ECON-122
Introduction to Econometrics

Agnieszka Postepska

July 21st, 2011

Omitted Variable Bias

In a two variable population model, the bias caused by omission of the second explanatory variable can be computed as follows:

Expression for the bias:

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \frac{\sum_i (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{\sum_i (x_{1i} - \bar{x_1})^2} = \beta_2 \hat{\delta}_1$$

where δ_1 correlation coefficient from a simple regression of x_{2i} on x_{1i}

Two cases are possible:

- **Upward (positive) bias**: $E(\hat{\beta}) - \beta > 0$
- **Downward (negative) bias**: $E(\hat{\beta}) - \beta < 0$

If x_1 and x_2 are uncorrelated, such that, $\beta_2 = 0$ the model is misspecified but the estimates are still unbiased.
If x_1 and x_2 are correlated, such that, $\beta_2 \neq 0$ the model is misspecified and the estimates are biased.

More than two explanatory variables in the model

- how can we extend the formula for the bias to allow for as many $x's$ as we want?
- Consider the following model:

$$wage = \beta_0 + \beta_1 education + \beta_2 ability + \beta_3 experience + u$$

- and again, we observe education and experience but not ability, so we estimate the following model:

$$\tilde{wage} = \tilde{\beta}_0 + \tilde{\beta}_1 education + \tilde{\beta}_3 experience$$

- does exclusion of ability influence estimates of both coefficients, β_1 and β_2?
  $\Rightarrow$ unfortunately, generally yes - even if the variable is uncorrelated with the omitted variable
- **unless a variable is uncorrelated with ALL RHS variables** (including the omitted variable), **its coefficient will be bias if the model is misspecified**

Expression for the bias in multiple regression model

Consider the true model:

$$y = X\beta + Z\alpha + u$$

Then:

$$\begin{aligned} \hat{\beta}_{OLS} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + Z\alpha + u) \\ &= \beta + (X'X)^{-1}X'Z\alpha + (X'X)^{-1}X'u \\ \Rightarrow E(\hat{\beta}_{OLS}) &= \beta + (X'X)^{-1}X'Z\alpha \end{aligned}$$

Expression for the bias with any number of regressors:

$$E(\hat{\beta}_{OLS}) - \beta = (X'X)^{-1}X'Z\alpha$$

And $X'Z$ is the variance-covariance matrix of included and omitted regressors which explains why omitted variable generally biases all coefficient.

Many regressors in one model-introduction to multicollinearity

- Recall the formula for the error variance:

$$Var(\hat{\beta}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- The variance of an individual coefficient can be written as (derivation on the blackboard):

$$Var(\hat{\beta}_j^{OLS}|\mathbf{X}) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

- 3 elements of the variance:
  - σ^2 - feature of the population - the only way to decrease σ is to add explanatory variables to the model - as this takes things out of the error term
  - SST_j - total sample variation in x_j - this we can increase by increasing the sample size - and we want a lot of variation in $x's$ as this increases the precision of estimates
  - R_j^2 - linear relationship between all explanatory variables in the model

Closer look at R_j^2

- this R_j^2 is obtained from regressing all independent variables on x_j - so here x_j acts as y in normal setting)
- in other words, R_j^2 comes from the following model:

$$x_{ji} = \alpha X_{-ji} + u_i$$

- if the explanatory variables in the model are highly correlated, such that R_j^2 is high, then $(1 - R_j^2)$ is small causing the variance of $\hat{\beta}_j$ to be large - thus we get imprecise estimate
- when R_j^2 is "large" or "close to 1" we have **multicollinearity** problem
- not well defined concept and no easy way to correct it...

Include or not to include...

- no easy answer
- usually data is the main constraint
- one can run some specification tests but it really should be the theory not the data that tells us what to include in a model
- also, often we are facing the unbiasedness vs. efficiency trade-off
  - if theory tells us we should include two very correlated variables it might be better to exclude one and get a biased (why biased?) but more precise (why more precise?) estimates
  - it is often the case that researcher must scarify some bias for the sake of precision
- one solution to this problem is to group variables - one cannot identify partial effects then but gets more precise estimates
- GOOD NEWS: precision of one estimate is not affected by multicollinearity among other, **uncorrelated** variables (why?)

Treatment of qualitative regressors

- **indicator variables** are used to capture qualitative variables and characteristics
- common examples: gender (male/female), presence of children (yes/no), employment status(employed/unemployed), marital status (married/divorced/separated/single/widowed), countries of birth etc.
- include in the model as any other regressor but be careful with interpretation:
  - for gender example - include female dummy - takes on value 1 for females and 0 for men
  - the omitted category is always the reference group
  - reference group is the group to which every other group is compared in estimation

Interaction terms

- sometimes we might also suspect that some characteristic changes the influence of another variable - for example effect of schooling on wages is different for women than for men - **interaction term** captures this effect in the model:

$$wage_i = \beta_0 + \beta_1 schooling_i + \beta_2 female_i + \beta_3 female_i * schooling_i + u_i$$

- Interpretation:
  - β_0 - effect of schooling on wages on men with 0 years of education
  - β_2 - effect of schooling on wages for females with 0 years of education
  - $\beta_1 + \beta_3$ - effect of one more year of edu on wages for women
  - β_1 - effect of one more year of edu on wages for men