CROSSTAB Example #5

SUDAAN Statements and Results Illustrated

- Accounting for multiple imputation of variables
- Taylor series linearization method
- BRR method with Fay’s adjustment
- SUBPOPX
- SETENV

Input Data Set(s): NHANES3.SAS7bdat

Example

Among adults aged 20 and older, use the NHANES III Multiply Imputed Dataset to estimate some descriptive statistics on the self-rating of health status and activity level compared to others.

Solution

This example uses data from the 1988-1994 NHANES III. NCHS and CDC have provided a Multiply Imputed Dataset constructed from these studies so that the user can compute estimates that account for the imputation of several key survey measures. The multiply imputed dataset and associated documentation can be obtained from the NCHS website.

The following CROSSTAB example was run in two parts. In the first run, the estimates were computed using the Taylor Series linearization method (Exhibit 1), and in the second run, the estimates were computed using the BRR method with Fay’s adjustment (Exhibit 5). The appropriate BRR replicate weights, adjusted using Fay’s method, can also be found on the multiply imputed dataset.

This example was run in SAS-Callable SUDAAN, and the SAS program and *.LST files are provided for each run.
options pagesize=70 linesize=80;
libname in "c:\903winbetatest\nhanes3";
proc format;
  value health 1="1=Excel" 
    2="2=Very Good" 
    3="3=Good" 
    4="4=Fair" 
    5="5=Poor";
  value activ 1="1=More Active" 
    2="2=Less Active" 
    3="3=Same";
data mi1; set in.nh3mi1;
  proc sort data=mi1; by sdpstra6 sdppsu6;
data mi2; set in.nh3mi2;
  proc sort data=mi2; by sdpstra6 sdppsu6;
data mi3; set in.nh3mi3;
  proc sort data=mi3; by sdpstra6 sdppsu6;
data mi4; set in.nh3mi4;
  proc sort data=mi4; by sdpstra6 sdppsu6;
data mi5; set in.nh3mi5;
  proc sort data=mi5; by sdpstra6 sdppsu6;

PROC CROSSTAB DATA=mi1 filetype=sas MI_COUNT=5 DESIGN=WR;
NEST SDPSTRA6 SDPPSU6 / MISSUNIT;
WEIGHT WTPFQX6;
SUBPOP X HSAGEIR >= 20;
CLASS   HAB1MI HAT28MI;
TABLES  HAB1MI*HAT28MI;
SETENV  ROWWIDTH=8 LB LWIDTH=9 COLWIDTH=8 DECWIDTH=2;
PRINT  NSUM="SampSize" COLPER="COL%" SECOL="SE COL%" ROWPER="ROW%" SEROW="SE ROW%"
/ NSUMFMT=F7.0;
rformat hab1mi health.;
rformat hat28mi activ.;
RTITLE "SELF RATING OF HEALTH STATUS vs. ACTIVITY"
"VARIANCES CALCULATED USING TAYLOR LINEARIZATION (WR)"
RFOOTNOTE "NHANES-III MULTIPLY IMPUTED DATA, ADULTS (20+)"

In the example above (Exhibit 1), the SAS datasets NH3MI1—NH3MI5 are derived from the IMP1.DAT,…,IMP5.DAT files supplied with the NHANES III public use documentation for the multiply imputed dataset. This example uses the shortcut MI_COUNT=5 to indicate the five files that are used by SUDAAN. The output from this example is illustrated below (beginning with Exhibit 2).
Exhibit 2. First Page of SUDAAN Output (SAS *.LST File)

SUDAAN
Software for the Statistical Analysis of Correlated Data
Copyright Research Triangle Institute December 2011
Release 11.0

DESIGN SUMMARY: Variances will be computed using the Taylor Linearization Method, Assuming a With Replacement (WR) Design
Sample Weight: WTPFQX6
Stratification Variables(s): SDPSTRA6
Primary Sampling Unit: SDPPSU6

Processing data for set 1 of imputed variables:
Processing data for set 2 of imputed variables:
Processing data for set 3 of imputed variables:
Processing data for set 4 of imputed variables:
Processing data for set 5 of imputed variables:

Processing data for set 1 of imputed variables:
Number of observations read : 33994 Weighted count : 251097002
Observations in subpopulation : 18825 Weighted count : 177180670
Denominator degrees of freedom : 49

Processing data for set 2 of imputed variables:
Number of observations read : 33994 Weighted count : 251097002
Observations in subpopulation : 18825 Weighted count : 177180670
Denominator degrees of freedom : 49

