

Chapter 6

Specification: Choosing the Independent Variables

Masao Ogaki

Department of Economics,
Ohio State University

October 30, 2007

Omitted Variables

An *omitted variable* means that an important explanatory variable that has been left out of a regression equation.

The bias caused by a leaving a variable out of a equation is called *omitted variable bias*.

Suppose that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \quad (1)$$

If you omit X_2 ,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i^* \quad (2)$$

where ϵ_i^* equals

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i} \quad (3)$$

As long as X_{1i} and X_{2i} are correlated, this causes Classical Assumption III to be violated in the Regression (??).

Assumption III: All explanatory variables are uncorrelated with the error term.

This lead to a bias in the OLS estimator:

$$E(\hat{\beta}_1) \neq \beta_1 \quad (4)$$

Exercises

(a) Consider a production function that states that output (Y) depends on the amount of labor (X_1) and capital (X_2) used. Suppose that labor and capital are positively correlated. What would happen to the slope coefficient on labor if data on capital were unavailable and X_2 was omitted?

(b) In the example of Studenmund's Section 6.1.2, use the t -test for the null hypothesis that its coefficient is zero. Use either one-tailed or two-tailed test. Which one do you think is better? Why? In terms of \bar{R}^2 , which regression do you prefer, (6.8) or (6.9)? Do you think that PB_t is an omitted variable in Regression (6.9)? Why?

Irrelevant Variables

An *irrelevant variable* means that an explanatory variable that has been included in a regression equation but does not belong there.

The addition of an irrelevant variable does not cause bias, but it increases the variances of the estimated coefficients of the other variables.

Suppose that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \quad (5)$$

but the researcher includes an irrelevant variable

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^* \quad (6)$$

$$\epsilon_j^* = \epsilon_j - \beta_2 X_{2j} \quad (7)$$

X_2 is an irrelevant variable if β_2 is zero. In this case, the stochastic error does not change, and including an irrelevant variable does not cause bias. However, the variance for $\hat{\beta}_1$ increases.

Four important specification criteria:

- ▶ *Theory*: Is the variable's place in the equation unambiguous and theoretically sound?
- ▶ *t*-test: Is the variable's estimated coefficient significant in the expected direction?
- ▶ \bar{R}^2 : Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
- ▶ *Bias*: Do other variables' coefficients change significantly when the variable is added to the equation?

Exercises

- (a) In the example of Section 6.2.2, evaluate R_t as a possible irrelevant variable according to the four criteria above.
- (b) In the example of Section 6.1.2., evaluate PB_t as a possible irrelevant variable according to the four criteria above.

Specification Searches

In practice, we need to search for a decent specification.

Example: in the example in Studenmund's Section 6.5, the college GPA is the dependent variable. Do we include the high school GPA ($HGPA_i$), the math section of the SAT test ($MSAT_i$), the verbal section of the SAT test ($VSAT_i$),

$SAT_i = MSAT_i + VSAT_i$, the i th student's estimate of the average number of hours spent studying per course per week in college (HRS_i), the natural log of the number of full courses that the i th student has completed in college ($lnEX_i$), etc.?

- ▶ We should not be searching and searching for the best-looking equation
- ▶ *Data mining*: searching for the best specification by exhausting all possible equations that can be estimated
- ▶ Data mining in econometrics is a bad thing: this invalidates inference

Recommendation:

- ▶ Try a sensible specification first, and only change it if this is what the data suggest

Example: SAT scores may be deemed irrelevant given the high school GAP: leave them out, and perhaps try a regression with them included to see if they are significant

- ▶ Then if t -values or \bar{R}^2 suggest they are relevant, change the original plan

Exercise

Compare Regression (6.18) and (6.19). Do you think that SAT_i should be included in the regression? Explain.