Econometrics: Endogeneity

Diego López Tamayo [*]   Based on MOOC by Erasmus University Rotterdam

Contents

> "There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

[*] El Colegio de México, diego.lopez@colmex.mx

Endogeneity

What is endogeneity?

```r
dataset4 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset4.csv")
```

Dataset 4

Simulated data on performance of 1000 participants of an Engineering MOOC. Performance is measured by Grade Point Average. Background variables are gender, whether participant followed a preparatory mathematics MOOC, and whether the participant received an email invitation for this preparatory MOOC. Variables: - GPA: Grade point average scale 0 to 10, 10 being the best - PARTICIPATION: 0/1 variable indicating participation (1) in preparatory MOOC or not (0) - GENDER: 0/1 variable indicating gender: male (1), female (0) - EMAIL: 0/1 variable indicating whether participant received email invitation for preparatory course (0: no invitation, 1: invitation)

Ordinary least squares (OLS) is a great tool to uncover relationships in economics and business. But we must be aware that this tool does not always work. There are circumstances where OLS breaks down. These circumstances relate to the **difference between correlation and causality**. Luckily, econometrics also has the solution. But we before we discuss this, let's consider a motivating example.

Suppose we want to explain:
- the monthly number of departing flights at an airport (y) - using the number of travel insurances sold in the month before. (x)

What kind of relationship would you expect if you regress flights as the variable y on a constant, and insurances as the variable x? Most likely we will obtain a positive relationship. Suppose OLS yields:

$$y = 10,000 + .25x + e$$

How should we interpret the obtained coefficients? What does the estimate .25 really mean? Suppose we have 4,000 travel insurances sold in the month before:

- **Correct:** $4,000$ insurances sold $\rightarrow$ expected number of flights $= 10,000 + .25\text{Ö}4,000 = 11,000$. Because High x tends to go together with high y. The identified correlation yields adequate predictions. Note that this statement merely relies on a correlation.

- **Incorrect:** Selling 4,000 additional insurances causes $.25\text{Ö}4,000 = 1,000$ additional flights. The regression does not identify a causal impact! A third variable (travel demand) affects y (flights) and x (insurances).

This example shows that we cannot always interpret least squares estimation results, as causal effects. However, identifying causal effects is one of the main goals of econometrics.

Ordinary least squares requires some assumptions for it to correctly estimate causal effects. One important assumption is that **explanatory variables are exogenous**. The violation of this assumption is called endogeneity.

In the following sections you will:

- Understand/recognize endogeneity
- Know the consequences of endogeneity
- Estimate parameters under endogeneity
- Know the intuition of the new estimator
- Test assumptions underlying this new estimator

Sources of endogeneity

Let us start by studying the source of endogeneity.

The formal assumption that we violate is the assumption that explanatory variables X in the linear model are non-stochastic. (Assumption A2) Explanatory variables are non-stochastic.

Literally speaking, non-stochastic means that if you would obtain new data only the y values would be different and the values for X would stay the same. This is like a *controlled experiment* where the researcher determines the experimental conditions coded in X. This assumption is crucial for the OLS estimator to be consistent. Consistent means that the estimator b converges to the true coefficient beta when the data set grows larger and larger. $b \to \beta$ for $n \to \infty$.

In economics however, controlled experiments are rare. X variables are often the consequence of an economic process, or of individual decision making. In our example, the travelers together determine the number of insurances sold. From the researcher's point of view, the X variables should therefore be seen as stochastic.

Once we allow X to be stochastic, we acknowledge that we would get different X values in a new data set. And if variables are stochastic, they can also be correlated with other variables, even with variables that are not included in the model!

In the context of our example, the number of insurances will be correlated with the travel demand. Although travel demand is difficult to observe and not included in the model, it does influence the number of flights. In the model, travel demand is therefore part of the error term epsilon. As a consequence, the X variable, insurances sold, is correlated with the error term ϵ.

- If X is endogenous → there is another variable(s) that affect y and X.
- OLS does not properly estimates β (inconsistent)

Usually, this correlation is due to an omitted factor.

Now let's consider three possible sources of endogeneity in more detail.

1. Endogeneity is often due to an **omitted variable**. In our example, the omitted variable was travel demand. Let's consider this situation formally.

