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.

Exhibit 1. SAS-Callable SUDAAN Code (DESIGN=WR)

```
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;

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 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)

```
                              S U D A A N
           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.
-----------------------------------------
Self-rating
  of health
  status            Frequency          Value
-----------------------------------------
Ordered
  Position:
  1                  2823.600        1=Excel
Ordered
  Position:
  2                  4388.200    2=Very Good
Ordered
  Position:
  3                  6741.000         3=Good
Ordered
  Position:
  4                  3834.800         4=Fair
Ordered
  Position:
  5                  1037.400         5=Poor
-----------------------------------------
```

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.
-----------------------------------------------
Compare own
  activity
  level to
  others            Frequency            Value
-----------------------------------------------
Ordered
  Position:
  1                  5938.200    1=More Active
Ordered
  Position:
  2                  4275.000    2=Less Active
Ordered
  Position:
  3                  8611.800           3=Same
-----------------------------------------------
```

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.

Exhibit 4. HAB1MI*HAT28MI Crosstabulation (DESIGN=WR)

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
For Subpopulation: HSAGEIR >= 20

SELF RATING OF HEALTH STATUS vs. ACTIVITY
VARIANCES CALCULATED USING TAYLOR LINEARIZATION (WR)

Results for Summary Over All Imputations
by: Self-rating of health status, Compare own activity level to others.

Self-rating of health status		Compare own activity level to others			
		Total	1=More Active	2=Less Active	3=Same
Total	SampSize	18825	5938	4275	8612
	COL%	100.00	100.00	100.00	100.00
	SE COL%	0.00	0.00	0.00	0.00
	ROW%	100.00	33.23	22.13	44.64
	SE ROW%	0.00	0.74	0.67	0.69
1=Excel	SampSize	2824	1369	368	1087
	COL%	20.77	29.55	12.40	18.38
	SE COL%	0.70	1.25	1.02	0.87
	ROW%	100.00	47.28	13.21	39.50
	SE ROW%	0.00	1.68	1.17	1.65
2=Very Good	SampSize	4388	1668	759	1961
	COL%	30.53	34.64	25.04	30.19
	SE COL%	0.70	1.26	1.27	0.91
	ROW%	100.00	37.70	18.16	44.14
	SE ROW%	0.00	1.51	0.97	1.31
3=Good	SampSize	6741	1969	1441	3331
	COL%	32.51	26.40	35.07	35.79
	SE COL%	0.70	1.19	1.18	0.74
	ROW%	100.00	26.98	23.88	49.14
	SE ROW%	0.00	0.86	1.01	0.90
4=Fair	SampSize	3835	797	1143	1895
	COL%	12.85	8.10	19.01	13.34
	SE COL%	0.58	0.55	1.08	0.72
	ROW%	100.00	20.95	32.72	46.33
	SE ROW%	0.00	0.93	1.43	1.51
5=Poor	SampSize	1037	136	563	338
	COL%	3.34	1.30	8.48	2.30
	SE COL%	0.17	0.16	0.68	0.25
	ROW%	100.00	12.97	56.24	30.80
	SE ROW%	0.00	1.52	2.96	2.86

NHANES-III MULTIPLY IMPUTED DATA, ADULTS (20+)

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)

```
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;
REPWGT WTPQRP1-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+)";
```

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

```
                              S U D A A N
           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.
-----------------------------------------------
Self-rating
  of health
  status            Frequency          Value
-----------------------------------------------
Ordered
  Position:
  1                  2823.600        1=Excel
Ordered
  Position:
  2                  4388.200    2=Very Good
Ordered
  Position:
  3                  6741.000         3=Good
Ordered
  Position:
  4                  3834.800         4=Fair
Ordered
  Position:
  5                  1037.400         5=Poor
-----------------------------------------------
```

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.
-------------------------------------------------
Compare own
  activity
  level to
  others            Frequency            Value
-------------------------------------------------
Ordered
  Position:
  1                  5938.200    1=More Active
Ordered
  Position:
  2                  4275.000    2=Less Active
Ordered
  Position:
  3                  8611.800           3=Same
-------------------------------------------------
```

Exhibit 8. HAB1MI*HAT28MI Crosstabulation (DESIGN=BRR)

Variance Estimation Method: BRR Using Multiply Imputed Data
For Subpopulation: HSAGEIR >= 20

SELF RATING OF HEALTH STATUS vs. ACTIVITY
VARIANCES CALCULATED VIA REPLICATION (BRR) WITH FAY ADJUSTMENT

Results for Summary Over All Imputations
by: Self-rating of health status, Compare own activity level to others.

Self-rating of health status		Compare own activity level to others			
		Total	1=More Active	2=Less Active	3=Same
Total	SampSize	18825	5938	4275	8612
	COL%	100.00	100.00	100.00	100.00
	SE COL%	0.00	0.00	0.00	0.00
	ROW%	100.00	33.23	22.13	44.64
	SE ROW%	0.00	0.67	0.61	0.63
1=Excel	SampSize	2824	1369	368	1087
	COL%	20.77	29.55	12.40	18.38
	SE COL%	0.61	1.13	0.97	0.67
	ROW%	100.00	47.28	13.21	39.50
	SE ROW%	0.00	1.38	1.12	1.38
2=Very Good	SampSize	4388	1668	759	1961
	COL%	30.53	34.64	25.04	30.19
	SE COL%	0.61	1.15	1.16	0.90
	ROW%	100.00	37.70	18.16	44.14
	SE ROW%	0.00	1.54	0.89	1.36
3=Good	SampSize	6741	1969	1441	3331
	COL%	32.51	26.40	35.07	35.79
	SE COL%	0.54	1.04	1.01	0.66
	ROW%	100.00	26.98	23.88	49.14
	SE ROW%	0.00	0.78	0.85	0.84
4=Fair	SampSize	3835	797	1143	1895
	COL%	12.85	8.10	19.01	13.34
	SE COL%	0.51	0.54	0.91	0.68
	ROW%	100.00	20.95	32.72	46.33
	SE ROW%	0.00	0.92	1.22	1.32
5=Poor	SampSize	1037	136	563	338
	COL%	3.34	1.30	8.48	2.30
	SE COL%	0.18	0.13	0.61	0.23
	ROW%	100.00	12.97	56.24	30.80
	SE ROW%	0.00	1.21	2.54	2.40

NHANES-III MULTIPLY IMPUTED DATA, ADULTS (20+)

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.