Midterm

Jacob Jameson

Due 8PM on March 9, 2023

1. Answer the following questions about the outcome and treatment effect we're interested in.

a. Using potential outcomes notation, write and interpret the two potential outcomes for an individual in this study. [2 points]

Define: $D_i \in \{0,1\}$ as the treatment indicator (winning the lottery for voucher) for student i:

- When student i is treated $\implies D_i = 1$
- When student i is not-treated $\implies D_i = 0$

Define $Y_i(D_i)$ as the outcome (test scores) for student i as a function of D_i:

- When student i is treated we observe $Y_i(1)$ (the test scores on a treated student)
- When student i is not-treated we observe $Y_i(0)$ (the test scores on an untreated student)

Thus, the impact of the treatment for student i is:

$$\tau_i = Y_i(1) - Y_i(0)$$

Where this represents the causal effect of the being assigned to the treatment group (winning the lottery). Note that these represent the counterfactual outcomes of student i, as we would not be able to observe both outcomes where student i either won or lost the lottery.

b. Using potential outcomes notation, write out the average treatment effect of attending a private school on normalized test scores, all else equal. [2 points]

We can estimate the average treatment effect (ATE) as the difference between the average outcome on normalized test scores in the treated population and the average outcome on normalized test scores in the control population: $\tau^{ATE} = Y(1) - Y(0)$.

We can conveniently do this is through a regression: $Y_i = \alpha + \tau D_i + \varepsilon_i$. We can do this because, thanks to randomization, we can assume that $E[\varepsilon|D_i] = 0$ and therefore that $\hat{\tau} = \hat{\tau}^{ATE}$.

2. Answer the following questions about the assumptions.

a. What is the Conditional Independence Assumption generally? What does it assume in this example? Explain in plain English using variable names. [3 points]

The Conditional Independence Assumption:

$$E[Y_i|X_i, D_i = 1] - E[Y_i|W_i, D_i = 0] = E[Y_i(1) - Yi(0)|X_i]$$

frequently also denoted as $(Y_1, Y_0) \perp D|X$, essentially means that if we control for characteristics X, the treatment assignment is independent of the potential outcomes. In equation (1), the authors' main estimating equation for the impact of receiving the voucher, this means that controlling for Normalized baseline Telugu score, Normalized baseline math score, Both parents have completed at least primary school, At least one parent has completed grade, Scheduled caste, and Household asset index, receiving a voucher should be independent of the potential outcome of normalized test scores.

b. How might we check if that assumption holds? [2 points]

We can check that this assumption holds, by ensuring that the treatment effect does not depend on covariates X_i.

Let $\beta W \equiv E[Y_i|X_i, D_i = 1] - E[Y_i|W_i, D_i = 0]$

If the Conditional Independence Assumption holds, this difference in means will give the conditional average treatment effect (CATE).

3. The table below shows coefficients and standard errors from a regression of treatment assignment on baseline characteristics, as well as the p-value from an F test of their joint significance. The regression includes cohort fixed effects. What do you conclude from this table? Justify your conclusion. [3 points]

From this table I conclude that the CIA holds. It appears that our baseline characteristics are not independently or jointly significantly associated with treatment assignment. We can determine this by looking at the size of the standard errors on each of these coefficients in the regression and by looking at the p-value from an F test of their joint significance (0.87). This seems to match with intuition, that being that the lottery systems seems like a pretty solid randomization design and so I would not have thought that treatment assignment would have been correlated with some of these observable baseline characteristics.

4. Suppose some students who participated in the lottery dropped out during the next 4 years after the voucher was offered. As a result, we are unable to observe their math normalized test scores. Would this threaten the internal validity of the study? Justify your answer. [3 points]

This attrition does have potential to threaten the internal validity of this study. In an experiment such as this, differential rates of attrition between treatment and control groups can skew results and generate bias on the impact of the treatment. In the paper, the authors fine that the four-year attrition rate was 15% and 19% in the treatment and control groups, and that these differences are statistically significant. The authors find no difference in observable characteristics between the attritors across the treatment and control arm, but despite the series of robustness checks, we cannot state for sure whether or not this created imbalance on unobservables. So I would say there is potentially a threat to the internal validity of the study.

