Survey Data Analysis with Stata

Jeff Pitblado
Associate Director, Statistical Software
StataCorp LP

JSM 2010

Outline

This workshop's materials have been posted to the following website for your convenience.

http://www.stata.com/users/jpitblado/2010jsm/

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

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.

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

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

2 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`**.

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

The standard header for a Stata estimation command contains a title and some information about the sample.

- **`proportion`** reports the sample size.
- **`svy: proportion`** also reports the number of strata, primary sampling units (PSU), the estimated population size, and the design degrees of freedom.

  `svy` reports the number of strata and PSUs, even for multistage designs, to show where the design degrees of freedom come from.

$$\mathrm{df} = N_{\mathrm{PSU}} - N_{\mathrm{strata}}$$

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. proportion sex
Proportion estimation                Number of obs    =   10351
```

	Proportion	Std. Err.	[95% Conf.	Interval]
sex				
Male	.4748333	.0049085	.4652117	.484455
Female	.5251667	.0049085	.515545	.5347883

```
. svy: proportion sex
(running proportion on estimation sample)
Survey: Proportion estimation
Number of strata =      31       Number of obs    =      10351
Number of PSUs   =      62       Population size  =  117157513
                                 Design df        =         31
```

	Proportion	Linearized Std. Err.	[95% Conf.	Interval]
sex				
