Violating the exogeneity assumption of the linear regression model: Implications for the sampling distribution of the OLS estimator

Econometrics for minor Finance, Lecture 7

Tunga Kantarcı, Fall 2025

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

In an earlier lecture we showed

$$\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x\right] = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, \mathrm{E}\left[u_i \mid x\right]}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

If we have endogeneity, for whatever reason, that is if

$$\mathrm{E}\left[u_i \mid x_i\right] \neq 0$$

the OLS estimator

$$\hat{\beta}$$

is biased since the second term of the sum does not disappear.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

The preceding slide proves theoretically that if the exogeneity assumption is violated, the OLS estimator is biased. We can also demonstrate this bias using simulation. We studied three cases that lead to endogeneity. Let us consider the omitted variable case to study how it affects the sampling distribution of the OLS estimator.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

Consider the linear model

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

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

Suppose we do not observe x_{2i} so that it enters the error and 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: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

x_{2i} is omitted from the model, and we end up with endogenity:

$$\mathrm{E}\left[u_i^* \mid x_{1i}\right] \neq 0$$

in the regression model

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

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

It is easy to imagine that this has an implication for the sampling distribution of

$$\hat{\beta}_1$$

as the OLS estimator of

$$\beta_1$$

as the population coefficient of

$$x_{1i}$$

Let's check the mean of this sampling distribution.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

Regress y, the true model, only on x_1, which is not what the true model asks us to do. In this case the OLS estimator is

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n x_{1i} y_i}{\sum_{i=1}^n x_{1i}^2} \\
&= \frac{\sum_{i=1}^n x_{1i}(x_{1i}\beta_1 + x_{2i}\beta_2 + u_i)}{\sum_{i=1}^n x_{1i}^2} \\
&= \frac{\beta_1 \sum\limits_{i=1}^n x_{1i}^2 + \beta_2 \sum_{i=1}^n x_{1i}x_{2i} + \sum_{i=1}^n x_{1i}u_i}{\sum_{i=1}^n x_{1i}^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2} + \frac{\sum_{i=1}^n x_{1i}u_i}{\sum_{i=1}^n x_{1i}^2}
\end{aligned}
$$

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

$$\hat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2} + \frac{\sum_{i=1}^{n} x_{1i} u_i}{\sum_{i=1}^{n} x_{1i}^2}$$

Take the expectation conditional on regressors:

$$\begin{aligned}
\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x_1, x_2\right] &= \mathrm{E}\left[\beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2} + \frac{\sum_{i=1}^{n} x_{1i} u_i}{\sum_{i=1}^{n} x_{1i}^2} \,\middle|\, x_1, x_2\right] \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2} + E\left[\frac{\sum_{i=1}^{n} x_{1i} u_i}{\sum_{i=1}^{n} x_{1i}^2} \,\middle|\, x_1, x_2\right]
\end{aligned}$$

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

$$
\begin{aligned}
\mathrm{E}\left[\left.\frac{\sum_{i=1}^{n} x_{1i} u_i}{\sum_{i=1}^{n} x_{1i}^2}\,\right|\, x_1, x_2\right] &= \frac{1}{\sum_{i=1}^{n} x_{1i}^2}\mathrm{E}\left[\left.\sum_{i=1}^{n} x_{1i} u_i \,\right|\, x_1, x_2\right] \\
&= \frac{1}{\sum_{i=1}^{n} x_{1i}^2}\sum_{i=1}^{n} x_{1i}\mathrm{E}\left[u_i \mid x_1, x_2\right] \\
&= 0
\end{aligned}
$$

if

$$\mathrm{E}[u_i \mid x_1, x_2] = 0$$

that if the exogeneity assumption holds.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

We obtain

$$\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x_1, x_2\right] = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2}$$

This shows that if we regress y on x_1 alone, but the true model contains x_2, the bias in

$$\hat{\beta}_1$$

is

- β_2 times
- a term capturing the linear association between x_1 and x_2 in the sample.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

$$\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x_1, x_2\right] = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2}$$

In two cases the estimator is unbiased. First, if

$$\beta_2 = 0$$

meaning that x_2 has no effect if it enters the true model.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

$$\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x_1, x_2\right] = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2}$$

Second, if

$$\frac{\sum_{i=1}^{n} x_{1i} x_{2i}}{\sum_{i=1}^{n} x_{1i}^2} = 0$$

meaning that there is no correlation between x_1 and x_2 in the sample. Realize that the stated expression is the OLS estimate of the coefficient of x_1 from the regression of x_2 on x_1.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

