Microeconometrics Module

Lecture 3: Endogeneity

Swapnil Singh

Lietuvos Bankas and KTU

Course Link

Simple Linear Regression

- Assume that there are two variables – y and x – and we are interested in understanding the relationship between them
  - y = wage and, x = education
  - y = diabetes and, x = whether smoking or not
- More importantly, we are interested in knowing whether x has a **causal** effect on y
- Three things to consider:
  1. How to allow other factors to affect y?
  2. What is the functional relationship between y and x?
  3. Under which conditions we can claim causality?

Simple Linear Regression

- Bite the bullet and write:

$$y = \beta_0 + \beta_1 x + u$$

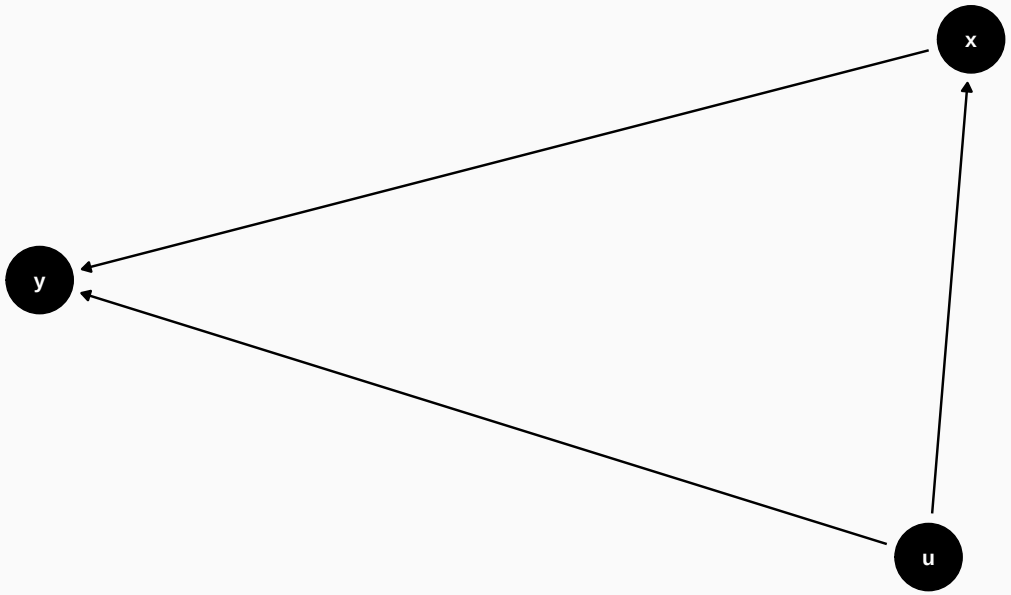
  where u is the error term
- Essentially, we are saying that
  1. Other factors affect y additively
  2. Parametric relationship is linear
     - Note that even though parametric relationship is linear, it can capture non-linear relationship between y and x
- But still there is an open question about causality
- For this, we put structure on the relationship between x and u

Simple Linear Regression

- Two assumptions: (1) $\mathbb{E}(u) = 0$, and (2) $\mathbb{E}(u|x) = \mathbb{E}(u) = 0$
- The first assumption is innoccuous
- The second assumption is the most important
- The violation of second assumption implies endogeneity problem, and we cannot claim causal effect. Why?

Simple Linear Regression

- Two assumptions: (1) $\mathbb{E}(u) = 0$, and (2) $\mathbb{E}(u|x) = \mathbb{E}(u) = 0$
- The first assumption is innoccuous
- The second assumption is the most important
- The violation of second assumption implies endogeneity problem, and we cannot claim causal effect. Why?



Example: Omitted Variable Bias

```
#preliminaries (just FYI, as I am displaying
    code for the first time)
rm(list=ls())
gc()
setwd(dirname(rstudioapi::
    getActiveDocumentContext()$path))
Sys.setenv(lang='en')
```

Example: Omitted Variable Bias

```
1
2     # Setting up the data
3     set.seed(123)
4     N <- 1000
5     X1 <- rnorm(N, 50, 10) # Explanatory variable
6     U <- rnorm(N, 0, 5)    # Unobserved variable
7     X2 <- 0.5 * X1 + rnorm(N,1,6)      # Another
        explanatory variable
8     Y <- 2 + 1 * X1 + 1.5 * X2 + U # Outcome
        variable
```

