Title

stata.com

Poststratification — Poststratification for survey data

Description

Poststratification is a method for adjusting the sampling weights, usually to account for underrepresented groups in the population.

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

Poststratification involves adjusting the sampling weights so that they sum to the population sizes within each poststratum. This usually results in decreasing bias because of nonresponse and underrepresented groups in the population. Poststratification also tends to result in smaller variance estimates.

The `svyset` command has options to set variables for applying poststratification adjustments to the sampling weights. The `poststrata()` option takes a variable that contains poststratum identifiers, and the `postweight()` option takes a variable that contains the poststratum population sizes.

In the following example, we use an example from Levy and Lemeshow (2008) to show how poststratification affects the point estimates and their variance.

▷ Example 1: Poststratified mean

Levy and Lemeshow (2008, sec. 6.6) give an example of poststratification by using simple survey data from a veterinarian's client list. The data in `poststrata.dta` were collected using simple random sampling without replacement. The `totexp` variable contains the total expenses to the client, `type` identifies the cats and dogs, `postwgt` contains the poststratum sizes (450 for cats and 850 for dogs), and `fpc` contains the total number of clients ($850 + 450 = 1300$).

```
. use https://www.stata-press.com/data/r18/poststrata
. svyset, poststrata(type) postweight(postwgt) fpc(fpc)
Sampling weights: <none>
             VCE: linearized
      Poststrata: type
Post. pop. sizes: postwgt
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: fpc
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =  1                  Number of obs   =    50
Number of PSUs   = 50                  Population size = 1,300
N. of poststrata =  2                  Design df       =    49
```

	Mean	Linearized std. err.	[95% conf.	interval]
totexp	40.11513	1.163498	37.77699	42.45327

The mean total expenses is \$40.12 with a standard error of \$1.16. In the following, we omit the poststratification information from `svyset`, resulting in mean total expenses of \$39.73 with standard error \$2.22. The difference between the mean estimates is explained by the facts that expenses tend to be larger for dogs than for cats and that the dogs were slightly underrepresented in the sample ($850/1{,}300 \approx 0.65$ for the population; $32/50 = 0.64$ for the sample). This reasoning also explains why the variance estimate from the poststratified mean is smaller than the one that was not poststratified.

```
. svyset, fpc(fpc)
Sampling weights: <none>
             VCE: linearized
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: fpc
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =  1                  Number of obs   = 50
Number of PSUs   = 50                  Population size = 50
                                       Design df       = 49
```

	Mean	Linearized std. err.	[95% conf.	interval]
totexp	39.7254	2.221747	35.26063	44.19017

◁

Methods and formulas

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

Suppose that you used a complex survey design to sample m individuals from a population of size M. Let P_k be the set of individuals in the sample that belong to poststratum k, and let $I_{P_k}(j)$ indicate if the jth individual is in poststratum k, where

$$I_{P_k}(j) = \begin{cases} 1, \text{ if } j \in P_k \\ 0, \text{ otherwise} \end{cases}$$

Also let L_P be the number of poststrata and M_k be the population size for poststratum k.

If w_j is the unadjusted sampling weight for the jth sampled individual, the poststratification adjusted sampling weight is

$$w_j^* = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} w_j$$

where $\widehat{M}_k$ is

$$\widehat{M}_k = \sum_{j=1}^{m} I_{P_k}(j) w_j$$

The point estimates are computed using these adjusted weights. For example, the poststratified total estimator is

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

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 poststratified total is

$$z_j(\widehat{Y}^P) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left(y_j - \frac{\widehat{Y}_k}{\widehat{M}_k} \right) \tag{1}$$

where $\widehat{Y}_k$ is the total estimator for the kth poststratum,

$$\widehat{Y}_k = \sum_{j=1}^{m} I_{P_k}(j) w_j y_j$$

For the poststratified ratio estimator, the score variable is

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

where $\widehat{X}^P$ is the poststratified total estimator for item x_j. For regression models, the equation-level scores are adjusted as in (1). These score variables were derived using the method described in [SVY] **Variance estimation** for the ratio estimator and are a direct result of the methods described in Deville (1999), Demnati and Rao (2004), and Shah (2004).

References

Demnati, A., and J. N. K. Rao. 2004. Linearization variance estimators for survey data. *Survey Methodology* 30: 17–26.

Deville, J.-C. 1999. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* 25: 193–203.

Levy, P. S., and S. A. Lemeshow. 2008. *Sampling of Populations: Methods and Applications*. 4th ed. Hoboken, NJ: Wiley.

Shah, B. V. 2004. Comment [on Demnati and Rao (2004)]. *Survey Methodology* 30: 29.

Also see

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

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

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

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

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