5. Answer the following questions about the ITT.

a. Using potential outcome notation, write the ITT as an average treatment effect (conditional if necessary). Show how we can re-write it to be in terms of outcomes we can observe in the data. State where you use the CIA. [3 points]

Recall that the average treatment effect, $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$, gives us the average effect of the treatment, but in the case of this experiment

$$ITT = E[Yi(TreatmentOffered) - Yi(NotOffered)]$$

b. Write down a regression specification to estimate the intent-to-treat (ITT) effect, adjusting for all available appropriate controls and optional valid controls. Identify the coefficient of interest. Estimate this ITT using regression and report the coefficient and standard error of the estimate. Explain why some controls are required and others are options. [3 points]

The regression specification used by the authors to estimate the intent-to-treat (ITT) effect, adjusting for all available appropriate controls and optional valid controls was:

$$T_{isv}(Y_n) = \beta_0 + \beta_1 \cdot T_{isv}(Y_0) + \beta_2 \cdot Voucher_i + \beta_{Z_i} \cdot Z_i + \beta_{X_i} \cdot X_i + \varepsilon_{isv}$$

The coefficient of interest in this regression is β_2, which provides an unbiased estimate of the impact of winning a voucher on test scores (the ITT estimate) since the voucher was assigned by lottery.

It appears my data is extremely limited and I do not have access to roughly half of the observable baseline characteristics, or geography. This means that I cannot do the same analysis as the authors, and cannot cluster my standard errors at the village level.

```r
data.1 <- read_dta("api115_midterm_part1.dta")

mdls <- list(
reg.1 <- lm(math_4 ~ voucher + math_0 +
            hh_asset + female, data = data.1),

reg.2 <- lm(telegu_4 ~ voucher +
            hh_asset + female, data = data.1)
)

# LaTeX output of 2 Models
stargazer(
  mdls, type = "latex", title = 'ITT Effects',
  column.labels = c('Model 1',
                    'Model 2'),
  dep.var.labels = c('Year 4 Math Scores', 'Year 4 Telegu Scores'),
  style = 'qje')
```

The coefficient and standard error I get on the effect of the voucher on normalized math scores 4 years later is 0.164 (0.043) and for normalized Telegu scores $-$0.010 (0.073). Some controls are requires, such as baseline test scores since test scores are highly correlated over time, but other controls such as female might not be necessary because we are balanced on this feature after randomization and we don't expect to be correlated with outcomes.

c. Interpret the coefficient with appropriate units -what does this estimate tell you? [3 points]

I will interpret for model 1 since this is the model that includes the relevant baseline test score. The coefficient on voucher is 0.164 for model 1. This indicates that the impact of winning the voucher lottery on normalized math test scores is a 0.164 unit increase. This is also statistically significant. This estimate tells me that the impact of receiving the voucher was beneficial for students' math scores, controlling for baseline match scores, gender, and socioeconomic indicators.

Table 1: ITT Effects

	Year 4 Math Scores	Year 4 Telegu Scores
	Model 1	Model 2
	(1)	(2)
voucher	0.164***	−0.010
	(0.043)	(0.073)
math_0	0.476***	
	(0.020)	
hh_asset	0.087***	0.003
	(0.021)	(0.036)
female	0.178***	0.124*
	(0.040)	(0.069)
Constant	0.036	−0.105*
	(0.032)	(0.054)
N	900	900
R^2	0.404	0.004
Adjusted R^2	0.402	0.0003
Residual Std. Error	0.603 (df = 895)	1.034 (df = 896)
F Statistic	151.961*** (df = 4; 895)	1.083 (df = 3; 896)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

6. Answer the following questions about the TOT.

a. Imagine you compare the baseline and final normalized test scores only for those who were assigned to and selected into treatment. The coefficient you get is 0.34. What could be wrong if we want to interpret this coefficient as a TOT? [5 points]

The issue with interpreting this coefficient as a TOT is that there may be selection bias issue. If it turns out that those students who were received a voucher and then decided to attend a private school had a higher level of intrinsic motivation to improve themselves compared to the group that received a voucher and decided not to attend a private school, then we may be overstating the TOT with the coeffecient 0.34.

