Survey Data Analysis with Stata

Jeff Pitblado

Associate Director, Statistical Software

StataCorp LP

JSM 2010

Outline

Why survey data?

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

Types of data

Simple random sample (*SRS*) data

Observations are "independently" sampled from a data generating process.

- Typical assumption: 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'll use this term to distinguish this data from survey data.

Types of data

Correlated data

Individuals are assumed not independent.

Cause:

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

Treatment:

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

Types of data

Survey data

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

Distinguishing characteristics:

- Complex nature under which individuals are sampled
- Make inferences about the fixed population
- Sample variability is attributed to the survey design

Survey data characteristics

Standard data

- Estimation commands for standard data:
  - `proportion`
  - `regress`
- We'll refer to these as *standard estimation commands*.

Survey data

- Survey estimation commands are governed by the `svy` prefix.
  - `svy: proportion`
  - `svy: regress`
- `svy` requires that the data is `svyset`.

Survey data characteristics

Example: **`proportion`** and **`svy: proportion`**

Survey data characteristics

Single-stage syntax

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

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

Survey data characteristics

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*.

Survey data characteristics

Sampling weight

The reciprocal of the probability for 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.

Survey data characteristics

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 to *SRS*.

Finite population correction (FPC)

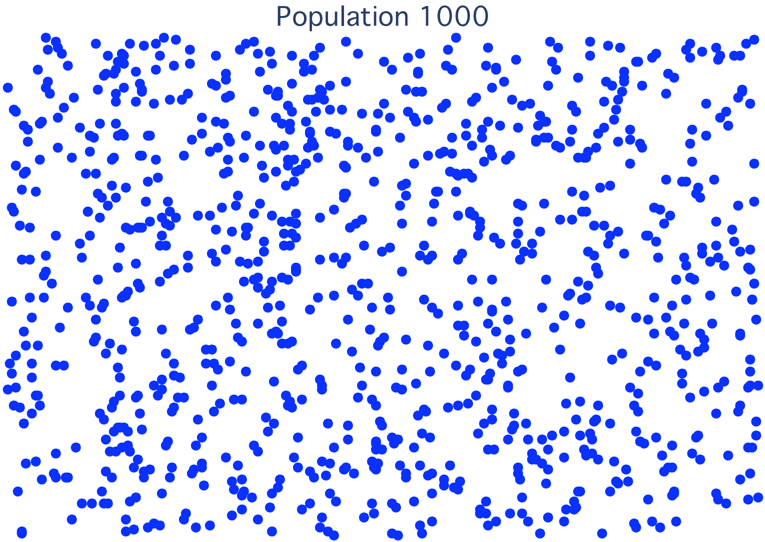
An adjustment applied to the variance due to sampling without replacement.

