ECON 1123 Review Session

Please Pick up Exam Packets at the Front!

Slides at github.com/cjleggett/1123-section

Logistics

- 9am on Tuesday, May 9
  - Please come a few minutes early!
- Science Center D: Last Names A-Q
- Science Center A: Last Names R-Z
- Bring:
  - A pen (not a pencil)
  - A simple calculator
  - 2 double-sided sheets of notes

Format

1. Read through section description, glance at tables, ask questions.
2. For each problem:
    A. Read the problem out loud
    B. Talk through thought process
    C. Write expected answer
    D. Ask Questions
3. Repeat for next section

Takeaways

- Correct answers can be much shorter than the example ones
- Key question types: “Why might a simple OLS be biased?” -> OVB
- Cheat Sheet Ideas

Part A

03:00

A1

Question

- Interpret the coefficient on *After Medicaid expansion* in regression (1)

Thoughts

- Interpretation
- Causal or Not?
- X variable + units
- Y variable + units
- Holding Constant

Response

- **Expanding Medicare is associated, on average, with a decrease of 13.74 bankruptcies per 1000 adults.**
- Could also use difference:
  - **On average, counties in which medicare has been expanded have 13.74 fewer bankruptcies per 1000 adults than counties in which it has not been expanded.**

Questions

A2

Question

- Use the omitted variable bias formula to explain why the coefficient on *After Medicaid expansion* changes in regression (2) compared with regression (1).

Thoughts

- OVB Problem
- Given Omitted Variable: Unemployment Rate
- General form:
  - State omitted variable
  - Sign β_2
  - Sign γ_1
  - Multiply the two to sign the bias
  - Overestimate or Underestimate

1) Response

- The coefficient in regression (1) could be biased because it omits the variable of unemployment rate.
- β_2 is positive because when the unemployment rate is higher, more people will be forced to declare bankruptcy
- γ_1 could be negative, because the country was still recovering from 2008 at that point so the unemployment rate was going down in the years after some states expanded Medicare
- This means the sign of the bias is negative.
- Therefore, since $\alpha_1 < 0$ we have overstated the causal effect of expanded Medicare on bankruptcies.
- This matches what we see in regression (2) as the value is less negative.

1) Questions

A3

Question

- All the regressions in Table A2 use clustered standard errors, where the clustering is at the state level. What is the advantage of using standard errors clustered by state instead of each of the following:
    - a) Heteroskedasticity-robust standard errors
    - b) Standard errors clustered at the county level

Thoughts

- Why do we use clustered standard errors?
- Definitions

Response

- **a)** We use clustered standard errors to account for serial correlation in the dataset. This is necessary because it is unreasonable to think that two data points taken from the same county one year apart are independent.

- **b)** We cluster at the state level rather than the county level because we always cluster at the level of policy change, and the expansion of medicare occurs at the state level.

Questions

A4

Question

- State bankruptcy laws, such as asset exemptions, differ significantly across states, but have not changed since the early 1900s. For example, Texas has the most debtor-friendly asset exemption laws in the U.S., allowing households to protect the full value of their homes from creditors in bankruptcy proceedings. Delaware has the least debtor-friendly laws, with no homestead exemption. Discuss if state bankruptcy laws are arguably a source of omitted variable bias (i) in regression (2)? (ii) in regression (4)?

Thoughts

- What’s different between regressions (1) and (4)?
- What’s being controlled for?

Response

- **i)** State bankruptcy laws are feasibly a source of OVB in regression (1) because we do not control for state, and bankruptcy laws could clearly have an effect on how many people file for bankruptcy, and it's reasonable to think a state's bankruptcy laws could be correlated with its Medicare expansion.

- **ii)** They are not a source of OVB in regression (2) because this regression includes county-fixed effects, and these laws are constant within each county.

Questions

A5

Question

- Column 6 presents estimates of the dynamic causal effect of the Medicaid expansions, by replacing *After Medicaid expansion* with three separate variables: $\Delta Expanded_t$, $\Delta Expanded_{t-1}$, and $Expanded_{t-2}$. Note that $Expanded_t$ is an indicator that equals 1 in the first year that a state expanded Medicaid (which is 2014 in our sample) and 0 otherwise, and that $\Delta Expanded_t = Expanded_t - Expanded_{t-1}$

  A. What is the long run cumulative effect of Medicaid expansion?

  B. Graph the cumulative impulse response function (i.e. cumulative dynamic multipliers) in response to expanding Medicaid, by year, from year t through t + 2.

  C. Do you have enough information in the table to report the 95% confidence intervals for the cumulative impulse response function? Explain why or why not.

