Midterm

Andrew Shao - as13381 - 005

Due Oct 25, 2023

This midterm must be turned in on Brightspace by Oct 25, 2023. 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 not** consult with others. All work must be independent.

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
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. Comment your code!

Problem 1 (25 points)

A cafe is testing out a promotion set to determine which pastry goes well with their new espresso blend. Customers are told that the promotion set is \$5 for a cup of espresso and a random pastry item. After receiving the promotional set, they are asked to rate the product. There are two types of pastries: a sweet scone and a savory bagel, customers are randomly assigned to receive either type. Let $D_i = 1$ if the customer receives the bagel (the "treatment") and $D_i = 0$ if they receive the scone. Let Y_i denote the observed rating from the ith customer.

Part a (12 points)

In your own words, explain what the following quantities represent in this setting and indicate whether this quantity is observable without making assumptions: (4 points each)

1. $Y_i(1)$ This is the outcome, or observed rating, for individual i if they were assigned to the treatment which is that the individual receives the bagel. This quantity is directly observable if the individual is indeed assigned to treatment. If not, it isn't observable.
2. $E(Y_i(1)|D_i = 0)$ This is the expected (average) outcome, or rating in this setting, for treated individuals in the world where the steps to physically assign these individuals to the treatment (bagel as opposed to scone) have not occurred. This isn't observable as these are counterfactuals.
3. $E(Y_i|D_i = 0)$ This is the average rating, or expected outcome, for individuals if they receive the scone, or assigned to the control group. This is not directly observable either unless you were to survey the entire population and assign all of them to the control, which is almost certainly not the case in this scenario. This value can be estimated, however, assuming the observed sample is representative of the population.

Part b (4 points)

Suppose we have 6 customers who bought the set this morning, the observed randomization and potential outcomes are:

Customer	D_i	$Y_i(1)$	$Y_i(0)$
1	1	5	5
2	1	9	5
3	0	8	6
4	0	4	1
5	1	8	5
6	0	7	5

Write down the individual treatment effects (ITE) and observed outcome for each customer.

Customer	D_i	$Y_i(1)$	$Y_i(0)$	ITE	Observed Outcome
1	1	5	5	0	5
2	1	9	5	4	9
3	0	8	6	2	6
4	0	4	1	3	1
5	1	8	5	3	8
6	0	7	5	2	5

ITE $= Y_i(1) - Y_i(0)$ \ Observed outcome is equal to $Y_i(1)$ if $D_i = 1$, else $Y_i(0)$.

Part c (4 points)

Estimate the difference in means (treatment - control) in this case using the table in part b, assuming consistency holds. Is this quantity equal to a causal effect in this case? Why or why not?

Difference in means estimate:

$$\frac{5+9+8}{3} - \frac{6+1+5}{3} = \frac{22}{3} - \frac{12}{3} = \frac{10}{3}$$

ATE:

$$\frac{0+4+2+3+3+2}{6} = \frac{7}{3}$$

Part d (5 points)

The cafe hired a new barista who is very considerate. She asks each customer whether they prefer sweet or savory things, and then gives them their preferred pastry item with their espresso. Is it possible to estimate the average treatment effect of getting the bagel on ratings with data collected after this new barista was hired? Why or why not?

No, there is now both a positivity violation as well as a ignorability violation. The positivity violation occurs because the assignment to treatment and control is not random as individuals are assigned based on their taste preferences. Thus only people who prefer savory things have a positive (which in this case is equal to 1) probability of being assigned the treatment of a bagel. Those who prefer sweet things all have probability of zero, which is therefore a positivity violation since we are trying to infer on people of all tastes, not just people with savory preference. The ignorablity violation occurs because there is now a confounding factor of individuals' taste preferences. As mentioned previously, treatment assignment is affected by this covariate, and the outcome is also naturally affected by this variable as individuals who prefer sweet things will obviously rate higher because they received a scone which they probably like and the individuals who prefer savory things will do the same for the bagel. In effect, outcomes on both treatment and control groups will likely be overestimated.

Problem 2 (25 points)

The STAR (Student–Teacher Achievement Ratio) Project is a four-year longitudinal study examining the effect of class size in early grade levels on educational performance and personal development (whether they finish high school). A longitudinal study is one in which the same participants are followed over time. This particular study lasted from 1985 to 1989 and involved 11,601 students. During the four years of the study, students were randomly assigned to small classes, regular-sized classes, or regular-sized classes with an aid. In all, the experiment cost around $12 million. Even though the program stopped in 1989 after the first kindergarten class in the program finished third grade, the collection of various measurements (e.g., performance on tests in eighth grade, overall high-school GPA) continued through to the end of participants' high-school attendance.