Male	.4793502	.005734	.4676557	.4910447
Female	.5206498	.005734	.5089553	.5323443

Notice that **proportion** reports different proportion values than **svy: proportion**. This is due to the survey design characteristics that were **svyset** when the dataset was created. ◁

2.1 Single stage designs

Single-stage syntax

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

- 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*.

▷ Example

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

◁

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.

▷ Example
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.

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

◁

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*.

▷ Example

- States of the union are typically used as strata in national surveys in the US.
- Demographic information like age group, gender, and ethnicity.

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

◁

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.

❑ Note

- We will see that the FPC affects the number of components in the linearized variance estimator for multistage designs.
- We can use **`svyset`** to specify an *SRS* design.

❑

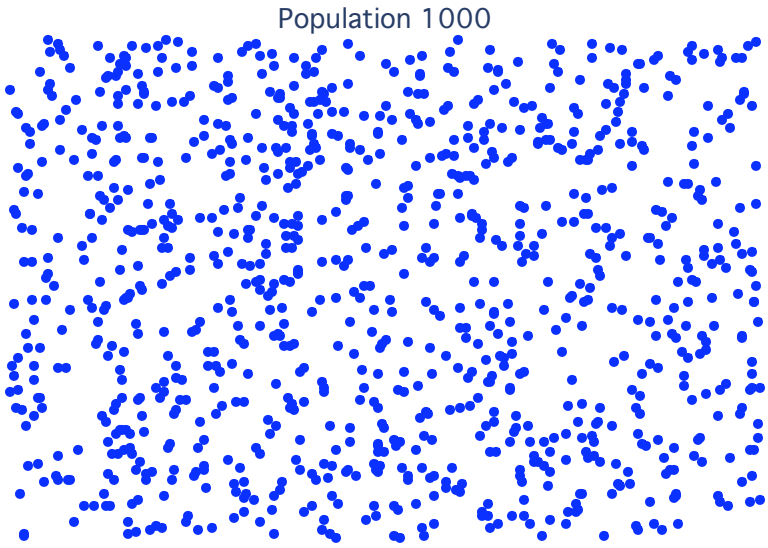
▷ Example: **`svyset`** for single-stage designs

1. **`auto`** – specifying an *SRS* design
2. **`nmihs`** – the National Maternal and Infant Health Survey (1988) dataset came from a stratified design
3. **`fpc`** – a simulated dataset with variables that identify the characteristics from a stratified and without-replacement clustered design

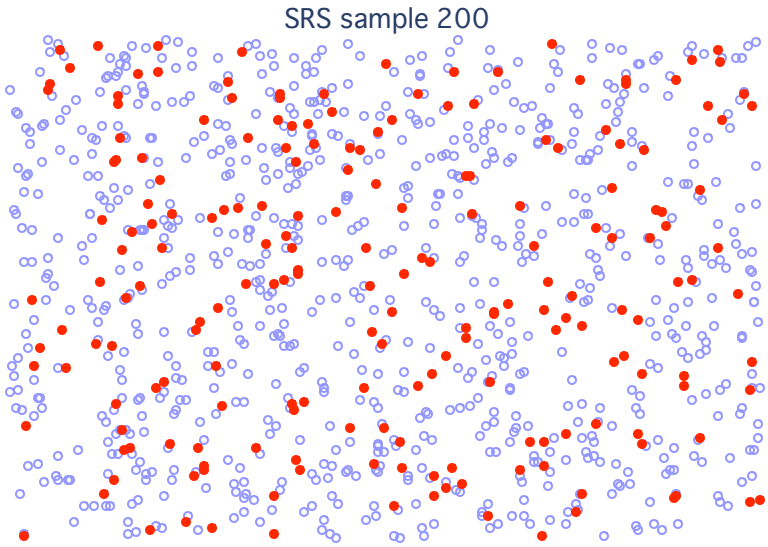
```
*** The auto data that ships with Stata
. sysuse auto
(1978 Automobile Data)
. svyset _n
      pweight: <none>
