Prof. Will Dobbie
Advanced Quantitative Methods II: Econometric Methods (API 210)

Harvard Kennedy School
Harvard University

MIDTERM EXAM

9 Mar, 2023

I - INSTRUCTIONS

- You will have until 7pm to complete this exam.
- BY 7pm, save your exam as a **PDF** and submit it and your **R script** to Canvas.
- 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 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.	SC

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 one 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 [60 points] – Randomized Control Trials

The file `api210_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.)
female       Is student female?
hh_asset     Household asset index
===========================================================================================
```

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. **[4 points]**

   Potential Outcome Notation:
   Let Yi be the outcome of interest, in this case the math exam score four years after treatment, ie variable math_4

   Let Di be the variable that denotes whether the student i received treatment or not, in this case Di = 1 if they won the lottery and Di = 0 if they did not win the lottery

   Therefore,
   Yi(1) = Math score of student i if they won the lottery to private school
   Yi(0) = Math score of student i if they did not win the lottery to private school

   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]**

   Average Treatment Effect (ATE) = E[Yi(1) – Yi(0) | D1]

2. Answer the following questions about the assumptions.

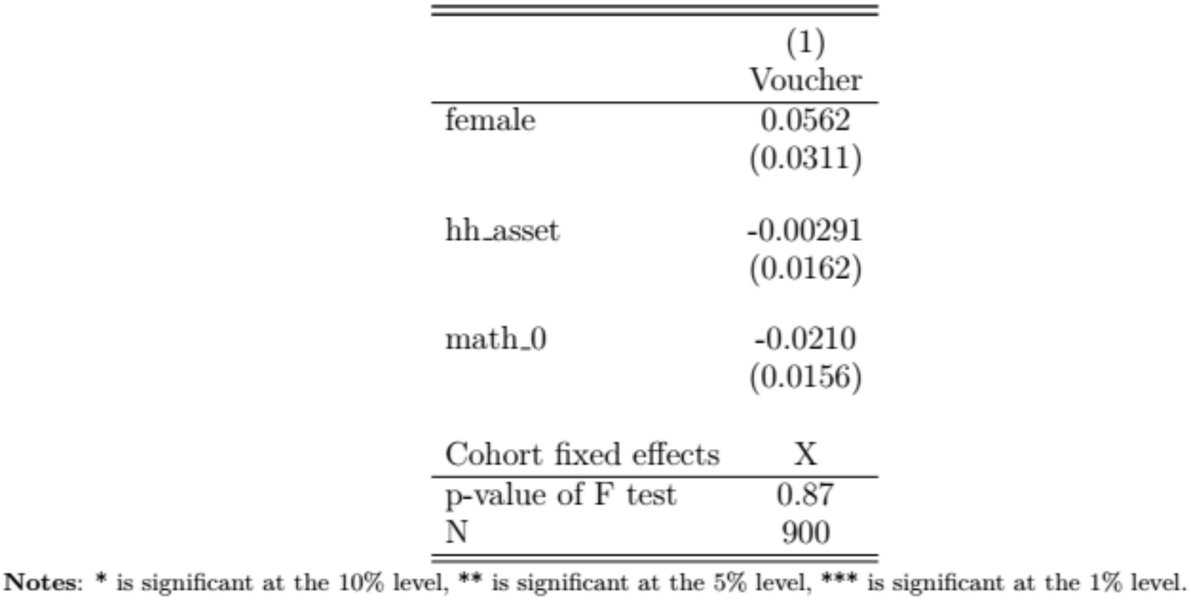
   a. What is the Conditional Independence Assumption generally? What does it assume in this example? Explain in plain English using variable names. **[4 points]**

   The Conditional Independence Assumption assumes that treatment assignment for any one is independent of the potential outcomes, and that treatment assignment is as good as random, given observed variables. In this example, controlling for gender, family income, and baseline scores allows the researchers to isolate the effect of attending a private school on their outcomes (future test scores). This assumption helps control for any potential confounding variables and attribute any observed difference in the final outcome to the treatment itself (winning the lottery).

