Microconometrics I - Problem Set 2

Bruno Ferman
TA: Arthur Botinha
Due: August 27th, 2024

For this problem set, you have a dataset with information on students who participated in the lotteries for the military-run school discussed in Problem set 1 ("base_lotteries.dta"). For each student, you have information on some covariates (collected before the first lottery), their lottery status (whether they entered the lottery, and whether they won a spot), and test scores at the end of 7th, 8th, and 9th grade. We are in a setting with perfect compliance – that is, if a student win the lottery, he/she enrolls in the military-run school. You can find the description of each variable in this data in the end of this problem set.

Question 1: Data cleaning

Every student in the lotteries has a test score between 0 and 10 for Math (`mt`) and Portuguese (`lt`) in the end of the 7th, 8th and 9th grade. Unfortunately, the difficulty of the tests varies considerably between years and so does the score distribution. You should, in each exercise, "standardize" (subtract the mean and divide by the standard deviation) the scores using the statistics from the relevant control group.

(a) (5 points) What are the costs and benefits of standardizing the scores, as opposed to using the raw ones?

(b) (5 points) Why is it reasonable to consider only the control statistics for the standardization, instead of the ones from the whole sample?

Question 2: Balance.

Note: for this question, I want you to present all the results you will discuss in one table (rather than having one separate table for each item).

(a) (5 points) Construct a table that compares the baseline covariates between treated and control students. For each lottery, this table should show the mean and standard error of the control group, the difference between treated and control group (with the standard error for this difference), and the p-value of these differences.

(b) (10 points) Would a few covariates presenting statistically significant differences between treated and controls at 10% be a strong indication that the lottery was not fair? Discuss how you would proceed in this situation, and which additional information you should include in this balance table. Implement that with the dataset.

(c) (5 points) Do the same as in the two previous items, but comparing students who participated in the first lottery (A) with students that participated in the second lottery (B). Do you have evidence that students who participated in the second lottery are similar to those who participated in the first one?

(d) (10 points) Suppose you want to estimate $\tilde{\beta} = \mathbb{E}[Y_i^9(1,1) - Y_i^9(0,0)]$. Consider an estimator that compares the average of the first-lottery winners with the first lottery-losers after excluding those who were admitted in the second lottery (because for those students we would observe $Y_i^9(0,1)$). In light of the results of the previous item, would that be a good idea?

(e) (5 points) Suppose you did not find any significant difference in the covariates considered in item c. In this case, would it be a good idea to consider the comparison described in item d in this case?

Question 3: Treatment effects on math test scores / alternative inference procedures

Again, you should present all results you will discuss in this question in a single table.

(a) (5 points) Construct a table with estimators for $\beta = \mathbb{E}[Y_i^8(1) - Y_i^8(0)]$ and their standard errors for math test scores. Consider two different specifications, one without covariates and another one with baseline covariates. How these two specifications compare?

(b) (10 points) Calculate p-values for both specifications using a permutation test. Consider both $|\hat{\beta}|$ and $|\hat{\beta}|/\hat{\sigma}_{\hat{\beta}}$ as test statistics. How do these permutation test p-values compare to the ones from the precious item? (Note: you cannot use a canned permutation test command from Stata or R).

(c) (5 points) In general terms, discuss the advantage of using $|\hat{\beta}|/\hat{\sigma}_{\hat{\beta}}$ relative to $|\hat{\beta}|$.

(d) (10 points) Calculate p-values for both using a wild bootstrap with null imposed. How do these permutation test p-values compare to the ones from the precious item? (Note: you cannot use a canned wild bootstrap command from Stata or R).

(e) (10 points) Based on the questions from the previous problem set, propose an estimator for $\tilde{\beta} = \mathbb{E}[Y_i^9(1,1) - Y_i^9(0,0)]$ for math test scores. Present in the table the estimate and standard error. Explain which procedure you used to compute the standard errors in this case.

Question 4: Multiple outcomes

(a) (10 points) Now we want you to estimate the treatment effects $\beta = \mathbb{E}[Y_i^8(1) - Y_i^8(0)]$ considering both math and portuguese test scores. Explain the different possible ways to take into account that you have more than one outcome.

(b) (10 points) Implement at least two alternatives using the data.

Question 5: Attrition

Suppose now at the end of grade 8 students can take an official exam that would count for university admission once they finish high school. This exam is voluntary, so we should expect students that are more interested in going to the university would be more likely to take it. Information from this exam (which includes students who did not participate in the lottery) is avaiable in the dataset "voluntary_exam.dta".

(a) (10 points) Estimate the causal effect of enrolling in the military-run school on the probability of taking this exam.

(b) (10 points) Run a balance table as in Question 1, but conditional on non-attriters.

(c) (10 points) Suppose now you want to estimate the effect of enrolling in the military-run school on the test score in this exam. Explain why the results from the previous items may be problematic in this case. In particular, discuss why a comparison between lottery winners and losers might not be valid, and discuss the direction in which you expect this comparison would be biased.

(d) (10 points) Discuss different alternatives you could use in this setting, being careful about the assumptions you would need in each of these alternatives.

(e) (10 points) Implement one alternative solutions using the available data (it is ok to use a canned command in this case).

Question 6: External Validity

(a) (10 points) In question 2, you (probably) considered a robust standard error (without a cluster structure), given that we had a student-level randomization. Discuss what are the implications in terms of the target parameter we are implicitly considering in this case. In particular, what would happen if there are some unobserved shocks that affect the military-run school (for example, a large construction nearby)?

(b) (10 points) Suppose now there are J military-run schools, all of them with lotteries. Discuss the differences between considering robust standard errors and standard errors clustered at the school level in this case.

(c) (5 points) For the previous item, what would happen if J is very small?

Dictionary

Dictionary for "base_lotteries.dta":

- **student_code**: Student unique id.
- **sex**: Sex
  - 1 Male
  - 2 Female
- **race**: Race
  - 1 White
  - 2 Brown
  - 3 Black
  - 4 Yellow
  - 5 Indigenous
- **schooling_mother**: Highest Schooling Degree of Mother/Responsible
  - 1 <5th grade
  - 2 5th-9th grade
  - 3 9th grade-HS
  - 4 HS-College
  - 5 College
  - 6 >College
  - 7 Unknown
- **schooling_father**: Highest Schooling Degree of Father/Responsible
  - 1 <5th grade
  - 2 5th-9th grade
  - 3 9th grade-HS
  - 4 HS-College
  - 5 College
  - 6 >College
  - 7 Unknown
- **ever_failed**: Ever Failed a School Year?
  - 1 Never
  - 2 Yes, Once
  - 3 Yes, More than Once
- **score_h_g**: Scores of subject h (Math, mt or Portuguese, lp) in g-th Grade (7, 8 or 9)
- **lottery_A**: Applied to the 8th Grade Lottery (A)
- **lottery_B**: Applied to the 9th Grade Lottery (B)
- **won_lottery_A**: Won 8th Grade Lottery (A)
- **won_lottery_B**: Won 9th Grade Lottery (B)