Title

survey — Introduction to survey commands

Description

The *Survey Data Reference Manual* organizes the commands alphabetically, making it easy to find individual command entries if you know the name of the command. This overview organizes and presents the commands conceptually, that is, according to the similarities in the functions that they perform.

The following list of commands may have been updated since the release of Stata 9. For an updated list, type the following in an up-to-date Stata:

```
. help survey
```

Survey design tools

`svyset`	Declare survey design for dataset
`svydes`	Describe survey data

Descriptive statistics

`svy: mean`	Estimation of population and subpopulation means
`svy: proportion`	Estimation of population and subpopulation proportions
`svy: ratio`	Estimation of population and subpopulation ratios
`svy: total`	Estimation of population and subpopulation totals
`svy: tabulate oneway`	One-way tables for survey data
`svy: tabulate twoway`	Two-way tables for survey data

Regression models

`svy: regress`	Linear regression for survey data
`svy: ivreg`	Instrumental variables regression for survey data
`svy: intreg`	Interval regression for survey data
`svy: logistic`	Logistic regression, reporting odds ratios, for survey data
`svy: logit`	Logistic regression, reporting coefficients, for survey data
`svy: probit`	Probit regression for survey data
`svy: mlogit`	Multinomial logistic regression for survey data
`svy: ologit`	Ordered logistic regression for survey data
`svy: oprobit`	Ordered probit regression for survey data
`svy: poisson`	Poisson regression for survey data
`svy: nbreg`	Negative binomial regression for survey data
`svy: gnbreg`	Generalized negative binomial regression for survey data
`svy: heckman`	Heckman selection model for survey data
`svy: heckprob`	Probit regression with selection for survey data

Survey data analysis tools

`svy`	Overview of the `svy` prefix command
`svy brr`	Balanced repeated replication for survey data
`svy jackknife`	Jackknife estimation for survey data
svy postestimation	Overview of postestimation commands for survey data analysis
`estat`	Postestimation statistics for survey data
`ml for svy`	Maximum pseudolikelihood estimation for survey data
`svymarkout`	Mark observations for exclusion based on survey characteristics

Survey data concepts

variance estimation	Variance estimation for survey data
subpopulation estimation	Subpopulation estimation for survey data
direct standardization	Direct standardization of means, proportions, and ratios
poststratification	Poststratification for survey data

Remarks

Remarks are presented under the headings

Overview
Survey design tools
Descriptive statistics
Regression models
Survey data analysis tools
Survey data concepts

Overview

Stata's facilities for survey data are centered around the `svy` prefix command. Once the design characteristics of a survey dataset are identified with the `svyset` command, the `svy` prefix can be used with supported estimation commands in essentially the same way as the corresponding command for nonsurvey data. For example, where you would normally use the `regress` command to fit a linear regression using nonsurvey data, use `svy: regress` for your survey data.

Why should you use the `svy` prefix command rather than, say, the `mean` command for means or `regress` for linear regression? To answer this question, we need to discuss some of the characteristics of survey design and survey data collection because these characteristics affect how we must perform our analysis if we want to "get it right".

Survey data are characterized by the following:

1. sampling weights, also called probability weights—`pweights` in Stata's syntax
2. cluster sampling
3. stratification

These factors arise from the design of the data collection procedure. Here's a brief description of how these design features affect the analysis of the data:

1. *Sampling weights.* In sample surveys, observations are selected through a random process, but different observations may have different probabilities of selection. Weights are equal to (or proportional to) the inverse of the probability of being sampled. Various postsampling

adjustments to the weights are sometimes made, as well. A weight of w_j for the jth observation means, roughly speaking, that the jth observation represents w_j elements in the population from which the sample was drawn.

Omitting weights from the analysis results in estimates that may be biased, sometimes seriously so. Sampling weights also need to be taken in account when estimating standard errors, and for purposes of testing and inference.

2. *Clustering.* Individuals are not sampled independently in almost all survey designs. Collections of individuals (for example, counties, city blocks, or households) are typically sampled as a group, known as a *cluster.*

   There may also be further subsampling within the clusters. For example, counties may be sampled, then city blocks within counties, then households within city blocks, and then finally persons within households. The clusters at the first level of sampling are called *primary sampling units* (PSUs)—in this example, counties are the PSUs. In the absence of clustering, the PSUs are defined to be the individuals or, equivalently, clusters each of size one.