b. Now re-estimate your regression equation from 5a, but this time exclude any lottery winners who did not attend a private school. Report the coefficient. What might be wrong with this estimate as a measure of the TOT? [5 points]

```
mdls <- list(
reg.1b <- lm(math_4 ~ voucher + math_0 +
            hh_asset + female,
            data = filter(data.1, !(voucher == 1 & attended == 0))),

reg.2b <- lm(telegu_4 ~ voucher +
            hh_asset + female,
           data = filter(data.1, !(voucher == 1 & attended == 0)))
)

# LaTeX output of 2 Models
stargazer(
  mdls, type = "latex", title = 'TOT Effects',
  column.labels = c('Model 1',
                    'Model 2'),
  dep.var.labels = c('Year 4 Math Scores', 'Year 4 Telegu Scores'),
  style = 'qje')
```

I am going to only interpret Model 1, which is the math scores model. The coefficient on voucher is 0.185, which is the impact of receiving the voucher and attending private school on future math scores when we exclude those that received the voucher and did not attend private school. It is still incorrect to use this estimate as a measure of the TOT for the same reasons mentioned earlier. We are biasing this coefficient by dropping observations that got the voucher and did not attend because there may be unobservable differences between these groups of people so we are inducing selection bias.

c. Now calculate the effect of attending a private school by using 2SLS or the Wald estimator. Interpret this estimate. [5 points]

```
ivreg.1 <- ivreg(math_4 ~ attended + math_0 +
            hh_asset + female | voucher + math_0 +
            hh_asset + female, data = data.1)

# LaTeX output of 2SLS
stargazer(
  ivreg.1, type = "latex", title = '2SLS Model',
  column.labels = c('Model 1'),
  dep.var.labels = c('Year 4 Math Scores'),
  style = 'qje')
```

Table 2: TOT Effects

	Year 4 Math Scores Model 1	Year 4 Telegu Scores Model 2
	(1)	(2)
voucher	0.185^{***}	0.027
	(0.044)	(0.076)
math_0	0.473^{***}	
	(0.020)	
hh_asset	0.086^{***}	-0.012
	(0.021)	(0.037)
female	0.176^{***}	0.114
	(0.041)	(0.070)
Constant	0.036	-0.100^{*}
	(0.032)	(0.054)
N	868	868
R^2	0.409	0.003
Adjusted R^2	0.406	0.00001
Residual Std. Error	0.599 (df = 863)	1.030 (df = 864)
F Statistic	149.110^{***} (df = 4; 863)	1.004 (df = 3; 864)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Table 3: 2SLS Model

	Year 4 Math Scores Model 1
attended	0.184***
	(0.048)
math_0	0.474***
	(0.020)
hh_asset	0.084***
	(0.021)
female	0.175***
	(0.040)
Constant	0.037
	(0.032)
N	900
R^2	0.406
Adjusted R^2	0.404
Residual Std. Error	0.602 (df = 895)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

In the 2SLS model the impact of attending private school was a 0.184 increase on future standardized math scores.

d. Are the estimates in 6.a, 6.b, and 6.c different from the ITT you calculated in question 5? Why? In your answer, reference both the reasons for any differences and the direction they influence relevant estimates. [5 points]

Yes, the estimates in 6.a, 6.b, and 6.c are different from the ITT I calculated in question 5, which was 0.164. Why? Because in 5 we are estimating the ITT (i.e. the effect of winning the voucher), vs. in 6 we are trying to get at the TOT (i.e. the effect of attending the private school). The differences stem from the fact that the ITT estimates the average effect of offering the treatment on outcomes, or the effect on everyone who was offered the treatment, whether or not they received it. The TOT estimates the average effect of the actual treatment on outcomes, or the effect only on those who received the full treatment. It is not surprising that the TOT estimates were all slightly larger than the ITT estimates because we are excluding those that did not attend private school but did receive the voucher so we are likely introducing an upward bias on the estimate for the reasons I have already mentioned earlier in these problems.