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

   As done below and in the paper, a good test of CIA is to compare the treatment and control groups across a variety of characteristics (such as gender, household assets, baseline scores). If the randomization is done correctly, or if CIA holds, then these outcomes should have no systematic difference in both groups. In that sense, the balance table is necessary but still not sufficient to check the validity of the CIA assumption. We would also need to explore the actual randomization process itself.

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. **[4 points]**



	(1) Voucher
female	0.0562 (0.0311)
hh_asset	-0.00291 (0.0162)
math_0	-0.0210 (0.0156)
Cohort fixed effects	X
p-value of F test	0.87
N	900

Notes: * is significant at the 10% level, ** is significant at the 5% level, *** is significant at the 1% level.

According to the table, there is evidence that the group receiving vouchers for private school and the group that did not receive vouchers were similar at baseline in terms of gender, household index, and math scores. This conclusion is supported by the high p-value (0.87) from the joint F-test in the table, or the statistically insignificant coefficients.
Since we do not have any statistically significant coefficients, we can say that the randomization was done adequately (provided we have no reasons to question the procedural implementation of the randomization algorithm etc).

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. **[4 points]**

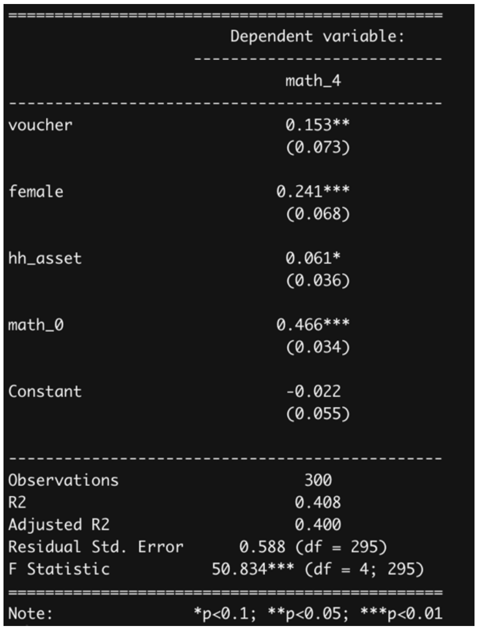
To ensure internal validity, dropout characteristics and reasons must be examined to avoid introducing bias and limiting generalizability. For instance, if there is some underlying reason behind the students dropping out (eg systemic factors such as gender, income etc), then our study's internal validity is compromised and excluding dropouts may introduce selection bias and overestimate private school impact on the outcome of interest.

5. Answer the following questions about the ITT.

a. 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. **[6 points]**

Regression Specification:

$$Math\_{}_{4} = \beta_0 + \beta_1{*}voucher + \beta_2{*}female + \beta_3{*}hh\_asset + \beta_4{*}Math\_{}_{0} + \varepsilon$$



