

# Endogeneity

AECN 396/896-002

# Before we start

## Learning objectives

Understand how endogeneity problems arise

## Table of contents

1. Selection Bias
2. Reverse Causality
3. Measurement Error

# Endogeneity

## Endogeneity

$E[u|x_k] \neq 0$  (the error term is not correlated with any of the independent variables)

## Endogenous independent variable

If the error term is, **for whatever reason**, correlated with the independent variable  $x_k$ , then we say that  $x_k$  is an endogenous independent variable.

- Omitted variable
- Selection
- Reverse causality
- Measurement error

# Omitted Variable

## True Model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ability + u$$

## Incorrectly specified (your) model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + v \quad (u + \beta_3 ability)$$

# Selection Bias

---

## Research Question

Does a soil moisture sensor reduce water use for farmers?

## Data

Observational (non-experimental) data on soil moisture sensor adoption and irrigation amount

## Model of interest

$$irrigation = \beta_0 + \beta_1 sensor + u$$

- *irrigation*: the amount of irrigation by the farmer
- *sensor*: dummy variable that indicates whether the farmer has adopted soil moisture sensor or not

## Question

Is *sensor* endogenous (is *sensor* correlated with the error term)?

Farmers do not just randomly adopt a soil moisture sensor, they consider available information to determine it is beneficial for them to adopt it or not.

### Adoption (selection) equation

$$sensor = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v$$

### Question

What would be variables that farmers look at when they decide whether they should get a soil moisture sensor or not?

### Question

Are any of the variables listed above also affect irrigation demand?

## Example

Soil quality/type (hard to accurately measure)

- farmers whose fields are sandy are more likely to adopt a soil moisture sensor (this is just a conjecture)
- farmers whose fields are sandy are likely to use more water

## Key

Soil quality/type affect **both** the decision of soil moisture sensor adoption and irrigation.

- *sensor* is a function of soil quality/type
- *irrigation* is a function of soil quality/type, which is in the error term uncontrolled for



$$irrigation = \beta_0 + \beta_1 sensor(\text{soil type}) + u \quad (= \beta_s \text{soil type} + v)$$

where  $v$  include all the unobservable variables except soil type.

So,  $sensor$  and the error term in the irrigation model are correlated through soil type, leading to biased estimation of the impact of a sensor.

## Important

Selection bias is a form of omitted variable bias

## Reasoning

If you accurately measure the common factors in the two equations, you can simply include them explicitly in the main model.

For example,

$$irrigation = \beta_0 + \beta_1 sensor(\text{soil type}) + \beta_s \text{soil type} + u$$

This will get the common factor (soil type) out of the error in the main model, which means the adoption variable and the error term are no longer correlated in the main model.

## Note

We call this kind of omitted variable bias "**selection bias**" because of the underlying mechanism through which the variable of interest (here, adoption of a technology) is made endogenous.

- In this example, farmers self-selected into the adoption of a soil moisture sensor

# Reverse Causality

---

## Research Question

Does a particular type of medical treatment improve health?

## Data

Observational (non-experimental) **cross-sectional** data on a particular type of medical treatment and health

## Model

$$health = \beta_0 + \beta_1 treatment + u$$

- health: indicator of the health of patients
- treatment: dummy variable that indicates whether the patient is treated or not

This model basically compares the health of patients who have and have not had the treatment (no before-after comparison, yes this is dumb).

## Question

Is *treatment* endogenous? (Is *treatment* correlated with the error term?)

### Key

Whether patients get the treatment or not is not randomized, rather it is determined by doctors (like in the real world).

### Question

How do doctors decide whether to put their patients under a medical treatment?

### Answer

Patients' health condition!!!

## Selection (treatment decision) model

$$\text{treatment} = \beta_0 + \beta_1 \text{health} + u$$

- `treatment` is affected by `health`
- `health` is affected by `treatment`

## Consequence

`treatment` is endogenous because it is a function of health itself!

## Reverse Causality

This type of endogeneity problem is called **reverse causality** because the independent variable of interest is causally affected by the dependent variable even though your interest is in the estimation of the impact of the independent variable on the dependent variable.

# Another reverse causality example

## Context

- Under the Clean Water Act, some of those who discharge wastes into water (e.g., oil refinery) need to comply with water quality criteria of their discharges set under the law.
- EPA (Environmental Protection Agency) can take enforcement actions (e.g., financial penalties) to those who violate the requirements.

## Research Question

Are enforcement actions effective in improving the water quality of waster discharges?

