Workplace and Employee Survey - Estimation

The initial sample selection determines the design weight of each unit. During the survey process the initial design weights may undergo several adjustments, which strive to maintain the representativity of the sample. For WES two adjustments are made, one to compensate for complete non-response and one to diminish the influence of stratum jumpers on estimates. To adjust for non-response one multiplies the initial design weights of responding units by a ratio of all sampled units to all responding units within each stratum. This process is predicated on the assumption that respondents and non-respondents behave alike. Since non-response exists mainly amongst the smaller units, this assumption is not unreasonable.

Adjusting for stratum jumpers is more complex as there are at least three methods for dealing with this problem in general. One can either decrease the design weight of the stratum jumper and distribute the difference over the remaining units within the stratum, or one can reduce its values, or one can remove the unit entirely and treat it as non-response. We selected the first option where we targeted approximately 30 employers for a design weight adjustment.

The use of the design weights, whether initial or adjusted, results in unbiased yet sometimes inefficient estimates. To improve the efficiency of the estimation process, one can benchmark, or calibrate, the sample to a set of known or efficiently estimated population totals. In WES this is done using total employment estimated by SEPH at the industry by region level, at which the WES estimates are forced to agree with the SEPH estimates. The resulting adjustment factors are applied to the adjusted design weights. Benchmarking is of the most benefit in situations where the calibration variable (in WES, it is employment) is highly correlated with the variables of interest.

The product of the adjusted design weight and the calibration factor is the final workplace weight. The final linked weight is obtained by adjusting the workplace weight for live employers with no responding employees before applying the calibration factor. The final employee weight accounts for selection of employees and additional non-response of employees. These final weights are used for computing statistics such as totals, means, regression coefficients, etc. To estimate the variance of these statistics, one has to use software packages that allow the user to specify the survey design. If one uses products such as SAS without suitably transforming the survey weights, the resulting underestimation of the variance may be quite severe.

Variance estimation

There are many avenues open to the analysts wishing to produce appropriate variance estimates. One is to use the Statistics Canada Generalized Estimations System (GES) that will handle the estimation of totals, means, and ratios for a variety of designs. The use of GES by external researchers may be financially prohibitive given its licensing costs.

A second option is by far the most general and the easiest to put into practice. It involves the use of bootstrap weights. Bootstrap is a statistical technique whereby one uses a re-sampling technique to generate a number of sets of weights that, if used correctly, capture the variability of a wide variety of statistics. The idea is to compute a large number of "bootstrap" estimates and then calculate their variance.

Once the bootstrap weights are computed, they can be specified in the weight statement in any SAS procedure that has one. To estimate the variance of a statistic, one has to produce an estimate based on each set of bootstrap weights. Then one uses the variability among these bootstrap estimates to produce an appropriate variance estimate of the desired statistic.

The use of Bootstrap weights for variance estimation

When one computes the variances for estimates based on samples coming from finite populations, one has to account for the sampling design. This is not easily done in most statistical analysis software packages. Although most of them do allow the use of weights, they do not use them in the proper manner thus often resulting in the underestimation of the variance. This could have dire consequences for hypothesis testing and for constructing confidence intervals.

Over the years statistical agencies have developed systems to deal with finite populations but most of them lack the flexibility needed to do data analysis. This is where BOOTSTRAP comes in. It is a technique based on re-sampling. One uses the original sample, from which one selects a simple random sample with replacement of as many units as one has at the outset. This procedure is repeated many times to guarantee convergence. This leads to several set of bootstrap weights. In WES, the mean bootstrap methodology is used, where each set of bootstrap weights is in fact obtained as an average of many (in WES, it is 50) sets of bootstrap weights.

