MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2023

Problem Set 6

Due: Friday, May 12 (in recitation)

1. The Canvas Module for Pset 6 contains a version of the Krueger (1999) data set with information on randomly assigned class size, student and teacher characteristics, and school identifiers.

   (a) Table VII in Krueger (1999) reports OLS and 2SLS estimates of class size effects. What are the instrument(s)? What's the motivation for 2SLS in this application?

   (b) Replicate the estimates in the first row of this table (for kindergarteners). Using the estimated first stage that goes with the 2SLS estimates in the replication to explain why 2SLS estimates are close to the corresponding OLS estimates.

2. The Catholic Church runs a large network of American parochial schools offering a low-cost private alternative to traditional public schools. Economists and educators have long debated the relative merits of Catholic schools and public schools. This debate motivates Evans and Schwab (1995) to estimate the effects of attending a Catholic high school on high school graduation and college attendance. Let Y_i be a dummy variable indicating college attendance and let CHS_i be a dummy variable indicating Catholic high school attendance. A linear model for the causal effect of Catholic school attendance on college-going is:

$$Y_i = \beta_0 + \beta_1 CHS_i + \beta_2 X_i + u_i, \tag{1}$$

   where the controls, X_i, include variables like gender, race, family income, and parental education.

   (a) Equations like (1), with a dummy dependent variable, are called *linear probability models*. Why does this name make sense?

   (b) Evans and Schwab (1995) reports 2SLS estimates of this model using a dummy for being Catholic as an instrument for CHS_i. Call the Catholic instrument Z_i and use this notation to describe the first and second stage equations that generate the 2SLS estimates reported in the last column of Table VI in Evans and Schwab (1995). Specifically, write out the two equations, using Greek for parameters, and briefly explain which variables appear in each one and the roles they play.

   (c) Under what assumptions does this 2SLS procedure capture the causal effects of Catholic high school attendance on college-going? Are these plausible?

3. This question explores the IV strategy used in Angrist & Krueger (1991) to estimate the economic returns to schooling.

   In most US states, children enter kindergarten in the calendar year in which they turn 5 years old: a child who is born in January 2015 and another who is born in December 2015 will both enter kindergarten in September 2020. States also enforce compulsory attendance laws. If the state dropout age is 16, for example, students can drop once they turn 16, whether or not they've finished the school year.

   (a) Consider young workers Sanjay and Emma, both living in North Carolina, where students enter kindergarten in the year they turn 5 and the dropout age is 16. Sanjay was born in January 2000, while Emma was born in December 2000.

      i. Which of these two children was older when they entered kindergarten, and by how much?

      ii. Suppose Sanjay and Emma both left school in the quarter of the year they turned 16 (the first quarter is Jan-March, the second is April-June, etc). Assuming no one repeats a grade, how many years of schooling will will each have completed upon leaving school (excluding their year in kindergarten)?

   (b) Download the AK91 data from the MM resources page under the section for Chapter 6. This file contains over 300,000 observations on men born 1930-39 in the 1980 Census. Variables include quarter-of-birth (QOB), year of birth (YOB), years of schooling, and log weekly wages.

i. Construct a single variable combing year and quarter of birth (e.g., men born in the first quarter of 1930 can be coded 30.00, while those born QOB II can be coded 30.25 etc.). Plot average schooling against this variable. How many points does your plot have? Highlight cohorts born in quarter 1 in one color and cohorts born in quarter 4 in another color (consider using Stata's `graph twoway` command to accomplish this).

ii. Make a similar plot for average weekly wages, again highlighting cohorts born in quarter 1 and quarter 4. In what sense do the average schooling and average wage plots move together?

4. Continue working with the Angrist & Krueger (1991) data.

   (a) Replicate MM Table 6.5, which shows OLS and IV estimates of the returns to schooling using alternative quarter-of-birth instruments, with and without year-of-birth controls. Report your replication results in a table format similar to Table 6.5.

   (b) Add to this a set of 2SLS estimates using 3 QOB plus 3 QOB * 9 YOB dummies as instruments.

      i. What controls are necessary here?

      ii. What do the extra instruments buy you?

      iii. How do the plots in Q3b relate to this 2SLS strategy?

5. Review the Ashenfelter-Rouse (1998) study summarized in LN12 and posted in Module E. As a reminder, this study uses data collected from identical twins to estimate the economic returns to schooling. The idea here is to compare education and income within pairs of twins, thereby controlling for their shared family background and similar (perhaps identical) genetic heritage.

   (a) Download data file ar98.dta from the Pset6 Canvas page. This file has data on wages and schooling for 340 twin pairs (680 total observations). Regress log wages (*lwage*) on years of schooling (*educ*), age (*age*), age-squared (*age2*), and dummies for female (*female*) and white (*white*). Interpret the estimated schooling and age coefficients.

