SPI 200, Spring 2025: Final Exam

Professor Rocio Titiunik

Instructions

Read each of the following instructions carefully before starting the exam.

- You have **3 hours** to complete this exam. Failure to submit answers within the time allocated to the exam will result in a grade of zero points.
- Your final answers should contain code, output, and written answers combined in a final report produced with R Markdown, and printed to a PDF or HTML file. No other file formats are permitted.
- When you are done, submit a single file to Canvas.
- You are not allowed to communicate with any human being during the exam, you must solve the exam on your own.
- You are not allowed to use generative language chatbots (Gemini, chatGPT, etc)
- You are not allowed to use the internet, except to search for programming tips if you get stuck. If you search the internet, you must include in your submission a disclosure of how many times you used the internet and what you searched for.
- The exam is open book. You are allowed to use all the materials from class: all problem sets with their respective solutions, all precept exercises, all materials on Canvas. No need to disclose use of any of these.
- Separate each question and subquestion using headers. Failure to do so will result in a 1 point penalty.
- This exam has a total of 6 questions. The last question is a bonus question, meaning that any points you get from answering this question are added to the numerator of your percentage grade, but not to the denominator.
- Each question is worth as many points as the number of sub-questions it contains. For example, Question 1 is worth 2 points, because it has two sub-questions.
- The denominator of your percentage grade is 15 points, corresponding to the total points in questions 1 through 5 (excluding the bonus question).

Introduction

In this exercise, we analyze the impact of placing students on academic probation on their future academic performance. The exercise is based on the following article:

Lindo, Jason M., Nicholas J. Sanders, and Philip Oreopoulos, 2010. "Ability, gender, and performance standards: Evidence from academic probation."*American economic journal: Applied Economics* 2(2): 95-117.

The authors study a policy at a Canadian university that places students on academic probation when their grade point average (GPA) falls below a threshold. The probation treatment enforces a standard for the student's future academic performance: a student who is placed on probation in a given term must improve her GPA in the next term according to campus-specific standards, or face suspension. Thus, in this study, the unit of analysis is the student, the treatment or intervention is placing the student on probation. The outcome we analyze is the GPA obtained by the student in the term immediately after he was placed on probation (`nextGPA`).

Students come from three different campuses. In campuses 1 and 2, the cutoff for placing a student on probation is 1.5; in campus 3 the cutoff is 1.6. **You must use this information to correctly answer some of the questions below.**

Name	Description
`nextGPA`	student GPA in term immediately after he/she was placed on probation
`priorGPA`	student GPA in term immediately before he/she was placed on probation
`left_school`	`1` if student dropped out of school after probation decision, `0` otherwise
`hsgrade_pct`	the student's average GPA in standard high school classes
`age_at_entry`	the student's age when they started college
`male`	`1` if student is male, `0` otherwise
`english`	`1` if student speaks English at home, `0` otherwise
`loc_campus1`	`1` if student is enrolled in campus 1, `0` otherwise
`loc_campus2`	`1` if student is enrolled in campus 2, `0` otherwise
`loc_campus3`	`1` if student is enrolled in campus 3, `0` otherwise

Question 1 (2 points)

1. Load the data into an object named `data`. How many observations are there?
2. For every variable in the dataset, report the number of missing observations.

Question 2 (2 points)

1. Add to the dataset a new variable called `treatment`, equal to 1 if the student was placed on probation, and 0 otherwise.
2. How many treated and control students are there?

Question 3 (3 points)

For this question, consider two covariates: `english` and `hsgrade_pct`.

1. Calculate the difference-in-means between treated and control students for each of the two covariates.

2. Manually construct a 95 percent confidence interval for these two differences, and compare it to the one produced by the `t.test()` function.
3. Are these covariates balanced between treated and control students? Provide a brief explanation based on the previous calculations.

Question 4 (4 points)

1. Create a new dataset keeping only students whose GPA before the probation treatment (`priorGPA`) is 0.2 points above or below the probation cutoff. That is, only keep students who are within a 0.2 point window of the probation cutoff. Keep in mind that the cutoffs are campus-specific. Name this new dataset `subset`.
2. Report the number of treated and control students in this new dataset.
3. Repeat the analysis carried out in Question 3, now using the newly created `subset` dataset. That is, compute the difference in means for the `english` and `hsgrade_pct` covariates across treated and control students, and manually compute a 95 percent confidence interval for these two differences.
4. Are treated and control students in the subset more similar in terms of these two covariates? Briefly explain if that is the case, and provide an explanation as to *why*.

Question 5 (4 points)

1. Suppose that you are given the choice to estimate the average treatment effect of the probation treatment on future GPA (`nextGPA`) using either the full data, or the subset of the data that you created in Question 4. You would like to interpret the estimated effect as the **causal** average effect of the probation treatment on future GPA. Which one of the two datasets do you choose to estimate the treatment effect? Why? Briefly justify your choice.
2. Using your chosen dataset (i.e. either `data` or `subset`), estimate the average effect of the treatment, and provide a 95 percent confidence interval for it. You may use any method you like to compute the effect and the confidence interval.
3. Interpret the average treatment effect computed above, as well as the confidence interval.
4. Regarding the average treatment effect, are you comfortable giving a causal interpretation for it? Would you be more comfortable if the probation treatment had been randomized? Why/why not?

BONUS Question 6 (3 points)

The GPA in the term after probation (`nextGPA`) is only observed for students who decide to continue at the university after they are placed on probation. The variable `left_school` contains an indicator equal to 1 if the student dropped out after being placed on probation.

1. Use a linear regression to predict `left_school` using the `treatment` variable. Use the `lm()` function to fit the model, and use `summary()` to print out the results.
2. Interpret the estimated coefficient of the `treatment` variable. Based on the p-value, is it significant at the 5% level?
3. What does this result imply about the validity of the conclusions you found in Question 5?