

Violating the exogeneity assumption of the linear regression model: When does it happen?

Econometrics for minor Finance, Lecture 7

Tunga Kantarcı, Fall 2025

Linear regression model: Model assumption: Error is exogenous

When we introduced the linear regression model, we assumed
exogeneity:

$$E[u_i | x_i] = 0$$

Linear regression model: Model assumption: Error is exogenous: Implications

$$E[u_i | x_i] = 0$$

has a number of implications.

Linear regression model: Model assumption: Error is exogenous: Implications: Implication one

First, if

$$E[u_i | x_i] = 0$$

then

$$\begin{aligned} E[u_i x_i] &= E_{x_i} [E[u_i x_i | x_i]] \\ &= E_{x_i} [x_i E[u_i | x_i]] \\ &= 0 \end{aligned}$$

using the LIE in the first equality.

Linear regression model: Model assumption: Error is exogenous: Implications: Implication two

Second, if

$$E[u_i | x_i] = 0$$

then

$$\begin{aligned} E[u_i] &= E_{x_i} [E[u_i | x_i]] \\ &= 0. \end{aligned}$$

by the LIE in the first equality. This says that if the average of u_i at all slices of the population determined by the values of x_i is zero, then the average of these zero conditional means must also be zero.

Linear regression model: Model assumption: Error is exogenous: Implications: Implication three

Third, if

$$E[u_i | x_i] = 0$$

then

$$\begin{aligned}\text{Cov}[u_i, x_i] &= E[u_i x_i] - E[u_i] E[x_i] \\ &= 0 - 0 E[x_i] \\ &= 0\end{aligned}$$

using the above results. That is, the exogeneity condition just says that u_i are x_i are uncorrelated.

Linear regression model: Model assumption: Error is endogenous

In this lecture we violate the exogeneity assumption. That is, we have

$$E[u_i | x_i] \neq 0$$

Linear regression model: Model assumption: Error is endogenous: When does this happen?

When does

$$E[u_i | x_i] \neq 0$$

happen?

Linear regression model: Model assumption: Error is endogenous: The case of omitted variable

Assume that the true linear model in the population is

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

Assume that

$$E[u_i | x_{1i}] = 0$$

and

$$E[u_i | x_{2i}] = 0$$

Hence, the model is correctly specified.

Linear regression model: Model assumption: Error is endogenous: The case of omitted variable

Assume also, however, that we actually do not observe x_{2i} so it enters the error, and hence we have

$$y_i = x_{1i}\beta_1 + u_i^*$$

and

$$u_i^* = x_{2i}\beta_2 + u_i$$

Linear regression model: Model assumption: Error is endogenous: The case of omitted variable

Taking the expectation conditional on x_{1i} of

$$u_i^* = x_{2i}\beta_2 + u_i$$

we have

$$\begin{aligned} E[u_i^* | x_{1i}] &= E[x_{2i}\beta_2 | x_{1i}] + E[u_i | x_{1i}] \\ &= \beta_2 E[x_{2i} | x_{1i}] \end{aligned}$$

Linear regression model: Model assumption: Error is endogenous: The case of omitted variable

If

$$\beta_2 \neq 0$$

and

$$E[x_{2i} | x_{1i}] \neq 0$$

we have

$$E[u_i^* | x_{1i}] = \beta_2 E[x_{2i} | x_{1i}] \neq 0$$

- $\beta_2 \neq 0$ means that x_{2i} should enter the true model.
- $E[x_{2i} | x_{1i}] \neq 0$ means that x_{1i} and x_{2i} are correlated.

Linear regression model: Model assumption: Error is endogenous: The case of omitted variable

$$E[u_i^* | x_{1i}] \neq 0$$

means that in the regression

$$y_i = x_{1i}\beta_1 + u_i^*$$

x_{1i} is correlated with u_i^* due to leaving x_{2i} in u_i^* . The **exogeneity assumption is violated for u_i^*** .

Linear regression model: Model assumption: Error is endogenous: The case of measurement error

Consider the linear model

$$y_i = \beta x_i^* + u_i$$

Suppose x_i^* is the true variable but we do not observe it. What we observe is x_i , a noisy version of x_i^* with unobserved measurement error ω_i so that

$$x_i = x_i^* + \omega_i$$

Since we observe only x_i , replace x_i^* in the model to obtain

$$y_i = \beta x_i - \beta \omega_i + u_i$$

Define

$$u_i^* := -\beta \omega_i + u_i$$

so that we have

$$y_i = \beta x_i + u_i^*$$

x_i is correlated with u_i^* due to ω_i . The exogeneity assumption is violated for u_i^* .

