Title

stata.com

Variance estimation — Variance estimation for survey data

Description Remarks and examples References Also see

Description

Stata's suite of estimation commands for survey data use the most commonly used variance estimation techniques: bootstrap, balanced repeated replication, jackknife, successive difference replication, and linearization. The bootstrap, balanced repeated replication, jackknife, and successive difference replication techniques are known as replication methods in the survey literature. We stick with that nomenclature here, but note that these techniques are also known as resampling methods. This entry discusses the details of these variance estimation techniques.

Also see Cochran (1977), Wolter (2007), and Shao and Tu (1995) for some background on these variance estimators.

Remarks and examples

stata.com

Remarks are presented under the following headings:

Variance of the total

This section describes the methods and formulas for `svy: total`. The variance estimators not using replication methods use the variance of a total as an important ingredient; this section therefore also introduces variance estimation for survey data.

We will discuss the variance estimators for two complex survey designs:

1. The stratified single-stage design is the simplest design that has the elements present in most complex survey designs.
2. Adding a second stage of clustering to the previous design results in a variance estimator for designs with multiple stages of clustered sampling.

Stratified single-stage design

The population is partitioned into groups called *strata*. Clusters of observations are randomly sampled—with or without replacement—from within each stratum. These clusters are called *primary sampling units* (PSUs). In single-stage designs, data are collected from every member of the sampled PSUs. When the observed data are analyzed, sampling weights are used to account for the survey design. If the PSUs were sampled without replacement, a finite population correction (FPC) is applied to the variance estimator.

The `svyset` syntax to specify this design is

`svyset` *psu* `[pweight=`*weight*`], strata(`*strata*`) fpc(`*fpc*`)`

The stratum identifiers are contained in the variable named *strata*, PSU identifiers are contained in variable *psu*, the sampling weights are contained in variable *weight*, and the values for the FPC are contained in variable *fpc*.

Let $h = 1, \ldots, L$ count the strata and (h, i) denote the ith PSU in stratum h, where $i = 1, \ldots, N_h$ and N_h is the number of PSUs in stratum h. Let (h, i, j) denote the jth individual from PSU (h, i) and M_{hi} be the number of individuals in PSU (h, i); then

$$M = \sum_{h=1}^{L} \sum_{i=1}^{N_h} M_{hi}$$

is the number of individuals in the population. Let Y_{hij} be a survey item for individual (h, i, j); for example, Y_{hij} might be income for adult j living in block i of county h. The associated population total is

$$Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} Y_{hij}$$

Let y_{hij} denote the items for individuals who are members of the sampled PSUs; here $h = 1, \ldots, L$; $i = 1, \ldots, n_h$; and $j = 1, \ldots, m_{hi}$. The number of individuals in the sample (number of observations) is

$$m = \sum_{h=1}^{L} \sum_{i=1}^{n_h} m_{hi}$$

The estimator for Y is

$$\widehat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}$$

where w_{hij} is a sampling weight, and its unadjusted value for this design is $w_{hij} = N_h/n_h$. The estimator for the number of individuals in the population (population size) is

$$\widehat{M} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

The estimator for the variance of $\widehat{Y}$ is

$$\widehat{V}(\widehat{Y}) = \sum_{h=1}^{L}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\overline{y}_h)^2 \tag{1}$$

where y_{hi} is the weighted total for PSU (h,i),

$$y_{hi} = \sum_{j=1}^{m_{hi}} w_{hij}y_{hij}$$

and $\overline{y}_h$ is the mean of the PSU totals for stratum h:

$$\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$$

The factor $(1-f_h)$ is the FPC for stratum h, and f_h is the sampling rate for stratum h. The sampling rate f_h is derived from the variable specified in the `fpc()` option of `svyset`. If an FPC variable is not `svyset`, then $f_h = 0$. If an FPC variable is set and its values are greater than or equal to n_h, then the variable is assumed to contain the values of N_h, and f_h is given by $f_h = n_h/N_h$. If its values are less than or equal to 1, then the variable is assumed to contain the sampling rates f_h.

If multiple variables are supplied to `svy: total`, covariances are also computed. The estimator for the covariance between $\widehat{Y}$ and $\widehat{X}$ (notation for X is defined similarly to that of Y) is

$$\widehat{\mathrm{Cov}}(\widehat{Y},\widehat{X}) = \sum_{h=1}^{L}(1-f_h)\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\overline{y}_h)(x_{hi}-\overline{x}_h)$$

Stratified two-stage design

The population is partitioned into strata. PSUs are randomly sampled without replacement from within each stratum. Clusters of observations are then randomly sampled—with or without replacement—from within the sampled PSUs. These clusters are called *secondary sampling units* (SSUs). Data are then collected from every member of the sampled SSUs. When the observed data are analyzed, sampling weights are used to account for the survey design. Each sampling stage provides a component to the variance estimator and has its own FPC.