Thoughts

- What is long-run cumulative effect?
- What is Cumulative Impulse Response Function?
- What do we need to form a 95% CI?

Response

A. -15.74 bankruptcies per 1000 adults

B. (0, -4.796), (1, -10.089), (2, -15.74) (Draw on board) (Label Axes!)

C. Yes, we can calculate the 95% CI because we have an estimate of the coefficient and its standard error in regression (6)

Questions

A6

Question

- Regression (5) adds county specific linear time trends, as suggested in Chapter 5 of Angrist and Pischke (2015). Explain the motivation for adding these terms to the regression. What do you conclude from regression (5) about the internal validity of regression (4)?

Thoughts

- Why would we add linear time trends?
- Are the results the same or different?

Response

- We add linear time trends to the regression to account for changes over time within each county. This will account for potential omitted variables that vary over time differently between counties.

- We conclude that regression (4) was not internally valid and suffered from OVB because when we added these fixed effects, the coefficient changed drastically and is no longer significant.

Questions

A7

Question

- Explain briefly how measurement error in bankruptcy rates would affect the estimates reported in Table A2.

Thoughts

- Measurement error in **dependent** variable
- Bias?

Response

- The standard errors of the estimates will increase, and rightfully so, as there is more uncertainty in the data.

Questions

Part B

05:00

B8

Question

- Using regression (1), can you reject the hypothesis that Hajj visas are randomly assigned? Report your F statistic, its p-value, the 5% critical value for the test, and number of restrictions for this test.

Thoughts

- Where do we find F-stat, # restrictions and p-value?
- Where do we find critical value?
- When do we reject?

Response

- F-statistic: .769
- p-value: .573
- # restrictions: 5
- 5% critical value: 2.21
- Because the p-value is greater than .573, we cannot reject the null hypothesis that Hajj visas are randomly assigned.

Questions

B9

Question

- Provide a reason why the coefficient on *Hajj* in regression (2) is plausibly biased. Use the omitted variable bias formula to give the direction of the bias.

Thoughts

- Same OVB Steps!

Response

- A possible omitted variable could be individual wealth
- β_2 is positive because wealthy individuals are more likely to value education and therefore may believe girls should be educated.
- γ_1 is positive because wealthy individuals can pay for an expensive tour package allowing them to attend Hajj even if they do not win the lottery.
- Therefore the sign of the bias is positive.
- Since α_1 is positive, this means we have **overstated** the effect of Hajj.
- (Note this is different from the conclusion in the given solutions!)

Questions

B10

Question

- Is *success* likely to be a valid instrument for *Hajj* in regression (5)? Explain.

Thoughts

- Requirements for Valid Instrument
- Relevance
- Exogeneity
  - As good as randomly assigned
  - Only affects outcome through treatment

Response

- I do believe this is a valid instrument because it satisfies relevance and exogeneity
- Relevance: Our F-statistic is $(\frac{.855}{.0123})^2 \approx 4832$, which is much higher than 23.1
- Exogeneity:
  - As good as randomly assigned: We read that there is a randomly lottery and probability of being chosen is only potentially affected by group size and place of departure, which we control for in this regression. Also in column 1, our F-stat shows we cannot reject the null that they are randomly assigned.
  - Winning the lottery should not affect beliefs about girls' education in any way other than attending Hajj, as there are no other benefits/drawbacks to being chosen or not.

Questions

B11

Question

- All the regressions in Table B3 include the same set of controls (listed below). Which of these controls are needed in order for success to be a valid instrument in regression (5)? Explain.
  - a) Place of departure × Party size fixed effects
  - b) Demographic controls (listed below Table B3)

Thoughts

- Why do we add controls in a 2SLS regression?

Response

- We only need controls that allow the instrument to be as good as randomly assigned.
- a) Region and size controls are necessary because lotteries take place at a regional level, and larger groups may have a slightly smaller chance of being chosen, so once you control for these, the instrument is as good as random.
- b) Demographic controls are **not** necessary because they should have no effect on winning the lottery, and we show this is plausible in regression (1)

Questions

B12

Question

- Regression (5) estimates a local average treatment effect (LATE). In your judgment is the LATE estimated in (5) greater than, less than, or equal to the average treatment effect? Explain.

