Lab 4: Weights and Survey Commands

Simon Halliday

ECO311, Fall 2016

Revision Quiz

In your notes for your exercises for this lab, try to answer the following quiz questions based on Lab 3 *before class* and *without checking the lab*! In your exercises, have three categories for each question, "M" for my answer, "TPS" for think-pair-share, and "N" for after notes.

a. Explain one method you would sue to export your regression analysis or summary statistics to a publishable output (MS Word, Excel, etc).

b. Explain in English what the following set of commands do:

```
gen someschool = 1
replace someschool = 0 if w1_r_b7 == 25
replace someschool = . if w1_r_b7 == .
```

c. Explain what you think the following commands do and explain why we might use the variable defined on the third line of the commands, `lnpcy`:

```
egen hhsize = count(pid), by(hhid)
gen pcy = hh_income / hhsize
gen lnpcy = ln(pcy)
```

d. Explain what you would do to complete the following tasks:

   i. Tag one member of each household

   ii. Temporarily save your data in memory

   iii. Get rid of all the household members who haven't been tagged

   iv. Find the average of the variable `pcy` defined above

   v. Go back to the data in memory before you eliminated the household members.

e. What process would you go through to visualize the relationship between years of education and the variable `pcy` we defined above?

When you get to class, check your answers with your neighbors (Think-Pair-Share). Once you have checked with your neighbors, go back to the lab to check what you didn't remember from the lab notes.

Doing Stata

By the end of this lab, you should have:

- Better understand survey designs and sampling
- Understand what a *survey weight* is
- Be able to correct frequency tables and summary statistics for survey weights
- Understand how to use the specifications of a survey's strata and clusters to set up survey commands, `svy`
- Understand how the various aspects of survey design affect the *precision* of statistical estimates
- Know how to run regressions using the survey design
- Know how to define subsamples in the population once a survey design has been set up

Introduction

Before you start this lab session remember to do the following:

- Direct your working directory to the Lab4 folder (or Labs folder if you prefer)
- Make sure that your directory structure is consistent with Lab1 (CommandFiles, log, cmdlog, etc) so you can use *relative file paths*.
- Open a `log` file to track your work:

  ```
  log using ./ProcessingAnalysis/CommandFiles/logs/
      yourlastname_lab4.log
  ```

  Be sure to specify the file extension `.log`
- Open a cmdlog file

  ```
  cmdlog using ./ProcessingAnalysis/CommandFiles/cmdlogs/
      yourlastname_lab4.do
  ```

After you have followed these steps, copy the NIDS data into your OriginalData folder for Lab4 and then open the data.

Preparation

Run the following commands, which we'll need for variables we want to use in this lab.

```
gen death = w1_h_c1 == 1
replace death =. if w1_h_c1 == . | w1_h_c1 < 0
recode w1_r_b7 (−9/−3 24 =.) (25 = 0) (13 16 = 10) (14 17
     = 11) (15 = 12) (18 = 13) (19 = 14) (20 = 15) (21 22
    = 16) (23 = 17), gen(edyears)
gen electricity = w1_h_d22 == 1
replace electricity = . if  w1_h_d22 == . |  w1_h_d22 < 0
egen hh_one = tag(hhid)
gen pcminc = w1_hhincome / w1_hhsizer
```

Surveys and Weights

Each household (person) interviewed for the survey "represents" some larger group of households (people) in the total population. To make your results representative of the population, you tell Stata to use the weight provided in the data set. Stata uses this weight to weigh some observations more heavily than others. Stata distinguishes between four types of weights – `fweights`, `pweights`, `aweights` and `iweights`.

To read about these different types of weights type

```
help weight
```

You should see information similar to the following:

- `fweights` – frequency weights
  - represent the number of duplicated observations – hence ‘frequency’
  - used for grouped data
  - therefore have to be integers
  - an fweight of 30 means there are 30 observations with the same values on all variables
- aweights – analytic weights
  - these are weights that are inversely proportional to the *variance* of an observation
  - used for heteroscedasticity corrections
  - the scale of the weight is arbitrary: it is only relative weights that matter, Stata will rescale these to sum to one
  - common use is when observations represent averages and the weights are the number of elements that gave rise to the average
  - aweight of 30 means that data is the average of 30 observations

REMINDER *heterescedasticity* occurs when the variances of variables differ or are ‘heterogeneous’, which is the opposite of them being homoscedastic which we assume for OLS estimation.