## Data

Annual data on

- water quality measures of waster discharges by individual firms
- enforcement actions taken on firms by EPA

### Model of Interest

$$\text{water quality} = \beta_0 + \beta_1 \text{enforcement actions} + u$$

### Selection (enforcement decision) model

$$\text{enforcement actions} = \beta_0 + \beta_1 \text{water quality} + u$$

- water quality is affected by enforcement actions
- enforcement actions is affected by water quality

### Consequence

enforcement actions is endogenous because it is a function of water quality itself!



# Measurement Error

---

# Measurement Error (ME)

## Definition

Inaccuracy in the values observed as opposed to the actual values

## Examples

- reporting errors (any kind of survey has the potential of mis-reporting)
  - household survey on income and savings
  - survey on rice yield by farmers in developing countries
- the use of estimated values
  - spatially interpolated weather conditions (precipitation)
  - imputed irrigation costs

## Question

What are the consequences of having measurement errors in variables you use in regression?

# ME in the Dependent Variable

## True Model

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

with MLR.1 through MLR.6 satisfied ( $u$  is not correlated with any of the independent variables).

## Measurement Errors

The difference between the observed ( $y$ ) and actual values  $y^*$

$$e = y - y^*$$

## Estimable Model

Plugging the second equation into the first equation, your model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v, \quad \text{where } v = (u + e)$$

## Estimable Model

Plugging the second equation into the first equation, your model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v, \text{ where } v = (u + e)$$

## Question

What are the conditions under which OLS estimators are unbiased?

## Answer

$$E[e|x_1, \dots, x_k] = 0$$

So, as long as the measurement error is uncorrelated with the independent variables, OLS estimators are still unbiased.

# ME in Independent Variables

## True Model

Consider the following general model

$$y = \beta_0 + \beta_1 x_1^* + u$$

with MLR.1 through MLR.6 satisfied.

## Measurement Errors

The difference between the observed ( $x_1$ ) and actual values ( $x_1^*$ )

$$e_1 = x_1 - x_1^*$$

## Estimable Model

Plugging the second equation into the first equation,

$$y = \beta_0 + \beta_1 x_1 + v, \text{ where } v = (u - \beta e_1)$$

### Estimable Model

Plugging the second equation into the first equation,

$$y = \beta_0 + \beta_1 x_1 + v, \text{ where } v = (u - \beta e_1)$$

### Question

What are the conditions under which OLS estimators are unbiased?

### Answer

$$E[e_1 | x_1] = 0$$

Unfortunately, this never holds.

## Classical errors-in-variables (CEV)

The correctly observed variable ( $x_1^*$ ) is uncorrelated with the measurement error ( $e_1$ ):

$$\text{Cov}(x_1^*, e_1) = 0$$

## Under CEV

The incorrectly observed variable ( $x_1$ ) must be correlated with the measurement error ( $e_1$ ):

$$\begin{aligned}\text{Cov}(x_1, e_1) &= E[x_1 e_1] - E[x_1]E[e_1] \\ &= E[(x_1^* + e_1)e_1] - E[x_1^* + e_1]E[e_1] \\ &= E[x_1^* e_1 + e_1^2] - E[x_1^* + e_1]E[e_1] \\ &= \sigma_{e_1}^2 = \sigma_{e_1}^2\end{aligned}$$

So, the mis-measured variable ( $x_1$ ) is always correlated with the measurement error ( $e_1$ ).

### Question

So, what is the direction of the bias?

### Note

The sign of the bias on  $x_1$  is the sign of the correlation between  $x_1$  and  $v = (u - \beta e_1)$ .

### Bias?

- Correlation between  $x_1$  and  $u$  is zero
- The sign of the correlation between  $x_1$  and  $e_1$  is positive (see the previous slide), which means that the sign of the correlation between  $x_1$  and  $-\beta e_1$  is the sign of  $-\beta$ .
  - if  $\beta > 0$ , then the sign of the bias is negative
  - if  $\beta < 0$ , then the sign of the bias is positive



## Bias?

- Correlation between  $x_1$  and  $u$  is zero
- The sign of the correlation between  $x_1$  and  $e_1$  is positive (see the previous slide), which means that the sign of the correlation between  $x_1$  and  $-\beta e_1$  is the sign of  $-\beta$ .
  - if  $\beta > 0$ , then the sign of the bias is negative
  - if  $\beta < 0$ , then the sign of the bias is positive

## Attenuation Bias

- So, the bias is such that your estimate of the coefficient on  $x_1$  is biased toward 0.
- In other words, your estimated impact of a mis-measured independent variable will look less influential than it actually is

(Imagine you mislabeled the treatment status of your experiment)