DS-UA 201
Causal Inference
Instructor: Parijat Dube

Homework 3-100 points

Due: July 13, 2024

General Instructions

This homework must be turned in on Gradescope by July 13, 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 - *CATE using GOTV* 20 points

Consider again the GOTV data from last problem set by Gerber, Green and Larimer (APSR, 2008). Although it is not specified in the paper, it is highly possible that the authors created subgroups based on the turnout history for 5 previous primary and general elections (number of times the individual voted), and number of registered voters in the household. In this problem, we will create subgroups based on the turnout history, and investigate the CATE(conditional average treatment effect) and the effect modifications in each subgroup. We denote the turnout history/number of times voted as a covariate X_i for individual i.

Part a. Data preparation (5 points):

Construct a new dataset for this problem using individual dataset from the last problem set.

1. Create a new column num_voted to represent the number of times the individual has voted in previous 5 elections by summing the variables g2000, p2000, g2002, p2002 and p2004 (exclude g2004 because the experiment filtered out people who didn't vote in g2004), the resulting column should be an integer ranging from [0,5]
2. In the following problems, we are using the individual data with num_voted as different subgroups. To simplify the problem, we investigate only the "Neighbor" treatment effect. Construct a cleaner dataset with {id, hh_id, hh_size, num_voted, voted, treatment} as columns and filter out treatment groups besides {Neighbor, Control}.

Homework 3-100 points

3. Construct a household-level dataset by taking the means of hh_size, num_voted, and voted in each household (the other variables are all equal within the same household and can simply be left as they are). Round the mean of num_voted **up** to the nearest integer. Your resulting dataset should have one household per row, and hh_id, hh_size, num_voted, voted, and treatment as columns. The variable num_voted should have only values 0, 1, 2, 3, 4, 5.
4. Report number of households in each subgroup for both treatment and control, what do you observe?

Part b. CATE for subgroups (6 points)

We define conditional average treatment effect as the ATE for different subgroups defined by the "num_voted" variable:

$$\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x], x \in \{0, 1, 2, 3, 4, 5\}$$

Since treatment was randomized at the household level, positivity and ignorability hold both unconditionally, and conditionally, within each subgroup. For each subgroup:

1. Estimate the CATE and report the variance of your estimates.
2. Construct a 95% confidence interval around your estimates.
3. What conclusions can you draw from these statistics?

You can skip subgroups that either do not have members in them or do not have any treated/control members.

Part c. Effect modification (6 points)

Suppose we want to estimate whether there is a difference in effects for two extreme groups, individuals who always vote($X_i = 5$) and individuals who never vote($X_i = 0$), we construct an estimator $\hat{\Delta}$to estimate the difference. As we saw in class, we can estimate this difference as:

$$\hat{\Delta} = \hat{\tau}(0) - \hat{\tau}(5)$$

1. Calculate the variance of $\hat{\delta}$ and construct a 95% confidence interval around it, can we say that there's significant difference in the treatment effect for people who always vote and people who never vote?
2. Combine your observations with conclusions from part b, comment about your findings.

Part d. Sample sizes and significance effect (3 points)

In the experiment, the authors claimed no significant differences between groups, one possible reason may be that the sample size for each subgroup is too small. This is a practical problem we may encounter in experimental designs when we are testing multiple hypothesis or we are having too many subgroups. Explain in your own words why having more hypothesis/subgroups would make significant effect harder to detect for each group, assuming the overall sample size is fixed.

Homework 3-100 points

Problem 2 - *Stratification using GOTV* 25 points

In this question we will be using the same household-level dataset that you constructed in part a of Problem 1.

Part a (5 points):

Compute the ATE of the "Neighbors" treatment using the standard difference-in-means estimator, i.e., $\hat{\tau} = \bar{Y}_t - \bar{Y}_c$. Provide standard errors and 95% confidence intervals for your estimates.

Part b (10 points):

Now compute the same ATE but with the stratification estimator that is defined as the weighted mean of the stratum CATEs that you computed in the previous problem:

$$\hat{\tau}_{\text{block}} = \sum_{x=0}^{5} \hat{\tau}(x) \frac{N_x}{N}.$$

Compute variance and 95% confidence intervals for this estimator as well using the stratified variance estimator defined as:

$$\widehat{\text{Var}}(\tau_{\text{block}}) = \sum_{x=0}^{5} \widehat{\text{Var}}(\tau(x)) \left(\frac{N_x}{N}\right)^2$$

Comment on the difference between the ATE estimates you obtained here and in part a and their variances. What is it due to?

