Seminar 9

Solutions to Endogeneity: Two Stage Least Squares

1. Import the Stata data file `"Card"` from the e-course platform.

   (a) In Table 15.1 from slide 20 of Lecture 6, the difference between the IV and OLS estimates of the return to education is economically important. Obtain the reduced form residuals, $\hat{v_2}$ from the reduced form regression *educ* on $nearc4, exper, exper^2, black, smsa, south, reg662, ..., reg669$. Use these to test whether *educ* is exogenous; that is, determine if the difference between OLS and IV is statistically significant.

   (b) Estimate the equation by 2SLS, adding $nearc2$ as an instrument. Does the coefficient on *educ* change much?

   (c) Test the single overidentifying restriction from part (b).

2. The purpose of this exercise is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Import the Stata data file `"wage2"` from the e-course platform.

   (a) Use a 2SLS routine to estimate the equation

   $$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 black + u,$$

   where *sibs* is the IV for *educ*.

   (b) Now, manually carry out 2SLS. That is, first regress $educ_i$ on $sibs_i, exper_i, tenure_i$ and $black_i$ and obtain the fitted values, $\hat{educ_i}, i = 1, ..., n$. Then, run the second stage regression $log(wage)$ on $\hat{educ_i}, exper_i$, $tenure_i$, and $black_i, i = 1, ..., n$. Verify that the $\hat{\beta_j}$ are identical to those obtained from part (a), but that the standard errors are somewhat different. The standard errors obtained from the second stage regression when manually carrying out 2SLS are generally inappropriate.

   (c) Now, use the following two-step procedure, which generally yields inconsistent parameter estimates of the β_j, and not just inconsistent standard errors. In step one, regress $educ_i$ on $sibs_i$ only and

obtain the fitted values, say $\tilde{educ}_i$. (Note that this is an incorrect first stage regression). Then, in the second step, run the regression of $log(wage_i)$ on $\tilde{educ}_i, exper_i, tenure_i$, and $black_i, i = 1, ..., n$. How does the estimate from this incorrect, two-step procedure compare with the correct 2SLS estimate of the return to education?

3. Import the Stata data file `"htv"` from the e-course platform.

   (a) Run a simple OLS regression of $log(wage)$ on *educ*. Without controlling for other factors, what is the 95% confidence interval for the return to another year of education?

   (b) The variable *ctuit*, in thousands of dollars, is the change in college tuition facing students from age 17 to age 18. Show that *educ* and *ctuit* are essentially uncorrelated. What does this say about *ctuit* as a possible IV for *educ* in a simple regression analysis?

   (c) Now, add to the simple regression model in part (a) a quadratic in experience and a full set of regional dummy variables for current residence and residence at age 18. Also include the urban indicators for current and age 18 residences. What is the estimated return to a year of education?

   (d) Again, using *ctuit* as a potential IV for *educ*, estimate the reduced form for *educ*. [Naturally, the reduced form for *educ* now includes the explanatory variables in part (c)]. Show that *ctuit* is now statistically significant in the reduced form for *educ*.

   (e) Estimate the model from part (c) by IV, using *ctuit* as an IV for *educ*. How does the confidence interval for the return to education compare with the OLS CI from part (c)?

   (f) Do you think the IV procedure from part (e) is convincing?