svy estimation — Estimation commands for survey data

Description

Survey data analysis in Stata is essentially the same as standard data analysis. The standard syntax applies; you just need to also remember the following:

- Use svyset to identify the survey design characteristics.
- Prefix the estimation commands with “svy:”.

For example,

```
. use http://www.stata-press.com/data/r13/nhanes2f
. svyset psuid [pweight=finalwgt], strata(stratid)
. svy: regress zinc age c.age#c.age weight female black orace rural
```

See [SVY] svyset and [SVY] svy.

The following estimation commands support the svy prefix:

Descriptive statistics

- mean [R] mean — Estimate means
- proportion [R] proportion — Estimate proportions
- ratio [R] ratio — Estimate ratios
- total [R] total — Estimate totals

Linear regression models

- cnsreg [R] cnsreg — Constrained linear regression
- etregress [TE] etregress — Linear regression with endogenous treatment effects
- glm [R] glm — Generalized linear models
- intreg [R] intreg — Interval regression
- nl [R] nl — Nonlinear least-squares estimation
- regress [R] regress — Linear regression
- tobit [R] tobit — Tobit regression
- truncreg [R] truncreg — Truncated regression

Structural equation models

- sem [SEM] sem — Structural equation model estimation command

Survival-data regression models

- stcox [ST] stcox — Cox proportional hazards model
- streg [ST] streg — Parametric survival models
Binary-response regression models

- `biprobit` [R] `biprobit` — Bivariate probit regression
- `cloglog` [R] `cloglog` — Complementary log-log regression
- `hetprobit` [R] `hetprobit` — Heteroskedastic probit model
- `logistic` [R] `logistic` — Logistic regression, reporting odds ratios
- `logit` [R] `logit` — Logistic regression, reporting coefficients
- `probit` [R] `probit` — Probit regression
- `scobit` [R] `scobit` — Skewed logistic regression

Discrete-response regression models

- `clogit` [R] `clogit` — Conditional (fixed-effects) logistic regression
- `mlogit` [R] `mlogit` — Multinomial (polytomous) logistic regression
- `mprobit` [R] `mprobit` — Multinomial probit regression
- `ologit` [R] `ologit` — Ordered logistic regression
- `oprobit` [R] `oprobit` — Ordered probit regression
- `slogit` [R] `slogit` — Stereotype logistic regression

Poisson regression models

- `gnbreg` Generalized negative binomial regression in [R] `nbreg`
- `nbreg` [R] `nbreg` — Negative binomial regression
- `poisson` [R] `poisson` — Poisson regression
- `tnbreg` [R] `tnbreg` — Truncated negative binomial regression
- `tpoisson` [R] `tpoisson` — Truncated Poisson regression
- `zinb` [R] `zinb` — Zero-inflated negative binomial regression
- `zip` [R] `zip` — Zero-inflated Poisson regression

Instrumental-variables regression models

- `ivprobit` [R] `ivprobit` — Probit model with continuous endogenous regressors
- `ivregress` [R] `ivregress` — Single-equation instrumental-variables regression
- `ivtobit` [R] `ivtobit` — Tobit model with continuous endogenous regressors

Regression models with selection

- `heckman` [R] `heckman` — Heckman selection model
- `heckoprobit` [R] `heckoprobit` — Ordered probit model with sample selection
- `heckprobit` [R] `heckprobit` — Probit model with sample selection

Menu

Statistics > Survey data analysis > ...

