

- You should **submit your answers on Canvas, including your do-file**.
- The deadline of submitting this Stata Assignment is **April 22, 11:59pm**. No late submission will be accepted.
- This Stata Assignment will be graded on three scales: 0%, 5%, and 10%. (For example, if you are able to answer more than half of the questions correctly, you will get the full credit for this assignment.)
- If your do-file does not run, we will subtract 2.5%.
- 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 2022
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 reasonably 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.

|                      |  |
|----------------------|--|
| <b>idcode</b>        | survey id  |
| <b>age</b>           | age  |
| <b>race</b>          | race, can take three values, <i>white</i> , <i>black</i> or <i>other</i> |
| <b>married</b>       | = 1 if is currently married, = 0 otherwise                               |
| <b>never_married</b> | = 1 if never married, = 0 otherwise                                      |
| <b>grade</b>         | current grade completed  |
| <b>collgrad</b>      | = 1 if graduated from college, = 0 otherwise                             |
| <b>south</b>         | = 1 if lives in southern states, = 0 otherwise                           |
| <b>smsa</b>          | = 1 if lives in standard metropolitan statistical area, = 0 otherwise    |
| <b>c_city</b>        | = 1 if lives in central city, = 0 otherwise                              |
| <b>industry</b>      | industry, use <code>tab industry</code> to see the categories            |
| <b>occupation</b>    | occupation, use <code>tab occupation</code> to see the categories        |
| <b>union</b>         | = 1 if is in a union, = 0 otherwise                                      |
| <b>wage</b>          | hourly wage, measured in \$  |
| <b>hours</b>         | hours worked per week  |
| <b>ttl_exp</b>       | total work experience, measured in years                                 |
| <b>tenure</b>        | 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 `live in urban areas` 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 `wagecopy` taking the same values as the variable `wage`, so that we can modify the wage data without losing 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. In `wagecopy`, replace `wagecopy` with 0 for workers that earned strictly less than \$3.35 an hour.
  - (e) How many observations are in this dataset?
  - (f) How many missing observations are in `wagecopy`?
  - (g) Generate a variable called `lnwagecopy` which is the natural logarithm of `wagecopy`.
  - (h) How many missing observations are in `lnwagecopy`? 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 two categories of the variable `collgrad` and their frequencies. What is the proportion of the category `not 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`? (Hint: after running the `summarize` command, you can find the sample variance by `r(Var)`).
  - (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 a college degree. The (population) variance of a Bernoulli random variable 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) Among unmarried people, what is the fraction of those who were married before? You should report a number between 0 and 1. (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 unmarried college graduates with strictly more than 8 years of experience?

(Hint: use the variables `married`, `collgrad`, `t1l_exp`, and `hours`.)

- (d) Among those living in urban areas, what is the fraction of laborers? Please report a number between 0 and 1. (Hint: use the variables `occupation` and `urban`. In addition, missing values should not be counted in your calculation.)
- (e) Use the variable `wagecopy`. Among unionized workers, what is the fraction of those who earn strictly more than \$6.5 an hour? Please report a number between 0 and 1.

4. This exercise refers to the following model:

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

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.

- (a) How many observations were dropped?
- (b) Use the `regress` command to estimate the 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. To find  $\hat{\beta}_0$ , check the row labeled by `_cons` and the column labeled by `Coef`.)
- (c) You are now asked to compute the same estimates using a different 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`.  
  
On a related note, 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  $\beta_1$  and the one in 4(b)?