Econometrics 120C: Threats to the Validity of a Regression Study

Kaspar Wüthrich

References: Stock and Watson Ch 6 and 9, EVH Section F

This lecture **will be recorded** and made available asynchronously via Canvas.

Introduction

- It is hard to resists the temptation of using regression analysis to estimate causal effects based on the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

  where u_i contains all other possible variables that determine Y_i.
- The biggest hurdle to causal inference is that variables in u_i are possibly correlated with X_i.
- Note that such correlation means that the OLS assumption $E[u_i \mid X_i] = 0$ is incorrect.
- Here we look at different scenarios, all of which render OLS inconsistent.

Threats to internal validity

There are many possible reasons for why X_i could be correlated with u_i

1. Omitted variable bias (OVB)
2. Measurement error
3. Simultaneous causality
4. Sample selection bias
5. (Functional form misspecification)
6. ...

Each of these, if present, leads to a violation of the key assumption that $E[u_i \mid X_i] = 0$. The consequence is that the OLS estimator $\hat{\beta}_1$ is generally inconsistent for the true parameter of interest β_1.

OVB: general principle

Suppose the true model is

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

where we assume that $E[u_i \mid X_{1i}, X_{2i}] = 0$. When X_{2i} is omitted, we have

$$Y_i = \beta_0 + \beta_1 X_{1i} + e_i, \text{ where } e_i = \beta_2 X_{2i} + u_i \tag{3}$$

In this case, the probability limit of the OLS estimator based on the "short" model (3) is

$$\begin{aligned}
\hat{\beta}_1^{short} \xrightarrow{p} \frac{Cov(Y_i, X_{1i})}{Var(X_{1i})} &= \frac{Cov(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, X_{1i})}{Var(X_{1i})} \\
&= \frac{\beta_1 Var(X_{1i}) + \beta_2 Cov(X_{2i}, X_{1i})}{Var(X_{1i})} \\
&= \beta_1 + \beta_2 \frac{Cov(X_{2i}, X_{1i})}{Var(X_{1i})} \neq \beta_1
\end{aligned}$$

OVB: general principle

Note that $Cov(X_{2i}, X_{1i})/Var(X_{1i})$ is nothing else than the population regression coefficient γ_1 in the following auxiliary regression model

$$X_{2i} = \gamma_0 + \gamma_1 X_{1i} + r_i.$$

Therefore, we often say that:

$$\text{short} = \text{long} + \text{effect of omitted} \times \text{regression of omitted on included}$$

OVB: discussion

- The OLS estimator $\hat{\beta}_1^{short}$ based on the "short" regression model (3) will generally not be consistent for the true β_1.
- The bias can be written as

$$\beta_2 \frac{Cov(X_{2i}, X_{1i})}{Var(X_{1i})} = \beta_2 Corr(X_{1i}, X_{2i}) \frac{\sqrt{Var(X_{2i})}}{\sqrt{Var(X_{1i})}}.$$

- Therefore, the sign of the bias depends on the correlation between omitted (X_{2i}) and included (X_{1i})
- This is a very useful insight since it allows us to gauge the sign of the bias of OLS even if we do not observe X_{2i} in our data set.
- Note that the bias is zero (i) if $Corr(X_{1i}, X_{2i}) = 0$ and/or if (ii) $\beta_2 = 0$. How can you interpret these two cases?

OVB: schooling example

Labor economists are very often interested in estimating returns to education. We usually think about wages as being determined by ability and schooling (abstracting from other characteristics):

$$\underbrace{wage_i}_{=Y_i} = \beta_0 + \beta_1 \underbrace{schooling_i}_{=X_{1i}} + \beta_2 \underbrace{ability_i}_{=X_{2i}} + u_i$$

Unfortunately, ability is very hard to measure and almost always unobserved (i.e., not in our data set). Thus, we can only estimate the short model:

$$wage_i = \beta_0 + \beta_1 schooling_i + e_i$$

The omitted variable bias tells us that

$$\hat{\beta}_1^{short} \xrightarrow{p} \beta_1 + \beta_2 Corr(schooling_i, ability_i) \frac{\sqrt{Var(ability_i)}}{\sqrt{Var(schooling_i)}}$$

One would expect that $\beta_2 > 0$ and $Corr(schooling_i, ability_i) > 0$. Therefore, $\hat{\beta}_1^{short}$ overestimates the wage returns.

