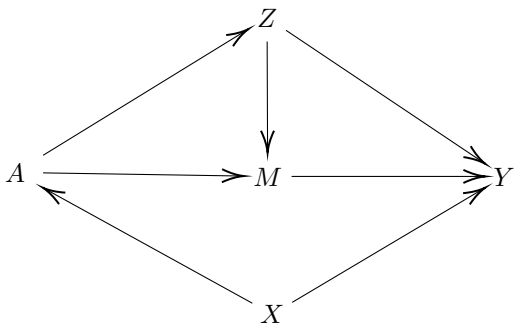
Problem Set 2

Erin Choi - eyc321 - Section 2

Due Oct 26, 2022

Problem 1 (25 Points Total)

Consider the following Directed Acyclic Graph:



Part A (15 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? Explain why?

The colliders are M and Y. A and Z are both causes of M, so their arrows collide into M, making it a collider. Similarly, Z, M, and X are all causes of Y, and their arrows also collide into Y, so Y is a collider.

The non-colliders are A, Z, and X. A and Z each have only one direct cause (i.e. one arrow pointing to them) - A is directly caused only by X, and Z is directly caused only by A. X has no arrows pointing to it, so no other variables cause X in this DAG. No arrows collide into A, Z, or X, so they are all 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 why. Also, indicate if we should or should not condition on Z and explain why or why not.

In estimating the effect of A (treatment) on Y (outcome), X is a confounder that causes both A and Y. We do want to control for the fork at X to close the backdoor path from A to Y (A <- X -> Y).

We should not control for Z because it is a post-treatment covariate that is a mediator in the effect of A on Y. Controlling on Z would block the causal paths from A to Y that travel through Z, which we do not want to do since we need to keep all causal paths open to estimate the effect.

Part C (5 points)

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

1. Y <- Z -> M

   Condition on Z, as it is a confounder.

2. Y <- Z <- A -> M

   Condition on A, as it is a confounder (there is a fork). Controlling for Z would also block this path.

3. Y <- X -> A -> M

   Condition on X, as it is a confounder (there is a fork). We can also control for A to block the path.

4. Y <- X -> A -> Z -> M

   Condition on X, as it is a confounder. Controlling for Z or A would also block this path.

Problem 2 (75 Points Total)

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 (20 Points Total)

Construct a new dataset for this problem using the individual level dataset provided below.

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]. (5 points)

```
# import libraries
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
# read in excel data
GOTV <- read_excel("gotv_individual.xlsx")

# create num_voted column = sum of g2000, p2000, g2002, p2002 and p2004
GOTV$num_voted <- GOTV$g2000 + GOTV$p2000 + GOTV$g2002 + GOTV$p2002 + GOTV$p2004
# unique(GOTV$num_voted) yields 3 0 1 2 5 4 (all [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 $Neighbor, Control$. (5 points)

```r
# subset only needed columns
GOTV_clean <- subset(GOTV, select=c("hh_id", "hh_size", "num_voted",
                                    "voted", "treatment"))

# filter by treatment groups
# unique(GOTV$treatment) yields "Civic Duty" "Hawthorne" "Control" "Self" "Neighbors"
GOTV_clean <- filter(GOTV_clean, treatment == "Control" | treatment=="Neighbors")
# unique(GOTV_clean$treatment) yields "Control" "Neighbors"
```

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. (5 points)

```r
# take means of all numeric columns, grouping by hh_id
GOTV_hh <- GOTV_clean %>%
  group_by(hh_id, treatment) %>%
  summarise(hh_size = mean(hh_size),
    num_voted = mean(num_voted),
    voted = mean(voted))
```

```
## `summarise()` has grouped output by 'hh_id'. You can override using the
## `.groups` argument.
```

```r
# round num_voted up to nearest int
GOTV_hh$num_voted <- ceiling(GOTV_hh$num_voted)
# unique(GOTV_hh$num_voted) yields 2 3 5 4 1 0
```

4. Report number of households in each subgroup for both treatment and control, what do you observe? (5 points)

```r
# total number of households
cat(paste("Control:", nrow(GOTV_hh[GOTV_hh$treatment == 'Control', ]),
          "\nTreatment:", nrow(GOTV_hh[GOTV_hh$treatment == 'Neighbors', ])))
```

