# The Endogeneity Problem

AGEC 317 Extra Material

Michael Black

## What is endogeneity?

When we perform OLS regression, we make several assumptions to arrive at the best linear unbiased estimator (BLUE). If those assumptions are violated, then our estimates are not BLUE - they are *biased.* Why does bias in models matter? Regression is used to inform decision-making, and if we are using biased numbers when making decisions, we could make terrible decisions!

For example, suppose you estimate own-price elasticity for some product, and you determine that demand for the product is elastic. With this information, you put the product on sale, because reductions in price for elastic goods will increase revenue (from AGEC 105). Now suppose your model is biased, and demand for the product is actually *inelastic.* Then the sales promotion will **lower** your revenues. Bad decision, informed by a bad model. But really, the culprit is you for believing a bad model.

Officially, endogeneity is when the independent variables are correlated with the error term. This violates the OLS assumption that the error term is, on average, zero (given the values of the independent variables). That violation causes bias in the estimated model.

## Math Example

Suppose we have the following true model:

$$wage_i = \beta_0 + \beta_1 education_i + \beta_2 ability_i + \varepsilon_i$$

We want to estimate this model, but we can't! How can we measure the ability of an individual? Instead, we estimate:

$$wage_i = \beta_0 + \beta_1 education_i + u_i$$

where $u_i = \beta_2 ability_i + \varepsilon_i$. Are education an ability correlated? Yes! That means $education_i$ and $u_i$ are correlated, which is the definition of endogeneity. Let's show the relationship of education and ability as:

$$ability_i = \delta_0 + \delta_1 education_i + e_i$$

Okay, now let's put everything together into the equation we are going to estimate. Remember we are estimating:

$$wage_i = \beta_0 + \beta_1 education_i + u_i$$

where

$$u_i = \beta_2(\delta_0 + \delta_1 education_i + e_i) + \varepsilon_i$$

Now suppose we run the estimated model, so we get:

$$wage_i = \hat{\beta}_0 + \hat{\beta}_1 education_i + u_i$$

We think that:

$$\frac{\partial wage_i}{\partial education_i} = \hat{\beta}_1$$

But in reality, remember the real model has ability in the $u_i$ error term. So the real model is:

$$wage_i = \beta_0 + \beta_1 education_i + \beta_2(\delta_0 + \delta_1 education_i + e_i) + \varepsilon_i$$

And then the *real* effect of education on wage is:

$$\frac{\partial wage_i}{\partial education_i} = \beta_1 + \beta_2\delta_1$$

Which means the data is telling us the marginal effect of education on wage is:

$$\hat{\beta}_1$$

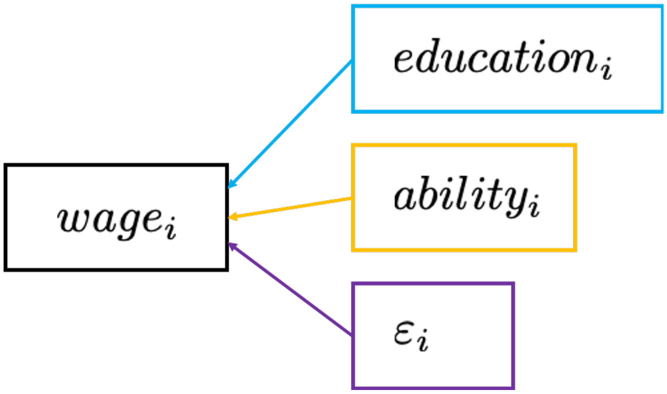
But the true effect is:

$$\beta_1 + \beta_2\delta_1$$

Our goal when we model is to make sure our predictions are correct. We want our estimation to equal the true model. Is $\hat{\beta}_1 = \beta_1$? Nope! We have bias of the magnitude $\beta_2 \times \delta_1$. That's the wedge between our answer and the correct answer.

