Problem Set 2: Answer Key (answers in **bold)**

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.

Include school fixed effects because random assignment was within school. The estimated effect of small-class assignment is 15.157 points (robust SE 1.972; $N = 5{,}786$): students assigned to small classes scored about 15 points higher than those assigned to regular classes.

(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).

Robust standard errors do not allow for within-classroom error correlation. We expect within-classroom error correlation because of classroom-level shocks (e.g., teacher effects). Since the treatment is assigned at the level of the classroom, there is a strong case for clustering at the classroom level.

(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.

Conditional independence: Assignment was random within schools, so assignment should be independent of the outcome error conditional on school. Functional form: Because school is a discrete variable, including a complete set of school fixed effects allows $E[\epsilon_i \mid X_i]$ to take a different value in every school. The conditional expectation is therefore saturated, so there is no substantive functional-form restriction.

(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?

The relevant comparison is within schools. The estimated small-class coefficients in school-adjusted balance regressions are:

Characteristic	Coefficient	Robust SE	N
Black	**-0.0057**	**0.0065**	**6,322**
Boy	**0.00002**	**0.0139**	**6,325**
Free lunch	**-0.0085**	**0.0118**	**6,301**

These small differences are consistent with conditional randomization, but cannot prove independence from unobserved student characteristics.

(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?

This decomposition would be an example of mediation analysis. Random assignment alone does not identify this decomposition. One approach is to estimate the effect of small-class assignment on teacher quality and combine that with an externally obtained estimate of the causal effect of teacher quality on test scores. This requires additional assumptions. You should not attempt to identify the "direct effect" of class size by adding measures of teacher quality to a regression of test scores on class size and school fixed effects (this would be an example of "bad control"). Note that in practice, teachers were supposed to be randomized to classrooms under STAR, so teacher quality should not be a mediator.

(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?

The observed test-taking rate is 3.7 percentage points higher among students assigned to small classes. The test-taking indicator is itself an outcome, but the ACT/SAT *score* is observed only for those who take a test. This is differential attrition out of the sample (differential selection into the sample with observed scores). Conditioning on taking the test can undo the comparability created by randomization. For example, if small classes induce relatively low-scoring students to take a test, the mean score among

small-class test-takers can be pulled down, even if small classes raise every student's score. The sign of the bias is not established by the participation rates alone.

(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.

The fraction is $(0.437 - 0.400)/0.437 \simeq 0.085$**, or about 8.5% of the observed small-class test-takers. Trimming the *bottom* 8.5% of small-class scores yields the upper bound on the treatment effect; trimming the *top* 8.5% yields the lower bound.**

(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?

Linear truncation removes the lowest-scoring marginal fraction of small-class test-takers. The authors motivate it as a point estimate under strong assumptions that the additional test-takers are the lowest-scoring students and that class-size treatment does not change the relevant score ranking. Trimming the lowest-scoring treated test-takers is also the procedure used to construct the upper Lee bound under monotonic selection. However, the regression-adjusted estimate in Table 8 need not coincide exactly with a Lee bound.

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)?

(i) The OLS coefficient (robust SE) is 0.0521 (0.0029), using 3,010 observations. This implies that an additional year of schooling is associated with approximately 5.2% higher hourly wages. This is an association, not necessarily a causal effect.

(ii) More educated mothers may provide greater educational support or resources, leading their children to obtain more schooling. These family-background advantages may also increase earnings independently of the individual's schooling. This could generate upward omitted-variable bias in the OLS estimate. Regressing `educ` on `motheduc` gives a coefficient (robust SE) of 0.3734 (0.0160), consistent with this argument: individuals with more educated mothers obtain significantly more schooling.

(iii) Restricting both regressions to the 2,657 observations with nonmissing mother's education gives:

Specification ($N = 2{,}657$)	Schooling coefficient	Robust SE
Without mother's education	**0.0510**	**0.0031**
With mother's education	**0.0435**	**0.0035**

The schooling coefficient falls when mother's education is included. This is consistent with the omitted-variable-bias argument in (ii): some of the positive OLS association between schooling and wages may reflect family-background advantages.

(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.

(i) Living near a four-year college may reduce travel and living costs and increase schooling. The unconditional first-stage coefficient is 0.8290 (robust SE 0.1067; $N = 3{,}010$): men living near a college completed about 0.83 more years of schooling. This provides strong evidence of first-stage relevance.

(ii) The independence assumption requires $Z_i \perp \epsilon_i$, where ϵ_i contains earnings determinants other than schooling. Families may sort into college areas according to resources or educational preferences, while local schools and labor-market opportunities may differ. Proximity could therefore be associated with earnings through channels other than the individual's schooling.

(iii) The unconditional differences in parental education by `nearc4` are:

Characteristic	N	Coefficient	Robust SE
Mother's education	**2,657**	**0.594**	**0.136**
Father's education	**2,320**	**1.135**	**0.170**

Both differences are statistically significant. They raise concerns about the independence assumption because parental resources and networks could affect earnings outside the schooling channel. These associations do not by themselves establish a violation: that conclusion also requires a link between parental education and the unobserved causal error.

