

Economics 326  
Methods of Empirical Research in Economics  
Lecture 19: Linear regression without strong  
exogeneity

Vadim Marmer  
University of British Columbia

May 5, 2010

## 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.  $E(U_i|X_i) = 0$ , the linear regression model defines the **CEF** of  $Y$  conditional on  $X$ :

## 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.  $E(U_i|X_i) = 0$ , the linear regression model defines the **CEF** of  $Y$  conditional on  $X$ :

$$CEF_Y(X_i) \equiv E(Y_i|X_i)$$

## 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.  $E(U_i|X_i) = 0$ , the linear regression model defines the **CEF** of  $Y$  conditional on  $X$ :

$$\begin{aligned} CEF_Y(X_i) &\equiv E(Y_i|X_i) \\ &= E(\beta_0 + \beta_1 X_i + U_i|X_i) \end{aligned}$$

## 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.  $E(U_i|X_i) = 0$ , the linear regression model defines the **CEF** of  $Y$  conditional on  $X$ :

$$\begin{aligned} \text{CEF}_Y(X_i) &\equiv E(Y_i|X_i) \\ &= E(\beta_0 + \beta_1 X_i + U_i|X_i) \\ &= \beta_0 + \beta_1 X_i + E(U_i|X_i) \end{aligned}$$

## 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.  $E(U_i|X_i) = 0$ , the linear regression model defines the **CEF** of  $Y$  conditional on  $X$ :

$$\begin{aligned} CEF_Y(X_i) &\equiv E(Y_i|X_i) \\ &= E(\beta_0 + \beta_1 X_i + U_i|X_i) \\ &= \beta_0 + \beta_1 X_i + E(U_i|X_i) \\ &= \beta_0 + \beta_1 X_i. \end{aligned}$$

## Weak exogeneity

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$
$$EU_i = 0$$

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

$$E(U_i X_i) = 0.$$

## Weak exogeneity

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$
$$EU_i = 0$$

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

$$E(U_i | X_i) = 0.$$

- ▶ In this case,

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

## Weak exogeneity

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$
$$EU_i = 0$$

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

$$E(U_i | X_i) = 0.$$

- ▶ In this case,

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

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

## Linear regression as a misspecified CEF

- ▶ Suppose that

$$E(Y_i|X_i) = g(X_i),$$

where  $g$  is some unknown *nonlinear* function. Thus, the **true** CEF is  $g(X_i) \neq \beta_0 + \beta_1 X_i$ .

## Linear regression as a misspecified CEF

- ▶ Suppose that

$$E(Y_i|X_i) = g(X_i),$$

where  $g$  is some unknown *nonlinear* function. Thus, the **true** CEF is  $g(X_i) \neq \beta_0 + \beta_1 X_i$ .

- ▶ Define

$$V_i = Y_i - E(Y_i|X_i),$$

so we can write the true model as

$$\begin{aligned} Y_i &= g(X_i) + V_i, \\ E(V_i|X_i) &= 0. \end{aligned}$$

## Linear regression as a misspecified CEF

- ▶ Write

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

## Linear regression as a misspecified CEF

- ▶ 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

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

## Linear regression as a misspecified CEF

- ▶ 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  $\beta_0$  and  $\beta_1$  so that  $E(U_i) = 0$  and  $E(X_i U_i) = 0$ ?  
If yes, how can we interpret such  $\beta_0$  and  $\beta_1$ ?

## Linear regression as a misspecified CEF

---

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

# Linear regression as a misspecified CEF

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

► Note that

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

# Linear regression as a misspecified CEF

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

► Note that

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

# Linear regression as a misspecified CEF

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

► Note that

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

# Linear regression as a misspecified CEF

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

► Note that

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

and

$$E(U_i X_i) = E[(V_i + g(X_i) - \beta_0 - \beta_1 X_i) X_i]$$

# Linear regression as a misspecified CEF

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

► Note that

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

and

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

# Linear regression as a misspecified CEF

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

► Note that

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

and

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

# Linear regression as a misspecified CEF

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

- Note that

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

and

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

- Thus, to have  $E(U_i) = E(U_i X_i) = 0$ , we need to find  $\beta_0$  and  $\beta_1$  such that

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

## Linear approximation of the CEF

- ▶ Consider the following approximation problem:

$$\min_{b_0, b_1} E (g (X_i) - b_0 - b_1 X_i)^2 .$$

## Linear approximation of the CEF

- ▶ Consider the following approximation problem:

$$\min_{b_0, b_1} E (g (X_i) - b_0 - b_1 X_i)^2 .$$

- ▶ We are approximating the CEF by linear functions.

