Regression II
Vanderbilt University
Prof. Sean P. Corcoran

November 4, 2021

Lecture 9 In-Class Exercise

Exercise 1. This exercise will use instrumental variables to estimate the earnings returns to years of education. The data are adapted from David Card (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling," in *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp.* Ed. L.N. Christophides, E.K. Grant, and R. Swidinsky, 201-222. Toronto: University of Toronto Press.

1. Use the *card.dta* dataset from the Wooldridge textbook website. (This file is also saved on the class Github):

   ```
   use "http://fmwww.bc.edu/ec-p/data/wooldridge/card",clear
   ```

2. Estimate a simple OLS regression of the log wage *lwage* on years of education *educ*. Interpret the slope coefficient on *educ*.

3. Calculate the IV estimator of the slope on years of education, using proximity to a 4-year college as an instrument (*nearc4*). Do this in four ways:

   (a) Divide $\hat{\sigma}_{ZY}$ by $\hat{\sigma}_{ZX}$ (the method of moments estimator) and interpret the result. Hint: you can get a sample covariance matrix using `corr` in Stata.

   (b) Calculate the Wald estimate by dividing the reduced form coefficient ($\hat{\rho}$) by the first stage ($\hat{\phi}$). Interpret $\hat{\rho}$ and $\hat{\phi}$ individually, as well as their ratio. Compare your result to 3(a).

   (c) Calculate the 2SLS estimate manually by estimating the first stage, obtaining predicted values, and the using the predicted values in the second stage. Interpret your results and compare to parts 3(a-b). Compare the standard error of your slope to 3(a).

   (d) Use `ivregress` to obtain the 2SLS estimate. Get the first stage F-statistic using `estat firststage`. Interpret your results and compare to parts 3(a-c). Compare the standard error of your slope coefficient to 3(c)

4. Repeat parts 2 (OLS) and 3(c-d), but include covariates in the model: years of work experience, Black, the SMSA variables, South, married, and region dummies (*reg661-reg669*, excluding *reg661*).

5. Repeat part 3(d) with covariates, and follow `ivregress` with the Durbin-Wu-Hausman test for endogeneity: `estat endog`

6. Repeat part 3(d) with covariates, but include a second instrument: proximity to a 2-year college (*nearc2*). Get the first stage F-statistic using `estat firststage`. Interpret your results and compare to part 3(d).

7. Following part 6, conduct the test for endogeneity again, as well as the overid test `estat overid` (since the model is now over-identified). Interpret.

8. Use `ivregress` again as in part 6, but use the LIML estimation method.

Exercise 2. You have an instrument Z that you believe satisfies the exclusion restriction. That is, Z affects an outcome Y, but only through X. You decide to test this proposition by putting Z directly into your main estimating equation. You suspect that—if the exclusion restriction holds—then the estimated coefficient on Z should be zero:

$$Y = \beta_0 + \beta_1 X + \gamma Z + u$$

Let's test this idea with simulated data. In the code below, y depends on x and an unobserved variable w. (We are creating it but pretending it is unobserved). x depends on z and w, so x is correlated with the unobserved variable. z is a valid instrument because it affects x but does not affect y other than through x.

```
clear
set obs 1000
gen z = rnormal()
gen w = rnormal()
gen x = -2*z + 2*w + rnormal()
gen y = 5*x + 10*w + rnormal()
corr y x z w
```

1. Estimate an OLS regression of y on x and note the OVB.

2. Estimate a 2SLS regression using `ivregress`. Compare your slope coefficient to your OLS estimate and to the true (known) slope.

3. Now try the "diagnostic" regression at top (y regressed on both x and z). What do you find?

What's going on here? By the Frisch-Waugh-Lovell theorem, the coefficient on z in this regression is the same as the coefficient on $\tilde{z}$, the residual after regressing z on x:

```
reg y x z

// FWL
reg z x
predict ztilde, r
reg y ztilde
```

You might think that $\tilde{z}$—the residual variation in z—should be unrelated to y and thus will have a zero coefficient in the "diagnostic" regression. But we know x is related to both z *and* w (by construction). Rearranging the definition of x in this example: $z = -0.5x + w + 0.5e$. If you regress z on x alone, w is in the residual. Someone with a high w will have a higher z than that predicted by x (a larger positive residual). If you then regress y on these residuals, you'll see a significant association because y is related to the omitted variable w. $\tilde{z}$ is endogenous! You don't want to run a regression like this.

2SLS works because it uses predicted values from a regression of x on z. By construction, these predicted values are *uncorrelated* with w.

Let's place a real research question on this example, borrowing from Exercise 1. Let y = earnings, x = education, and w = unobserved ability. Let the instrument z = distance to the nearest 4-year college. Suppose you regress z on x and get the residual. If the instrument works in the direction we think it does, the coefficient on x will be negative. (Persons with more education tend to live closer to a college). Now think about a positive residual: this is someone who—for a given x—lives further from colleges than predicted. Who might this be? Someone with high unobserved ability w. A negative residual would be someone who—for a given x—lives closer to colleges than predicted. The residual reflects the effect of w.

Your (bad) diagnostic would regress earnings on education (x) and the distance to colleges (z). *Holding x constant,* someone who lives further from colleges (a higher z) is likely to be someone with a higher w. (They got their education despite being someone further from college). z is endogenous once you've conditioned on x.

It's OK to regress y on z alone—the reduced form—since z is exogenous. It's only when you try to include both x and z together that it becomes a problem.

Exercise 3. This exercise is taken from Murnane & Willett chapter 10. It considers the relationship between educational attainment and civic engagement, as measured using voter registration. The data are from High School & Beyond and adapted from Dee (2004), "Are There Civic Returns to Education?" *Journal of Public Economics*, 88(9-10), 1697–1720.

1. Use the *ch10_dee.dta* dataset from Github:

   ```
   use https://github.com/spcorcor18/LPO-8852/raw/main/data/ch10_dee.dta
   ```

2. Estimate a simple OLS regression of voter registration (*register*) on *college*, the indicator of attendance at a 2- or 4-year college by 1984, when the study sample was around age 28. Interpret the slope coefficient on *college*.

3. Calculate the IV estimator of the slope on *college*, using *distance*, the number of miles between the respondent's high school and the nearest 2-year college. Do this in four different ways: method of moments, dividing the reduced form coefficient by the first stage, manual 2SLS, and `ivregress` (see Exercise 1 part 3). Following *ivregress*, get the first-stage F-statistic.

4. Try the "bad diagnostic" explored in Exercise 2. That is, regress *register* on both *college* and the instrument. What do you find?

5. Repeat parts 2 (OLS) and 3 (IV), but include the race/ethnicity covariates: Black, Hispanic, other race.

6. If the instrument *distance* is exogenous, then interactions between the instrument and other exogenous covariates must be too. Add to your list of instruments interactions between *distance* and the three race/ethnicity covariates. Interpret your first stage coefficients and obtain the first stage F statistic. Note you now have an over-identified model since the number of instruments exceeds the number of endogenous regressors.

7. The interactions in part (6) allow you to also include interactions in your second stage. If *college* is endogenous in your main model, then interactions between *college* and the race/ethnicity variables are endogenous too. You need at least as many instruments in an IV model as you have endogenous variables. Fortunately, your interactions in part (6) make this possible. Using `ivregress 2sls`, calculate the IV estimator of the slope on *college* and the slopes on interactions between *college* and Black, Hispanic, and other race/ethnicity. Use as instruments *distance* and its interactions with these three covariates. Note with four instruments, there will be four first stage regressions.