Thoughts

- LATE vs ATE Question!
- Eliminate cases where LATE = ATE
- Identify compliers (grasshoppers)
- State whether treatment effect is bigger or smaller for compliers
- State whether LATE > ATE or LATE < ATE

Response

- First, we'll eliminate the cases where LATE = ATE:
    - There is intrinsic heterogeneity in the first stage, as rich people can afford to attend Hajj whether or not they lottery.
    - There is intrinsic heterogeneity in the treatment effect, because some people may be so set in their ways nothing will change their minds.
    - It's reasonable to believe there's correlation between how rich someone is and how set in their ways they are, so $cov(\pi_1, \beta_1) \neq 0$
- The compliers are poor people, and non-compliers are rich people, because rich people can afford to get a visa even if they lose the lottery, so the instrument affects them less.
- I believe the treatment effect is larger for poor people than rich people, because rich people have the means to travel to other countries with more gender equality anyway, while Hajj may be the only time this could happen for someone without means to travel.
- Therefore, I conclude the LATE is greater than the ATE.

Questions

B13

Question

- One hypothesis is that the Hajj affects participants' beliefs by increasing their exposure to people from different countries and sects, and to members of the opposite gender outside their family. If this hypothesis is correct, pilgrims who travel in smaller parties, and thus have more opportunity to interact with non-Pakistanis, may experience larger treatment effects.
    - a) Using Column 6 of Table B3, what is the effect of the Hajj for a pilgrim in a small party?
    - b) What additional information would you need in order to compute a 95% confidence interval for the predicted effect in part (a)?
    - c) Can you reject the hypothesis that the effect of the Hajj pilgrimage is the same for large parties and small parties?

Thoughts

- Interaction Terms
- Calculating Standard Errors of Sums

Response

A. .0104 + .0356 = .0460, so the effect of attendance for a small party is increasing probability of responding that girls should go to school by 4.6 percentage points.

B. We would need the standard error of this, which we could get from the variance of the sum of the two coefficients. Since $Var(X, Y) = var(X) + var(Y) + 2cov(X, Y)$, and we already can calculate $var(X)$ and $var(Y)$ from the standard errors, the missing information is the covariance between the two coefficients.

C. The 95% CI is .0356 +- 1.96 * .0267, or [-.0167, .0878]. Since this interval includes 0, we cannot reject the null hypothesis that the effect is the same on large and small parties.

Questions

Part C

05:00

C14

Question

- Consider the regression in column 1 of Table C2
    - a) Interpret the coefficient on UnionWin in column 1 of Table C2.
    - b) As a general matter, provide a definition in words for a 95% confidence interval.
    - c) Calculate the 95% confidence interval for the coefficient on UnionWin in column 1 of Table C2.
    - d) Are the end points on your 95% confidence interval large in a real world sense? In particular, do they allow you to rule out (at the 5% level) the estimated 15% union wage premium described in introduction to Part C?

Thoughts

- Interpreting logarithmic coefficients
- CI definition and how it’s calculated
- How do we interpret CIs?

Response

A. A union barely winning an election is associated with a 2.6% decrease in employee wages.

B. A 95% CI is an interval that contains the true value in 95% of repeated samples.

C. $-.026 \pm 1.96 \times .017$ -> [-.059, .00732]

D. The endpoints are -5.9% and 0.73%, so the lower end seems large in a real-world sense, but the upper end seems very small. This interval does rule out the apparent 15% increase in wages we read about.

Questions

C15

Question

15) The regression in column 1 of Table C2 is of the following form:

$$Y_i = \beta_0 + \beta_1 UnionWin_i + \beta_2(VoteShare_i - 50\%) + \beta_3(VoteShare_i - 50\%) \times UnionWin_i + v_i$$

Consider the following modified version of this regression that replaces $VoteShare_i - 50\%$ with $VoteShare_i$:

$$Y_i = \alpha_0 + \alpha_1 UnionWin_i + \alpha_2 VoteShare_i + \alpha_3 VoteShare_i \times UnionWin_i + u_i$$

What is the relationship between β_1 and α_1?

Thoughts

- Oh no!
- I remember deriving this in class, but I don't remember the steps.
- Think about what β_1 represents: the difference at the threshold between winning and losing.

Response

- $Y(50,W) = \alpha_0 + \alpha_1 + 50\alpha_2 + 50\alpha_3$
- $Y(50,L) = \alpha_0 + 50\alpha_2$
- $Y(50,W) - Y(50,L) = \alpha_1 + 50\alpha_3$
- So we can conclude that $\beta_1 = \alpha_1 + 50\alpha_3$

Questions

C16

Question

- Explain what Figure C1 shows and how it relates to the key assumption needed for the regressions in Table C2 to measure the causal effect of unionization on workers' wages.

Thoughts

- McCrary Test
- What’s the null hypothesis and what’s the critical value?
- When is this a problem?

