Stata Assignment 1

Econ 120B Spring 2020 · Xinwei Ma
Department of Economics, UCSD

- You should **submit your answers on Gradescope, including your do-file**. We will also manage regrade requests using Gradescope.
- The deadline of submitting this Stata Assignment is **May 3, 11:59pm**. No late submission will be accepted.
- This Stata Assignment will be graded on five scales: 0%, 25%, 50%, 75%, and 100%. If your do-file does not run, we will subtract 25%.
- Name your do-file with your PID, such as `A12345678.do`. Start your do-file with the following (also include your name and PID in your do-file)

```
/*******************************************************************************
ECON 120B, Spring 2020
Stata Assignment 1

Name:
PID:
*******************************************************************************/

clear all // clear the environment/memory
set more off
sysuse nlsw88 // load the built-in dataset nlsw88
```

  Please make sure your do-file is clearly documented to help us understand your code.

- `nlsw88` is a built-in dataset that comes with Stata. It is an extract from the 1988 round of the National Longitudinal Survey of Mature and Young Women. Following is a summary of the variables in this dataset.

`idcode`	survey id
`age`	age
`race`	race, can take three values, *white*, *black* or *other*
`married`	= 1 if is currently marries, = 0 otherwise
`never_married`	= 1 if never married, = 0 otherwise
`grade`	current grade completed
`collgrad`	= 1 if graduated from college, = 0 otherwise
`south`	= 1 if lives in southern states, = 0 otherwise
`smsa`	= 1 if lives in standard metropolitan statistical area, = 0 otherwise
`c_city`	= 1 if lives in central city, = 0 otherwise
`industry`	industry, use `tab industry` to see the categories
`occupation`	occupation, use `tab occupation` to see the categories
`union`	= 1 if is in a union, , = 0 otherwise
`wage`	hourly wage, measured in $
`hours`	hours worked per week
`ttl_exp`	total work experience, measured in years
`tenure`	current job tenure, measured in years

  More information on the original data can be found here:

  https://www.bls.gov/nls/orginal-cohorts/mature-and-young-women.htm

1. In this exercise you will re-label variables and create some new variables which will be used later.

   (a) Re-label the variable `smsa` to "lives in urban area" so that it is more informative. Note that SMSA stands for "standard metropolitan statistical area."

   (b) Re-name the variable `smsa` to `urban`.

   (c) Generate a new variable called `wageofc` taking the same values as the variable `wage`, so that we can modify the wage data without loosing the original variable.

   (d) The minimum wage in 1988 was $3.35 an hour. Let's say our fictional bosses at the Bureau of Labor Statistics will be mad if they see evidence of minimum wage law violations in the dataset. Re-classify those earning below minimum wage as "volunteers." To be more specific, In `wageofc`, replace `wageofc` with 0 for workers that earned strictly less than $3.35 an hour. Note that we often find evidence of statutes not being followed in datasets.

   (e) How many observations are in this dataset?

   (f) How many non-missing observations are in `wageofc`?

   (g) Generate a variable called `lnwageofc` which is the natural logarithm of `wageofc`.

   (h) How many non-missing observations are in `lnwageofc`? Why does this make sense?

2. In this exercise, you are asked to compute some simple summary statistics using the binary variable `collgrad`, contained in the dataset.

   (a) Use the command `tabulate` to show the categories of the variable `collgrad` and their frequencies. What is the relative frequency of the category *college grad*? Please report a number between 0 and 1.

   (b) Use the same command, this time specifying the option `nolabel`, to visualize the numeric values corresponding to the different categories of `collgrad`. Which numeric value corresponds to the label *college grad*?

   (c) Use the command `summarize` to compute the sample mean of `collgrad`. After executing `summarize`, Stata stores temporarily the sample mean in the object `r(mean)`. To see this, generate a scalar variable `collgrad_mean` equal to `r(mean)`, by typing `scalar collgrad_mean = r(mean)` in the line just after the command `summarize`. Finally, display the variable value by typing `display collgrad_mean`, and verify that the value displayed is the same as the one returned by the command `summarize`. What is the sample mean of `collgrad`? What is its relation to your answer in 2(a)?

