Lecture 5: Omitted Variables, Dummy Variables and Multicollinearity

R.G. Pierse

1 Omitted Variables

Suppose that the 'true' model is

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \quad , \quad i = 1, \cdots, n \tag{1.1}$$

where $\beta_3 \neq 0$ but that the researcher mistakenly omits the variable X_3 and estimates the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \, . \tag{1.2}$$

What is the effect of omitting X_3? The OLS estimator of β_2 in (1.2) is

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n x_{2i} y_i}{\sum_{i=1}^n x_{2i}^2}$$

and, substituting into (1.1)

$$\widehat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^n x_{2i} x_{3i}}{\sum_{i=1}^n x_{2i}^2} + \frac{\sum_{i=1}^n x_{2i} u_i}{\sum_{i=1}^n x_{2i}^2} \, .$$

Taking expectations,

$$\mathrm{E}(\widehat{\beta}_2) = \beta_2 + \frac{\sum_{i=1}^n x_{2i} x_{3i}}{\sum_{i=1}^n x_{2i}^2} \, .$$

In general, the second term will not be zero so that $\mathrm{E}(\widehat{\beta}_2) \neq \beta_2$ and the estimator is *biased*. Only in the special case that

$$\widehat{\mathrm{cov}}(X_2, X_3) = \frac{\sum_{i=1}^n x_{2i} x_{3i}}{n-1} = 0$$

will $\widehat{\beta}_2$ be unbiased. This is where the omitted variable is completely uncorrelated with the included regressors.

Consider the residuals from the regression (1.2) which are

$$\begin{aligned} e_i &= y_i - \widehat{\beta}_2 x_{2i} \\ &= x_{3i}\beta_3 - \frac{x_{2i}\sum_{i=1}^n x_{2i}x_{3i}}{\sum_{i=1}^n x_{2i}^2} + \frac{x_{2i}\sum_{i=1}^n x_{2i}u_i}{\sum_{i=1}^n x_{2i}^2}\,. \end{aligned}$$

Note that $\mathrm{E}(e_i) \neq 0$, *even in the special case* that $\sum_{i=1}^n x_{2i}x_{3i} = 0$.

Misspecification due to omitting a variable leads to *biased* estimators and to residuals which will exhibit a systematic pattern. This will often be reflected in evidence of *significant serial correlation*.

2 Including Redundant Variables

Suppose the reverse situation to that of the last section. The 'true' model is now

$$Y_i = \beta_1 + \beta_2 X_{2i} + u_i \tag{2.1}$$

but the researcher mistakenly includes the *redundant* variable X_3 and estimates the model

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i\,. \tag{2.2}$$

What is the effect of including X_3 when in fact $\beta_3 = 0$? Estimating (2.2) results in the estimators

$$\widehat{\beta}_2 = \frac{\sum y_i x_{2i} \sum x_{3i}^2 - \sum y_i x_{3i} \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}$$

and

$$\widehat{\beta}_3 = \frac{\sum y_i x_{3i} \sum x_{2i}^2 - \sum y_i x_{2i} \sum x_{2i}x_{3i}}{\sum x_{2i}^2 \sum x_{3i}^2 - \left(\sum x_{2i}x_{3i}\right)^2}\,.$$

and it is easy to show that

$$\mathrm{E}(\widehat{\beta}_2) = \beta_2 \quad \text{and} \quad \mathrm{E}(\widehat{\beta}_3) = \beta_3 = 0\,.$$

In this case, the estimators are *unbiased*. However, because a redundant variable has been included, the estimated variances will be larger than those of the $BLUE$ estimators from the 'true' model (2.1) and estimated standard errors will be larger than they should be.

Thus *including an irrelevant variable is far less serious an econometric problem than excluding a relevant variable*.

3 Dummy Variables

Sometimes there may be variables that affect the dependent variable but that cannot be readily quantified. One way of incorporating such effects into the regression model is by the use of *dummy variables*. The non-quantifiable effect is represented by a variable that takes the value either of one or zero; one representing the presence of the effect and zero its absence. Examples of effects that are often proxied by dummy variables are wars, seasonal effects or dichotomous variables such as gender.

3.1 Dichotomous dummies

Suppose a researcher suspects that the relationship between earnings and age is affected by a person's sex. This could be tested by running the regression:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

where Y_i is earnings, X_i is age, and D_i is a dummy variable taking the value 1 for a woman, and 0 for a man. Then a significant t-ratio for $\widehat{\beta}_3$ would indicate that gender is a significant factor in the relationship. Considering the two groups of observations $D_i = 1$ and $D_i = 0$ separately, we have that:

$$Y_i = (\beta_1 + \beta_3) + \beta_2 X_i + u_i \quad , \quad D_i = 1$$

and

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad , \quad D_i = 0\,.$$

Thus the dummy coefficient β_3 allows a different intercept term for women and for men. If the estimated coefficient $\widehat{\beta}_3$ were found to be significantly positive it would indicate that, on average, women earn $\widehat{\beta}_3$ more than men at all ages. (*Vice versa* for a significant negative coefficient.)

Suppose instead, that the researcher suspected that gender affects, not the intercept but the *slope* of the relationship between age and earnings. Consider the alternative regression

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i X_i + u_i$$

where $D_i X_i$ is the product of the dummy variable and age, and is treated as a separate variable. Considering the two groups of observations $D_i = 1$ and $D_i = 0$ separately, we have that:

$$Y_i = \beta_1 + (\beta_2 + \beta_3) X_i + u_i \quad , \quad D_i = 1$$

and

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad , \quad D_i = 0\,.$$

Now a significant value for the dummy coefficient β_3 represents a different slope term for the two groups.

