Introductory Econometrics
Endogeneity

by Hieu Nguyen

Fall 2024

1.

Suppose that you wish to estimate the effect of class attendance on student performance. A model to explain standardized outcome on a final exam (stndfnl) in terms of percentage of classes attended (attnd), prior college Grade Point Average (priGPA), and American College Testing score (ACT) is:

$$\text{stndfnl} = \beta_0 + \beta_1 \text{attnd} + \beta_2 \text{priGPA} + \beta_3 \text{ACT} + \epsilon.$$

(a) Why might attnd be suspected to be endogenous in the model?

(b) Let dist be the distance from the students' living quarters to the lecture hall. Do you think dist is uncorrelated with ϵ?

(c) Assuming that dist and ϵ are uncorrelated, what other assumption must dist satisfy in order to be a good instrument for attnd?

(d) Suppose we add the interaction term priGPA $\cdot$ attnd to the model:

$$\text{stndfnl} = \beta_0 + \beta_1 \text{attnd} + \beta_2 \text{priGPA} + \beta_3 \text{ACT} + \beta_4 \text{priGPA} \cdot \text{attnd} + u.$$

If attnd is correlated with ϵ, then, in general, so is priGPA $\cdot$ attnd. What might be a good instrument candidate for priGPA $\cdot$ attnd?

2.

The data in `fertil2.gdt` includes, for a sample of women in Botswana during 1988, information on the number of children, years of education, age, and religious and economic status variables.

(a) Estimate this model by OLS and briefly comment on results:

$$\text{children} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{age} + \beta_3 \text{age}^2 + \epsilon.$$

If 100 women receive another year of education, how many fewer children are they expected to have?

(b) In lecture #10, we discussed why we might suspect educ to be endogenous in this model. We also suggested `frsthalf` (a dummy variable equal to one if the woman was born in the first six months of a year) to be a good candidate for an instrument for educ. Show its relevance via a first stage regression. Assume that `frsthalf` is uncorrelated with the error term ϵ. Now estimate the model from part (a) by using `frsthalf` as an instrument for educ (= IV estimator, 2SLS). Compare the estimated effect of education with the OLS estimate. Which of the estimators is consistent?

(c) Add the binary explanatory variables `electric`, `tv`, and `bicycle` to the model and assume these are exogenous as well. Estimate the equation by 2SLS directly in Gretl and compare the estimated coefficient of educ with part (b) and with the OLS estimate. Interpret the output of the Hausman test.

3.

A researcher estimated by OLS two specifications of a regression model:

$$y = \alpha + \beta x_1 + \epsilon,$$

$$y = \tilde{\alpha} + \tilde{\beta} x_1 + \tilde{\gamma} x_2 + \tilde{\epsilon}$$

Explain theoretically under what circumstances the following will be true. If some case cannot be true, explain why.

(a) $\hat{\beta} = \hat{\tilde{\beta}}$.

(b) β is statistically significant (at the 5

(c) $\tilde{\beta}$ is statistically significant (at the 5% level) but β is not.

(d) If $\hat{\epsilon_i}$ and $\hat{\tilde{\epsilon_i}}$ are the estimated residuals from the two equations,

$$\sum_{i=1}^{n} \hat{\epsilon_i}^2 \geq \sum_{i=1}^{n} \hat{\tilde{\epsilon_i}}^2.$$