Problem Set 2

Q1: Project STAR

Project STAR was a famous experiment designed to provide evidence on the effects of smaller classes. The experiment was conducted in 79 Tennessee elementary schools in the mid-1980s. Within each of these schools, students were randomly assigned to one of three class types: small classes (13–17 students), regular classes (22–25 students), or regular classes with a teacher's aide. Students entered the experiment in kindergarten or, in some cases, in later grades. Random assignment was carried out separately within each school and entry grade. For this question, we will focus on students who entered the experiment in kindergarten and compare small and regular classes.

Some parts of this question require you to work with a subset of the Project STAR data. To access this dataset in Stata enter the following command on one line:

```
use tscorek stark black white boy freelunk schidkn sck if stark==1 using

"http://fmwww.bc.edu/ec-p/data/stockwatson/star_sw", clear
```

The dataset contains 6,325 observations on kindergarten students in small or regular classes. The variables are:

- Kindergarten test score (`tscorek`)
- Small-class indicator (`sck`)
- School (`schidkn`)
- Various student characteristics (`black`, `boy`, `freelunk`).

Note that `sck` is the class type students initially attended, rather than their randomized assignment. However, Krueger (1999) obtained original assignment records for 1,581 kindergarten students in 18 participating schools and found that 99.7% attended their assigned type. We will therefore treat initial class type attended as equivalent to initial assignment.

(a) Using the Project STAR data, estimate the causal effect of assignment to a small class on kindergarten test scores. Report and interpret your estimate. Explain your choice of regression specification. For now, use robust standard errors.

(b) Explain why it would make sense to cluster standard errors at the classroom level. (Note that classroom indicators are not included in this data extract and so you cannot do this).

(c) When estimating a causal treatment effect using regression with covariates, two key assumptions are conditional independence of treatment and the causal error given the covariates, and correct specification of the conditional mean of the causal error as a function of the covariates. Explain why both assumptions are plausible in this setting.

(d) To assess the plausibility of the conditional independence assumption, conduct "balance checks" of whether small-class assignment is associated with predetermined student characteristics conditional on school. Report results for at least two characteristics. What do you think the evidence establishes about the plausibility of the conditional independence assumption?

(e) Suppose it turned out that more-effective teachers were assigned to smaller classes. In that case, being assigned to a small class could benefit students for one of two reasons. First, for a given quality of teacher, students perform better in smaller classes. Second, students in smaller classes get better teachers. Suppose you had good data on teacher quality, how would you go about trying to decompose the overall effect into the effects attributable to these two channels?

(f) Krueger and Whitmore (2001) report that 43.7% of students initially assigned to a small class took the ACT or SAT, compared with 40.0% of students assigned to a regular class (Figure 4 - these numbers refer to the on-grade-level sample). Why would it be potentially misleading to compare average ACT/SAT scores among test-takers, even if initial assignment to small or regular class was randomized?

(g) Suppose that assignment to a small class cannot cause any student who would otherwise take the ACT/SAT to *stop* taking it. Under this monotonic-selection assumption, calculate the fraction of observed small-class test-takers that would be trimmed to construct Lee (2009) bounds, using the rates in part (f). Which end of the score distribution would be trimmed for the upper and lower bounds? For this question, abstract from the restriction to on-grade-level students and assume that these participation rates came from an unconditionally randomized experiment.

(h) Krueger and Whitmore (2001) use a "linear truncation" procedure to adjust for sample selection bias (Table 8). Explain this procedure and the assumptions under which they interpret it as a point estimate. How does this procedure relate to Lee (2009) bounds?

