

Survey Data Analysis with Stata

Jeff Pitblado

Associate Director, Statistical Software

StataCorp LP

JSM 2011



Outline

- 1 Types of data
- 2 Survey data characteristics
- 3 Variance estimation
- 4 Estimation for subpopulations
- 5 Postestimation
- 6 Summary



Why survey data?

- Collecting data can be expensive and time consuming.
- Consider how you would collect the following data:
 - Smoking habits of teenagers
 - Birthweights for expectant mothers with high blood pressure
- Using stages of clustered sampling can help cut down on expense and time.



Types of data

Simple random sample (SRS) data

Observations are independently sampled from a data-generating process.

- Typical assumption: data are independent and identically distributed (iid).
- Make inferences about the data generating process.
- Sample variability is explained by the statistical model attributed to the data generating process.

Standard data

We will use this term to distinguish this type of data from survey data.



Types of data

Correlated data

Individuals are assumed not to be independent.

Causes:

- Observations are taken over time
- Random-effects assumptions
- Cluster sampling

Treatment:

- Time-series models
- Longitudinal/panel data models
- **vce(cluster ...)** option



Survey data

Individuals are sampled from a fixed population according to a survey design.

Distinguishing characteristics of this type of data include

- the complex nature under which individuals are sampled,
- inferences made about the fixed population, and
- sample variability attributed to the survey design.



Survey data characteristics

Standard data

- Estimation commands for standard data include
 - **proportion**, and
 - **regress**.
- We will refer to these as *standard estimation commands*.

Survey data

- Estimation commands for survey data are governed by the **svy** prefix, for example
 - **svy: proportion**, and
 - **svy: regress**.
- **svy** requires that the data be **svyset**.



Survey data characteristics

Example: `proportion` and `svy: proportion`



Single-stage syntax

```
svyset [psu] [weight] [, strata(varname) fpc(varname) ]
```

Syntax elements:

- Primary sampling units (PSU)
- Sampling weights – **pweight**
- Strata
- Finite population correction (FPC)



Sampling unit

An individual or collection of individuals from the population that can be selected for observation.

- Sampling groups of individuals is synonymous with cluster sampling.
- Cluster sampling usually results in inflated variance estimates compared to SRS.



Sampling weight

The reciprocal of the probability that an individual to be sampled.

- Probabilities are derived from the survey design.
 - sampling units
 - strata
- Typically considered to be the number of individuals in the population that a sampled individual represents.
- Reduces bias induced by the sampling design.



Strata

In stratified designs, the population is partitioned into well-defined groups called strata.

- Sampling units are independently sampled from within each stratum.
- Stratification usually results in smaller variance estimates compared with SRS.



Finite population correction (FPC)

An adjustment applied to the variance due to sampling without replacement.

- Sampling without replacement from a finite population reduces sampling variability.