Measurement error: setup and example

- Suppose we want to estimate

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

  but instead of the true X_i we only observe a noisy measurement $\tilde{X}_i$
- Example:
  - Y_i: indicator for lung cancer
  - X_i: true cigarette consumption
  - $\tilde{X}_i$: self-reported cigarette consumption

Measurement error: theory

- Written in terms of $\tilde{X}_i$, the population regression equation becomes

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i \\ &= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (X_i - \tilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \tilde{X}_i + v_i \end{aligned}$$

  where $v_i = [\beta_1 (X_i - \tilde{X}_i) + u_i]$.
- Thus, the population regression model written in terms of $\tilde{X}_i$ has an error that contains $(X_i - \tilde{X}_i)$. If $(X_i - \tilde{X}_i)$ is correlated with $\tilde{X}_i$ then $\hat{\beta}_1$ will be inconsistent.
- In general, the size and direction of the bias depend on the correlation of $\tilde{X}_i$ and $(X_i - \tilde{X}_i)$ and this correlation depends, in turn, on the specific nature of the measurement error.

Measurement error: example (classical measurement error)

- For example, suppose that $\tilde{X}_i = X_i + w_i$, where the measurement error w_i is purely random (i.e., independent of u_i and X_i) with mean zero and variance σ_w^2.
- Even in this "ideal" case, some algebra (show this!) shows that

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

- Because $\frac{\sigma_X^2}{\sigma_X^2+\sigma_w^2} \leq 1$, $\hat{\beta}_1$ will be biased towards 0.
- Extreme case 1: if measurement error is so large that no information about X_i remains, i.e., $\sigma_w^2 \to \infty$, then $\hat{\beta}_1 \xrightarrow{p} 0$
- Extreme case 2: if there is no measurement error, i.e., $\sigma_w^2 = 0$, then $\hat{\beta}_1 \xrightarrow{p} \beta_1$

Simultaneity: theory

- So far, we have assumed that causality runs from X_i to Y_i. But what if causality also runs from Y_i to X_i?
- If so, causality runs backwards as well as forward, that is, there is simultaneous causality. This will again lead to inconsistency of OLS.
- Consider a simple setup with two variables X_i and Y_i. Accordingly, there are two equations

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4}$$
$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{5}$$

  Simultaneity leads to correlation between X_i and the error term u_i in (4).
- To see this, imagine that u_i is negative, which decreases Y_i. However, this lower value of Y_i affects the value of X_i through equation (5), and if γ_1 is positive, a low value of Y_i will lead to a low value of X_i. Thus, if γ_1 is positive, X_i and u_i will be positively correlated.

Simultaneity: example

Let us revisit the police spending and crime example from the previous set of slides. In this case

$$\begin{aligned} crime_i &= \beta_0 + \beta_1 spending_i + u_i \\ spending_i &= \gamma_0 + \gamma_1 crime_i + v_i \end{aligned}$$

where we would expect $\beta_1 < 0$ and $\gamma_1 > 0$.

Sample selection bias: definition

- Sample selection occurs when the availability of the data is influenced by a selection process that is related to the value of the dependent variable.
- This selection process can introduce correlation between X_i and u_i.
- Sample selection generally leads to inconsistency of the OLS estimator.

Sample selection bias: example (wage regression)

- A sample selection problem occurs because only individuals who have jobs have wages (by definition).
- The factors that determine whether someone has a job are similar to the factors that determine how much that person earns when employed.
- Thus, the fact that someone has a job suggests that, all else equal, u_i for that person is positive.
- As a consequence, the simple fact that someone has a job, and thus appears in the data set, provides information that u_i is positive, at least on average, and could be correlated with regressors.