# WTADJX Example #1

## *SUDAAN Statements and Results Illustrated*

- Raking, raking to a size-variable
- Nearly pseudo-optimal calibration approach
- ADJUST = POST;  POSTWGT
- CALVARS
- CLASS; VAR

## *Input Data Set(s):  DAWN.SAS7bdat*

## *Example*

Using manufactured sample and frame data inspired by the  Drug Abuse Warning Network (DAWN) survey and its public-use data set http://www.samhsa.gov/data/2k11/DAWN/2k9DAWNED/HTML/DAWN2k9ED.htm), we estimate the annual  number of drug-related emergency department visits in the U.S. and by census region in an increasingly efficient manner using raking to a size variable and nearly quasi-optimal calibration.

## *Solution*

WTADJUST and WTADJX can employ raking (also called "raking-ratio adjustment" and "iterative proportional fitting") and nearly pseudo-optimal calibration to reduce the standard errors of estimated totals in the absence of nonresponse. Traditional raking to frame counts is effectively a feature of WTADJUST in SUDAAN 10 (a different computation method is used, but the results are the same).  SUDAAN 11 also possesses the ability to compute standard errors that correctly incorporate the impact of raking.

The procedure WTADJX introduced in SUDAAN 11 also allows raking to the size-variable totals, which in this case are frame emergency-department visits. WTADJX can also be used for nearly pseudo-optimal calibration (Kott 2011). This often reduces standard errors further than raking does.

The two forms of raking can additionally be employed when the population values  whether counts or size variables  used to benchmark the weights come from a more reliable source than the frame, when the latter is subject to undercoverage (missing population units) or overcoverage (duplicate records). *Nearly pseudo-optimal calibration*, by contrast, *should not be used to adjust for frame errors.*

In most of what follows, we will assume that there is no nonresponse or frame errors. A note at the end of this discussion addresses frame errors. We assume the SAS-callable version SUDAAN is being used.

The following variables from the DAWN dataset are of interest in this example:

| Variable | Definition |
|---|---|
| RECORD | |
| STRATUM | |
| BIG_N | Population size in stratum |
| N | Sample size in stratum |
| W | Weight (BIG_N/N) |
| REGION | East = 1; South = 2; Midwest = 3; West = 4 |
| PUBLIC | Yes = 1 (the alternative is a privately owned hospital) |
| METRO | Yes = 1 (i.e., located in urban area) |
| FRAME_VISITS | Number of previous-year emergency-department visits recorded on the frame |
| ER_VISITS | Annual drug-related emergency-department (room) visits collected on the survey |

In *Exhibit 1*, we download the data and create some more variables. We create a new frame-visit overall size variable Z, by dividing FRAME_VISITS by 1000. This is for convenience, although sometimes it helps when running WTADJX to reduce the size of relatively large variables.

We also create two frame-visit size calibration variables, PUBLICZ and METROZ, by multiplying the corresponding original variables (PUBLIC and METRO) by Z. For use in nearly pseudo-optimal calibration, we create W1 = W - 1 and multiples of Z(W - 1): such as PUBLICZW1, METROZW1 and ZW1 itself.

**Exhibit 1. Downloading the Data and Creating Some Variables**

```
LIBNAME IN \\rtints29\sudaan\phil\;
DATA R; SET IN.DAWN;
Z = FRAME_VISITS/1000;
W1 = W - 1;
ZW1 = Z*W1;

PUBLICZ = PUBLIC*Z;
PUBLICZW1 = PUBLIC*Z*W1;
METROZ = METRO*Z;
METROZW1 = METRO*Z*W1;
```

*Exhibit 2* shows us what the estimated total number of drug-related emergency-department visits (VAR ER_VISITS), by region (CLASS REGION) is without any weight adjustments.

The sampling design is stratified simple random sampling without replacement (DESIGN STRWOR;
NEST STRATUM). To incorporate the impact of finite population correction in our standard-error determination we include a TOTCNT BIG_N statement.

Since we are interested in estimating totals and their standard errors, we include the line:

```
OUTPUT TOTAL SETOTAL/FILENAME = OUT0 REPLACE;
```

The second part of this line outputs TOTAL and SETOTAL onto OUT0, which we will print later.

