Specification Bias

a. Specification mistake - suppose an important variable, $X_2$, is left out of the regression model.

The true model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

But, you assume:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \nu_i$$

(What CRM assumptions have been violated? Assumption #1 and Assumption #3.)

b. What happens - Verbally.

Your model assumes only $X_1$ causes $Y$ to change, **but** in truth, the variable $X_2$ also causes $Y$ to change.

The effects of $X_2$ on $Y$ are not accounted for in your model.

As a result, the effect of $X_2$ on $Y$ gets *tangled up* with the effect of $X_1$ on $Y$. We can't get a clear picture of how changes in $X_1$ affect changes in $Y$.

c. What happens - Mathematically.

The estimator that you use is:

$$\hat{\alpha}_i = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}$$

This will be **biased**. To show this take the expected value of the estimator and use the *true expected value of $Y$* when evaluating:

First, insert the true expected value of $Y_i$

$$E [ \hat{\alpha}_i ] = \frac{\sum x_{1i}(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}{\sum x_{1i}^2}$$

and:

$$E [ \hat{\alpha}_i ] = \beta_1 \frac{\sum x_{1i} X_{1i}}{\sum x_{1i}^2} + \beta_2 \frac{\sum x_{1i} X_{2i}}{\sum x_{1i}^2} = \beta_1 + \beta_2 \frac{\sum x_{1i} X_{2i}}{\sum x_{1i}^2}.$$ 

which says that the expected value of $\hat{\alpha}_i$ equals the true effect of $X_1$ on $Y$, $\hat{\alpha}_1$, plus the bias due to model misspecification.

The bias due to model misspecification is made up of two parts:

1. $\hat{\alpha}_2$ - the true effect of $X_2$ on $Y$; and
2. the relationship between $X_2$ and $X_1$.

**Moral of this story:**

*Leaving out an important independent variable can lead to biased parameter estimates.*