Once the bootstrap weights are computed, they can be specified in the weight statement in any SAS procedure that has one. To estimate the variance for a desired statistic, one has to produce an estimate based on each set of bootstrap weights. Then one computes the variability among these bootstrap estimates to produce an appropriate variance estimate of the desired statistic. Below are two examples of how this can be achieved for totals and for correlation coefficients

Depending on your analysis you would use either the wkp_bsw1-wkp_bsw100 (workplace bootstrap weights), emp_bsw1-emp_bsw100 (employee bootstrap weights) or lnk_bsw1-lnk_bsw100 (linked bootstrap weights). SPSS users will use wkp_b1-wkp_b100, emp_b1-emp_b100, lnk_b1_lnk_b100. The following example looks at workplace information.

```
      PROC SUMMARY DATA = WES NWAY;
      CLASS DOM_IND;
      VAR WKP_FINAL_WT WKP_BSW1-WKP_BSW100;
      WEIGHT TTL_EMP;
      OUTPUT OUT = ESTIM (DROP = _FREQ_ _TYPE_)
      SUM = EMPL WKP_BSW1-WKP_BSW100;
      RUN;
      PROC TRANSPOSE DATA = ESTIM
      OUT = T_ESTIM (DROP = _NAME_ RENAME = (COL1 = ESTIM));
VAR WKP_BSW1-WKP_BSW100;
BY DOM_IND;
      RUN;
      PROC SUMMARY DATA = T_ESTIM NWAY;
      CLASS DOM_IND;
      VAR ESTIM;
      OUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
      CSS = VAR;
      RUN;
      DATA ESTIM;
      MERGE ESTIM (KEEP = DOM_IND EMPL)
      VAR;
      BY DOM_IND;
      CV = ROUND (SQRT(50 / 100 * VAR) / EMPL, 0.01);
      RUN;
```

The first SUMMARY procedure uses a trick that allows one to compute all necessary estimates in one simple step. This can only be done when one is producing estimates for a single variable. The trick is to specify the bootstrap weights as the analysis variables and to use the analysis variable as the weight. The estimates are computed at the domain industry level specified by the class statement.

After estimates have been computed, transposed and renamed, another SUMMARY procedure is used to compute their variance (actually, their corrected sum of squares, or CSS in SAS). And finally, multiplying the CSS by 50 / 100 produces the correct design variance. The denominator (100) is the normal adjustment *n* that yields the classical variance. The numerator (50) reflects the fact that each set of bootstrap weights has been averaged over 50 iterations, resulting in an average bootstrap weight. Therefore, the adjustment injects back the variability that has been lost by using the average.

The next example illustrates the use of bootstrap weights for computing correlation coefficients. Here, one has to use a macro to compute individual coefficients, as one cannot easily use the above trick.

```
%MACRO COR_COEF;
            %DO I = 1 %TO 100;
                  PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
VAR TTL_EMP CBA_EMP;
BY DOM_IND;
WEIGHT WKP_BSW&I;
                  RUN;
                  DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP
            = CORR));
SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
                  RUN;
                  PROC DATASETS FORCE NOLIST;
APPEND BASE = ESTIM DATA = CORRS;
QUIT;
                  RUN;
            %END;
      %MEND;
      %COR_COEF;
      PROC SUMMARY DATA = ESTIM NWAY;
CLASS DOM_IND;
AR CORR;
VOUTPUT OUT = VAR (DROP = _FREQ_ _TYPE_)
CSS = VAR;
      RUN;
      PROC CORR DATA = BOOT OUTP = CORRS NOPRINT;
VAR TTL_EMP CBA_EMP;
BY DOM_IND;
WEIGHT WKP_FINAL_WT;
      RUN;
      DATA CORRS (KEEP = DOM_IND CBA_EMP RENAME = (CBA_EMP =
EST_CORR));
SET CORRS (WHERE = (_TYPE_ = 'CORR' & _NAME_ = 'TTL_EMP'));
      RUN;
      DATA ESTIM;
MERGE VAR CORRS;
BY DOM_IND;
CV = ROUND(SQRT(50 / 100 * VAR) / EST_CORR * 100, 0.01);
      RUN;
```

The macro COR_COEF computes correlation coefficients based on each set of bootstrap weights. The example here treats two continuous variables but may be easily extended to multiple variables both continuous and categorical. After estimates have been computed, the corrected sum of squares is produced along with a correlation coefficient that is based on the final weights.

The two files are then merged, the corrected sum of squares is adjusted and a CV is computed. Similar steps should be followed for computing variances of regression estimates, principal components, and other statistic. With the exception of totals of a single variable the computations cannot be done in one step. To reduce computing time per iteration it is recommended that the initial data set be reduced to the analysis variables.