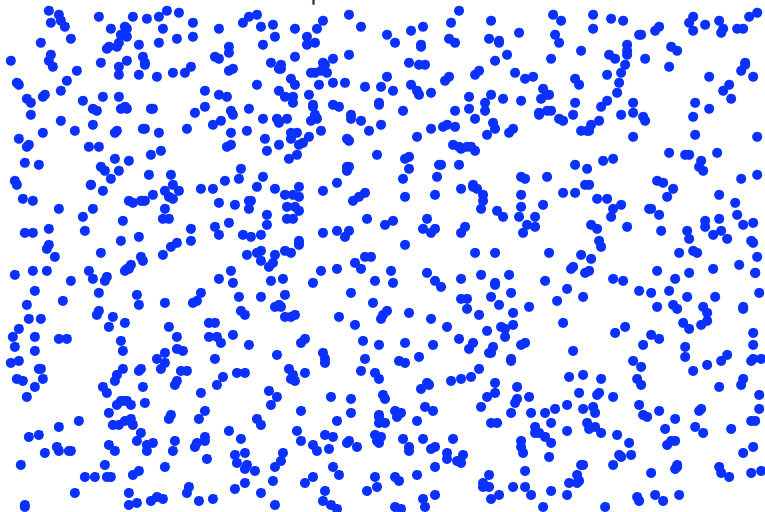
Survey data characteristics

Example: **svyset** for single-stage designs



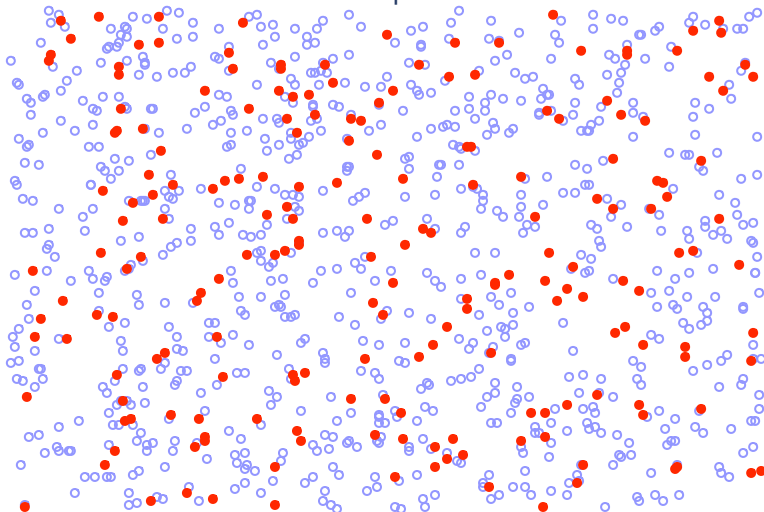
Survey data characteristics

Population 1000



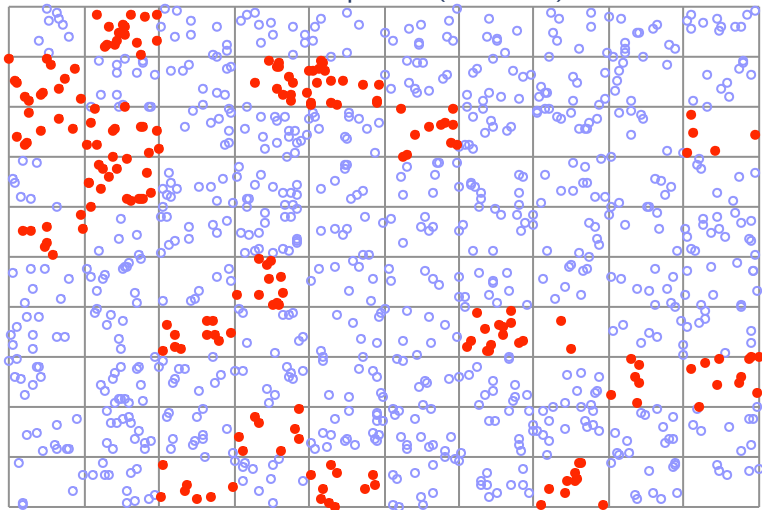
Survey data characteristics

SRS sample 200



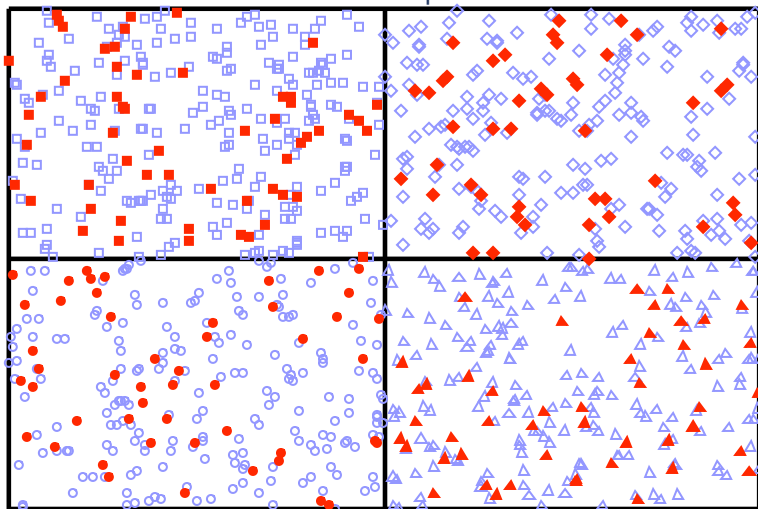
Survey data characteristics

Cluster sample 20 (208 obs)



