MIDTERM EXAM

I - INSTRUCTIONS

* You will have from 7am to 8pm to complete this exam.
* By 8pm, save your exam and submit it and your **script** to Canvas.
* You may use any programming language.
* The exam is out of 100 points. If you get bogged down on something you may want to move on to gather points from questions you understand more quickly.
* In open-answer questions please be concise.
* If you have a clarification question you may email the entire teaching team. We will do our best to respond.
* If you’re unsure of how to answer a question, just use your best judgment and explain why you did what you did.

II – EXAM POLICIES

Although collaboration in the problem sets is permitted, exams are different. There is no collaboration allowed. This exam is open-book, open-notes, open-Internet, but you **may not speak to each other or anyone else, in person or on the Internet or in any form.**

The standard rules for plagiarizing apply – you must type your own answers and code.

Before you begin, please read this excerpt from the [HKS Honor Code](https://www.hks.harvard.edu/educational-programs/academic-calendars-policies/student-handbook/general-regulations-and-1) and write your initials in the box below:

*Cheating on assignments or exams, plagiarizing or misrepresenting the ideas or language of someone else as one’s own, falsifying data, or any other instance of academic dishonesty violates the standards of our community, as well as the standards of the wider world of learning and affairs. Using someone else’s words or concepts without attribution is a serious violation of the Academic Code. It is the student’s responsibility to learn and use the proper forms of citation. If students submit work either not their own or without clear attribution to the original source, including but not limited to the Internet, they will be subject to discipline by the HKS Administrative Board, ranging from a warning to required withdrawal or expulsion from Harvard Kennedy School.*

|  |  |
| --- | --- |
|  | **Your Initials** |
| I certify that my work in this exam complies with the Harvard Kennedy School Academic Code and that I did not communicate with anyone during the exam. | |  |

**Introduction:**

In this exam, you will be building on the results from Muralidharan and Sundararaman (2015):

Karthik Muralidharan, Venkatesh Sundararaman, The Aggregate Effect of School Choice: Evidence from a Two-Stage Experiment in India, The Quarterly Journal of Economics, Volume 130, Issue 3, August 2015, Pages 1011–1066, <https://doi.org/10.1093/qje/qjv013>

As in the original paper, we are interested in exploring the effect of private school vouchers on educational outcomes. We will examine two hypothetical eligibility rules for private schools and heterogeneity in treatment effects.

You are encouraged to refer to the original paper to help understand the context, though the data you will use is completely simulated so the results in the paper may or may not hold.

**Part I [50 points] – Randomized Control Trials**

The file api115\_midterm\_part1.dta contains simulated cross-sectional data of normalized test scores of students from private and public schools in India. Suppose parents were offered a voucher to send their children to private schools, and the vouchers were allocated through a random lottery. The lottery was run for different cohorts of students and the number of applicants was growing cohort by cohort. Assume that the program faced the typical problem of excess demand; vouchers can only be assigned to a limited number of children per cohort. None of the lottery losers were able to attend private schools. While all of the lottery winners were eligible to enroll in a private school, some lottery winners opted to attend a public school instead. You are interested in the causal effect of attending a private school.

The data sample consists of lottery participants and contains the following variables:

=========== ===============================================================================

variable label

=========== ===============================================================================

studentid Student ID

voucher Indicates whether the student was offered a private school voucher

attended Indicates whether the student attended a private school

math\_4 Math exam score (normalized) four years after the vouchers offered (in std. dev.)

math\_0 Math exam score (normalized) in the year before vouchers offered (in std. dev.)

telugu\_4 Telugu exam score (normalized) four years after vouchers offered (in std. dev.)

female Is student female?

hh\_asset Household asset index

===================================================================================