   Sampling by cluster implies a sample-to-sample variability of the resulting estimator that is usually greater than that obtained through sampling individually, and this variability must be accounted for when estimating standard errors, testing, or performing other inference.

3. *Stratification.* In surveys, different groups of clusters are often sampled separately. These groups are called *strata*. For example, the 254 counties of a state might be divided into two strata, say, urban counties and rural counties. Then ten counties might be sampled from the urban stratum, and fifteen from the rural stratum.

   Sampling is done independently across strata; the stratum divisions are fixed in advance. Thus strata are statistically independent and can be analyzed as such. When the individual strata are more homogenous than the population as a whole, the homogeneity can be exploited to produce smaller (and honestly so) estimates of standard errors.

To put it succinctly: it is important to use sampling weights in order to get the point estimates right. We must consider the weighting, clustering, and stratification of the survey design to get the standard errors right. If our analysis ignores the clustering in our design, we would likely produce standard errors that are smaller than they should be. Stratification can be used to get smaller standard errors for a given total amount of data.

▷ Example 1: A preview of survey data analysis with Stata

We have (fictional) data on American high school seniors (12th graders), and the data were collected according to the following multistage design. In the first stage, counties were independently selected within each state. In the second stage, schools were selected within each chosen county. Within each chosen school, a questionaire was filled out by every attending high school senior. We've entered all the information into a Stata dataset called `multistage.dta`. The survey design variables are as follows:

1. `state` contains the stratum identifiers
2. `county` contains the first-stage sampling units
3. `ncounties` contains the total number of counties within each state
4. `school` contains the second-stage sampling units
5. `nschools` contains the total number of schools within each county
6. `sampwgt` contains the sampling weight for each sampled individual

Here we load multistage.dta into memory and use svyset with the above variables to declare that this data is survey data.

```
. use http://www.stata-press.com/data/r9/multistage
. svyset county [pw=sampwgt], strata(state) fpc(ncounties) || school, fpc(nschools)
      pweight: sampwgt
          VCE: linearized
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

Now that the data are svyset, we can use the svy estimation commands to perform our analysis. In the following, we estimate the mean of weight (in lbs.) for each subpopulation identified by the categories of the sex variable (male and female).

```
. svy: mean weight, over(sex)
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      50          Number of obs    =     4071
Number of PSUs   =     100          Population size  = 8.0e+06
                                    Design df        =       50
         male: sex = male
       female: sex = female
```

Over	Mean	Linearized Std. Err.	[95% Conf.	Interval]
weight				
male	175.4809	1.116802	173.2377	177.7241
female	146.204	.9004157	144.3955	148.0125

Based on the above results, we are 95% confident that the average weight of male high school seniors is between 173.2 and 177.7 pounds.

Here we use the test command to test the hypothesis that the average male is 30 pounds heavier than the average female; however, based on the results we cannot reject this hypothesis at the 5% level.