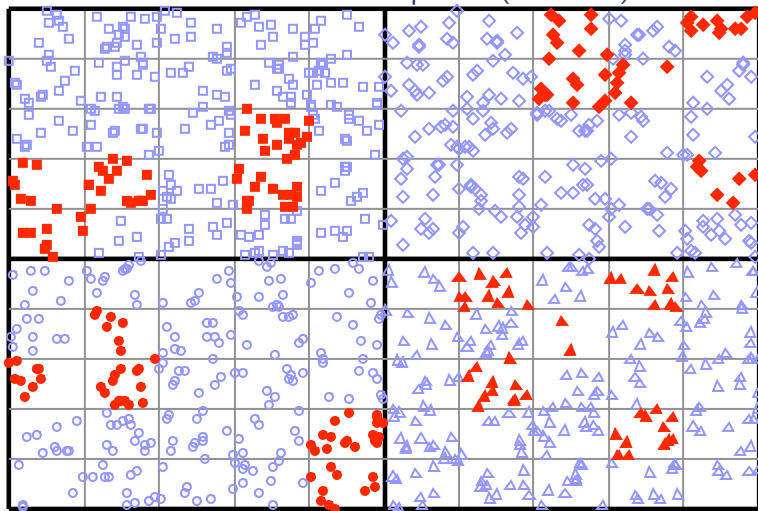
Survey data characteristics

Stratified sample 198



Survey data characteristics

Stratified-cluster sample 20 (215 obs)



Survey data characteristics

Purpose

Study the smoking habits of teenagers in the U.S.

Survey design

- 1 Use state for strata, and counties are the PSUs.
- 2 The second-stage units are high schools, randomly selected within each sampled county.
- 3 Stratifying on gender, the final-stage units are high school seniors, randomly selected within each sampled high school.



Multistage syntax

```
svyset psu [weight] [, strata(varname) fpc(varname) ]  
    [|| ssu [, strata(varname) fpc(varname) ]]  
    [|| ssu [, strata(varname) fpc(varname) ]] ...
```

- Stages are delimited by “||”.
- SSU stands for secondary or subsequent sampling units.
- FPC is required at stage s for stage $s + 1$ to play a role in the linearized variance estimator.



Multiple stages of cluster sampling

- ❶ PSUs are independently selected within each stratum.
 - ❷ Second-stage units are independently selected within each sampled PSU.
 - ❸ Third-stage units are independently selected within each sampled second-stage unit.
- Sampling units are independently selected within each sampled SSU.
 - Stratification is also allowed at each sampling stage.



High school senior data

- 1 Counties are randomly selected within each state.
- 2 High schools are randomly selected within each sampled county.
- 3 Female and male seniors are randomly selected within each sampled high school.



Survey data characteristics

Example: **svyset** for a multistage design



FPC variables

- `ncounties` is the number of counties within each category of state.
- `nschools` is the number of high schools within state county.
- `nseniors` is the number of high school seniors within state county school sex.



Poststratification

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

- adjusts weights to sum to the poststratum sizes in the population,
- reduces bias due to nonresponse and underrepresented groups, and
- can result in smaller variance estimates.

Syntax

```
svyset ... poststrata(varname) postweight(varname)
```



Survey data characteristics

Example: **svyset** for poststratification



Strata with a single sampling unit

Big problem for variance estimation

- Consider a sample with only one observation.
- **svy** reports missing standard-error estimates by default.

Finding these lonely sampling units

Use **svydescribe**

- to describe the strata and sampling units and
- to help find strata with a single sampling unit.



