Applied Exercise #3

– Due Friday 03/21/08 –

This exercise examines the return to education on wages in an instrumental variables framework, using college proximity as an instrument for education. It will be helpful to read the paper by David Card (1995) in the reader as you do this problem set. Feel free to work in groups. Each student must hand in their own solution to the problem set using own words to explain and interpret the results.

Data: The data are drawn from the NLS Young Men Cohort. We have a sample of 14-24 year old men for whom the data contains information about family background, living arrangements, and region of residence in 1966. A follow up survey in 1976 provides information on educational attainment and earnings for the same individuals (for details see Card 1995). Access the data on bspace site for Econ C142: problem set 3/ nls.dta (STATA 9 format). The observational unit of the data is the individual. The data set has 3010 observations.

The key variables for this exercise are:

- **lwage76** – natural log of hourly wage in 1976
- **age76** – age of the person in 1976
- **black** – indicator variable equal to 1 if the person is black, zero otherwise
- **ed76** – the individual’s completed years of schooling in 1976
- **daded** – father’s completed years of schooling; set equal to sample mean if missing
- **nodaded** – father’s education imputed
- **mommed** – mother’s completed years of schooling; set equal to sample mean if missing
- **nomommed** – mother’s education imputed
- **famed** – father and mother educational class 1-9
- **momdad14** – lived with both parents at age of 14
- **sinmom14** – lived with single mother at age of 14
- **step14** – lived with a step parent at age of 14
- **nearc4** – grew up near a 4 year college
nearc2 – grew up near a 2 year college  
kww – Knowledge of the world of work test score  
south66 – lived in the South in 1966  
smsa66 – lived in a metropolitan area in 1966  
reg661-reg669 – dummies for region of residence in 1966  
south76 – lived in the South in 1976  
smsa76 – lived in a metropolitan area in 1976

Question 1: Omitted variables bias and proxy variables:
   a) Explain how omitting unobservable information about “ability” could bias the OLS estimate of the return of education on wages. What is your expectation of the direction of the bias? Explain how you could use a proxy variables approach to reduce the omitted variables bias.
   b) Generate a measure for potential experience, as exper=age76-ed76-6 and estimate two regression models: First, regress log-wage on education, experience, experience squared, black, and region indicators. Second, use the same regression models and add family background characteristics, like mother’s and father’s education as proxy variables for unobservable ability. Interpret the difference in the coefficient on education in both specifications.
   c) Now add the kww test score as proxy for ability. How does the regression coefficient on education change in comparison to the models estimated in (a)? Check the correlation coefficient between the kww test score and education. Could this correlation be a problem if there is measurement error in reported education?
   d) Summarizing the results from b) and c), what do proxy variables tell you about the severity of the ability bias problem in this application?

Question 2: Instrumental variables estimation
   a) Explain how the omitted variables problem in the wage equation could be adressed with an instrumental variables estimator. What are the conditions for a valid instrumental variable for education? Argue why college proximity might be a good instrument for education. What are potential disadvantages of college proximity as instrument?
   b) For the wage and education example, explain the difference between a proxy variable and an instrumental variable.
c) Regress ed76 on nearc4, experience, experience squared, black, region- and family background indicators. How can you use this result to assess the relevance of college proximity as an instrument for education?

d) Estimate the instrumental variables regression using the STATA ivreg command with nearc4 as an instrument for ed76, and the same additional regressors as in c). Compare the return to education to the results you found in Question 1. How can you interpret the difference?

e) Use the predicted residuals from the reduced form regression you estimated in c) to construct a test of endogeneity: regress lwage76 on ed76, the other regressors, and the predicted residuals from the reduced form. Interpret the coefficient on the predicted residuals variable. What does this test tell you about the endogeneity of education in the log wage equation?

[Note: an alternative to perform a Hausman-Wu test for endogeneity, is available in STATA by typing ivendog after the running the IV regression. ivendog is a program which you have to install separately on your computer, type search ivendog and follow the instructions. This does not work on the Socrates server.]

Question 3: Two stage least squares

a) Now use the two indicators for proximity of a 4-year college and a 2-year college as instrumental variables for education. Start with the estimation of the reduced form equation by regressing ed76 on both instruments and the additional regressors. Then construct the two-stage estimator using two different methods: i) use the ivreg command and, ii) estimate an OLS regression of log-wages where you replace ed76 by the predicted values from the reduced form equation. (Make sure that you use the same variables as additional regressors in all models.)

Compare the coefficient on education from i) and ii), are they the same? Why do the standard errors for education in i) and ii) differ? Which are the correct standard errors?

b) Now perform a test for overidentifying restrictions: regress the predicted residuals from the reduced form equation on all exogenous variables in the model (both instruments and the additional regressors). Then look at the n*R-squared from this regression. How can you interpret the overidentification test? What does it tell you about the validity of the instruments?

[Note: an alternative to perform a Sargan test for endogeneity, is available in STATA by typing overid after the running the IV regression. overid is a program which you]
have to install separately on your computer, type `search overid` and follow the instructions. This does not work on the Socrates server.]