### Exhibit 2. Estimating the Unadjusted Totals by Region

```
PROC DESCRIPT  DATA = R DESIGN = STRWOR; NEST STRATUM; WEIGHT W; TOTCNT BIG_N;
    CLASS REGION;  VAR ER_VISITS ;
    OUTPUT TOTAL SETOTAL/FILENAME = OUT0 REPLACE;
RUN;
```

The strata are almost completely cross-classified by region, public/private, and metro/non-metro – but not quite. There is only one hospital in the sample in the East (REGION = 1), private (PUBLIC = 0), and not urban (METRO = 0), but it is *not* in its own stratum.

*Raking to Frame Counts*

From the frame we can get population totals for the number hospitals by region as well as the number of urban and public hospitals. This allows us to perform traditional raking to these totals using PROC WTADJUST with DESIGN = POST, putting the known totals in a POSTWGT statement

In the MODEL statement within the code in *Exhibit 3*, the "dependent variable" is simply _ONE_ since the entire sample is being weight adjusted.

We include frame totals for PUBLIC, METRO, and every region (in their proper order) in the POSTWGT statement and employ the no-intercept (NOINT) option in the MODEL statement.

Since we are interested in standard errors of drug-related emergency-department totals (by region), the rest of the syntax closely follows DESCRIPT, except SE_TOTAL replaces SETOTAL.

We output the estimated totals and their standard errors onto OUT1 for later comparison.

### Exhibit 3. Estimating the Totals by Region by Raking to Frame Counts

```
PROC WTADJUST DATA = R DESIGN = STRWOR ADJUST = POST; NEST STRATUM; WEIGHT W;
  TOTCNT BIG_N;
  CLASS REGION;   VAR ER_VISITS ;
  MODEL _ONE_ = PUBLIC METRO REGION/NOINT;
  POSTWGT  1642 856    489    1636    3124    1051.00;     /* Computed  from the frame */
   OUTPUT TOTAL SE_TOTAL/FILENAME=OUT1 REPLACE;
RUN;
```

Let us briefly mention two tables from the standard output of WTADJUST. The first is in *Exhibit 4* may be a bit off putting since it reveals that no *beta* is significant. That is as it should be since what is being fit here is a response (or coverage) model, and there is no nonresponse (or coverage errors).

### Exhibit 4. Estimated *Betas* from Raking

| Independent Variables and Effects | Beta Coeff. | SE Beta | Lower 95% Limit Beta | Upper 95% Limit Beta | T-Test B=0 | P-value T-Test B=0 | Respondent Sample Size | Nonrespondent Sample Size |
|---|---|---|---|---|---|---|---|---|
| PUBLIC | -0.01 | 0.04 | -0.10 | 0.07 | -0.28 | 0.7786 | . | . |
| METRO | -0.00 | 0.00 | -0.00 | 0.00 | -0.47 | 0.6390 | . | . |
| REGION | | | | | | | | |
| 1 | 0.01 | 0.04 | -0.06 | 0.08 | 0.25 | 0.8000 | 159 | 0 |
| 2 | 0.01 | 0.03 | -0.05 | 0.07 | 0.28 | 0.7768 | 54 | 0 |
| 3 | 0.00 | 0.00 | -0.00 | 0.00 | 0.29 | 0.7726 | 78 | 0 |
| 4 | 0.00 | 0.00 | -0.00 | 0.00 | 0.46 | 0.6439 | 55 | 0 |

### Exhibit 5. Comparing Estimated Totals to Control Totals

| Independent Variables and Effects | Sum of Original Weights Over Respondents | Sum of Trimmed Weights Over Respondents | Sum of Final Adjusted Weights Over Respondents | Control Totals | Final Weight Sum Minus Controls | Original Unequal Weighting Effect | Trimmed Unequal Weighting Effect |
|---|---|---|---|---|---|---|---|
| PUBLIC | 1647.11 | 1647.11 | 1642.00 | 1642.00 | 0.00 | . | . |
| METRO | 856.00 | 856.00 | 856.00 | 856.00 | 0.00 | . | . |
| REGION | | | | | | | |
| 1 | 489.00 | 489.00 | 489.00 | 489.00 | 0.00 | 2.7157 | 2.7157 |
| 2 | 1636.00 | 1636.00 | 1636.00 | 1636.00 | -0.00 | 1.1379 | 1.1379 |
| 3 | 3124.00 | 3124.00 | 3124.00 | 3124.00 | -0.00 | 1.0423 | 1.0423 |
| 4 | 1051.00 | 1051.00 | 1051.00 | 1051.00 | -0.00 | 1.4846 | 1.4846 |

More important is *Exhibit 5*. It tells us that our calibration targets (Control Totals) have been reached. This is crucially important when when running WTADJUST and should always be checked. If the values of the "Final Weight Sum Minus Controls" column are not zero or close to it, then the calibration has failed. The same applies for WTADJX when the number of calibration variables is not greater than the number of model variables.

