Economics 475: Econometrics

Homework #4

This homework is Monday, February 13th.

Your Midterm Exam will occur on Wednesday, February 15th.

1. A large number of regressions investigating why some counties experience higher murder rates. These regressions typically estimate equations similar to:

(1) $M_i = \beta_0 + \beta_1 P_i + \beta_2 U_i + e_{1i}$

where M is the number of murders per 100,000 residents, P is the number of policemen per 100,000 residents, U is the unemployment rate, i indexes counties, and e_{1i} is mean zero, variance σ_1^2.

a. What signs do you expect β_1 and β_2 to take?

b. Many have argued that crime is not an exogenous variable. Indeed, one might think of murders being determined simultaneously with police presence. Consider the simultaneous system of equations:

(2) $M_i = \beta_0 + \beta_1 P_i + \beta_2 U_i + e_{1i}$

(3) $P_i = \alpha_0 + \alpha_1 M_i + \alpha_2 Inc_i + e_{2i}$

where Inc_i is the county's level of per capita income.

What are the reduced form equations for M and P?

c. If equations (2) and (3) describe the murder rate, what is the covariance between e_1 and P? What is the covariance between e_1 and U? Given these covariances, what will happen to an OLS estimate of (2)? Specifically, what will $\hat{\beta}_1$ and $\hat{\beta}_2$ be relative to their true values?

d. Are structural equations (1) and (2) over, exactly, or underidentified?

e. When I solve for the reduced form equations for M and P, I get:

(3) $M_i = \Pi_0 + \Pi_1 Inc_i + \Pi_2 U_i + w_i$

(4) $P_i = \Pi_3 + \Pi_4 Inc_i + \Pi_5 U_i + v_i$

where the Π's are functions of the α's and β's and the w's and v's are functions of the random error terms and the α's and β's. After using OLS to estimate equations (3) and (4), I find: $\hat{\Pi}_0 = .01$, $\hat{\Pi}_1 = -5$, $\hat{\Pi}_2 = 12$, $\hat{\Pi}_3 = 8$, $\hat{\Pi}_4 = 7$, $\hat{\Pi}_5 = 1$

What are your ILS estimates of β_0, β_1, β_2, α_0, α_1, α_2?

2. Perhaps the most frequently estimated regression is known as a Mincer Earnings Equation which expresses the natural log of wages as a function of individual observables including things like gender, age, experience and education. Economists have used the Mincer Earnings Equation to estimate the returns to education; that is the percent increase in wages given another year of education. However, this estimation is commonly criticized as having omitted variable bias; namely individuals going to school longer likely have characteristics that simultaneously make them better students and lead to higher pay. Thus, the coefficient on education is probably biased.

a. If one estimates the regression:

$$\ln(Wage_i) = \beta_0 + \beta_1 Educ_i + \varepsilon_i$$

but one omits variables such as ability and motivation, in what direction will OLS' estimate of β_1 be biased? What assumptions are you making in order to identify the direction of this bias?

b. Economists have long sought an instrumental variable that could be used to eliminate the bias from the regression in part a. What characteristics does such an instrument require? Some possible instruments suggested for this problem have been: 1) the number of siblings an individual has; 2) the distance from the nearest college an individual lives; 3) the education of an individual's parents. Comment on if these are appropriate or not.

c. One famous idea for an instrument was proposed by Joshua Angrist and Alan Krueger in a 1991 paper published by the Quarterly Journal of Economics. Before introducing this instrument, open the data set entitled "NEW7080.dta." This is the original data used by Angrist and Krueger and contains 247,199 observations of men born between 1920 and 1929 from the 1970 U.S. Census. Using this data estimate the equation:

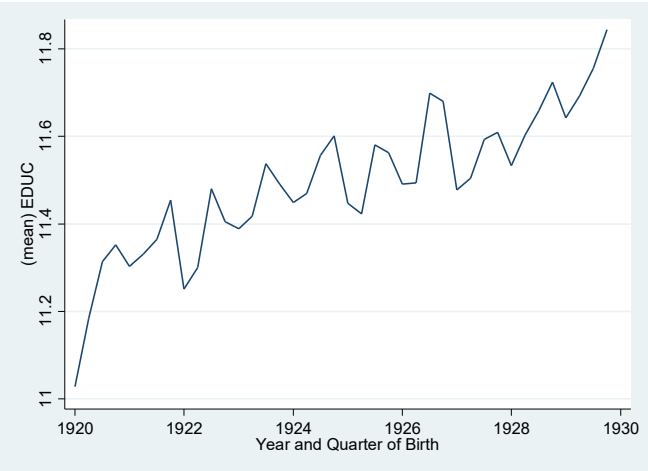
$$LWKLYWGE = \beta_0 + \beta_1 EDUC_i + \beta_2 BLACK_i + \beta_3 MARRIED_i + \beta_4 SMSA_i + \beta_5 NEWENG_i + \beta_6 MIDATL_i + \beta_7 ENOCENT_i + \beta_8 WNOCENT_i + \beta_9 SOATL_i + \beta_{10} ESOCENT_i + \beta_{11} WSOCENT_i + \beta_{12} MT_i + \beta_{13} YR20_i + \beta_{14} YR21_i + \beta_{15} YR22_i + \beta_{16} YR23_i + \beta_{17} YR24_i + \beta_{18} YR25_i + \beta_{19} YR26_i + \beta_{20} YR27_i + \beta_{21} YR28_i + \beta_{23} AGE_i + \beta_{24} AGEQSQ_i$$

In this case, the dependent variable is the natural log of weekly wages, EDUC is the years of education, BLACK and MARRIED are dummy variables, SMSA is a dummy variable indicating if an individual lives in a city, the next 8 variables are location dummy variables (e.g., NEWENG = new England); AGE and AGESQ are age and age squared, and the dummy variables starting with YR indicate the year the individual was born.

What is your estimate of β_1? How do you interpret this number?

d. Angrist and Krueger argue that the quarter-of-birth of an individual might be correlated with their education. Their argument has to do with the fact that individuals are required to attend school until the age of 16 (in many states). Someone born at the beginning of the year (quarter 1) will reach the age of 16 at an earlier point in their grade than someone born later in the year (say quarter 4). Thus, among two students dropping out of school at age 16, one will have more school than the other because they were born earlier in the year.

As evidence, they present this graph:



In this graph, the lowest points within a year are the first quarter of the year and the highest are the fourth. I made this graph using your data set and the following commands:
gen y = YOB + 0*QTR1 + .25*QTR2 + .5*QTR3 + .75*QTR4
collapse EDUC, by(y)
label variable y "Year and Quarter of Birth"
line EDUC y

Comment on the quarter of birth as an instrument.

e. From the graph in part d, it is clear that education is a function of the quarter of birth and the year of birth (there is more education for people born later in the decade). Angrist and Krueger propose as the instruments all possible dummy variables that represent year and quarter of birth (i.e., one dummy variable for 1920 quarter 1, another for 1920 quarter 2, etc.). Fortunately, these variables were included in your data set entitled QTR120, QTR121, QTR122, etc.

Using these instruments, estimate your first stage regression (don't forget the other exogenous variables from part c). What do you find? Evaluate if these are good instruments or not.

f. Estimate equation c using the instruments developed from the first stage in part e. What do you find? Do your results change relative to those found in part c?

3. Suppose you want to test whether girls who attend a girls' high school do better in math than girls who attend coed schools. You have a random sample of senior high school girls and measure the variable *score*, an outcome of a mathematics standardized test. Let *girlhs* be a dummy variable indicating whether a student attends a girls' high school. Consider the regression $Score_i = B_0 + B_1 Girlhs_i + \varepsilon_i$.

a. Suppose that parental support and motivation are unmeasured factors in ε. How does this fact impact estimates of B_1?

b. Consider the variable *Numgirl* where *Numgirl* is the number of girls' high schools within a 20 mile radius of the observation's home. Under what conditions could *Numgirl* be used as a valid IV for *Girlhs*.

4. Describe the data you will use in your final project. If possible, show me a regression from this data.