Processing data for set 3 of imputed variables:
Number of observations read : 33994 Weighted count : 251097002
Observations in subpopulation : 18825 Weighted count : 177180670
Denominator degrees of freedom : 49

Processing data for set 4 of imputed variables:
Number of observations read : 33994 Weighted count : 251097002
Observations in subpopulation : 18825 Weighted count : 177180670
Denominator degrees of freedom : 49

Processing data for set 5 of imputed variables:
Number of observations read : 33994 Weighted count : 251097002
Observations in subpopulation : 18825 Weighted count : 177180670
Denominator degrees of freedom : 49

There are 18,825 adults ages 20 and older in each of the 5 multiply imputed datasets (Exhibit 2).
### Exhibit 3. CLASS Variable Frequencies

Frequencies and Values for CLASS Variables
Results for Summary Over All Imputations

by: Self-rating of health status.
--------------------------------------------

<table>
<thead>
<tr>
<th>Self-rating of health status</th>
<th>Frequency</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered Position: 1</td>
<td>2823.600</td>
<td>1=Excel</td>
</tr>
<tr>
<td>Ordered Position: 2</td>
<td>4388.200</td>
<td>2=Very Good</td>
</tr>
<tr>
<td>Ordered Position: 3</td>
<td>6741.000</td>
<td>3=Good</td>
</tr>
<tr>
<td>Ordered Position: 4</td>
<td>3834.800</td>
<td>4=Fair</td>
</tr>
<tr>
<td>Ordered Position: 5</td>
<td>1037.400</td>
<td>5=Poor</td>
</tr>
</tbody>
</table>

### Exhibit 3. CLASS Variable Frequencies-cont.

Frequencies and Values for CLASS Variables
Results for Summary Over All Imputations

by: Compare own activity level to others.
----------------------------------------------

<table>
<thead>
<tr>
<th>Compare own activity level to others</th>
<th>Frequency</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordered Position: 1</td>
<td>5938.200</td>
<td>1=More Active</td>
</tr>
<tr>
<td>Ordered Position: 2</td>
<td>4275.000</td>
<td>2=Less Active</td>
</tr>
<tr>
<td>Ordered Position: 3</td>
<td>8611.800</td>
<td>3=Same</td>
</tr>
</tbody>
</table>

In this example, the variable HAB1MI holds the multiply imputed response for “Would you say your health in general is excellent, very good, good, fair or poor?” and the variable HAT28MI holds the multiply imputed response for “Compared with most men/women your age, would you say that you are more active, less active or about the same?” These categorical variables were defined in CROSSTAB using the CLASS statement. The above “Frequency” output (Exhibit 3) represents the average frequency of these multiply imputed variables on the five NH3MI1—NH3MI5 datasets.
The table displayed in **Exhibit 4** is the summary over all imputations. This table shows, for example, that 47.28% of those adults who rated their health as “excellent” also believe that they are more active than other men/women their age. In comparison, only 12.97% of those adults who rated their health as “poor” also believe they are more active than other men/women their age. The standard errors of these statistics are 1.68 and 1.52, respectively.
The following replicates the example above, but uses the BRR (with Fay Adjustment) method for computing the variances *(Exhibit 5)*.

**Exhibit 5. SAS-Callable SUDAAN Code (DESIGN=BRR)**

```sas
options pagesize=70 linesize=80;
libname in "c:\903winbetatest\nhanes3";
proc format;
  value health 1="1=Excel"
               2="2=Very Good"
               3="3=Good"
               4="4=Fair"
               5="5=Poor";
  value activ 1="1=More Active"
               2="2=Less Active"
               3="3=Same";

data mi1; set in.nh3mi1;
  proc sort data=mi1; by sdpstra6 sdppsu6;

data mi2; set in.nh3mi2;
  proc sort data=mi2; by sdpstra6 sdppsu6;

data mi3; set in.nh3mi3;
  proc sort data=mi3; by sdpstra6 sdppsu6;

data mi4; set in.nh3mi4;
  proc sort data=mi4; by sdpstra6 sdppsu6;

data mi5; set in.nh3mi5;
  proc sort data=mi5; by sdpstra6 sdppsu6;

PROC CROSSTAB DATA=mi1 filetype=sas MI_COUNT=5 DESIGN=BRR;
  WEIGHT WTPFQX6;
  REPWT WTQPQP1-WTPQRP52 / ADJFAY=2.0408;
  SUBPOPX HSAGEIR >= 20;
  CLASS HAB1MI HAT28MI;
  TABLES HAB1MI*HAT28MI;
  SETENV ROWWIDTH=8 LBLWIDTH=9 COLWIDTH=8 DECWIDTH=2;
  PRINT NSUM="SampSize" COLPER="COL%" SECOL="SE COL%" ROWPER="ROW%" SEROW="SE ROW%"
       / NSUMFMT=F7.0;
  rformat hab1mi health.;
  rformat hat28mi activ.;
  RTITLE "SELF RATING OF HEALTH STATUS vs. ACTIVITY"
       "VARIANCES CALCULATED VIA REPLICATION (BRR) WITH FAY ADJUSTMENT";
  RFOOTNOTE "NHANES-III MULTIPLY IMPUTED DATA, ADULTS (20+);";
```