Part c (10 points):

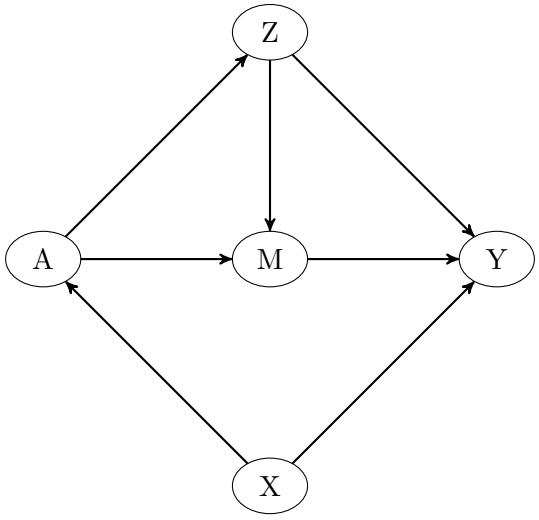
Now Divide the data set into 6 strata in such a way that each of the strata have same proportion of Treated and Control observations. You can do so by creating a new variable called "group" with values $0, 1, 2, 3, 4, 5$ and randomly assigning each value to $N_t/6$ treated units and $N_c/6$ control units. You may exclude enough treated and control units from the data to make N_t and N_c divisible by 6.

Compute the ATE by applying the estimator $\hat{\tau}_{block}$ to these newly created strata. Provide variance estimates and 95% confidence intervals for these ATE estimates as well using the stratified variance estimator. Is the variance of this estimator much different from that of $\hat{\tau}$ you computed in part A? Why do you think this is the case?

Problem 3 - *Directed Acyclic Graphs (DAGs)* 15 points

Consider the following Directed Acyclic Graph:

Homework 3-100 points



Part a (5 points)

Of the five variables in the graph, 2 are colliders and 3 are non colliders. Which variables are colliders and which are non-colliders?

Part b (5 points)

Suppose that we wanted to estimate the effect of A on Y . Indicate if we should or should not condition on X, and explain why, and indicate if we should or should not condition on Z and explain why.

Part c (5 points)

Suppose that we wanted to estimate the effect of M on Y . List all the backdoor paths between M and Y, and indicate which variable we should condition on to close each path. There may be multiple valid options for each path.

Problem 4 - *TRCs and Racial Attitudes* 25 points

In new democracies and post-conflict settings, Truth and Reconciliation Commissions (TRCs) are often tasked with investigating and reporting about wrongdoing in previous governments. Depending on the con- text, institutions such as TRCs are expected to reduce hostilities (e.g. racial hostilities) and promote peace.

In 1995, South Africa's new government formed a national TRC in the aftermath of apartheid. [Gibson 2004] uses survey data collected from 2000-2001 to examine whether this TRC promoted inter-racial reconciliation. The outcome of interest is respondent racial attitudes (as measured by the level of agreement with the prompt: "I find it difficult to understand the customs and ways

of [the opposite racial group]".) The treatment is "exposure to the TRC" as measured by the individual's level of self-reported knowledge about the TRC.

You will need to use the trc_data.dta file for this question. The relevant variables are

- RUSTAND - Outcome: respondent's racial attitudes (higher values indicate greater agreement)
- TRCKNOW - Treatment dummy (1 = if knows about the TRC, 0 = otherwise)
- age - Respondent age (in 2001)
- female - Respondent gender
- wealth - Measure of wealth constructed based on asset ownership (assets are fridge, floor polisher, vacuum cleaner, microwave oven, hi-fi, washing machine, telephone, TV, car)
- religiosity - Self-reported religiosity (7 point scale)
- ethsalience - Self-reported ethnic identification (4 point scale)
- rcblack - Respondent is black
- rcwhite - Respondent is white
- rccol - Respondent is coloured (distinct multiracial ethnic group)
- EDUC - Level of education (9 point scale)

Part a (4 points)

Estimate the average treatment effect of TRC exposure on respondents' racial attitudes under the assumption that TRC exposure is ignorable. Report a 95% confidence interval for your estimate and interpret your results.

Part b (5 points)

Examine whether exposed and nonexposed respondents differ on the full set of observed covariates using a series of balance tests. Briefly discuss, in which ways do exposed and nonexposed respondents differ?

Part c (8 points)

Now assume that TRC exposure is conditionally ignorable given the set of observed covariates:

1. Use an additive logistic regression model to estimate the propensity score for each observation.
2. With this model, construct inverse propensity of treatment weights (IPTW) for each observation.

3. Use the propensity score to construct an IPW estimator and report the point estimate for the ATE.

4. Plot the histograms of the propensity scores in treatment and control.

Part d (8 points)

Using a pairs bootstrap (resampling individual rows of the data with replacement), obtain estimate for the standard error of your IPTW estimator for the ATE. Compute a 95% confidence interval and interpret your findings. (you should report **estimate, Std, 95% CI lower, 95% CI upper**, for interpretation, compare your results in Part C/D to your estimate from Part A and briefly discuss your findings.)

Problem 5 - *Stratified Study* 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.

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 $\mathrm{E}[\hat{\tau}] = \tau$:

Homework 3-100 points

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.