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
*** National Maternal and Infant Health Survey
. webuse nmihs
. svyset [pw=finwgt], strata(stratan)
      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
*** Simulated data
. webuse fpc
. svyset psuid [pw=weight], strata(stratid) fpc(Nh)
      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: stratid
         SU 1: psuid
        FPC 1: Nh
```

◁

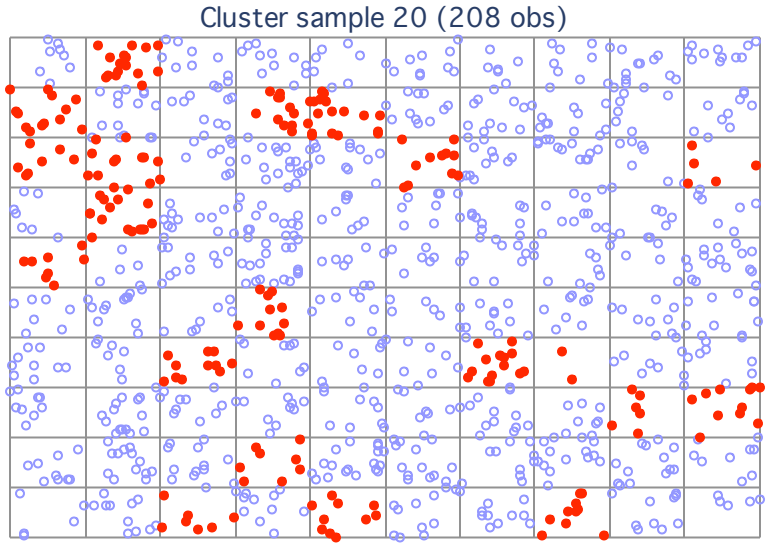
Below is a visual representation of a hypothetical population. Suppose each blue dot represents an individual.



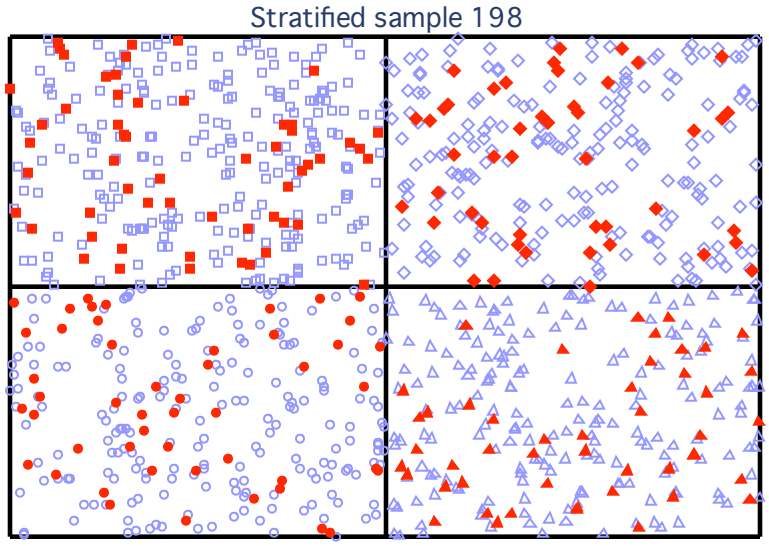
The following shows a 20% simple-random-sample. The solid symbols identify sampled individuals.



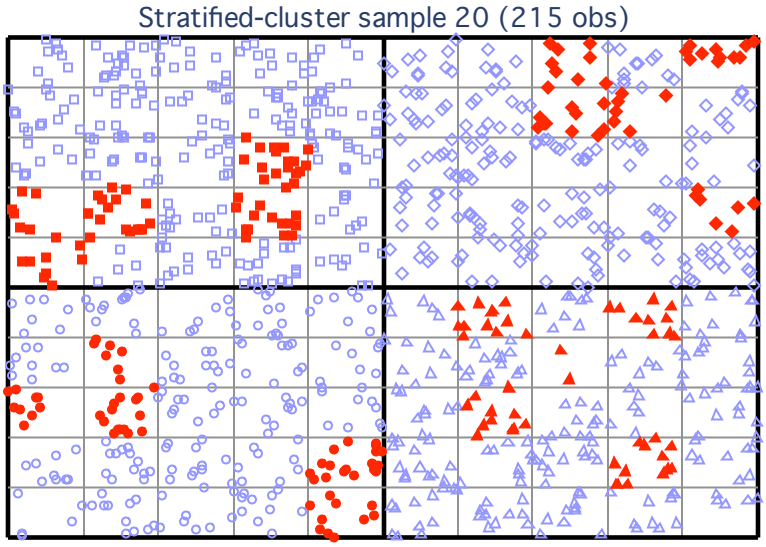
Here we partition the population into small blocks, then sample 20% of the blocks. Not all blocks contain the same number of individuals, so the sample size is a random quantity.



Here we partition the population into four big regions, then perform a 20% sample within each region. The sample size is not exactly 20% of the population size due to unbalanced regions and rounding.



Here we re-establish the smaller blocks within the four regions, then sample 20% of the blocks within each region.



2.2 Multistage designs

Let's use an example to introduce/motivate multistage designs.

▷ Example

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.

◁

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

❑ Note

svyset will note that it is disregarding subsequent stages when an FPC is not specified for a given stage.

❑

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.

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

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.

