DS-UA 201
Causal Inference
Instructor: Parijat Dube

Homework 2-100 points

Due: June 29, 2024

General Instructions

This homework must be turned in on Gradescope by June 29th, 2024, 11:59pm. It must be your own work, and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone. Your homework submission must be written and submitted using Rmarkdown. No handwritten solutions will be accepted. You should submit:

1. A compiled PDF file named yourNetID solutions.pdf containing your solutions to the problems.
2. A .Rmd file containing the code and text used to produce your compiled pdf named yourNetID solutions.Rmd. Note that math can be typeset in Rmarkdown in the same way as Latex.

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

1. Label each question part(e.g. 3.a).
2. Do not include written answers as code comments.
3. The code used to obtain the answer for each question part should accompany the written answer.

Problem 1 - *Changing Minds on Gay Marriage?* 30 points

In this exercise, we analyze the data from two experiments in which households were canvassed for support on gay marriage. Note that the original study was later retracted due to allegations of fabricated data. In this exercise, however, we analyze the original data while ignoring the allegations.

Canvassers were given a script leading to conversations that averaged about twenty minutes. A distinctive feature of this study is that gay and straight canvassers were randomly assigned to households, and canvassers revealed whether they were straight or gay in the course of the conversation. The experiment aims to test the "contact hypothesis," which contends that out-group hostility (towards gay people in this case) diminishes when people from different groups interact with one another. The data file is gay.csv, which is a CSV file. The below table presents the names and descriptions of the variables in this data set.

Name	Description
`study`	Study (1 or 2)
`treatment`	Treatment assignment: `No contact`, `Same-Sex Marriage Script by Gay Canvasser`, `Same-Sex Marriage Script by Straight Canvasser`, `Recycling Script by Gay Canvasser`, and `Recycling Script by Straight Canvasser`
`wave`	Survey wave (1-7). Note that Study 2 lacks wave 5 and 6.
`ssm`	Support for gay marriage (1 to 5). Higher scores indicate more support.

Homework 2-100 points

Each observation of this data set is a respondent giving a response to a four-point survey item on same-sex marriage. There are two different studies in this data set, involving interviews during seven different time periods (i.e., seven waves). In both studies, the first wave consists of the interview before the canvassing treatment occurs.

1. **(4 points)** Using the baseline interview wave before the treatment is administered, examine whether randomization was properly conducted. Base your analysis on the three groups of study 1: "same-sex marriage script by gay canvasser," "same-sex marriage script by straight canvasser" and "no contact." Briefly comment on the results.

2. **(4 points)** The second wave of the survey was implemented two months after canvassing. Using study 1, estimate the average treatment effects of gay and straight canvassers on support for same-sex marriage, separately. Give a brief interpretation of the results.

3. **(5 points)** The study contained another treatment that involves contact, but does not involve using the gay marriage script. Specifically, the authors used a script to encourage people to recycle. What is the purpose of this treatment? Using study 1 and wave 2, compare outcomes from the treatment "same-sex marriage script by gay canvasser" to "recycling script by gay canvasser." Repeat the same for straight canvassers, comparing the treatment "same-sex marriage script by straight canvasser" to "recycling script by straight canvasser." What do these comparisons reveal? Give a substantive interpretation of the results.

4. **(5 points)** In study 1, the authors reinterviewed the respondents six different times (in waves 2 to 7) after treatment, at two-month intervals. The last interview, in wave 7, occurs one year after treatment. Do the effects of canvassing last? If so, under what conditions? Answer these questions by separately computing the average effects of straight and gay canvassers with the same-sex marriage script for each of the subsequent waves (relative to the control condition).

5. **(4 points)** The researchers conducted a second study to replicate the core results of the first study. In this study, same-sex marriage scripts are given only by gay canvassers. For study 2, use the treatments "same-sex marriage script by gay canvasser" and "no contact" to examine whether randomization was appropriately conducted. Use the baseline support from wave 1 for this analysis.

6. **(3 points)** For study 2, estimate the treatment effects of gay canvassing using data from wave 2. Are the results consistent with those of study 1?

7. **(5 points)** Using study 2, estimate the average effect of gay canvassing at each subsequent wave and observe how it changes over time. Note that study 2 did not have a fifth or sixth wave, but the seventh wave occurred one year after treatment, as in study 1. Draw an overall conclusion from both study 1 and study 2.

