# Analysis of Survey Data in Stata

Jeff Pitblado

Stata User Group Meeting, 2006 Italy

# Contents



---

# Why survey data?

**note:** Collecting a simple random sample of individuals from a population that is spread out across a wide geographic area can be expensive and time consuming.

**note:** Using complex survey designs can help cut the cost and duration required to collect the data.

### 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.

---

# 1 Types of data

**note:** We typically use random number generators to help us choose a simple random sample.

**note:** Flipping an coin, tossing dice; I tend to think of these to explain the 'iid' assumption.

**Types of data**

**Simple random sample data – SRS data**
Observations are "independently" sampled from a data generating process.

- Typical assumption: iid – independent and identically distributed
- 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.

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

Treatment:

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

---

### Types of data

### Survey data

Individuals are sampled from a fixed population according to a survey design. Distinguishes itself from other forms of data:

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

---

# 2 Survey data characteristics

### 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 dataset is **`svyset`**.

---

**note:** **`proportion`** is using a stub with an index to identify the age groups because the labels on **`agegrp`** are not valid Stata names.

**note:** **`svy: proportion`** also reports the estimated population size, and the design degrees of freedom.

**note:** We show the number of Strata and PSUs, even for multistage designs, to show where the design degrees of freedom come from: number of PSUs - number of Strata.

**Survey data characteristics**

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

---

**note:** This syntax is slightly different from Stata 8 (and previous); *`psu`* has moved from being an option.

**Survey data characteristics**

**Single-stage syntax**

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

- PSU – primary sampling units
- pweight – sampling weights
- Strata
- FPC – finite population correction

---

**note:** Examples of sampling units:

- High schools for sampling from the population of 12th graders.
- Hospitals for sampling from the population of newborns.

**note:** I might slip and say cluster instead of sampling unit or SSU; they are synonymous.

**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*.

---

**note:** If there are 100 hospitals in our population, and we choose 5 of them, the sampling weight is 20=100/5. Thus a sampled hospital represents 20 hospitals in the population.

**note:** Sampling weights correct for over/under sampling of sections in the population. Many times this over/under sampling is on purpose.

**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.

---

**note:** Examples of strata:

- States for national surveys that require that every one of the united states is represented.
- Demographic information like age group, gender, and ethnicity.

**note:** Although, there is potential for improving efficiency by reducing sampling variability, it is usually not very practicle to stratify on demographic information.

**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*.

---

**note:** In a future slide, we will see that the FPC affects the number of components in the linearized variance estimator for multistage designs.

**Survey data characteristics**

**Finite population correction**
An adjustment applied to the variance due to sampling without replacement.

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

---

**note:** We can use `svyset` to specify an SRS design.

**note:** The National Maternal and Infant Health Survey (1988) dataset came from a stratified design.

**Survey data characteristics**
Example: `svyset` for single-stage designs

---

**note:** Let's introduce/motivate multistage designs using an example.

**Survey data characteristics**

*Example* 1. Purpose: Study the smoking habits of teenagers in the US.

Multistage 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 highschool seniors, randomly selected within each sampled high school.

---

**note:** Notice that the stage specific information is delimited by the double OR bars.

**note:** `svyset` will note when it is disregarding subsequent stages if an FPC is not specified for a given stage.

**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.

---

**note:** Recall the example motivating our discussion of multistage designs.

**Survey data characteristics**

*Example* 2. Smoking habits of teenagers in the US.

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.

Example: **`svyset`** for a multistage design

**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`

---

**note:** Recall that I said it is usually not vey practicle to stratify on demographic information such as age group, gender, and ethnicity. Their frequency distribution is usually available from census data. We can use the frequency distribution for poststratification.

**Survey data characteristics**

**Poststratification**

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

- Adjusted weights 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`*`)`

---

**note:** A veterinarian has 1300 clients, 450 cats and 850 dogs. He would like to estimate the average annual expenses of his clientele but only has enough time to gather information on 50 randomly selected clients.

**note:** This is an SRS design, the sampling weight is 26=1300/50.

**note:** As we can see, the dog clients are (on average) twice as expensive as cat clients.

**Survey data characteristics**

Example: `svyset` for poststratification

---

**Intermission**

Let's take a short break

---

# 3 Strata with a single PSU

**note:** How do we get stuck with strata that have only 1 PSU?

- A very bad design.
- Missing data can cause entire PSUs to be dropped from the analysis, possibly leaving a single PSU in the estimation sample.

**Strata with a single sampling unit**

**Problem**

- This is a big issue for variance estimation:
  - Consider a sample with only 1 observation.
  - `svy` reports missing standard errors estimates.

**Solution**

- Use `svydes`:
  - Describe the strata and sampling units.
  - Helps find strata with a single sampling unit.
- Drop the strata with a single sampling unit.

- Somehow combine the singleton units into another stratum.