```
. test [weight]male - [weight]female = 30
Adjusted Wald test
 ( 1)  [weight]male - [weight]female = 30
       F(  1,    50) =    0.23
            Prob > F =    0.6353
```

◁

Survey design tools

Before using svy, first take a quick look at [SVY] **svyset**. Use the svyset command to specify the variables that identify the survey design characteristics and default method for standard error estimation. Once set, svy will automatically use these design specifications until they are cleared or changed, or a new dataset is loaded into memory.

The svydes command describes the survey design and is useful in, among other things, tracking down strata with only one sampling unit.

Descriptive statistics

svy: mean, svy: ratio, svy: proportion, and svy: total produce estimates of finite-population means, ratios, proportions and totals. svy: mean, svy: ratio, and svy: proportion can also estimate standardized means, ratios, and proportions. Estimates for multiple subpopulations can be obtained using the over() option.

svy: tabulate can be used to produce one-way and two-way tables with survey data and can also produce tests of independence for two-way contingency tables.

Regression models

Many commands in Stata are used to fit regression models to data, for example regress for linear regression, poisson for Poisson regression, logistic for logistic regression, etc. A subset of these *estimation commands* are supported by svy, that is, they may be prefixed by svy: in order to produce results appropriate for complex survey data. Whereas poisson is used with standard, nonsurvey data, svy: poisson is used with survey data. In what follows we refer to any estimation command unprefixed by svy: as the standard command. A standard command prefixed by svy: is referred to as a svy command.

Most standard commands (and all standard commands supported by svy) allow pweights and the cluster(*varname*) option, where *varname* corresponds to the *psu* variable that you svyset. If your survey data exhibit only sampling weights and/or first-stage clusters, you can get by with using the standard command with pweights and/or cluster(). Your parameter estimates will always be identical to those you would have obtained from the svy command, and the standard command uses the same robust (linearization) variance estimator as the svy command with a similarly svyset design.

Most standard commands are also fit using maximum-likelihood methodology. When used with independently distributed, nonweighted data, the likelihood to be maximized is reflective of the joint probability distribution of the data given the chosen model. With complex survey data, however, this interpretation of the likelihood is no longer valid, as survey data are either weighted, not independently distributed, or both. With survey data, (valid) parameter estimates are obtained using the independence-assuming likelihood and weighting if necessary. Since the probabilistic interpretation no longer holds, the likelihood here is instead called a *pseudolikelihood.* See Skinner (1989, section 3.4.4) for a discussion of maximum pseudolikelihood estimators.

Below we highlight the other features of svy commands.

1. svy commands handle stratified sampling, but none of the standard commands do. Since stratification usually makes standard errors smaller, ignoring stratification is usually conservative. So, not using svy with stratified sample data is not a terrible thing to do. However, to get the smallest possible "honest" standard-error estimates for stratified sampling, use svy.

2. svy commands use t statistics with $n - L$ degrees of freedom to test the significance of coefficients, where n is the total number of sampled PSUs (clusters) and L is the number of strata in the first stage. Some of the standard commands use t statistics, but most use z statistics. If the standard command uses z statistics for its standard variance estimator, then it also uses z statistics with the robust (linearization) variance estimator. Strictly speaking, t statistics are appropriate with the robust (linearization) variance estimator; see [P] **_robust** for the theoretical rationale. But, using z rather than t statistics only yields a nontrivial difference when there is a small number of clusters (< 50). If a regression model command uses t statistics and the `cluster()` option is specified, then the degrees of freedom used are the same as that of the svy command (in the absence of stratification).

3. svy commands produce an adjusted Wald test for the model test, and `test` can be used to produce adjusted Wald tests for other hypotheses after svy commands. Only unadjusted Wald tests are available if the svy prefix is not used. The adjustment can be important when the degrees of freedom $n - L$ are small relative to the dimension of the test. (If the dimension is one, then the adjusted and unadjusted Wald tests are identical.) This fact along with point 2 make it important to use the svy command if the number of sampled PSUs (clusters) is small (< 50).

