“How can I incorporate multiply imputed data sets into my analysis using SUDAAN?”

Authored by

stats writer

July 1, 2024

Using SUDAAN, a statistical software package designed for analyzing complex survey data, one can easily incorporate multiply imputed data sets into their analysis. This process involves combining the results from multiple imputations to obtain a single, valid estimate of the desired parameter. SUDAAN offers various options for handling imputed data sets, such as using the "MIANALYZE" procedure to combine estimators from each imputed data set. This allows for the proper adjustment of standard errors and confidence intervals to account for the uncertainty introduced by imputation. By incorporating multiply imputed data sets into their analysis, researchers can potentially improve the accuracy and reliability of their findings, particularly when dealing with missing data in large survey studies.

How can I use multiply imputed data sets in SUDAAN? | SUDAAN FAQ

**One of the new features of SUDAAN 9 is its ability to use multiply imputed
data sets. According the SUDAAN 9 Language Manual (page 91), SUDAAN does not accept the data sets stacked into a single data set as the SAS
proc
mi creates. Rather, SUDAAN accepts multiply imputed data sets in two
forms: either the individually imputed data sets or a single data set with
variables in the same file with different numeric suffixes. For our
examples below, we will use the NHANES III data. The NHANES III multiply imputed data sets can be found at http://www.cdc.gov/nchs/nhanes.htm about half way**

down
the page. You need to have each of the imputed data sets sorted by strata
and PSU just as you have to have a non-imputed data set sorted. You can
use the macro shown below (with minor changes for the names of the data sets) to
sort the data sets, or you can have multiple calls to proc sort. Once the
data sets are sorted, you need to either add one more option, mi_count,
to the proc statement, or a mi_file statement. If the data sets are sequentially numbers, such as nh3mi1, nh3mi2, etc., you can use the
mi_count option and indicate the number of imputed data sets. On the
data option, specify the first of the imputed data sets. Remember
that the variables in each of the imputed data sets need to be in the same
order, of the same type, etc. SUDAAN will issue a warning if the number of
cases differs between the imputed data sets.

NOTE: The examples of the use of these options and statements in the
SUDAAN 9 Manual (page 91) show the use of quotes around a file path
specification. This will work only for stand-alone SUDAAN, not the
SAS-callable version.

```
%MACRO srt(NUMBER);
PROC SORT DATA=nh3mi&NUMBER;
by sdpstra6 sdppsu6;
run;
%mend srt;

%srt(1);
%srt(2);
%srt(3);
%srt(4);
%srt(5);

proc descript data = NH3MI1 filetype = sas mi_count = 5
design = wr;
nest sdpstra6 sdppsu6 / missunit;
weight WTPFQX6 ;
```

```
var TCPMI;
setenv colwidth = 19;
setenv decwidth = 3;
print nsum wsum mean semean / nohead;
run;
```

```
Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
Results for Summary Over All Imputations
by: Variable, One.

----------------------------------------------------------
| | |
| Variable | | One
| | | 1 |
----------------------------------------------------------
| | | |
| Serum | Sample Size | 28012.000 |
| cholesterol | Weighted Size | 235771269.750 |
| (mg/dL) | Mean | 194.357 |
| | SE Mean | 0.577 |
----------------------------------------------------------
```

In the example below, the mi_files statement is used

instead of the
mi_count option. As before, the first of the imputed data sets is
listed on the data option on the proc statement. The rest of the
files are listed on the mi_files statement.

```
proc regress data = nh3mi1 filetype = sas design = wr;
nest sdpstra6 sdppsu6 / missunit;
weight WTPFQX6 ;
class HAN6SRMI ;
model BMPWSTMI = HAM5MI HAN6SRMI HSSEX;
mi_files nh3mi2 nh3mi3 nh3mi4 nh3mi5 ;
run;
```

Frequencies and Values for CLASS Variables
Results for Summary Over All Imputations
by: Beer/wine/liquor (recode).

```
-------------------------------------
Beer/wine/l-
iquor
(recode) Frequency Value
-------------------------------------
```

Ordered

Position:

1 11230.000 1

Ordered

Position:

2 5546.600 2

Ordered

Position:

3 3273.400 3

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data

SE Method: Robust (Binder, 1983)

Working Correlations: Independent

Link Function: Identity

Response variable BMPWSTMI: Waist circumference (cm)

Results for Summary Over All Imputations

by: Independent Variables and Effects.

Independent

Variables and Beta Lower 95% Upper 95%
Effects Coeff. SE Beta Limit Beta Limit Beta T-Test B=0

Intercept 44.05 4.54 34.90 53.20 9.71
How tall are you
without shoes-
inchs 0.74 0.06 0.62 0.86 12.41
Beer/wine/liquor
(recode)
1 4.23 0.38 3.46 4.99 11.13
2 0.91 0.41 0.07 1.75 2.20
3 0.00 0.00 0.00 0.00 .
Sex -2.84 0.49 -3.83 -1.84 -5.75

Independent P-value
Variables and T-Test DDF
Effects B=0 Beta

Intercept 0.0000 43.516
How tall are you

without shoes-
inchs 0.0000 43.551
Beer/wine/liquor
(recode)
1 0.0000 45.861
2 0.0342 37.224
3 . 49.000
Sex 0.0000 43.755

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable BMPWSTMI: Waist circumference (cm)
Results for Summary Over All Imputations
by: Contrast.