The `svyset` syntax to specify this design is

`svyset` *psu* `[pweight=`*weight*`], strata(`*strata*`) fpc(`*fpc*$_1$`) || `*ssu*`, fpc(`*fpc*$_2$`)`

The stratum identifiers are contained in the variable named *strata*, PSU identifiers are contained in variable *psu*, the sampling weights are contained in variable *weight*, the values for the FPC for the first sampling stage are contained in variable *fpc*$_1$, SSU identifiers are contained in variable *ssu*, and the values for the FPC for the second sampling stage are contained in variable *fpc*$_2$.

The notation for this design is based on the previous notation. There still are L strata, and (h,i) identifies the ith PSU in stratum h. Let M_{hi} be the number of SSUs in PSU (h,i), M_{hij} be the number of individuals in SSU (h,i,j), and

$$M = \sum_{h=1}^{L}\sum_{i=1}^{N_h}\sum_{j=1}^{M_h i} M_{hij}$$

be the population size. Let Y_{hijk} be a survey item for individual (h, i, j, k); for example, Y_{hijk} might be income for adult k living in block j of county i of state h. The associated population total is

$$Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} \sum_{j=1}^{M_{hi}} \sum_{k=1}^{M_{hij}} Y_{hijk}$$

Let y_{hijk} denote the items for individuals who are members of the sampled SSUs; here $h = 1, \ldots, L$; $i = 1, \ldots, n_h$; $j = 1, \ldots, m_{hi}$; and $k = 1, \ldots, m_{hij}$. The number of observations is

$$m = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} m_{hij}$$

The estimator for Y is

$$\widehat{Y} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{m_{hij}} w_{hijk} y_{hijk}$$

where w_{hijk} is a sampling weight, and its unadjusted value for this design is

$$w_{hijk} = \left(\frac{N_h}{n_h}\right)\left(\frac{M_{hi}}{m_{hi}}\right)$$

The estimator for the population size is

$$\widehat{M} = \sum_{h=1}^{L} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \sum_{k=1}^{m_{hij}} w_{hijk}$$

The estimator for the variance of $\widehat{Y}$ is

$$\begin{aligned} \widehat{V}(\widehat{Y}) = & \sum_{h=1}^{L} (1 - f_h) \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \overline{y}_h)^2 \\ & + \sum_{h=1}^{L} f_h \sum_{i=1}^{n_h} (1 - f_{hi}) \frac{m_{hi}}{m_{hi} - 1} \sum_{j=1}^{m_{hi}} (y_{hij} - \overline{y}_{hi})^2 \end{aligned} \tag{2}$$

where y_{hi} is the weighted total for PSU (h, i); $\overline{y}_h$ is the mean of the PSU totals for stratum h; y_{hij} is the weighted total for SSU (h, i, j),

$$y_{hij} = \sum_{k=1}^{m_{hij}} w_{hijk} y_{hijk}$$

and $\overline{y}_{hi}$ is the mean of the SSU totals for PSU (h, i),

$$\overline{y}_{hi} = \frac{1}{m_{hi}} \sum_{j=1}^{m_{hi}} y_{hij}$$

Equation (2) is equivalent to (1) with an added term representing the increase in variability because of the second stage of sampling. The factor $(1 - f_h)$ is the FPC, and f_h is the sampling rate for the first stage of sampling. The factor $(1 - f_{hi})$ is the FPC, and f_{hi} is the sampling rate for PSU (h, i). The sampling rate f_{hi} is derived in the same manner as f_h.

If multiple variables are supplied to `svy: total`, covariances are also computed. For estimated totals $\widehat{Y}$ and $\widehat{X}$ (notation for X is defined similarly to that of Y), the covariance estimator is

$$\widehat{\mathrm{Cov}}(\widehat{Y}, \widehat{X}) = \sum_{h=1}^{L}(1 - f_h)\frac{n_h}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \overline{y}_h)(x_{hi} - \overline{x}_h) + \sum_{h=1}^{L} f_h \sum_{i=1}^{n_h}(1 - f_{hi})\frac{m_{hi}}{m_{hi} - 1}\sum_{j=1}^{m_{hi}}(y_{hij} - \overline{y}_{hi})(x_{hij} - \overline{x}_{hi})$$

On the basis of the formulas (1) and (2), writing down the variance estimator for a survey design with three or more stages is a matter of deriving the variance component for each sampling stage. The sampling units from a given stage pose as strata for the next sampling stage.