When calibration does fail, this table can help us understand why by revealing which calibration targets could not be reached.

Also new in SUDAAN 11 is the standard error table in *Exhibit 6* produced because we asked for it with the VAR ER_VISITS and CLASS REGION statements.

### Exhibit 6. Estimated Mean, Totals, and Standard Errors

| Variable | | REGION | | | | |
|---|---|---|---|---|---|---|
| | | Total | 1 | 2 | 3 | 4 |
| ER_VISITS | Mean | 852.67 | 1498.47 | 1067.28 | 438.26 | 1449.93 |
| | SE Mean | 55.26 | 85.42 | 148.73 | 33.10 | 211.44 |
| | Total | 5371839.99 | 732749.44 | 1746076.79 | 1369139.81 | 1523873.95 |
| | SE Total | 348146.09 | 41772.02 | 243320.44 | 103401.73 | 222228.51 |

*Raking to Frame Size-Variable Totals*

Since drug-related emergency-department visits are likely to be nearly linearly related to the emergency-department visits on the frame, a more efficient version of raking calibrates the weights not to the total number of hospitals by ownership, urbanicity, and region but to the total number of frame visits to public, urban, and regional hospitals. This calibration can be done in WTADJX.

The MODEL statement in *Exhibit 7* is the same as we just used with WTADJUST, but a CALVARS statement is added that contains the new calibration targets: PUBLICZ, METROZ, and REGION * Z, the last creates the four regional frame-visit variables for us. This is possible because REGION is in a CLASS statement.
The associated totals, which are provided on the frame, appear in the POSTWGT statement.

Since NOINT is in the MODEL statement, it must also be in the CALVARS statement.

We output the estimated drug-related emergency-department total and its standard error onto OUT2 for later comparison.

**Exhibit 7. Estimating the Totals by Region by Raking to Frame Size Variables**

```
PROC WTADJX DATA = R DESIGN = STRWOR ADJUST = POST; NEST STRATUM; WEIGHT W;
    TOTCNT BIG_N;
    CLASS REGION;   VAR ER_VISITS ;
    MODEL _ONE_ =  PUBLIC METRO REGION/NOINT;
    CALVARS  PUBLICZ METROZ REGION*Z/NOINT;
    POSTWGT    58000   44000    22000   43000   33000 36000;
    OUTPUT TOTAL SE_TOTAL/FILENAME=OUT2 REPLACE;
RUN;
```

*Two Examples of Nearly Pseudo-Optimal Calibration*

In the run of WTADJX in Exhibit 7, the weight adjustment was a function of PUBIC, METRO, and the four regions. Kott (2011) argues that replacing each with a instrumental variable of the form Variable * Z * W-1 will usually result in smaller standard errors. This *nearly pseudo-optimal* approach to calibration is performed in *Exhibit 8*.

We output the estimated drug-related emergency-department total and its standard error onto OUT3 for later comparison.

**Exhibit 8. Estimating the Totals by Region with Nearly Pseudo-optimal calibration**

```
PROC WTADJX DATA = R DESIGN = STRWOR ADJUST = POST; NEST STRATUM;     WEIGHT W;
  TOTCNT BIG_N;
  CLASS REGION;   VAR ER_VISITS ;
  MODEL _ONE_ =   PUBLICZW1 METROZW1 REGION*ZW1/NOINT;
  CALVARS  PUBLICZ METROZ REGION*Z/NOINT;
  POSTWGT  58000   44000    22000   43000   33000 36000;  /* Computed  from the frame */
  OUTPUT TOTAL SE_TOTAL/FILENAME = OUT3 REPLACE;
RUN;
```

Finally, note that the implied prediction model relating the survey variables to the calibration variables does not have an intercept. We introduce one in *Exhibit 9* by adding _ONE_ to the CALVARS statement and ZW1 to the MODEL statement. Because they differ, we still use the NOINT option in the MODEL and CALVARS steps.

We output the estimated drug-related emergency-department totals and their standard errors onto OUT4 for later comparison.