## Linear approximation of the CEF

- ▶ Consider the following approximation problem:

$$\min_{b_0, b_1} E (g (X_i) - b_0 - b_1 X_i)^2 .$$

- ▶ 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.

## Linear approximation of the CEF

---

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

$$MSE(b_0, b_1) = E(g(X_i) - b_0 - b_1 X_i)^2.$$

## Linear approximation of the CEF

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

$$MSE(b_0, b_1) = E(g(X_i) - b_0 - b_1 X_i)^2.$$

- ▶ Let  $\beta_0$  and  $\beta_1$  denote the solution:  
 $(\beta_0, \beta_1) = \arg \min_{b_0, b_1} MSE(b_0, b_1).$

## Linear approximation of the CEF

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

$$MSE(b_0, b_1) = E(g(X_i) - b_0 - b_1 X_i)^2.$$

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

## Linear approximation of the CEF

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

$$MSE(b_0, b_1) = E(g(X_i) - b_0 - b_1 X_i)^2.$$

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

$$\frac{\partial MSE(\beta_0, \beta_1)}{\partial b_0} = -2E(g(X_i) - \beta_0 - \beta_1 X_i) = 0.$$

## Linear approximation of the CEF

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

$$MSE(b_0, b_1) = E(g(X_i) - b_0 - b_1 X_i)^2.$$

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

$$\frac{\partial MSE(\beta_0, \beta_1)}{\partial b_0} = -2E(g(X_i) - \beta_0 - \beta_1 X_i) = 0.$$

$$\frac{\partial MSE(\beta_0, \beta_1)}{\partial b_1} = -2E[(g(X_i) - \beta_0 - \beta_1 X_i) X_i] = 0.$$

# Linear regression as the best linear approximation of the CEF

- ▶ We have

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

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

# Linear regression as the best linear approximation of the CEF

- ▶ We have

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

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

- ▶ With  $(\beta_0, \beta_1) = \arg \min_{b_0, b_1} E (g(X_i) - b_0 - b_1 X_i)^2$ ,

$$E U_i = 0 \text{ and } E (U_i X_i) = 0.$$

# Linear regression as the best linear approximation of the CEF

- ▶ We have

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

- ▶ With  $(\beta_0, \beta_1) = \arg \min_{b_0, b_1} E (g(X_i) - b_0 - b_1 X_i)^2$ ,

$$E U_i = 0 \text{ and } E (U_i X_i) = 0.$$

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

## Misspecification and heteroskedasticity

- ▶ We have

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

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

## Misspecification and heteroskedasticity

- ▶ We have

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

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

- ▶ Suppose that the "true" error  $V_i$  is homoskedastic:  
 $E(V_i^2 | X_i) = \sigma_V^2$  for all  $X_i$ .

## 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:  
 $E(V_i^2 | X_i) = \sigma_V^2$  for all  $X_i$ .
- ▶  $U_i$  is **heteroskedastic** if  $g(X_i) \neq \beta_0 + \beta_1 X_i$ :

$$E(U_i^2 | X_i) = E\left[(V_i + g(X_i) - \beta_0 - \beta_1 X_i)^2 | X_i\right]$$

## 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:  
 $E(V_i^2 | X_i) = \sigma_V^2$  for all  $X_i$ .
- ▶  $U_i$  is **heteroskedastic** if  $g(X_i) \neq \beta_0 + \beta_1 X_i$ :

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

## 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:  
 $E(V_i^2 | X_i) = \sigma_V^2$  for all  $X_i$ .
- ▶  $U_i$  is **heteroskedastic** if  $g(X_i) \neq \beta_0 + \beta_1 X_i$ :

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