Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

10 August, 2018

Assignment 1

You can already start working on Assignment 1

Exercise 1 of the assignment only requires knowledge of OLS
(stuff you've learned in EMET2007 and/or STAT2008)

Deadline: 29 August at 12:00pm (noon, mid-day)

Reminder: my deadlines are *very* sharp!
(If you submit at 12:01pm you will receive a mark of zero!)

No extensions given under any circumstances!

Note: I do not offer any help on solving the assignment!

Computer tutorials

Starting today, Simon Mishricky will be teaching the tutorials

Simon is friendly and clever, should be fun and educational

Roadmap

At the end of the last lecture, we looked at three important statements:

- A statistical analysis is **internally valid** if statistical inferences about causal effects are valid for the population being studied.
- We aim to have estimates that are internally valid.
- The OLS estimator is internally valid if it is unbiased and consistent.
- The OLS estimator is unbiased and consistent only under OLS Assumption 1.

Important question to ask:
When is OLS Assumption 1 not satisfied?

Roadmap

Assumption (OLS Assumption 1)

The error term u_i is conditionally mean independent (CMI) of X_i

$$E[u_i|X_i] = E[u_i] = \mu_u.$$

Assumption 1 says that X_i is not informative about the expected value of u_i

This would be guaranteed if X_i and u_i were independent

When would they be independent?
For example, if X_i and/or u_i are purely random

But are they?

OLS Assumption 1 is also sometimes called the *Exogeneity Assumption*

Whenever OLS Assumption 1 doesn't hold, we deal with the problem of endogeneity

There are essentially 5 reasons for why the exogeneity assumption might fail

1. Omitted variable bias
2. Wrong functional form
3. Errors-in-variables bias
4. Sample selection bias
5. Simultaneous causality bias

Let's talk about these in turn…

Roadmap

Consider the following multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

According to this model, we should regress Y on X_1 and X_2 to obtain OLS estimates of β_1 and β_2

But suppose we only have data on X_{1i} and Y_i

We know (for some reason) that the variable X_{2i} should also be included in the model but the data set does not contain it

Therefore, all we can do is regress Y_i on X_{1i}

How does this affect the OLS estimator for β_1?

To find out, rewrite the model as follows

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \\ &= \beta_0 + \beta_1 X_{1i} + (\beta_2 X_{2i} + u_i) \\ &= \beta_0 + \beta_1 X_{1i} + v_i, \end{aligned}$$

where $v_i := \beta_2 X_{2i} + u_i$ denotes a new error term

In general, $v_i \neq u_i$ (because $\beta_2 \neq 0$)

The last equation now looks like a simple linear regression model in which the error term is called v_i

Given $Y_i = \beta_0 + \beta_1 X_{1i} + v_i$, the OLS estimator of β_1 is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1)}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

The first factor in the numerator can be expanded like

$$Y_i - \bar{Y} = \beta_1(X_{1i} - \bar{X}_1) + (v_i - \bar{v})$$

and plug in to get (after some simplifications)

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(v_i - \bar{v})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

Typically, the argument now would be that X_{1i} and error term v_i are uncorrelated so that $\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(v_i - \bar{v})$ is almost zero

Big problem here:
this particular error term is not uncorrelated with X_{1i}

Recall that v_i is not just any random error

It also contains X_{2i} because $v_i := \beta_2 X_{2i} + u_i$

It has two components

- u_i which is purely random and uncorrelated with X_{1i}
- X_{2i} which is an omitted regressor which could well be correlated with X_{1i}

If X_{1i} and X_{2i} are correlated with each other than the error term v_i will be correlated with X_{1i}

This will lead to bias in the OLS estimate $\hat{\beta}_1$

Going back to our previous result

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(v_i - \bar{v})}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2}$$

If we are interested in the expected value of $\hat{\beta}_1$, $\mathrm{E}[\hat{\beta}_1 | X_{1i}, X_{2i}]$, the second term on the rhs will not be equal to zero

Instead, we get ...

$$
\begin{aligned}
&\mathrm{E}\left[\hat{\beta}_1 | X_{1i}, X_{2i}\right] \\
&= \beta_1 + \frac{\sum_{i=1}^{n} \mathrm{E}\left[(X_{1i} - \bar{X}_1)(v_i - \bar{v}) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n} \mathrm{E}\left[(X_{1i} - \bar{X}_1)(\beta_2 (X_{2i} - \bar{X}_2) + (u_i - \bar{u})) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&\qquad + \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1) \mathrm{E}\left[(u_i - \bar{u}) | X_{1i}, X_{2i}\right]}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{n} (X_{1i} - \bar{X}_1)^2} \\
&\simeq \beta_1 + \beta_2 \frac{\mathrm{E}[(X_{1i} - \mu_{X_1})(X_{2i} - \mu_{X_2})]}{\mathrm{Var}(X_{1i})} \\
&= \beta_1 + \beta_2 \frac{\mathrm{Cov}(X_{1i}, X_{2i})}{\mathrm{Var}(X_{1i})},
\end{aligned}
$$