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

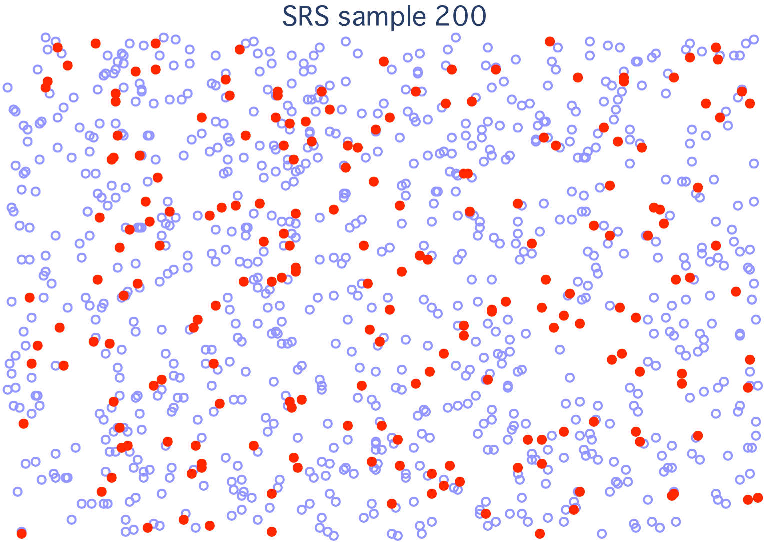
Survey data characteristics

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

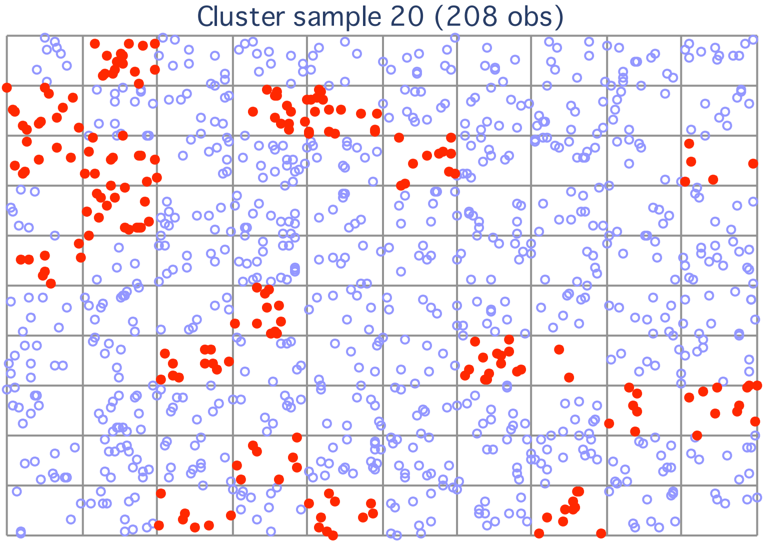
Survey data characteristics



Survey data characteristics

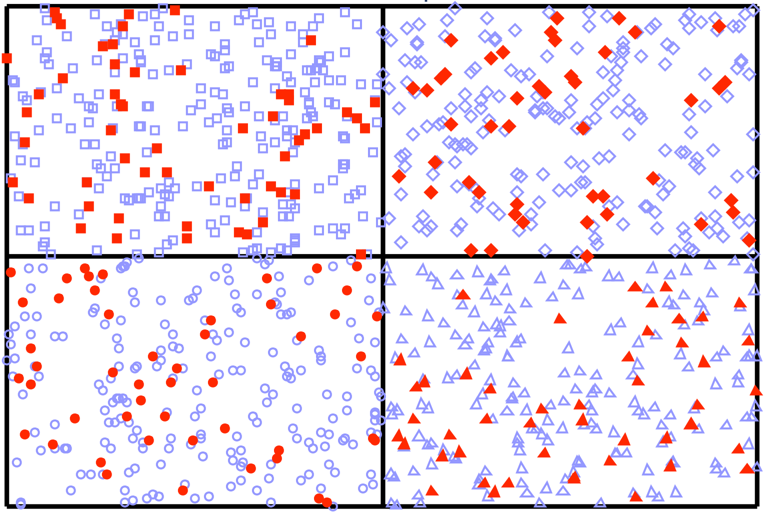


Survey data characteristics

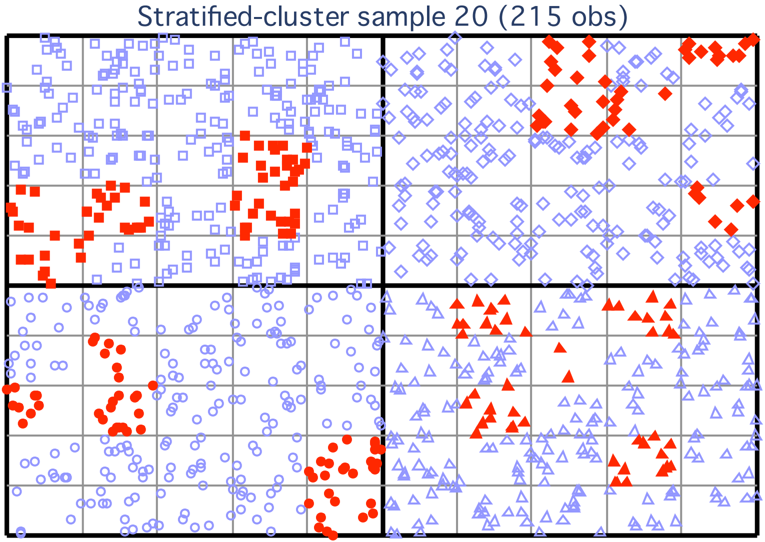


Survey data characteristics

Stratified sample 198



Survey data characteristics



Survey data characteristics

Purpose

Study the smoking habits of teenagers in the US.