Dialog boxes for all statistical estimators that support `svy` can be found on the above menu path. In addition, you can access survey data estimation from standard dialog boxes on the `SE/Robust` or `SE/Cluster` tab.
Remarks and examples

Remarks are presented under the following headings:

- Overview of survey analysis in Stata
- Descriptive statistics
- Regression models
- Health surveys

Overview of survey analysis in Stata

Many Stata commands estimate the parameters of a process or population by using sample data. For example, `mean` estimates means, `ratio` estimates ratios, `regress` fits linear regression models, `poisson` fits Poisson regression models, and `logistic` fits logistic regression models. Some of these estimation commands support the `svy` prefix, that is, they may be prefixed by `svy:` 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 not prefixed 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 `pweight`s and the `vce(cluster clustvar)` option, where `clustvar` corresponds to the PSU variable that you `svyset`. If your survey data exhibit only sampling weights or first-stage clusters (or both), you can get by with using the standard command with `pweight`s, `vce(cluster clustvar)`, or both. 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. When used with independently distributed, nonweighted data, the likelihood to be maximized reflects 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, because survey data are weighted, not independently distributed, or both. Yet for survey data, (valid) parameter estimates for a given model can be obtained using the associated likelihood function with appropriate weighting. Because the probabilistic interpretation no longer holds, the likelihood here is instead called a pseudolikelihood, but likelihood-ratio tests are no longer valid. See Skinner (1989, sec. 3.4.4) for a discussion of maximum pseudolikelihood estimators.

Here we highlight the other features of `svy` commands:

- `svy` commands handle stratified sampling, but none of the standard commands do. Because 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`.

- `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 yields a nontrivial difference only when there is a small number of clusters ($< 50$). If a regression model command uses $t$ statistics and the `vce(cluster clustvar)` option is specified, then the degrees of freedom used is the same as that of the `svy` command (in the absence of stratification).
• \texttt{svy} commands produce an adjusted Wald test for the model test, and \texttt{test} can be used to produce adjusted Wald tests for other hypotheses after \texttt{svy} commands. Only unadjusted Wald tests are available if the \texttt{svy} prefix is not used. The adjustment can be important when the degrees of freedom, \( n - L \), is 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 the point made in the second bullet make using the \texttt{svy} command important if the number of sampled PSU s (clusters) is small (< 50).

• \texttt{svy: regress} differs slightly from \texttt{regress} and \texttt{svy: ivregress} differs slightly from \texttt{ivregress} in that they use different multipliers for the variance estimator. \texttt{regress} and \texttt{ivregress} (when the \texttt{small} option is specified) use a multiplier of \( \left\{ \frac{(N-1)}{(N-k)} \right\} \left\{ \frac{n}{n-1} \right\} \), where \( N \) is the number of observations, \( n \) is the number of clusters (PSUs), and \( k \) is the number of regressors including the constant. \texttt{svy: regress} and \texttt{svy: ivregress} use \( n/\left( n-1 \right) \) instead. Thus they produce slightly different standard errors. The \( \left( \frac{N-1}{N-k} \right) \) is ad hoc and has no rigorous theoretical justification; hence, the purist \texttt{svy} commands do not use it. The \texttt{svy} commands tacitly assume that \( N \gg k \). If \( \left( \frac{N-1}{N-k} \right) \) is not close to 1, you may be well advised to use \texttt{regress} or \texttt{ivregress} so that some punishment is inflicted on your variance estimates. Maximum likelihood estimators in Stata (for example, \texttt{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 (for example, \texttt{svy: logit}) produce the same standard errors as the corresponding maximum likelihood commands (for example, \texttt{logit}), but \( p \)-values are slightly different because of the point made in the second bullet.

• \texttt{svy} commands can produce proper estimates for subpopulations by using the \texttt{subpop()} option. Using an \texttt{if} restriction with \texttt{svy} or standard commands can yield incorrect standard-error estimates for subpopulations. Often an \texttt{if} restriction will yield the same standard error as \texttt{subpop()}; most other times, the two standard errors will be slightly different; but sometimes—usually for thinly sampled subpopulations—the standard errors can be appreciably different. Hence, the \texttt{svy} command with the \texttt{subpop()} option should be used to obtain estimates for thinly sampled subpopulations. See [SVY] \texttt{subpopulation estimation} for more information.

• \texttt{svy} commands handle zero sampling weights properly. Standard commands ignore any observation with a weight of zero. Usually, this will yield 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, \texttt{svy} and standard commands will produce different standard errors, but usually this difference is very small.