Response

- Figure C1 shows evidence of manipulation of the running variable.
- This is backed up by the fact that the McCrary test statistic is 9.997, which is larger than the critical value of 1.96, so we reject the null that there is no manipulation.
- This threatens our regression discontinuity design because it provides evidence that what side of the threshold we are on is not as good as randomly assigned. There are significantly more unions who barely lose than barely win elections, and the types of companies where these votes occurred could plausibly be different on either side of the election.

Questions

Part D

05:00

D17

Question

- Table D2 reports two types of standard errors below the estimated regression coefficients: heteroskedasticity-robust standard errors in parentheses, and Newey-West standard errors in square brackets.
  - a) Briefly explain the difference between Newey-West standard errors and heteroskedasticity-robust standard errors.
  - b) Explain which set of standard errors is preferred, and why.

Thoughts

- Definition of HR vs Newey-West Standard Errors
- When do we use one or the other?

Response

A. HR standard errors account for heteroskedasticity, while Newey-West standard errors also account for autocorrelation of the error term.

B. It depends. If we have included enough lags of Y in our model (which we know is the case if we select our model using BIC) then we can use HR standard errors. If we include too few lags of Y, or if we are doing multi-step ahead forecasting, we should use Newey-West standard errors.

Questions

D18

Question

- Choose a preferred forecasting model among regressions (1)-(5) in Table D2. Explain your reasoning.

Thoughts

- How do we decide on a model?
- What info is given to us?
- Read the question carefully!

Response

- I chose regression (3) because it has the lowest BIC.

Questions

D19

Question

- The bottom panel of Table D2 reports both the root mean squared error (RMSE) and a GARCH(1,1) forecast standard deviation for 2018w18. Consider the following statement:

  “The forecast intervals constructed using the GARCH(1,1) estimate of the root mean squared forecast error (RMSFE) are narrower than the intervals constructed with the RMSE. Therefore, the GARCH(1,1) estimate of the RMSFE is preferred.”

  Do you agree with the logic in the statement? Explain.

Thoughts

- Always a tough one.
- Make sure you understand the statement
- RMSE vs GARCH, how do we choose?

Response

- I disagree with the logic of the statement but agree with its conclusion.
- I disagree with the logic because we should choose our RMSFE estimate based on which one will more accurately describe our uncertainty, which is not necessarily the one that is smaller.
- I agree with the conclusion that we should use the GARCH estimate because RMSE assumes the volatility is constant throughout time, and GARCH estimates the volatility for a specific point in time.

Questions

D20

Question

- The recent values of ILI outpatient visits are given below Table D1.
    - a) Using model (6), compute the forecast of ILI outpatient visit growth (percent per week) in the U.S. for 2018w18, which will be released tomorrow.
    - b) Using model (7), compute the forecast of ILI outpatient visit growth (percent per week) in the Taiwan for 2018w18, which will be released tomorrow.
    - c) Using model (6) compute the forecast of ILI outpatient visit growth (percent per week) in the U.S. for 2018w19, which will be released next week.

Thoughts

- Plug and chug
- Part c) is a bit tricky: 2 steps ahead

Response

A. .0272 + (.382)(-5.4079) + (.0117)(-5.5521) = -2.10358 %

B. -.00186 + (.130)(-5.4079) + (-.283)(-5.5521) = 0.86636 %

C. We can use our above predictions to make a prediction two steps ahead:
.0272 + (.382)(-2.10358) + (.0117)(0.86636) = -0.76623 %

Questions

D21

Question

- Figure D2 plots the Quandt Likelihood Ratio (QLR) statistic, computed for regression (6) from Table D2. The maximum value from the figure is 4.461.
    - a) What are the critical value at the 5% significance level and number of restrictions for this test?
    - b) What do you conclude from the QLR test?

Thoughts

- How many restrictions do we have?
- Where do I find critical values?
- What’s my null hypothesis? Do I reject?

Response

A.

- There are 3 restrictions in this test.
- The critical value for this test is 4.71

B. Because the QLR statistic is less than the critical value, we do not reject the null hypothesis that there is no break in stationarity.

Questions

Conclusions

Takeaways

- Correct answers can be much shorter than the example ones
- Key question types: “Why might a simple OLS be biased?” -> OVB
- Cheat Sheet Ideas

Final Advice

- Some finals are all about memorization, on these there's a high return to all-nighters
- This final allows a HUGE cheatsheet, meaning there's a low return to all-nighters
- This final requires that you think critically and carefully read about new information, meaning there's a negative return to all-nighters
- Get a good night's sleep, eat/drink something before arriving, and try to relax as much as you can!

ECON 1123 Review Session

Slides at github.com/cjleggett/1123-section