Econometrics A: Problemset 1

LAURA MAYORAL
Institute for Economic Analysis and Barcelona GSE

January 2021

Deadline: January 21th before 15:00. Please submit your answers electronically –scanned or typed, as you prefer–, to your TA, sanghyun.park@insead.edu.

1. Read Chapter 2 in Mostly Harmless Econometrics (MHE) and answer the following questions.

(1) You are interested in measuring the impact of treatment D on some outcome Y in a population. You compute the observed difference in average Y as follows:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

(a) Show that this equation can be written as the sum of two components, the average treatment effect on the treated and the selection bias. Explain in your own words the meaning of each of these terms.

(b) Using the above-mentioned decomposition, explain why randomisation makes the selection bias to be equal to zero.

(c) Also, explain why randomisation makes the average treatment effect on the treated to be equal to $E(Y_{1i}) - E(Y_{0i})$.

2. You are interested in measuring the effect of a new anti-cancer drug. In your experiment patients self-select themselves to the new treatment. After a few months you measure the average health Y of all the patients and compute the difference:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$$

(1) Do you expect the selection bias to be equal to zero or different from zero?

(2) In the latter case, do you think the selection bias would be positive or negative?

(3) Using the answers to the previous questions discuss whether $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ is a good measure of the average causal effect of the new drug or whether it overestimates or underestimates the true causal effect.

(4) Propose an alternative experiment so that you can obtain a better estimate of the average causal effect of the drug.

3. A researcher wants to assess the impact of alcohol consumption during pregnancy on newborns' weight. To that effect, she employs survey data where women declare their weekly alcohol intake. The weight of the babies is also recorded.

(1) What do you think about this procedure? do you think that selection can be an issue in this case? If your answer is affirmative, do you think that it will lead to an overestimation or to an underestimation of the effect of alcohol consumption on newborns' weight? justify your answer.

(2) Alternatively, the researcher is planning to run an experiment where pregnant women are randomly assigned to the "treatment" of interest. Describe how this have to be done. Do you think an ethics committee would approve of such an experiment?

Note: You can read https://www.theatlantic.com/health/archive/2013/08/thinking-about-pregnancy-like-an-economist/278874/here a newspaper article written by an economist on the benefits of knowing econometrics when it comes to interpreting medical advice.

4. Properties of expectations, variances and covariances

Let X_1, X_2 and X_3 be three random variables such that $E(X_1) = 2$; $E(X_2) = 3$; $E(X_3) = 0$, $Var(X_1) = 2$; $Var(X_2) = 4$; $Var(X_3) = 4$; $cov(X_1, X_2) = -1; cov(X_2, X_3) = 2; cov(X_1, X_3) = 0.4$; $E(X_2)|X_1) = 2$. (Read Handout0.pdf if you need additional information).

i) Compute:

(1) $E(3X_1 + 0.5X_2 + 4X_3 + 7)$

(2) $Var(X_1 - X_3)$

(3) $Corr(3X_1 + 2, 1/6X_3 + 2)$

5. Computer Practise

(1) Use Stata to answer the following questions

(a) Empirical papers typically start by presenting a **table of summary statistics**. This table typically includes the number of observations, mean, standard deviation, minimum and maximum values, kurtosis, skewness, among other statistics. Load in Stata the dataset mroz_ps0.dta; See Wooldridge p. 59 for a description of the data. Produce a table summarizing the main variables you'll find there.
Note: you can use different STATA commands to create tables of summary statistics that are easily exportable to other documents, see for instance tabstat or latabstat (for latex output).

(b) For reasons that we will see later on in the course, many times we'll model variables in logs. Generate a new variable (lwage) that is the log of the variable *wage*. Note: notice that $log(0) = -\infty$. Thus, when variables contain zeros, we typically add a small quantity, .1 for instance, and then compute the log. Add a label to the new variable (for instance, label the new variable: log of (wage+0.1)).

(c) Plot a scatter plot showing the relationship between lwage (Y axis) and exper (X axis). What do you observe?

(d) Add to the previous scatter the line that best fits the data. The slope of the line gives you an idea of the type of (linear) relationship between those variables. What do you see now? Is the relationship positive or negative?

(e) Add to the scatter plot above the quadratic fit (instead of the linear one) between these two variables. What do you observe now? How would you interpret this result?