The second equality holds because
$v_i := \beta_2 X_{2i} + u_i$ and $\bar{v} = \beta_2 \bar{X}_2 + \bar{u}$

The third equality holds because X_{1i} and X_{2i} can be treated as constants

The fourth equality holds because of exogeneity
(we will learn that this is OLS Assumption 1 in the multiple linear regression model next week)

To get the asymptotic result just replace sample averages by population averages

This shows that the expected value of $\hat{\beta}_1$ is not equal to β_1

The OLS estimator $\hat{\beta}_1$ is therefore not unbiased

What is the bias equal to?

This bias term is $\text{Cov}(X_{1i}, X_{2i})/\text{Var}(X_{1i})$

Intuitively, this bias is proportional to the covariance between X_{1i} and X_{2i} and inversely proportional to the variance of X_{1i}

The omitted variables bias could be positive or negative:
it depends on the sign of the covariance between X_{1i} and X_{2i}

If you do not like the mathematics of it, maybe you prefer to understand it intuitively

If you omit X_{2i} from the estimation, then the estimate of β_1 will be biased

The reason for this is that the estimator $\hat{\beta}_1$ is doing two jobs at the same time:

- it captures the direct effect of X_{1i} on Y
  (this is what you *want* to capture; it's the effect β_1)
- but it also captures the indirect effect that X_{2i} has through its covariance with X_{1i}
  (this creates the bias)

Roadmap

Recall the generic regression response function from last week:

$$Y_i = f(X_i, u_i)$$

In the linear model, we simply force $f(\cdot)$ to be linear in all arguments

$$f(X_i, u_i) = \beta_0 + \beta_1 X_i + u_i$$

(this is easily generalized to a model with additional regressors)

But what if $f(\cdot)$ isn't actually linear?

For example, what would the OLS estimator $\hat{\beta}_1$ estimate, if $f(\cdot)$ was a higher order polynomial in X_i?

Answering this question requires some complicated math, I'll just give you the answer: $\hat{\beta}_1$ would be biased and inconsistent!

In other words, when the actual relationship between Y_i and X_i is nonlinear, then the OLS estimator is not internally valid

Example: You estimate the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

and afterwards an oracle tells you that the true association between X_i and Y_i is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}^8 + \beta_3 \log(X_{2i}) + u_i$$

Your estimate of β_1 will be biased

Remedy: if you know the actual functional form, then you could just throw X_{2i}^8 and $\log(X_{2i})$ into the regression

The problem in practice is that you never really now the correct functional form

If you don't know the actual functional specification of $f(\cdot)$ then all you can do is cross your fingers!

Turns out, that's what most people do when they run regressions!

Another problem associated with functional form concerns the dependent variable Y_i

Even if $f(\cdot)$ was indeed linear, you could still have a misspecified functional form if Y_i is a categorical or a binary variable

We will study this after the midterm break (book chapter 11)

Roadmap

Elements of the data $X_{1i}, \ldots, X_{ki}, Y_i$ may be measured imprecisely

How does this create problems?

Example: Causal effect of heights on earnings

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i are hourly wages and X_i are heights

(for simplicity, we ignore other regressors)

Heights may be measured imprecisely

That would be an example of X_i being measured with error

Instead of

- X_i (actual/true height)

  we observe

- $\tilde{X}_i$ (reported height)

Most generally, we permit $X_i \neq \tilde{X}_i$ (for some i)

The relationship between the two is thought of as

$$\tilde{X}_i := X_i + w_i$$

In words: the *reported* height is equal to the *true* height plus some unobserved error term

Reported height is a noisy measurement of true height

Why would this cause any problems?

At least two ways to think about the discrepancy between $\tilde{X}_i$ and X_i:

- $\tilde{X}_i$ deviates from X_i for systematic reasons
  Example: small persons tend to overstate their heights while tall persons tend to report accurately
- $\tilde{X}_i$ deviates from X_i completely at random
  (referred to as *classical measurement error*)

Which one of these two will result in bias OLS estimates of the causal effect of height on earnings?

It seems obvious that systematic misreporting will bias the OLS estimate

In the given example (small persons tend to exaggerate heights), what is the bias?

Would OLS using reported heights result in an overestimate or an underestimate of the actual causal effect of heights on earnings?

