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.

Since $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$
$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$ = $E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1) + E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)$
= $E(Y_{1i} - Y_{0i}|D_i = 1) + E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)$
= The average treatment effect on treated+The selection bias

$E(Y_i|D_i = 1)$ = the average Y_i among treated (can be observed)
$E(Y_i|D_i = 0)$ = the average Y_i among not treated (can be observed)
$E(Y_{1i}|D_i = 1)$ = the average Y_i among treated (can be observed)
$E(Y_{0i}|D_i = 1)$ = the average Y_i among treated if they were not treated (cannot be observed; counterfactual)
$E(Y_{1i} - Y_{0i}|D_i = 1)$ = the average treatment effect on treated
$E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)$ = the difference in Y_i between treated and not treated if both were not treated (cannot be observed)

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

Under randomization, D_i is independent of Y_{0i}.
In other words, $E(Y_{0i}|D_i = 1) = E(Y_{0i}|D_i = 0) = E(Y_{0i})$.
Thus, we have $E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0) = 0$

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

Under randomization, D_i is independent of Y_{0i} and D_i is independent of Y_{1i}.
In other words, $E(Y_{0i}|D_i = 0) = E(Y_{0i}|D_i = 1)$ and $E(Y_{1i} - Y_{0i}|D_i = 1) = E(Y_{1i} - Y_{0i})$.
$E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$
$=E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0)$
$=E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)$
$=E(Y_{1i} - Y_{0i}|D_i = 1)$
$=E(Y_{1i} - Y_{0i})$

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?

It is likely to be different from zero.

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

In the latter case, do you think the selection bias would be positive or negative? The selection bias is likely to be negative. Imagine that patients who cannot be cured by existing treatments are more likely to apply for the new treatment due to risk. In this case, participants' health conditions are likely to be worse than those who do not apply. In other words, $E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0) < 0$ (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.

It is likely to underestimate the true causal effect since the selection bias is negative.

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

Randomization. In other words, we should randomly select patients to be treated among population of the study (i.e., people with cancer) so that the expectation of the selection bias becomes zero.

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.

Since alcohol consumption during pregnancy is an stigmatised behavior, it's likely that women that drink a lot tend to underreport their actual consumption. This might lead to an overestimation of the effect of light drinking, because if heavy drinking has an effect on newborns' weight, women that drink a lot would report moderate levels of consumption and low newborns' weight.

(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?

It would be unethical to carry out such a randomised experiment, as women would be exposed to an unknown risk. This highlights the fact that experiments cannot be always carried out, sometimes due to budget constraints but also because sometimes they are unethical or unfeasible.

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

$E(3X_1 + 0.5X_2 + 4X_3 + 7) = 3E(X_1) + 0.5E(X_2) + 4E(X_3) + 7 = 3 * 2 + 0.5 * 3 + 4 * 0 + 7 = 14.5$

(2) $Var(X_1 - X_3)$

$Var(X_1 - X_3) = Var(X_1) + Var(X_3) - 2Cov(X_1, X_3) = 2 + 4 - 2 * 0.4 = 5.2$

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

$Corr(3X_1+2, X_3/6+2) = Cov(3X_1+2, X_3/6+2)/\sqrt{Var(3X_1 + 2)Var(X_3/6 + 2)}$
$= Cov(3X_1, X_3/6)/\sqrt{Var(3X_1)Var(X_3/6)}$
$= 0.5 * Cov(X_1, X_3)/(0.5 * \sqrt{Var(X_1)Var(X_3)})$
$= 0.4/\sqrt{8} = 0.2/\sqrt{2}$

6. 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).

TABLE 1. Summary Statistics

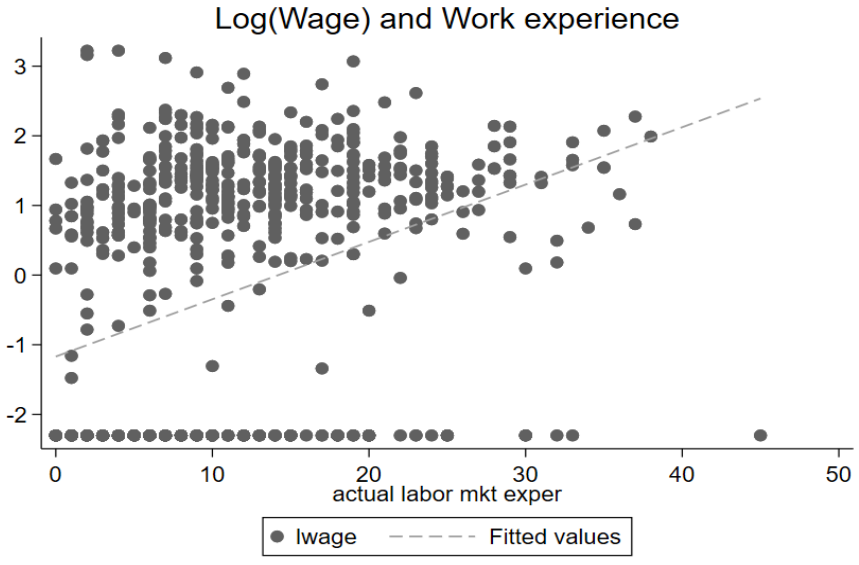
stats	**wage**	**exper**	**educ**	**age**	**kidslt6**	**kidsge6**
N	753	753	753	753	753	753
mean	2.374565	10.63081	12.28685	42.53785	.2377158	1.353254
sd	3.241829	8.06913	2.280246	8.072574	.523959	1.319874
min	0	0	5	30	0	0
max	25	45	17	60	3	8
kurtosis	15.79665	3.70137	3.744087	1.981077	8.254322	3.809829
skewness	2.777771	.9605118	.021034	.150879	2.309519	.9077226

(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)).

```
gen lwage=log(wage+0.1)
label variable lwage "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?
It is difficult to see the relationship through scatter plot without fitted line.

(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?
According to the fitted line, experience and wage have a positive relationship.



(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?

According to the quadratic fit, there is inverted U shaped relationship between experience and wage. In other words, the wage increases with experience up to certain point, but, beyond that point, wage decreases with experience.