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.

Survey data characteristics

Multistage syntax

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

- Stages are delimited by "||"
- SSU – secondary/subsequent sampling units
- FPC is required at stage s for stage $s+1$ to play a role in the linearized variance estimator

Survey data characteristics

Multiple stages of cluster sampling

1. PSUs are independently selected within each Stratum.
2. SSUs are independently selected within each sampled PSU.
3. . . .

- Sampling units are independently selected within each sampled SSU.
- Stratification is also allowed at each sampling stage.

Survey data characteristics

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

Survey data characteristics

FPC variables

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

Survey data characteristics

Poststratification

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

- Adjusts weights to sum to the poststratum sizes in the population
- Reduces bias due to nonresponse and underrepresented groups
- 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 1 observation
- **`svy`** reports missing standard error estimates by default

Finding these lonely sampling units

Use **`svydes`**:

- Describes the strata and sampling units
- Helps find strata with a single sampling unit

Strata with a single sampling unit

Example: **`svydes`**

Strata with a single sampling unit

Handling lonely sampling units

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

Certainty units

- 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.

Variance estimation

Stata has five variance estimation methods for survey data:

- Linearization
- Balanced repeated replication (BRR)
- The jackknife
- The bootstrap, with replicate weights
- Successive difference replication (SDR), with replicate weights

The bootstrap and SDR methods were added in the Stata 11.1 update.

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

Variance estimation

Total estimator – Stratified two-stage design

- y_{hijk} – observed value from a sampled individual
- Strata: $h = 1, \ldots, L$
- PSU: $i = 1, \ldots, n_h$
- SSU: $j = 1, \ldots, m_{hi}$
- Individual: $k = 1, \ldots, m_{hij}$

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

Variance estimation

Total estimator – Stratified two-stage design

- y_{hijk} – observed value from a sampled individual
- Strata: $h = 1, \ldots, L$
- PSU: $i = 1, \ldots, n_h$
- SSU: $j = 1, \ldots, m_{hi}$
- Individual: $k = 1, \ldots, m_{hij}$

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

Variance estimation

Total estimator – Stratified two-stage design

- y_{hijk} – observed value from a sampled individual
- Strata: $h = 1, \ldots, L$
- PSU: $i = 1, \ldots, n_h$
- SSU: $j = 1, \ldots, m_{hi}$
- Individual: $k = 1, \ldots, m_{hij}$

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

Variance estimation

Example: **`svy: total`**

Linearized variance for regression models

- Model is fit using estimating equations.
- $\widehat{G}()$ is a total estimator, use Taylor expansion to get $\widehat{V}(\widehat{\beta})$.

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

$$\widehat{V}(\widehat{\beta}) = D\widehat{V}\{\widehat{G}(\beta)\}|_{\beta=\widehat{\beta}}D'$$

Variance estimation

Linearized variance for regression models

- Model is fit using estimating equations.
- $\widehat{G}()$ is a total estimator, use Taylor expansion to get $\widehat{V}(\widehat{\beta})$.

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

$$\widehat{V}(\widehat{\beta}) = D\widehat{V}\{\widehat{G}(\beta)\}\big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} D'$$

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 – sampling weight for individual j, in the first PSU of stratum h.
- H_r is a Hadamard matrix for r replications; $H_r' H_r = r\mathrm{I}$.
- Fay's adjustment f; $f = 0$ by default.