Page 6 of 9
Exhibit 6.  First Page of SUDAAN Output (SAS *.LST File)

SUDAAN  
Software for the Statistical Analysis of Correlated Data  
Copyright Research Triangle Institute December 2011  
Release 11.0

DESIGN SUMMARY: Variances will be computed using the Balanced Repeated Replication (BRR) Method  
Sample Weight: WTPFQX6  
Replicate Sample Weights:  
WTPQRP1  WTPQRP2  WTPQRP3  WTPQRP4  WTPQRP5  WTPQRP6  WTPQRP7  
WTPQRP8  WTPQRP9  WTPQRP10  WTPQRP11  WTPQRP12  WTPQRP13  WTPQRP14  
WTPQRP15  WTPQRP16  WTPQRP17  WTPQRP18  WTPQRP19  WTPQRP20  WTPQRP21  
WTPQRP22  WTPQRP23  WTPQRP24  WTPQRP25  WTPQRP26  WTPQRP27  WTPQRP28  
WTPQRP29  WTPQRP30  WTPQRP31  WTPQRP32  WTPQRP33  WTPQRP34  WTPQRP35  
WTPQRP36  WTPQRP37  WTPQRP38  WTPQRP39  WTPQRP40  WTPQRP41  WTPQRP42  
WTPQRP43  WTPQRP44  WTPQRP45  WTPQRP46  WTPQRP47  WTPQRP48  WTPQRP49  
WTPQRP50  WTPQRP51  WTPQRP52  
Multiplier Associated with Replicate Weights: 2.04

Processing data for set 1 of imputed variables:
Processing data for set 2 of imputed variables:
Processing data for set 3 of imputed variables:
Processing data for set 4 of imputed variables:
Processing data for set 5 of imputed variables:

Processing data for set 1 of imputed variables:
Number of observations read : 33994  Weighted count :251097002  
Observations in subpopulation : 18825  Weighted count:177180670  
Denominator degrees of freedom : 52

Processing data for set 2 of imputed variables:
Number of observations read : 33994  Weighted count :251097002  
Observations in subpopulation : 18825  Weighted count:177180670  
Denominator degrees of freedom : 52

Processing data for set 3 of imputed variables:
Number of observations read : 33994  Weighted count :251097002  
Observations in subpopulation : 18825  Weighted count:177180670  
Denominator degrees of freedom : 52

Processing data for set 4 of imputed variables:
Number of observations read : 33994  Weighted count :251097002  
Observations in subpopulation : 18825  Weighted count:177180670  
Denominator degrees of freedom : 52

Processing data for set 5 of imputed variables:
Number of observations read : 33994  Weighted count :251097002  
Observations in subpopulation : 18825  Weighted count:177180670  
Denominator degrees of freedom : 52
### Exhibit 7. Class Variable Frequencies

**Frequencies and Values for CLASS Variables**  
Results for Summary Over All Imputations  
by: Self-rating of health status.

<table>
<thead>
<tr>
<th>Ordered Position:</th>
<th>Frequency</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2823.600</td>
<td>1=Excel</td>
</tr>
<tr>
<td>2</td>
<td>4388.200</td>
<td>2=Very Good</td>
</tr>
<tr>
<td>3</td>
<td>6741.000</td>
<td>3=Good</td>
</tr>
<tr>
<td>4</td>
<td>3834.800</td>
<td>4=Fair</td>
</tr>
<tr>
<td>5</td>
<td>1037.400</td>
<td>5=Poor</td>
</tr>
</tbody>
</table>

### Exhibit 7. Class Variable Frequencies-cont.

**Frequencies and Values for CLASS Variables**  
Results for Summary Over All Imputations  
by: Compare own activity level to others.