4. `svy: regress` differs slightly from `regress` and `svy: ivreg` differs slightly from `ivreg` in that they use different multipliers for the variance estimator. `regress` and `ivreg` use a multiplier of $\{(N-1)/(N-k)\}\{n/(n-1)\}$, where N is the number of observations, n is the number of clusters (PSUs), and k is the number of regressors including the constant. `svy: regress` and `svy: ivreg` use $n/(n-1)$ instead. Thus they produce slightly different standard errors. The $(N-1)/(N-k)$ is ad hoc and has no rigorous theoretical justification; hence, the purist svy commands do not use it. The svy commands tacitly assume that $N \gg k$. If $(N-1)/(N-k)$ is not close to 1, you may be well advised to use `regress` or `ivreg` so that some punishment is inflicted on your variance estimates. Note that maximum likelihood estimators in Stata (e.g., `logit`) do no such adjustment, but rely on the sensibilities of the analyst to ensure that N is reasonably larger than k. Thus the maximum pseudolikelihood estimators (e.g., `svy: logit`) produce exactly the same standard errors as the corresponding maximum likelihood commands (e.g., `logit`), but p-values are slightly different because of point 2.

5. svy commands can produce proper estimates for subpopulations through use of the `subpop()` option. Use of an *if* restriction with svy or standard commands can yield incorrect standard error estimates for subpopulations. Often an *if* restriction will yield exactly the same standard error as `subpop()`; most other times, the two standard errors will be slightly different; but, in some cases—usually for thinly sampled subpopulations—the standard errors can be appreciably different. Hence, the svy command with the `subpop()` option should be used to obtain estimates for thinly sampled subpopulations. See [SVY] **subpopulation estimation** for more information.

6. svy commands handle zero sampling weights properly. Standard commands ignore any observation with a weight of zero. Usually, this will yield exactly the same standard errors, but sometimes they will differ. Sampling weights of zero can arise from various postsampling adjustment procedures. If the sum of weights for one or more PSUs is zero, svy and standard commands will produce different standard errors, but usually this difference is very small.

7. You can `svyset iweights` and let these weights be negative. Negative sampling weights can arise from various postsampling adjustment procedures. If you want to use negative sampling weights, then you must `svyset iweights` instead of `pweights`; no standard command will allow negative sampling weights.

8. The svy commands compute finite population corrections (FPC).

9. After a svy command, estat effects will compute the design effects DEFF and DEFT and the misspecification effects MEFF and MEFT.
10. svy commands can perform variance estimation that accounts for multiple stages of clustered sampling.
11. svy commands can perform variance estimation that accounts for poststratification adjustments to the sampling weights.

Survey data analysis tools

Stata's suite of survey-data commands is governed by the svy prefix command. svy runs the supplied estimation command while accounting for the survey design characteristics in the point estimates and the variance estimator. The three available variance estimation methods are balanced repeated replication (BRR), the jackknife, and first-order Taylor linearization. By default, svy computes standard errors using the linearized variance estimator—so called because it is based on a first-order Taylor series linear approximation. In the nonsurvey context, we refer to this variance estimator as the *robust* variance estimator, otherwise known in Stata as the Huber/White/sandwich estimator; see [P] **_robust**.

The svy brr and svy jackknife prefix commands can be used with those commands that may not be fully supported by svy but are compatible with BRR and the jackknife; see [SVY] **svy brr** and [SVY] **svy jackknife**.

All the standard postestimation commands (e.g., estat, lincom, nlcom, test, testnl) are also available after svy.

estat has specific subroutines for use after svy. estat svyset reports the survey design settings used to produce the current estimation results. estat effects and estat lceffects report a table of design and misspecification effects for point estimates and linear combinations of point estimates, respectively. estat size reports a table of sample and subpopulation sizes after svy: mean, svy: proportion, svy: ratio, and svy: total.

The ml command can be used to fit a pseudolikelihood model. When maximum pseudolikelihood is carried out using ml, the weighting during estimation and postestimation linearization is performed automatically, provided that the user specifies the appropriate survey options to ml; see [R] **ml** for details.

svymarkout is a programmer's command that resets the values in a variable that identifies the estimation sample, dropping observations for which any of the survey-characteristic variables contain missing values. This tool is most helpful for developing estimation commands that use ml to fit models using maximum pseudolikelihood.

Survey data concepts