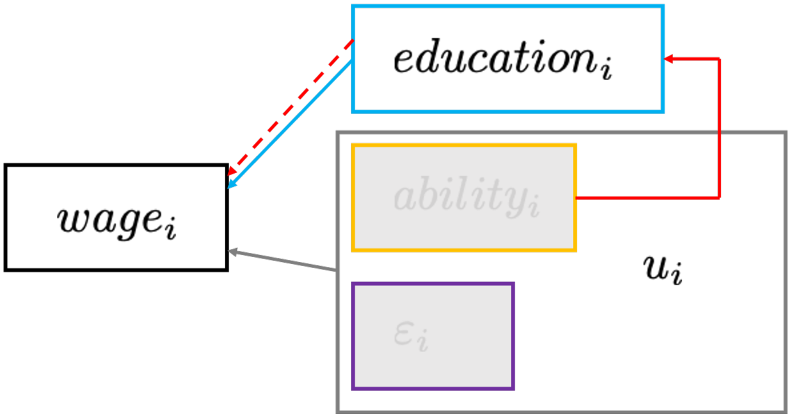
## Visual Example

Okay, that's a lot of math. What is going on visually? Here is the true model in visual form:



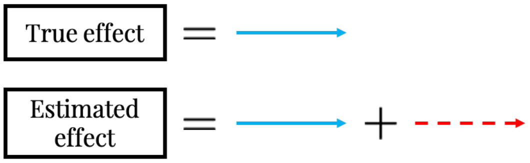
We think that wage is a function of eduction, ability, and some random effect. In other words, we think that we can predict a person's wage if we know their education and ability, but our prediction may be off by some bit (the error).

Unfortunately, we cannot measure ability, so we don't include it in the model we estimate. We only use information on education and wage to estimate the model:



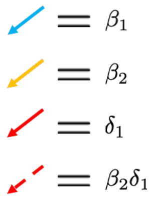
UH OH! We only included $education_i$, so the new error term $u_i$ contains the original random error $\varepsilon_i$ **and** $ability_i$. The insidious $ability_i$ variable is still affecting $education_i$! That's the solid red line. That effect of $ability_i$ on $education_i$ leaks through to the outcome variable, $wage_i$. That's the dotted red line.

So here is the difference between the **true** effect of education on wage, and what we **estimated**:



## Connecting the math and the visual

Each arrow of the figures above corresponds to a coefficient! Specifically:

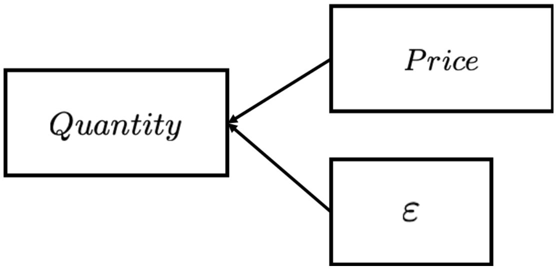


## Simultaneity

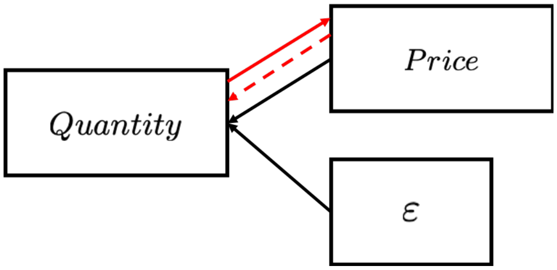
Another source of bias is simultaneity. This occurs when the dependent variable affects the independent variable, *and* vice versa. Suppose we are trying to estimate a classic demand curve, where:

$$Quantity = \beta_0 + \beta_1 Price + \varepsilon$$

Our economic intuition says that $\beta_1$ should be negative, right? That's the Law of Demand: as price increases, quantity demanded decreases, *ceteris paribus*. That equation can be visually represented as



But wait! Sure, price affects the quantity demanded, but couldn't suppliers observe the quantity demanded and change the price? For example, suppose Blue Bell ice cream is selling for $100 per gallon at HEB. Hardly anyone (except the true heroes among us) would buy the ice cream at that price. So quantity demanded would be very low. However, because quantity demanded is so low, Blue Bell could respond by lowering the price. That means quantity demanded has an effect on price! Visually:



The effect of quantity demanded on price is the solid red line, which leaks back into the effect of price on quantity as the dotted red line. This actually continues forever in a never-ending feedback loop. If you estimated the model above, your estimate of the effect of price on quantity demanded would be very biased.

In fact, the bias from simultaneity (a type of endogeneity) is so severe that estimated models often predict **upward-sloping** demand curves, which is totally nonsensical for most goods.

**Key message: just because you can get results from a regression does not mean the regression is good. It is very easy to create bad models in economics.**