**Exhibit 9. Estimating the Totals by Region with Nearly Pseudo-optimal calibration and an intercept**

```
PROC WTADJX DATA = R DESIGN = STRWOR ADJUST = POST;  NEST STRATUM; WEIGHT W;
   TOTCNT BIG_N;
   CLASS REGION;  VAR ER_VISITS
   MODEL _ONE_ = W1  PUBLICZW1 METROZW1 REGION*ZW1/NOINT;
   CALVARS  _ONE_ PUBLICZ METROZ REGION*Z/NOINT;
   POSTWGT 6300   58000   44000    22000   43000   33000 36000;
   OUTPUT TOTAL SE_TOTAL/FILENAME = OUT4 REPLACE;
RUN;
```

*Comparing the Results*

In *Exhibit 10*, we combine the output, totals and coefficients of variation (CVs) from each set of estimates.
The estimates from DESCRIPT are labeled DESTOTAL and DESCV.
The estimates from WTADJUST, which used traditional raking, are labeled RAK1TOTAL and RAK1CV.
The estimates from the first WTADJX, which raked to frame-visit totals are labeled DESTOTAL and DESCV.
The estimates from the second WTADJX, which used nearly pseudo-optimal calibration are labeled QO1TOTAL and QO1CV.
The estimates from the third WTADJX, which used nearly pseudo-optimal calibration and added an intercept are labeled QO2TOTAL and QO2CV.

The estimated totals and their CV's are then printed in *Exhibit 11*.

**Exhibit 10. Comparing the Results (Code)**

```
DATA OUT0; SET OUT0;
DESCV = SETOTAL/TOTAL;
DESTOTAL = TOTAL;
RUN;

DATA OUT1; SET OUT1;
RAK1CV = SE_TOTAL/TOTAL;
RAK1TOTAL = TOTAL;
RUN;

DATA OUT2; SET OUT2;
RAK2CV = SE_TOTAL/TOTAL;
RAK2TOTAL = TOTAL;
RUN;

DATA OUT3; SET OUT3;
QO1CV = SE_TOTAL/TOTAL;
QO1TOTAL = TOTAL;
RUN;

DATA OUT4; SET OUT4;
QO2CV = SE_TOTAL/TOTAL;
QO2TOTAL = TOTAL;
RUN;

DATA C; MERGE  OUT0 OUT1 OUT2 OUT3 OUT4; BY VARIABLE REGION;

DESCV = ROUND(DESCV * 100, .01);
RAK1CV=  ROUND(RAK1CV * 100, .01);
RAK2CV=  ROUND(RAK2CV * 100, .01);
QO1CV = ROUND(QO1CV * 100, .01);
QO2CV = ROUND(QO2CV * 100, .01);

PROC PRINT; ID  REGION; VAR DESTOTAL RAK1TOTAL RAK2TOTAL QO1TOTAL  QO2TOTAL;
PROC PRINT; ID  REGION; VAR  DESCV RAK1CV RAK2CV QO1CV QO2CV; RUN;
```

**Exhibit 11. Comparing the Results (Output)**

| REGION | DESTOTAL | RAK1TOTAL | RAK2TOTAL | Q01TOTAL | Q02TOTAL |
|---|---|---|---|---|---|
| 0 | 5376256.13 | 5371839.99 | 5526307.12 | 5519244.82 | 5531363.83 |
| 1 | 732956.71 | 732749.44 | 785406.82 | 787581.77 | 787026.39 |
| 2 | 1750451.22 | 1746076.79 | 1836788.16 | 1832654.77 | 1833783.13 |
| 3 | 1369022.76 | 1369139.81 | 1425517.22 | 1426593.01 | 1433667.24 |
| 4 | 1523825.45 | 1523873.95 | 1478594.92 | 1472415.27 | 1476887.07 |

| REGION | DESCV | RAK1CV | RAK2CV | Q01CV | Q02CV |
|---|---|---|---|---|---|
| 0 | 6.47 | 6.48 | 2.16 | 1.91 | 1.87 |
| 1 | 5.67 | 5.71 | 3.32 | 3.27 | 3.28 |
| 2 | 13.92 | 13.94 | 3.49 | 2.02 | 1.95 |
| 3 | 7.55 | 7.55 | 3.23 | 3.22 | 3.26 |
| 4 | 14.58 | 14.58 | 5.77 | 5.69 | 5.61 |

Traditional raking, which calibrates to population counts, does little to either the estimated totals or their CVs. In fact, it makes the estimated totals slightly worse (their CVs increase). Calibrating to the frame-visit totals, by contrast, increases the estimated totals everywhere but in the West. Since all the estimated totals are nearly unbiased, such differences are due the random nature of the sample. More importantly, CV's drop sharply when calibrating to frame-visit totals, with nearly pseudo-optimal calibration (with or without adding an intercept) faring a bit better than raking as theory predicts. The impact of adding an intercept decreases CVs overall but not in the East or Midwest.