```
=============================================
                    Dependent variable:
                ---------------------------
                          math_4
---------------------------------------------
voucher                  0.153**
                        (0.073)

female                  0.241***
                        (0.068)

hh_asset                 0.061*
                        (0.036)

math_0                  0.466***
                        (0.034)

Constant                 -0.022
                        (0.055)

---------------------------------------------
Observations              300
R2                       0.408
Adjusted R2              0.400
Residual Std. Error  0.588 (df = 295)
F Statistic       50.834*** (df = 4; 295)
=============================================
Note:           *p<0.1; **p<0.05; ***p<0.01
```

Our primary coefficient of interest is β_1 which in this case turns out to be 0.153
This can be interpreted as, "All other things constant, the causal effect of getting the voucher on increase in test scores 4 years later is +0.153 SD ". This is statistically significant at the 5% level.

Controls are important to avoid bias in the treatment effect β_1 due to factors that affect both the treatment and control group, such as gender, household index, and previous math scores. Optional controls like location, parents education, health parameters could be included since they may also be affecting the outcome but we may not have data on those parameters, or may risk overfitting our model since they may not be as directly influential as the main controls.

b. Interpret the coefficient with appropriate units - what does this estimate tell you? **[4 points]**

As mentioned above, this estimate can be interpreted as, "All other things constant, the causal effect of getting the voucher on increase in test scores 4 years later is +0.153 SD ". This is statistically significant at the 5% level. Intuitively, this coefficient is telling us that students who received the private school voucher have significantly higher test scores 4 years later, holding constant all other factors.

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.

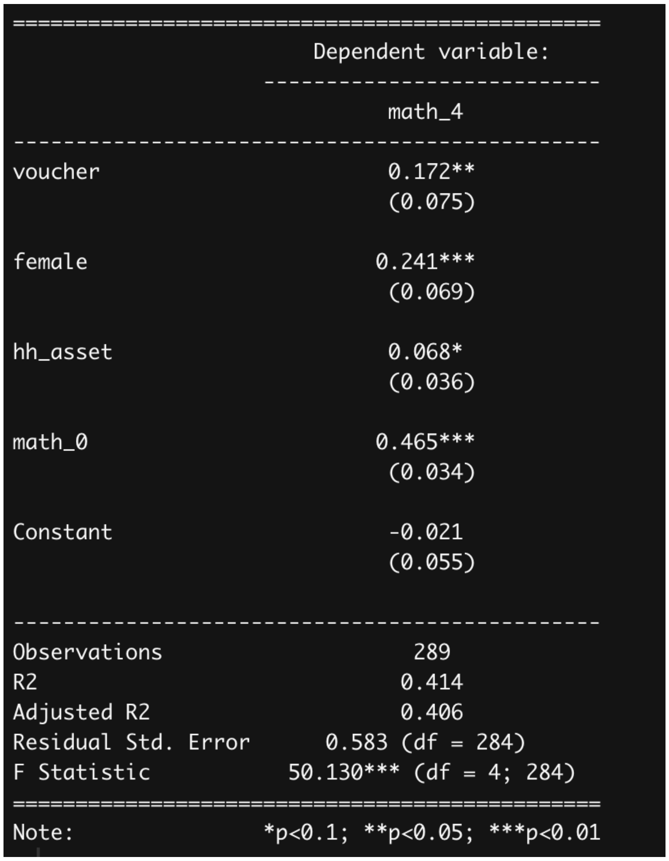
      What could be wrong if we want to interpret this coefficient as a TOT? **[6 points]**

      The comparison of baseline and final test scores for only those who opted for private schools could lead to a biased estimate of TOT due to selection bias. The 0.34 coefficient may not be solely driven by private schools if only certain types of students opt into private schools (eg students of richer parents that value education and support them academically and otherwise).

      Students who remain in private schools may have these other resources that support their attendance, leading to an overestimation of the impact of private schools on test scores.

   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? **[8 points]**

      (Table below)
      The coefficient of interest here is 0.172, which implies that the students who received the voucher and attended the private school had 0.172 SDs higher test scores 4 years later than comparable students who did not. This coefficient is also statistically significant at the 5% level.
      However, this estimate runs into the issues discussed above – which means that it could be an overestimate of the true impact of treatment due to the positive selection bias discussed above.



```
===============================================
                        Dependent variable:
                    ---------------------------
                               math_4
-----------------------------------------------
voucher                        0.172**
                               (0.075)

female                         0.241***
                               (0.069)

hh_asset                       0.068*
                               (0.036)

math_0                         0.465***
                               (0.034)

Constant                       -0.021
                               (0.055)

-----------------------------------------------
Observations                     289
R2                              0.414
Adjusted R2                     0.406
Residual Std. Error        0.583 (df = 284)
F Statistic            50.130*** (df = 4; 284)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

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

The Wald estimator of attending private school is 0.1709 SD. This can be interpreted as the causal impact of winning the lottery and attending private school, all other factors constant, is +0.17 SD.

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, these estimates are different from ITT. To summarise,

Question 5: ITT estimates for students including those who won the lottery vouchers but did not attend private schools

Question 6a: Estimate of attending private schools over time for only those in private schools (includes no comparison with public school trends)

Question 6b: Estimate of attending private schools over time for those who received vouchers and attended (introduced selection bias terms as mentioned above)

Question 6c: Estimate of attending private schools over time for the group that attended private schools after winning vouchers compared to those who did not win the vouchers.