Q2: Returns to Education

We use the Card (1993) data, drawn from the National Longitudinal Survey of Young Men, to estimate the causal effect of an additional year of schooling on earnings. The extract contains 3,010 men, initially surveyed in 1966, with schooling and wages measured in 1976. College proximity provides a possible source of variation in schooling costs. Load the data in Stata (enter the following on one line):

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

The variables needed are:

- `educ`: years of schooling in 1976; `lwage`: log hourly wage (wage measured in cents) in 1976.
- `nearc4`, `nearc2`: indicators for living near a four-year or two-year college, respectively, in 1966. The numbers refer to college type, not distance in kilometers.
- `black`: indicator for being Black; `smsa66`: indicator for metropolitan residence in 1966; `reg661--reg669`: nine region indicators for 1966.
- `motheduc`, `fatheduc`: mother's and father's years of schooling;

Use heteroskedasticity-robust standard errors throughout. Always report sample sizes as part of your regression output.

(a) We begin by estimating the OLS relationship between years of schooling and log wages.

(i) Regress log wages (`lwage`) on years of schooling (`educ`). Report the coefficient and standard error on `educ`, and interpret the coefficient.
(ii) Explain why the OLS coefficient might not identify the causal effect of schooling on earnings. In particular, describe a potential omitted-variable-bias problem involving mother's education (`motheduc`). Use the data to assess whether mother's education is associated with an individual's schooling.
(iii) Reestimate the OLS relationship from (i), first without and then with mother's education as an additional control. Use the same observations in both regressions. Report the coefficient and standard error on schooling in each specification. What do your findings imply about the omitted-variable-bias argument in (ii)?

(b) Consider using `nearc4` as an instrument for education, without controlling for any covariates.

(i) Explain why living near a four-year college might affect educational attainment.

Estimate the first-stage regression and assess whether the first-stage relevance condition (IV-FS in the lecture notes) is satisfied.

(ii) Explain why the independence assumption (IV-I in the lecture notes) might fail in this setting. In particular, consider why individuals living near four-year colleges might differ from those living farther away in ways that affect their earnings independently of their educational attainment.

(iii) Use the data to investigate whether `nearc4` is associated with mother's and father's education. Explain what your findings imply about the plausibility of the independence assumption.

(c) We now consider using `nearc4` as an instrument conditional on the following covariates:

$$X = \{\text{black, smsa66, reg661–reg668}\}.$$

(i) Explain why controlling for these variables might make the conditional independence assumption (IVX-CIA in the lecture notes) more plausible than the unconditional independence assumption considered in (b). Give an example of a reason why the conditional independence assumption might still fail.

(ii) Estimate the first-stage regression conditional on X. Report the coefficient and standard error on `nearc4`. What do you conclude about the conditional first-stage relevance condition (IVX-FS in the lecture notes)?

(iii) Repeat the balance checks from (b), now controlling for X. Compare the results with those from (b). What do you conclude about the plausibility of the conditional independence assumption?

(iv) When identification using an instrument relies on conditional independence given X, a linear IV regression may also require a functional-form assumption. Explain what this assumption requires when X enters the regression additively. For the X in this example, how could you specify the regression to avoid imposing a functional-form restriction on $E[\epsilon \mid X]$?

(v) Estimate the reduced-form relationship between nearc4 and log wages, conditional on X. Report and interpret the coefficient on nearc4. Using the first-stage estimate from (ii), calculate the IV estimate as the ratio of the reduced-form and first-stage coefficients. Verify that you obtain the same estimate using 2SLS.

(vi) Compare your IV estimate with the OLS estimate using the same controls. Interpret both estimates. Under what assumptions can the IV estimate be given a causal interpretation? What do you conclude from the difference between the two estimates?

(d) Report the first-stage F-statistic from part (c)(ii). Explain whether weak instruments appear to be a concern and discuss the traditional $F > 10$ rule of thumb.

(e) Another potential instrument for education is `nearc2`, an indicator for whether an individual lived near a two-year college in 1966.

   (i) Using the same covariates as in (c), estimate the first stage and the IV coefficient using `nearc2` instead of `nearc4`. Report the coefficient and standard error from each regression, together with the first-stage F-statistic.

