jy3440_solutions

Candice Yao - jy3440 - Section 005

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, which is the rating a customer will give if they receive the bagel. While this quantity is irrelevant to what a customer actually received in real life, it is observable if the customer actually received the bagel($D_i = 1$). However, for any customer who receive the scone($D_i = 0$), $Y_i(1)$ is an unobservable, counterfactual quantity. Therefore, $Y_i(1)$ is only observable conditioning on the assumption that customer i received the bagel.
2. $E(Y_i(1)|D_i = 0)$
   - This is the expectation of the rating of a customer who actually received the scone if they receive the bagel(in an alternative universe). This is an unobservable quantity since we cannot observe the counterfactual - we are only able to observe their rating given that they were already assigned to the scone.
3. $E(Y_i|D_i = 0)$
   - This is the expectation of the rating of a customer given that they received the bagel. This is an observable quantity, and it can be computed by taking the means of the ratings of the customers who actually received the bagel.

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	ITE	Observed outcome
1	5-5=0	5
2	9-5=4	9

Customer	ITE	Observed outcome
3	8-6=2	6
4	4-1=3	1
5	8-5=3	8
6	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?

- $\frac{(5+9+8)}{3} - \frac{(6+1+5)}{3} = \frac{10}{3}$

  Assuming that consistency(i.e. SUTVA: no spillover + single version of treatment) holds, this computed quantity is equal to a causal effect, because the treatment assignment satisfies the definition of an experiment. In other wods, the probability of treatment assignment for each individual is under total control by the researcher, and the researcher randomizes the assignment of pastry regardless of whatever pre-treatment covariates are associated with each individual. Given experimental condition, the ignorability assumption and the positivity assumption are automatically fulfilled. However, there is a caveat concerning positivity, since violation can arise out of pure chance especially given how small the sample size is(e.g. in an extreme case, all bagels can be accidentally assigned to men and all scones accidentally assigned to women). It is also important to note that the causal effect estimated from this six people might not be generalizable to the entire interested subjects(customers of this cafe) since the sample size is very small.

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?

- In this case, we are no longer able to estimate the ATE because what the barista does violates the ignorability assumption. If we give bagels to people who are originally fond of bagels, we are assigning the treatment based on potential outcomes - people will get bagels because they will be more likely to rate higher if they receive bagels.

- Besides, the positivity assumption might also be violated when preference of sweet vs. savory things is underlied by some pre-treatment covariates(e.g. gender or age). Since for every individual who has a preference, their possibility of receiving the treatment or control is fixed(either 1 or 0), so at each level of the covariate there will be imbalanced possibility to be assigned to the treatment or control

- Overall, I argue that it is impossible to estimate ATE after this new barista was hired since two necessary assumptions is violated.

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.

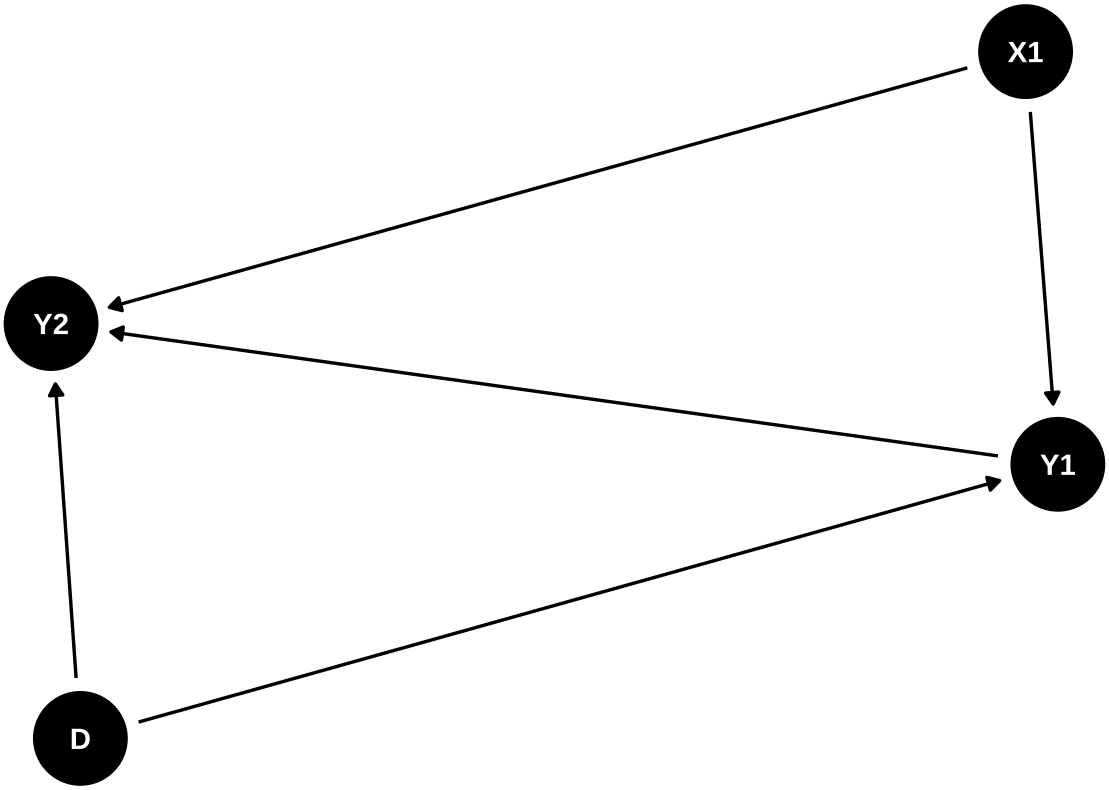
```r
#install.packages("dagitty")
#install.packages("ggdag")
library(dagitty)
library(ggdag)
```

```
##
## Attaching package: ’ggdag’

## The following object is masked from ’package:stats’:
##
##     filter
```

```r
g <- dagitty("dag{
  D -> Y1; D -> Y2; X1 -> Y1; X1 -> Y2; Y1 -> Y2;
  D [classize]
  X1 [sex]
  Y1 [gpa]
  Y2 [grad]
}")

ggdag(g) + theme_dag()
```