All but the last stage must be sampled without replacement to get nonzero variance components from each stage of clustered sampling. For example, if $f_h = 0$ in (2), the second stage contributes nothing to the variance estimator.

Variance for census data

The point estimates that result from the analysis of census data, in which the entire population was sampled without replacement, are the population's parameters instead of random variables. As such, there is no sample-to-sample variation if we consider the population fixed. Here the sampling fraction is one; thus, if the FPC variable you `svyset` for the first sampling stage is one, Stata will report a standard error of zero.

Certainty sampling units

Stata's `svy` commands identify strata with an FPC equal to one as units sampling with certainty. To properly determine the design degrees of freedom, certainty sampling units should be contained within their own strata, one for each certainty unit, in each sampling stage. Although the observations contained in certainty units from a given sampling stage play a role in parameter estimation, they contribute nothing to the variance for that stage.

Strata with one sampling unit

By default, Stata's `svy` commands report missing standard errors when they encounter a stratum with one sampling unit. Although the best way to solve this problem is to reassign the sampling unit to another appropriately chosen stratum, there are three automatic alternatives that you can choose from, in the `singleunit()` option, when you `svyset` your data.

`singleunit(certainty)` treats the strata with single sampling units as certainty units.

`singleunit(scaled)` treats the strata with single sampling units as certainty units but multiplies the variance components from each stage by a scaling factor. For a given sampling stage, suppose that L is the total number of strata, L_c is the number of certainty strata, and L_s is the number of strata with one sampling unit, and then the scaling factor is $(L - L_c)/(L - L_c - L_s)$. Using this scaling factor is the same as using the average of the variances from the strata with multiple sampling units for each stratum with one sampling unit.

`singleunit(centered)` specifies that strata with one sampling unit are centered at the population mean instead of the stratum mean. The quotient $n_h/(n_h - 1)$ in the variance formula is also taken to be 1 if $n_h = 1$.

Ratios and other functions of survey data

Shah (2004) points out a simple procedure for deriving the linearized variance for functions of survey data that are continuous functions of the sampling weights. Let θ be a (possibly vector-valued) function of the population data and $\widehat{\theta}$ be its associated estimator based on survey data.

1. Define the jth observation of the score variable by

$$z_j = \frac{\partial \widehat{\theta}}{\partial w_j}$$

   If $\widehat{\theta}$ is implicitly defined through estimating equations, z_j can be computed by taking the partial derivative of the estimating equations with respect to w_j.

2. Define the weighted total of the score variable by

$$\widehat{Z} = \sum_{j=1}^{m} w_j z_j$$

3. Estimate the variance $V(\widehat{Z})$ by using the design-based variance estimator for the total $\widehat{Z}$. This variance estimator is an approximation of $V(\widehat{\theta})$.

Revisiting the total estimator

As a first example, we derive the variance of the total from a stratified single-stage design. Here you have $\widehat{\theta} = \widehat{Y}$, and deriving the score variable for $\widehat{Y}$ results in the original values of the variable of interest.

$$z_j(\widehat{\theta}) = z_j(\widehat{Y}) = \frac{\partial \widehat{Y}}{\partial w_j} = y_j$$

Thus you trivially recover the variance of the total given in (1) and (2).

The ratio estimator

The estimator for the population ratio is

$$\widehat{R} = \frac{\widehat{Y}}{\widehat{X}}$$

and its score variable is

$$z_j(\widehat{R}) = \frac{\partial \widehat{R}}{\partial w_j} = \frac{y_j - \widehat{R}\, x_j}{\widehat{X}}$$

Plugging this into (1) or (2) results in a variance estimator that is algebraically equivalent to the variance estimator derived from directly applying the delta method (a first-order Taylor expansion with respect to y and x)

$$\widehat{V}(\widehat{R}) = \frac{1}{\widehat{X}^2}\{\widehat{V}(\widehat{Y}) - 2\widehat{R}\,\widehat{\mathrm{Cov}}(\widehat{Y}, \widehat{X}) + \widehat{R}^2\,\widehat{V}(\widehat{X})\}$$

A note about score variables

The functional form of the score variable for each estimation command is detailed in the *Methods and formulas* section of its manual entry; see [R] **total**, [R] **ratio**, and [R] **mean**.

Although Deville (1999) and Demnati and Rao (2004) refer to z_j as the *linearized variable*, here it is referred to as the *score variable* to tie it more closely to the model-based estimators discussed in the following section.