If these two conditions do not hold, the OLS estimator is subject to what we call the omitted variable bias. The statement

$$\mathrm{E}\left[\hat{\beta}_1 \,\middle|\, x_1, x_2\right] = \beta_1 + \beta_2 \frac{\sum\limits_{i=1}^{n} x_{1i} x_{2i}}{\sum\limits_{i=1}^{n} x_{1i}^2}$$

is the omitted variable bias formula.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

Let's demonstrate this bias using simulation.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased

Suppose that we do not observe x_{2i} so that it enters the error. The model becomes

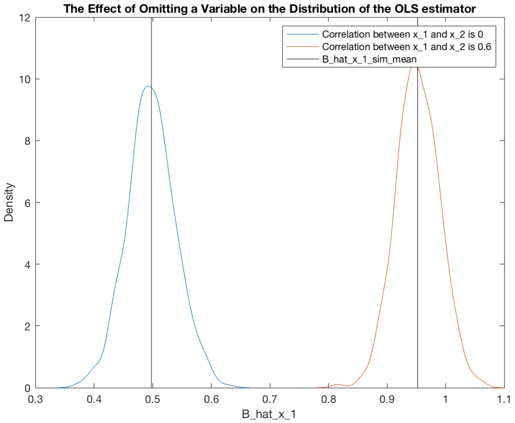
$$y_i = \beta_1 x_{1i} + u_i^*$$

where

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

Assume that the true value of β_1 is 0.5. Consider two cases. In the first case, the correlation between the two regressors is 0. In the second case, it is 0.6. Using Monte Carlo simulation, let's check the sampling distribution of $\hat{\beta}_1$ in these two cases.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased



Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

Regress wage on educ but ignore exper because it is, say, unobserved.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

```
. regress wage educ

      Source |       SS           df       MS      Number of obs   =       997
-------------+----------------------------------   F(1, 995)       =    251.46
       Model |  7842.35455         1  7842.35455   Prob > F        =    0.0000
    Residual |  31031.0745       995  31.1870095   R-squared       =    0.2017
-------------+----------------------------------   Adj R-squared   =    0.2009
       Total |   38873.429       996  39.0295472   Root MSE        =    5.5845

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.135645   .0716154    15.86   0.000     .9951106     1.27618
       _cons |  -4.860424   .9679821    -5.02   0.000    -6.759944   -2.960903
------------------------------------------------------------------------------
```

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

Regress wage on educ and exper, and observe that

$$\hat{\beta}_{educ}$$

increases. This suggests that

$$\hat{\beta}_{educ}$$

has downward bias when exper is ignored in the regression. How do we reach this conclusion?

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

```
. regress wage educ exper

      Source |       SS           df       MS      Number of obs   =       997
-------------+----------------------------------   F(2, 994)       =    172.32
       Model |  10008.3629         2  5004.18147   Prob > F        =    0.0000
    Residual |  28865.0661       994  29.0393019   R-squared       =    0.2575
-------------+----------------------------------   Adj R-squared   =    0.2560
       Total |   38873.429       996  39.0295472   Root MSE        =    5.3888

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.246932   .0702966    17.74   0.000     1.108985    1.384879
       exper |   .1327808   .0153744     8.64   0.000     .1026108    .1629509
       _cons |  -8.833768   1.041212    -8.48   0.000    -10.87699   -6.790542
------------------------------------------------------------------------------
```

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

In the regression we have ignored *exper*. We suspect that

$$\hat{\beta}_{educ}$$

is biased. That is, we suspect that

$$\hat{\beta}_{educ}$$

would change if we control for *exper* in the regression. Do you expect

$$\hat{\beta}_{educ}$$

to have an upward or downward bias?

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

Use the omitted variable bias formula to form an expectation:

$$\mathrm{E}\left[\hat{\beta}_{educ} \,\middle|\, educ, exper\right] = \beta_{educ} + \beta_{exper} \frac{\sum_{i=1}^{n} educ_i \; exper_i}{\sum_{i=1}^{n} educ_i^2}$$

We would expect effect of exper on wage, that is,

$$\beta_{exper}$$

to be positive, and the effect of exper on educ, that is,

$$\frac{\sum_{i=1}^{n} educ_i \; exper_i}{\sum_{i=1}^{n} educ_i^2}$$

to be negative. Therefore, $\hat{\beta}_{educ}$ should have downward bias when we ignore *exper* in the true regression.