*A Note on Bounds*

It these examples, default bounding and centering parameters were used. For the centering parameter (CENTER), the default 1 in this context; this is, when ADJUST = POST. The default for the upper bound (UPPERBD) is virtually infinite ($10^{20}$), while the default for the lower bound (LOWERBD) is 0.

Many would prefer to set LOWERBD L, where L = 1/W and W is the original weight. This prevents the calibrated weight (WTFINAL) from falling below 1. Were one to use this lower bound, however, the default value for the centering parameter, which must always be between the lower and upper bounds, would no longer be 1 for those records with W = 1 (in fact, it would become virtually infinite: $(10^{20} + 1)/2$ ). To avoid that, one could remove those records from the data (and not allow their weights to change) before running WTADJUST or WTADJX with LOWERBD L. The calibration targets in POSTWGT would need to be changed accordingly.

Under nearly pseudo-optimal calibration, when W = 1, all the instrumental variables are 0. As a result for calibrated weight for such a record will equal its centering parameter. This means one can use the code

IF W = 1 THEN L = 0; ELSE L = 1/W;

and keep the centering parameter default at 1 while being assured that WTFINAL never falls below 1. That is done in *Exhibit 12* in what was otherwise the last WTADJX run. The results are in *Exhibit 13*.

**Exhibit 12. Adding a Lower Bound to Nearly Quasi-optimal Calibration**

```
DATA R; SET R;
IF W = 1 THEN L = 0; ELSE L = 1/W;

PROC WTADJX DATA = R DESIGN = STRWOR ADJUST = POST;
NEST STRATUM; WEIGHT W;
   TOTCNT BIG_N;
   CLASS REGION;  VAR  ER_VISITS ;
   LOWERBD  L;
   MODEL _ONE_ = W1  PUBLICZW1 METROZW1
REGION*ZW1/NOINT;
   CALVARS  _ONE_ PUBLICZ METROZ REGION*Z/NOINT;
   POSTWGT 6300   58000   44000    22000   43000   33000 36000;
   OUTPUT TOTAL SE_TOTAL/FILENAME = OUT5 REPLACE;
RUN;

DATA OUT5; SET OUT5;
QO3TOTAL = TOTAL;
QO3CV = SE_TOTAL/TOTAL;
QO3CV=  ROUND(QO3CV * 100, .01);
RUN;

PROC PRINT; ID REGION; VAR QO3TOTAL QO3CV; RUN;
```

**Exhibit 13. The New Results**

| REGION | QO3TOTAL | QO3CV |
|---|---|---|
| 0 | 5532038.22 | 1.87 |
| 1 | 787534.82 | 3.29 |
| 2 | 1833932.29 | 1.95 |
| 3 | 1433680.29 | 3.26 |
| 4 | 1476890.82 | 5.61 |

Not surprisingly, the results are very close to what we had before (for QO2TOTAL and QO2CV).

If we so desired, we could make additional changes to the bounds without affecting the large-sample unbiasedness of the resulting estimates. Strictly speaking, doing so changes raking to "generalized raking" (Deville *et al.* 1993). SUDAAN 11 allows the user to see the impact of such changes on standard errors.

*A Note on Frame Coverage Errors*

As noted above, the two forms of raking, but *not* nearly pseudo-optimal calibration, can be used when the population totals for the calibration variables come from a reliable outside source while the frame itself suffers from coverage errors due to missing records (undercoverage), duplication (overcoverage), or both. When the the design features with-replacement sampling in its first or only stage, no change is needed to compute standard errors appropriately.

When, as in our data set, sampling is without replacement and *there is only undercoverage in the frame*, the the design statement should be supplemented with the word VARNONADJ; for example,

```
PROC WTADJUST DATA = R DESIGN = STRWOR ADJUST = POST VARNONADJ; NEST STRATUM;
  WEIGHT W;
```

would replace the first line of code in our traditional raking example. This is because adjustments for undercoverage are perfectly analogous to adjustments for nonresponse when ADJUST = POST. They are both versions of missingness. (Although it may be tempting to change the lower bound in the previous code from the default of 0 to 1 by adding the statement LOWER BD 1, WTADJUST will not converge if that statement were added given our data.)

When the sampling is without replacement in the first or only stage, and there is some duplication in the data set, then SUDAAN cannot compute standard errors properly. An advisable strategy in that situation is to compute conservative measures of the standard errors by putting an analogous with-replacement design after DESIGN = [whatever] . In our example, it would be DESIGN = STRWR.