```
. webuse seniors
. svyset county [pw=sampwgt], strata(state) fpc(ncounties)
    || school, fpc(nschools)
    || _n, strata(gender) fpc(nseniors)
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
     Strata 3: gender
         SU 3: <observations>
        FPC 3: nseniors
```

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`

◁

2.3 Poststratification

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*`)`

❑ Note

Recall that I said it is usually not vey practical to stratify on demographic information such as age group, gender, and ethnicity. However we can usually poststratify on these variables using the frequency distribution information available from census data.

❑

▷ Example: **svyset** for poststratification

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. Thus we have an *SRS* design, the sampling weight is $26 = 1300/50$.

Notice that the dog clients are (on average) twice as expensive as cat clients. We can use the above frequency distribution of dogs and cats to poststratify on animal type.

```
*** Cat and dog data from Levy and Lemeshow (1999)
. webuse poststrata
. bysort type: sum totexp
```

```
-> type = dog
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      totexp |         32    49.85844    8.376695      32.78       66.2
```

```
-> type = cat
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      totexp |         18    21.71111    8.660666       7.14      39.88
```

Here are the mean estimates with poststratification:

```
. svyset [pw=weight], poststrata(type) postweight(postwgt) fpc(fpc)
      pweight: weight
          VCE: linearized
   Poststrata: type
   Postweight: postwgt
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: fpc
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       1          Number of obs    =      50
Number of PSUs   =      50          Population size  =    1300
N. of poststrata =       2          Design df        =      49
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
totexp	40.11513	1.163498	37.77699	42.45327

Here are the mean estimates without poststratification:

```
. svyset _n [pw=weight]
      pweight: weight
          VCE: linearized
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
. svy: mean totexp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       1          Number of obs    =      50
Number of PSUs   =      50          Population size  =    1300
                                    Design df        =      49
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
totexp	39.7254	2.265746	35.17221	44.27859

◁

2.4 Strata with a single sampling unit

How do we get stuck with strata that have only one sampling unit?

- Missing data can cause entire sampling units to be dropped from the analysis, possibly leaving a single sampling unit in the estimation sample.
- Certainty units
- Bad design

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

- Example: **svydes**

  The NHANES2 data has 31 strata, each containing 2 PSUs.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svydes
Survey: Describing stage 1 sampling units
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	2	380	165	190.0	215
2	2	185	67	92.5	118
3	2	348	149	174.0	199
4	2	460	229	230.0	231
5	2	252	105	126.0	147
6	2	298	131	149.0	167
7	2	476	206	238.0	270
8	2	338	158	169.0	180
9	2	244	100	122.0	144
10	2	262	119	131.0	143
11	2	275	120	137.5	155
12	2	314	144	157.0	170
13	2	342	154	171.0	188
14	2	405	200	202.5	205
15	2	380	189	190.0	191
16	2	336	159	168.0	177
17	2	393	180	196.5	213
18	2	359	144	179.5	215
20	2	285	125	142.5	160
21	2	214	102	107.0	112
22	2	301	128	150.5	173
23	2	341	159	170.5	182
24	2	438	205	219.0	233
25	2	256	116	128.0	140
26	2	261	129	130.5	132
27	2	283	139	141.5	144
28	2	299	136	149.5	163
29	2	503	215	251.5	288
30	2	365	166	182.5	199
31	2	308	143	154.0	165
32	2	450	211	225.0	239
31	62	10351	67	167.0	288

Some variables in this dataset have enough missing values to cause us the lonely PSU problem.

```
*** Mean high density lipids (mg/dL)
. svy: mean hdresult
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =      31        Number of obs    =      8720
Number of PSUs   =      60        Population size  =  98725345
                                  Design df        =        29
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
hdresult	49.67141	.	.	.

```
Note: missing standard error because of stratum with single
      sampling unit.
```

Use **`if e(sample)`** after estimation commands to restrict **`svydes`**'s focus on the estimation sample. The **`single`** option will further restrict output to strata with one sampling unit.

```
*** Restrict to the estimation sample
. svydes if e(sample), single
Survey: Describing strata with a single sampling unit in stage 1
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

			#Obs per Unit		
Stratum	#Units	#Obs	min	mean	max
1	1*	114	114	114.0	114
2	1*	98	98	98.0	98
	2				

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