```
## Control: 99999
## Treatment: 20000
```

```r
# number of households in each subgroup (0 to 5 previous votes)
for (i in 0:5) {
  cat(paste("\n",i,"Votes:"))
  cat(paste("\nControl:", nrow(GOTV_hh[GOTV_hh$treatment == 'Control'
                                       & GOTV_hh$num_voted==i, ])))
  cat(paste("\nTreatment:", nrow(GOTV_hh[GOTV_hh$treatment == 'Neighbors'
                                        & GOTV_hh$num_voted==i, ])))
}
```

```
##
##  0 Votes:
## Control: 74
## Treatment: 14
##  1 Votes:
## Control: 3238
## Treatment: 646
##  2 Votes:
## Control: 25397
## Treatment: 5126
##  3 Votes:
## Control: 45511
## Treatment: 9078
##  4 Votes:
## Control: 22901
## Treatment: 4521
##  5 Votes:
## Control: 2878
## Treatment: 615
```

There are 99999 households in the control group versus 20000 in the Neighbors treatment group. The control group is much larger than (almost 5 times the size of) the treatment group. This difference in numbers is also evident in each subgroup (based on number of previous votes, 0 to 5). The number of control households in each subgroup is much greater than treatment households. These differences can lead to potential bias in estimating the treatment effect.

Part B. CATE for subgroups (25 points total)

We define conditional average treatment effects 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. (5 points)

```r
for (i in 0:5) {
  cat(paste("\n",i,"Votes:"))
  nam <- paste("CATE", i, sep = "")
  assign(nam, mean(GOTV_hh$voted[GOTV_hh$treatment == 'Neighbors' & GOTV_hh$num_voted==i])
         - mean(GOTV_hh$voted[GOTV_hh$treatment == 'Control' & GOTV_hh$num_voted==i]))
  cat(paste("\nCATE:", mean(GOTV_hh$voted[GOTV_hh$treatment == 'Neighbors'
```

```
                                          & GOTV_hh$num_voted==i])
            - mean(GOTV_hh$voted[GOTV_hh$treatment == 'Control'
                                          & GOTV_hh$num_voted==i])))

  nam <- paste("var", i, sep = "")
  assign(nam, var(GOTV_hh$voted[GOTV_hh$treatment=="Neighbors" & GOTV_hh$num_voted==i])
         /sum(GOTV_hh$treatment=="Neighbors" & GOTV_hh$num_voted==i)
        + var(GOTV_hh$voted[GOTV_hh$treatment=="Control" & GOTV_hh$num_voted==i])
        /sum(GOTV_hh$treatment=="Control" & GOTV_hh$num_voted==i))

  cat(paste("\nVariance:", var(GOTV_hh$voted[GOTV_hh$treatment=="Neighbors"
                                          & GOTV_hh$num_voted==i])
            /sum(GOTV_hh$treatment=="Neighbors" & GOTV_hh$num_voted==i)
            + var(GOTV_hh$voted[GOTV_hh$treatment=="Control" & GOTV_hh$num_voted==i])
            /sum(GOTV_hh$treatment=="Control" & GOTV_hh$num_voted==i)))
}
```

```
##
##  0 Votes:
## CATE: 0.103281853281853
## Variance: 0.017602911851735
##  1 Votes:
## CATE: 0.0837668463568907
## Variance: 0.000296983031699667
##  2 Votes:
## CATE: 0.0640186990759893
## Variance: 3.4999680017872e-05
##  3 Votes:
## CATE: 0.0908311168158508
## Variance: 2.57776065616704e-05
##  4 Votes:
## CATE: 0.101828776204867
## Variance: 5.53403844751974e-05
##  5 Votes:
## CATE: 0.0459571536994036
## Variance: 0.000387211209651595
```

2. Construct a 95% confidence interval around your estimates. (5 points)

```
# 0 votes
se0 = sqrt(var0)
ci0 = c(CATE0-qnorm(.975)*se0, CATE0+qnorm(.975)*se0)
paste("CI for 0 votes: [", ci0[1], ",", ci0[2], "]")
```

```
## [1] "CI for 0 votes: [ -0.156758260936605 , 0.363321967500311 ]"
```

```
# 1 vote
se1 = sqrt(var1)
ci1 = c(CATE1-qnorm(.975)*se1, CATE1+qnorm(.975)*se1)
paste("CI for 1 vote: [", ci1[1], ",", ci1[2], "]")
```

```
## [1] "CI for 1 vote: [ 0.0499904035791325 , 0.117543289134649 ]"
```

```r
# 2 votes
se2 = sqrt(var2)
ci2 = c(CATE2-qnorm(.975)*se2, CATE2+qnorm(.975)*se2)
paste("CI for 2 votes: [", ci2[1], ",", ci2[2], "]")
```