The variables of interest are:

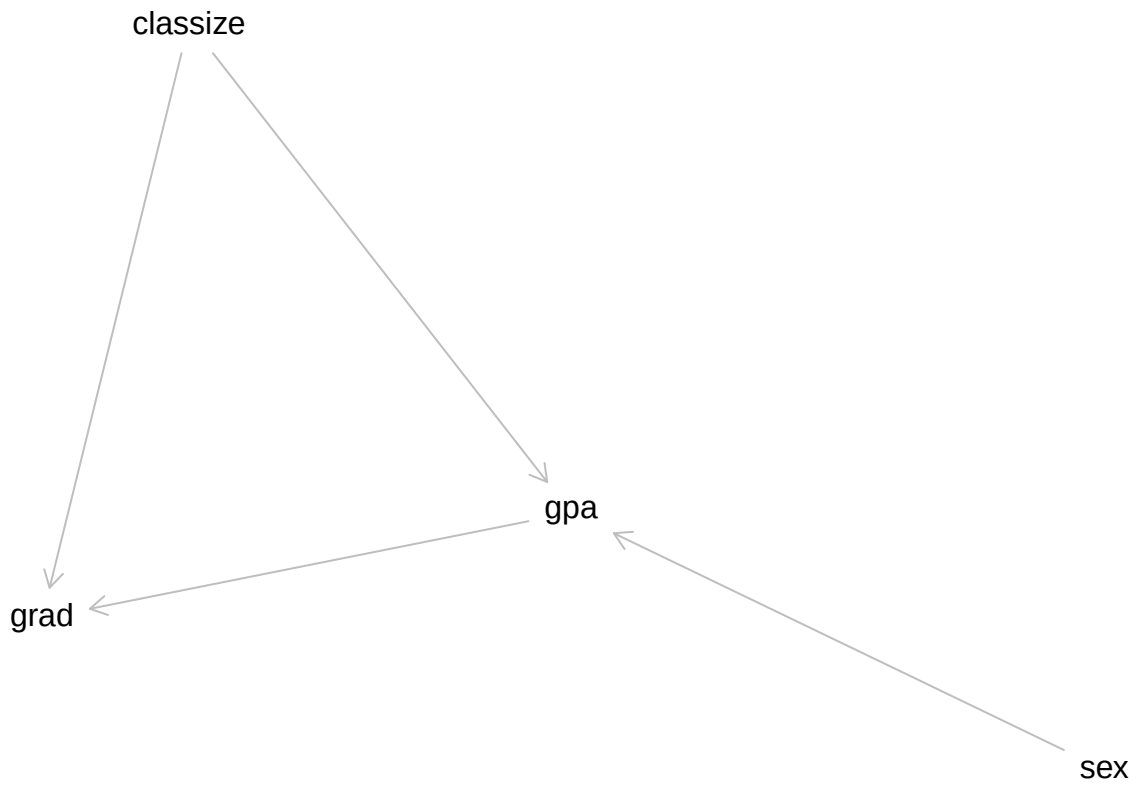
1. classsize - Treatment variable - size of class before the fourth grade.
2. sex
3. race
4. g4math - total scaled score for the math portion of the fourth-grade standardized test
5. g4reading total scaled score for the reading portion of the fourth-grade - standardized test
6. gpa - high school gpa
7. grad - finish high school, 1 yes, 0 no

Part a (8 points)

