Final exam

This homework must be turned in on Brightspace by December 13th 2022, 11:59pm (Paris time). Your homework submission must be written and submitted using Jupyter Notebook. No handwritten solutions will be accepted. You should submit:

- A compiled PDF file named `yourNetID_solutions.pdf` containing your solutions to the problems.
- A `.ipynb` file containing the code and text used to produce your compiled pdf named `yourNetID_solutions.ipynb`. Note that math can be typeset in markdown cells (in the jupyter notebook) in the same way as Latex.

Late finals will not be accepted, so start early and plan to finish early. Remember that exams often take longer to finish than you might expect.

Please make sure your answers are clearly structured in the jupyter notebook file:

- Label each question (e.g. 1.1 for exercise 1, question 1).
- Do not include written answers as code comments.
- The code used to obtain the answer for each question part should accompany the written answer.

You may use your notes, books, and internet resources to answer the questions below. However, you are to work on the exam by yourself. You are prohibited from corresponding with any human being regarding the exam (unless following the procedures below). Professor Judith Abécassis will answer clarifying questions during the exam. She will not answer statistical or computational questions until after the exam is over. If you have a question, ask by email. If your question is a clarifying one, she will reply. Please do not post on Brightspace.

Exercise 1 Instrumental variables (30 points)

Patients who get surgery, for example for orthopaedic reasons, are often advised by the doctors, subsequently to surgery, to get physiotherapy, that is, a series of exercises to help rehabilitation and more complete recovery. However, the costs of physiotherapy may often deter patients from following it. It is therefore important to try to show the potential benefits of physiotherapy, so that more patients can become convinced to follow it.

In the period of 4 years, three cooperating hospitals randomly assigned each of the 537 eligible patients, who had gone through an orthopeadic operation, in one of two groups : patients in the first group, ($Z_i = 1$), were offered the opportunity to get physiotherapy at 50% reduced hospital fees; for patients assigned in the second group, physiotherapy was available at the standard cost. For each patient, the recorded variables, in addition to assignment Z_i, are: whether or not the patient got physiotherapy, $T_i^{obs} = 1$ for yes, 0 for no; an assessment of the patients recovery 3 months after surgery, $Y_i^{obs} = 1$ for satisfactory, 0 for unsatisfactory or poor. The assessment of this studys data was done by physicians blinded to both the assignment Z_i and the taking (or not) of physiotherapy by the patient. The table below gives the counts, n_{zty}, of patients assigned $Z_i = z$ and with physiotherapy-taking status $T_i^{obs} = t$ and outcome $Y_i^{obs} = y$.

	Z	T^{obs}	Y^{obs}	n
0	0	0	0	185
1	0	0	1	123
2	0	1	0	9
3	0	1	1	41
4	1	0	0	37
5	1	0	1	20
6	1	1	0	26
7	1	1	1	96

Question 1 (5 points)

Estimate the intention-to-treat (ITT) effect of offering the discount on the improvement of recovery, $E[Y(Z = 1)] - E[Y(Z = 0)]$, using a difference-in-means estimator. Also estimate the standard error and the asymptotic 95% confidence interval. Explain why, the ITT effect can be different from the contrast that compares outcomes Y^{obs} of the patients who take vs. do not take physiotherapy.

Be aware that the input data is aggregated, so you should either used weighted estimators (for the mean and standard error). You can use Python code for the computations, and in that case manually create the input dataframe from the given table.

Question 2 (4 points)

In plain language of this setting, and using the potential treatment notation, what are the four possible strata defined by the instrument and the treatment values?

Question 3 (6 points)

In plain language of this setting, and in terms of potential outcomes, state the four assumptions under which the randomizer Z_i is an "instrument", and the local ATE is non-parametrically identified. Discuss their plausibility.

Question 4 (5 points)

Which of the assumptions from the question 3 is/are enough to estimate the proportion of "never-takers", i.e. patients who would not take physiotherapy whether or not they had been offered the discount in this study? Under this/these assumption(s), report estimates of the proportions of the groups defined in question 2.

Question 5 (5 points)

Under assumptions from question 3, estimate the local ATE. In which group defined in question 2 is this treatment effect estimated? You can use the python function `IV2SLS` to provide the standard error and a 95% confidence interval for your estimate.

Question 6 (5 points)

Discuss briefly (i) the clinical and (ii) the health policy implications of the difference between your estimates in question 5 vs. question 1.

Exercise 2 Regression Discontinuity Design (30 points)

In this problem you will be analyzing a dataset from a 2011 paper by Carpenter and Dobkin. The full citation for the paper is:

> Carpenter, C., & Dobkin, C. (2011). The minimum legal drinking age and public health. Journal of Economic Perspectives, 25(2), 133-56.

This paper examines evidence linking the legal alcohol drinking age in the US (21) to increased likelihood of accidents, hospitalization, and health hazards in general. The main identification strategy employed by the authors is a sharp Regression Discontinuity Design (RDD), where age is the running variable, and 21 is the cutoff.

The dataset contains 80 observations, where each unit is an age group, and values are collected over 4 US states.

The dataset is `ER.csv` and it contains five variables:

- age – The age of the unit, where the decimal indicates month of the year
- all – The total number of ER admissions
- injury – The total number of ER admissions due to injury
- illness – The total number of ER admissions due to viral illness
- alcohol – An adjusted index of how many ER admissions were linked to alcohol consumption

Question 1 (5 points)

Preprocess the data by creating a centralized version of the running variable, and a binary variable indicating the treatment.

Question 2 (10 points)

Estimate the effect of being legally able to purchase alcohol (age ≥ 21) on the `all`, `injury`, and `alcohol` variables using an RDD with bandwidth = 1. For each of the three outcomes report point estimates and 95% confidence intervals. Repeat the analysis for bandwidth = 0.5 years, and bandwidth = 2 years. Discuss and interpret your results. Which outcome variable seems to be associated with the largest effect? Does bandwidth selection influence results? You can use the model `wls` from package `statsmodels.formula.api`, as in the recitation.

Question 3 (5 points)

Using the entire dataset (no need to aggregate), create and show RDD plots that visualize the discontinuity for each of the three outcome variables used in Question 1. The plots should display observed points, regression lines, and vertical lines to indicate the bandwidth.

Question 4 (10 points)

Conduct a placebo RDD analysis using the `illness` variable as outcome: since viral illnesses are not caused by alcohol consumption, we have no reason to expect that being legally able to drink will have an effect on this variable. Report both RDD estimates and 95% CIs, and make a RDD plot for this outcome variable. Is there a treatment effect and is it statistically significant? What does this suggest about the plausibility of the RDD assumptions?