Suppose that the true model for a variable y contains two blocks of explanatory variables, X1 and X2. And that in this true model, all assumptions are satisfied $y = X_1\beta_1 + X_2\beta_2 + \eta$ but we do not observe X_2 and perform OLS on $y = X_1\beta_1 + \epsilon$. The error term in the second model is:

$$\epsilon = X_2\beta_2 + \eta$$

From this relationship we can see that in the second model X_1 will be correlated with epsilon if X_1 and X_2 are correlated and β_2 does not equal 0: $Cov(X_1, X_2)\beta_2 \neq 0$ notice that $Cov(X_1, \eta) = 0$ due to orthogonality.

$$Cov(X_1, \epsilon) = Cov(X_1, X_2\beta_2 + \eta) = Cov(X_1, X_2)\beta_2 + Cov(X_1, \eta)$$

When thinking about whether certain variables in a model are endogenous, it is good to think about potential omitted variables. If you can think of an omitted variable that is related to the included variables, and the dependent variable, you will have endogeneity.

Suppose we run a regression to explain a student's grade using only the number of attended lectures. What omitted variable leads to endogeneity here? There are many possible ommited factors:

- Difficulty of exam? Probably NOT correlated with attendance.
- Motivation of the students? Probably correlates with attendance and affects grade.
- Compulsory attendance yes/no? Does not directly impact the grade

The omission of the motivation of students does lead to endogeneity. Highly motivated students are likely to attend many lectures and obtain high grades. So a regression of grades on attendance will not show the true impact of attendance. It will partly capture the unobserved motivation as well.

2. A second cause of endogeneity is **strategic behavior**.

Consider a model in which you explain the demand for products using only its price. If the salesperson strategically sets high prices when a high demand is expected, high demand will often go together with high prices! A simple regression may then yield a positive price coefficient. This is of course not the true impact of price. **Price is endogenous in this regression** as it correlates with the market information, which in turn, determines demand.

3.A third reason for endogeneity, is **measurement error**.

Suppose that we have a variable y, say, salary, That depends on a factor that is difficult to measure. For example, intelligence. Let's denote the intelligence by x^*. We can obtain a noisy measurement of intelligence, for example through an IQ test. The test score is called x and is equal to the true intelligence plus the measurement error.

$$x = x^* + measurementerror$$

> To summarize, endogeneity is a common and serious challenge in econometrics as OLS is not useful under endogeneity.

Formalizing endogeneity

We will show that such measurement error leads to endogeneity in a model that explains why using the IQ test score x in the salary example:

We want to explain the income yi of an individual $i = 1, ..., n$ using the individual's intelligence x_i^*. Suppose that the true relationship between these two variables is:

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

where β gives the impact of intelligence on income. Furthermore, suppose that this model satisfies all the standard assumptions of the linear model. However, the intelligence x_i^* cannot be observed directly. We can only observe a test score that equals the true intelligence plus a measurement error, that is $(1) x = x_i^* + w_i$ The measurement error process w_i satisfies the following conditions:
- Mean zero $E(w_i) = 0$ - Constant variance $Var(w_i) = \sigma_w^2$ - Zero correlation across individuals: $Cov(w_i, w_j) = 0$ $\forall i \neq j$ - Uncorrelated with unexplained income and true intelligene: $Cov(w_i, u_i) = 0$ and $Cov(w_i, x_i^*) = 0$

We have data on (y_i, x_i) for $i = 1, ..., n$ Suppose we ignore measurement error and simply apply OLS to:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

By definition $\epsilon_i = -\beta w_i + u_i$ we can show this just by equalizing the true relation and the estimated model:

$$\alpha + \beta x_i^* + u_i = \alpha + \beta x_i + \epsilon_i$$

Which can be rewritten as:

$$\epsilon_i = \beta(x_i^* - x_i) + u_i = -\beta w_i + u_i$$

Using this $(1) x = x_i^* + w_i$ equation and $(2) \epsilon_i = -\beta w_i + u_i$ we can show that the covariance between x_i and ϵ_i is $Cov(x_i, \epsilon) = -\beta \sigma_w^2$

$$Cov(x_i, \epsilon) = Cov(x_i^* + w_i, -\beta w_i + u_i) = \beta Cov(x_i^*, w_i) + Cov(x_i^*, u_i) - \beta Cov(w_i, w_i) + Cov(w_i, u_i)$$

Where we know by definition that $Cov(x_i^*, w_i) = 0$, $Cov(x_i^*, u_i) = 0$ and $Cov(w_i, u_i) = 0$

$$Cov(x_i, \epsilon) = -\beta\sigma_w^2$$

So x_i is endogenous if the $Cov(x_i, \epsilon) \neq 0$ in this case using the last resutl, it means that the variance of the measurement error $\sigma_w^2 \neq 0$ and that the true impact of inteligence on income $\beta \neq 0$.

Consequences