• You can \texttt{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 \texttt{svyset iweights} instead of \texttt{pweights}; no standard command will allow negative sampling weights.

• The \texttt{svy} commands compute finite population corrections (FPCs).

• After a \texttt{svy} command, \texttt{estat effects} will compute the design effects DEFF and DEFT and the misspecification effects MEFF and MEFT.

• \texttt{svy} commands can perform variance estimation that accounts for multiple stages of clustered sampling.

• \texttt{svy} commands can perform variance estimation that accounts for poststratification adjustments to the sampling weights.

• Some standard options are not allowed with the \texttt{svy} prefix. For example, \texttt{vce()} and weights cannot be specified when using the \texttt{svy} prefix because \texttt{svy} is already using the variance estimation and sampling weights identified by \texttt{svyset}. Some options are not allowed with
survey data because they would be statistically invalid, such as noskip for producing optional likelihood-ratio tests. Other options are not allowed because they change how estimation results are reported (for example, nodisplay, first, plus) or are not compatible with svy’s variance estimation methods (for example, irls, mse1, hc2, hc3).

- Estimation results are presented in the standard way, except that svy has its own table header: In addition to the sample size, model test, and $R^2$ (if present in the output from the standard command), svy will also report the following information in the header:
  a. number of strata and PSUs
  b. number of poststrata, if specified to svyset
  c. population size estimate
  d. subpopulation sizes, if the subpop() option was specified
  e. design degrees of freedom

Descriptive statistics

Use svy: mean, svy: ratio, svy: proportion, and svy: total to estimate finite population and subpopulation means, ratios, proportions, and totals, respectively. You can also estimate standardized means, ratios, and proportions for survey data; see [SVY] direct standardization. Estimates for multiple subpopulations can be obtained using the over() option; see [SVY] subpopulation estimation.

Example 1

Suppose that we need to estimate the average birthweight for the population represented by the National Maternal and Infant Health Survey (NMIHS) (Gonzalez, Krauss, and Scott 1992).

First, we gather the survey design information.

- Primary sampling units are mothers; that is, PSUs are individual observations—there is no separate PSU variable.
- The finalwgt variable contains the sampling weights.
- The stratan variable identifies strata.
- There is no variable for the finite population correction.

Then we use svyset to identify the variables for sampling weights and stratification.

```
   . use http://www.stata-press.com/data/r13/nmihs
   . svyset [pweight=finwgt], strata(stratan)
      pweight: finwgt
      VCE: linearized
      Single unit: missing
      Strata 1: stratan
      SU 1: <observations>
      FPC 1: <zero>
```

Now we can use svy: mean to estimate the average birthweight for our population.
. svy: mean birthwgt
(running mean on estimation sample)
Survey: Mean estimation
Number of strata = 6  Number of obs = 9946
Number of PSUs = 9946  Population size = 3895562
Design df = 9940

<table>
<thead>
<tr>
<th></th>
<th>Linearized</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Std. Err.</td>
<td>[95% Conf. Interval]</td>
</tr>
<tr>
<td>birthwgt</td>
<td>3355.452</td>
<td>6.402741</td>
<td>3342.902</td>
</tr>
</tbody>
</table>

From these results, we are 95% confident that the mean birthweight for our population is between 3,343 and 3,368 grams.

Regression models

As exhibited in the table at the beginning of this manual entry, many of Stata’s regression model commands support the svy prefix. If you know how to use one of these commands with standard data, then you can also use the corresponding svy command with your survey data.

Example 2

Let’s model the incidence of high blood pressure with a dataset from the Second National Health and Nutrition Examination Survey (NHANES II) (McDowell et al. 1981). The survey design characteristics are already svyset, so we will just replay them.

```
. use http://www.stata-press.com/data/r13/nhanes2d
. svyset

   pweight: finalwgt
   VCE: linearized
   Single unit: missing
   Strata 1: strata
   SU 1: psu
   FPC 1: <zero>
```

Now we can use svy: logistic to model the incidence of high blood pressure as a function of height, weight, age, and sex (using the female indicator variable).
The odds ratio for the `female` predictor is 0.63 (rounded to two decimal places) and is significantly less than 1. This finding implies that females have a lower incidence of high blood pressure than do males.