Linearized/robust variance estimation

The regression models for survey data that allow the `vce(linearized)` option use *linearization*-based variance estimators that are natural extensions of the variance estimator for totals. For general background on regression and generalized linear model analysis of complex survey data, see Binder (1983); Cochran (1977); Fuller (1975); Godambe (1991); Kish and Frankel (1974); Särndal, Swensson, and Wretman (1992); and Skinner (1989).

Suppose that you observed $(Y_j, \mathbf{x}_j)$ for the entire population and are interested in modeling the relationship between Y_j and $\mathbf{x}_j$ by the vector of parameters $\boldsymbol{\beta}$ that solve the following estimating equations:

$$G(\boldsymbol{\beta}) = \sum_{j=1}^{M} S(\boldsymbol{\beta}; Y_j, \mathbf{x}_j) = 0$$

For ordinary least squares, $G(\boldsymbol{\beta})$ is the normal equations

$$G(\boldsymbol{\beta}) = X'Y - X'X\boldsymbol{\beta} = 0$$

where Y is the vector of outcomes for the full population and X is the matrix of explanatory variables for the full population. For a pseudolikelihood model—such as logistic regression—$G(\boldsymbol{\beta})$ is the first derivative of the log-pseudolikelihood function with respect to $\boldsymbol{\beta}$. Estimate $\boldsymbol{\beta}$ by solving for $\widehat{\boldsymbol{\beta}}$ from the weighted sample estimating equations

$$\widehat{G}(\boldsymbol{\beta}) = \sum_{j=1}^{m} w_j S(\boldsymbol{\beta}; y_j, \mathbf{x}_j) = 0 \tag{3}$$

The associated estimation command with `iweights` will produce point estimates $\widehat{\boldsymbol{\beta}}$ equal to the solution of (3).

A first-order matrix Taylor-series expansion yields

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left\{\frac{\partial \widehat{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{-1} \widehat{G}(\boldsymbol{\beta})$$

with the following variance estimator for $\widehat{\boldsymbol{\beta}}$:

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = \left[\left\{\frac{\partial \widehat{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{-1} \widehat{V}\{\widehat{G}(\boldsymbol{\beta})\} \left\{\frac{\partial \widehat{G}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}^{-\mathrm{T}}\right]\Bigg|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = D\widehat{V}\{\widehat{G}(\boldsymbol{\beta})\}\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} D'$$

where D is $(X_s' W X_s)^{-1}$ for linear regression (where W is a diagonal matrix of the sampling weights and X_s is the matrix of sampled explanatory variables) or the inverse of the negative Hessian matrix from the pseudolikelihood model. Write $\widehat{G}(\boldsymbol{\beta})$ as

$$\widehat{G}(\boldsymbol{\beta}) = \sum_{j=1}^{m} w_j \mathbf{d}_j$$

where $\mathbf{d}_j = s_j \mathbf{x}_j$ and s_j is a residual for linear regression or an equation-level score from the pseudolikelihood model. The term *equation-level score* means the derivative of the log pseudolikelihood with respect to $\mathbf{x}_j \boldsymbol{\beta}$. In either case, $\widehat{G}(\widehat{\boldsymbol{\beta}})$ is an estimator for the total $G(\boldsymbol{\beta})$, and the variance estimator $\widehat{V}\{\widehat{G}(\boldsymbol{\beta})\}\big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}}$ is computed using the design-based variance estimator for a total.

The above result is easily extended to models with ancillary parameters, multiple regression equations, or both.

The bootstrap

The bootstrap methods for survey data used in recent years are largely due to McCarthy and Snowden (1985), Rao and Wu (1988), and Rao, Wu, and Yue (1992). For example, Yeo, Mantel, and Liu (1999) cite Rao, Wu, and Yue (1992) with the method for variance estimation used in the National Population Health Survey conducted by Statistics Canada.

In the survey bootstrap, the model is fit multiple times, once for each of a set of adjusted sampling weights that mimic bootstrap resampling. The variance is estimated using the resulting replicated point estimates.

Let $\widehat{\boldsymbol{\theta}}$ be the vector of point estimates computed using the sampling weights for a given survey dataset (for example, $\widehat{\boldsymbol{\theta}}$ could be a vector of means, ratios, or regression coefficients). Each bootstrap replicate is produced by fitting the model with adjusted sampling weights. The adjusted sampling weights are derived from the method used to resample the original survey data.

According to Yeo, Mantel, and Liu (1999), if n_h is the number of observed PSUs in stratum h, then $n_h - 1$ PSUs are sampled with replacement from within stratum h. This sampling is performed independently across the strata to produce one bootstrap sample of the survey data. Let r be the number of bootstrap samples. Suppose that we are about to generate the adjusted-weight variable for the ith bootstrap replication and w_{hij} is the sampling weight attached to the jth observation in the ith PSU of stratum h. The adjusted weight is

$$w^*_{hij} = \frac{n_h}{n_h - 1} m^*_{hi} w_{hij}$$

where m^*_{hi} is the number of times the ith cluster in stratum h was resampled.