```
*** Specifying variables for more information
. svydes hdresult, single
Survey: Describing strata with a single sampling unit in stage 1
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
```

			#Obs with	#Obs with	#Obs per included Unit		
Stratum	#Units included	#Units omitted	complete data	missing data	min	mean	max
1	1*	1	114	266	114	114.0	114
2	1*	1	98	87	98	98.0	98
	2						

◁

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.

2.5 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.

3 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.

❑ Note

- Linearization
  - Stata's **vce(robust)** for complex data
  - The default variance estimation method for **svy**.
- Replication methods
  - Motivation
    * Linearization can have poor performance in datasets with a small number of sampling units.
    * Due to privacy concerns, data providers are reluctant to release strata and sampling unit information in public-use data. Thus some datasets now come packaged with weight variables for use with replication methods.
  - Concept
    * Think of a replicate as a copy of the point estimates.
    * The idea is to resample the data, computing replicates from each resample, then using the replicates to estimate the variance.

❑

3.1 Linearization

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

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

- f_h is the sampling fraction for stratum h in the first stage.
- f_{hi} denotes a sampling fraction in the second stage.
- Remember that the design degrees of freedom is

$$\mathrm{df} = N_{\mathrm{PSU}} - N_{\mathrm{strata}}$$

▷ Example: **`svy: total`**

Let's use our (imaginary) survey data on high school seniors to estimate the number of smokers in the population.

```
. webuse seniors
. svyset
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
     Strata 3: gender
         SU 3: <observations>
        FPC 3: nseniors
*** Estimate number of seniors who have smoked
. svy: total smoked
(running total on estimation sample)
Survey: Total estimation
Number of strata =      50        Number of obs    =     10559
Number of PSUs   =     100        Population size  =  20992929
                                  Design df        =        50
------------------------------------------------------------------------------
             |             Linearized
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoked |    8347260   331155.1       7682115     9012404
------------------------------------------------------------------------------
*** Use first stage without FPC
. svyset county [pw=sampwgt], strata(state)
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: <zero>
. svy: total smoked
(running total on estimation sample)
Survey: Total estimation
Number of strata =      50        Number of obs    =     10559
Number of PSUs   =     100        Population size  =  20992929
                                  Design df        =        50
------------------------------------------------------------------------------
             |             Linearized
             |      Total   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoked |    8347260   346853.4       7650584     9043935
------------------------------------------------------------------------------
```

◁

Linearized variance for regression models

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

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

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

ML models

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

Least squares regression

- $\widehat{G}()$ is the normal equations
- s_j is a residual
- D is the inverse of the weighted outer product of the predictors—including the intercept

$$D = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$$

▷ Example: **`svy: logit`**

Here is an example of a logistic regression, modeling the incidence of high blood pressure as a function of some demographic variables.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svyset
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
*** Model high blood pressure on some demographics
. describe highbp height weight age female
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
highbp          byte   %8.0g                  1 if BP > 140/90, 0 otherwise
height          float  %9.0g                  height (cm)
weight          float  %9.0g                  weight (kg)
age             byte   %9.0g                  age in years
female          byte   %8.0g                  1=female, 0=male
. svy: logit highbp height weight age female
(running logit on estimation sample)
Survey: Logistic regression
Number of strata   =        31                 Number of obs      =      10351
Number of PSUs     =        62                 Population size    =  117157513
                                               Design df          =         31
                                               F(   4,     28)    =     178.69
                                               Prob > F           =     0.0000
------------------------------------------------------------------------------
             |             Linearized
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |  -.0316386   .0058648    -5.39   0.000    -.0435999   -.0196772
      weight |   .0511574   .0031191    16.40   0.000     .0447959     .057519
         age |   .0492406   .0023624    20.84   0.000     .0444224    .0540587
      female |  -.3215716   .0884387    -3.64   0.001    -.5019435   -.1411998
       _cons |  -2.858968   1.049395    -2.72   0.010    -4.999224   -.7187117
------------------------------------------------------------------------------
```

◁

3.2 Balanced repeated replication (BRR)

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

- The idea is to resample the data, compute replicates from each resample, then use the replicates to estimate the variance.
- 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.
- We can find a balanced subset by finding a Hadamard matrix of order r.
- When the dataset contains replicate weight variables, you do not need to worry about Hadamard matrices.

