# Econometrics A

Xinyu Liu

January 15, 2021

## 1.

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

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

**1. 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.**

The first crucial step is to express $Y_i$ as follows:

$$Y_i = \begin{cases} Y_{1i} & D_i = 1 \\ Y_{0i} & D_i = 0 \end{cases}$$

where $Y_{1i}, Y_{0i}$ represent the potential outcomes if treated or controlled for individual $i$. Therefore the expression from the question can be rewritten in the following:

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\ \text{subtract and add: } &= \underbrace{E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)}_{\text{treatment effect}} + \underbrace{E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)}_{\text{selection bias}} \end{aligned}$$

What I did is to subtract and add a counterfactual term $E(Y_{0i}|D_i = 1)$, which stands for the expectation of the treated individual $i$, had the person not been treated. The treatment effect essentially require us to compare the post-treatment with the counterfactual for the treated individual. However, what we can observed are only sample means corresponding to $E(Y_{1i}|D_i = 1)$ and $E(Y_{0i}|D_i = 0)$. If we use them as the estimate of treatment effect, we will bring in the selection bias, which is the second component in the preceding expression: $E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)$. If the sample is not randomized selected, then $E(Y_{0i}|D_i = 1)$ can be different from $E(Y_{0i}|D_i = 0)$, which implies that if the treated individuals were not treated, their average status are still different from the control group average.

**2. Using the above-mentioned decomposition, explain why randomisation makes the selection bias to be equal to zero.**

If there is proper randomization, then the treatment $D_i$ is independent from the potential outcome $Y_{0i}$, therefore I can drop out the conditional expectation:

$$E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0) = E(Y_{0i}) - E(Y_{0i}) = 0$$

Thus the selection bias will be equal to zero once there is randomization.

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

This is another consequence of independence. Randomization makes $D_i$ to be independent from $Y_{0i}$ and $Y_{1i}$, therefore:

$$\begin{aligned} E(Y_i|D_i = 1) - E(Y_i|D_i = 0) &= E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) \\ &= E(Y_{1i}) - E(Y_{0i}) \end{aligned}$$

## 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?**

I expect that the selection bias will be different from zero, because patients in this experiment self-select themselves to the new treatment. It's reasonable to believe that patients who choose to take the drug have different characteristics than patients who opt out.

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

Further, I tend to believe that the selection bias will be negative. As more severe patients will be more willing to try new drugs, and these patients on average have worse health conditions than those who are in the control.

**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.**

Firstly, I argue that it is not a good measure of the average treatment effect of the new drug, as it doesn't satisfy the assumption that there is no selection bias. More specifically, the treatment group is heterogeneous from the control group. Secondly, I argue that this estimate will likely to underestimate the effect of the new drug, because the selection bias is negative. I.e. suppose the treated patients were untreated, there average health condition will be lower than those in the control, corresponding to their worse health status from the first place.

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

This biggest issue with previous experiment is that patients are not randomly assigned to the two groups, in other words, treatment is endogenous. To obtain a better estimate of ATE, I propose that all patients be randomly assigned to two groups, the control group receives placebos and the treatment group receives the new drug. Although this will give a better estimate for the drug effect, but I am personally concerned that this arrangement is more or less unethical.

## 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.**

I believe there is selection bias issue unresolved using the aforementioned method. There are two major problems with the selection. First, the sample who fill in the survey may not necessarily represent the population well. Second, even if we assume that the survey randomly select people, still it is not a experiment, let along meeting the requirement of randomization. The women participated in the survey choose their alcohol consumption endogenously. I.e. their alcohol consumption may be related to some unobserved variables such as their tolerance of alcohol, which will also affect the weight of the new born. Therefore, it is not likely that this survey will correctly estimate the impact.

Furthermore, I think the impact can be negatively estimated. I believe on average, alcohol is bad

for health therefore can potentially negatively impact the development of the new born, the more the women drinks, the lighter the baby may becomes. However, when conducting the survey, people who are capable of drinking report higher volume of alcohol consumption, but because the are better at digesting the alcohol, their baby may be less affected by this habit, reducing the magnitude of the estimated impact.

