Lecture 18: Misspecification

Economics 326 — Introduction to Econometrics II

Vadim Marmer, UBC

Strong exogeneity and the CEF

- Consider the linear regression model

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

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

$$\begin{aligned}\mathrm{CEF}_Y(X_i) &\equiv \mathrm{E}[Y_i \mid X_i] \\ &= \mathrm{E}[\beta_0 + \beta_1 X_i + U_i \mid X_i] \\ &= \beta_0 + \beta_1 X_i + \mathrm{E}[U_i \mid X_i] \\ &= \beta_0 + \beta_1 X_i.\end{aligned}$$

Weak exogeneity

- Consider the linear regression model with $\mathrm{E}[U_i] = 0$:

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

- Suppose the errors are only **weakly exogenous**:

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

- In this case,

$$\mathrm{CEF}_Y(X_i) \neq \beta_0 + \beta_1 X_i.$$

- **Question:** What does the econometrician estimate when running a linear regression and the regressors are *not* strongly exogenous?

Misspecified CEF and the regression error

- Suppose that

$$\mathrm{E}[Y_i \mid X_i] = g(X_i),$$

  where g is some unknown *nonlinear* function. Thus, the **true** CEF is $g(X_i) \neq \beta_0 + \beta_1 X_i$.
- Define $\varepsilon_i = Y_i - \mathrm{E}[Y_i \mid X_i]$, so the true model is $Y_i = g(X_i) + \varepsilon_i$ with $\mathrm{E}[\varepsilon_i \mid X_i] = 0$.
- Adding and subtracting $\beta_0 + \beta_1 X_i$:

$$\begin{aligned}Y_i &= g(X_i) + \varepsilon_i \\ &= \beta_0 + \beta_1 X_i \\ &\quad + \underbrace{g(X_i) - \beta_0 - \beta_1 X_i}_{\text{specification error}} + \varepsilon_i\end{aligned}$$

- Rearranging, $Y_i = \beta_0 + \beta_1 X_i + U_i$, where the regression error U_i combines the true error and the **specification error**:

$$U_i = \varepsilon_i + \underbrace{g(X_i) - \beta_0 - \beta_1 X_i}_{\text{specification error}}.$$

- Can we find β_0 and β_1 so that X_i is uncorrelated with U_i, i.e., $\mathrm{E}[U_i] = 0$ and $\mathrm{E}[X_i U_i] = 0$?

Conditions for weak exogeneity

- Denote the specification error by $\Delta(X_i) = g(X_i) - \beta_0 - \beta_1 X_i$, so $U_i = \varepsilon_i + \Delta(X_i)$.
- Since $\mathrm{E}[\varepsilon_i \mid X_i] = 0$, the law of iterated expectations gives

$$\begin{aligned} \mathrm{E}[\varepsilon_i] &= \mathrm{E}[\mathrm{E}[\varepsilon_i \mid X_i]] = 0, \\ \mathrm{E}[\varepsilon_i X_i] &= \mathrm{E}[X_i \mathrm{E}[\varepsilon_i \mid X_i]] = 0. \end{aligned}$$

- Therefore

$$\begin{aligned} \mathrm{E}[U_i] &= \mathrm{E}[\varepsilon_i] + \mathrm{E}[\Delta(X_i)] = \mathrm{E}[\Delta(X_i)], \\ \mathrm{E}[U_i X_i] &= \mathrm{E}[\varepsilon_i X_i] + \mathrm{E}[\Delta(X_i) X_i] = \mathrm{E}[\Delta(X_i) X_i]. \end{aligned}$$

- The conditions $\mathrm{E}[U_i] = 0$ and $\mathrm{E}[U_i X_i] = 0$ reduce to conditions on the **specification error** alone:

$$\begin{aligned} \mathrm{E}[\Delta(X_i)] &= 0, \\ \mathrm{E}[\Delta(X_i) X_i] &= 0. \end{aligned}$$

  That is, we need β_0 and β_1 such that the specification error has **mean zero** and is **uncorrelated with** X_i.

Linear approximation of the CEF

- Consider the following approximation problem:

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

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

Regression as best linear approximation

- Let $(\beta_0, \beta_1) = \arg\min_{b_0, b_1} \mathrm{MSE}(b_0, b_1)$, where

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

- The first-order conditions are:

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

- OLS chooses β_0 and β_1 so that X_i is **uncorrelated with the specification error** $\Delta(X_i) = g(X_i) - \beta_0 - \beta_1 X_i$, yielding the best linear approximation of the CEF in the MSE sense.
- Since $U_i = \varepsilon_i + \Delta(X_i)$, this also gives

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

Misspecification and heteroskedasticity

- Recall $U_i = \varepsilon_i + \Delta(X_i)$, where $\Delta(X_i) = g(X_i) - \beta_0 - \beta_1 X_i$ is the specification error.
- Suppose the true error ε_i is homoskedastic: $\mathrm{E}\left[\varepsilon_i^2 \mid X_i\right] = \sigma_\varepsilon^2$ for all X_i.
- When the specification error is nonzero, U_i is **heteroskedastic**:

$$\begin{aligned}
\mathrm{E}\left[U_i^2 \mid X_i\right] &= \mathrm{E}\left[\left(\varepsilon_i + \Delta(X_i)\right)^2 \mid X_i\right] \\
&= \mathrm{E}\left[\varepsilon_i^2 \mid X_i\right] + \Delta(X_i)^2 + 2\,\Delta(X_i)\,\mathrm{E}\left[\varepsilon_i \mid X_i\right] \\
&= \sigma_\varepsilon^2 + \Delta(X_i)^2 .
\end{aligned}$$