PPOL502-01, Spring 2016
Course Notes #5: Omitted Variables and Irrelevant Variables in Multiple Regression

---

I. LET THE PROBLEMS BEGIN: OMITTED VARIABLES

- The police, crimes, and criminals example we talked about provides an opportunity to examine the problem of *omitted variables*. What's the problem?

- Remember MLR4—the assumption that E(u|X) = 0 (i.e., that the *X*s in the model and all other unobserved factors that are associated with *Y* were uncorrelated)?

- If this assumption *doesn't* hold, we should be concerned about the coefficient estimates that we obtained: in particular, they will be biased estimates of the true population parameters.

- The following gives you some framework for thinking more precisely about the bias in the coefficients that will result when key factors are omitted from the model. We'll start with a simple analysis, then generalize from that.

THE OMITTED VARIABLES FRAMEWORK

- Suppose we can characterize the true relationship in the population between some variable of interest, *Y*, and variables X_1 and X_2:

  "True" Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

- Suppose you are particularly interested in the relationship between X_1 and *Y*.

- However, for some reason (restricted data, laziness) you are not able to estimate this equation with data on X_1 and X_2. So instead you estimate a "restricted" model:

  Restricted Model: $Y = \gamma_0 + \gamma_1 X_1 \qquad + \eta$

- What can you know about your estimate of γ_1, that is, the association between X_1 and *Y* that you obtain from this restricted estimation?

  -- Wooldridge shows you the math on pp. 89-91 (and we will look at in Section III of these notes). The bottom line is that your estimate of γ_1 will be **a *biased*** estimate of β_1 ***if*** you do not include variables in the equation that should be in the "true" model (i.e., X_2 in this case):

$$E[\hat{\gamma}_1] \;=\; \beta_1 \;+\; \beta_2 * \left(\frac{S(X_1, X_2)}{S^2(X_1)} \right)$$

  -- where the last term in parenthesis is the covariance between X_1 and X_2, divided by the variance of X_1.

- This simple formula can provide powerful insight, especially for policy work, where we often aren't able to measure or include the variables that we think matter for explaining some dependent variable of interest, *Y*.

- You do *not* need to know the specific numerical values of the terms in the above expression in order for this to be useful. Instead, you only need to know (or in most cases make an educated guess) about the *signs* of those relationships. This process is called "signing the bias."

$$E[\hat{\gamma}_1] = \beta_1 + \beta_2 * \left(\frac{S(X_1, X_2)}{S^2(X_1)}\right)$$

- Think of the elements in the second term above in this way:

  $\beta_2 =$ partial correlation between X_2 and Y
  (i.e., how are X_2 [the omitted variable] and Y related?)

  -- Think back to the formula for the estimated slope coefficient for the regression with only one independent variable: $\hat{\beta}_1 = \dfrac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$

  -- [The formula for the actual multiple regression coefficient for β_2 is different, but the basic idea is the same]

  $\left(\dfrac{S(X_1, X_2)}{S^2(X_1)}\right) \approx$ correlation between X_1 and X_2 (i.e., how are X_1 and X_2 related?)

- Using this simplification, you can predict the direction of bias on the coefficient in the equation you actually estimate by multiplying together the last two terms: this give you the sign of the bias.

- HINT: when signing the bias, it is helpful to draw a vertical line to mark the slope estimate from the restricted regression, and what you predict it would be in the true model.

- When we sign the bias, we'll use the terms "upward" and "downward" biased, instead of "positive" or "negative" bias, for reasons discussed by Wooldridge on p. 92-93.

If the sign of the relationship between X_2 and Y (holding other Xs fixed), (i.e., β_2) **is…**	**And the sign of the relationship between X_1 and X_2,** (i.e., $S(X_1, X_2)$) **is…**	**Then the coefficient estimate γ_1 is biased….**

- Under what conditions will γ_1 be an unbiased estimator of β_1?

- Example 1: The police-criminals-crime example:

"True" model: $CRIME1 = \beta_0 + \beta_1 POLICE + \beta_2 CRIMINALS + u$

Restricted model: $CRIME1 = \gamma_0 + \gamma_1 POLICE + \eta$

- Actual estimated restricted model: $CRIME1HAT = 0.9849 + 0.7934 POLICE$

- *CRIMINALS* is an omitted variable in the estimated model (Regression 1-A). How is the coefficient on *POLICE* biased? Walk through the steps:

a. Holding POLICE fixed, what's the hypothesized relationship between *CRIMINALS* and *CRIME1*? (X_2 and Y)

b. What's the hypothesized relationship between *POLICE* and *CRIMINALS*? (X_1 and X_2)

c. So, is the coefficient estimate on POLICE in the restricted model biased ***upward*** or ***downward****?*

d. What was the actual estimated equation with both *POLICE* and *CRIMINALS*? (Reg 1-B)

e. Was your analysis in (c) correct?

