The University of Texas at Austin
ECO 348K (Advanced Econometrics)
Prof. Jason Abrevaya
Fall 2016

PROBLEM SET #3 (due Thursday, October 13th, 8am)

1. Refer to the last page of the "Instrumental variables, part II" lecture notes, which is Table 5 from the Angrist and Evans 1998 *American Economic Review* paper entitled "Children and their parents' labor supply: evidence from exogenous variation in family size." Explain what the -5.18 Wald estimate for the $Hours/week$ variable means, and exactly the variation that is being used to identify this causal effect. Given the reported standard error of 1.00, what can you say about the statistical significance of the estimated causal effect?

2. (Wooldridge Problem 15.2) Suppose that you wish to estimate the effect of class attendance on student performance. A basic model is

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u,$$

where the variables are defined as in Chapter 6 (see Example 6.3).

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

   (b) Assuming that *dist* and u are uncorrelated, what other assumption must *dist* satisfy to be a valid IV for *atndrte*?

   (c) Suppose we add the interaction term $priGPA \cdot atndrte$ to the model:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \cdot atndrte + u.$$

   If *atndrte* is correlated with u, then in general so is $priGPA \cdot atndrte$. What might be a good IV for $priGPA \cdot atndrte$? (Hint: If $E(u|priGPA, ACT, dist) = 0$, as happens when $priGPA$, ACT, and *dist* are all exogenous, then any function of $priGPA$ and *dist* is uncorrelated with u.)

3. (Wooldridge Problem 15.7) The following is a simple model to measure the effect of a school choice program on standardized test performance:

$$score = \beta_0 + \beta_1 choice + \beta_2 faminc + u_1,$$

where *score* is the score on a statewide test, *choice* is a binary variable indicating whether a student attended a "choice school" in the last year, and $faminc$ is family income. The IV for *choice* is *grant*, the dollar amount granted to students to use for tuition at choice schools. The grant amount differed by family income level, which is why we control for $faminc$ in the equation.

(a) Even with $faminc$ in the equation, why might *choice* be correlated with u_1?

(b) If within each income class, the grant amounts were assigned randomly, is *grant* uncorrelated with u_1?

(c) Write the reduced form equation for *choice*. What is needed for *grant* to be partially correlated with *choice*? (Note: The term "reduced-form equation" refers to an equation in which a variable is related to a set of exogenous variables. Read the Wooldridge chapter to fully understand how he uses this terminology.)

(d) Write the reduced form equation for *score*. Explain why this is useful. (Hint: How do you interpret the coefficient on *grant*?)

4. (Wooldridge Problm 15.C8) Use the dataset **401ksubs.dta** for this question. The equation of interest is a linear probability model:

$$pira = \beta_0 + \beta_1 p401k + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u.$$

The goal is to test whether there is a tradeoff between participating in a 401(k) plan and having an individual retirement account (IRA). *pira* is an indicator for having an IRA, and $p401k$ is an indicator for participation in a 401(k) plan. The (causal) parameter of interest is β_1.

(a) Estimate the equation by OLS and discuss the estimated slope on $p401k$.

(b) For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be the problem with OLS?

(c) The variable $e401k$ is a binary variable equal to one if a worker is *eligible* to participate in a 401(k) plan. Explain what is required for $e401k$ to be a valid IV for $p401k$. Do these assumptions seem reasonable?

(d) Estimate the reduced form for $p401k$, and verify that $e401k$ has a positive partial correlation with $p401k$. What is the p-value associated with testing for a significant partial correlation? (Since the reduced form is also a linear probability model, remember to use heteroskedasticity-robust standard errors.)

(e) Now, estimate the structural equation by IV and compare the estimate of β_1 with the OLS estimate. (Again, use heteroskedasticity-robust standard errors.) How do the standard errors for the OLS and IV estimators of β_1 compare to each other?

5. (Wooldridge Problem 15.C11, adapted from the data from the paper Rouse (1998, Quarterly Journal of Economics, "Private school vouchers and student achievement: an evaluation of the Milwaukee parental choice program")) Use the dataset **voucher.dta** for this question. Attendance at a choice school was paid for by a voucher, which was determined by a lottery among those who applied. This question asks you to do a cross-sectional analysis where winning the lottery for a voucher acts as an instrumental variable for attending a choice school. Actually, because we have multiple years of data on each student, we construct two variables. The first, *choiceyrs*, is the number of

years from 1991 to 1994 that a student attended a choice school; this variable ranges from zero to four. The variable *selectyrs* indicates the number of years a student was selected for a voucher. If the student applied for the program in 1990 and received a voucher, then $selectyrs = 4$; if he or she applied in 1991 and received a voucher, then $selectyrs = 3$; and so on. The outcome of interest is *mnce*, the student's percentile score on a math test in 1994.

(a) Of the 990 students in the sample, how many were never awarded a voucher? How many had a voucher available for four years? How many students actually attended a choice school for four years?

(b) Run a simple regression of *choiceyrs* on *selectyrs*. Are these variables related in the direction you expected? How strong is the relationship? Is *selectyrs* a sensible IV candidate for *choiceyrs*?

(c) Run a simple regression of *mnce* on *choiceyrs*. What do you find? Is this what you expected? What happens if you add the variables *black*, *hispanic*, and *female*?

(d) Why might *choiceyrs* be endogenous in an equation such as

$$mnce = \beta_0 + \beta_1 choiceyrs + \beta_2 black + \beta_3 hispanic + \beta_4 female + u_1?$$

(e) Esimate the equation in part (d) by instrumental variables, using *selectyrs* as the IV for *choiceyrs*. Does using IV produce a positive effect of attending a choice school? What do you make of the coefficients on the other explanatory variables?

(f) To control for the possibility that prior achievement affects participating in the lottery (as well as predicting attrition), add *mnce*90—the math score in 1990—to the equation in part (d). Estimate the equation by OLS and IV, and compare the results for β_1. For the IV estimate, how much is each year in a choice school worth on the math percentile score? Is this a practically large effect?

(g) Why is the analysis from part (f) not entirely convincing? (Hint: Compared with part (e), what happens to the number of observations, and why?)

(h) The endogenous variables *choiceyrs*1, *choiceyrs*2, and so on are dummy variables indicating the different number of years a student could have been in a choice school (from 1991 to 1994). The dummy variables *selectyr*1, *selectyrs*2, and so on have a similar definition, but for being selected from the lottery. Estimate

$$mnce = \beta_0 + \beta_1 choiceyrs1 + \cdots + \beta_4 choiceyrs4 + \beta_5 black + \beta_6 hispanic + \beta_7 female + u_1$$

by IV, using as instruments the four *selectyrs* dummy variables.

(i) (Extra—not in Wooldridge) Use the four *selectyrs* dummy variables as IV's for the single endogenous variable *choiceyrs* in the model of part (d). How does your 2SLS estimate of β_1 compare to that found in part (d)? Now do the GMM (instrumental variables) estimator, and report the p-value for the overidentification test.

(j) (Extra—not in Wooldridge) If you wanted to predict *mnce* based upon the explanatory variables in part (d) and also the *selectyrs* variable, but you were <u>not</u> interested in the causal effect of *choiceyrs*, what regression would you run?