Econometric Analysis of Cross Section and Panel Data

Lecture 6: Endogeneity

Zhian Hu

Central University of Finance and Economics

Fall 2024

This Lecture

- Hansen (2022): Chapter 12

Overview

- We say that there is endogeneity in the linear model

$$Y = X'\beta + e$$

if β is the parameter of interest **but**

$$\mathbb{E}[Xe] \neq 0$$

- To distinguish from the regression and projection models, we will call it a structural equation and β a structural parameter.

Overview

- Endogeneity cannot happen if the coefficient is defined by linear projection.

$$\begin{aligned} Y &= X'\beta^* + e^* \\ \mathbb{E}\left[Xe^*\right] &= 0 \end{aligned}$$

- Under endogeneity, the projection coefficient β^* does not equal the structural parameter β.

$$\begin{aligned} \beta^* &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] \\ &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[X\left(X'\beta + e\right)\right] \\ &= \beta + \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[Xe] \neq \beta \end{aligned}$$

Overview

- **Endogeneity implies that the least squares estimator is inconsistent for the structural parameter.**
- Under i.i.d. sampling, least squares is consistent for the projection coefficient.

$$\widehat{\beta} \xrightarrow[p]{} \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] = \beta^* \neq \beta$$

- The inconsistency of least squares is typically referred to as **endogeneity bias** or **estimation bias due to endogeneity**.
  - This is an imperfect label as the actual issue is inconsistency, not bias.

Endogeneity bias is often caused by

- Measurement error in the regressor
- Simultaneous equations bias
- Choice variables as regressors / missing variables

Example: Measurement error in the regressor

- Suppose that (Y, Z) are joint random variables, $\mathbb{E}[Y \mid Z] = Z'\beta$ is linear.

$$Y = Z'\beta + e$$

  - Z is not observed.
  - Instead we observe $X = Z + u$ where u is a $k \times 1$ measurement error,
  - u is **independent** of e and Z.
- The model $X = Z + u$ with Z and u independent and $\mathbb{E}[u] = 0$ is known as **classical measurement error**.
  - This means that X is a noisy but unbiased measure of Z.

Example: Measurement error in the regressor

- By substitution we can express Y as a function of the observed variable X.

$$Y = Z'\beta + e = (X - u)'\beta + e = X'\beta + v$$

  where $v = e - u'\beta$.
- This means that (Y, X) satisfy the linear equation

$$Y = X'\beta + v$$

Example: Measurement error in the regressor

- The error v is not a projection error.

$$\mathbb{E}[X\nu] = \mathbb{E}\left[(Z+u)\left(e - u'\beta\right)\right] = -\mathbb{E}\left[uu'\right]\beta \neq 0$$

  if $\beta \neq 0$ and $\mathbb{E}\left[uu'\right] \neq 0$. Then least squares estimation will be inconsistent.

- We can calculate the form of the projection coefficient (which is consistently estimated by least squares). For simplicity suppose that $k = 1$. We find

$$\beta^* = \beta + \frac{\mathbb{E}[X\nu]}{\mathbb{E}\left[X^2\right]} = \beta\left(1 - \frac{\mathbb{E}\left[u^2\right]}{\mathbb{E}\left[X^2\right]}\right)$$

- Since $\mathbb{E}\left[u^2\right]/\mathbb{E}\left[X^2\right] < 1$ the projection coefficient shrinks the structural parameter β towards zero. This is called **measurement error bias** or **attenuation bias**.

Example: Simultaneous equations bias

- The variables Q and P (quantity and price) are determined jointly by the demand equation
$$Q = -\beta_1 P + e_1$$
and the supply equation
$$Q = \beta_2 P + e_2$$
- Assume that $e = (e_1, e_2)'$ satisfies $\mathbb{E}[e] = 0$ and $\mathbb{E}\left[ee'\right] = I_2$ (the latter for simplicity).
- The question is: if we regress Q on P, what happens?

Example: Simultaneous equations bias

- It is helpful to solve for Q and P in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix} \begin{pmatrix} Q \\ P \end{pmatrix} = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix}$$

- So

$$\begin{aligned} \begin{pmatrix} Q \\ P \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \left(\frac{1}{\beta_1 + \beta_2} \right) \\ &= \begin{pmatrix} (\beta_2 e_1 + \beta_1 e_2) / (\beta_1 + \beta_2) \\ (e_1 - e_2) / (\beta_1 + \beta_2) \end{pmatrix} \end{aligned}$$

- The projection of Q on P yields $Q = \beta^* P + e^*$ with $\mathbb{E}[Pe^*] = 0$ and the projection coefficient is

$$\beta^* = \frac{\mathbb{E}[PQ]}{\mathbb{E}[P^2]} = \frac{\beta_2 - \beta_1}{2} \neq \beta_1 \quad \text{or} \quad \beta_2$$

Example: Choice variables as regressors

- Take the classic wage equation

$$\log(wage) = \beta education + e$$

  with β the average causal effect of education on wages.
- If wages are affected by unobserved ability, and individuals with high ability self-select into higher education, then e contains unobserved ability, so education and e will be positively correlated. Hence education is endogenous.
- The positive correlation means that the linear projection coefficient β^* will be upward biased relative to the structural coefficient β.
- Thus least squares (which is estimating the projection coefficient) will tend to over-estimate the causal effect of education on wages.