- Example 2: the Wages, IQ, and Education equation

  "True" model: $WAGES = \beta_0 + \beta_1 IQ + \beta_2 EDUC + u$

  Restricted model: $WAGES = \gamma_0 + \gamma_1 IQ + \eta$

  - Actual estimated restricted model: $WAGEShat = 116.99 + 8.30IQ$

  - *EDUC* is an omitted variable in the estimated model. How is the coefficient on IQ biased?

    a. What's the hypothesized relationship between *EDUC* and *WAGES* (holding IQ fixed)? (X_2 and Y, controlling for X_1)

    b. What's the hypothesized relationship between *IQ* and *EDUC*? (X_1 and X_2)

    c. So, is the coefficient estimate on IQ in the restricted model biased ***upward*** or ***downward****?*

    d. What was the actual estimated equation with both *IQ* and *EDUC*?

    e. Was your predicted bias in (c) correct?

- Example 3: Cigs, family income, and birth weight

  "True" model: $BWGHT = \beta_0 + \beta_1 CIGS + \beta_2 FAMINC + u$

  Restricted model: $BWGHT = \gamma_0 + \gamma_1 CIGS + \eta$

  - Estimated restricted model: $BWGHThat = 119.77 - 0.514CIGS$

  - *FAMINC* is an omitted variable in the estimated model. How is the coefficient on CIGS biased?

    a. What's the hypothesized relationship between *FAMINC* and *BWGHT*? (X_2 and Y)

    b. What's the hypothesized relationship between *CIGS* and *FAMINC*? (X_1 and X_2)

    c. So, is the coefficient estimate on *CIGS* in the restricted model biased *upward* or *downward?*

II. OMITTED VARIABLES – ADDITIONAL COMMENTS

- The power of "omitted variable thinking" isn't so much in adding single variables or blocks of variables to a regression and watching how the coefficient estimates change (though we will do this to build your intuition and skills).

- The really useful thing to get out of this logic is to understand how your estimated coefficients may be affected by things you can't measure at all. This requires:

  (1) THINKING about what variables are associated with the dependent variable of interest;

  (2) THINKING about whether you have measures of those constructs in your data set (a measured variable with a name that looks promising may or may not be what you think it is);

  (3) If you *don't* have measures of the things (*Xs*) that you think affect *Y,* then

  > (a) THINKING about whether and how these omitted *Xs* are correlated with *Y*,
  >
  > (b) THINKING about whether and how these omitted *Xs* are correlated with *X*s that are actually in the estimated model,
  >
  > (c) THINKING AND DESCRIBING how the coefficients that you actually estimate may be affected by your inability to measure all the things that you think matter.

- This ____________________________ process indicates that you don't want to just add in variables to a regression model until you get the (statistically significant or insignificant) results that you "want" (for policy or political or pleasing-the-boss reasons).

- Instead, you want to:
  (1) describe the "true" model that you think applies,
  (2) estimate the best, complete model that you can, and then
  (3) discuss the interpretation of what you've estimated.

- Often, you'll see reports of a series of regressions. It may seem that what's going on it a search for the best model. In fact, what's often happening is that the author is showing the consequences of leaving out important, unmeasured factors from the model.

III. AN ALTERNATIVE (AND MORE GENERAL) WAY TO THINK ABOUT SIGNING THE BIAS USING AUXILIARY MODELS

- The method for signing the bias discussed in part I of these notes is derived from the following model. We show it here, to be more specific, and also to provide the framework for thinking about omitted variable bias in a more general sense (i.e., when more than one variable may be omitted from the true model). The setup is:

"True" Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

Restricted Model: $Y = \gamma_0 + \gamma_1 X_1 \qquad + \eta$

Auxiliary Model: $X_2 = \delta_0 + \delta_1 X_1 + \tau$

Substitute results form the auxiliary model into the "true" model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (\delta_0 + \delta_1 X_1 + \tau) + u$$
$$Y = (\beta_0 + \beta_2 \delta_0) + (\beta_1 + \beta_2 \delta_1) X_1 + (\beta_2 \tau + u)$$
$$Y = \quad \gamma_0 \quad + \gamma_1 X_1 \quad + \eta$$

<u>NOTES:</u>

1. The "True" and "Auxiliary" models are *conceptual* models, which we seldom are able to run. Why have them? Because they provide a formal structure that can help us think about what is missing from the model we actually estimate.

2. From the last line above, you can see: $\gamma_1 = \beta_1 + \beta_2 \delta_1$
   What is deltaonehat? It's the coefficient of the simple regression of X_2 on X_1, so:
   $\hat{\delta}_1 = \dfrac{\text{cov}\,ariance(X_1, X_2)}{\text{var}\,iance(X_1)} = \dfrac{s_{x1x2}}{s_{x1}^2}$. This maps back onto our formula from earlier.

IV. OMITTED VARIABLE BIAS – *k* regressors

- So far, we've only looked at examples with two explanatory variables in the "true" population regression model: X_1 and X_2. In reality, there may be many variables in this model.

"True model" $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \ldots \beta_k X_k + u$

Restricted model: $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 \quad + \gamma_4 X_4 + \ldots \gamma_k X_k + \varphi$

  - Does the basic intuition that we've already talked about still apply?