Problem 2 - *Election Fraud in Russia* 20 points

In this exercise, we use the rules of probability to detect election fraud by examining voting patterns in the 2011 Russian State Duma election. The State Duma is the federal legislature of Russia. The

Homework 2-100 points

ruling political party, United Russia, won this election, but to many accusations of election fraud, which the Kremlin, or Russian government, denied. Some protesters highlighted irregular patterns of voting as evidence of election fraud. In particular, the protesters pointed out the relatively high frequency of common fractions such as $1/4, 1/3$, and $1/2$ in the official vote shares.

Variable	*Description*
N	total number of voters in a precinct
turnout	total turnout in a precinct
votes	total number of votes for the winner in a precinct

Note: The results of each election are stored in a data frame. The RData file fraud. RData contains data on four elections: the 2007 and 2011 Russian Duma elections, the 2012 Russian presidential election, and the 2011 Canadian election.

We analyze the official election results, contained in the russia2011 data frame in the RData file fraud.RData, to investigate whether there is any evidence for election fraud. The RData file can be loaded using the load( ) function. Besides russia2011 , the RData file contains the election results from the 2003 Russian Duma election, the 2012 Russian presidential election, and the 2011 Canadian election, as separate data frames. The table above presents the names and descriptions of variables used in each data frame. Note: Part of this exercise may require computationally intensive code.

1. **(5 points)** To analyze the 2011 Russian election results, first compute United Russia's vote share as a proportion of the voters who turned out. Identify the 10 most frequently occurring fractions for the vote share. Create a histogram that sets the number of bins to the number of unique fractions, with one bar created for each uniquely observed fraction, to differentiate between similar fractions like $1/2$ and $51/100$. This can be done by using the breaks argument in the hist () function. What does this histogram look like at fractions with low numerators and denominators such as $1/2$ and $2/3$?

2. **(10 points)** The mere existence of high frequencies at low fractions may not imply election fraud. Indeed, more numbers are divisible by smaller integers like 2,3 , and 4 than by larger integers like 22,23 , and 24 . To investigate the possibility that the low fractions arose by chance, assume the following probability model. The turnout for a precinct has a binomial distribution, whose size equals the number of voters and success probability equals the turnout rate for the precinct. The vote share for United Russia in this precinct is assumed to follow a binomial distribution, conditional on the turnout, where the size equals the number of voters who turned out and the success probability equals the observed vote share in the precinct. Conduct a Monte Carlo simulation under this alternative assumption (1000 simulations should be sufficient). What are the 10 most frequent vote share values? Create a histogram similar to the one in the previous question. Briefly comment on the results you obtain. Note: This question requires a computationally intensive code. Write a code with a small number of

simulations first and then run the final code with 1000 simulations.

3. **(5 points)** To judge the Monte Carlo simulation results against the actual results of the 2011 Russian election, we compare the observed fraction of observations within a bin of certain size with its simulated counterpart. To do this, create histograms showing the distribution of question 2's four most frequently occurring fractions, i.e., $1/2, 1/3, 3/5$, and $2/3$, and compare them with the corresponding fractions' proportion in the actual election. Briefly interpret the results.

Problem 3 - 35 points

Gerber, Green and Larimer randomly assigned households to receive a mailing encouraging them to turn out to vote before the Michigan 2006 primary election (Gerber, Green and Larimer (APSR, 2008)). We will be using the individual data obtained from the experiment. Each row in the dataset represent an individual record, where p2000 represents whether the individual had voted in August 2000, g2000 represents whether the individual had voted in November 2000 (same for p2002,g2002,p2004). Each individual belongs to a household specified by hh_id.

Part a. (4 points) Data preparation : In order to analyze the GOTV data we will need to reproduce the household-level dataset of the original paper

(a) Recode the variable "sex" by changing the character to float (i.e. "female" → 1., "male" → 0).

(b) Recode the variable "yob" into a new variable called "age" by subtracting yob from the year the experiment took place, 2006.

(c) Group the data into households, i.e., create a new dataframe where each row is a household with a unique hh id, and each column is the the mean value of each of the other individual-level variables in that household. (Hint: you may consider using dplyr.)

(d) In the paper, the authors analyzed households rather than individual. Why did they do this?