Strata with a single sampling unit

Example: **svydescribe**



Handling lonely sampling units

- Drop them from the estimation sample.
- **svyset** one of the ad hoc adjustments in the **singleunit()** option.
- Somehow combine them with other strata.



- Sampling units that are guaranteed to be chosen by the design.
- Certainty units are handled by treating each one as its own stratum with an FPC of 1.



Stata has five variance estimation methods for survey data:

- linearization
- balanced repeated replication (BRR)
- jackknife
- bootstrap
- successive difference replication (SDR)



Variance estimation

Linearization

A method for deriving a variance estimator using a first order Taylor approximation of the point estimator of interest.

- Foundation: Variance of the total estimator

Syntax

```
svyset ... [vce(linearized) ]
```

- Delta method
- Huber/White/robust/sandwich estimator



Total estimator—Stratified two-stage design

- y_{hijk} is the observed value from a sampled individual.
- strata: $h = 1, \dots, L$
- PSU: $i = 1, \dots, n_h$
- SSU: $j = 1, \dots, m_{hi}$
- individual: $k = 1, \dots, m_{hij}$

$$\hat{Y} = \sum w_{hijk} y_{hijk}$$
$$\hat{V}(\hat{Y}) = \sum_h (1 - f_h) \frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2 +$$
$$\sum_h f_h \sum_i (1 - f_{hi}) \frac{m_{hi}}{m_{hi} - 1} \sum_j (y_{hij} - \bar{y}_{hi})^2$$



Total estimator—Stratified two-stage design

- y_{hijk} is the observed value from a sampled individual.
- strata: $h = 1, \dots, L$
- PSU: $i = 1, \dots, n_h$
- SSU: $j = 1, \dots, m_{hi}$
- individual: $k = 1, \dots, m_{hij}$

$$\begin{aligned}\hat{Y} &= \sum w_{hijk} y_{hijk} \\ \hat{V}(\hat{Y}) &= \sum_h (1 - f_h) \frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2 + \\ &\quad \sum_h f_h \sum_i (1 - f_{hi}) \frac{m_{hi}}{m_{hi} - 1} \sum_j (y_{hij} - \bar{y}_{hi})^2\end{aligned}$$



Total estimator—Stratified two-stage design

- y_{hijk} is the observed value from a sampled individual.
- strata: $h = 1, \dots, L$
- PSU: $i = 1, \dots, n_h$
- SSU: $j = 1, \dots, m_{hi}$
- individual: $k = 1, \dots, m_{hij}$

$$\begin{aligned}\hat{Y} &= \sum w_{hijk} y_{hijk} \\ \hat{V}(\hat{Y}) &= \sum_h (1 - f_h) \frac{n_h}{n_h - 1} \sum_i (y_{hi} - \bar{y}_h)^2 + \\ &\quad \sum_h f_h \sum_i (1 - f_{hi}) \frac{m_{hi}}{m_{hi} - 1} \sum_j (y_{hij} - \bar{y}_{hi})^2\end{aligned}$$



Variance estimation

Example: `svy: total`

Linearized variance for regression models

- The model is fit using estimating equations.
- $\hat{G}()$ is a total estimator; use Taylor expansion to get $\hat{V}(\hat{\beta})$:

$$\hat{G}(\beta) = \sum_j w_j s_j \mathbf{x}_j = \mathbf{0}$$

$$\hat{V}(\hat{\beta}) = D \hat{V}^{-1} \{ \hat{G}(\beta) \} |_{\beta=\hat{\beta}} D'$$



Linearized variance for regression models

- The model is fit using estimating equations.
- $\hat{G}()$ is a total estimator; use Taylor expansion to get $\hat{V}(\hat{\beta})$:

$$\hat{G}(\beta) = \sum_j w_j s_j \mathbf{x}_j = \mathbf{0}$$

$$\hat{V}(\hat{\beta}) = D\hat{V}^{-1}\{\hat{G}(\beta)\}|_{\beta=\hat{\beta}}D'$$



Variance estimation

Example: `svy: logit`

Variance estimation

Balanced repeated replication

For designs with two PSUs in each of L strata.