```
## [1] "CI for 2 votes: [ 0.05242344877571 , 0.0756139493762686 ]"
```

```r
# 3 votes
se3 = sqrt(var3)
ci3 = c(CATE3-qnorm(.975)*se3, CATE3+qnorm(.975)*se3)
paste("CI for 3 votes: [", ci3[1], ",", ci3[2], "]")
```

```
## [1] "CI for 3 votes: [ 0.0808800558622282 , 0.100782177769473 ]"
```

```r
# 4 votes
se4 = sqrt(var4)
ci4 = c(CATE4-qnorm(.975)*se4, CATE4+qnorm(.975)*se4)
paste("CI for 4 votes: [", ci4[1], ",", ci4[2], "]")
```

```
## [1] "CI for 4 votes: [ 0.0872483849864242 , 0.11640916742331 ]"
```

```r
# 5 votes
se5 = sqrt(var5)
ci5 = c(CATE5-qnorm(.975)*se5, CATE5+qnorm(.975)*se5)
paste("CI for 5 votes: [", ci5[1], ",", ci5[2], "]")
```

```
## [1] "CI for 5 votes: [ 0.00738960365033044 , 0.0845247037484767 ]"
```

3. What conclusion can you draw from these statistics? (15 points)

The confidence interval for the subgroup with 0 previous votes ([-0.1567583, 0.3633220]) contains 0, so the treatment has no statistically significant effect on households with no previous voting history. In contrast, the confidence intervals for all other subgroups (with 1-5 votes in their previous voting history) do not contain 0; the entire interval for each of these subgroups is above 0. Therefore, for all 5 subgroups of households that have a voting turnout history, we reject the null hypothesis that the treatment has no effect (there is no difference in effect between treatment and control groups) at alpha=0.05. We can conclude that there is convincing evidence that the treatment did have an effect on households with prior turnout history but did not on households with no turnout history. The 95% confidence intervals for subgroups of households with turnout history indicate that, if this experiment were repeated many times and many confidence intervals were constructed, about 95% of the confidence intervals would contain the true value of the CATE.

Part C. Effect Modification (15 points total)

Suppose we want to estimate whether these 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. We can estimate this difference as:

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

Calculate the variance of $\hat{\Delta}$ and construct a 95% confidence interval around it. Can we say that there's a significant difference in the treatment effect for people who always vote and people who never vote? (15 points)

```
diff_ate <- CATE0 - CATE5
paste("Difference in CATEs:", diff_ate)
```

```
## [1] "Difference in CATEs: 0.0573246995824497"
```

```
se_diff <- sqrt(se0^2 + se5^2)
cat(paste("Variance of the difference:", se0^2 + se5^2,
          "\nStandard error:", se_diff))
```

```
## Variance of the difference: 0.0179901230613866
## Standard error: 0.134127264422214
```

```
ci_diff <- c(diff_ate - qnorm(.975)*se_diff, diff_ate + qnorm(.975)*se_diff)
paste("95% confidence interval: [", ci_diff[1], ",", ci_diff[2], "]")
```

```
## [1] "95% confidence interval: [ -0.205559908029971 , 0.320209307194871 ]"
```

The 95% confidence interval for the difference in CATEs contains 0, so at alpha=0.05, we fail to reject the null hypothesis that there is no difference in the treatment effect for households that never vote and those that always vote. Thus there is no convincing evidence of a significant difference between the treatment effect for households that never vote versus always vote.

Part D (15 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.

With a fixed overall sample size, having more and more subgroups would mean the size of each subgroup can only get smaller and smaller. A smaller sample leads to a larger variance/standard error, as calculating standard error involves division by the sample size. This then leads to the calculation of a larger and less precise confidence interval around the CATE/point estimate for each subgroup, making it easier for the confidence interval to contain 0 and harder for us to reject the null hypothesis of no difference, and it becomes more difficult to detect a significant effect for each group. Additionally, the point estimate itself may be inaccurate due to random noise. Smaller samples are more likely to be impacted by random noise since one observation in a small sample holds more weight than in a larger sample and can easily skew the distribution of results, especially if it is a more extreme outcome. A potentially inaccurate point estimate along with an imprecise confidence interval constructed around it makes it harder to detect a significant effect and to most accurately estimate the true effect.