   (b) Consider the regression model

$$\ln Y_{if} = \alpha' X_{if} + \beta S_{if} + \gamma A_f + \epsilon_{if},$$

   where f stands for family and subscript $i = 1, 2$ indexes twins in family f. The vector X_{if} includes the covariates from part (a), including a constant, while A_f is an *unobserved* ability variable assumed to be fixed within families.

      i. Explain why the regression of log wages on X_{if} and S_{if} without control for A_{if} is likely to generate a biased estimate of β.

      ii. Show (mathematically) that a regression of the within-family difference in log wages on the corresponding difference in schooling eliminates ability bias. What's the key assumption that makes this work?

   (c) Run the regression suggested by 4b(ii). Note that there are 340 unique twin differences, and that these are already computed for you (be sure to delete the redundant differences). Check your results by comparing them with those reported in the first two columns of MM Table 6.2. Do the first-differenced results align with your expectations about the direction of ability bias in the undifferenced model?

   (d) Using the difference in cross-sibling reports of educational attainment as an instrument for the difference in own reports of educational attainment, construct 2SLS estimates of the returns to schooling for both models, thereby completing your replication of MM Table 6.2. What do these results suggest about the relative importance of measurement error and ability bias in OLS estimates of the economic returns to schooling?

6. The MM website contains the youth mortality data used to make MM Table 5.2. Data sources and methods are given in the Empirical Notes section of the book.

(a) Use these data to replicate results in the table for estimated MLDA effects on all-cause and motor vehicle accident (MVA) mortality with and without state-specific trends and with and without weighting by state population (hint: replication code is archived on the book website). You should limit the sample to years before 1983 (included) throughout this problem. Report these in a table formatted like the original.

(b) The posted data include mortality for 15-17 year olds and 21-24 year olds. Estimate the effects of the fraction legal ($LEGAL1820_{st}$) as defined in MM on these death rates (this variable is the fraction legal among 18-20 year olds, so here you are asking how fraction legal drinking among 18-20 year olds affects mortality in other age groups). What do you expect here and how does this work out?

(c) As an alternative to the fraction legal ($LEGAL_{st}$) defined in the book, code a dummy variable indicating states and years allowing any under-21s to drink. Re-estimate the Table 5.2 models with this dummy variable replacing $LEGAL_{st}$ and report these new results as additional rows below your replication results. Compare results from the two specifications: are they consistent in magnitude and direction?

(d) (more challenging) Set the MLDA problem up as a staggered-adoption event-study design, with 3 leads and 10 lags around the adoption date. Here adoption is defined as allowing any under-21s to drink, as in part (c), and the 3rd lead captures treatment effects 3 or more years before adoption, while the 10th lag captures treatment effects 10 or more years after adoption. Focus on age 18-20 MVA mortality in 1970-1983, and drop drop Illinois (FIPS code 17) and Michigan (FIPS code 26), yielding a sample that has only adoptions.

i. Plot unweighted and population-weighted event-study estimates with confidence bands including a zero for the reference year. Do the event-study estimates show evidence of confounding trends? Does weighting matter?

ii. Comment on the dynamics seen in your estimated treatment effects. Do mortality declines induced by outlawing youth drinking appear to be lasting or transitory?

7. (extra credit) So far, we've discussed two scenarios where IV methods may be helpful: OVB problems and possible attenuation bias from measurement error. A third type of IV application uses instrumental variables methods to estimate supply and demand elasticities in simultaneous equations models (SEMs).

(a) Read Lecture Note 15 and the Angrist, Graddy, and Imbens (2000) paper on the Fulton Fish Market (posted in Module F on Canvas). Focus on Section 5 of the paper, which uses data on offshore weather conditions to construct 2SLS estimates of a linear equation describing the daily demand for fish. Using SEM theory, explain why weather instruments identify demand elasticities rather than supply elasticities in the market for fish.

(b) Use the fish.dta set (used in Pset 5) to replicate the demand elasticity estimates reported at the end of LN15.[1] Present your results with a brief explanation: what is the unit of observation? What are the endogenous variables, instrumental variables, and exogenous covariates? Label the first-stage, reduced form, and 2SLS estimates in your output. How are these different types of parameter estimates related? How and why do 2SLS and OLS estimates of demand elasticities differ?

(c) Use Stata's `reshape` command to change the data structure from time series of prices and quantities for Asians and whites to panel data that stacks the ethnic groups in a format where ethnicity is identified by a dummy variable rather than by distinct variables for Asians and Whites. Use this stacked data set to compute separate demand elasticities for Asians and Whites and test whether they differ. Discuss your results in light of the fact that Asian buyers appear to pay less than others for the same fish. (Stata hint: this requires a 2SLS specification that includes ethnicity interactions in both the first and second stages.)

[1] For this, it will help to know that stormy3=(speed3>18)*(wave3>4.5); mixed3=(1-stormy3)*(speed3>15)*(wave3>3); stormy2=(speed2>12)*(wave2>5.5); mixed2=(1-stormy2)*(speed2>10)*(wave2>3).