3.2 One-off dummies

Dummies are also used to pick up the effects of one-off events such as wars or strikes. A dummy variable that takes a non-zero value in *only one single observation*, allows the regression to explain that observation perfectly, so that $e_i = 0$ for the dummy observation. That observation will have no influence on the other estimated parameters. One-off dummies can thus be used to eliminate the effect of *outliers* on the regression.

3.3 Interactive dummies

When there is more than one dummy variable, then it is possible that the combined effect of both dummies is more than the sum of the individual effect of each alone. This can be allowed for by considering interactive dummies. Suppose there are two dummies D_1 and D_2. Then the product $D_1 D_2$ represents the interaction between the two and is only non-zero when both D_1 and D_2 are non-zero. If this product is significant in addition to the individual dummies, then this means that interactive effects are important.

3.4 Seasonal dummies

Many economic variables are highly seasonal. The seasonal effects in the dependent variable are often proxied by a set of seasonal dummies for the s seasons

$$\begin{aligned} D_i &= 1 \text{ in season } i \quad , \quad i = 1, \cdots, s-1 \\ &= 0 \text{ otherwise} \end{aligned}$$

Note that only $s-1$ seasonal dummies are needed. This is because $\sum_{i=1}^{s} D_i = 1$ and so would be *perfectly collinear* with the intercept. In general, when there are q categories into which the dependent variable can fall, then only $q-1$ dummy variables are needed. This avoids the so-called *dummy variable trap*.

3.5 Pooled Time Series-Cross-Sectional data

Dummy variables are very useful in dealing with data sets that combine time-series with cross-sectional data such as the model

$$Y_{it} = \beta_1 + \beta_2 X_{it} + u_{it} \quad , \quad i = 1, \cdots, n \quad t = 1, \cdots, T.$$

This model assumes that the parameters β_1 and β_2 are the same over all time periods and all cross-sectional units. Suppose we suspect that the intercept term

may differ over cross-sectional units. Then we can estimate the model

$$Y_{it} = \beta_1 + \beta_2 X_{it} + \sum_{j=1}^{n-1} \delta_j D_{it}^j + u_{it}$$

where the $n-1$ variables D^j, $j = 1, \cdots, n-1$ take the value 1 where $i = j$, and zero otherwise. The hypothesis that the intercept is the same for all cross-sections can be tested by a joint test of

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_{n-1} = 0\,.$$

4 Multicollinearity

One of the assumptions of the Classical model is that

The variables X_j are *not perfectly collinear.* (A4′)

Perfect collinearity occurs when there is one or more variables X_m such that

$$X_{mi} = \sum_{j \neq m} c_j X_{ji} \quad , \quad i = 1, \cdots, n$$

where c_j are fixed constants. Consider the special case where $k = 3$:

$$X_{3i} = c_1 + c_2 X_{2i}$$

or, subtracting sample means,

$$x_{3i} = c_2 x_{2i}$$

so that

$$\sum x_{2i} x_{3i} = c_2 \sum x_{2i}^2 = \frac{1}{c_2} \sum x_{3i}^2$$

and the squared sample correlation coefficient between X_{2i} and X_{3i} is

$$r_{23}^2 = \frac{(\sum x_{2i} x_{3i})^2}{\sum x_{2i}^2 \sum x_{3i}^2} = 1\,.$$

What happens if we attempt to estimate the model by OLS ? The formulae for the variances of the OLS estimators can be written as

$$Var(\widehat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \quad \text{and} \quad Var(\widehat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

so that, in this case, the variance is infinite and the *OLS estimators cannot be computed.*

What happens if the collinearity is 'high' but less than perfect? Then clearly the variance of OLS estimator will also be 'high'. This is known as the problem of *multicollinearity.* However, as Kmenta (1986) has stated

> Multicollinearity is a question of degree and not of kind. The meaningful distinction is not between the presence and the absence of multicollinearity, but between its various degrees.

4.1 Recognising Multicollinearity

Two accepted symptoms of multicollinearity are (i) 'High R^2 but few significant t ratios' and (ii) 'High correlations among regressors'. However, both these measures can be affected by a simple renormalisation of the regression. For example, consider the simple model

$$\widehat{Y} = \widehat{a}_1 X_1 + \widehat{a}_2 X_2$$

where

$$\begin{bmatrix} \widehat{a}_1 \\ \widehat{a}_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad , \quad Var \begin{bmatrix} \widehat{a}_1 \\ \widehat{a}_2 \end{bmatrix} = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

so that $t(\widehat{a}_1) = 2$ and $t(\widehat{a}_2) = 1$ and the latter is insignificant. This model can be renormalised as

$$\widehat{Y} = \widehat{b}_1 (X_1 - X_2) + \widehat{b}_2 X_2$$

where

$$\begin{bmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{bmatrix} = \begin{bmatrix} \widehat{a}_1 \\ \widehat{a}_1 + \widehat{a}_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \quad , \quad Var \begin{bmatrix} \widehat{b}_1 \\ \widehat{b}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0.1 \\ 0.1 & 0.2 \end{bmatrix}$$

and both renormalised coefficients now have significant t-ratios. The renormalisation has no effect on R^2 but reduces the correlation between the regressors from -0.9 to 0.22. This example shows that neither of the standard ways of recognising multicollinearity is foolproof.

Multicollinearity is essentially a *sample* problem. The theoretical properties of OLS estimators are not affected by multicollinearity. In practice, however, it may make inference difficult in a particular sample. For example it will be very difficult to find a significant coefficient on an economic variable in a sample in which that variable is not changing significantly. The reason is that the variable will appear highly collinear with a constant. Similarly, if all regressors are moving in a similar way, it will be hard to attribute the effects of each one individually even though, jointly, they should be significant.

4.2 Degrees of Freedom

A related problem to that of multicollinearity is that of insufficient degrees of freedom. In the limit