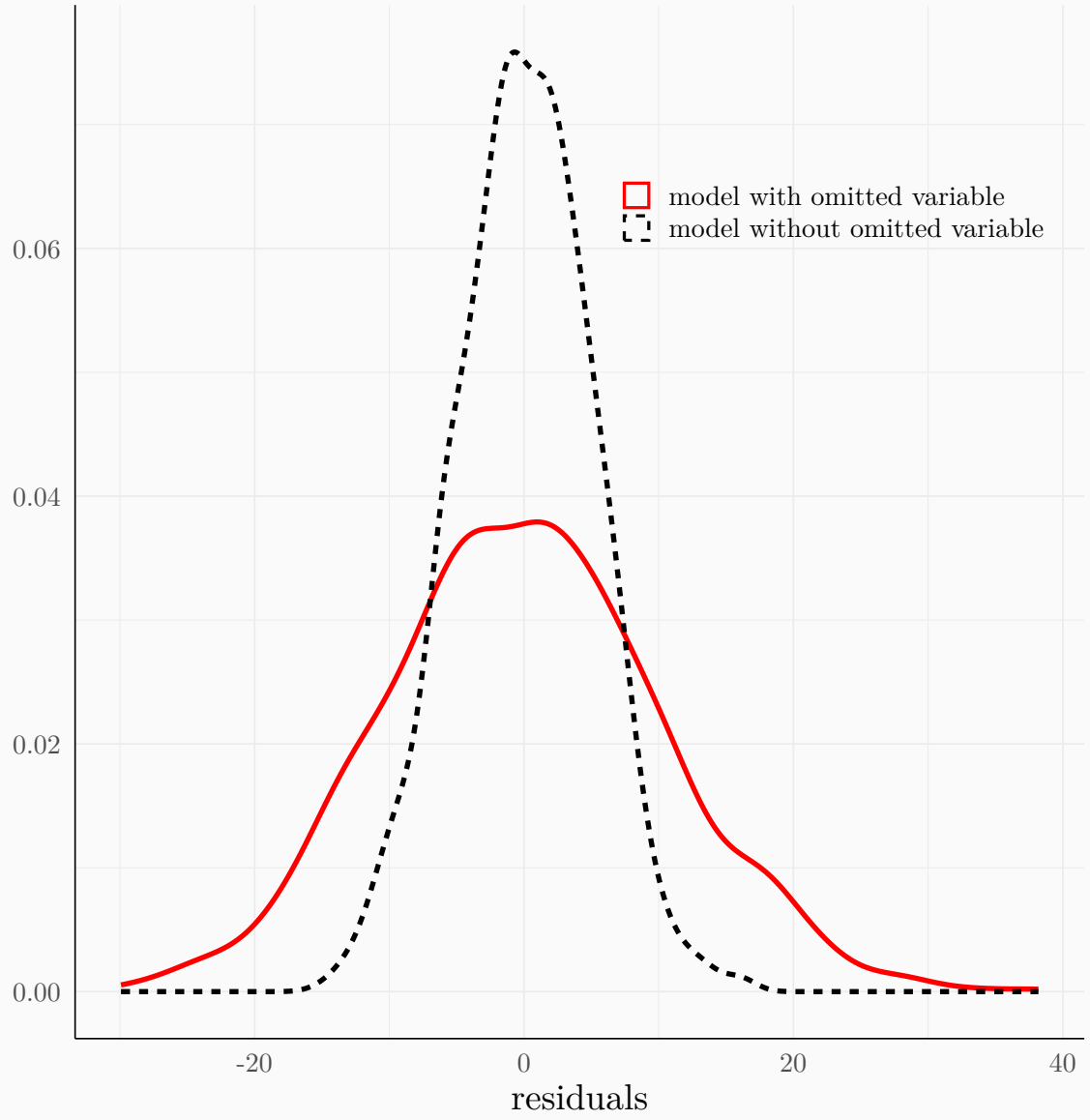
Example: Omitted Variable Bias

Dependent variable:	Y	
	(1)	(2)
X1	1.777***	1.032***
	(0.033)	(0.021)
X2		1.524***
		(0.027)
Constant	2.184	−0.031
	(1.675)	(0.822)
Observations	1,000	1,000
R^2	0.747	0.939
Note:	*p<0.1; **p<0.05; ***p<0.01	

Point of Caution

- Endogeneity is a conceptual issue
  - Cannot test it by using residuals after running the OLS regression

Point of Caution



Point of Caution

- Endogeneity is a conceptual issue
  - Cannot test it by using residuals after running the OLS regression
- You cannot compute error term, i.e. u, but you will always get residuals after running the regression
- Remember, source of endogeneity is $\mathbb{E}(u|x) \neq 0$