

**IR602 B1: Quantitative Analysis for International Affairs**  
**Frederick S. Pardee School of Global Studies**  
**Problem Set 3**

***General Instructions for Problem Sets***

Please conduct and document all of your work using a Stata .do ("DO-file") file. You can open a new .do file with Ctrl-9 on a PC and CMD-N on a Mac.

***Submitting Your Assignment***

Please email your assignment to the IR602 e-mail account at **ir602.2021@gmail.com**. You do not need to submit any additional write-up, unless you would like further feedback.

**Problem 1: Derivation of OLS**

- a. Open Stata, and generate a data set with 100 observations containing the following variables:
  1. **height\_mother**: a random variable with mean 165 cm and standard deviation 10 cm.
  2. **noise**: a normal random variable with mean 5 cm and standard deviation 10 cm.

Remember to set the seed of the dataset before generating the variables above. Set the **seed** of the dataset to 6023.

Give proper **labels** to each of the variables. **Summarize** the variables (provide the mean, standard deviation, minimum, and maximum for each variable), and generate **histograms** for each of the variables.

- b. Generate the height of the child (variable "child height"), **hc**, variable which is given by the true relationship defined below:

$$h_c = 10 + 0.5 \cdot h_m + \varepsilon_c$$

where **hm** is the height of the mother and  $\varepsilon_c$  is the noise term. Summarize this new variable and generate the histogram to show the distribution of child height. What is the mean child height that you obtain?

- c. The **scatter** command in Stata is used to generate a scatter plot for a set of data points in Stata, and the **lfit** command in Stata can be used to plot the line of best fit for the set of points. These two commands can be combined into a single command on the same graph as follows:

***graph twoway (scatter y x)(lfit y x)***

where **y** is the dependent variable and **x** is the independent variable.

Use the combined graph command above to generate a scatter plot of maternal height on the x-axis and child height on the y-axis and the line of best fit for the data.

- d. We will now estimate the slope of the line of best fit that you have drawn in part c, denoted  $\widehat{\beta}_1$ :

The **egen** command in Stata is very useful for generating special functions and extensions of variables that go beyond just the standard **gen** command. One of the functions that we can use with the **egen** command is the **mean** function, which is as follows:

$$\mathbf{egen\ y = mean(x)}$$

which generates a new variable  $y$  that takes on the average of all of the individual  $i$  values in the variable  $x$ , i.e.

$$y = \bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

- Use the **egen** and **mean** syntax above to generate  $\overline{h_m}$  and  $\overline{h_c}$ , which are variables of the mean of the mother's height and child's height, respectively.
- Use the **gen** command to generate a variable that takes the difference of each individual mother's height  $\overline{h_{m,i}}$  and  $\overline{h_m}$ , i.e.  $h_{m,i} - \overline{h_m}$ .
- Use the **gen** command to generate a variable that takes the difference of each individual child's height,  $h_{c,i}$  and  $\overline{h_c}$ , i.e.  $h_{c,i} - \overline{h_c}$ .
- Use the **gen** command to generate a variable that is the square of  $h_{m,i} - \overline{h_m}$  from (ii), i.e.  $(h_{m,i} - \overline{h_m})^2$ .
- Use the **gen** command to generate a variable that is the product of  $h_{m,i} - \overline{h_m}$  from (ii), and  $h_{c,i} - \overline{h_c}$  from (iii), i.e.  $(h_{c,i} - \overline{h_c})(h_{m,i} - \overline{h_m})$ .

Another function that we can use with the **egen** command is the **sum** function, as follows:

$$\mathbf{egen\ y = sum(x)}$$

which generates a new variable  $y$  that is the sum of all of the values in the variable  $x$ , i.e.

$$y = \sum_{i=1}^N x_i$$

- Using the **egen** and **sum** commands in Stata, generate the sum of the  $(h_{m,i} - \overline{h_m})^2$  variable from (iv), i.e.

$$\sum_{i=1}^{100} (h_{m,i} - \overline{h_m})^2$$

- Using the **egen** and **sum** commands in Stata, generate the sum of the  $(h_{c,i} - \overline{h_c})(h_{m,i} - \overline{h_m})$  variable from (v), i.e.

$$\sum_{i=1}^{100} (h_{c,i} - \overline{h_c})(h_{m,i} - \overline{h_m})$$

- Generate the variable  $\widehat{\beta}_1$ , which is the ratio of (vii) and (vi) as follows:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{100} (h_{c,i} - \overline{h_c})(h_{m,i} - \overline{h_m})}{\sum_{i=1}^{100} (h_{m,i} - \overline{h_m})^2}$$

- Summarize the variable  $\widehat{\beta}_1$ , which is your estimate for the slope of the line that you drew in part c. What is its mean? How does your mean compare to the true value of the relationship between mother's height and child height from part b. above?

- f. We will now estimate the intercept of the line of best fit that you have drawn in part c, denoted  $\hat{a}$ . Using the variables that you have generated in the previous sections, generate  $\hat{a}$  as follows:

$$\hat{a} = \overline{h_c} - \widehat{\beta}_1 \cdot \overline{h_m}$$

Summarize the variable  $\hat{a}$ , which is your estimate for the intercept of the line that you drew in part c. What is its mean? How does this estimate of  $\hat{a}$  compare with the true value of the constant in the relationship from part b. above?

- g. The Stata command **reg** runs the OLS regression of variable  $y$  on variable  $x$  as follows:

**reg y x**

Run the OLS regression of child height  $h_c$  on maternal height  $h_m$ . What is the value of the coefficient that you obtain on maternal height? What is the value of the constant? How do these values compare to the estimates of the slope and intercept that you calculated in parts e. and f. above?

- h. Provide an interpretation for this estimate  $\widehat{\beta}_1$ .
- i. Provide an interpretation for the estimated constant  $\hat{a}$ .

### Problem 2: The Effect of Smoking on Health

A researcher is interested in estimating the effect that the number of cups of coffee that a pregnant mother drinks has on her baby's health. A common measure of a child's health and birth is her weight. The researcher decides to estimate the following regression:

$$W_i = a + \beta C_i + \varepsilon_i$$

where:

$W_i$ : the weight at birth of mother  $i$ 's baby (in ounces)

$C_i$ : the daily number of cups of coffee drunk by mother  $i$  during the pregnancy

$\varepsilon_i$ : the error term

- Would you expect the estimate of  $a$  to be positive or negative? What does  $a$  represent?
- Would you expect the estimate of  $\beta$  to be positive or negative? What does  $\beta$  represent?
- Give an interpretation of  $\varepsilon_i$ . Give examples of 3 factors that may end up in  $\varepsilon_i$ .
- Suppose the following estimates are obtained from a sample of 1,388 births:

$$\widehat{W}_i = 119.77 - 0.514C_i$$

To obtain a predicted birth weight of 125 ounces, what would  $C_i$  have to be? Is this value of  $C_i$  sensible? Why or why not?