Here we use the `subpop()` option to model the incidence of high blood pressure in the subpopulation identified by the `female` variable.

Because the odds ratio for the `age` predictor is significantly greater than 1, we can conclude that older females are more likely to have high blood pressure than are younger females.

**Health surveys**

There are many sources of bias when modeling the association between a disease and its risk factors (Korn, Graubard, and Midhune 1997; Korn and Graubard 1999, sec. 3.7). In cross-sectional health surveys, inference is typically restricted to the target population as it stood when the data were collected. This type of survey cannot capture the fact that participants may change their habits over time. Some health surveys collect data retrospectively, relying on the participants to recall the status of risk factors as they stood in the past. This type of survey is vulnerable to recall bias.
Longitudinal surveys collect data over time, monitoring the survey participants over several years. Although the above biases are minimized, analysts are still faced with some tough choices/situations when modeling time-to-event data. For example:

- **Time scale.** When studying cancer, should we measure the time scale by using the participant’s age or the initial date from which data were collected?
- **Time-varying covariates.** Were all relevant risk factors sampled over time, or do we have only the baseline measurement?
- **Competing risks.** When studying mortality, do we have the data specific to cause of death?


### Example 3: Cox’s proportional hazards model

Suppose that we want to model the incidence of lung cancer by using three risk factors: smoking status, sex, and place of residence. Our dataset comes from a longitudinal health survey: the First National Health and Nutrition Examination Survey (NHANES I) (Miller 1973; Engel et al. 1978) and its 1992 Epidemiologic Follow-up Study (NHEFS) (Cox et al. 1997); see the National Center for Health Statistics website at http://www.cdc.gov/nchs/. We will be using data from the samples identified by NHANES I examination locations 1–65 and 66–100; thus we will `svyset` the revised pseudo-PSU and strata variables associated with these locations. Similarly, our `pweight` variable was generated using the sampling weights for the nutrition and detailed samples for locations 1–65 and the weights for the detailed sample for locations 66–100.

```
. use http://www.stata-press.com/data/r13/nhefs
. svyset psu2 [pw=swgt2], strata(strata2)
pweight: swgt2
VCE: linearized
Single unit: missing
Strata 1: strata2
SU 1: psu2
FPC 1: <zero>
```

The lung cancer information was taken from the 1992 NHEFS interview data. We use the participants’ ages for the time scale. Participants who never had lung cancer and were alive for the 1992 interview were considered censored. Participants who never had lung cancer and died before the 1992 interview were also considered censored at their age of death.
Although `stset` warns us that it is a “probable error” to have 5,126 observations with missing event times, we can verify from the 1992 NHEFS documentation that there were indeed 9,281 participants with complete information.

For our proportional hazards model, we pulled the risk factor information from the NHANES I and 1992 NHEFS datasets. Smoking status was taken from the 1992 NHEFS interview data, but we filled in all but 132 missing values by using the general medical history supplement data in NHANES I. Smoking status is represented by separate indicator variables for former smokers and current smokers; the base comparison group is nonsmokers. Sex was determined using the 1992 NHEFS vitality data and is represented by an indicator variable for males. Place-of-residence information was taken from the medical history questionnaire in NHANES I and is represented by separate indicator variables for rural and heavily populated (more than 1 million people) urban residences; the base comparison group is urban residences with populations of fewer than 1 million people.

```
. svy: stcox former_smoker smoker male urban1 rural
  (running stcox on estimation sample)

Survey: Cox regression
Number of strata = 35 Number of obs = 9149
Number of PSUs = 105 Population size = 151327827
Design df = 70
F(5, 66) = 14.07 Prob > F = 0.0000

_linearized_ Haz. Ratio Std. Err. t P>|t| [95% Conf. Interval]
former_smoker 2.788113 .6205102 4.61 0.000 1.788705 4.345923
smoker 7.849483 2.593249 6.24 0.000 4.061457 15.17051
male 1.187611 .3445315 0.59 0.555 .6658757 2.118142
urban1 .8035074 .3285144 -0.54 0.594 .3555123 1.816039
rural 1.581674 .5281859 1.37 0.174 .8125799 3.078702
```