The variance estimation methods used by Stata are discussed in [SVY] **variance estimation**.

See [SVY] **subpopulation estimation** for an explanation of why you should use the subpop() option instead of the *if* and *in* options.

The weight adjusting methods for direct standardization and poststratification are discussed in [SVY] **direct standardization** and [SVY] **poststratification**.

For more detailed introductions to complex survey data analysis, see Scheaffer, Mendenhall, and Ott (1996), Stuart (1984), Williams (1978), and Levy and Lemeshow (1999). Advanced treatments and discussion of important special topics are given by Cochran (1977), Korn and Graubard (1999), Särndal, Swensson, and Wretman (1992), Shao and Tu (1995), Skinner, Holt, and Smith (1989), Thompson (2002), and Wolter (1985).

Acknowledgments

Many of the svy commands were developed in collaboration with John L. Eltinge, Bureau of Labor Statistics. We thank him for his invaluable assistance.

We thank Wayne Johnson of the National Center for Health Statistics for providing the NHANES II dataset.

We thank Nicholas Winter, Department of Government, Cornell University, for his diligent efforts to keep Stata up to date with mainstream variance estimation methods for survey data, and for providing versions of `svy brr` and `svy jackknife`.

William Gemmell Cochran (1909–1980) was born in Rutherglen, Scotland, and educated at the Universities of Glasgow and Cambridge. He accepted a post at Rothamsted before finishing his doctorate. Cochran emigrated to the United States in 1939 and worked at Iowa State, North Carolina State, Johns Hopkins, and Harvard. He made many major contributions across several fields of statistics, including experimental design, the analysis of counted data, sample surveys and observational studies, and was author or co-author (with Gertrude M. Cox and George W. Snedecor) of various widely used texts.

References

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

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

——. 1982. *Contributions to Statistics*. New York: Wiley.

Eltinge, J. L. and W. M. Sribney. 1996a. svy1: Some basic concepts for design-based analysis of complex survey data. *Stata Technical Bulletin* 31: 3–6. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 208–213.

——. 1996b. svy4: Linear, logistic, and probit regressions for survey data. *Stata Technical Bulletin* 31: 26–31. Reprinted in *Stata Technical Bulletin Reprints*, vol. 6, pp. 239–245.

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

Garrett, J. M. 2001. sxd4: Sample size estimation for cluster designed samples. *Stata Technical Bulletin* 60: 41–45. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 387–393.

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

Gonzalez J. F., Jr., N. Krauss, and C. Scott. 1992. Estimation in the 1988 National Maternal and Infant Health Survey. In *Proceedings of the Section on Statistics Education, American Statistical Association*, 343–348.

Hansen, M. and F. Mosteller. 1987. William Gemmell Cochran. *Biographical Memoirs, National Academy of Sciences* 56: 60–89.

Johnson, W. 1995. Variance estimation for the NMIHS. Technical document. Hyattsville, MD: National Center for Health Statistics.

Kish, L. and M. R. Frankel. 1974. Inference from complex samples. *Journal of the Royal Statistical Society* B 36: 1–37.

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

——. 1999. *Analysis of Health Surveys*. New York: Wiley.

Kott, P. S. 1991. A model-based look at linear regression with survey data. *The American Statistician* 45: 107–112.

Levy, P. and S. Lemeshow. 1999. *Sampling of Populations.* 3rd ed. New York: Wiley.

McDowell, A., A. Engel, J. T. Massey, and K. Maurer. 1981. Plan and operation of the Second National Health and Nutrition Examination Survey, 1976–1980. *Vital and Health Statistics* 15(1). Hyattsville, MD: National Center for Health Statistics.

McDowell, A. and J. Pitblado. 2002. From the help desk: It's all about the sampling. *Stata Journal* 2: 190–201.

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

Scheaffer, R. L., W. Mendenhall, and L. Ott. 1996. *Elementary Survey Sampling.* 5th ed. Boston: Duxbury Press.

Scott, A. J. and D. Holt. 1982. The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* 77: 848–854.

Shao, J. 1996. Resampling methods for sample surveys (with discussion). *Statistics* 27: 203–254.

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.

Skinner, C. J., D. Holt, and T. M. F. Smith, eds. 1989. *Analysis of Complex Surveys.* New York: Wiley.

Stuart, A. 1984. *The Ideas of Sampling.* 3rd ed. New York: Macmillan.

Thompson, S. K. 2002. *Sampling.* 2nd ed. New York: Wiley.

Watson, G. S. 1982. William Gemmell Cochran 1909–1980. *Annals of Statistics* 10: 1–10.

Williams, B. 1978. *A Sampler on Sampling.* New York: Wiley.

Wolter, K. M. 1985. *Introduction to Variance Estimation.* New York: Springer.

Also See

Complementary: [R] **estat**, [R] **jackknife**, [R] **lincom**, [R] **ml**, [R] **nlcom**, [R] **predict**, [R] **predictnl**, [R] **test**, [R] **testnl**

Background: [P] **_robust**