Consider the variables $sex, classize, gpa$, and $grad$ Draw a DAG representing the causal relationship between them in this experiment.

```
library(dagitty)
# ?dagitty(): See the help file
g <- dagitty('dag {
    classize -> gpa
    classize -> grad
    gpa -> grad
    sex -> gpa
}')
plot(g)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```



Part b (10 points)

Suppose in the experiment, the researcher found out the CATE for female students is different from CATE for male students. We want to know whether these two CATEs are statistically different from each other. Can we conclude anything about this from the fact that one of them is statistically different from zero and the other is not? Why or why not?

We can't conclude that these two CATEs are statistically different from each other. We can only conclude that the treatment has a statistically significant effect in the subgroup that has a CATE different from zero. It could be the case that one CATE is statistically significantly different from zero and the other isn't but the two could be close in value, as their confidence intervals could overlap significantly however only one excludes zero and not by much. Generally to conclude whether two CATEs are significantly different you must perform a statistical test or construct confidence intervals.

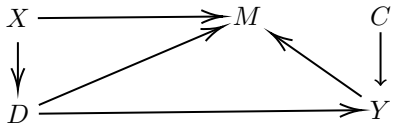
Part c (7 points)

Imagine we wanted to estimate the effect of class size on finishing high school in this experiment. What would be necessary for you to control to estimate an unbiased treatment effect? How would you estimate the treatment effect? Explain your answer.

There are no confounders in this experiment, so there is nothing we need to control for. Since the class size was presumably assigned randomly, all of these variables besides race and sex should presumably be post-treatment and race and sex shouldn't affect treatment.

Problem 3 (25 points)

Consider the following Directed Acyclic Graph:



Part a (15 points)

List all of the paths from D to Y. On each path, identify confounders and colliders.

There is one path, which goes directly from D to Y. There is 1 collider, M, as there are arrows from both the treatment and outcome pointing to M. There are no confounders as no arrow points to both D and Y.

Part b (10 points)

Are there any variables that we should condition on in order to identify the causal effect of D on Y? Explain.

There are no variables that we need to condition on to identify the causal effect. We explicitly cannot condition on M, as it's a collider. If we condition on M it will open a backdoor path and cause a spurious correlation between D and Y which would induce bias. Conditioning on C isn't necessary, however doing so will reduce the standard error of our causal effect estimate.

4 Design Your Study (25 points)

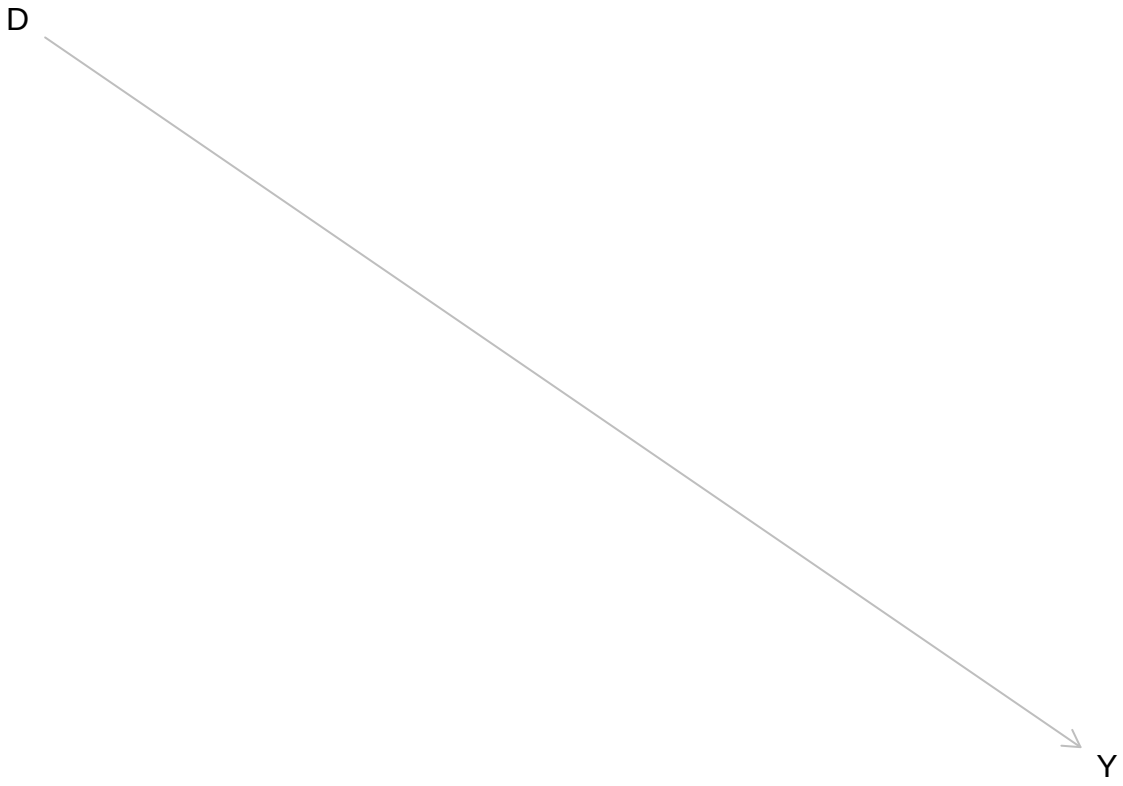
Design your own study from start to finish. Choose an *interesting* question that we have not mentioned in class. Answer the following questions: (1) Explain the effect you wish to estimate in words and why you think it's interesting. Carefully explain both your treatment, outcome, and the research question you wish to answer. (2) What is the "ideal experiment" for your question? (3) Draw the ideal experiment in a DAG. Can you estimate the effect of your treatment on your outcome? Is it identifiable and how do we know? (4) If you were to collect observational data on this topic, what potential confounders and mediators would exist? Please explain them in words. (5) Draw out a DAG that corresponds to this observational study. Please include at least one confounder and one mediator. (6) Using the DAG you drew in question 5, can you estimate the impact of your treatment on your outcome? Is the effect identifiable? Explain why or why not.

*Note: You cannot reuse an example we went over in class nor an example you used in a previous problem set.

Question: Do gun control laws affect gun violence incidence rate? This is interesting because gun violence is a big issue in the US (and other places as well) and gun control has long been a commonly cited but highly controversial method to stop, or at least reduce gun violence in the nation. My treatment would be requiring annual gun license recertification (demonstrate proper usage and pay a fee). The outcome would be the rate of gun violence (non-accidental shooting incidents per capita). The ideal experiment would be to randomly select from a group of cities of identical, characteristics and demographics to impose these measures (the treatment group) and have no such measure for the rest of the cities (control group). All the cities will share the exact same laws and protocols besides the treatment.