- In the case of this study, the class size is the treatment(D) and the outcomes are educational performance(GPA) and personal development(graduate from high school or not). Therefore there should be two arrows going out from the treatment to both of the outcome variables. Next, while gender could have been a potential confounder, its association with the assignment of treatment is broken since students are randomly assigned to classes of different sizes, so there should be two edges going out from gender to the two outcomes, assuming that one's gender may affect their outcome of treatment given how the society systematically bias against women. Last, one's GPA may affect their possibility of graduating from high school since one with a lower gpa can be more likely to decide to quit school, so there should be an edge between Y1(gpa) and Y2(grad).

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?

- This is insufficient for us to draw conclusions because the significance of one CATE and the insignificance of another could be caused by (a) random noises, which is possible to arise for smaller subgroups, or (b) the two subgroups are in fact different. For us to tell if the ostensible difference here is by pure chance or by a systematic difference in the subgroups, we will need to directly compute the difference in CATES and the its asymptotic 95% confidence interval to assess if the two groups' difference in CATEs are statistically significant. If the 95% confidence interval computed does not include zero, we can conclude that it is very unlikely that the difference in CATE occurs due to pure chance, and we reject the null hypothesis that there is no difference between treatment effects across genders at

$a = 0.05$. On the other hand, if the confidence interval includes zero, we are unable to reject the null, and the difference could be caused by random noises.

Part c (7 points)

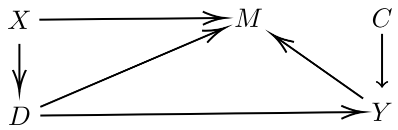
Imagine we wanted to estimate the effect of class size on finishing high school in this experiment. What would be necessary for you to control to estimate an unbiased treatment effect? How would you estimate the treatment effect? Explain your answer.

- Since this is an experiment and the students are randomly assigned into classes of different class size, the observed pre-treatment covariates' effects on treatment assignment were already accounted for. This means that students of all races or genders should be equally likely to be treated. Hence, if the assignment is truly randomized, gender and race will not be confounders and we don't need to control for them to get an unbiased treatment effect.
- To estimate average treatment effect(ATE), we can first calculate the means of outcome for students assigned to each of the three class size, and than we compute the differences(and their confidence intervals) between the means to assess if the differences are statistically significant.

  Additionally, while controlling for gender and race are not necessary for us to estimate unbiased treatment effect, it is still useful for us to estimate CATE(onditional average treatment effect) and account for heterogeneous treatment effects. Besides, it is notable that in real-life circumstances, even a designed experiment has chances of failing to account for unobserved confounders, which may also lead to biases.

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.

paths:

- D->M<-Y(non-causal)
  - confounder: none; collider: M
- D<-X->M<-Y(non-causal)
  - confounder:none; collider: M
- D->Y(causal)
  - confounder: none; collider: none

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 is no confounder we need to condition on in order to identify the causal effect because there is no variable on the paths that is casually linked to both D and Y.
  - D and Y collide at M, and we don't want to control for collider; X has casual paths to D and M, but since Y and X collides at M, the path from X to Y is also not causal(if it is, then X will be a confounder and we would need to control for it).