  -- *ALL* the coefficients in the estimated model generally will be biased, even if one of the *Xs* in the model is *uncorrelated* with X_3. An exception is when *all Xs* in the model are uncorrelated with X_3.

  -- Usually, you are interested in particular coefficients in the model (not all of them). As a *rough approximation* of the sign of the bias, you can still use the basic intuition that we already talked about (see Wooldridge p. 91-92 for further discussion).

- Or, we could be more precise by using the language of the auxiliary model:

"True model" $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + u$

Restricted model: $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 \quad + \varphi$

Auxiliary for X_3: $X_3 = \delta_0 + \delta_1 X_1 + \delta_2 X_2 + \tau$

Auxiliary for X_4: $X_4 = \kappa_0 + \kappa_1 X_1 + \kappa_2 X_2 + \theta$

Substitute in:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\delta_0 + \delta_1 X_1 + \delta_2 X_2 + \tau) + \beta_4(\kappa_0 + \kappa_1 X_1 + \kappa_2 X_2 + \theta) + u$$

$$Y = (\beta_0 + \beta_3\delta_0 + \beta_4\kappa_0) + (\beta_1 + \beta_3\delta_1 + \beta_4\kappa_1)X_1 + (\beta_2 + \beta_3\delta_2 + \beta_4\kappa_2)X_2 + (\beta_3\tau + \beta_4\theta + u)$$

$$Y = \quad \gamma_0 \quad + \quad \gamma_1 X_1 \quad + \quad \gamma_2 X_2 \quad + \quad \varphi$$

[Remember that the model that is actually estimated (the restricted model) cannot distinguish among each of the elements in parentheses!]

- The magnitude of the bias is determined by the magnitudes of the omitted relationships (e.g., the magnitude of the bias of γ_1 is determined by the magnitudes of β_3, δ_1, β_4, and κ_1)

Empirical Example: birthweight

TRUE MODEL:

Dependent Variable: bwght

```
. reg bwght cigs faminc motheduc
------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.4633487   .0927471    -5.00   0.000    -.6452888   -.2814086
      faminc |   .0914712   .0324594     2.82   0.005     .0277961    .1551462
    motheduc |   .0142561   .2579877     0.06   0.956    -.4918334    .5203456
       _cons |   116.8349   3.137782    37.23   0.000     110.6795    122.9902
------------------------------------------------------------------------------
```

RESTRICTED MODEL:

Dependent Variable: bwght

```
. reg bwght cigs faminc
------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.4641368   .0916108    -5.07   0.000    -.6438478   -.2844259
      faminc |    .092252   .0292113     3.16   0.002     .0349488    .1495553
       _cons |   116.9982   1.050233   111.40   0.000      114.938    119.0585
------------------------------------------------------------------------------
```

AUXILIARY MODEL:

Dependent Variable: motheduc

```
. reg motheduc cigs faminc
------------------------------------------------------------------------------
    motheduc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.0552854   .0095485    -5.79   0.000    -.0740165   -.0365542
      faminc |   .0547756   .0030447    17.99   0.000      .048803    .0607483
       _cons |    11.4605   .1094649   104.70   0.000     11.24577    11.67524
------------------------------------------------------------------------------
```

BIAS of CIGS coeff = -.4641368 – (-0.4633487) = -.0007881

"True" model: $BWGHT = \beta_0 + \beta_1 CIGS + \beta_2 FAMINC + \beta_3 MOTHEDUC + u$

Restricted model: $BWGHT = \gamma_0 + \gamma_1 CIGS + \gamma_2 FAMINC + \eta$

Auxiliary model: $MOTHEDUC = =\delta_0 + \delta_1 CIGS + \delta_2 FAMINC + \tau$

Substitute in:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3(\delta_0 + \delta_1 X_1 + \delta_2 X_2 + \tau) + u$$

$$Y = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)X_1 + (\beta_2 + \beta_3\delta_2)X_2 + (\beta_3\tau + u)$$

$$Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \varphi$$

BIAS of CIGS coeff = $\beta_3\delta_1$ **=** (.0142561)(-.0552854) = **-.00078815**

V. INCLUDING IRRELEVANT VARIABLES IN THE REGRESSION

- Should we include an explanatory variable X_k in the regression that we predict to have *no* partial effect on Y (i.e., $\beta_k = 0$)?

- Thinking about the omitted variable formula, there is no harm done (in terms of biased coefficients on the other variables in the model) when we do include such irrelevant variables.

- In terms of efficiency—the estimated standard errors on coefficients—however, there is a cost to including these irrelevant variables. Will talk about std errors in course notes7-8.

VI. ADDING REGRESSORS TO REDUCE THE ERROR VARIANCE

- Should we include variables (Z_k) that are related to Y, but are *not* related to the Xs in the model? Leaving such Z_k out of the model will not bias the coefficients of the other variables in the model, so we're o.k. there.

- However, including such Z_k in the model will often have an added benefit: it will contribute to the explanatory power of the model, reduce the error variance (i.e., $\hat{\sigma}_2$). So, there are advantages to include variables in a regression that are correlated with Y, but not with any of the other independent variables. Will talk about std errors in course notes 7-8.