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]**

   My main takeaway from this analysis is that private schools seem to have a positive benefit on mathematics test scores for children over a 4 year time duration, all other factors constant. This could be possibly due to improved resources at private schools, or better teacher quality, or a number of such other factors. However, a number of students who receive the voucher and still cannot attend the school due to other considerations.

Part II [40 points] – Instrumental Variables

The file `api210_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. **[8 points]**

   In this context, an IV approach is necessary over a naïve observational regression since the basic regression does not help us estimate the causal impact of the voucher. Even if the regression controls for observable factors such as gender, baseline score, household income etc. we risk omitted variable vias.
   To understand this, we can consider an example of household education levels. This variable is not captured anywhere in our dataset, and it is easy to understand how this may impact our outcome of interest: Parents that are better educated themselves may value education more,

and invest more resources and time into their childrens learning (for example, helping out with homework etc). As a result, our true estimate of the voucher due to these omitted variables will be lower than the estimate we get through the regression (overestimation).
An instrumental variable approach addresses these concerns effectively.

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]**

   (As discussed in class lecture slides) The four assumptions for an IV with heterogenous treatment effects are:

   1. *Relevance: The instrument must impact treatment status, which means that receiving a voucher for a private school leads to attendance at the nearest private school, while not receiving a voucher result in not attending a private school.*
   2. *Exclusion Restriction: The exclusion restriction states that the instrument only affects the outcome through treatment. Receiving a voucher to attend a private school should not have a direct impact on test scores but should only influence performance by encouraging attendance at the private school.*
   3. *Independence: The independence condition requires that the instrument is randomly assigned. Therefore, vouchers must be distributed randomly to children.*
   4. *Monotonicity: Monotonicity implies that the instrument affects everyone in the same way. Thus, receiving a voucher must have either a weakly positive impact on test scores for everyone or no impact at all, but not a negative impact. In other words, receiving a voucher should lead to better test scores or no change, but not worse test scores.*

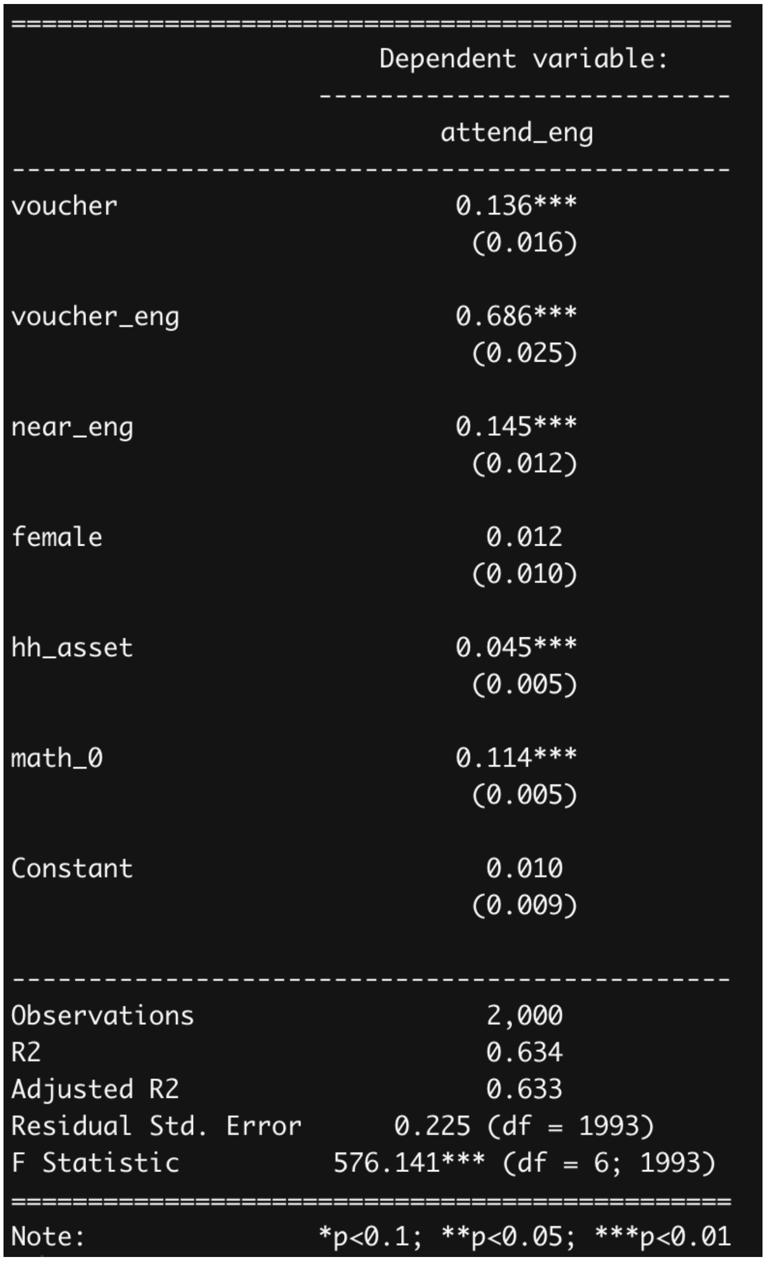
   b. Do you think they are reasonable? Why or why not? How can you check? **[4 points]**

