Homework 1

CS396-4 Causal Inference

January 14, 2022

Instructions

This assignment is due on Thursday, Jan 20 at 11:59pm CST. Late assignments will be accepted, but with a 14.3% (1/7th) penalty per day late. If your assignment is less than 24 hours late, we'll grade it and you'll receive 85.7% of those points; if it's less than 48 hours late, you'll receive 71.4% of those points. If it's more than 6 days late, you'll receive no points.

Your answers must be uploaded to Canvas as a single pdf document; you should edit the LaTeX source for this pdf to add in your answers. This is an individual assignment – you are welcome to discuss the problems with your classmates, but you must solve each part and write each answer on your own.

0.1 (1 point)

By submitting this assignment, you affirm that you have neither given nor received any unauthorized aid on this assignment, and that the solutions shown here are wholly you own. Any violation of Northwestern's academic integrity policies will result in you receiving a 0 on this assignment and a report to your dean's office. If you're ever worried about whether you are at risk of violating these policies, please ask – we can help you follow the rules, but we can't retract a report of suspected cheating.

1 Simpson's Paradox with synthetic data

For this question, you will need to run the code provided in `lecture2demo.py`. It uses the Python libraries `numpy`, `pandas`, and `statsmodels`. If you have trouble running this code, make a post on CampusWire and we'll provide help getting set up. If you are familiar with either `virtualenv` or `conda`, you may find it helpful to use such an environment to manage dependencies.

1.1 (1 point)

Fill in the tables by running the provided code with the indicated arguments. The code returns the mean and standard deviation of `repeats` samples. The first cell in each table is filled in for you. For example, in the top left table, the first cell's value is computed by running:

```
python lecture2demo.py observed --repeats 10 --c_dim 10 --ols "y ~ a"
```

You will use these tables to answer the next few questions.

observed(ols = ”y $\sim$ a”, . . .)

repeats	c_dim = 10	c_dim = 500
10	-1.625 ± 0.454	
100		
1000		

observed(ols = ”y $\sim$ a + c”, . . .)

repeats	c_dim = 10	c_dim = 500
10	0.409 ± 0.625	
100		
1000		

randomized(ols = ”y $\sim$ a”, . . .)

repeats	c_dim = 10	c_dim = 500
10	0.480 ± 0.387	
100		
1000		

randomized(ols = ”y $\sim$ a + c”, . . .)

repeats	c_dim = 10	c_dim = 500
10	0.517 ± 0.213	
100		
1000		

1.2 (1 point)

The true causal effect of A on Y is 0.5. Which table has mean results furthest away from that value? Why?

observed data and using y ~ a because we are not taking into account the effect of C variable on outcome Y, and also we're not using randomized trials and in the observed case the result is affected by the C variable so it will not be 0.5

1.3 (1 point)

in general, randomized y~a low c_dim is accurate, y~a+c high c_dim also accurate. if c_dim too high without c, will be unstable.

How do the mean and standard deviation of the results change as you increase `c_dim` and `repeats`? What explains these trends?

increasing repeats pushes the mean closer to 0.5 while the variance is not affected. increasing c_dim makes the mean very off and also has high variance. because the distribution of 1 coin flip making a difference is very tease out.

1.4 (1 point)

Compare the `c_dim = 500` columns in the top right table (**observed(ols = ”y $\sim$ a + c”, . . .)**) and the bottom left table (**randomized(ols = ”y $\sim$ a”, . . .)**). This is the only comparison (for the same value of `c_dim`) where a column in a **observed** table has lower variance and a mean closer to 0.5 than a **randomized** table. Why does this happen for this comparison? Why doesn't it happen anywhere else?

because in the observed case we take into account the confounding variable whereas in the randomized case not taking into account the c variable while having a high c_dim will cause high variance in mean and stdev. Important is taking into the c when choosing ols method.

why not anywhere else? look at cdim10, dont see it there. why cdim controls this behavior?

2 Simpson's Paradox in expectation

		A=0	A=1	
C=0		$x_1 = 0.93$	$x_2 = 0.87$	
	0.931	(81/87)	(234/270)	0.867
C=1		$x_3 = 0.73$	$x_4 = 0.69$	
	0.73	(192/263)	(55/80)	0.6875
Both		$x_5 = 0.78$	$x_6 = 0.83$	
	0.78	(273/350)	(289/350)	0.8257

Table 1: Simpson's Paradox, as covered in lecture. C is patient age, A is one of two drugs. Each cell shows average (binary) recovery rate Y. We've named the cells x_i to make them easier to reference below.

Consider Table 1, which we saw in lecture (for example, slide 9 of lecture 1). We said that this table shows Simpson's paradox because if you don't know the causal structure of the data, you can't tell which drug ($A = 0$ or 1) is better. If C were a mediator (like a side effect), we should compare x_5 against x_6 to see which drug is best. But if C were a confounder (like age), we should compare x_1 against x_2 and compare x_3 against x_4.

