# Omitted Variable Bias

Eduard Bukin

# Omitted Variable Bias

In multiple regression, the Ceteris Paribus is achieved by introducing control variables.

> ⚠ **Warning**
>
> Having **bad controls** / insufficient / not right controls leaves us with the Selection Bias.

In the context of regression analysis, selection bias is called **OVB - Omitted Variable Bias**.

# Long Model

- A regression model that we wish to have.

$$Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l$$

where:

- $Y_i$ is the outcome variable;
- $P_i$ is the key variable of interest;
- $A_i$ is the **omitted variable**;
- $\alpha^l$ , $\beta^l$ are true regression coefficients;
- $\gamma$ is the effect of omitted variable in long;
- $e_i^l$ true error terms.

# Short model

- Is a model that we actually have, which omits one important variable ($A_i$) from the long model.

$$Y_i = \alpha^s + \beta^s P_i + e_i^s$$

where:

- $Y_i$ is the outcome variable;
- $P_i$ is the key variable of interest;
- $\alpha^s$ , $\beta^s$ are the estimates of regression coefficients in the short model;
- $e_i^s$ error terms.

# Omitted Variable Bias (1)

Omitting variable $A_i$ in the short model causes **bias** of $\beta^s$.

$$\beta^s = \beta^l + \text{OVB}$$

We can measure Omitted Variable Bias ($\text{OVB}$) as:

$$\text{OVB} = \beta^s - \beta^l$$

# Omitted Variable Bias happens when:

1. $P_i$ and $A_i$ relates to each other:
    - $E[A_i|P_i] \neq 0$ ;
2. $A_i$ and $Y_i$ relates to each other:
    - $E[Y_i|A_i] \neq 0$ in the long regression or $\gamma \neq 0$;

# Auxiliary regression

- Is a regression of **omitted variable** ($A_i$) on treatment $P_i$ and other regressors in short (if any).

$$A_i = \pi_0 + \pi_1 P_i + u_i$$

- Auxiliary regression helps us to calculate the $\text{OVB}$.

# Omitted Variable Bias (2) the key

With:

- **Long:** $Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l;$
- **Short:** $Y_i = \alpha^s + \beta^s P_i + e_i^s;$
- **Auxiliary:** $A_i = \pi_0 + \pi_1 P_i + u_i;$

We can measure Omitted Variable Bias as:

$$\text{OVB} = \beta^s - \beta^l$$

$$\text{OVB} = \pi_1 \times \gamma$$

# Math behind the OVB

- **Long:** $Y_i = \alpha^l + \beta^l P_i + \gamma A_i + e_i^l;$
- **Short:** $Y_i = \alpha^s + \beta^s P_i + e_i^s;$
- **Auxiliary:** $A_i = \pi_0 + \pi_1 P_i + u_i;$

- Let us substitute $A_i$ in the long with Auxiliary regression:

$$\Rightarrow Y_i = \alpha^l + \beta^l P_i + \gamma\{\pi_0 + \pi_1 P_i + u\} + e_i^l$$

$$\Rightarrow Y_i = \underbrace{\alpha^l + \gamma\pi_0}_{\alpha^s} + \underbrace{(\beta^l + \gamma\pi_1)}_{\beta^s} P_i + \underbrace{e_i^l + \gamma u_i}_{e_i^s}$$

- We obtain our short regression, where every estimate is biased.

# Why OVB formula is important (1)

- Presence of OVB in regression renders all our estimates biased/useless.

- **Omitted Variable** - means that we cannot have it in the regression, we can’t use data.
- Having knowledge of mathematics behind OVB, we can **make an educated guess about consequences of the variable omission: the BIAS** (Angrist & Pischke, 2014)

# How to check the OVB (2)

1. Write down Short, Long and Auxiliary regressions
2. Justify potential signs of $\pi_1$ and $\gamma$;
3. Conclude how the OV biases our regression based on the formula: $\text{OVB} = \pi_1 \times \gamma$.
4. OBV can bias estimates:
   - upwards ($\text{OVB} > 0$): increasing the effect of $P_i$
   - downwards ($\text{OVB} < 0$): decreeing the effect of $P_i$
   - rendering the effect of $P_i$ insignificant

# How to resolve the OVB?

- No solution!
  - Proxies;
  - Research design (Panel Regression/DiD, RDD);
- Acknowledge presence of the OVB;
- Discuss the bias;

# Example 1. Education and Experience

# Mincer equation

In 1970, Jacob Mincer in his work Schooling, Experience, and Earnings (Mincer, 1974) attempted to quantify the premium of schooling on wage. He used the following regression equation:

$$\log \text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \epsilon_i$$

**Prove that omitting experience causes OBV!**

# Step 1. Write long, short and auxiliary regressions:

Long: $\log \text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \epsilon_i$

Short: $\log \text{wage}_i = \beta_0^s + \beta_1^s \text{educ}_i + \epsilon_i^s$

Auxiliary: $\text{exper}_i = \rho_0 + \rho_1 \text{educ}_i + u_i$

# Step 2. Hypothesize about crucial effects:

> Use literature and other empirical research to reinforce your claims.

1. Effect of **experience** on **wage** ($\beta_2$)

$$\beta_2 > 0$$

- More years of experience, higher wage

2. Effect of **education** on **experience** ($\rho_1$)

$$\rho_1 < 0$$

- More time person spend in education, less time is left to work and gain experience.