From the above results, we can see that both former and current smokers have a significantly higher risk for developing lung cancer than do nonsmokers.
Technical note

In the previous example, we specified a sampling weight variable in the calls to both `svyset` and `stset`. When the `svy` prefix is used with `stcox` and `streg`, it identifies the sampling weight variable by using the data characteristics from both `svyset` and `stset`. `svy` will report an error if the `svyset pweight` variable is different from the `stset pweight` variable. The `svy` prefix will use the specified `pweight` variable, even if it is `svyset` but not `stset`. If a `pweight` variable is `stset` but not `svyset`, `svy` will note that it will be using the `stset pweight` variable and then `svyset` it.

The standard `st` commands will not use the `svyset pweight` variable if it is not also `stset`.

Example 4: Multiple baseline hazards

We can assess the proportional-hazards assumption across the observed race categories for the model fit in the previous example. The race information in our 1992 NHEFS dataset is contained in the `revised_race` variable. We will use `stphplot` to produce a log-log plot for each category of `revised_race`. As described in [ST] `stcox PH-assumption tests`, if the plotted lines are reasonably parallel, the proportional-hazards assumption has not been violated. We will use the `zero` option to reset the risk factors to their base comparison group.

```
. stphplot, strata(revised_race) adjust(former_smoker smoker male urban1 rural) > zero legend(col(1))
  failure _d: lung_cancer
  analysis time _t: age_lung_cancer
  weight: [pweight=swgt2]
```

As we can see from the graph produced above, the lines for the black and white race categories intersect. This indicates a violation of the proportional-hazards assumption, so we should consider using separate baseline hazard functions for each race category in our model fit. We do this next, by specifying `strata(revised_race)` in our call to `svy: stcox`. 

![Graph showing multiple baseline hazards](image-url)
```stata
.svy: stcox former_smoker smoker male urban1 rural, strata(revised_race)
(running stcox on estimation sample)
Survey: Cox regression
Number of strata  = 35  Number of obs  = 9149
Number of PSUs   = 105  Population size = 151327827
Design df        = 70  F( 5, 66) = 13.95  Prob > F = 0.0000

Linearized
|   | Haz. Ratio | Std. Err. | t    | P>|t|       | [95% Conf. Interval] |
|---|------------|-----------|------|-----------|----------------------|
|former_smoker| 2.801797   | .6280352  | 4.60 | 0.000     | 1.791761 4.381201    |
smoker       | 7.954921   | 2.640022  | 6.25 | 0.000     | 4.103709 15.42038    |
|male         | 1.165724   | .3390339  | 0.53 | 0.600     | .6526527 2.082139    |
|urban1       | .784031    | .3120525  | -0.61| 0.543     | .3544764 1.73412     |
rural        | 1.490269   | .5048569  | 1.18 | 0.243     | .7582848 2.928851    |
```

Stratified by revised_race

References


Also see

[SVY] svy postestimation — Postestimation tools for svy
[SVY] estat — Postestimation statistics for survey data
[SVY] direct standardization — Direct standardization of means, proportions, and ratios
[SVY] poststratification — Poststratification for survey data
[SVY] subpopulation estimation — Subpopulation estimation for survey data
[SVY] variance estimation — Variance estimation for survey data

[U] 20 Estimation and postestimation commands
[SVY] svyset — Declare survey design for dataset
[SVY] svy — The survey prefix command