What is the definition of omitted variable bias and what are some examples of it?

Authored by

stats writer

April 19, 2024

Omitted variable bias is a statistical phenomenon where the exclusion of a relevant variable from a statistical model results in biased estimations of the relationship between the included variables. This bias occurs when the omitted variable is correlated with both the independent and dependent variables, leading to incorrect conclusions and interpretations.

For example, in a study examining the relationship between exercise and weight loss, if the variable of diet is not included in the model, the estimated effect of exercise on weight loss may be biased. This is because diet is a relevant variable that is correlated with both exercise and weight loss. Therefore, the omission of diet leads to an underestimation of the true effect of exercise on weight loss.

Another example can be seen in a study investigating the impact of education on income. If the variable of intelligence is not included in the model, the estimated effect of education on income may be biased. This is because intelligence is a relevant variable that is correlated with both education and income. Therefore, the omission of intelligence leads to an overestimation of the true effect of education on income.

In summary, omitted variable bias is a statistical error that can occur when a relevant variable is excluded from a model, resulting in biased estimations of the relationship between the included variables. It is important to carefully consider and include all relevant variables in statistical models to avoid this bias and ensure accurate conclusions.

Omitted Variable Bias: Definition & Examples

Omitted variable bias occurs when a relevant explanatory variable is not included in a regression model, which can cause the coefficient of one or more explanatory variables in the model to be biased.

An omitted variable is often left out of a regression model for one of two reasons:

1. Data for the variable is simply not available.

2. The effect of the explanatory variable on the response variable is unknown.

In order for the omitted variable to actually bias the coefficients in the model, the following two requirements must be met:

1. The omitted variable must be correlated with one or more explanatory variables in the model.

2. The omitted variable must be correlated with the response variable in the model.

The Effects of Omitted Variable Bias

Suppose we have two explanatory variables, A and B, and one response variable, Y. Suppose we fit a simple linear regression model with A as the only explanatory variable and we leave B out of the model.

If B is correlated with A *and* correlated with Y, then it will cause the coefficient estimate of A to be biased. The following diagram shows how the coefficient estimate of A will be biased, depending on the nature of the relationship with B:

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with Y	Positive Bias	Negative Bias
B is negatively correlated with Y	Negative Bias	Positive Bias

Example: Omitted Variable Bias

Suppose we want to study the effect that square footage has on house price so we fit the following simple linear regression model:

House price = B0 + B1(square footage)

Suppose we find the estimated model to be:

House price = 40,203.91 + 118.31(square footage)

The way we would interpret the coefficient for square footage is that *each additional one unit increase in square footage is associated with an increase in house price of $118.31, on average.*

Based on the fact that *age* is negatively correlated with both the explanatory variable and the response variable in the model, we would expect the coefficient estimate

for square footage to be positively biased:

	A and B are positively correlated	A and B are negatively correlated
B is positively correlated with Y	Positive Bias	Negative Bias
B is negatively correlated with Y	Negative Bias	Positive Bias

Suppose we find data for house age and then include it in the model. The model then becomes:

House price = B0 + B1(square footage) + B2(age)

Suppose we find the estimated model to be:

House price = 123,426.20 + 81.06(square footage) - 1,291.04(age)

Note that the coefficient estimate for square footage went significantly down, which means it *was* positively biased in the previous model.

The way we would interpret the coefficient for square footage in this model is that *each additional one unit increase in square footage is associated with an*

average increase in house price of $81.06, assuming age is held constant.

What to Do About Omitted Variable Bias

Unfortunately omitted variable bias occurs often in the real world because there are usually some variables that *should* be included in a regression model but aren't because data for them isn't available or the relationship between them and the response variable is unknown.

If possible, you should try to include any and all relevant explanatory variables in a regression model so that you can understand the true relationship between the explanatory variables and the response variable.

Leaving relevant explanatory variables out of a model can significantly affect the interpretation of the model, as we saw in the previous example with house prices.

What is a Lurking Variable?

What is a Confounding Variable?