Title

stata.com

Calibration — Calibration for survey data

Description

Calibration is a method for adjusting the sampling weights, often to account for nonresponse and underrepresented groups in the population.

See [SVY] **Poststratification** for a discussion of a weight adjustment method that is a special case of the calibration methods discussed here.

See [SVY] **Direct standardization** for a similar method of adjustment that allows the comparison of rates that come from different frequency distributions.

Remarks and examples

stata.com

The standard application of calibration uses population totals to adjust the sampling weights. Population totals are typically taken from a census or other source separate from the survey. In a business survey, the frame might have the number of employees from an earlier time period for each establishment. In a household survey, counts of persons in groups defined by age, race or ethnicity, and gender may be published from a census, population projections, or a separate survey.

Calibration involves adjusting the sampling weights so that they more closely estimate known population totals. Calibration is more general than poststratification because the weight adjustments can be made across multiple group-identifier variables simultaneously and also includes population totals other than simple counts. Calibration usually results in decreasing bias because of nonresponse and underrepresented groups in the population. Much like poststratification, calibration also tends to result in smaller variance estimates.

The `svyset` command has the options `rake()` and `regress()` for applying calibration adjustments to the sampling weights. `rake()` specifies that the weights be adjusted via the raking-ratio method. `regress()` specifies that the weights be adjusted via linear regression. `rake()` and `regress()` produce the same weight adjustment as poststratification when they are used to adjust the sampling weights across the levels of a single group-identifier variable.

In the following example, we use a version of the data that Valliant and Dever (2018) resampled from the Survey of Mental Health Organizations (SMHO) (Manderscheid and Henderson 2002).

▷ Example 1: Population mean, using calibrated weights

Valliant and Dever (2018, sec. 4.3) give an example of calibration by using a stratified simple random sample of 120 hospitals from the SMHO. The sample is stratified by four hospital types, identified in the variable `hosptype`. The four levels of `hosptype` are 1–Psychiatric, 2–Residential or veterans, 3–General, and 5–Multi-service, substance abuse. Within each hospital type, a simple random sample of 30 hospitals was selected without replacement.

For each sampled hospital, the `eoycnt` variable contains the end-of-year patient counts, and the `beds` variable contains the number of beds. Our separate source for the auxiliary information about the population is the full SMHO. The population total for `eoycnt` is 505,345. For `beds`, the population totals within each of the four hospital types is given in the following table:

hosptype	beds
1	37,978
2	13,066
3	9,573
5	10,077

We can use this information to adjust the original sampling weights that are stored in `wt`. In the calibration model specification, we use the interaction between the categorical variable `hosptype` and the continuous variable `beds` to specify regressors for the number of beds for each hospital type. See **[U] 11.4.3 Factor variables** for details of the interaction specification.

```
. use https://www.stata-press.com/data/r17/smho
(Resampled SMHO data from Valliant & Dever, 2018)
. svyset [pw=wt], strata(hosptype)
>        regress(eoycnt i.hosptype#c.beds, noconstant
>                totals(eoycnt=505345
>                       1.hosptype#c.beds=37978
>                       2.hosptype#c.beds=13066
>                       3.hosptype#c.beds=9573
>                       5.hosptype#c.beds=10077))
Sampling weights: wt
             VCE: linearized
     Calibration: regress
     Single unit: missing
        Strata 1: hosptype
 Sampling unit 1: <observations>
           FPC 1: <zero>
```

The goal of the analysis is to estimate the mean expenditures per hospital in the population. The variable `exptot` contains the expenditures for each of the sampled hospitals, measured in millions of dollars.

```
. svy: mean exptot
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =   4            Number of obs   =        120
Number of PSUs   = 120            Population size = 889.062959
Calibration: regress              Design df       =        116
```

	Mean	Linearized std. err.	[95% conf.	interval]
exptotal	10.40552	.8560101	8.710082	12.10095

The estimated mean expenditures per hospital are $10.41 (in millions) with a standard error of 0.86.

In the following, we re-estimate the mean expenditures per hospital using just the original survey design characteristics.

```
. svyset [pw=wt], strata(hosptype)
Sampling weights: wt
             VCE: linearized
     Single unit: missing
        Strata 1: hosptype
 Sampling unit 1: <observations>
           FPC 1: <zero>
. svy: mean exptot
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =   4                  Number of obs    = 120
Number of PSUs   = 120                  Population size = 725
                                        Design df        = 116
```

	Mean	Linearized std. err.	[95% conf.	interval]
exptotal	9.939402	.9487877	8.060209	11.8186

The result is a mean expenditure per hospital of \$9.94 (in millions) with a standard error of 0.95. The difference between the mean estimates is explained by the fact that end-of-year patient counts and number of beds are strong predictors of hospital expenditures.