To accommodate privacy concerns, many public-use datasets contain replicate-weight variables derived from the "mean bootstrap" described by Yung (1997). In the mean bootstrap, each adjusted weight is derived from b bootstrap samples instead of one. The adjusted weight is

$$w^*_{hij} = \frac{n_h}{n_h - 1} \overline{m}^*_{hi} w_{hij}$$

where

$$\overline{m}^*_{hi} = \frac{1}{b} \sum_{k=1}^{b} m^*_{hik}$$

is the average of the number of times the ith cluster in stratum h was resampled among the b bootstrap samples.

Each replicate is produced using an adjusted-weight variable with the estimation command that computed $\widehat{\boldsymbol{\theta}}$. The adjusted-weight variables must be supplied to `svyset` with the `bsrweight()` option. For the mean bootstrap, b must also be supplied to `svyset` with the `bsn()` option; otherwise, `bsn(1)` is assumed. We call the variables supplied to the `bsrweight()` option *bootstrap replicate-weight variables* when $b = 1$ and *mean bootstrap replicate-weight variables* when $b > 1$.

Let $\widehat{\boldsymbol{\theta}}_{(i)}$ be the vector of point estimates from the ith replication. When the `mse` option is specified, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \frac{b}{r}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}'$$

Otherwise, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \frac{b}{r}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}'$$

where $\overline{\boldsymbol{\theta}}_{(.)}$ is the bootstrap mean,

$$\overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{r}\sum_{i=1}^{r}\widehat{\boldsymbol{\theta}}_{(i)}$$

BRR

BRR was first introduced by McCarthy (1966, 1969a, and 1969b) as a method of variance estimation for designs with two PSUs in every stratum. The BRR variance estimator tends to give more reasonable variance estimates for this design than the linearized variance estimator, which can result in large values and undesirably wide confidence intervals.

The model is fit multiple times, once for each of a balanced set of combinations where one PSU is dropped (or downweighted) from each stratum. The variance is estimated using the resulting replicated point estimates (replicates). Although the BRR method has since been generalized to include other designs, Stata's implementation of BRR requires two PSUs per stratum.

Let $\widehat{\boldsymbol{\theta}}$ be the vector of point estimates computed using the sampling weights for a given stratified survey design (for example, $\widehat{\boldsymbol{\theta}}$ could be a vector of means, ratios, or regression coefficients). Each BRR replicate is produced by dropping (or downweighting) a PSU from every stratum. This could result in as many as 2^L replicates for a dataset with L strata; however, the BRR method uses Hadamard matrices to identify a balanced subset of the combinations from which to produce the replicates.

A Hadamard matrix is a square matrix, H_r (with r rows and columns), such that $H_r'H_r = rI$, where I is the identity matrix. The elements of H_r are $+1$ and -1; -1 causes the first PSU to be downweighted and $+1$ causes the second PSU to be downweighted. Thus r must be greater than or equal to the number of strata.

Suppose that we are about to generate the adjusted-weight variable for the ith replication and w_j is the sampling weight attached to the jth observation, which happens to be in the first PSU of stratum h. The adjusted weight is

$$w_j^* = \begin{cases} fw_j, & \text{if } H_r[i,h] = -1 \\ (2-f)w_j, & \text{if } H_r[i,h] = +1 \end{cases}$$

where f is Fay's adjustment (Judkins 1990). By default, $f = 0$.

Each replicate is produced using an adjusted-weight variable with the estimation command that computed $\widehat{\boldsymbol{\theta}}$. The adjusted-weight variables can be generated by Stata or supplied to `svyset` with the `brrweight()` option. We call the variables supplied to the `brrweight()` option *BRR replicate-weight variables*.

Let $\widehat{\boldsymbol{\theta}}_{(i)}$ be the vector of point estimates from the ith replication. When the `mse` option is specified, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \frac{1}{r(1-f)^2}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}'$$

Otherwise, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \frac{1}{r(1-f)^2}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}'$$

where $\overline{\boldsymbol{\theta}}_{(.)}$ is the BRR mean,

$$\overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{r}\sum_{i=1}^{r}\widehat{\boldsymbol{\theta}}_{(i)}$$

The jackknife

The jackknife method for variance estimation is appropriate for many models and survey designs. The model is fit multiple times, and each time one or more PSUs are dropped from the estimation sample. The variance is estimated using the resulting replicates (replicated point estimates).