7. What would you say about the effectiveness of private schools based on your analysis? Make sure to highlight any key takeaways as well as any limitations of your analysis using non-technical language. [6 points]

Based on my analysis in this assignment (I am going to base my takeaways on my own analysis and not on the paper since they lead to slightly different results), I think that private schools appear to have positive associations with future test outcomes. Note that I am saying associations as I am purposefully not using causal langauge because I am not convinced (within this toy dataset) that my estimates are robust enough to use causal language. Across the several different specification we did, we saw that there was consistently a positive coeffecient on the voucher variable in our regression models.

Part II [34 points]-Instrumental Variables

1. Why is an IV approach necessary in this context instead of a naive observational regression (i.e. one that regresses math exam scores on attending a private school that teaches in English and all valid control variables)? Give one concrete example of what you might be concerned about. For full credit, be sure to explain in which direction you might expect bias and why. [7 points]

An IV approach is necessary in this context instead of a naïve observational regression, because in the naïve observational regression, attending a private school that teaches in English might be correlated with unobserved characteristics that affect math scores. For example, imagine that a family really values education. They may have aspirations of having their child attend college in Europe or the US one day. If this is the case they me bore likely to send their lottery winning kids to private schools where the language of instruction is English. These same families likely also invest more time into education at home, they probably read to their kids more often or encourage them to do homework. In this scenario the naive observational regression will overstate the effect of attending a private school that teaches in English on math scores, since the observed effect would capture both the direct effect of attending the school and the effect of having a family that values education more highly.

Using the suggested IV approach, we can address these potential sources of bias and better estimate the heterogeneity of the causal effect of private school attendance by the private school's language of instruction.

2. Answer the following questions about the IV assumptions.

a. Identify the four assumptions for IV with heterogeneous treatment effects. Explain what they mean in this context. [4 points]

- Relevance: The instrument must affect the treatment status. This means that the language of instruction of the nearest private school must be correlated with attending a private school that teaches in English.
- Independence: The instrument must be as good as randomly assigned, which in this case means that the language of instruction of the nearest private school must be as good as random. This may not hold for us.
- Exclusion: The instrument must affect the outcome only through the treatment, in this context it means that there are no spillover effects of having the nearest private school teach in English. The language of instruction should only affect math scores through its effect on attending a private school that teaches in English.
- Monotonicity: The instrument must affect everyone in the same way, which in this context means that being near a private school that teaches in English weakly increases the probability that you will attend an English speaking private school.

b. Do you think they are reasonable? Why or why not? How can you check? [4 points]

I think most of these assumptions are reasonable except I am skeptical about independence. I think for the same reasons we might think students in English speaking private schools do better on math (i.e. their families value education more), that families might choose to reside in areas with greater proximity to English instruction schools. If this is the case then the independence assumption is violated. In order to check this I may survey families to collect data on the reasons why they chose to reside in certain geographical locations and whether or not instruction language of the school was a factor.

3. We want to find the effect of attending a private school that teaches in English. Write down the first stage regression specification, including all available appropriate controls. Run the first stage of the IV and interpret the coefficient. What do your results say about the validity of the instrument? (Hint: Following Muralidharan and Sundararaman (2015), we are interested in the interactions of both languages of instruction with receiving a voucher. Note that in our data the variable near_tel = 1 - near_eng; therefore, do not include both in the regression. All other valid controls should be included) [4 points]

In the first stage of 2SLS, we regress our covariates on the instruments (again, non-endogenous covariates are their own instruments). Thus our first stage is :

$$atten\hat{d}\_eng = \beta_0 + \beta_1 voucher * near\_eng + \beta_3 X$$

```
data.2 <- read_dta("api115_midterm_part2.dta")


first_stage <- lm(attend_eng ~ voucher*near_eng +
                      hh_asset + female + math_0, data = data.2)

# LaTeX output of 2SLS
stargazer(
  first_stage, type = "latex", title = 'First Stage',
  column.labels = c('Model 1'),
  dep.var.labels = c('Attending a School in English'),
  style = 'qje')
```

Table 4: First Stage

	Attending a School in English Model 1
voucher	0.135^{***} (0.018)
near_eng	0.144^{***} (0.013)
hh_asset	0.044^{***} (0.006)
female	0.017 (0.011)
voucher:near_eng	0.699^{***} (0.028)
Constant	0.009 (0.010)
N	2,000
R^2	0.535
Adjusted R^2	0.534
Residual Std. Error	0.253 (df = 1994)
F Statistic	459.677^{***} (df = 5; 1994)
Notes:	***Significant at the 1 percent level. **Significant at the 5 percent level. *Significant at the 10 percent level.