(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?

(i) Controlling for predetermined region, metropolitan status and race compares men with more similar educational access and earnings opportunities, making conditional independence, $Z_i \perp \epsilon_i \mid X_i$, more plausible than unconditional independence. It may still fail if families living near colleges have more highly educated parents even within these geographic and demographic groups, and parental background affects earnings outside schooling.

(ii) The conditional first-stage coefficient is 0.4621 (robust SE 0.1164; $N =$ 3,010). Proximity remains predictive of schooling after conditioning on X, supporting conditional first-stage relevance, although its estimated association is smaller than without controls.

(iii) The balance comparisons are shown below, with robust standard errors in parentheses. Each row uses the same observations before and after adjustment; missing parental education produces different sample sizes across rows.

Characteristic	N	Unconditional	Conditional on X
Mother's education	2,657	0.594 (0.136)	0.016 (0.142)
Father's education	2,320	1.135 (0.170)	0.333 (0.179)

Conditioning substantially reduces both observed imbalances, supporting the plausibility of conditional independence relative to unconditional independence. Father's education remains somewhat imbalanced. Balance on observed characteristics cannot establish conditional independence from the unobserved causal error, and non-rejection can reflect limited power.

(iv) The functional-form assumption requires the conditional mean of the causal error to be correctly represented by the included functions of X: $E[\epsilon_i \mid X_i] = \alpha_0 + \beta_2 X_i$ for the additive specification. Separate region indicators and race and metropolitan indicators impose additivity; they do not saturate the

model over all combinations. Including indicators for every observed region-by-race-by-metropolitan cell allows a separate conditional error mean in each cell, eliminating the functional-form restriction with respect to these discrete variables. This does not establish conditional independence or remove within-cell confounding.

(v) The conditional reduced-form coefficient is 0.05429 (robust SE 0.01795; $N = 3{,}010$**). Men living near a four-year college have about 5.4% higher wages conditional on** X**; this association is not automatically causal. The conditional IV estimate is**

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{IV}\,\boldsymbol{X}} = \frac{\widehat{\boldsymbol{r}}_{\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X}}}{\widehat{\boldsymbol{r}}_{\boldsymbol{D},\boldsymbol{Z}|\boldsymbol{X}}} = \frac{\mathbf{0.0542865}}{\mathbf{0.4620650}} \simeq \mathbf{0.1175.}$$

With the same controls and sample, this ratio agrees with 2SLS in this just-identified model. The IV standard error must come from the 2SLS procedure, which accounts for estimation of the first stage, rather than an ordinary regression treating fitted schooling as fixed.

(vi) Using the same X **and all 3,010 observations:**

Estimator	Schooling coefficient	Robust SE
OLS	**0.0367**	**0.0029**
IV	**0.1175**	**0.0428**

The estimates imply approximately 3.7% and 11.8% higher wages per additional year of schooling, respectively, using the log-point approximation. Under the maintained common-effect model, IV identifies the causal return if conditional first-stage relevance, conditional independence and the functional-form assumption hold. IV exceeding OLS does not establish that OLS is biased downward or that these IV assumptions hold; sampling uncertainty, measurement error or failures of identification could contribute to the difference.

(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.

The robust first-stage statistic is $F(1, 2998) = 15.77$**, equal to the squared robust first-stage** t**-statistic. It exceeds the traditional** $F > 10$ **rule of thumb, but this threshold does not guarantee reliable conventional IV inference, especially under heteroskedasticity.**

(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?

(i) With the same additive controls X and 3,010 observations as in (c), the estimates are:

Statistic	`nearc4`	`nearc2`
First-stage coefficient	**0.4621**	**0.0900**
First-stage robust SE	**0.1164**	**0.1035**
First-stage F	**15.77**	**0.76**
IV coefficient	**0.1175**	**0.4542**
IV robust SE	**0.0428**	**0.5135**

(ii) Conditional on X, `nearc2` has an extremely weak estimated first stage. Its IV estimate is much larger than the `nearc4` estimate but extremely imprecise, illustrating how a nearly zero estimated first stage can produce an unstable IV estimate. The results do not establish that the population first stage is exactly zero. Weak first-stage relevance and failure of the independence assumption are distinct issues: a weak first stage need not imply a failure of independence. Conventional IV confidence intervals and tests can be unreliable with weak instruments, even with robust standard errors.

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?

(i) The true causal effect is $\beta = 0$.

(ii) Independence and unit variances of the three shocks imply

$$\mathbf{Var}(\boldsymbol{D}) = \boldsymbol{\gamma^2 + \rho^2 + (1 - \rho^2 - \gamma^2) = 1}, \qquad \mathbf{Cov}(\boldsymbol{D, \epsilon}) = \boldsymbol{\rho = 0.75.}$$

$$\boldsymbol{\beta_{\mathrm{OLS}} = \beta + \frac{\mathrm{Cov}(D, \epsilon)}{\mathrm{Var}(D)} = 0.75.}$$

Treatment is correlated with the causal error, so OLS does not identify the true effect. Its population coefficient remains 0.75 as γ changes.