# Step 3. Write down an OBV formula

$$\text{OVB} = \beta_2 \times \rho_1$$

Given our previous hypotheses:

- $\beta_2 > 0 = +$
- $\rho_1 < 0 = -$

$$\text{OVB} = (+) \times (-) < 0$$

- Omitting experience in short regression might cause a downward bias on the estimated effect of education. As a result, we may:
  - underestimate the effect of education.
  - find the effect of education insignificant or negative.

# Wage and Education

Supposed that we have estimates equation:

$$\log \text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \epsilon_i$$

- Could there be any other OVB in the wage-education relationship?

# Checking OBV based on the data

```r
library(tidyverse)
dta <- read_csv("wage.csv")
```

```r
short <- lm(log(wage) ~ educ, data = dta)
short
```

```

Call:
lm(formula = log(wage) ~ educ, data = dta)

Coefficients:
(Intercept)         educ
    5.97306      0.05984
```

```r
long <- lm(log(wage) ~ educ + exper, data = dta)
long
```

```

Call:
lm(formula = log(wage) ~ educ + exper, data = dta)

Coefficients:
(Intercept)         educ        exper
    5.50271      0.07778      0.01978
```

```r
aux <- lm(exper ~ educ, data = dta)
aux
```

```

Call:
lm(formula = exper ~ educ, data = dta)

Coefficients:
(Intercept)         educ
    23.7831      -0.9073
```

# Estimating the bias

```r
coef(aux)[["educ"]] * coef(long)[["exper"]]
```

```
[1] -0.01794277
```

## Checking the difference between long and short.

```r
coef(short)[["educ"]] - coef(long)[["educ"]]
```

```
[1] -0.01794277
```

# Example 2. Ability bias

Show how omitting ability biases the estimates of the effect of education on wages.

# The problem

Supposed that we have estimated the following regression:

$$\log \text{wage}_i = \beta_0^s + \beta_1^s \text{educ}_i + \beta_2^s \text{exper}_i + \beta_3^s \text{exper}_i^2 + \epsilon_i$$

What the other variables that are omitted and that may cause the bias to our estimates?

- Note that OV should correlate with $\log \text{wage}_i$ and $\text{educ}_i$;

These are other human capital related variables: age, ability, motivation.

# Show that omitting ability causes the OVB

1. Write short, long and auxiliary regression
   - short:
   - long:
   - auxiliary:
2. Write the OVB formula:
   - $\text{OVB} = \cdot$
3. Make Hypothesis about the effect of included on omitted and omitted on dependent:
   - Argument your statements.
4. Conclude about the bias:
   - $\text{OVB} = \cdot$

# Solution

1. Write short, long and auxiliary regression
   - short: $\log \text{wage}_i = \beta_0^s + \beta_1^s \text{educ}_i + \beta_2^s \text{exper}_i + \beta_3^s \text{exper}_i^2 + \epsilon_i$
   - long: $\log \text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \gamma \text{ability}_i + \epsilon_i$
   - auxiliary: $\text{ability}_i = \rho_0 + \rho_1 \text{educ}_i + \rho_2 \text{exper}_i + \rho_3 \text{exper}_i^2 + u_i$
2. Write the OVB formula:
   - $\text{OVB} = \beta_1^s - \beta_1 = \rho_1 \times \gamma$
3. Make Hypothesis about $\gamma$ and $\rho_1$.
   - $\rho_1 > 0$ as more years of education are usually associated with higher abilities;
   - $\gamma$ Higher abilities are usually rewarded with higher salary.
4. Conclude about the bias;
   - $\text{OVB} = (+) \times (+) > 0$

# 4. Conclude about the bias;

- We might have an upwards bias of the estimates in our regression.
- Specifically the effect of education is overestimated.
- In the long model we might observe a lower effect of education on wage.

# Check this conclusion empirically in R

1. Estimate the short model
2. Estimate the long model where instead of abilities the IQ level is used as a proxy.
3. Calculate the extent of the OVB

# Takeaways and homework

# Takeaways

1. OVB Formula (Short, Long and Auxiliary regressions)
2. Be ready to demonstrate how to use the OVB formula for making an educated guess about the direction of the bias during the exam.

# Homework

# Watch several videos about

Dale, S. B., & Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. The Quarterly Journal of Economics, 117(4), 1491-1527.

1. Video 1. Selection Bias: Will You Make More Going to a Private University? **From minute 6:30** to the end. https://youtu.be/6YrIDhaUQOE
2. Video 2. **From 47:22** 2017 AEA Cross-Section Econometrics. Part 2
   - Note that other videos and handouts are available here 2017 AEA Cross-Section Econometrics

# Do the OVB analysis on your own

You want to estimate the causal effect of union membership on employees’ wages. And you estimate the following regression equation:

$$\log \text{wage}_i = \beta_0 + \beta_1 \textbf{union} + \beta_2 \textbf{experience} + \beta_3 \textbf{experience}^2 \\ + \beta_4 \textbf{married} + \beta_5 \textbf{sex} + \beta_6 \textbf{hours per week} + \epsilon_i$$

Your colleagues suggest that you should include an individual’s education in the list of control variables as omitting such regressor biases the estimate.

1. Using OVB formula prove that omitting education causes/does not causes the OVB.
2. Calculate the extent of the OVB

# References

Angrist, J. D., & Pischke, J.-S. (2014). *Mastering'metrics: The path from cause to effect*. Princeton University Press.

Mincer, J. (1974). Schooling, experience, and earnings. Human behavior & social institutions no. 2.