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) My study aims to study the effect of having lived in a dorm for at least one year on one's social accomplishment throughout university. I think this question is particularly interesting because many universities including NYU require their students to live on campus in the freshman year. However, due to the COVID-19 pandemic, only a portion of the 2024 class have lived on campus while the other commuted all the time, thus enabling us to compare the social outcomes between the two groups. While some argue that dormitory life increases one's social capability, others might contend that living in a dorm could limit one's potential social opportunities. It will be particularly interesting to assess which stance holds, or is there no difference between where a college student lives at all.

- Treatment: $D_i = 1$ means that one has lived in a dorm for at least one year.
- Control: $D_i = 0$ means that one has never lived in a dorm throughout their university.
- Outcome: A student's number of friends made throughout the time span of their undergraduate. This quantity will be used as an indicator of one's social accomplishment. Here, we assume a rather narrow definition of social accomplishment, which is how capable one is in making new friends.

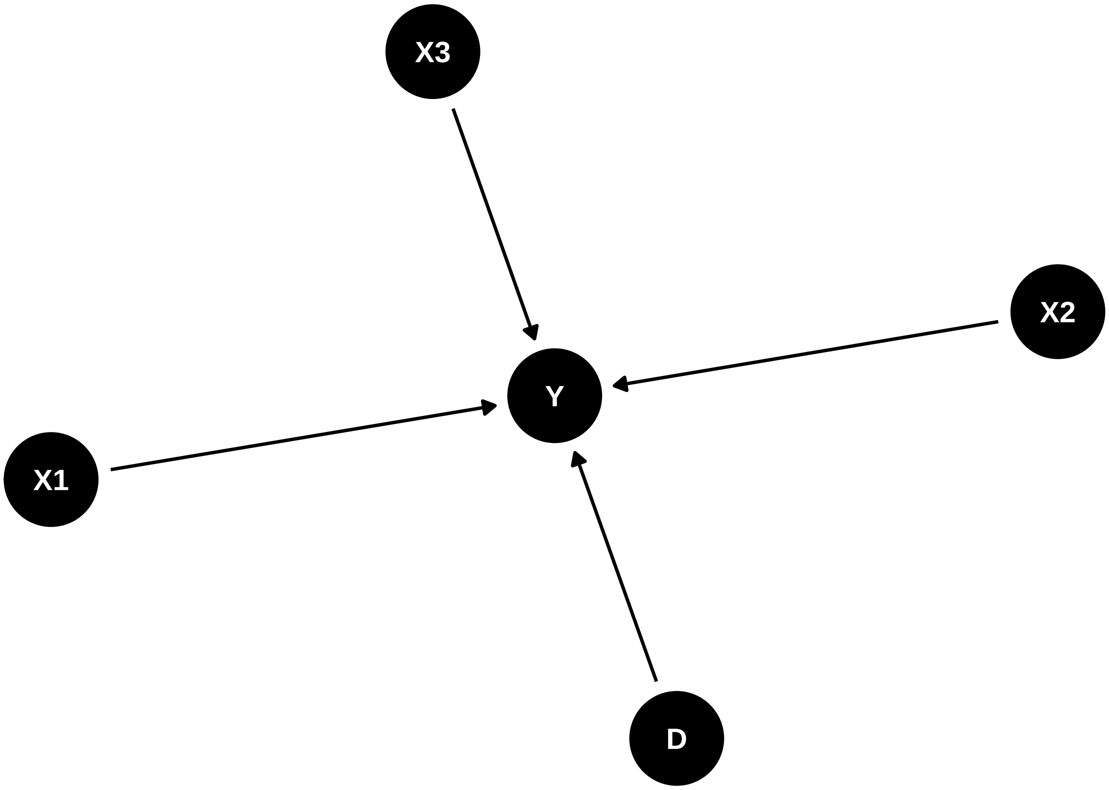
(2) The ideal experiment will be that I as the researcher can recruit a sample of students as participants, and randomly assign some to live in the dormitory for at least one year, and ask the rest to live off-campus throughout their university. In this case, randomization takes care of potential confounders like a student's nationality, race, and family income(assuming that these variables are what could affect one's decision of where to live and their capability of making new friends and there are no other unobserved variables). In an ideal experimental condition, I will be able to ascertain that treatment is assigned randomly at all levels of each of these confounders.

(3)

```r
g1 <- dagitty("dag{
  D -> Y; X1 -> Y; X2 -> Y; X3 -> Y
  D [have lived in dorm for at least one year]
  X1 [family income]
  X2 [nationality]
  X3 [race]
  Y [friends made since college]

}")

ggdag(g1) + theme_dag()
```