1. Answer the following questions about the outcome and treatment effect we’re interested in.
   1. Using potential outcomes notation, write and interpret the two potential outcomes for an individual in this study. **[2 points]**
   2. Using potential outcomes notation, write out the average treatment effect of attending a private school on normalized test scores, *all else equal*. **[2 points]**
2. Answer the following questions about the assumptions.
   1. What is the Conditional Independence Assumption generally? What does it assume in this example? Explain in plain English using variable names. **[3 points]**
   2. How might we check if that assumption holds? **[2 points]**
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]**

Table

Description automatically generated

1. 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]**
2. Answer the following questions about the ITT.
   1. 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]**
   2. 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]**
   3. Interpret the coefficient with appropriate units - what does this estimate tell you? **[3 points]**
3. Answer the following questions about the TOT.
   1. 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]**

* 1. 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]**
  2. Now calculate the effect of attending a private school by using 2SLS or the Wald estimator. Interpret this estimate. **[5 points]**
  3. 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]**

1. 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]**

**Part II [34 points] – Instrumental Variables**

The file api115\_midterm\_part2.dta contains the same sort of simulated cross-sectional data as in Part I, but with variables specifying whether they attended a private school that teaches in English or Telugu and the language of instruction of their nearest private school. All public schools in the study teach using Telugu and all private schools in the study either teach using Telugu or English. Some students who were randomly provided a school voucher were living near a private school that uses English as their language of instruction while others were living near a private school that uses their state’s native language, Telugu. You can therefore use the language of instruction of the nearest private school interacted with the receipt of the randomly assigned voucher as an instrumental variable for the language of instruction of the private school that the student attends. As before, the outcome variable is math scores four years after vouchers were offered to the students’ cohort normalized relative to the distribution of public school students by subject and grade. You are interested in estimating the heterogeneity of the causal effect of private school attendance by the private school’s language of instruction. In particular, we would like to focus on the effect of attending a private school that teaches using English.

Note that you **do not** need to consider recent work questioning the interpretation of two-staged least squares as LATE when there are covariates (e.g. Blandhol et al., 2022).

The data sample consists of lottery participants and contains the following variables:

=========== ===============================================================================

variable label

=========== ===============================================================================

studentid Student ID

near\_eng Does the nearest private school to the student teach in English?

near\_tel Does the nearest private school to the student teach in Telugu?

voucher Indicates whether the student was offered a private school voucher

attend\_eng Indicates whether the student attended a private school that teaches in English

math\_4 Math exam score (normalized) four years after the vouchers offered (in std. dev.)

math\_0 Math exam score (normalized) in the year before vouchers offered (in std. dev.)

female Is student female?

hh\_asset Household asset index

===================================================================================

1. Why is an IV approach necessary in this context instead of a naïve 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]**
2. Answer the following questions about the IV assumptions.
   1. Identify the four assumptions for IV with heterogeneous treatment effects. Explain what they mean in this context. **[4 points]**
   2. Do you think they are reasonable? Why or why not? How can you check? **[4 points]**
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]**

1. 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]**
2. 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]**
3. Provide (1) the percentage of the population that are compliers and (2) the average baseline math score of compliers. **[3 points]**
4. 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.
   1. 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]**
   2. Explain and give an example of how a violation of the exclusion restriction for IV may generate this discrepancy. **[1 point]**
   3. Explain and give an example of how a violation of the exogeneity/independence assumption for IV may generate this discrepancy. **[1 point]**

**Part III [16 points] – Regression Discontinuity**

You do not need to use any data to answer the questions in this part of the exam. Suppose the siblings of private school students automatically receive a scholarship to attend their sibling’s private school themselves if they score above the 70th percentile on a standardized private school entrance exam. Otherwise, they are ineligible. Assume some but not all siblings opt to take the exam, but all siblings who score above the 70th percentile threshold and receive a scholarship choose to attend their sibling’s private school. You can use a regression discontinuity design around the 70th percentile of the entrance exam’s scores to estimate the effect of attending private schools on test scores for a sample of students who take the private school entrance exam

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]**
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]**
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]**
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]**

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]**