The adjusted sampling weight for the ith replicate is

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

BRR variance formulas

- $\widehat{\theta}$ – point estimates
- $\widehat{\theta}_{(i)}$ – ith replicate of the point estimates
- $\overline{\theta}_{(.)}$ – average of the replicates

Default variance formula:

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

Mean squared error (MSE) formula:

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

Variance estimation

Example: **`svy brr: logit`**

Variance estimation

The jackknife

A replication method for variance estimation. 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]
```

Variance estimation

Delete-1 jackknife replicate weights

- w_{hij} – sampling weight for individual j in PSU i of stratum h.
- Dropping PSU i^* from stratum h^*.
- n_{h^*} replicates from stratum h^*.

The adjusted sampling weight is

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

Variance estimation

Delete-k jackknife replicate weights

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

The adjusted sampling weight is

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

Variance estimation

Jackknife variance formulas

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

Example: **`svy jackknife: logit`** #1

Variance estimation

The bootstrap

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

- Resample the observed data by adjusting the sampling weights.
- Requires replicate weight variables.

Syntax

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

Variance estimation

Bootstrap variance formulas

- $\widehat{\theta}$ – point estimates
- $\widehat{\theta}_{(i)}$ – ith replicate of the point estimates
- $\overline{\theta}_{(.)}$ – average of the replicates
- b – number of bootstrap samples used to generate each replicate weight variable, default is **`bsn(1)`**

Default variance formula:

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

Mean squared error (MSE) formula:

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

Successive difference replication – SDR

Replication method designed for systematic samples where the observed sampling units are ordered.

- Resample the observed data by adjusting the sampling weights.
- Requires replicate weight variables.

Syntax

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

SDR variance formulas

- $\widehat{\theta}$ – point estimates
- $\widehat{\theta}_{(i)}$ – ith replicate of the point estimates
- $\overline{\theta}_{(.)}$ – average of the replicates
- f – sampling fraction from **fpc()** option

Default variance formula:

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

Mean squared error (MSE) formula:

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

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.
- Eliminate 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)
```

Example: **`svy jackknife: logit`** #2

Estimation for subpopulations

Focus on a subset of the population

- Subpopulation variance estimation:
  - Assumes the same survey design for subsequent data collection.
  - 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.
  - The **`if`** and **`in`** restrictions.

Estimation for subpopulations

Total from *SRS* data

- Data is $y_1, \ldots, y_n$ 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:

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

- Sampling weight and subpopulation size:

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

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.

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

Variance of a restricted-sample total

Sample n_S without replacement from the subpopulation of N_S values.

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

Estimation for subpopulations

Example: `svy, subpop()`

New in Stata 11

- **margins** computes predictive margins and marginal effects using the current regression results.
- **estat gof** is now available after **svy: logistic**, **svy:logit**, and **svy:probit**.
- **estat cv** computes the coefficient of variation for each of your point estimates. This includes means, totals, ratios, and regression coefficients.

Effects of the survey design

Design effects

Compare the sample variability between the survey design and a hypothetical *SRS* design of the same sample size.

- $\widehat{V}_{db}$ – design based variance estimate
- $\widehat{V}_{srs}$ – simple random sample variance estimate
- $\widehat{V}_{srswr}$ – simple random sample with replacement

$$\text{DEFF} = \frac{\widehat{V}_{db}}{\widehat{V}_{srs}}, \qquad \text{DEFT} = \sqrt{\frac{\widehat{V}_{db}}{\widehat{V}_{srswr}}}$$

Effects of the survey design

Misspecification effects

Compare the design based variance estimate to the variance from a misspecified model fit (no weighting or other design characteristics).

- $\widehat{V}_{db}$ – design based variance estimate
- $\widehat{V}_{msp}$ – misspecified variance estimate

$$\text{MEFF} = \frac{\widehat{V}_{db}}{\widehat{V}_{msp}}, \qquad \text{MEFT} = \sqrt{\text{MEFF}}$$

Effects of the survey design

Example: **`estat effects`**

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. Postestimation:
   - Use **`margins`** for predictive margins and marginal effects
   - **`estat`** has several **`svy`** specific features:
     - goodness-of-fit test after logistic regression
     - coefficient of variation
     - design effects

References

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

StataCorp. 2009.
Survey Data Reference Manual: Release 11.
College Station, TX: StataCorp LP.