(iii) Z is independent of ϵ by construction. Also, $\mathrm{Cov}(Z, D) = \gamma$, so the first-stage relevance condition holds when $\gamma \neq 0$, including all four strengths below.

(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.

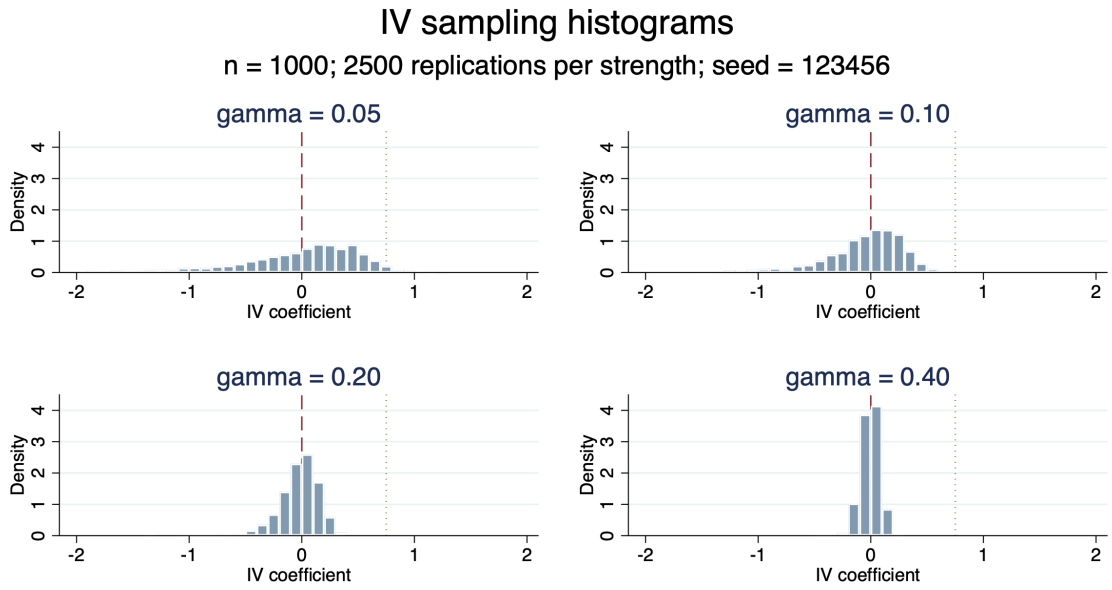
Draw histograms showing the sampling distribution of the IV estimator for each value of γ. Use the same horizontal range, [-2,2], and the same bin width (0.1) for all four histograms. Indicate the true effect 0 and the population OLS coefficient 0.75. The following stata code will generate a histogram 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).



Histograms restrict estimates to [-2,2], as in the question's command.
Bin width = 0.10. Lines: true effect = 0; population OLS = 0.75.
All 2,500 draws per strength enter the summary table.

The histograms follow the example command: bins have width 0.10 and start at -2, axes are common, and reference lines mark 0 and 0.75. Only estimates in $[-2, 2]$ enter each histogram, so bar heights are normalized over that restricted sample. The table below instead uses all 2,500 replications per strength.

All 2,500 replications per strength are retained, including extreme IV estimates. The completed table reports realized Monte Carlo results; the first-stage F-statistics use the homoskedastic first-stage convention above.

γ	Mean OLS	Median IV	Mean first-stage F	IV rejection rate (%)
0.05	0.7496	0.0830	3.50	8.48
0.10	0.7497	0.0005	11.09	6.76
0.20	0.7497	0.0001	42.69	4.72
0.40	0.7496	0.0001	191.79	5.00

(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?

(i) Yes. Mean OLS is about 0.7496–0.7497 for every strength, agreeing with the population coefficient of 0.75 rather than the causal effect of zero. Changing γ leaves the population OLS coefficient unchanged; small differences between simulated means reflect Monte Carlo variation.

(ii) With $\gamma = 0.40$, IV is concentrated near zero; its median is 0.0001. As the instrument weakens, the distribution becomes more dispersed and asymmetric, with much heavier tails. For $\gamma = 0.05$, the median is 0.0830, shifted toward the OLS benchmark of 0.75, and 11.88% of estimates fall outside $[-2, 2]$. The weakest-IV distribution is not simply centered at the OLS coefficient. Median bias is smaller at each stronger instrument in this run, but this is an observed simulation pattern rather than a general monotonicity result.

(iii) The empirical rejection rate is 8.48% when mean first-stage F is 3.50, falling to 6.76% when F is 11.09. For the stronger instruments, mean F-statistics are 42.69 and 191.79 and rejection rates are 4.72% and 5.00%, close to the nominal 5% level. Thus conventional IV inference overrejects for the weaker instruments in this experiment. Rejection rates are not monotone across all four strengths. The rate remains above 5% when mean F exceeds 10, illustrating that the traditional rule of thumb does not guarantee reliable inference.

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.