❑

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

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

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

❑ Note

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

❑

BRR variance formulas

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

❑ Note

- The default variance formula uses deviations of the replicates from their mean.
- The MSE formula uses deviations of the replicates from the point estimates.
- **BRR *** is clickable, taking you to a short help file informing you that you used the MSE formula for BRR variance estimation.

❑

▷ Example: **`svy brr: logit`**

Let's revisit the previous logistic model fit, but use BRR for variance estimation.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2brr
. svyset [pw=finalwgt], vce(brr) mse brrweight(brr_*)
      pweight: finalwgt
          VCE: brr
          MSE: on
    brrweight: brr_1 brr_2 brr_3 brr_4 brr_5 brr_6 brr_7 brr_8 brr_9 brr_10
               brr_11 brr_12 brr_13 brr_14 brr_15 brr_16 brr_17 brr_18 brr_19
               brr_20 brr_21 brr_22 brr_23 brr_24 brr_25 brr_26 brr_27 brr_28
               brr_29 brr_30 brr_31 brr_32
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
. svy: logit highbp height weight age female
(running logit on estimation sample)
BRR replications (32)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
................................
Survey: Logistic regression                     Number of obs      =      10351
                                                Population size    =  117157513
                                                Replications       =         32
                                                Design df          =         31
                                                F(   4,     28)    =     173.94
                                                Prob > F           =     0.0000
------------------------------------------------------------------------------
             |                BRR *
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |  -.0316386   .0058774    -5.38   0.000    -.0436255   -.0196516
      weight |   .0511574   .0031267    16.36   0.000     .0447806    .0575343
         age |   .0492406   .0023449    21.00   0.000     .0444581     .054023
      female |  -.3215716   .0897343    -3.58   0.001    -.5045859   -.1385574
       _cons |  -2.858968   1.044318    -2.74   0.010    -4.988868    -.729067
------------------------------------------------------------------------------
```

◁

3.3 Jackknife

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]
```

❑ 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.
- In the delete-1 jackknife, each PSU is represented by a corresponding replicate.
- The delete-k jackknife is only supported if you already have the corresponding replicate weight variables for **`svyset`**.

❑

For completeness, here is how the sampling weights are adjusted to produce the jackknife replicate weights.

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

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

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

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

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
- $m_h = (n_h - k)/c_h k$ – delete-k

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

- The default variance formula uses deviations of the replicates from their mean.
- The MSE formula uses deviations of the replicates from the point estimates.
- **`Jknife *`** is clickable, taking you to a short help file informing you that you used the MSE formula for jackknife variance estimation.
- Make sure to specify the correct multiplier when you **`svyset`** jackknife replicate weight variables.

❑

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

Here we are again with our now familiar logistic model fit, using the delete-1 jackknife variance estimator.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2
. svyset
      pweight: finalwgt
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: psu
        FPC 1: <zero>
. svy jknife, mse: logit highbp height weight age female
(running logit on estimation sample)
Jackknife replications (62)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
............
Survey: Logistic regression
Number of strata   =        31                  Number of obs      =      10351
Number of PSUs     =        62                  Population size    =  117157513
                                                Replications       =         62
                                                Design df          =         31
                                                F(   4,     28)    =     178.53
                                                Prob > F           =     0.0000
------------------------------------------------------------------------------
             |             Jknife *
      highbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |  -.0316386   .0058674    -5.39   0.000    -.0436052   -.0196719
      weight |   .0511574   .0031203    16.40   0.000     .0447936    .0575213
         age |   .0492406   .0023634    20.83   0.000     .0444204    .0540607
      female |  -.3215716    .088471    -3.63   0.001    -.5020093   -.1411339
       _cons |  -2.858968   1.049924    -2.72   0.011    -5.000302    -.717633
------------------------------------------------------------------------------
```

◁

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]`

3.4 Bootstrap

Bootstrap variance formulas

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

3.5 Successive difference replication (SDR)

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

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

Default variance formula:

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

Mean squared error (MSE) formula:

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

3.6 Replicate weights

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

One final look at our logistic model fit, using replicate weight variables. Notice that the stratum and multiplier information is saved as variable characteristics.