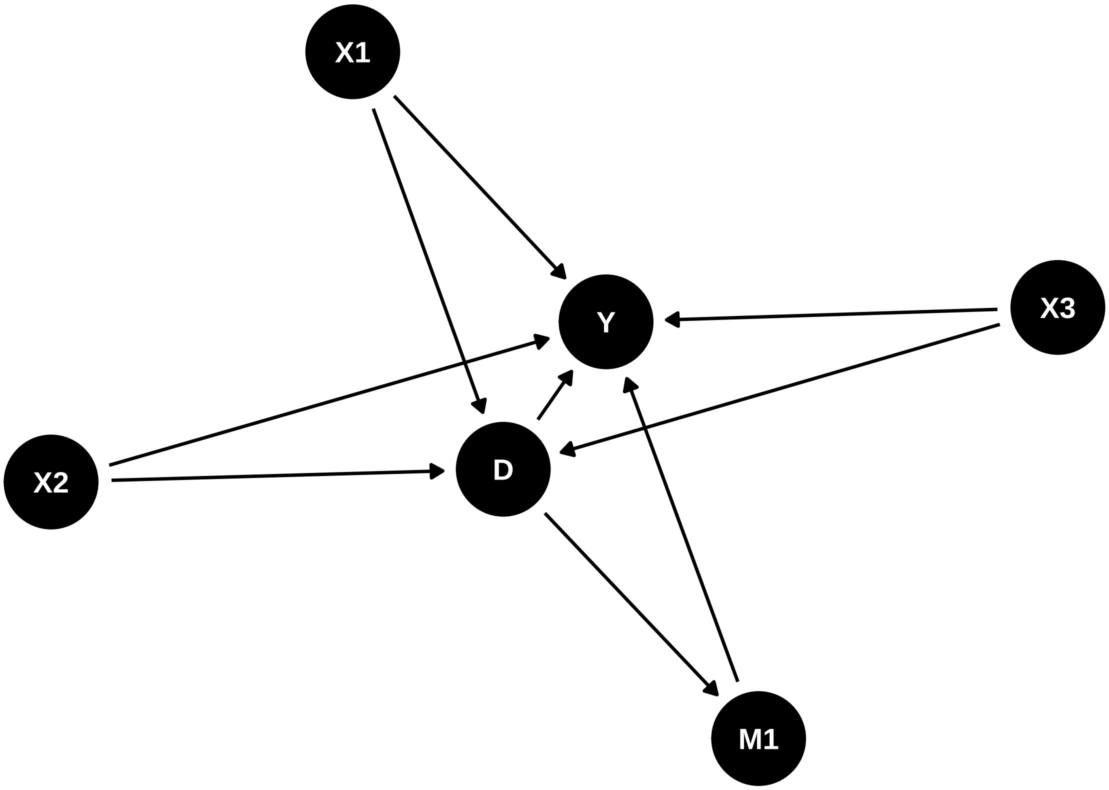
- Yes, I can estimate the effect of my treatment on the outcomes. I can do so by computing the difference in means of the outcomes(i.e. the friends made throughout college) between my treatment group and my control group, and construct the confidence interval to assess its statistical significance.
- We know that the ATE is identifiable because in an ideal experiment, we guarantee the randomization of treatment assignment at all level of potential observed confounders(and there is no unobserved confounders). Consequently, there will be no confounder lying on the path from the treatment to the outcome(which means no backdoor path), and the effect on the outcome can be attributed to the difference in treatment versus control.

(4) If you were to collect observational data on this topic, what potential confounders and mediators would exist? Please explain them in words.

- In an observational study, I don't have the capability and power to assign students to different residential situations. Instead, I can only collect information of students' residential behaviors throughout college and the number of new friends they have made since college. In this case, all three of the covariates listed above(race, nationality, and family income) can be confounders that both affect one's decision of where to live, and the number of new friends they could make since college. For example, being an international student can make one more inclined to live off-campus at a community where other international students live. At the same time, their cultural background can make it harder for them to make new friends in the U.S.. In this case, nationality is a confounder. Besides, coming from a wealthier family could increase the likelihood of one choosing to live on-campus due to how expensive dorms can be in NYC.
- One mediator could be the access to school-held social events. Basically, living in a dorm will give a student more access and convenience to take part in those events and thus having more chances to make new friends, since dorms are often located close to the campus.

(5)

```
g2 <- dagify(
  Y ~ D + M1 + X1 + X2 + X3,
  D ~ X1 + X2 + X3,
  M1 ~ D,
  labels = c(
    D = "have lived in dorm",
    X1 = "family income",
    X2 = "nationality",
    X3 = "race",
    M1 = "Access to social event",
    Y = "friends made since college"
  )
)
ggdag(g2) + theme_dag()
```



- With the above DAG, I can estimate the treatment effect of D on Y if conditioning on all observed confounders(X1 - family income, X2 - nationality, and X3 - race). The effect is identifiable because if I close all backdoor paths(control all observed confounders), the difference in outcomes between groups can be attributed to the treatment. Since we care about the total effect of the treatment(including direct and indirect effects), we should not control for the mediator(M1 - assess to school-held social event) since it is on one of the causal paths going from D to Y(D->M1->Y).