Let $\widehat{\boldsymbol{\theta}}$ be the vector of point estimates computed using the sampling weights for a given survey design (for example, $\widehat{\boldsymbol{\theta}}$ could be a vector of means, ratios, or regression coefficients). The dataset is resampled by dropping one or more PSUs from one stratum and adjusting the sampling weights before recomputing a replicate for $\widehat{\boldsymbol{\theta}}$.

Let w_{hij} be the sampling weight for the jth individual from PSU i in stratum h. Suppose that you are about to generate the adjusted weights for the replicate resulting from dropping k PSUs from stratum h. The adjusted weight is

$$w^*_{abj} = \begin{cases} 0, & \text{if } a = h \text{ and } b \text{ is dropped} \\ \dfrac{n_h}{n_h - k}w_{abj}, & \text{if } a = h \text{ and } b \text{ is not dropped} \\ w_{abj}, & \text{otherwise} \end{cases}$$

Each replicate is produced by using the adjusted-weight variable with the estimation command that produced $\widehat{\boldsymbol{\theta}}$. For the delete-one jackknife (where one PSU is dropped for each replicate), adjusted weights can be generated by Stata or supplied to `svyset` with the `jkrweight()` option. For the delete-k jackknife (where $k > 1$ PSUs are dropped for each replicate), the adjusted-weight variables must be supplied to `svyset` using the `jkrweight()` option. The variables supplied to the `jkrweight()` option are called *jackknife replicate-weight variables*.

The delete-one jackknife

Let $\widehat{\boldsymbol{\theta}}_{(h,i)}$ be the point estimates (replicate) from leaving out the ith PSU from stratum h. The pseudovalue for replicate (h,i) is

$$\widehat{\boldsymbol{\theta}}^{*}_{h,i} = \widehat{\boldsymbol{\theta}}_{(h,i)} + n_h\{\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_{(h,i)}\}$$

When the `mse` option is specified, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \sum_{h=1}^{L}(1-f_h)\,m_h\sum_{i=1}^{n_h}\{\widehat{\boldsymbol{\theta}}_{(h,i)} - \widehat{\boldsymbol{\theta}}\}\{\widehat{\boldsymbol{\theta}}_{(h,i)} - \widehat{\boldsymbol{\theta}}\}'$$

and the jackknife mean is

$$\overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\widehat{\boldsymbol{\theta}}_{(h,i)}$$

where f_h is the sampling rate and m_h is the jackknife multiplier associated with stratum h. Otherwise, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \sum_{h=1}^{L}(1-f_h)\,m_h\sum_{i=1}^{n_h}\{\widehat{\boldsymbol{\theta}}_{(h,i)} - \overline{\boldsymbol{\theta}}_h\}\{\widehat{\boldsymbol{\theta}}_{(h,i)} - \overline{\boldsymbol{\theta}}_h\}', \qquad \overline{\boldsymbol{\theta}}_h = \frac{1}{n_h}\sum_{i=1}^{n_h}\widehat{\boldsymbol{\theta}}_{(h,i)}$$

and the jackknife mean is

$$\overline{\boldsymbol{\theta}}^{*} = \frac{1}{n}\sum_{h=1}^{L}\sum_{i=1}^{n_h}\widehat{\boldsymbol{\theta}}^{*}_{h,i}$$

The multiplier for the delete-one jackknife is

$$m_h = \frac{n_h - 1}{n_h}$$

The delete-k jackknife

Let $\widetilde{\boldsymbol{\theta}}_{(h,d)}$ be one of the point estimates that resulted from leaving out k PSUs from stratum h. Let c_h be the number of such combinations that were used to generate a replicate for stratum h; then $d = 1, \ldots, c_h$. If all combinations were used, then

$$c_h = \frac{n_h!}{(n_h - k)!k!}$$

The pseudovalue for replicate (h,d) is

$$\widetilde{\boldsymbol{\theta}}^{*}_{h,d} = \widetilde{\boldsymbol{\theta}}_{(h,d)} + c_h\{\widehat{\boldsymbol{\theta}} - \widetilde{\boldsymbol{\theta}}_{(h,d)}\}$$

When the `mse` option is specified, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \sum_{h=1}^{L}(1-f_h)\,m_h\sum_{d=1}^{c_h}\{\widetilde{\boldsymbol{\theta}}_{(h,d)} - \widehat{\boldsymbol{\theta}}\}\{\widetilde{\boldsymbol{\theta}}_{(h,d)} - \widehat{\boldsymbol{\theta}}\}'$$

and the jackknife mean is

$$\overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{C}\sum_{h=1}^{L}\sum_{d=1}^{c_h}\widetilde{\boldsymbol{\theta}}_{(h,d)}, \quad C = \sum_{h=1}^{L}c_h$$

