Midterm

Yiyun (Leo) Yao - yy3959 - (Recitation) 002

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 potential outcome given the individuals received the treatment. In this specific case, it is the potential outcome given the customers receive the bagel. So yes, this quantity is observable without making assumptions.

2. $E(Y_i(1)|D_i = 0)$

This is the expected outcome for the ith individual if the individuals did not receive the treatment. In this specific case, it is the expected outcome for the ith customer's rating of the bagel given the customers received the scone. So no, this quantity is not observable without making assumptions because we cannot observe the customer rate the bagel if they receive the scone. We can only observe the outcome for the treatment the customer actually received, in this case, the treatment is the scone.

3. $E(Y_i|D_i = 0)$

This is the expected outcome in the condition of the individual being in the control group and not receiving the treatment. In this specific case, it is the expected rating of the bagel given the customers received the bagel. And yes, this quantity is observable without making assumptions.

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_i	$Observed_i$
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

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?

The difference in means (treatment - control) in this case is $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$. Assuming consistency holds, $E(Y_i|D_i = 1) = E(Y_i(1)|D_i = 1)$ and $E(Y_i|D_i = 0) = E(Y_i(0)|D_i = 0)$. Therefore,

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \frac{1}{n_t} \sum_{i:D_i=1} Y_i - \frac{1}{n_c} \sum_{i:D_i=0} Y_i = \frac{5+9+8}{3} - \frac{6+1+5}{3} = \frac{10}{3}$$

But the causal effect is

$$E[Y_i(1)] - E[Y_i(0)] = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - \frac{1}{n} \sum_{i=1}^{n} Y_i(0) = \frac{5+9+8+4+8+7}{6} - \frac{5+5+6+1+5+5}{6} = \frac{7}{3}$$

So the difference in means $\frac{10}{3}$ is not equal to the causal effect $\frac{7}{3}$ in this case. It could be that any of the three assumptions (SUTVA, ignorability, positivity) are not met. For the difference in means we are missing the counterfactual data, the treated value in a world where treatment is not assigned to the unit, compare to the causal effect.

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, it is not possible to estimate the average treatment effect in this case because the new barista is so considerate that causes violation of the assumptions to be necessarily satisfied to identify the average treatment effect. The act of asking customers which flavor do they prefer does not satisfy ignorability, making the treatment assignment dependent on the potential outcome, as customers will give higher ratings if they are asked to choose what they prefer because they simply like that food flavor they prefer more no matter if it is a bagel or a scone (in this case, it's the bagel). The positivity assumption is potentially not satisfied either. Customers who love sweet things and hate savory things would have a zero chance of receiving savory things because the new barista will only give them what they like. These customers will not have a positive probability of receiving the food with the flavor they hate. Same thing for customers who love savory things and hate sweet things. The new barista will assign the treatment to these customers as they wish.

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.

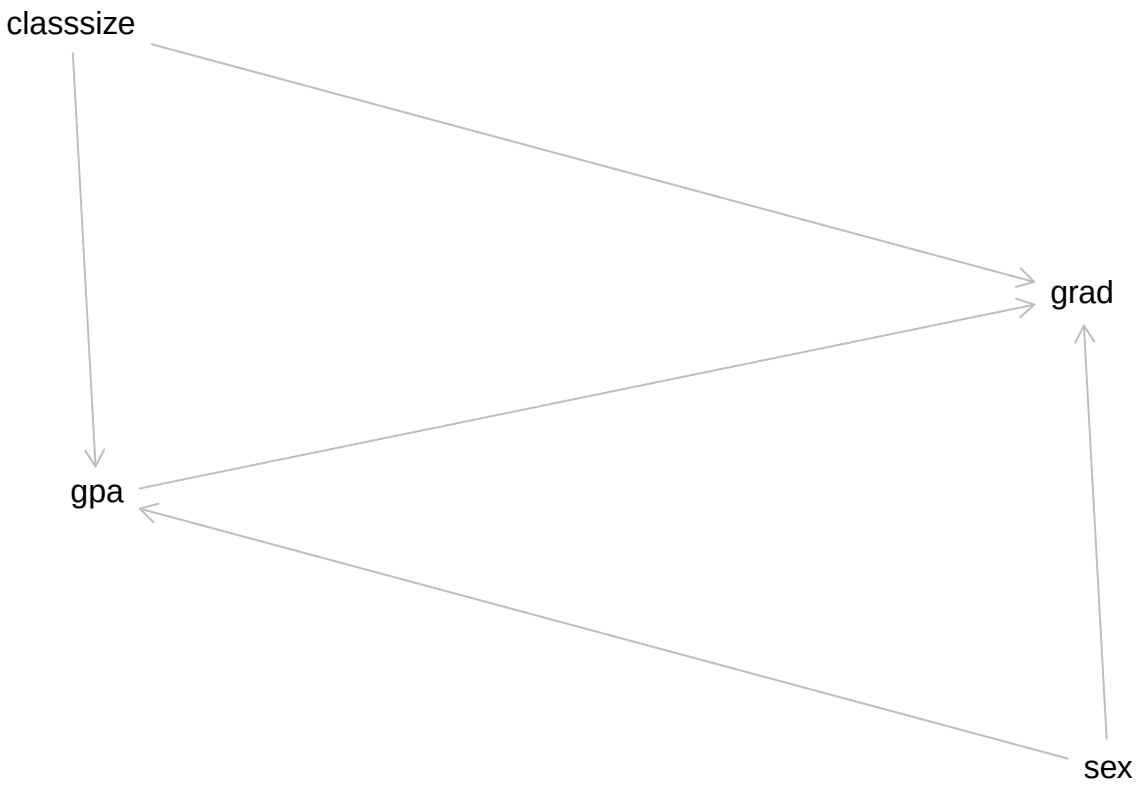
The variables of interest are:

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)
# Define the DAG using dagitty()
dag <- dagitty("dag {
  grad <- gpa <- classsize
  grad <- classsize
  gpa <- sex
  grad <- sex
}")
plot(graphLayout(dag))
```



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?

No, we can't conclude anything about whether the two CATEs (for female and for male) are statistically different from each other from the fact that one of them is statistically different from zero and the other is not because we are only observing the two CATEs themselves which is not how hypothesis testing work. If we want to determine whether the difference between the CATE for female and the CATE for male students is statistically different, we must estimate the difference between CATEs and perform a hypothesis test on that.

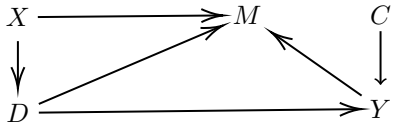
Part c (7 points)

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.

In this experiment, we don't really need to control anything to estimate an unbiased treatment effect since there is no confounders or colliders at all. Although conditioning on sex and race would help reduce standard error and make our conclusion more precise because they singly affect graduation (has nothing to do with class size), it is not necessary for us to estimate an unbiased treatment effect.

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.

1. $D \rightarrow Y \leftarrow C$: There is no confounder or collider in this path.
2. $D \rightarrow M \leftarrow Y \leftarrow C$: M is a collider.
3. $D \leftarrow X \rightarrow M \leftarrow Y \leftarrow C$: M is a collider and X is a confounder.

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.

No, we don't need to condition any variables in this DAG in order to identify the causal effect of D on Y. The first path is direct path so we don't need to control anything. The second path is a backdoor path. In this backdoor path, M is a collider. We again don't need to control anything because colliders by themselves always close backdoor paths. The third path is also a backdoor path. In this backdoor path, M is a collider and X is a confounder. For the same reasoning, we don't need to control anything here either becuase colliders close backdoor paths. If we condition on M, we open up this backdoor path that has been closed by M before. Such a control will introduce new patterns of bias since X will be a confounder that jointly determines D and Y.

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.

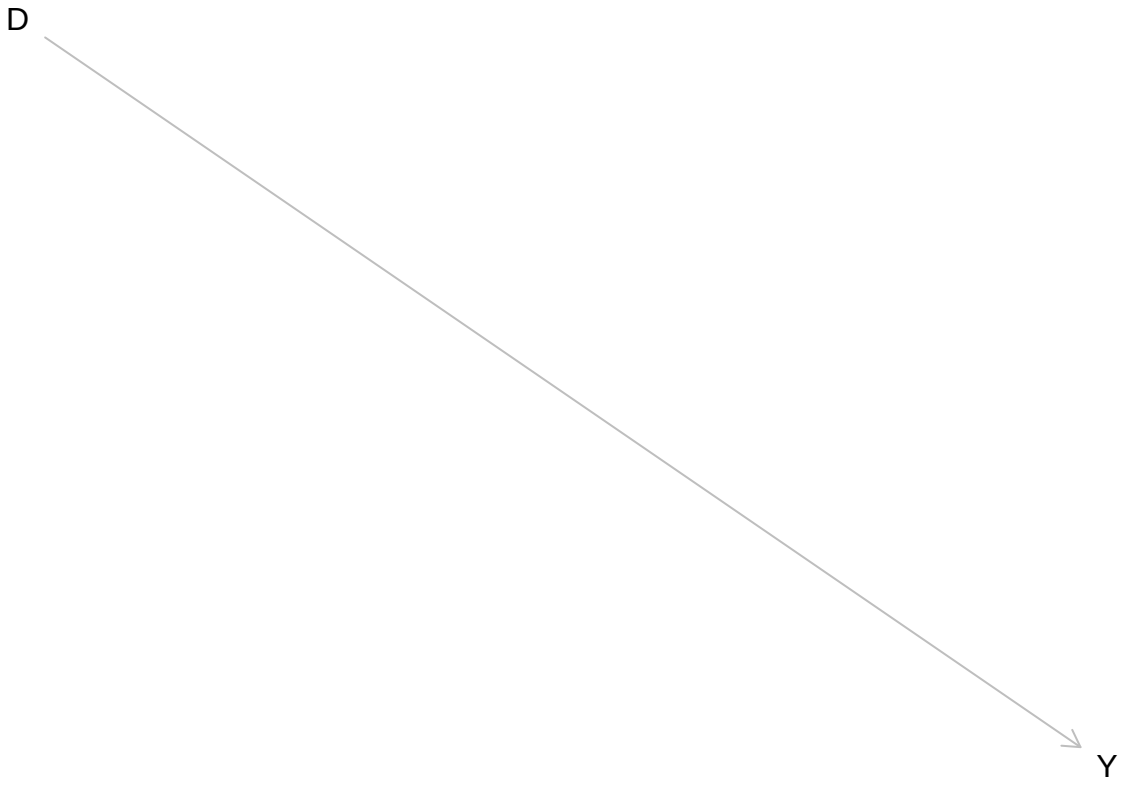
*Note: You cannot reuse an example we went over in class nor an example you used in a previous problem set.

