Introductory Econometrics

Lecture 19: Linear regression without strong exogeneity

Instructor: Ma, Jun

Renmin University of China

November 8, 2021

Strong exogeneity and the conditional expectation function (CEF)

- Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + U_i.$$

- When the errors are strongly exogenous, i.e. $\mathrm{E}\left[U_i \mid X_i\right] = 0$, the linear regression model defines the CEF of Y conditional on X:

$$\begin{aligned}
CEF_Y\left(X_i\right) &= \mathrm{E}\left[Y_i \mid X_i\right] \\
&= \mathrm{E}\left[\beta_0 + \beta_1 X_i + U_i \mid X_i\right] \\
&= \beta_0 + \beta_1 X_i + \mathrm{E}\left[U_i \mid X_i\right] \\
&= \beta_0 + \beta_1 X_i
\end{aligned}$$

Weak exogeneity

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i, \\ \mathrm{E}[U_i] &= 0 \end{aligned}$$

- Suppose the errors are only weakly exogenous:

$$\mathrm{E}[U_i X_i] = 0.$$

- In this case,

$$CEF_Y(X_i) \neq \beta_0 + \beta_1 X_i.$$

- Question: What does the econometrician estimates when he runs a linear regression and the regressors are *not* strongly exogenous?

Linear regression as a misspecified CEF

- Suppose that
$$\mathrm{E}\left[Y_i \mid X_i\right] = g\left(X_i\right),$$
where g is some unknown nonlinear function. Thus, the true CEF is $g\left(X_i\right) \neq \beta_0 + \beta_1 X_i$.
- Define
$$V_i = Y_i - \mathrm{E}\left[Y_i \mid X_i\right],$$
so we can write the true model as
$$\begin{aligned} Y_i &= g\left(X_i\right) + V_i, \\ \mathrm{E}\left[V_i \mid X_i\right] &= 0. \end{aligned}$$

- Write

$$
\begin{aligned}
Y_i &= g(X_i) + V_i \\
&= \beta_0 + \beta_1 X_i - \beta_0 - \beta_1 X_i + g(X_i) + V_i
\end{aligned}
$$

or

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$

where

$$U_i = V_i + g(X_i) - \beta_0 - \beta_1 X_i.$$

- Can we find β_0 and β_1 so that $\mathrm{E}[U_i] = 0$ and $\mathrm{E}[X_i U_i] = 0$? If yes, how can we interpret such β_0 and β_1?

$$U_i = V_i + g(X_i) - \beta_0 - \beta_1 X_i.$$

- Note that

$$\begin{aligned} \mathrm{E}[U_i] &= \mathrm{E}[V_i + g(X_i) - \beta_0 - \beta_1 X_i] \\ &= \mathrm{E}[V_i] + \mathrm{E}[g(X_i) - \beta_0 - \beta_1 X_i] \\ &= \mathrm{E}[g(X_i) - \beta_0 - \beta_1 X_i], \end{aligned}$$

  and

$$\begin{aligned} \mathrm{E}[U_i X_i] &= \mathrm{E}[(V_i + g(X_i) - \beta_0 - \beta_1 X_i) X_i] \\ &= \mathrm{E}[V_i X_i] + \mathrm{E}[(g(X_i) - \beta_0 - \beta_1 X_i) X_i] \\ &= \mathrm{E}[(g(X_i) - \beta_0 - \beta_1 X_i) X_i]. \end{aligned}$$

- Thus, to have $\mathrm{E}[U_i] = \mathrm{E}[U_i X_i] = 0$, we need to find β_0 and β_1 such that

$$\begin{aligned} \mathrm{E}[g(X_i) - \beta_0 - \beta_1 X_i] &= 0 \\ \mathrm{E}[(g(X_i) - \beta_0 - \beta_1 X_i) X_i] &= 0. \end{aligned}$$

Linear approximation of the CEF

- Consider the following approximation problem:

$$\min_{b_0, b_1} \mathrm{E}\left[\left(g\left(X_i\right) - b_0 - b_1 X_i\right)^2\right].$$

- We are approximating the CEF by linear functions.
- Among the linear functions, we are looking for the best linear approximation in the mean squared error (MSE) sense.

$$\min_{b_0,b_1} MSE\left(b_0, b_1\right),$$

$$MSE\left(b_0, b_1\right) = \mathrm{E}\left[\left(g\left(X_i\right) - b_0 - b_1 X_i\right)^2\right].$$

- Let β_0 and β_1 denote the solution: $\left(\beta_0, \beta_1\right) = \arg\min_{b_0,b_1} MSE\left(b_0, b_1\right)$.
- The first-order conditions are:

$$\begin{aligned}
\frac{\partial MSE\left(\beta_0, \beta_1\right)}{\partial b_0} &= -2 \cdot \mathrm{E}\left[g\left(X_i\right) - \beta_0 - \beta_1 X_i\right] = 0. \\
\frac{\partial MSE\left(\beta_0, \beta_1\right)}{\partial b_1} &= -2 \cdot \mathrm{E}\left[\left(g\left(X_i\right) - \beta_0 - \beta_1 X_i\right) X_i\right] = 0.
\end{aligned}$$

Linear regression as the best linear approximation of the CEF

- We have

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + U_i, \\
U_i &= V_i + g\left(X_i\right) - \beta_0 - \beta_1 X_i.
\end{aligned}
$$

- With $(\beta_0, \beta_1) = \arg\min_{b_0, b_1} \mathrm{E}\left[\left(g\left(X_i\right) - b_0 - b_1 X_i\right)^2\right]$,

$$
\mathrm{E}\left[U_i\right] = 0 \text{ and } \mathrm{E}\left[U_i X_i\right] = 0.
$$

- Thus, the linear regression model gives us the best linear approximation of the CEF (in the MSE sense).

Misspecification and heteroskedasticity

- We have

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i, \\ U_i &= V_i + g(X_i) - \beta_0 - \beta_1 X_i. \end{aligned}$$

- Suppose that the "true" error V_i is homoskedastic: $\mathrm{E}\left[V_i^2 \mid X_i\right] = \sigma_V^2$ for all X_i.
- U_i is heteroskedastic if $g(X_i) \neq \beta_0 + \beta_1 X$:

$$\begin{aligned} \mathrm{E}\left[U_i^2 \mid X_i\right] &= \mathrm{E}\left[(V_i + g(X_i) - \beta_0 - \beta_1 X_i)^2 \mid X_i\right] \\ &= \mathrm{E}\left[V_i^2 + (g(X_i) - \beta_0 - \beta_1 X)^2 + \right. \\ &\quad \left. +2V_i (g(X_i) - \beta_0 - \beta_1 X_i) \mid X_i\right] \\ &= \sigma_V^2 + (g(X_i) - \beta_0 - \beta_1 X)^2 . \end{aligned}$$