<td>14.11</td>
<td>30.11</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>217436</td>
<td>6.78</td>
<td>23.84</td>
<td>32.73</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>912201</td>
<td>28.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HS Grad</td>
<td>Married</td>
<td>538829</td>
<td>16.81</td>
<td>62.54</td>
<td>25.50</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>123769</td>
<td>3.86</td>
<td>14.37</td>
<td>28.96</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>198944</td>
<td>6.21</td>
<td>23.09</td>
<td>29.95</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>861542</td>
<td>26.88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Some College</td>
<td>Married</td>
<td>849023</td>
<td>26.49</td>
<td>74.82</td>
<td>40.17</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>131638</td>
<td>4.11</td>
<td>11.60</td>
<td>30.80</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>154135</td>
<td>4.81</td>
<td>13.58</td>
<td>23.20</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1134795</td>
<td>35.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>College Grad</td>
<td>Married</td>
<td>531324</td>
<td>16.58</td>
<td>62.23</td>
<td>25.17</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>127924</td>
<td>3.99</td>
<td>14.98</td>
<td>29.64</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>194601</td>
<td>6.07</td>
<td>22.79</td>
<td>29.37</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>853849</td>
<td>26.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>Married</td>
<td>2113440</td>
<td>65.94</td>
<td>427364</td>
<td>664297</td>
</tr>
<tr>
<td></td>
<td>Never married, member unmarried couple</td>
<td>427364</td>
<td>13.33</td>
<td>664297</td>
<td>20.73</td>
</tr>
<tr>
<td></td>
<td>Divorced, Widowed, Separated</td>
<td>664297</td>
<td>20.73</td>
<td>20.73</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>3205101</td>
<td>100.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>I_EDUCATION(Post-Imputation value of EDUCATION)</th>
<th>I_MARITAL_STATUS(Post-Imputation value of MARITAL_STATUS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>Percent</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>Less than HS</td>
<td>155492</td>
</tr>
<tr>
<td>HS Grad</td>
<td>570408</td>
</tr>
<tr>
<td>Some College</td>
<td>531324</td>
</tr>
<tr>
<td>College Grad</td>
<td>853718</td>
</tr>
<tr>
<td>Total</td>
<td>2110943</td>
</tr>
</tbody>
</table>
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.

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>
<thead>
<tr>
<th>education</th>
<th>Std Error of Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Respondents</td>
</tr>
<tr>
<td>Less than HS</td>
<td>0.2943</td>
</tr>
<tr>
<td>HS Grad</td>
<td>0.4547</td>
</tr>
<tr>
<td>Some College</td>
<td>0.4506</td>
</tr>
<tr>
<td>College Grad</td>
<td>0.4812</td>
</tr>
</tbody>
</table>
### Table 5. Standard Error of Percent for Marital Status

<table>
<thead>
<tr>
<th>marital_status</th>
<th>Respondents</th>
<th>PROC HOTDECK</th>
<th>SAS Macro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married</td>
<td>0.4640</td>
<td>0.4507</td>
<td>0.4540</td>
</tr>
<tr>
<td>Never married, member unmarried couple</td>
<td>0.3553</td>
<td>0.3422</td>
<td>0.3508</td>
</tr>
<tr>
<td>Divorced, Widowed, Separated</td>
<td>0.3676</td>
<td>0.3598</td>
<td>0.3583</td>
</tr>
</tbody>
</table>

### Table 6. Standard Error for Percent of combined distribution

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

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.
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>
<thead>
<tr>
<th>Table 7. SAS Hotdeck Macro</th>
<th>Table 8. PROC HOTDECK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donor ID</td>
<td>Times contributed</td>
</tr>
<tr>
<td>12009031968</td>
<td>1</td>
</tr>
<tr>
<td>1200903023</td>
<td>1</td>
</tr>
<tr>
<td>1200903049</td>
<td>1</td>
</tr>
<tr>
<td>12009034511</td>
<td>1</td>
</tr>
<tr>
<td>12009034544</td>
<td>1</td>
</tr>
<tr>
<td>12009035339</td>
<td>1</td>
</tr>
<tr>
<td>12009038197</td>
<td>1</td>
</tr>
<tr>
<td>12009038236</td>
<td>1</td>
</tr>
<tr>
<td>12009038331</td>
<td>1</td>
</tr>
<tr>
<td>12009041403</td>
<td>1</td>
</tr>
<tr>
<td>12009044811</td>
<td>1</td>
</tr>
<tr>
<td>12009053256</td>
<td>1</td>
</tr>
</tbody>
</table>
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


ACKNOWLEDGEMENTS

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.

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