(1) I want to estimate the effect of assigning a coach to a team affects their performance in League of Legends. League of Legends is a popular MOBA video game. Two teams play against each other with 5 people on each team. I find it really interesting because I personally love playing this game and is eager to discover how to improve my skills and rank. The treatment is assigning a coach to the team. The potential outcome is the outcome for an individual under a potential treatment. In this case, the potential outcome is the win rate of the team under treatment or control, which is whether the team has a coach or not. I want to see if teams perform better if they have a coach compared to teams without a coach.

(2) The "ideal experiment" for my question is completely randomized. The unit of analysis is a group of teams made up by players selected randomly in the North America League of Legends server. The coaches should also be selected randomly from the professional League of Legends competitions in North America so that they would have approximately the same level of knowledge and skills in the field, and they should be randomly assigned to teams for treatment. We need to make sure that other factors are held constant as well. For example, these players should be at the same rank. They should be around the same age, say they are all college students. Keeping them at the same age help reduce the bias caused by young people's faster response time. Then, we randomly assign treatment and control to our units of analysis. Using coin flips, tossing a head means the player will be put into the team with a coach, and tossing a tail means the player will be put into the team without a coach, With this experiment design, all assumptions which are required in order to ensure a causal treatment effect are satisfied. SUTVA is satisfied because we have no interference as each team that's assigned a coach will receive coaching separately, which cannot affect the state of other teams. We also have single value of treatment since the coaches are randomly selected from the same league. The quality of the treatment is kept at the same level for all teams that receive the treatment. Ignorability is satisfied because the random assignment makes the treatment assignment independent of the potential outcome. Also, the positivity assumption is met because we use coin flips to decide whether our units receive treatment. That means each unit has a 50% chance of receiving treatment and a 50% chance of receiving control, which is above 0.

(3)