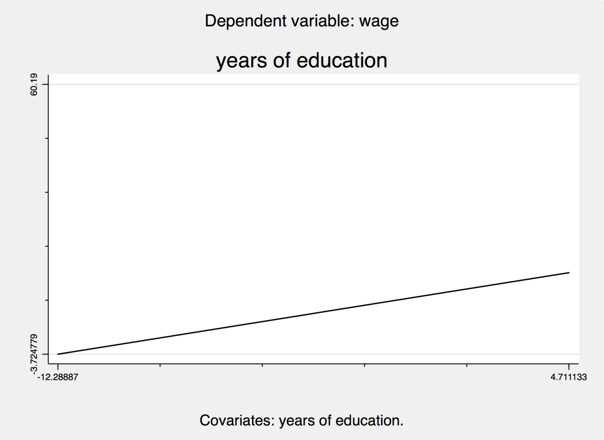
Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

The fitted line from the regression of wage on educ. The slope is

$$\hat{\beta}_{educ}$$

and it is biased because we ignore exper.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example



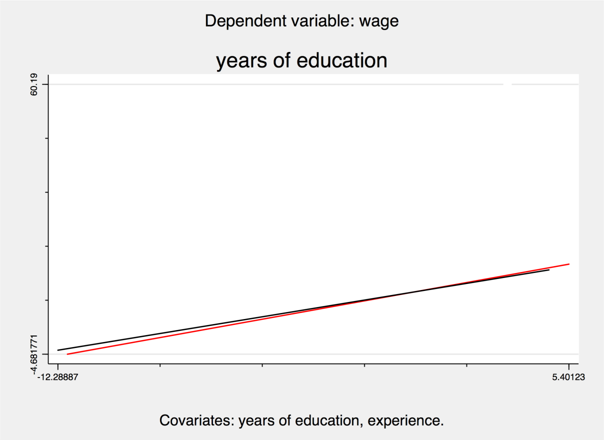
Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example

Adding the fitted line from the regression of wage on educ after partialling out the effect of exper: red line. The slope is

$$\hat{\beta}_{educ}$$

and it is unbiased. The difference in the slopes is the size of the bias due to ignoring exper in the regression.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: OLS estimator is biased: Example



Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: Efficiency losses its meaning

Efficiency becomes meaningless because OLS is no longer estimating the true parameter. Its distribution is centered at a biased mean, so the variance around that wrong mean tells you nothing about how well it estimates the true

$$\beta$$

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in finite samples: Normality loses its meaning

The OLS estimator Normality remains mathematically valid, but it becomes useless for inference about the true

$$\beta$$

because the estimator is centered at the wrong mean.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in large samples: OLS estimator is inconsistent

In an earlier lecture we showed that

$$\text{plim } \hat{\beta} = \beta + \underbrace{\text{plim } \left[\left(\frac{1}{n} \sum_{i=1}^{n} x_i x_i' \right)^{-1} \right]}_{\left(\mathrm{E}\left[x_i x_i' \right] \right)^{-1}} \underbrace{\text{plim } \frac{1}{n} \sum_{i=1}^{n} x_i u_i}_{\mathrm{E}[x_i u_i]}$$

If we have endogeneity, that is

$$\mathrm{E}\left[u_i x_i \right] \neq 0$$

the OLS estimator

$$\hat{\beta}$$

is inconsistent: second term does not disappear, and the estimator is not converging to the true value when the sample size increases.

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in large samples: Asymptotic efficiency loses its meaning

Asymptotic efficiency is defined only for estimators that consistently estimate the same true parameter. When exogeneity fails, OLS is inconsistent and converges to a pseudo-true value rather than the true

$$\beta$$

Because the estimator is centered on the wrong limit, comparing its asymptotic variance to that of other estimators is meaningless: the variance describes precision around a biased target, not around the true parameter

$$\beta$$

Linear regression model: Model assumption: Error is endogenous: Implications for the sampling distribution of the OLS estimator in large samples: Asymptotic normality loses its meaning

When exogeneity fails, OLS is inconsistent and does not converge to the true parameter

$$\beta$$

Although OLS remains asymptotically normal, the limiting distribution is centered at the wrong value. Therefore, asymptotic normality becomes useless for inference about the true parameter, because the estimator is not converging to

$$\beta$$

in the first place.