Analysis of the Omitted Variable Bias

Omitted variable bias: In a linear regression when a variable which has an effect on the independent and dependent variable is omitted this is called the *omitted variable bias.* This bias affects the exogeneity assumption of linear regression.

Example:
If you were to study what factors determine the price of a car (such as features, safety, accidents, etc.) and omit the cars age, then the regression is going to give you biased estimates, so two cars with the exact same variables may have different prices if their age is different.

The omitted variable will bias your coefficients given that:

1. It is correlated with the dependent variable.
2. Is correlated with one or more other explanatory variables.

We suppose the correct model is:

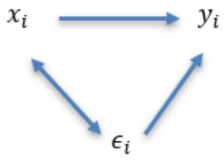
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 q_i + \epsilon_i,$$

Where,

$$x_i = \lambda_1 q_i + \mu_i,$$

Because of measurement inaccuracy, we cannot observe the q_i. We try to find the relationship between x_i and y_i by a misspecified model where x_i and ϵ_i are linked:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



Model without omitted variable

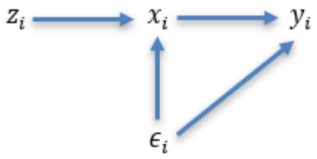
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Model with omitted variable

$$= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} x_i q_i}{\sum_{i=1}^{n} {x_i}^2} + \frac{\sum_{i=1}^{n} x_i \epsilon_i}{\sum_{i=1}^{n} {x_i}^2}$$

Instrumental variable approach/two-stage least square estimator (2SLS)

One of the most widely used approaches to solve endogeneity issues such as omitted variable bias is to use the two-stage least squares estimator (2SLS) and instrument variable approach. By introducing an instrumental variable, we hope to isolate the effect of x on y.



1. First the effect of an instrumental variable (*zi*) on *xi* is calculated.

$$x_i = \pi_0 + \pi_1 z_i + v_i$$

2. The second step consists of using this estimation *x̂i* to predict y, using these two steps makes the exogeneity assumption consistent since *x̂i* is exogenous.

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

$$\hat{y}_i = \beta_0 + \beta_{IV}\hat{x}_i + e_i$$

Simulation studies with varying sample sizes

	n=50			n=500			n=5000		
	Model_True	*Model_Reduced*	*Model_IV*	*Model_True*	*Model_Reduced*	*Model_IV*	*Model_True*	*Model_Reduced*	*Model_IV*
β_1	*0.95102****	*1.3236****	*0.6226**	*1.01368****	*1.23089****	*1.0213****	*1.00246****	*1.24878****	*0.98671****
	(16.709)	*(9.819)*	*(2.154)*	*(64.596)*	*(37.84)*	*(13.27)*	*(188.581)*	*(117.48)*	*(62.36)*
β_2	*1.09350****	-	-	*1.00056****	-	-	*0.99283****	-	-
	(16.507)	-	-	*(43.416)*	-	-	*(133.193)*	-	-
β_0	*0.01653*	*16.0701****	*37.295****	*0.56740*	*19.11760****	*25.420****	*1.11334****	*18.54657****	*26.41589****
	(0.009)	*(3.933)*	*(4.259)*	*(0.917)*	*(19.52)*	*(10.98)*	*(5.596)*	*(58.04)*	*(55.57)*
Methods	*Y=f(X, Q)*	*Y=f(X)*	*Y=f(X̂)*	*Y=f(X, Q)*	*Y=f(X)*	*Y=f(X̂)*	*Y=f(X, Q)*	*Y=f(X)*	*Y=f(X̂)*
R^2	*0.9511*	*0.6676*	*0.4804*	*0.9462*	*0.7419*	*0.2612*	*0.9416*	*0.7341*	*0.7018*