ECONOMETRICS MADE SIMPLE

Omitted Variable Bias

The hidden variable that quietly changes your regression result

A simple question

Suppose your regression finds that people with more years of education earn higher wages. Does the education coefficient measure only the effect of education? Not necessarily. People with more education may also differ in ability, family support, or opportunities.

Meet the missing variable

Imagine ability affects wages and is also related to how much education a person gets. If you leave ability out of the regression, the education coefficient may partly pick up ability too.

THE STORY IN ONE PICTURE



Ability also affects wages. If ability is left out, part of its influence can be wrongly attributed to education.

The key idea

Omitted variable bias (OVB) occurs when a relevant variable is left out of a regression and, as a result, the estimated coefficient on an included variable is systematically distorted.

Memory line: **A missing factor can get credit - or blame - for the wrong thing.**

HOW OVB HAPPENS

What exactly goes wrong?

Compare the model we should estimate with the simpler model we actually run.

THE FULL MODEL (if ability matters)

$$\text{Wage} = b0 + b1\ \text{Education} + b2\ \text{Ability} + u$$

Here, b1 is the education effect holding ability constant, under the model assumptions.

THE SHORTER MODEL (ability is omitted)

$$\text{Wage} = a0 + a1\ \text{Education} + v$$

The new error term v now contains the omitted influence of ability. If ability is related to education, the error term is also related to education.

Two conditions to create omitted-variable bias

1 **It affects the outcome**

Ability must help explain wages after the other included factors are considered: b2 is not zero.

2 **It is related to education**

Ability must be correlated with education, after accounting for the other included variables.

When is the omitted variable NOT biasing this coefficient?

In the simple setup, if either condition fails - the omitted factor does not affect wage, or it is uncorrelated with education - it does not create omitted-variable bias in the education coefficient. Other problems with the model may still exist.

A friendly warning

A variable is not automatically a confounder just because it is important. It must also be related to the included regressor in the relevant model.

THE FORMULA + A WORKED EXAMPLE

How big is the bias?

In a simple regression with one included variable X and one omitted variable Z, the classic expression is:

$$\text{Bias in } b1\text{-hat} = b2 \times \text{Cov}(X, Z) / \text{Var}(X)$$

Read it as: the omitted variable's effect on Y, multiplied by how strongly it moves with X relative to X's own variation.

A made-up example with wages

Assume the true relationship is:

$$\text{Wage} = 10 + 2(\text{Education}) + 3(\text{Ability}) + u$$

For simplicity, wage is measured in thousands of currency units per month. These numbers are hypothetical and are used only to show the mechanics.

1. **True education effect**

   b1 = 2. An additional year of education is associated with 2 extra thousand currency units per month, holding ability constant in this hypothetical model.

2. **Ability also matters**

   b2 = 3. A one-unit increase in the ability measure is associated with 3 extra thousand currency units per month, holding education constant.

3. **Education and ability move together**

   Suppose Cov(Education, Ability) / Var(Education) = 0.5. This is just an illustrative value, not a real estimate.

Calculate the omitted-variable bias

$$\text{Bias} = 3 \times 0.5 = +1.5$$

Estimated coefficient = 2 + 1.5 = 3.5

The regression that omits ability may report 3.5 instead of the true education coefficient of 2. The estimate is biased upward because ability raises wages and is positively related to education.

INTERPRET + REMEMBER

Which direction does bias go?

The direction depends on two signs: (1) the omitted variable's effect on the outcome, and (2) its relationship with the included variable.

Effect of omitted Z on Y	Relationship of Z with X	Likely bias in X coefficient
Positive (+)	Positive (+)	**Upward (+)**
Positive (+)	Negative (-)	**Downward (-)**
Negative (-)	Positive (+)	**Downward (-)**
Negative (-)	Negative (-)	**Upward (+)**

How can researchers reduce the risk?

- **Use theory to choose controls**

  Include relevant confounders measured well, based on the research question and causal structure.

- **Improve the research design**

  Randomisation, natural experiments, valid instruments, or credible panel methods can help identify causal effects when their assumptions hold.

- **Be transparent**

  Explain which variables are omitted, why, and how the results may change under alternative specifications.

Exam-ready definition

Omitted variable bias occurs when a relevant variable is excluded from a regression and the omission causes an included coefficient to be systematically biased. In the simple case, this occurs when the omitted variable affects the dependent variable and is correlated with an included explanatory variable.

Quick check: If ability raises wages but is unrelated to education, does omitting ability bias the education coefficient? In the simple setup, no.