   1. *Relevance: I think this assumption holds true, since the voucher results in attending the nearest private school.*
   2. *Exclusion Restriction: While this seems to hold, there may be some scenarios where this condition breaks – for example, children who win the voucher may have improved beliefs of their own abilities. I would look into this by checking if there is an improvement in scores for children who won the lottery and chose to not attend private schools.*
   3. *Independence: As we can see from the balance tests and related discussions above, independence assumption seems reasonable in this context.*
   4. *Monotonicity: While monotonicity is a fair assumption, I would double check by seeing if all students who won the voucher indeed improve their scores. If the score of a child who attends primary school drops relative to his baseline, there could be factors in the school that actually regress the score (peers, adjustment issues, etc).*

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, $near\_eng_i$ should be included as a control, and all other valid controls should be included) **[8 points]**

$Attend\_eng = \beta_0 + \beta_1{*}voucher\_eng + \beta_2{*}near\_eng + \beta_3{*}female + \beta_4{*}hh\_asset + \beta_5{*}Math\_{}_0 + \varepsilon$

Where voucher_eng is a dummy for whether a student received the voucher AND the nearest private school is in English.

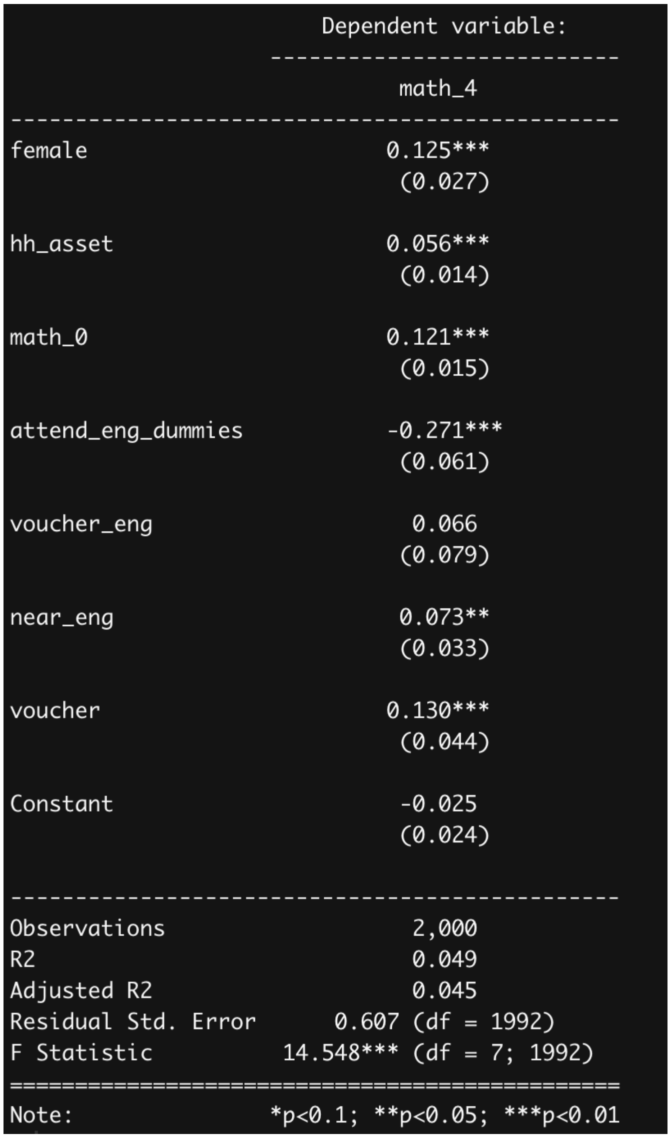


```
==============================================
                    Dependent variable:
                ---------------------------
                        attend_eng
----------------------------------------------
voucher                  0.136***
                          (0.016)

voucher_eng              0.686***
                          (0.025)

near_eng                 0.145***
                          (0.012)

female                     0.012
                          (0.010)

hh_asset                 0.045***
                          (0.005)

math_0                   0.114***
                          (0.005)

Constant                   0.010
                          (0.009)

----------------------------------------------
Observations               2,000
R2                         0.634
Adjusted R2                0.633
Residual Std. Error    0.225 (df = 1993)
F Statistic         576.141*** (df = 6; 1993)
==============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

The instrument voucher_eng is 0.686, which can be interpreted as that the children who receive the lottery vouchers when the nearest private school is in English are 68.6% more likely to attend the private school compared to those who do not. This is also statistically significant at the 1% level.

This confirms our relevance assumption stated above, and is reassuring the validity of our instrument.

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. **[6 points]**



