Specification Error: Omitted and Extraneous Variables

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised February 15, 2015

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. X2 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 β_1? If biased, how is it biased?

Note that b_1*

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

Formula for bivariate regression coefficient

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

Substitute the formula for Y from the correctly specified model

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

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

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

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

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

Simplify expression.

If your eyes glaze over when looking at equations, just make sure you get the conclusion. If X2 has mistakenly been omitted from the model, then, taking expectations, we get

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

Very Important: Hence, b_1^* is a biased estimator of β_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. In effect, x1 gets credit (or blame) for the effects of the variables that have been omitted from the model.

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. X2 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 1. I will construct a data set where b1 = 3, b2 = 2, and x1 and x2 have a correlation of .5. The standard deviation of x1 is 4 and the standard deviation of x2 is 4. We will see what happens if x2 is omitted from the model.

```
. clear all
. matrix input corr = (1,.5,0\.5,1,0\0,0,1)
. matrix input sds = (4\4\10)
. corr2data x1 x2 e, corr(corr) sd(sds) n(500)
(obs 500)
. gen y = 3*x1 + 2*x2 + e
. corr y x1 x2
(obs=500)

             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.7960   1.0000
          x2 |   0.6965   0.5000   1.0000

. corr y x1 x2, cov
(obs=500)

             |        y       x1       x2
-------------+---------------------------
           y |      404
          x1 |       64       16
          x2 |       56        8       16

. * Correct regression
. reg y x1 x2

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  755.44
       Model |      151696     2  75847.9998           Prob > F      =  0.0000
    Residual |  49899.9993   497  100.402413           R-squared     =  0.7525
-------------+------------------------------           Adj R-squared =  0.7515
       Total |  201595.999   499  403.999998           Root MSE      =   10.02

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          3   .1294885    23.17   0.000     2.745588    3.254412
          x2 |          2   .1294885    15.45   0.000     1.745588    2.254412
       _cons |  -4.41e-09   .4481125    -0.00   1.000    -.8804284    .8804284
------------------------------------------------------------------------------
```

```
. * Omitted variable bias
. reg y x1

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  861.41
       Model |      127744     1      127744           Prob > F      =  0.0000
    Residual |  73851.9991   498  148.297187           R-squared     =  0.6337
-------------+------------------------------           Adj R-squared =  0.6329
       Total |  201595.999   499  403.999998           Root MSE      =  12.178

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          4   .1362876    29.35   0.000     3.732231    4.267769
       _cons |   7.29e-08   .5446048     0.00   1.000    -1.070006    1.070006
------------------------------------------------------------------------------
```

We see that, when x2 is omitted from the model, the effect of x1 is over-estimated in this case. (In other situations it could be under-estimated). To confirm that Stata got it right,

$$b_1* = b_1 + b_2 \frac{\hat{C}ov(X_1, X_2)}{\hat{V}(X_1)} = 3 + 2\frac{8}{16} = 4$$

Example 2. Here is an example of a special case where omitting a variable does NOT result in omitted variable bias. I construct a data set similar to what we had before, except x1 and x2 are uncorrelated.

```
. clear all
. matrix input corr = (1,0,0\0,1,0\0,0,1)
. matrix input sds = (4\4\10)
. corr2data x1 x2 e, corr(corr) sd(sds) n(500)
(obs 500)
. gen y = 3*x1 + 2*x2 + e
. corr y x1 x2
(obs=500)

             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.6838   1.0000
          x2 |   0.4558   0.0000   1.0000

. * Correct regression
. reg y x1 x2

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  516.88
       Model |      103792     2  51896.0002           Prob > F      =  0.0000
    Residual |  49899.9994   497  100.402413           R-squared     =  0.6753
-------------+------------------------------           Adj R-squared =  0.6740
       Total |      153692   499         308           Root MSE      =   10.02

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          3   .1121403    26.75   0.000     2.779672    3.220328
          x2 |          2   .1121403    17.83   0.000     1.779672    2.220328
       _cons |  -4.71e-08   .4481125    -0.00   1.000    -.8804285    .8804284
------------------------------------------------------------------------------
```

```
. * X2 omitted but no bias in this case
. reg y x1

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  437.27
       Model |  71856.0006     1  71856.0006           Prob > F      =  0.0000
    Residual |  81835.9992   498  164.329316           R-squared     =  0.4675
-------------+------------------------------           Adj R-squared =  0.4665
       Total |      153692   499         308           Root MSE      =  12.819

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          3   .1434654    20.91   0.000     2.718128    3.281872
       _cons |   3.71e-08   .5732876     0.00   1.000     -1.12636     1.12636
------------------------------------------------------------------------------
```

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. X2 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 β_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 ($1 - R^2_{12}$) and N-K-1 will be smaller.