2.1 (1 point)

For $i = 1 \ldots 5$, define x_i as a conditional expectation involving Y, A, and C. For example, $x_6 = \mathbb{E}[Y \mid A = 1]$.

2.2 (1 point)

Consider the return statement in the `observed` function of `lecture2demo.py`:

```
smf.ols(ols, data=df).fit().params['a']
```

For both `ols = "y ~ a"` and `ols = "y ~ a + c"`, write out the value returned by this line in terms of the expectations you wrote out in your answer for 2.1. Explain your answer based on what `smf.ols` is doing.

You may assume that `c_dim = 2` and assume that C is always either 0 or 1. You may find it helpful to reference the `statsmodels` documentation here and here, as well as CampusWire post # 9 here.

2.3 (2 point)

Note that in Table 1, $p(C = 0) = 357/700 = 0.51$ and $p(C = 1) = 343/700 = 0.49$. Define $\Theta = p(C = 0) \cdot (x_1 - x_2) + p(C = 1) \cdot (x_3 - x_4) - (x_5 - x_6)$. What is the relationship between Θ and Simpson's Paradox? Explain.

2.4 (2 point)

Consider the data-generating process implemented in the `observed()` function in `lecture2demo.py`. Suppose we let `c_dim = 10`; we must update our definition of Θ to include $p(C = 1), p(C = 2), \ldots,$ and $p(C = 10)$. We can also show that $\mathbb{E}[Y \mid A = 1] \approx 3.055$ and $\mathbb{E}[Y \mid A = 0] \approx 4.505$. With `c_dim` $= 10$ and with the new definition of Θ, what is $\mathbb{E}[\Theta]$ for the `observed()` data-generating process?

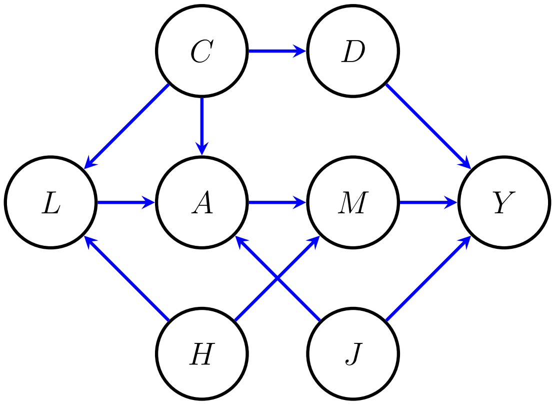
x5 = 4.5, x6 = 3, p(c=0) = p(c=1) = 0.1

2.5 (2 point)

Let C, A, and Y be binary variables where C is patient age, A is drug assignment, and Y is recovery. Suppose we have some dataset sampled from the distribution $p(C, A, Y)$ that represents a randomized trial. Assume that A is marginally randomized, such that each patient has an equal probability of receiving $A = 0$ or $A = 1$. Using the definition of Θ from 2.3 above, use the rules of probability to show that for this distribution, $\mathbb{E}[\Theta] = 0$.

3 d-Separation in graphical models

Consider the following DAG:



C>L<H: dec: {A, M, Y}
C>A<L: dec: {M, Y}
L>A<J: dec: {M, Y}
C>A<J: dec: {M, Y}
A>M<H: dec: {Y}
M>Y<J: dec: {}
D>Y<J: dec: {}
M>Y<D: dec: {}

3.1 (1 point)

List all collider ("head-to-head") nodes in the graph. For each, list all their descendants in the graph.

3.2 (2 points)

For each of the following parts, we write "Is $A \perp B \mid C$" to mean "Is A d-separated (and therefore conditionally independent) of B given C?" For each question, provide your explanation in terms of blocked and unblocked paths. For example, if we asked, "Is $L \perp D \mid C$?" it would not be enough to say "yes" – you must explain, e.g.:

> Yes, because all paths from L to D go through C or Y, and the path $L \leftarrow C \rightarrow D$ is blocked by conditioning on C, the path $L \rightarrow A \leftarrow C \rightarrow D$ is blocked by conditioning on C, and the path $L \rightarrow A \rightarrow M \rightarrow Y \leftarrow D$ is blocked at Y, a collider.

1. Is $H \perp J \mid Y$? Why?

   False: M>Y<J collider (observed) so unblocked. H>M>Y chain, observed at Y does not give info about H so unblocked. Both of these unblocked, is not d-sep.???

2. Is $H \perp J \mid L$? Why?

   True - H>L>A since L observed, is blocked. J>A<L is collider, parent observed so is blocked. Thus, H to J is d-sep

   True

3. Is $A \perp Y \mid J, M, C$? Why?

   False

4. Is $L \perp Y \mid A, M, C, D$? Why?