<table>
<thead>
<tr>
<th>Ordered Position:</th>
<th>Frequency</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5938.200</td>
<td>1=More Active</td>
</tr>
<tr>
<td>2</td>
<td>4275.000</td>
<td>2=Less Active</td>
</tr>
<tr>
<td>3</td>
<td>8611.800</td>
<td>3=Same</td>
</tr>
</tbody>
</table>
**Exhibit 8. HAB1MI^HAT28MI Crosstabulation (DESIGN=BRR)**

<table>
<thead>
<tr>
<th>Self-rating of health status vs. activity</th>
<th>Total</th>
<th>1=More</th>
<th>2=Less</th>
<th>3=Same</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE COL%</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>33.23</td>
<td>22.13</td>
<td>44.64</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>0.67</td>
<td>0.61</td>
<td>0.63</td>
</tr>
<tr>
<td>SampSize</td>
<td>18825</td>
<td>5938</td>
<td>4275</td>
<td>8612</td>
</tr>
<tr>
<td>COL%</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>47.28</td>
<td>13.21</td>
<td>39.50</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>1.38</td>
<td>1.12</td>
<td>1.38</td>
</tr>
<tr>
<td>SampSize</td>
<td>2824</td>
<td>1369</td>
<td>368</td>
<td>1087</td>
</tr>
<tr>
<td>COL%</td>
<td>20.77</td>
<td>29.55</td>
<td>12.40</td>
<td>18.38</td>
</tr>
<tr>
<td>SE COL%</td>
<td>0.61</td>
<td>1.13</td>
<td>0.97</td>
<td>0.67</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>47.28</td>
<td>13.21</td>
<td>39.50</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>1.38</td>
<td>1.12</td>
<td>1.38</td>
</tr>
<tr>
<td>SampSize</td>
<td>4388</td>
<td>1668</td>
<td>759</td>
<td>1961</td>
</tr>
<tr>
<td>COL%</td>
<td>30.53</td>
<td>34.64</td>
<td>25.04</td>
<td>30.19</td>
</tr>
<tr>
<td>SE COL%</td>
<td>0.61</td>
<td>1.15</td>
<td>1.16</td>
<td>0.90</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>37.70</td>
<td>18.16</td>
<td>44.14</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>1.54</td>
<td>0.89</td>
<td>1.36</td>
</tr>
<tr>
<td>SampSize</td>
<td>6741</td>
<td>1969</td>
<td>1441</td>
<td>3331</td>
</tr>
<tr>
<td>COL%</td>
<td>32.51</td>
<td>26.40</td>
<td>35.07</td>
<td>35.79</td>
</tr>
<tr>
<td>SE COL%</td>
<td>0.54</td>
<td>1.04</td>
<td>1.01</td>
<td>0.66</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>26.98</td>
<td>23.88</td>
<td>49.14</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>0.78</td>
<td>0.85</td>
<td>0.84</td>
</tr>
<tr>
<td>SampSize</td>
<td>3835</td>
<td>797</td>
<td>1143</td>
<td>1895</td>
</tr>
<tr>
<td>COL%</td>
<td>12.85</td>
<td>8.10</td>
<td>19.01</td>
<td>13.34</td>
</tr>
<tr>
<td>SE COL%</td>
<td>0.51</td>
<td>0.54</td>
<td>0.91</td>
<td>0.68</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>20.95</td>
<td>32.72</td>
<td>46.33</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>0.92</td>
<td>1.22</td>
<td>1.32</td>
</tr>
<tr>
<td>SampSize</td>
<td>1037</td>
<td>136</td>
<td>563</td>
<td>338</td>
</tr>
<tr>
<td>COL%</td>
<td>3.34</td>
<td>1.30</td>
<td>8.48</td>
<td>2.50</td>
</tr>
<tr>
<td>SE COL%</td>
<td>0.18</td>
<td>0.13</td>
<td>0.61</td>
<td>0.23</td>
</tr>
<tr>
<td>ROW%</td>
<td>100.00</td>
<td>12.97</td>
<td>56.24</td>
<td>30.80</td>
</tr>
<tr>
<td>SE ROW%</td>
<td>0.00</td>
<td>1.21</td>
<td>2.54</td>
<td>2.40</td>
</tr>
</tbody>
</table>

The above table (Exhibit 8) shows that the variance estimates computed using the BRR method are generally smaller than the variance estimates computed using the Taylor Series linearization method. This phenomenon is not true in general, and may be an indication that, for this particular example, the weight adjustments in the NHANES III data may actually be improving the precision of estimates.