Example 3. This is similar to the first example, except that x2 has no effect on y.

```
. * Extraneous variables
. clear all
. matrix input corr = (1,.5,0\.5,1,0\0,0,1)
. matrix input sds = (4\4\10)
. corr2data x1 x2 e, corr(corr) sd(sds) n(500)
(obs 500)
. gen  y = 3*x1 + e
. corr y x1 x2
(obs=500)

             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.7682   1.0000
          x2 |   0.3841   0.5000   1.0000


. * Correct regression
. reg y x1

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =  717.12
       Model |  71856.0006     1  71856.0006           Prob > F      =  0.0000
    Residual |  49899.9991   498  100.200801           R-squared     =  0.5902
-------------+------------------------------           Adj R-squared =  0.5893
       Total |      121756   499  243.999999           Root MSE      =   10.01

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          3   .1120277    26.78   0.000     2.779895    3.220105
       _cons |  -6.22e-08   .4476624    -0.00   1.000    -.8795398    .8795397
------------------------------------------------------------------------------

. * Extraneous variable added
. reg y x1 x2

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  2,   497) =  357.84
       Model |  71856.0006     2  35928.0003           Prob > F      =  0.0000
    Residual |  49899.9991   497  100.402413           R-squared     =  0.5902
-------------+------------------------------           Adj R-squared =  0.5885
       Total |      121756   499  243.999999           Root MSE      =   10.02

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |          3   .1294885    23.17   0.000     2.745588    3.254412
          x2 |   7.70e-09   .1294885     0.00   1.000    -.2544123    .2544123
       _cons |  -6.22e-08   .4481125    -0.00   1.000    -.8804285    .8804284
------------------------------------------------------------------------------
```

As you can see the coefficient for x1 did not change but the standard error increased and the t value went down.

Appendix: Another example of omitted variable bias

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

```
. use https://www3.nd.edu/~rwilliam/statafiles/reg01.dta, clear
. corr educ jobexp income, cov
(obs=20)

             |     educ   jobexp   income
-------------+---------------------------
        educ |    20.05
      jobexp | -2.61316  29.8184
      income |  37.0676  14.3108  95.8119


. reg income educ jobexp

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) =   46.33
       Model |  1538.22521     2  769.112605           Prob > F      =  0.0000
    Residual |  282.200265    17  16.6000156           R-squared     =  0.8450
-------------+------------------------------           Adj R-squared =  0.8267
       Total |  1820.42548    19  95.8118671           Root MSE      =  4.0743

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.933393   .2099494     9.21   0.000     1.490438    2.376347
      jobexp |   .6493654   .1721589     3.77   0.002     .2861417    1.012589
       _cons |  -7.096855   3.626412    -1.96   0.067    -14.74792    .5542052
------------------------------------------------------------------------------
```

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

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

Stata confirms that this is correct:

```
. reg income educ

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =   45.21
       Model |  1302.05369     1  1302.05369           Prob > F      =  0.0000
    Residual |  518.371789    18  28.7984327           R-squared     =  0.7152
-------------+------------------------------           Adj R-squared =  0.6994
       Total |  1820.42548    19  95.8118671           Root MSE      =  5.3664

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    1.84876   .2749479     6.72   0.000     1.271116    2.426404
       _cons |   2.137446   3.523734     0.61   0.552    -5.265645    9.540537
------------------------------------------------------------------------------
```

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

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

Stata again confirms this:

```
. reg income jobexp

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =    1.39
       Model |  130.495675     1  130.495675           Prob > F      =  0.2538
    Residual |   1689.9298    18  93.8849889           R-squared     =  0.0717
-------------+------------------------------           Adj R-squared =  0.0201
       Total |  1820.42548    19  95.8118671           Root MSE      =  9.6894

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      jobexp |   .4799311   .4070792     1.18   0.254    -.3753106    1.335173
       _cons |   18.34387   5.586783     3.28   0.004     6.606476    30.08127
------------------------------------------------------------------------------
```

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.