- `pweights` probability weights
  - these are sampling weights
  - these weights are the number of subjects in the full population that the sampled observation in your data represents
  - pweight of 30 means the observation had a 1 in 30 probability of being included in the sample and “represents” 30 subjects in the population
- `iweights` – importance weights
  - mostly used by programmers, we ignore them

Most Stata commands can deal with weighted data although there are some commands that do not allow all four types of weighs. The syntax for producing weighted results is the same in most Stata commands: you specify the weight variable inside square brackets at the end of the command but before the comma. Below are some examples:

```
mean edyears [w = w1_wgt]
tab w1_best_race [w = w1_wgt], sum(pcminc)
ci electricity if hh_one == 1 [w = w1_wgt]
sum pcminc if hh_one == 1 [w = w1_wgt]
```

In each of these examples, we did not specify the weight type so STATA uses the default type of weight for that command. The wrong weight type can produce very misleading standard errors. To see this let's estimate mean household income using fweights, aweights and pweights.

```
mean w1_hhincome [fw = w1_wgt]
mean w1_hhincome [aw = w1_wgt]
mean w1_hhincome [pw = w1_wgt]
```

Notice that the point estimates are the *same* for all types of weights but the *standard errors* of the estimates are vastly different. Rather than allowing Stata to use the default weight, it is safer to explicitly specify which weight type you would like Stat to use.

REMINDER What do we mean by *point estimate*? If you're not sure, ask the person next to you. If they're not sure too, then Google it.

For `pweights` contrast the output from following two commands:

```
mean edyears
mean edyears [pw = w1_wgt]
```

In Stata the default weight for the tab command followed by a single variable is a frequency weight. Frequency weights must be an integer so we will get an error message if we type

```
tab w1_best_race [w = w1_wgt]
```

The tab command does not accept pweights so we need to use aweights. To tell STATA that we want to use aweights we replace the w in the weights bracket with aw.

Type:

```
tab w1_best_race [aw = w1_wgt]
tab w1_hhprov if hh_one == 1 [aw = w1_wgt]
```

Compare these summary details to the *unweighted* details:

```
ci edyears
tab w1_best_race, sum(pcminc)
ci electricity if hh_one == 1
sum pcminc if hh_one == 1
tab w1_best_race if hh_one == 1
tab w1_hhprov if hh_one == 1
```

Exercise 1

a. Which provinces were oversampled?

b. Which provinces were undersampled?

The survey commands

Stata has developed a whole range of commands for summarizing and analyzing complex sample data – the `svy` commands. The `svy` commands make it easier for us to incorporate the *weighting* we looked at in the previous section, as well as the issue of survey *stratification* and *clustering*.

Type

```
help svy
```

to see a description of these commands.

In order to use the svy commands we first have to let Stata know which variables specify the sampling design. We use the svyset command to specify the weight variable (w1_wgt which we learned about in Lab 3), stratum variable (w1_hhdc) and the primary sampling unit or cluster variable (w1_hhcluster). To set up these variables as the survey design, type

```
svyset w1_hhcluster [pw = w1_wgt], strata(w1_hhdc)
```

We are now ready to use the svy commands. For a summary of the sampling design type

```
svydes
```

Previously we calculated a confidence interval for the number of households with electricity using the ci command. When using the ci command, we assumed that the sample was a *simple random sample* with epsem sampling.

REMINDER Check here for a reminder about what SRS means: http://en.wikipedia.org/wiki/Simple_random_sample. Also if you're not sure what 'epsem' means, Google it. I promise you learned the idea of it (if not the abbreviation) in ECO220 or MTH220.

Having learned the command for confidence intervals, we introduced weights to our ci command to adjust for differential probabilities of selection, non-response and post-stratification in the previous section. While using the weights meant we had *unbiased* estimates, the standard errors were still calculated on the assumption that the sample was an SRS. But, all is not lost! Stata helps us out by providing commands to cater for the complexity of survey design (with underlying fairly complicated econometrics we won't go into here). In Stata, we can use the svy commands. For example, to examine the means of some variable of interest, we use the command svy: mean to take the complex sample design into account in calculating our standard errors. Type

```
svy: mean electricity if hh_one == 1
```

Notice how the estimated proportion of households with electricity is the same as when we used the ci command with weights but the standard errors are quite different. To see the design effect for this estimate, type:

```
estat effects, deff
```

The design effect (*deff*) of 19.76 is the ratio of the variance calculated using the complex sample design to the variance calculated assuming SRS. There are 7011 households in our sample. The design effect of 19.76 means that, due to our complex sample design, our precision is *only as good as* an SRS of $7011/19.76 = 355$.