**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?**

The easiest way will be to randomly assign pregnant women into groups of different level of alcohol consumption, participants in each group on average should be homogeneous thus resolving the selection bias issue. Again, this may not be an ethical experiment because alcohol consumption will potentially do harm to the women and babies. Participants have no reason to obey the assignment and risk their health.

## 4. Properties of expectations, variances and covariances

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

$$\begin{aligned} E(3X_1 + 0.5X_2 + 4X_3 + 7) &= 3E(X_1) + 0.5E(X_2) + 4E(X_3) + 7 \\ &= 3 * 2 + 0.5 * 3 + 4 * 0 + 7 = 14.5 \end{aligned}$$

**2.**$Var(X_1 - X_3)$

$$\begin{aligned} Var(X_1 - X_3) &= Var(X_1) + Var(X_3) - 2Cov(X_1, X_3) \\ &= 2 + 4 - 2 * 0.4 = 5.2 \end{aligned}$$

**3.**$Corr(3X_1 + 2, \frac{1}{6}X_3 + 2)$

$$\begin{aligned} Corr(3X_1 + 2, \frac{1}{6}X_3 + 2) &= \frac{Cov(3X_1 + 2, \frac{1}{6}X_3 + 2)}{\sqrt{Var(3X_1 + 2)Var(\frac{1}{6}X_3 + 2)}} \\ &= \frac{Cov(3X_1, \frac{1}{6}X_3)}{\sqrt{Var(3X_1)Var(\frac{1}{6}X_3)}} \\ &= \frac{Cov(X_1, X_3)}{\sqrt{Var(X_1)Var(X_3)}} \\ &= \frac{0.4}{\sqrt{2 * 4}} = \frac{\sqrt{2}}{10} \approx 0.14 \end{aligned}$$

## 5. Computer Practise

**1. 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).**

The main variables used in the exmaple are: 'exper','educ','age','kidslt6','kidsge6'. I give their summary statistics as follows:

Table 1: Summary statistics

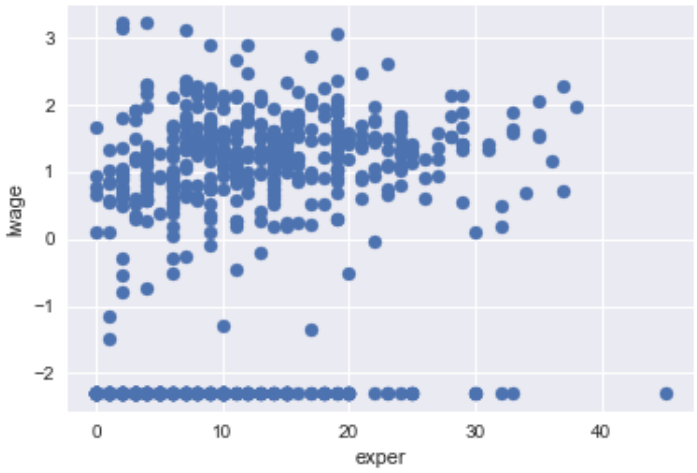
| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| wage | 753.0 | 2.4 | 3.2 | 0.0 | 0.0 | 1.6 | 3.8 | 25.0 |
| exper | 753.0 | 10.6 | 8.1 | 0.0 | 4.0 | 9.0 | 15.0 | 45.0 |
| educ | 753.0 | 12.3 | 2.3 | 5.0 | 12.0 | 12.0 | 13.0 | 17.0 |
| age | 753.0 | 42.5 | 8.1 | 30.0 | 36.0 | 43.0 | 49.0 | 60.0 |
| kidslt6 | 753.0 | 0.2 | 0.5 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 |
| kidsge6 | 753.0 | 1.4 | 1.3 | 0.0 | 0.0 | 1.0 | 2.0 | 8.0 |

**2. 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)).**

I process the data as described by the question, see the result from next question (see the code in Appendix for details).