- Compute replicates by dropping a PSU from each stratum.
- Find a balanced subset of the 2^L replicates. $L \leq r < L + 4$
- The replicates are used to estimate the variance.

Syntax

```
svyset ... vce(brr) [mse]
```



BRR replicate weights

- w_j is the sampling weight for individual j in the first PSU of stratum h .
- H_r is a Hadamard matrix for r replications; $H_r' H_r = rI$.
- Fay's adjustment is f ; $f = 0$ by default.

The adjusted sampling weight for the i th replicate is

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



BRR variance formulas

- $\hat{\theta}$ denotes the point estimates.
- $\hat{\theta}_{(i)}$ denotes the i th replicate of the point estimates.
- $\bar{\theta}_{(.)}$ denotes the average of the replicates.

Default variance formula:

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

Mean squared error (MSE) formula:

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



Variance estimation

Example: `svy brr: logit`

The jackknife

A replication method for variance estimation. It is not restricted to a specific survey design.

- Delete-1 jackknife: drop 1 PSU
- Delete- k jackknife: drop k PSUs within a stratum

Syntax

```
svyset ... vce(jackknife) [mse]
```



Delete-1 jackknife replicate weights

- w_{hij} is the sampling weight for individual j in PSU i of stratum h .
- Drop PSU i^* from stratum h^* .
- There are n_{h^*} replicates from stratum h^* .

The adjusted sampling weight is

$$w_{hij}^* = \begin{cases} 0 & , \text{ if } h = h^* \text{ and } i = i^* \\ \frac{n_h}{n_h - 1} w_{hij} & , \text{ if } h = h^* \text{ and } i \neq i^* \\ w_{hij} & , \text{ otherwise} \end{cases}$$



Delete- k jackknife replicate weights

- w_{hij} is the sampling weight for individual j in PSU i of stratum h .
- Drop k PSUs from stratum h^* .
- There are $c_{h^*} = \binom{n_{h^*}}{k}$ replicates from stratum h^* .

The adjusted sampling weight is

$$w_{hij}^* = \begin{cases} 0 & , \text{ if } h = h^* \text{ and } i \text{ is dropped} \\ \frac{n_h}{n_h - k} w_{hij} & , \text{ if } h = h^* \text{ and } i \text{ is not dropped} \\ w_{hij} & , \text{ otherwise} \end{cases}$$



Variance estimation

Jackknife variance formulas

- $\hat{\theta}_{(h,i)}$ is a replicate of the point estimates from stratum h , PSU i .
- $\bar{\theta}_h$ is the average of the replicates from stratum h .
- $m_h = (n_h - 1)/n_h$ is the delete-1 multiplier for stratum h .
- $m_h = (n_h - k)/c_h k$ is the delete- k multiplier for stratum h .

Default variance formula:

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

Mean squared error (MSE) formula:

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



Example: `svy jackknife: logit`

The bootstrap

Even less restrictive on the design and parameters than the delete-1 jackknife.

- Resample the observed data by adjusting the sampling weights.
- **svy bootstrap**: requires replicate-weight variables.

Syntax

```
svyset ... vce(bootstrap) bsrweight(varlist)  
          [bsn(#) mse]
```



Variance estimation

Bootstrap variance formulas

- $\hat{\theta}$ denotes point estimates.
- $\hat{\theta}_{(i)}$ is the i th replicate of the point estimates.
- $\bar{\theta}_{(.)}$ is the average of the replicates.
- b is the number of bootstrap samples used to generate each replicate weight variable, default is `bsn(1)`.

Default variance formula:

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

Mean squared error (MSE) formula:

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



Successive difference replication (SDR)

Replication method designed for systematic samples whose sampling units are ordered.

- Resample the observed data by adjusting the sampling weights.
- **svy sdr**: requires replicate-weight variables.