Linear regression model: Model assumption: Error is endogenous: The case of simultaneity

Consider the **simultaneous equations model** given by

$$y_{1i} = \alpha_1 y_{2i} + \beta_1 z_{1i} + u_{1i}$$

$$y_{2i} = \alpha_2 y_{1i} + \beta_2 z_{2i} + u_{2i}$$

In each equation the constant is ignored for simplicity. Assume that

$$E[u_{1i} \mid z_{1i}, z_{2i}] = 0$$

$$E[u_{2i} \mid z_{1i}, z_{2i}] = 0$$

which imply

$$E[u_{1i}] = 0$$

$$E[u_{2i}] = 0$$

Hence, z_{1i} and z_{2i} are uncorrelated with u_{1i} and u_{2i} . Suppose that the **interest lies in estimating α_1** in the first equation.

Linear regression model: Model assumption: Error is endogenous: The case of simultaneity

Solve the two equations for y_{2i} , in terms of z_{1i} , z_{2i} , u_{1i} , and u_{2i} .
First, replace y_{1i} in the equation for y_{2i} , and then solve for y_{2i} as

$$\begin{aligned}y_{2i} &= \alpha_2 y_{1i} + \beta_2 z_{2i} + u_{2i} \\&= \alpha_2 (\alpha_1 y_{2i} + \beta_1 z_{1i} + u_{1i}) + \beta_2 z_{2i} + u_{2i} \\&= \alpha_1 \alpha_2 y_{2i} + \beta_1 \alpha_2 z_{1i} + \alpha_2 u_{1i} + \beta_2 z_{2i} + u_{2i} \\(1 - \alpha_1 \alpha_2) y_{2i} &= \beta_1 \alpha_2 z_{1i} + \beta_2 z_{2i} + \alpha_2 u_{1i} + u_{2i} \\y_{2i} &= z_{1i} \frac{\beta_1 \alpha_2}{1 - \alpha_1 \alpha_2} + z_{2i} \frac{\beta_2}{1 - \alpha_1 \alpha_2} + u_{1i} \frac{\alpha_2}{1 - \alpha_1 \alpha_2} \\&\quad + u_{2i} \frac{1}{1 - \alpha_1 \alpha_2}\end{aligned}$$

assuming that $\alpha_1 \alpha_2 \neq 1$.

Linear regression model: Model assumption: Error is endogenous: The case of simultaneity

The parameter of interest was α_1 in the equation

$$y_{1i} = \alpha_1 y_{2i} + \beta_1 z_{1i} + u_{1i}$$

and we have just shown that

$$y_{2i} = z_{1i} \frac{\beta_1 \alpha_2}{1 - \alpha_1 \alpha_2} + z_{2i} \frac{\beta_2}{1 - \alpha_1 \alpha_2} + u_{1i} \frac{\alpha_2}{1 - \alpha_1 \alpha_2} + u_{2i} \frac{1}{1 - \alpha_1 \alpha_2}$$

Note that we need

$$E[y_{2i} u_{1i}] = 0$$

to hold to estimate α_1 with no bias. Does it hold?

Linear regression model: Model assumption: Error is endogenous: The case of simultaneity

$$y_{2i} = z_{1i} \frac{\beta_1 \alpha_2}{1 - \alpha_1 \alpha_2} + z_{2i} \frac{\beta_2}{1 - \alpha_1 \alpha_2} + u_{1i} \frac{\alpha_2}{1 - \alpha_1 \alpha_2} + u_{2i} \frac{1}{1 - \alpha_1 \alpha_2}$$

Multiply both sides with u_{1i} , take expectations, and use the earlier assumption that $E[z_{1i}u_{1i}] = 0$ and $E[z_{2i}u_{1i}] = 0$ to obtain

$$E[y_{2i}u_{1i}] = E[u_{1i}u_{1i}] \frac{\alpha_2}{1 - \alpha_1 \alpha_2} + E[u_{2i}u_{1i}] \frac{1}{1 - \alpha_1 \alpha_2}$$

If

$$\alpha_2 \neq 0, E[u_{2i}u_{1i}] = 0$$

or

$$\alpha_2 = 0, E[u_{2i}u_{1i}] \neq 0$$

we have

$$E[y_{2i}u_{1i}] \neq 0$$

Linear regression model: Model assumption: Error is endogenous: The case of simultaneity

Go back to the simultaneous equations model

$$y_{1i} = \alpha_1 y_{2i} + \beta_1 z_{1i} + u_{1i}$$

$$y_{2i} = \alpha_2 y_{1i} + \beta_2 z_{2i} + u_{2i}$$

We have just shown that

$$E[y_{2i}u_{1i}] \neq 0$$

meaning that, in the first equation, y_{2i} is correlated with u_{1i} due to simultaneity. The **exogeneity assumption is violated for u_{1i}** .