   (d) Repeat the steps of 2(c), this time to create a scalar variable, `collgrad_var`, containing the sample variance of `collgrad`. What is the sample variance of `collgrad`?

   (e) Compute the sample variance of `collgrad` without the summarize command, using only the variable `collgrad_mean`. (Hint: you can think of `collgrad` as drawn from a Bernoulli distribution with parameter p, where p is the probability of having graduated from college. The (population) variance of a Bernoulli is $p(1-p)$. What is the relation between p and the sample mean `collgrad_mean`? Finally, remember that the sample variance can be obtained starting from the formula of the population variance by replacing the population mean with the sample mean.)

3. The following problems provide more practice using conditional statements to tabulate and summarize variables.

   (a) How many unmarried people in the dataset were married before? (Hint: use the variables, `married` and `never_married`.)

   (b) What is the difference in average hours worked for married and unmarried workers? Please report a positive number. (Hint: use the variables `married` and `hours`.)

   (c) What is the average hours worked for married college graduates with strictly more than 10 years of experience? (Hint: use the variables `married`, `collgrad`, `ttl_exp`, and `hours`.)

(d) What fraction of laborers or craftsman that live in urban areas are black? Please report a number between 0 and 1. (Hint: use the variables `occupation`, `urban`, and `race`.)

(e) Using the variable `wageofc`, what fraction of workers that earn strictly more than \$7 an hour are in a union? Please report a number between 0 and 1. (Be careful about missing values.)

(f) Using the variable `lnwageofc`, what fraction of workers that earn strictly more than \$7 an hour are in a union? Please report a number between 0 and 1. (That is, you should compare the variable, `lnwageofc`, to $\ln 7$. Be even more careful about missing values.)

4. This exercise refers to the following model:

$$\texttt{wage}_i = \beta_0 + \beta_1 \texttt{grade}_i + u_i,$$

where the wage of individual i is regressed on his/her highest grade completed and a constant term. You are asked to compute the intercept and slope estimates in a variety of ways, and compare your results in each case. First, use the command

```
keep if !missing(wage, grade)
```

to drop people with missing `wage` or `grade` from the dataset. How many observations were dropped?

(a) Use the `regress` command to estimate the OLS coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. What is the value of $\hat{\beta}_0$? What is the value of $\hat{\beta}_1$? (Hint: type `regress wage grade`, the constant term will be added automatically to the regression.)

(b) You are now asked to compute the same estimates using the formulas we derived in the lecture. Adopt the following procedure:

- Compute the sample covariance between `wage` and `grade`, and the sample variance of `grade`, and save them in two scalars, `cov_wg` and `var_g`. (Hint: you can compute the variance-covariance matrix using the `corr` command, with the option `covariance`. For instance, if you type `corr wage grade, covariance`, the output will be a matrix containing the variance of `wage`, the variance of `grade` and the covariance between `wage` and `grade`; the three values will be stored in `r(Var_1)`, `r(Var_2)` and `r(cov_12)`, respectively. You can check the list of stored objects by typing `return list` just after running the `corr` command.)

- Generate the scalar `beta_1` equal to `cov_wg/var_g` and display it by typing `display beta_1`. What is the relation between this estimate for β_1 and the one in 4(a)?

- Create two scalars, `grade_mean` and `wage_mean`, equal to the sample means of `grade` and `wage`.

- compute your estimate for β_0 by typing `scalar beta_0 = wage_mean - beta_1 * grade_mean`, and then display `beta_0`. What is the relation between this estimate for β_0 and the one in 4(a)?

(c) Finally, you can compute $\hat{\beta}_1$ using a "centered" regression. For this part, Adopt the following procedure:

- Define a new variable, `wage_0` as `wage - wage_mean`, so that this new variable has a sample mean of 0. Similarly, define `grade_0` as `grade - grade_mean`. This is called "demeaning" or "centering" a variable.

- Regress the centered variable, `wage_0`, on the other centered variable, `grade_0`. What are the intercept and slope estimates in this new regression?