Syntax

```
svyset ... vce(sdr) sdrweight(varlist) [mse]
```



Variance estimation

SDR variance formulas

- $\hat{\theta}$ denotes the point estimates.
- $\hat{\theta}_{(i)}$ is the i th replicate of the point estimates.
- $\bar{\theta}_{(.)}$ is the average of the replicates.
- f is the sampling fraction from the `frpc()` option.

Default variance formula:

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

Mean squared error (MSE) formula:

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



Variance estimation

Replicate weight variable

A variable in the dataset that contains sampling weight values that were adjusted for resampling the data.

- Typically used to protect the privacy of the survey participants.
- Eliminates the need to **svyset** the strata and PSU variables.

Syntax

```
svyset ... brrweight(varlist) [fay(#)]  
svyset ... jkrweight(varlist [, ... multiplier(#)])  
svyset ... bsrweight(varlist) [bsn(#)]  
svyset ... sdrweight(varlist)
```



Focus on a subset of the population

- Subpopulation variance estimation
 - assumes the same survey design for subsequent data collection, and
 - requires the **subpop()** option.
- Restricted-sample variance estimation
 - assumes the identified subset for subsequent data collection,
 - ignores the fact that the sample size is a random quantity, and
 - requires using the **if** or **in** restrictions.



Estimation for subpopulations

Total from SRS data

- Data are y_1, \dots, y_n and S is the subset of observations.

$$\delta_j(S) = \begin{cases} 1, & \text{if } j \in S \\ 0, & \text{otherwise} \end{cases}$$

- Subpopulation (or restricted-sample) total:

$$\hat{Y}_S = \sum_{j=1}^n \delta_j(S) w_j y_j$$

- Sampling weight and subpopulation size:

$$w_j = \frac{N}{n}, \quad N_S = \sum_{j=1}^n \delta_j(S) w_j = \frac{N}{n} n_S$$



Estimation for subpopulations

Variance of a subpopulation total

Sample n without replacement from a population comprised of the N_S subpopulation values with $N - N_S$ additional zeroes.

$$\hat{V}(\hat{Y}_S) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \sum_{j=1}^n \left\{ \delta_j(S) w_j y_j - \frac{1}{n} \hat{Y}_S \right\}^2$$

Variance of a restricted-sample total

Sample n_S without replacement from the subpopulation of N_S values.

$$\tilde{V}(\hat{Y}_S) = \left(1 - \frac{n_S}{\hat{N}_S}\right) \frac{n_S}{n_S-1} \sum_{j=1}^n \delta_j(S) \left\{ y_j - \frac{1}{n_S} \hat{Y}_S \right\}^2$$



Estimation for subpopulations

Example: `svy, subpop()`



Working with estimation results

Most standard postestimation commands support **svy** results:

- **estat**
- **estimates**
- **lincom, nlcom**
- **predict, predictnl**
- **test, testnl**
- **margins, marginsplot, contrast, pwcompare**



Survey-specific features in `estat`

Archer–Lemeshow goodness-of-fit

- `estat gof`

Coefficient of variation

- `estat cv`

Design and misspecification effects

- `estat effects`
- `estat lceffects`

Survey design characteristics

- `estat svyset`



Marginal effects

Predictive margins and marginal effects

- **margins**

Graph results from **margins**

- **marginsplot**

Perform ANOVA-style tests on the effects of factor variables

- **contrast**

Perform pairwise comparisons of marginal means and slopes

- **pwcompare**



Effects of the survey design


Example: Postestimation

Summary

- 1 Use **svyset** to specify the survey design for your data.
- 2 Use **svydes** to find strata with a single PSU.
- 3 Choose your variance estimation method; you can **svyset** it.
- 4 Use the **svy** prefix with estimation commands.
- 5 Use **subpop()** instead of **if** and **in**.
- 6 Most standard postestimation commands support **svy** results.



References

-  Levy, P. S., and S. Lemeshow. 1999.
Sampling of Populations. 3rd ed.
New York: Wiley.
-  StataCorp. 2011.
Survey Data Reference Manual: Release 12.
College Station, TX: StataCorp LP.