---

**note:** Use **`if e(sample)`** after estimation commands to restrict **`svydes`**'s focus on the estimation sample.

**note:** Specifying variable names with **`svydes`** will result in more information about missing values.

**Strata with a single sampling unit**

Example: **`svydes`**

---

# 4 Certainty units

**note:** What are certainty PSUs? PSUs that are guaranteed to be chosen by the design.

**note:** How should we handle certainty PSUs? They can be handled by treating each of them as a stratum with an FPC of 1.

**Certainty units**

**Certainty units**

Sampling units that are guaranteed to be chosen by the design.

- Treat each certainty unit as a stratum with an FPC of 1.
- No contribution to the variance.
- Certainty PSUs are not counted in the design degrees of freedom.

---

**note:** Let's use the dataset from the previous example, and pretend that the strata with single PSUs are actually certainty PSUs.

**Certainty units**

Example: Certainty PSUs

---

# 5 Variance estimation

**note:** Why does Stata 9 now have 3 different methods for variance estimation?

- Due to privacy concerns, data providers are reluctant to release strata and PSU information in public use data.
- Some datasets now come packaged with weight variables for use with replication methods.

---

**note:** Linearization is Stata's **`robust`** method for complex data; this is the default method for variance estimation.

**note:** BRR and jackknife are replication methods; the idea is to resample the data, computing replicates from each resample, then using the replicates to estimate the variance.

**note:** Think of a replicate as a copy of the point estimates.

**Variance estimation**

Stata has three variance estimation methods for survey data:

- Linearization
- Balanced repeated replication
- The jackknife

---

**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

---

**note:** Prior to Stata 9, only the first stage information could be used to estimate the variance. Also, Stata would print a warning message about subsampling when an FPC was used.

**note:** $f_h$ is the FPC for stratum $h$ in the first stage.

**note:** $f_{hi}$ denotes an FPC for the second stage.

**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: $k = 1, \ldots, m_{hij}$

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

---

**note:** Remember that the design degrees of freedom is: number of PSUs - number of strata.

**Variance estimation**

Example: **`svy: total`**

---

**note:** ML models:

- $\widehat{G}()$ is the gradient.
- $s_j$ is an equation-level score.
- $D$ is the inverse Hessian matrix at the solution.

**note:** Least squares regression:

- $\widehat{G}()$ is the normal equations.
- $s_j$ is a residual.
- $D = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$; inverse of the weighted outer product of the predictors—including the intercept.

**Variance estimation**

**Linearized variance for regression models**

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

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

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

---

**note:** The next few examples we will fit a logistic regression, modeling the incidence of high blood pressure as a function of some demographic variables.

**note:** Here is the example using linearization.

**Variance estimation**

Example: **`svy: logit`**

---

**note:** Reminder: The idea is to resample the data, compute replicates from each resample, then use the replicates to estimate the variance.

**note:** Reminder: Think of a replicate as a copy of the point estimates.

**note:** Balance here means that stratum specific contributions to the variance cancel out. In other words, no stratum contributes more to the variance than any other.

**note:** We can find a balanced subset by employing a Hadamard matrix of order $r$; a Hadamard matrix is a type of orthogonal matrix, shown on the next slide.

### 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]
```

---

**note:** When the dataset contains replicate weight variables, you do not need to worry about Hadamard matrices.

**note:** For completeness, here is how the sampling weights are adjusted to produce BRR replicate weights.

**note:** These replicate weights are used to produce a copy of the point estimates (replicate). The replicates are then used to estimate the variance.

**note:** **`svy brr`** can employ replicate weight variables in the dataset, if you **`svyset`** them. Otherwise, **`svy brr`** will automatically adjust the sampling weights to produce the replicates; however, a Hadamard matrix must be specified.

**Variance estimation**

**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 $i$th replicate is

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

---

**note:** The default variance formula uses deviations of the replicates from their mean.

**note:** The MSE formula uses deviations of the replicates from the point estimates.

**Variance estimation**

**BRR variance formulas**

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

---

**note:** **`BRR *`** is clickable, taking you to a short help file informing you that you used the MSE formula for BRR variance estimation.

**Variance estimation**

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

---

**Intermission**

Let's take a short break

---

**note:** Reminder: The idea is to resample the data, compute replicates from each resample, then use the replicates to estimate the variance.

**note:** Reminder: Think of a replicate as a copy of the point estimates.

**Variance estimation**

**The jackknife**

A replication method for variance estimation. Not restricted to a specific survey design.

- Delete-1 jackknife: drop 1 PSU
- Delete-$d$ jackknife: drop $d$ PSUs within a stratum

**Syntax**

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

---

**note:** **`svy jackknife`** can employ replicate weight variables in the dataset, if you **`svyset`** them. Otherwise, **`svy jackknife`** will automatically adjust the sampling weights to produce the replicates using the delete-1 jackknife methodology.

**note:** Here is how the sampling weights are adjusted to produce delete-1 jackknife replicate weights.

**note:** Each PSU is represented by a corresponding replicate.

**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

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

---

**note:** The delete-$d$ jackknife is only supported if you already have the corresponding replicate weight variables for **`svyset`**.

**note:** Here is how the sampling weights are adjusted to produce delete-$d$ jackknife replicate weights.

**note:** These replicate weights are used to produce a copy of the point estimates (replicate). The replicates are then used to estimate the variance.

**Variance estimation**

**Delete-$d$ jackknife replicate weights**

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

The adjusted sampling weight is

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

---

**note:** The default variance formula uses deviations of the replicates from their mean.

**note:** The MSE formula uses deviations of the replicates from the point estimates.

**Variance estimation**

**Jackknife variance formulas**

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

---

**note:** **`Jknife *`** is clickable, taking you to a short help file informing you that you used the MSE formula for jackknife variance estimation.

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

---

**note:** The 1999-2000 NHANES data was published with jackknife replicate weight variables, but not the 2001-2002 NHANES data.

**note:** Make sure to specify the correct multiplier when you **`svyset`** your jackknife replicate weight variables.

**Variance estimation**

**Replicate weight variable**
A Stata variable that contains sampling weight values that were adjusted for resampling the data using BRR or the jackknife.

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

**Syntax**