The coeffecient on our instrument is 0.698598, meaning that receiving a voucher when the nearest private school teaches in English increases the probability that the student attends a private school that teaches in English.

4. Now estimate the causal effect on math scores of switching to a school that uses English as the language of instruction, assuming constant treatment effects. Interpret your estimate. Note that this may not match the results in the paper. [4 points]

```
ivreg(math_4 ~ attend_eng | voucher*near_eng,
                data = data.2) %>%
  stargazer(type = "latex", title = '2SLS: Constant Treatment Effects',
  column.labels = c('Model 1'),
  dep.var.labels = c('Normalized Future Math Scores'),
  style = 'qje')
```

Table 5: 2SLS: Constant Treatment Effects

	Normalized Future Math Scores Model 1
attend_eng	0.012 (0.052)
Constant	0.056^{***} (0.016)
N	2,000
R^2	−0.00003
Adjusted R^2	−0.001
Residual Std. Error	0.622 (df = 1998)

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

My statistically insgnificant estimate of 0.012 suggests that affect of attending an english speaking private school does not matter for future math scores.

5. Now, without assuming constant treatment effects, what population does your IV estimate apply to? What population(s) does it not apply to? Explain who these groups of people are in this context. [4 points]

```
ivreg(math_4 ~ attend_eng +  hh_asset + female + math_0 |
                  voucher*near_eng +  hh_asset + female + math_0, data = data.2 )%>%
  stargazer(type = "latex", title = '2SLS: Heterogeneous Treatment Effects',
  column.labels = c('Model 1'),
  dep.var.labels = c('Normalized Future Math Scores'),
  style = 'qje')
```

My IV estimate can be applied to students with the characteristics controlled for in my regression (female, houehold assets, baseline math scores)

The main result is the lack of any consistent evidence of heterogeneous effects along most student characteristics.

Table 6: 2SLS: Heterogeneous Treatment Effects

	Normalized Future Math Scores
	Model 1
attend_eng	0.004
	(0.052)
hh_asset	0.042^{***}
	(0.014)
female	0.122^{***}
	(0.027)
math_0	0.090^{***}
	(0.015)
Constant	-0.005
	(0.021)
N	2,000
R^2	0.036
Adjusted R^2	0.034
Residual Std. Error	0.611 (df = 1995)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

6. Provide (1) the percentage of the population that are compliers and (2) the average baseline math score of compliers. [3 points]

```r
reg <- lm(attend_eng ~  voucher*near_eng, data = data.2)
paste0(round(summary(reg)$coefficients[2],6)*100,'%')
```

```
## [1] "13.1428%"
```

13.14% are compliers. The average baseline math score of compliers is:

7. Now suppose that in later years, students with vouchers were randomly assigned to a private school using a lottery instead of being able to choose the school they use their voucher for. An analysis of the lottery finds that being randomly assigned to private schools that teach using English decreases the math scores for the students by 0.15 standard deviations.

a. How does the lottery estimand compare to your IV estimand in terms of the type of estimate and the population it applies to? [2 points]

My IV estimand represents the effect of attending an english speaking private school among the students who’s compliance is modified by the interaction being near an english speaking school. Therefore the effect we identified was for compliers only.

Tthe lottery estimand represents the effect of attending a private school that teaches using English on math scores for all students who received a voucher to attend a private school, regardless of whether they would have attended one that teaches in english or somethiing else.