REMINDER Do you think precision is increases or decreased with a smaller equivalent sample size? Why?

To examine the impact of the various aspects of the complex sample design on our estimates we will look at each component in turn. First, we will clear the survey design variables. Type

```
svyset, clear
```

Let us first look at the effect of weights. Type

```
svyset [pw = w1_wgt]
svy: mean electricity death pcinc if hh_one == 1
estat effects, deff
```

Exercise 2

Look at the design effects. What do they tell you about the impact of the weights on the *precision* of the estimates?

Next let's examine the *strata*. Type

```
svyset, strata(w1_hhdc)
svy: mean electricity death pcinc if hh_one == 1
estat effects, deff
```

Notice how the design effects are now less than 1 indicating a gain in *precision* (that is, a reduction in the size of the standard errors).

Exercise 3

a. For which variable is there a larger gain in precision?

b. Why do you think it gains more precision?

Third let's examine the effect of *clustering*. Type

CLUSTERING Quora has a really great discussion of clustering: http://www.quora.com/Why-do-we-use-clustering-in-statistical-analysis-Can-you-give-an-intuitive-explanation-or-intuitive-examples

```
svyset w1_hhcluster
svy: mean electricity death pcinc if hh_one==1
estat effects, deff
```

Exercise 4

a. What are the design effects?

b. For which variable is there the largest loss in precision?

c. For which variable is there the smallest loss in precision? Why?

d. Can you think of examples of other variables where clustering would have a large impact?

Finally let's put it altogether

```
svyset w1_hhcluster [pw = w1_wgt], strata(w1_hhdc)
svy: mean electricity death pcinc if hh_one == 1
estat effects, deff
```

If we want to look at the frequency distribution of electricity by toilet type, type

```
svy: tab w1_h_d20 electricity if hh_one == 1, row
```

Suppose we wanted to estimate the total number of households in South Africa with electricity. Then we would type

```
svy: total electricity if hh_one == 1
```

Exercise 5

a. Assuming we had an SRS design with epsem sampling (and no non-response etc.), what would the *inflation weight* per household be?

b. Using the weight we did above how many households would you estimate have electricity?

c. What is the difference?

Subpopulations with svy

We have been restricting our sample to one observation per household using an 'if' statement. This is OK as the household was the sampling unit. When we are using the svy commands to look at sub-populations (e.g. women only, adults) we cannot use if statements to restrict the sample. We need to use the subpop option. The first step is to create a dummy variable that takes the value 1 if the observation should be included in the estimation sample and 0 otherwise. We then add the subpop option to the svy prefix. To estimate the average education level of adult males type:

```
gen adultmale = w1_r_b4 == 1 & w1_r_b6 >= 18 & w1_r_b6 !=
     .
svy, subpop(adultmale): mean edyears
```

Regressions with svy

Let's re-run the regressions from Lab 2 taking the complex sample design into account. We will need to create a dummy variable to indicate that the individual is between the ages of 18 and 80 (inclusive).

```
gen age18to80 = w1_r_b6 >= 18 & w1_r_b6 <= 80
svy, subpop(age18to80): reg edyears w1_r_b6 i.w1_r_b4
svy, subpop(age18to80): reg edyears w1_r_b6 i.
    w1_best_race
xi: svy, subpop(age18to80): reg edyears i.w1_best_race*
    w1_r_b6
```

Exercise 6

1. Compare the results to those when you do not take the sample design into account (how would you do this?).
2. Notice that the standard errors on your regression coefficients are *larger* once the clustering and weights are taken into account. What does this mean (think about the word ‘precision’)?
3. How would you interpret the results of the regression you just ran? What do the coefficients mean?

Commands and options in Lab 4

weights	w =	aw =
pw =	fw =	iw =
svyset, strata()	estat svydes	svy: mean
svy: reg	svy: tab	svy: total
svy, subpop()		