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**;
- α^l , β^l are true regression coefficients;
- γ 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;
- α^s , β^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 β^s.

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

We can measure Omitted Variable Bias (OVB) as:

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

Omitted Variable Bias happens when:

1. P_i and A_i relates to each other:
   - $E[A_i|P_i] \neq 0$; or
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 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 tcheck the OVB (2)

1. Write down Short, Long and Auxiliary regressions
2. Justify potential signs of π_1 and γ;
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: $\log \text{exper}_i = \rho_0 + \rho_1 \text{educ}_i u_i$

Step 2. Hypothesize about crucial effects:

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

1. Effect of **experience** on **wage** (β_2)

$$\beta_2 > 0$$

- More years of experience, higher wage

2. Effect of **education** on **experience** (ρ_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")

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

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.

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.

DIY Example 1.

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

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

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.