Additional codes written in STATA and SAS showing how to use the WES bootstrap weights to perform a wide array of statistical analyses are included in \CODE. This set of codes is anchored in prior work by François Brisebois (SPSS and SAS macros for NPHS), Pierre Felx (SAS macros for WES), Tony Fang (STATA macro for WES) and Dominic Grenier (STATA and SAS macros for LSIC). The focus of these macros is not estimation of means, totals or ratios; these programs are rather primarily prepared with a view at illustrating the use of the WES bootstrap weights in statistical modelling. The codes allow the following types of analyses:

- linear regression
- T-test
- analysis of variance
- analysis of covariance
- logistic regression
- probit models
- multinomial logistic regression
- ordinal logit models
- ordinal probit models
- generalized estimating equation (GEE)
- generalized linear models (the entire family)
- goodness-of-fit, homogeneity and association tests using both the first- and second-order Rao-Scott corrections

The programs are flexible, easy to reproduce, easy to use and generalizable to any survey for which bootstrap weights are available.

Flexibility:
The programs are not provided as STATA ado files or as SAS macros to be saved in a macro library. The experienced users as well as those with less experience with STATA or SAS can, with minor work, adapt these codes to the particular problem at hand. They can easily expand or contract them. The less experienced users may want to use them as is, in their current formulation.

Ease of reproduction:
The same programming structure is repeated in every program. This pattern can be easily extended to or reproduced with other statistical models for which no explicit bootstrap codes are provided.

Ease of use:
First of all, the users prepare a data set with the relevant variables required by the models they want to fit. This dataset must be augmented with the bootstrap weights; depending on the type of analysis, the employee, linked or employer survey final weight are also included.

Then, on the Stata model command line, users have to specify the name of their own variables and the final weight they are going to use, as in the examples provided. These programs use the stub of the bootstrap weight variable, emp_bsw for the employee portion; for an analysis using the workplace portion, the stub would be wkp_bsw, for example.

In the SAS macros, at the beginning of the programs, users need to specify the survey final weight, the number of bootstrap weights, the number of iterations, the dataset they intend to use and the stub for the bootstrap weights.

To be specified at the beginning:

```
%let bsw = emp_bsw;/* in the employee file use emp_bsw, in the employer
                          replace that variable by wkp_bsw */
%let fwgt = emp_final_wt; /*Use the variable name for the final weight, e.g.
                              emp_final_wt for the employee file, wkp_final_wt for the
                              employer file*/
%let dsn=boot_data;/* this data set has the subset of relevant variables for the
                          analysis and the bootstrap weights*/
%let b=100;/* the number of bootstrap weights available in the file*/
%let iter=50;
```

To be specified at the end:

```
%linregress(boot_data,hr_waget,age) /* the number of items in this line depend
                                          on the models an other macro parameters
                                          needed. This line does a regression analysis,
                                          stating that hourly wage as a function of
                                          employee age based on the dataset
                                          boot_data.
```

Finally, the results are saved in a directory provided by the users, by replacing the path "c:\Documents and Settings\decayve\bootstrap_yves\res.dta" with their own path.

Generalizability:

These program files can be used with any survey that provides bootstrap weights. The unique aspect that makes it particular to WES is that in the computation of the variance, the fact that each bootstrap weight represents an average of 50 iterations was taken into account. In the STATA program, this was translated with the instruction “local iter = 50. With other surveys, users have only to replace 50 by 1 in that line. Also, if a particular survey provides 1000 bootstrap weights, just replace 100 by 1000 in the command “local bs = 100”. That is only what is needed.

The same can be accomplished with the SAS macros by replacing

"%let iter= 50" with "%let iter = 1".

When 1000 bootstrap weights are provided, replace

"%let b =100" with "%let b =1000".

Commercial packages such as SUDAAN and WesVar can be used to perform bootstrap variance estimation if the variance estimation approach is specified as BRR, and if the bootstrap weight variables are specified as BRR weights (D. Binder and G. Roberts, 2004, in "Statistical inference in survey data analysis: Where does the sample design fit in?"). With WES bootstrap weights, the results provided by these packages would have to be adjusted to account for the fact that each set of bootstrap weights has been averaged over 50 iterations, resulting in an average bootstrap weight. The codes provided herein take the iterations into account, rendering them therefore specific to WES. However, by setting the number of iterations equal to one, the generalizability to all surveys providing bootstrap weights to their users is achieved.