Otherwise, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = \sum_{h=1}^{L}(1-f_h)\,m_h \sum_{d=1}^{c_h}\{\widetilde{\boldsymbol{\theta}}_{(h,d)} - \overline{\boldsymbol{\theta}}_h\}\{\widetilde{\boldsymbol{\theta}}_{(h,d)} - \overline{\boldsymbol{\theta}}_h\}', \qquad \overline{\boldsymbol{\theta}}_h = \frac{1}{c_h}\sum_{d=1}^{c_h}\widetilde{\boldsymbol{\theta}}_{(h,d)}$$

and the jackknife mean is

$$\overline{\boldsymbol{\theta}}^* = \frac{1}{C}\sum_{h=1}^{L}\sum_{d=1}^{c_h}\widetilde{\boldsymbol{\theta}}^*_{h,d}$$

The multiplier for the delete-k jackknife is

$$m_h = \frac{n_h - k}{c_h k}$$

Variables containing the values for the stratum identifier h, the sampling rate f_h, and the jackknife multiplier m_h can be `svyset` using the respective suboptions of the `jkrweight()` option: `stratum()`, `fpc()`, and `multiplier()`.

Successive difference replication

Successive difference replication (SDR) was first introduced by Fay and Train (1995) as a method of variance estimation for annual demographic supplements to the Current Population Survey. This method is typically applied to systematic samples, where the observed sampling units are somehow ordered.

In SDR, the model is fit multiple times, once for each of a set of adjusted sampling weights. The variance is estimated using the resulting replicated point estimates.

Let $\widehat{\boldsymbol{\theta}}$ be the vector of point estimates computed using the sampling weights for a given survey dataset (for example, $\widehat{\boldsymbol{\theta}}$ could be a vector of means, ratios, or regression coefficients). Each SDR replicate is produced by fitting the model with adjusted sampling weights. The SDR method uses Hadamard matrices to generate these adjustments.

A Hadamard matrix is a square matrix, H_r (with r rows and columns), such that $H_r'H_r = rI$, where I is the identity matrix. Let h_{ij} be an element of H_r; then $h_{ij} = 1$ or $h_{ij} = -1$. In SDR, if n is the number of PSUs, then we must find H_r with $r \geq n + 2$.

Without loss of generality, we will assume the ordered PSUs are individuals instead of clusters. Suppose that we are about to generate the adjusted-weight variable for the ith replication and that w_j is the sampling weight attached to the jth observation. The adjusted weight is $w^*_j = f_{ji}w_j$, where f_{ji} is

$$f_{ji} = 1 + \frac{1}{2\sqrt{2}}(h_{j+1,i} - h_{j+2,i})$$

Here we assume that the elements of the first row of H_r are all 1.

Each replicate is produced using an adjusted-weight variable with the estimation command that computed $\widehat{\boldsymbol{\theta}}$. The adjusted-weight variables must be supplied to `svyset` with the `sdrweight()` option. We call the variables supplied to the `sdrweight()` option *SDR replicate-weight variables*.

Let $\widehat{\boldsymbol{\theta}}_{(i)}$ be the vector of point estimates from the ith replication, and let f be the sampling fraction computed using the FPC information `svyset` in the `fpc()` suboption of the `sdrweight()` option, where $f = 0$ when `fpc()` is not specified. When the `mse` option is specified, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = (1-f)\frac{4}{r}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \widehat{\boldsymbol{\theta}}\}'$$

Otherwise, the variance estimator is

$$\widehat{V}(\widehat{\boldsymbol{\theta}}) = (1-f)\frac{4}{r}\sum_{i=1}^{r}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}\{\widehat{\boldsymbol{\theta}}_{(i)} - \overline{\boldsymbol{\theta}}_{(.)}\}'$$

where $\overline{\boldsymbol{\theta}}_{(.)}$ is the SDR mean,

$$\overline{\boldsymbol{\theta}}_{(.)} = \frac{1}{r}\sum_{i=1}^{r}\widehat{\boldsymbol{\theta}}_{(i)}$$

Confidence intervals

In survey data analysis, the customary number of degrees of freedom attributed to a test statistic is $d = n - L$, where n is the number of PSUs and L is the number of strata. Under regularity conditions, an approximate $100(1-\alpha)\%$ confidence interval for a parameter θ (for example, θ could be a total, ratio, or regression coefficient) is

$$\widehat{\theta} \pm t_{1-\alpha/2,d}\,\{\widehat{V}(\widehat{\theta})\}^{1/2}$$