   (ii) Compare these results with those obtained using `nearc4`. What do you conclude about the usefulness of `nearc2` as an instrument?

Q3: Weak Instruments — Monte Carlo Exercise

In Q2, we found that proximity to a two-year college was a weak predictor of schooling and produced a highly imprecise IV estimate. We now use a Monte Carlo experiment to investigate how the sampling distribution of IV estimates and the reliability of conventional IV inference depend on instrument strength.

Consider the data-generating process

$$Y_i = \beta D_i + \epsilon_i, \qquad D_i = \gamma Z_i + \rho\epsilon_i + \sqrt{1-\rho^2-\gamma^2}\, u_i,$$

where Z_i, ϵ_i and u_i are mutually independent standard normal random variables. Here Y_i is the outcome, D_i is the treatment, Z_i is a candidate instrument, and ϵ_i is the causal outcome error. The parameters β, γ and ρ determine the causal effect, instrument strength and dependence between treatment and the outcome error. Set $\beta = 0$ and $\rho = 0.75$ throughout.

(a) Establish the following properties of the DGP.

   (i) What is the true causal effect of D on Y?

   (ii) Explain why OLS does not identify this causal effect. Derive the population regression coefficient of Y on D. Recall this is $\frac{Cov(Y,D)}{VarD}$.

   (iii) Explain why Z satisfies the IV independence assumption. Under what condition does it satisfy the first-stage relevance condition?

(b) For each $\gamma \in \{0.05, 0.10, 0.20, 0.40\}$, simulate 2,500 independent samples of size $n = 1000$. Set the random seed to 123456 before running the Monte Carlo replications for each value of γ. This ensures that your results are reproducible. In each replication:

   (i) Generate the data from the DGP above.

   (ii) Estimate OLS of Y on D, including a constant.

   (iii) Estimate the first stage of D on Z, including a constant, and record its usual F-statistic.

   (iv) Estimate IV using Z as the instrument for D, including a constant and using heteroskedasticity-robust standard errors.

   (v) Test the true null $H_0 : \beta = 0$ using the conventional two-sided IV Wald test, rejecting when $|\widehat{\beta}_{IV}/\widehat{SE}_{IV}| > 1.96$ (use the standard normal 97.5th percentile).

(vi) Record the estimates, IV standard error, first-stage F-statistic and rejection indicator needed below.

Draw histograms showing the sampling distribution of the IV estimator for each value of γ. Use the same horizontal range, [-2,2], and the same bin width (0.1) for all four histograms. Indicate the true effect 0 and the population OLS coefficient 0.75. The following stata code will generate a histogram that meets those requirements (this assumes you have saved your simulation results in a dataset containing `b_iv`, the IV estimate, and gamma, the instrument-strength parameter):

```
histogram b_iv if gamma==0.05 & inrange(b_iv,-2,2), ///
width(0.10) start(-2) xline(0 0.75)
```

Complete the following table using the results of your Monte Carlo experiment.

γ	Mean OLS	Median IV	Mean first-stage F	IV rejection rate (%)
0.05				
0.10				
0.20				
0.40				

Use all 2,500 replications to calculate each statistic. Report median IV because the sampling distribution of the just-identified IV estimator need not have a finite mean. The rejection rate refers to the true null at the nominal 5% level. The rejection rate is the percentage of the 2,500 replications in which you reject $H_0 : \beta = 0$ using the 5% significance test in part (v).

(c) Interpret the results.

(i) Do the Monte Carlo results for OLS agree with the theoretical results in (a)? Explain.

(ii) Describe how the sampling distribution of IV changes as the instrument becomes weaker. Discuss its location, dispersion and shape relative to the true causal effect and the OLS benchmark.

(iii) How do the empirical rejection rates vary with the first-stage F-statistic? What do the results imply about the reliability of conventional IV inference with weak instruments?

References

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics*, 114(2):497–532.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from project star. *The Economic Journal*, 111(468):1–28.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102.