```
                      Dependent variable:
                    ---------------------------
                              math_4
-----------------------------------------------
female                       0.125***
                              (0.027)

hh_asset                     0.056***
                              (0.014)

math_0                       0.121***
                              (0.015)

attend_eng_dummies          -0.271***
                              (0.061)

voucher_eng                   0.066
                              (0.079)

near_eng                      0.073**
                              (0.033)

voucher                      0.130***
                              (0.044)

Constant                     -0.025
                              (0.024)

-----------------------------------------------
Observations                  2,000
R2                            0.049
Adjusted R2                   0.045
Residual Std. Error      0.607 (df = 1992)
F Statistic          14.548*** (df = 7; 1992)
===============================================
Note:              *p<0.1; **p<0.05; ***p<0.01
```

The coefficient of voucher_eng here is a 0.066 SD, which is insignificant statistically. Based on this, I would be inclined to conclude that while receiving the voucher has strong effect on whether a student attends private school or not, there is no influence of attending a school with English as a mode of instruction on test scores four years later.

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]**

I think this estimate is the LATE or local average treatment effect. This is because:

1. *This estimate represents the impact of attending the nearest English or Telegu private schools for the compliers or the children who attended the school only on receiving the vouchers.*
2. *This estimate is not applicable to the defiers or the children who did not attend the English private school despite being given the voucher.*

6. 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]**

   The lottery estimand measures the ITT effect by comparing the outcomes of children assigned to private schools in the lottery with those who were not, without considering the actual school they attend. On the other hand, the IV estimand employs the voucher and proximity to English/Telugu language school as an instrument for the language of instruction, which may not accurately reflect parents' preferences, as they might opt for a English school despite its distance if they believe it is better for some reason.

   Moreover, when assuming variable treatment effects, the IV estimate represents the LATE, applicable solely to the compliers, i.e., those who attended the nearest English/Telugu private school after receiving the voucher, and similarly for Telugu private schools and vouchers, as described above.

   b. Explain and give an example of how a violation of the exclusion restriction for IV may generate this discrepancy. **[2 points]**

   A violation of the exclusion restriction may happen if there are additional paths through which the instrument can affect our treatment. For example children who learn in a private school that teachers in English may perform worse due to not being familiar with the language of instruction as compared to children learning in their native telegu, resulting in overall lower test scores.

   c. Explain and give an example of how a violation of the exogeneity/independence assumption for IV may generate this discrepancy. **[2 points]**

   A violation of the independence assumption would imply that the treatment variable is not randomly assigned across the treatment and control. This could result in systemic differences between the two groups that result in the discrepancy.