◁

Methods and formulas

The following discussion assumes that you are already familiar with the topics discussed in [SVY] **Variance estimation**.

Calibration methods adjust the sampling weights to minimize the difference between known population totals and their weighted estimates. For a full discussion on the motivation and derivation of the methods described here, see Deville and Särndal (1992); Deville, Särndal, and Sautory (1993); and Valliant (2002).

Suppose that you used a complex survey design to sample m individuals from a population of size M. Let $\mathbf{T}_a$ be a collection of population totals corresponding to a collection of auxiliary variables denoted by $\mathbf{a}$. The adjusted weights take on the form

$$w_j^* = w_j F(\mathbf{a}_j'\boldsymbol{\lambda})$$

where w_j are the original sampling weights, $F(z)$ is derived from a chosen calibration method, and $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers computed by solving the calibration equations

$$\sum_{j=1}^{m} w_j F(\mathbf{a}_j'\boldsymbol{\lambda})\mathbf{a}_j = \mathbf{T}_a$$

The linear regression method uses

$$F(z) = \begin{cases} 1+z & \text{if } z \in [L-1, U-1] \\ L & \text{if } z < L-1 \\ U & \text{if } z > U-1 \end{cases}$$

and corresponds to `svyset`'s option `regress()` with suboptions `ll(`L`)` and `ul(`U`)`. By default, $L = -\infty$ and $U = \infty$; otherwise, the only restriction is that $L < 1 < U$.

The raking-ratio method uses

$$F(z) = e^z$$

and corresponds to `svyset`'s option `rake()` specified without limits on the weight ratios. With limits on the weight ratios, the restrictions are $0 \leq L < 1 < U$. By default, $L = 0$; otherwise, U must be specified if a different value of L is specified. Therefore,

$$F(z) = \frac{L(U-1) + U(1-L)\exp(Az)}{U - 1 + (1-L)\exp(Az)}$$

where

$$A = \frac{U-L}{(1-L)(U-1)}$$

Point estimates are computed using the adjusted weights w_j^*. For example, the calibrated total estimator is

$$\widehat{Y}^C = \sum_{j=1}^{m} w_j^* y_j$$

where y_j is an item from the jth sampled individual.

For replication-based variance estimation, the replicate-weight variables are similarly adjusted to produce the replicate values used in the respective variance formulas.

The score variable for the linearized variance estimator of a calibrated total is taken directly from the residuals of a weighted linear regression of y_j on $\mathbf{a}_j$ using the adjusted weights w_j^*. Let these residuals be denoted by $z_j(\widehat{Y}^C)$. For the calibrated ratio estimator, the score variable is

$$z_j(\widehat{R}^C) = \frac{\widehat{X}^C z_j(\widehat{Y}^C) - \widehat{Y}^C z_j(\widehat{X}^C)}{(\widehat{X}^C)^2}$$

where $\widehat{X}^C$ is the calibrated total estimator for item x_j. For regression models, the equation-level scores are computed similarly to those of the calibrated total; that is, the adjusted scores are taken directly from the residuals of a weighted linear regression of each original score on $\mathbf{a}_j$ using the adjusted weights w_j^*.

References

Deville, J.-C., and C.-E. Särndal. 1992. Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87: 376–382. https://doi.org/10.1080/01621459.1992.10475217.

Deville, J.-C., C.-E. Särndal, and O. Sautory. 1993. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association* 88: 1013–1020. https://doi.org/10.2307/2290793.

Manderscheid, R. W., and M. J. Henderson, ed. 2002. *Mental Health, United States, 2002*. DHHS Publication No. SMA04-3938, Rockville, MD: Substance Abuse and Mental Health Services Administration.

Valliant, R. 2002. Variance estimation for the general regression estimator. *Survey Methodology* 28: 103–114.

Valliant, R., and J. Dever. 2018. *Survey Weights: A Step-by-Step Guide to Calculation*. College Station, TX: Stata Press.

Also see

[SVY] **Survey** — Introduction to survey commands

[SVY] **svy** — The survey prefix command

[SVY] **svyset** — Declare survey design for dataset

[SVY] **Direct standardization** — Direct standardization of means, proportions, and ratios

[SVY] **Poststratification** — Poststratification for survey data