Part b. (4 points) Validate Randomization :

Use the household dataset you obtained above, show that the experimental assignment is randomized at the household level by computing and showing the sample means of each of the variables: *p2000, g2000, p2002, g2002, p2004, hhsize, sex, and age* in each of the treatment groups. Are these means similar across groups? And if so what does that imply for randomization and ignorability?

Part c. (4 points) ATE :

Use the household dataset you obtained above, use the Neyman Estimator, denoted here as $\hat{\tau}$, to compute the average treatment effect for each treatment group comparing to the control group. Name and briefly explain two assumptions in this experiment that allow us to compute the ATE.

Homework 2-100 points

Part d. (10 points) Variance and Average HP testing :

Assuming that the experiment is a completely randomized experiment, give an estimate of the ATE variance of the treatment effect of the Neighbors treatment compared to the control group, using the Neyman variance estimator, denoted as $\hat{Var}[\tau]$. In addition, conduct a two-sided hypothesis test against the null that the ATE is 0, i.e.: $H_0 : \tau = 0$, with the alternative is $H_1 : \tau \neq 0$, using the Z-statistic as your test statistic, i.e.:

$$Z_n = \frac{\sqrt{n}(\hat{\tau} - \tau)}{\sqrt{\hat{Var}[\tau]}}$$

Report both the value of Z_n and the p-value for the test

Part e. (10 points) Randomization Inference :

Conduct a randomization inference hypothesis test on the experiment data for the sharp null hypothesis that $Y_i(neighbors) = Y_i(control)$ for all i. Using Z_n as defined before as your test statistic, follow the steps below:

(a) Simulate the value of Z_n under the sharp null for at least N = 1000 iterations.
(b) Plot the values you obtained as a histogram.
(c) Add a marker for the observed value of Z_n
(d) Report the two-sided p-value for the test

Part f. (3 points) Compare hypothesis tests:

Briefly comment on the difference between the p-value you obtained in parts d and e. Which is smaller? And what could this difference be due to?

Problem 4 - 15 points

Consider a study with N units. Each unit i in the sample belongs to one of G mutually exclusive strata. $G_i = g$ denotes that the ith unit belongs to stratum g. N_g denotes the size of stratum g and $N_{t,g}$ denotes the number of treated units in that stratum. Suppose that treatment is assigned via block-randomization. Within each stratum, $N_{t,g}$ units are randomly selected to receive treatment and the remainder receive control. Suppose that the proportion of treated units in each stratum, $\frac{N_{t,g}}{N_g}$ **is not the same** for all strata. After treatment is assigned, you record an outcome Y_i for each unit in the sample. Assume consistency holds with respect to the potential outcomes:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

Part a (4 points)

Show that the ATE: $\tau = E[Y_i(i) - Y_i(0)]$ is identified in this setting, i.e., show that τ equal to a function of the observed outcomes.

Homework 2-100 points

Part b (5 points)

Assume that $E[\hat{\tau}(g)|G_i = g, N_g = n_g] = \tau(g)$ and that $E[\frac{N_g}{N}] = Pr(G_i = g)$. Show that the stratified estimator:

$$\hat{\tau} = \sum_{g=1}^{G} \hat{\tau}(g)\frac{N_g}{N}$$

is unbiased for the ATE, i.e., show that $\mathrm{E}[\hat{\tau}] = \tau$:

Part c (6 points)

Instead of using the stratified difference-in-means estimator, your colleague suggests an alternative that assigns a weight to each unit and takes two weighted averages. Let $w(G_i) = Pr(D_i = 1|G_i)$ denote the known (constant) probability that unit i would receive treatment given its stratum membership G_i. The new estimator is:

$$\hat{\tau}_w = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{w(G_i)} - \frac{(1-D_i)Y_i}{1-w(G_i)}\right)$$

Assuming that $E[\frac{N_g}{N}] = Pr(G_i = g)$, show that $\hat{\tau}_w$ is unbiased i.e., show that $\mathrm{E}[\hat{\tau}_w] = \tau$.

Note: either showing that $\hat{\tau}_w$ is unbiased for $\tau = E[Y_i(1)-Y_i(0)]$ or for $\tau = \frac{1}{N}\sum_{i=1}^{N} E[Y_i(1)-Y_i(0)]$ will count as a valid answer.