```
*** Second National Health and Nutrition Examination Survey
. webuse nhanes2jknife
. svyset [pw=finalwgt], vce(jackknife) jkrweight(jkw_*)
      pweight: finalwgt
          VCE: jackknife
          MSE: off
    jkrweight: jkw_1 jkw_2 jkw_3 jkw_4 jkw_5 jkw_6 jkw_7 jkw_8 jkw_9 jkw_10
               jkw_11 jkw_12 jkw_13 jkw_14 jkw_15 jkw_16 jkw_17 jkw_18 jkw_19
               jkw_20 jkw_21 jkw_22 jkw_23 jkw_24 jkw_25 jkw_26 jkw_27 jkw_28
               jkw_29 jkw_30 jkw_31 jkw_32 jkw_33 jkw_34 jkw_35 jkw_36 jkw_37
               jkw_38 jkw_39 jkw_40 jkw_41 jkw_42 jkw_43 jkw_44 jkw_45 jkw_46
               jkw_47 jkw_48 jkw_49 jkw_50 jkw_51 jkw_52 jkw_53 jkw_54 jkw_55
               jkw_56 jkw_57 jkw_58 jkw_59 jkw_60 jkw_61 jkw_62
  Single unit: missing
     Strata 1: <one>
         SU 1: <observations>
        FPC 1: <zero>
. char list jkw_1[]
  jkw_1[jk_multiplier]:       .5
  jkw_1[jk_stratum]:          1
```

The standard error estimates in this example should match those of the previous (we **`svyset`** our data using the correct multipliers and stratum identifiers).

```
. svy, mse: logit highbp height weight age female
(running logit on estimation sample)
Jackknife replications (62)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
............
Survey: Logistic regression
Number of strata   =        31                Number of obs      =      10351
                                              Population size    =  117157513
                                              Replications       =         62
                                              Design df          =         31
                                              F(   4,     28)    =     178.53
                                              Prob > F           =     0.0000
```

highbp	Coef.	Jknife * Std. Err.	t	P>|t|	[95% Conf.	Interval]
height	-.0316386	.0058674	-5.39	0.000	-.0436052	-.0196719
weight	.0511574	.0031203	16.40	0.000	.0447936	.0575213
age	.0492406	.0023634	20.83	0.000	.0444204	.0540607
female	-.3215716	.088471	-3.63	0.001	-.5020093	-.1411339
_cons	-2.858968	1.049924	-2.72	0.011	-5.000302	-.717633

◁

4 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.

❑ 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.
- 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`**.
- We can illustrate the difference between these estimators with an *SRS* design.

❑

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:

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

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.

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

▷ Example: **`svy, subpop()`**

Suppose we want to estimate the mean birth weight for mothers with high blood pressure. The `highbp` variable (in the `nmihs` data) is an indicator for mothers with high blood pressure.

In the reported results, the subpopulation information is provided in the header. Notice that although the restricted sample results reproduce the same mean, the standard errors differ.

```
*** National Maternal and Infant Health Survey
. webuse nmihs
. svyset [pw=finwgt], strata(stratan)
      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
*** Focus: birthweight, mothers with high blood pressure
. describe birthwgt highbp
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
birthwgt        int    %8.0g                  Birthweight in grams
highbp          byte   %8.0g       hibp       High blood pressure: 1=yes,0=no
. label list hibp
hibp:
           0 norm BP
           1 hi BP
*** Subpopulation estimation
. svy, subpop(highbp): mean birthwgt
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       6         Number of obs    =     9953
Number of PSUs   =    9953         Population size  =  3898922
                                   Subpop. no. obs  =      595
                                   Subpop. size     = 186196.7
                                   Design df        =     9947
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    birthwgt |   3202.483   33.29493      3137.218    3267.748
--------------------------------------------------------------
*** Restricted sample estimation
. svy: mean birthwgt if highbp
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =       6          Number of obs    =     595
Number of PSUs   =     595          Population size  =  186197
                                    Design df        =     589
--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
    birthwgt |   3202.483    28.7201      3146.077     3258.89
--------------------------------------------------------------
```

◁

5 Postestimation

5.1 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.

❑ Note

Most of the postestimation commands that work after the standard estimation commands also work after **svy**. The available options may differ between standard and survey results. ❑

5.2 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

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

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

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

▷ Example: **`estat effects`**

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

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