Contrast Degrees
of P-value

Freedom Wald F Wald F

OVERALL MODEL 5 40326.90 0.0000

MODEL MINUS

INTERCEPT 4 182.83 0.0000

INTERCEPT . . .

HAM5MI 1 153.97 0.0000

HAN6SRMI 2 62.84 0.0000

HSSEX 1 33.04 0.0000

In the two examples below, we show that you can use either method of
correcting the standard errors, strata/PSUs or replicate weights.

```
proc crosstabs data = NH3MI1 filetype = sas mi_count =
5 design = wr;
nest sdpstra6 sdppsu6 / missunit;
weight WTPFQX6 ;
subgroups DMARETHN HAE7;
levels 2 2;
tables DMARETHN*HAE7;
setenv colwidth = 12;
```

run;

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
Results for Summary Over All Imputations
by: Race-ethnicity, Ever told had high cholesterol.

| | |
| Race-ethnicity | | Ever told had high cholesterol
| | | Total | 1 | 2 |

| | | | | |
| Total | Sample Size | 7830 | 2548 | 5282 |
| | Weighted Size | 94129643.85 | 31162209.08 | 62967434.77 |
| | Tot Percent | 100.00 | 33.11 | 66.89 |
| | Col Percent | 100.00 | 100.00 | 100.00 |
| | SE Col Percent | 0.00 | 0.00 | 0.00 |
| | Row Percent | 100.00 | 33.11 | 66.89 |
| | SE Row Percent | 0.00 | 0.82 | 0.82 |

| | | | | |

| 1 | Sample Size | 5378 | 1856 | 3522 |

| | Weighted Size | 84795202.29 | 28668926.11 | 56126276.18 |

| | Tot Percent | 90.08 | 30.46 | 59.63 |

| | Col Percent | 90.08 | 92.00 | 89.14 |

| | SE Col Percent | 0.69 | 0.62 | 0.79 |

| | Row Percent | 100.00 | 33.81 | 66.19 |

| | SE Row Percent | 0.00 | 0.90 | 0.90 |

| | | | | |

| 2 | Sample Size | 2452 | 692 | 1760 |

| | Weighted Size | 9334441.56 | 2493282.97 | 6841158.59 |

| | Tot Percent | 9.92 | 2.65 | 7.27 |

| | Col Percent | 9.92 | 8.00 | 10.86 |

| | SE Col Percent | 0.69 | 0.62 | 0.79 |

| | Row Percent | 100.00 | 26.71 | 73.29 |

| | SE Row Percent | 0.00 | 1.00 | 1.00 |

---

proc crosstabs data = NH3MI1 filetype = sas mi_count = 5 design = brr;
repwgt WTPQRP1 - WTPQRP52 / adjfay = 1.7;
weight WTPFQX6 ;
subgroups DMARETHN HAE7;
levels 2 2;
tables DMARETHN*HAE7;
setenv colwidth = 12;
print nsum wsum totper colper secol rowper serow;
run;

Variance Estimation Method: BRR Using Multiply Imputed Data
Results for Summary Over All Imputations
by: Race-ethnicity, Ever told had high cholesterol.

| | |
| Race-ethnicity | | Ever told had high cholesterol
| | | Total | 1 | 2 |

| | | | | |

Total	Sample Size	7830	2548	5282
	Weighted Size	94129643.85	31162209.08	62967434.77
	Tot Percent	100.00	33.11	66.89
	Col Percent	100.00	100.00	100.00
	SE Col Percent	0.00	0.00	0.00
	Row Percent	100.00	33.11	66.89
	SE Row Percent	0.00	0.64	0.64

1	Sample Size	5378	1856	3522
	Weighted Size	84795202.29	28668926.11	56126276.18
	Tot Percent	90.08	30.46	59.63
	Col Percent	90.08	92.00	89.14
	SE Col Percent	0.23	0.35	0.29
	Row Percent	100.00	33.81	66.19
	SE Row Percent	0.00	0.70	0.70

| | | | | |
| 2 | Sample Size | 2452 | 692 | 1760 |
| | Weighted Size | 9334441.56 | 2493282.97 | 6841158.59

|

| | Tot Percent | 9.92 | 2.65 | 7.27 |

| | Col Percent | 9.92 | 8.00 | 10.86 |

| | SE Col Percent | 0.23 | 0.35 | 0.29 |

| | Row Percent | 100.00 | 26.71 | 73.29 |

| | SE Row Percent | 0.00 | 0.92 | 0.92 |

To illustrate the use of the multiple imputed variables in a single data
file, we will create a small example data set and then use the mi_vars
statement.

```
data temp;
input x x1 x2 x3 y;
cards;
1 1 1 1 7
3 3 3 3 8
. 2 1 3 5
. 1 5 4 8
4 4 4 4 9
6 6 6 6 7
```

```
. 7 5 4 9
;
run;

proc regress data = temp filetype = sas design = wr;
weight _one_;
nest _one_;
model y = x;
mi_vars x1 x2 x3;
run;
```