```
svyset ... brrweight(varlist)
svyset ... jkrweight(varlist [, ... multiplier(#)])
```

---

**note:** The standard error estimates in this example should match those of the previous if we **`svyset`** our data correctly.

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

---

# 6 Estimation for subpopulations

**note:** Most of you are familiar with the **`if`** and **`in`** restrictions; however, you most likely want to use the **`subpop()`** option of **`svy`** instead of **`if`** and **`in`**.

**note:** As I mentioned earlier on, variability is governed by the survey design, so our variance estimates assume the design is fixed. The **`subpop()`** option assumes this too.

**note:** If we discourage you from using **`if`** and **`in`**, why does **`svy`** allow them?

- You might want to restrict your sample because of known defects in some of the variables.
- Researchers can use **`if`** and **`in`** to conduct simulation sudies by simulating survey samples from a population dataset without having to use **`preserve`** and **`restore`**.

**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 variable.
  - The **`if`** and **`in`** restrictions.

---

**note:** We can illustrate the difference between these estimators with an SRS design.

**Estimation for subpopulations**

Total from a simple random sample without replacement design:

- Data is $y_1, \ldots, y_n$ and $S$ is the subset of observations.

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

- Subpopulation (or restricted-sample) total:

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

- Sampling weight and subpopulation size:

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

---

**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.

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

**Variance of a restricted-sample total**

Sample $n_S$ without replacement from the subpopulation of $N_S$ values.

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

---

**note:** Suppose we want to estimate the mean birth weight for mothers with high blood pressure.

**note:** The **`highbp`** variable is an indicator for mothers with high blood pressure.

**note:** In the reported results, the subpopulation information is provided in the header.

**note:** Notice that although the restricted sample results reproduce the mean, the standard errors differ.

**Estimation for subpopulations**

Example: **`svy, subpop()`**

---

# 7 Effects of the survey design

**note:** Design effects give you a sense of how efficient your survey design is compared to the SRS design.

**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

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

---

**note:** Suppose we want to compare the mean birth weight between mothers with high blood pressure and mothers with normal blood pressure.

**note:** The labels on **highbp** cannot be used as identifiers, so I'll just define some labels that will better serve my purpose.

**note:** We'll use the **over()** option so that we can estimate the means for both subpopulations simultaneously. I now have a variance matrix that I can use to perform tests or compute linear combinations.

**note:** Use **estat effects** to display design effects for the entire set of point estimates.

**note:** Use the **meff** and **meft** options to get the misspecification effects. Note that Stata had to refit the model to get the misspecified variance. This extra model fit is only required the first time you specify **meff** or **meft**, **estat effects** posts the newly acquired information to **e()** for future reference.

**note:** Recall that we are interested in comparing the mean birth weight between mothers with high and normal blood pressure. We could use the **test** command to accomplish this, but we used **lincom** to compute the difference of the means and get a 95% CI instead.

**note:** By the way, most of the postestimation commands that work after the standard estimation commands also work after the **svy**. The available options may differ between standard and survey results.

**note:** Use **estat lceffects** to get design and misspecification effects for linear combinations.

**Effects of the survey design**

Example: **estat effects**

---

# 8 Summary

**Summary**

1. Use **`svyset`** to identify the survey design characteristics for your data.
2. Use **`svydes`** to find strata with a single sampling unit.
3. Each certainty unit deserves its own stratum with a sampling fraction of 1.
4. Choose your variance estimation method; you can even **`svyset`** it.
5. Use the **`svy`** prefix with estimation commands.
6. Use **`subpop()`** instead of **`if`** and **`in`**.
7. Use **`estat`** to get design effects.

---

**References**

# References


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

[2] StataCorp. 2005. *Survey Data Reference Manual: Release 9.* College Station, TX: StataCorp LP.