```
# Define the DAG using dagitty()
ideal <- dagitty("dag {
  Y <- D
}")
plot(graphLayout(ideal))
```

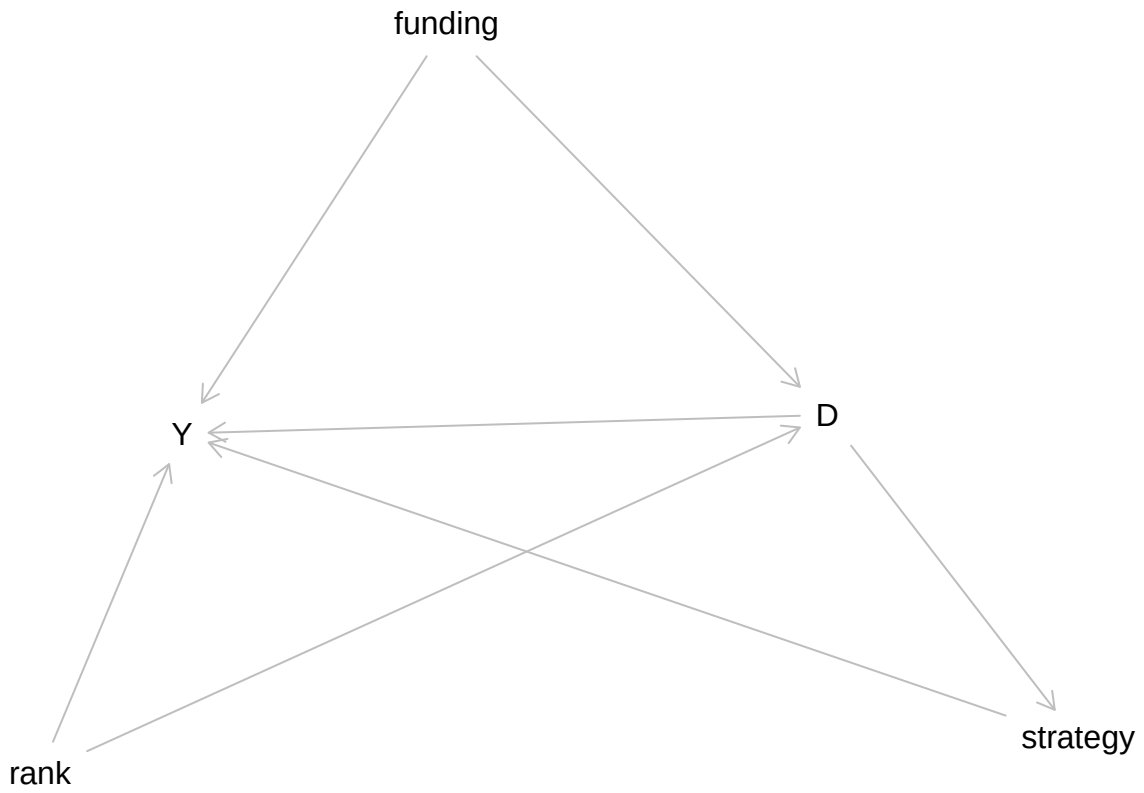


In this DAG, D is the assignment of a coach. Y is the performance of the team. Yes, we can estimate the effect of my treatment on the outcome and it is also identifiable because it is a direct path without any backdoor paths or conditioning.

(4) If I were to collect observational data on this topic, I would potentially encounter 2 confounders as well as 1 mediators. The 2 confounders are the team's fundings/sponsorships and the rank of the players on the team. They affect D and Y at the same time. The more funds the team has, the more likely they will hire a coach. The higher the rank is for players on the team, the more willing the coach would like to come to that team. At the same time, the more funds the team has, the better training condition the players will have, including better laptops and accessories, which results in higher win rate of the team. Also, obviously, the higher the rank is for the team, the higher the win rate is. The mediator is strategy because a coach can improve the team's strategy a lot, including better team cooperation, effective communication, pre-game planning as well as live decision-making. The coach can help the team study their opponent's strategy and develop counter-strategies. All of these strategies will lead to higher win rates.

(5)

```
# Define the DAG using dagitty()
observational <- dagitty("dag {
  Y <- D
  Y <- strategy <- D
  Y <- rank -> D
  Y <- funding -> D
}")
plot(graphLayout(observational))
```



(6) No, we can't estimate the impact of treatment on outcome and the effect is not identifiable because there are two confounders "funding" and "rank" on two backdoor paths. They jointly determines D and Y and confounds our ability to deduct the causal effect of D on Y. However, we can fix that by conditioning on these two variables "funding" and "rank", blocking the backdoor paths, leaving only the direct path and the path with mediator "strategy", and successfully enabling us to identify the causal effect of D on Y.