Omitted variable bias (as the name suggests) is the bias in the estimates of regression coefficients due to omitted variables. Consider the following regression model:

$$Y = \beta_0 + \beta_1 X + u$$

For omitted variable bias to occur, the omitted factor Z must be:

- A determinant of Y (i.e. Z is part of u), and

- Correlated with the X

Remember the OLS estimator:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)}$$

Plug in $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $\bar{Y} = \beta_0 + \beta_1 \bar{X}$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i + u_i - \beta_0 - \beta_1 \bar{X})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^{n} u_i (X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

So we have,

$$\hat{\beta_1} = \beta_1 + \frac{Cov(X, u)}{Var(X)}$$

The bias $\hat{\beta_1} - \beta_1$ depends on the sign of $Cov(X, u)$.

The direction of bias:

- $Cov(X, u) > 0$, then $\hat{\beta_1} > \beta_1$, so OLS estimator will be biased upwards.

- $Cov(X, u) < 0$, then $\hat{\beta_1} < \beta_1$, so OLS estimator will be biased downwards.

Note that if you are considering an omitted variable Z that enters u negatively, i.e., has a negative impact on Y, then the direction of the bias above should be flipped.

Example: Is Advertising Effective?

We are interested in examining whether higher advertising expenditures lead to higher sales. Consider the following simple regression model of sales revenue and advertising expenditure:

$$Sales = \beta_0 + \beta_1 Advertising + u$$

Let's consider how the following omitted variables could impact our estimates of the effect of advertising expenditures on sales.

- *Size of the firm:* Larger firms tend to have higher sales and higher advertising expenditures. If we omit firm size from our model, we would overestimate the impact of advertising expenditures on sales, as the effect of firm size on sales is mistakenly attributed to advertising expenditures. Note that here $Cov(X, u) > 0$ and so $\hat{\beta}_1 > \beta_1$.

- *Competition:* Some firms may face higher levels of competition, which may cause them to spend more on advertising but also result in lower sales. If we omit a measure of competition from our model, we would underestimate the impact of advertising expenditures on sales. Note that here $\hat{\beta}_1 < \beta_1$ because $Cov(X, Z) > 0$ but $Cov(X, u) < 0$ as Z enters u negatively.

- *A strong economy:* A strong economy may increase sales, but it is not necessarily correlated with advertising expenditures. Therefore, omitting information about the state of the economy from our model may not result in biased estimates of the effect of advertising expenditures on sales.

- *Consumer demographics:* If firms need to appeal to specific consumer demographics, they may use different advertising channels, which could have varying costs. For instance, firms targeting young consumers may use social media, which is less expensive than television advertising. However, omitting information about consumer demographics from our model would not introduce bias if consumer demographics are not correlated with the total value of sales.