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

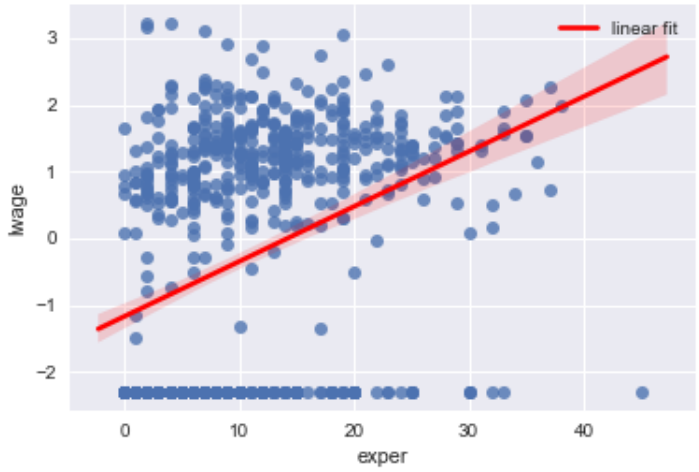
Figure 1: Scatter plot of lwage with exper




It seems that on average, the lwage is positively correlated with actual market experience. Also note that there are quite a few observation reporting very low income, this may impact the credibility of the conclusion.

**4. 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?**

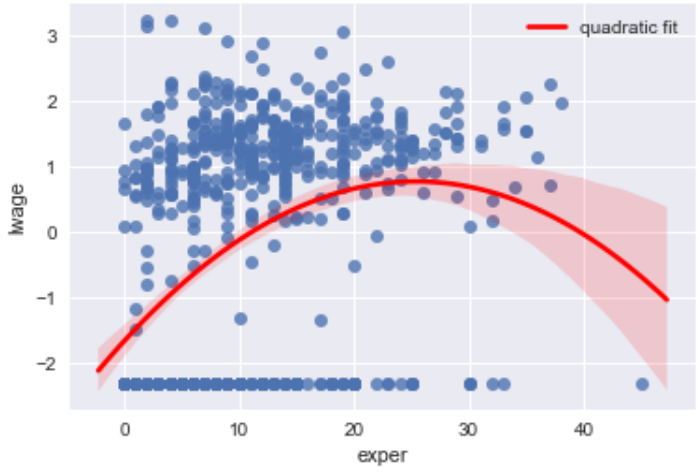
Figure 2: Scatter plot with fitted linear line of lwage with exper




The fitted line shows a positive correlation, which aligns with previous observation. This is intuitive because more experienced workers have higher human capital, therefore they are likely to be paid more.

**5. 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?**

Figure 3: Scatter plot with fitted quadratic curve of lwage with exper




Now the quadratic line adds a little more insights to the relationship: as experience gets to a certain level, lwage seems to peak and bend downwards. To be more clear, this part is less robust compared to the linear fit, and people can potentially argue that this is partially due to outliers (see the point at the right corner). However, this is also plausible in the sense that people with too much experience suffer from aging issues and will be applied a discount as employers tend to favor more energetic workers.

# Appendix

```
import pandas as pd
import numpy as np
import datetime as dt
import matplotlib.pyplot as plt

plt.style.use('seaborn')
df = pd.read_stata('mroz_ps0.dta')
print(df[['wage','exper','educ','age','kidslt6','kidsge6']].describe().T.applymap('{:,.1f}'.format).to_latex())
df['lwage']= np.log(df['wage']+0.1)
plt.scatter(df['exper'], df['lwage'])
plt.xlabel('exper')
plt.ylabel('lwage')
plt.savefig("p5qc")

plt.scatter(df['exper'], df['lwage'])
plt.xlabel('exper')
plt.ylabel('lwage')
plt.savefig("p5qc")

import seaborn as sns
ax = sns.regplot(x="exper", y="lwage", data=df,line_kws= {'label':'linear fit','color':'r'})
plt.legend()
plt.savefig("p5qd")

ax = sns.regplot(x="exper", y="lwage", data=df,order=2, line_kws= {'label':'quadratic fit','color':'r'})
plt.legend()
plt.savefig("p5qe")
```