Therefore both of these estimate the effect of attending an english instructed private school on math scores, but we get different effects because they are referring to different populations.

b. Explain and give an example of how a violation of the exclusion restriction for IV may generate this discrepancy. [1 point]

If there is a violation of the exclusion restriction assumption in the IV framework, this could generate a discrepancy between the IV estimand and the lottery estimand. For example if the interaction of voucher and near_eng affects future math scores not only through its effect on attend_eng but also through another pathway.The lottery estimand is not affected by biases due to the influence of the instrument on the outcome through channels other than the treatment variable.

c. Explain and give an example of how a violation of the exogeneity/independence assumption for IV may generate this discrepancy. [1 point]

If there is a violation of the exogeneity/independence assumption for IV , then the IV estimand may be biased, createing a discrepancy between the IV estimand and the lottery estimand. Suppose that students who live near English-speaking private schools are more likely to have families who value education highly and they invest more time and resources into their child’s education at home. In this case, both the instrument and the treatment variable may be correlated with unobserved factors that affect future math scores. The lottery estimand is not affected by such biases because it estimates the causal effect of attending an English-speaking private school on math scores for all students who were randomly assigned to receive a voucher.

Part III [16 points] -Regression Discontinuity

1. What are two necessary conditions for a regression discontinuity design to produce internally valid estimates of the impact of private school attendance? [4 points]

Two necessary conditions for a regression discontinuity design to produce internally valid estimates of the impact of private school attendance would be:

- Pre-treatment covariates should be balanced across the cutoff. This means that pre-treatment characteristics should be balanced between students a scored right below the cutoff score and right above the cutoff score.
- The density of X_i should be continuous at the cutoff point.

2. Describe how you would use local linear regression to generate the causal estimate of interest. Define your sample and variables clearly, describe the regression or regressions you would run, and identify which coefficient or combination of coefficients represents the causal estimate. [4 points]

We could consider the regression discontinuity an instrumental variable and estimate first and second stage regressions with a score $\geq$ 70th percentile being our instrument:

First stage:

$$D_i = \alpha + \gamma 1[X_i \geq c] + v_i$$

Second stage:

$$Y_i = \alpha + \tau \hat{D}_i + \varepsilon_i$$

Where the in the first stage we are estimating γ, the effect of being above the 70th percentile has on treatment (attending a private school). In the second stage, we are using the predicted treatment probability from our first stage regression to estimate τ which will be the effect that predicted treatment has on the outcome of interest (future test scores).

This causal effect is identified for compliers at the threshold (in our case we have full compliance). This means the effect that we are estimating is the effect of private school education at 70th percentile compared to non private school education at the 69th percentile.

3. Between the regression discontinuity and the lottery estimate of the causal effect of attending a private school from Part I, which do you think is more relevant to policy and why? [2 points]

I am not completely sure what is meant here by policy relevant. If in both cases we get the causal impact of attending private school then they are both equally relevant to policy because we should get roughly the same estimate and directional impact of private school education. If this question is asking is it easier to implement a lottery voucher system or a test threshold system where only siblings of private school enrollees can attend, I would think the lottery is easier to implement.

4. What is an alternative to the local linear approach? List some advantages and disadvantages of the local linear approach vs. the alternative approach? [3 points]

An alternative to the local linear approach in regression discontinuity is the fuzzy regression discontinuity design.

The advantages of the local linear approach include its simplicity, transparency, and flexibility. The disadvantages of the local linear approach include sensitivity to bandwidth and biased treatment effect estimates.

The advantages of the fuzzy regression discontinuity design include increased precision of the estimate of the policy. Disadvantages are the complexity of the design

5. Suppose that due to privacy concerns, you could only observe the grade quantiles in 5pp increments (65th, 70th, 75th, 80th quantile etc) where the actual quantile value would just be rounded down to highest bin (e.g. 54.5th quantile would appear as the 50th; 61th as the 60th). Consider a person from your sample who'd be at a 68.9th quantile. Would you likely over or underweight this individual if blindly applying the same RD methods? Also, can you still identify the marginal causal effect at the discontinuity? [3 points]

If we are rounding down, the individual at the 68.9th quantile would appear in the 65th quantile. Hence, if we applyed the same RD methods as before you would likely underweight this individual since their actual percentile score is closer to the cutoff threshold than we have data for. I believe that you can still estimate the marginal causal effect at the discontinuity, but the estimates precision will likely be reduces because of the loss of information from rounding.