What does not seem so obvious is that even the second type of measurement error results in biased OLS estimates

Completely random measurement error creates so-called *classical measurement error bias*

The formal result is

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2} = \beta_1 \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$$

In words: OLS is inconsistent

The bias term is $-\beta_1 \frac{\sigma_w^2}{\sigma_X^2 + \sigma_w^2}$

Recall that w_i was the random discrepancy between the true X_i and the reported $\tilde{X}_i$

The size of the bias depends on the relative variances of X_i and w_i

In the extreme case that the variance of w_i is very large, the OLS estimator converges in probability to zero, irrespective of what the true population parameter β_1 is equal to!

This demonstrates that random measurement error can be a big problem

On the other hand, if $\sigma_w^2 = 0$, $\hat{\beta}_1 \xrightarrow{p} \beta_1$ and life is good again

So far we've discussed how measurement error in X_i leads estimates that are not internally valid

But what if the measurement error is in Y_i instead?

Does this create problems?

Roadmap

Data are often missing

There are three ways to look at it:

1. Data are missing at random
2. Data are missing based on the value of one or more X
3. Data are missing based in part on the value of Y

Which of these lead to bias?

Data missing at random

Suppose I randomly sample 100 Canberrans and ask them to fill out a survey asking for their earnings and height

I want to regress earnings on heights

There's one catch:
my six year old daughter destroys the survey responses of 30 people (six year olds do such things for no good reason)

My daughter undertook her destructive efforts randomly

Luckily, my daughter did not introduce any bias:
her behavior is equivalent to me having sampled only 70 people in the first place

Effectively, I still have a random sample

It's just a bit smaller

This increases the standard errors of the OLS estimates

But OLS Assumption 1 still applies

Data missing based on X_i

Suppose I randomly sample 100 Canberrans and ask them to fill out a survey asking for their earnings and height

I still want to regress earnings on heights

There's another catch:
I only let people participate if they are at least 175cm tall

This still does not result in bias because people are still sampled randomly (as long as they are tall enough)

Obviously, I won't be able to learn anything about people who are smaller than 175cm

But my OLS estimates will be internally valid for the subset of the population that is at least 175cm tall

Data missing based on Y_i

Again, suppose I randomly sample 100 Canberrans and ask them to fill out a survey asking for their earnings and height

My daughter does not destroy any survey responses and I don't only ask tall people for a response

Still, there will be a missing data problem

How so?

Missing data problem:
I can only observe earnings for people who have a job

When I run a regression of earnings on heights, I have to exclude people who have no reported earnings because they do not have a job

Typical example from the past: house wifes don't have earnings

The problem is that whenever a person has no reported earnings, we cannot assume that this is a random event

There may be a systematic reason for why the person does not have any earnings

To demonstrate the problem, consider a slightly more modern example:

A couple's decision to have a baby

Both are young professionals (meaning: both working)

They decide that one of them will take a three year work leave to tend to the baby

Who should take the work leave?

Purely based on opportunity cost, the one with the lower earnings should stay at home

This example suggests that the subset of wage earners may be biased: people who earn relatively more tend to stay in the labor force

Conversely, people who earn relatively little tend to drop out (at least temporarily)

This is an example of a non-random sample

People *endogenously* select themselves into (and out of) the sample

The sample is not representative of the entire population

This is an example of (endogenous) *sample selection*

The resulting bias is called sample selection bias

Roadmap

So far we have always presumed that X_i causes Y_i

But what if the causation goes the other way?

Example: children's height and parent's earnings

You randomly sample 500 parents who have grown-up children; you conduct a survey asking for

- parents earnings
- childrens' heights

You regress earnings on heights

Is this sensible?

Regressing earnings on heights is informative about the covariance between the two

But you should not give this a *causal* interpretation! (Obvious?)

Unlike in the example in which we regressed a person's earnings on own height, a child's height cannot be causal for the earnings of the parent (I hope you agree!)

Instead, economic research has provided ample evidence that the children of well-earning parents grow taller
(because of better quality "production" inputs like nutritious food and good education)

This was an example of causality from Y_i to X_i

But how about causality in both directions at the same time?

Let's say I am interested in estimating the causal effect of lecture attendance on course outcomes

The primitive research question is:
Does lecture attendance improve grades?

Suppose I can accurately measure both lecture attendance and grades

What problem do you see with the following regression:

$$\text{Grades} = \beta_0 + \beta_1 \text{Attendance} + u_i$$

Is the OLS estimator $\hat{\beta}_1$ internally valid for the causal effect β_1?

Problem Solving Exercises

1. Properly define the **individual causal effect** in the more general model with k independent variables.
2. What does the individual causal effect in the model with k independent variables boil down to, once you impose linearity?
3. Derive the classical measurement error bias.
   (Hint: Frame the problem of classical measurement error as an omitted variable bias problem and apply the ovb results.)