

## Specification Error: Omitted and Extraneous Variables

*Omitted variable bias.* Suppose that the “correct” model is

$$y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

If we estimate

$$y = a + b_1 X_1 + b_2 X_2 + e$$

we know that  $E(b_1) = \beta_1$  and  $E(b_2) = \beta_2$  i.e. the regression coefficients are unbiased estimators of the population parameters.

Suppose, however, the researcher mistakenly believes

$$y = \alpha^* + \beta_1^* X_1 + \varepsilon^*$$

and therefore estimates

$$y = a^* + b_1^* X_1 + e^*$$

i.e.  $X_2$  is mistakenly omitted from the model. How does  $b_1$  (the regression estimate from the correctly specified model) compare to  $b_1^*$  (the regression estimate from the mis-specified model)? What is  $E(b_1^*)$ ? Is it a biased or unbiased estimator of  $\beta_1$ ? If biased, how is it biased?

Note that  $b_1^*$

$$= \frac{\hat{Cov}(X_1, Y)}{\hat{V}(X_1)}$$

Formula for bivariate regression coefficient

$$= \frac{\hat{Cov}(X_1, a + b_1 X_1 + b_2 X_2 + e)}{\hat{V}(X_1)}$$

Substitute the formula for Y from the correctly specified model

$$= \frac{\hat{Cov}(X_1, a) + b_1 \hat{Cov}(X_1, X_1) + b_2 \hat{Cov}(X_1, X_2) + \hat{Cov}(X_1, e)}{\hat{V}(X_1)}$$

Expectations rules:  
 $Cov(a+b,c+d) = Cov(a,c) + Cov(a,d) + Cov(b,c) + Cov(b,d)$

$$= \frac{0 + b_1 \hat{V}(X_1) + b_2 \hat{Cov}(X_1, X_2) + 0}{\hat{V}(X_1)}$$

Recall that  $Cov(\text{variable, constant}) = 0$ . Also, X's are uncorrelated with the residuals.

$$= b_1 + b_2 \frac{\hat{Cov}(X_1, X_2)}{\hat{V}(X_1)}$$

Simplify expression.

Taking expectations, we get

$$E(b_1^*) = \beta_1 + \beta_2 \frac{\sigma_{12}}{\sigma_1^2}$$

Hence,  $b_1^*$  is a biased estimator of  $\beta_1$ . Further, this bias will not disappear as sample size gets larger, so the omission of a variable from a model also leads to an inconsistent estimator.

Note that there are two conditions under which  $b_1^*$  will not be biased:

- $\beta_2 = 0$ . Of course, if  $\beta_2 = 0$ , this means that the model is not mis-specified, i.e.  $X_2$  does not belong in the model because it has no effect on  $Y$ .
- $\sigma_{12} = 0$ . That is, if the 2  $X$ 's are uncorrelated, then omitting one does not result in biased estimates of the effect of the other.

**EXAMPLE:** Consider our income/education/job experience example:

Covariance:

	EDUC	JOBEXP	INCOME
EDUC	20.050	-2.613	37.068
JOBEXP	-2.613	29.818	14.311
INCOME	37.068	14.311	95.812

Variable(s) Entered on Step Number

- 1.. JOBEXP
- 2.. EDUC

----- Variables in the Equation -----

Variable	B	SE B	Beta	T	Sig T
EDUC	1.933393	.209949	.884438	9.209	.0000
JOBEXP	.649365	.172159	.362261	3.772	.0015
(Constant)	-7.096855	3.626412		-1.957	.0670

Note that, when both EDUC and JOBEXP are in the equation,  $b_1 = 1.933393$ ,  $b_2 = .649365$ ,  $Cov(Educ, Jobexp) = -.2613$ ,  $V(Educ) = 20.05$ ,  $V(Jobexp) = 29.818$ . Hence, if we omit Jobexp from the model, the new coefficient  $b_1^*$  is

$$b_1 + b_2 \frac{\hat{Cov}(X_1, X_2)}{\hat{V}(X_1)} = 1.933393 + .649365 \frac{-2.613}{20.050} = 1.848765$$

SPSS confirms that this is correct:

Block Number 2. Method: Remove JOBEXP

----- Variables in the Equation -----

Variable	B	SE B	Beta	T	Sig T
EDUC	1.848760	.274948	.845723	6.724	.0000
(Constant)	2.137446	3.523734		.607	.5517

Or, if we instead omit EDUC from the equation, for  $b_2^*$  we get

$$b_2 + b_1 \frac{\hat{Cov}(X_1, X_2)}{\hat{V}(X_2)} = .649365 + 1.933393 \frac{-2.613}{29.818} = .479928616$$

SPSS again confirms this:

Variable	B	SE B	Beta	T	Sig T
JOBEXP	.479931	.407079	.267739	1.179	.2538
(Constant)	18.343871	5.586783		3.283	.0041

If we assume that the model with both EDUC and JOBEXP is correct, omitting one or the other results in the effects of the remaining variable being mis-estimated.

In more complicated models with omitted variables, it will continue to be the case that observed effects represent a confounding of the actual effect with other sources of association.

*Inclusion of extraneous variables.* Suppose that the “correct” model is

$$y = \alpha + \beta_1 X_1 + \varepsilon$$

If we estimate

$$y = \alpha + b_1 X_1 + e$$

we know that  $E(b_1) = \beta_1$ , i.e. the regression coefficients is an unbiased estimators of the population parameter.

Suppose, however, the researcher mistakenly believes

$$y = \alpha^* + \beta_1^* X_1 + \beta_2^* X_2 + \varepsilon^*$$

and therefore estimates

$$y = a^* + b_1^* X_1 + b_2^* X_2 + e^*$$

i.e.  $X_2$  is mistakenly added to the model. How does  $b_1$  (the regression estimate from the correctly specified model) compare to  $b_1^*$  (the regression estimate from the mis-specified model)? What is  $E(b_1^*)$ ? Is it a biased or unbiased estimator of  $\beta_1$ ? If biased, how is it biased?

Here is an informal proof: We can think of the “correct” model as being a special case of the “incorrect” model, where  $\beta_2 = 0$ . It will therefore be the case that  $E(b_1^*) = \beta_1$ , and  $E(b_2^*) = 0$ . Hence, addition of extraneous variables does not lead to biased coefficients. However, adding extraneous (or “junk”) variables to the model will result in inflated standard errors and all the problems they create. Recall that, in the two IV case,

$$s_{b_k} = \sqrt{\frac{1 - R_{y12}^2}{(1 - R_{12}^2) * (N - K - 1)}} * \frac{s_y}{s_{X_k}}$$

As the formula suggests, adding irrelevant variables will tend not to increase the numerator, because irrelevant variables will not substantially increase  $R^2$ . However, irrelevant variables will tend to increase the denominator. The tolerance will be smaller ( $1 - R_{12}^2$ ) and  $N - K - 1$  will be smaller.