Cochran (1977, sec. 2.8) and Korn and Graubard (1990) give some theoretical justification for using $d = n - L$ to compute univariate confidence intervals and p-values. However, for some cases, inferences based on the customary $n - L$ degrees-of-freedom calculation may be excessively liberal; the resulting confidence intervals may have coverage rates substantially less than the nominal $1 - \alpha$. This problem generally is of the greatest practical concern when the population of interest has a skewed or heavy-tailed distribution or is concentrated in a few PSUs. In some of these cases, the user may want to consider constructing confidence intervals based on alternative degrees-of-freedom terms, based on the Satterthwaite (1941, 1946) approximation and modifications thereof; see, for example, Cochran (1977, sec. 5.4) and Eltinge and Jang (1996).

Sometimes there is no information on n or L for datasets that contain replicate-weight variables but no PSU or strata variables. Each of `svy`'s replication commands has its own default behavior when the design degrees of freedom are not `svyset` or specified using the `dof()` option. `svy brr:` and `svy jackknife:` use $d = r - 1$, where r is the number of replications. `svy bootstrap:` and `svy sdr:` use $z_{1-\alpha/2}$ for the critical value instead of $t_{1-\alpha/2,d}$.

References

Binder, D. A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51: 279–292. https://doi.org/10.2307/1402588.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

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.

Eltinge, J. L., and D. S. Jang. 1996. Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology* 22: 157–165.

Fay, R. E., and G. F. Train. 1995. Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. In *Proceedings of the Government Statistics Section*, 154–159. American Statistical Association.

Fuller, W. A. 1975. Regression analysis for sample survey. *Sankhyā, Series C* 37: 117–132.

Godambe, V. P., ed. 1991. *Estimating Functions*. Oxford: Oxford University Press.

Judkins, D. R. 1990. Fay's method for variance estimation. *Journal of Official Statistics* 6: 223–239.

Kish, L., and M. R. Frankel. 1974. Inference from complex samples. *Journal of the Royal Statistical Society, Series B* 36: 1–37. https://doi.org/10.1111/j.2517-6161.1974.tb00981.x.

Kolenikov, S. 2010. Resampling variance estimation for complex survey data. *Stata Journal* 10: 165–199.

Korn, E. L., and B. I. Graubard. 1990. Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni t statistics. *American Statistician* 44: 270–276. https://doi.org/10.2307/2684345.

McCarthy, P. J. 1966. Replication: An approach to the analysis of data from complex surveys. In *Vital and Health Statistics*, ser. 2. Hyattsville, MD: National Center for Health Statistics.

——. 1969a. Pseudoreplication: Further evaluation and application of the balanced half-sample technique. In *Vital and Health Statistics*, ser. 2. Hyattsville, MD: National Center for Health Statistics.

——. 1969b. Pseudo-replication: Half-samples. *Revue de l'Institut International de Statistique* 37: 239–264. https://doi.org/10.2307/1402116.

McCarthy, P. J., and C. B. Snowden. 1985. The bootstrap and finite population sampling. In *Vital and Health Statistics*, 1–23. Washington, DC: U.S. Government Printing Office.

Rao, J. N. K., and C. F. J. Wu. 1988. Resampling inference with complex survey data. *Journal of the American Statistical Association* 83: 231–241. https://doi.org/10.2307/2288945.

Rao, J. N. K., C. F. J. Wu, and K. Yue. 1992. Some recent work on resampling methods for complex surveys. *Survey Methodology* 18: 209–217.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York: Springer.

Satterthwaite, F. E. 1941. Synthesis of variance. *Psychometrika* 6: 309–316. https://doi.org/10.1007/BF02288586.

——. 1946. An approximate distribution of estimates of variance components. *Biometrics Bulletin* 2: 110–114. https://doi.org/10.2307/3002019.

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

Shao, J., and D. Tu. 1995. *The Jackknife and Bootstrap*. New York: Springer.

Skinner, C. J. 1989. Introduction to part A. In *Analysis of Complex Surveys*, ed. C. J. Skinner, D. Holt, and T. M. F. Smith, 23–58. New York: Wiley.

Wolter, K. M. 2007. *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

Yeo, D., H. Mantel, and T.-P. Liu. 1999. Bootstrap variance estimation for the National Population Health Survey. In *Proceedings of the Survey Research Methods Section*, 778–785. American Statistical Association.

Yung, W. 1997. Variance estimation for public use files under confidentiality constraints. In *Proceedings of the Survey Research Methods Section*, 434–439. American Statistical Association.

Also see

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

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

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

[P] **_robust** — Robust variance estimates