Lecture 12

Specification Analysis

The consequences of

1) Omitting a Relevant Variable
2) Including an Irrelevant Variable

1) Assume the correct (true) model is:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad (1)$$

but instead you estimate the model

$$y_i = \beta_0 + \beta_1 X_{i1} + u_i . \qquad (2)$$

Now

$$\hat{\beta}_1 = \frac{\sum_1^n (X_{i1} - \bar{X}_1) y_i}{\sum_1^n (X_{i1} - \bar{X}_1)^2}$$

Therefore

$$\hat{\beta}_1 = \frac{\sum_1^n (X_{i1} - \bar{X}_1)(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i)}{SST_1}$$

where $SST_1 = \sum_1^n (X_{i1} - \bar{X}_1)^2$

Then

$$\hat{\beta}_1 = \beta_0 \frac{\sum_1^n (X_{i1} - \bar{X}_1)}{SST_1} + \beta_1 \frac{\sum_1^n (X_{i1} - \bar{X}_1) X_{i1}}{SST_1}$$

$$+ \beta_2 \frac{\sum_1^n (X_{i1} - \bar{X}_1) X_{i2}}{SST_1} + \frac{\sum_1^n (X_{i1} - \bar{X}_1) u_i}{SST_1}$$

$$= \beta_0 \cdot \frac{0}{SST_1} + \beta_1 \frac{SST_1}{SST_1} + \beta_2 \frac{\sum_1^n (X_{i1} - \bar{X}_1) X_{i2}}{SST_1}$$

$$+ \frac{\sum_1^n (X_{i1} - \bar{X}_1) u_i}{SST_1}$$

Now taking the expectation we have

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_1^n (X_{i1} - \bar{X}_1) X_{i2}}{SST_1} + \frac{\sum_1^n (X_{i1} - \bar{X}_1) E u_i}{SST_1}$$

$$= \beta_1 + \beta_2 \frac{\sum_1^n (X_{i1} - \bar{X}_1) X_{i2}}{SST_1}$$

$$= \beta_1 + \beta_2 \tilde{\delta}_1$$

where $\tilde{\delta}_1$ is the slope coefficient estimate obtained by applying ordinary least squares to the auxillary equation

$$X_2 = \delta_0 + \delta_1 X_1 + v.$$

The term $\beta_2 \tilde{\delta}_1$ is called the omitted variable bias.

Consider the Following table which tells us the bias of using ordinary least squares to estimate β_1 in the misspecified equation (2) when in fact X_2 should have been included (see eq. (1)).

Table 3.2 in Wooldridge

	$corr(X_1, X_2) > 0$	$Cov(X_1, X_2) < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

Now consider the case where the true model is

$$y_i = \beta_0 + \beta_1 X_{i1} + u_i . \qquad (3)$$

However, suppose that you instead estimate the following model that includes the irrelevant variable X_2

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i . \qquad (4)$$

Let $\hat{\beta}_1$ be the ordinary least squares estimator of β_1 in equation (3) and $\tilde{\beta}_1$ and $\tilde{\beta}_2$ be, respectively, the ordinary least squares estimators of β_1 and β_2 in equation (4).

Then it can be shown that

$E(\tilde{\beta}_2) = 0 \quad \therefore$ unbiased

$E(\tilde{\beta}_1) = \beta_1 \quad \therefore$ unbiased

but

$$Var(\tilde{\beta}_1) \geq Var(\hat{\beta}_1)$$

and the ordinary least squares estimator of β_1 obtained from the overspecified equation is <u>inefficient</u>. The inefficiency of $\tilde{\beta}_1$ is given by the equation

$$\frac{Var(\tilde{\beta}_1)}{Var(\hat{\beta}_1)} = \frac{1}{1 - r_{12}^2}, \quad 0 \leq r_{12}^2 \leq 1 \quad -1 \leq r_{12} \leq 1$$

where r_{12} is the sample correlation between X_1 and X_2, namely

$$r_{12} = \frac{\sum_{1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{\sqrt{\sum_{1}^{n}(X_{i1}-\bar{X}_1)^2 \sum_{1}^{n}(X_{i2}-\bar{X}_2)^2}}.$$

Note that $Var(\tilde{\beta}_1) = Var(\hat{\beta}_1)$ only when $r_{12} = 0$ and X_1 and X_2 are orthogonal to each other.