```
*** National Maternal and Infant Health Survey
. webuse nmihs
. svyset [pw=finwgt], strata(stratan)
      pweight: finwgt
          VCE: linearized
  Single unit: missing
     Strata 1: stratan
         SU 1: <observations>
        FPC 1: <zero>
*** Focus: birthweight between mothers with normal and high blood pressure
. describe birthwgt highbp
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
birthwgt        int    %8.0g                  Birthweight in grams
highbp          byte   %8.0g       hibp       High blood pressure: 1=yes,0=no
*** Labels that can also be used as Stata names
. label list hibp
hibp:
           0 norm BP
           1 hi BP
. label define bloodpressure 0 "BPnormal" 1 "BPhigh"
. label values highbp bloodpressure
```

We'll use the **`over()`** option so that we can estimate the means for both subpopulations simultaneously.

```
*** Focus: birthweight between mothers with normal and high blood pressure
. svy: mean birthwgt, over(highbp)
(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
     BPnormal: highbp = BPnormal
       BPhigh: highbp = BPhigh
------------------------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
birthwgt     |
    BPnormal |   3363.131   6.605511      3350.183    3376.079
      BPhigh |   3202.483   33.29483      3137.219    3267.748
------------------------------------------------------------------------------
```

We now have a variance matrix that we can use to perform tests or compute linear combinations, but first we'll use **`estat effects`** to display design effects for the entire set of point estimates.

```
*** Design effects
. estat effects
     BPnormal: highbp = BPnormal
       BPhigh: highbp = BPhigh
```

Over	Mean	Linearized Std. Err.	DEFF	DEFT
birthwgt				
BPnormal	3363.131	6.605511	1.17952	1.08606
BPhigh	3202.483	33.29483	1.14081	1.06809

We can also 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`** for a given set of estimation results, **`estat effects`** posts the newly acquired information to **`e()`** for future reference.

```
*** Misspecification effects require an extra model fit
. estat effects, meff meft
     BPnormal: highbp = BPnormal
       BPhigh: highbp = BPhigh
```

Over	Mean	Linearized Std. Err.	MEFF	MEFT
birthwgt				
BPnormal	3363.131	6.605511	.424756	.651733
BPhigh	3202.483	33.29483	.657389	.810796

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'll use **`lincom`** to compute the difference of the means and get a 95% CI instead. We follow that up with **`estat lceffects`** to get design and misspecification effects for our linear combination.

```
*** Focus: birthweight between normal and high blood pressure
*** Linear combinations
. lincom [birthwgt]BPnormal - [birthwgt]BPhigh
 ( 1)  [birthwgt]BPnormal - [birthwgt]BPhigh = 0
```

Mean	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	160.6476	34.28316	4.69	0.000	93.44569	227.8496

```
*** Effects for linear combinations
. estat lceffects [birthwgt]BPnormal - [birthwgt]BPhigh
 ( 1)  [birthwgt]BPnormal - [birthwgt]BPhigh = 0
```

Mean	Coef.	Std. Err.	DEFF	DEFT	MEFF	MEFT
(1)	160.6476	34.28316	1.16519	1.07944	.656975	.81054

◁

6 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

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

[2] StataCorp. 2009. *Survey Data Reference Manual: Release 11.* College Station, TX: StataCorp LP.