```
library(dagitty)
# ?dagitty(): See the help file
g <- dagitty('dag {
    D -> Y
}')
plot(g)
```

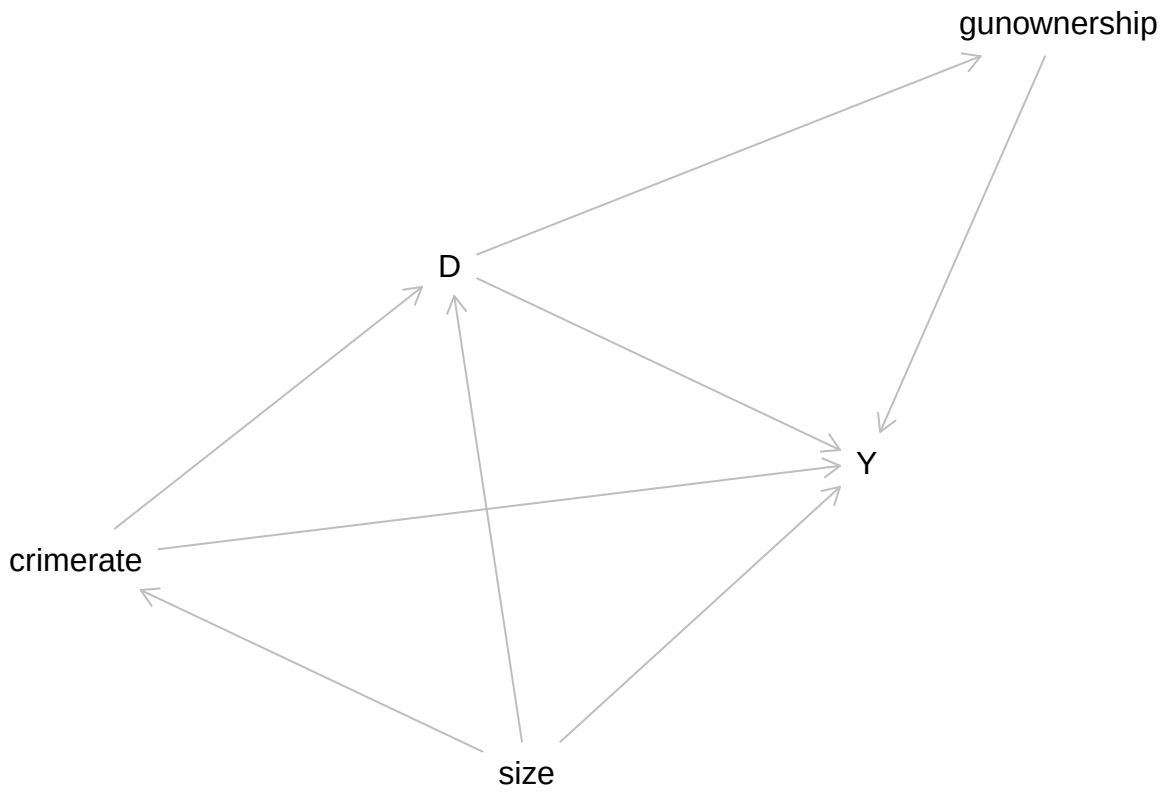
```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```



In the above DAG D is the treatment and Y the outcome. There are no other covariates as ideally we will have controlled for all of them and as a result we can estimate the causal effect of the treatment. If we were to conduct an observational study, we would have many other variables that would both be confounders and mediators. Potential confounders include overall crime rate, population, and resident income distribution. A potential mediator could be gun ownership rate (guns per capita). Cities of similar certain crime rates, population size, and wealth often have similar gun control laws. Gun ownership can be affected by gun control rates (making it harder to buy guns) and is cited to affect gun violence rate.

```
library(dagitty)
# ?dagitty(): See the help file
g <- dagitty('dag {
    D -> Y
    D -> gunownership
    gunownership -> Y
    crimerate -> D
    crimerate -> Y
    size -> D
    size -> Y
    size -> crimerate
}')
plot(g)
```

```
## Plot coordinates for graph not supplied! Generating coordinates, see ?coordinates for how to set you
```



We probably can't estimate the effect of the outcome because it is very difficult to control for all of these variables, as many of these variables are difficult to condition on due to sparsity. It may be impossible to stratify on all of these variables as there be a lack of cities which fall under certain categories or values.