Assignment 2

CS6480: Causal Inference and Learning
IIT-Hyderabad
Feb-Apr 2021

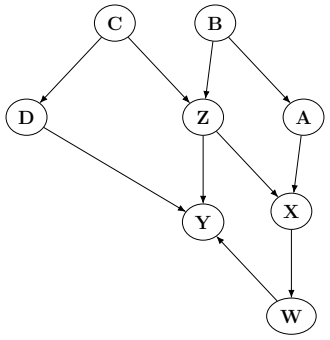
Max Marks: 40
Due: 11th Apr 2021 11:59 pm

Instructions

- Please use Piazza to upload your submission by the deadline mentioned above. Your submission should comprise of a single file, named `<Your_Roll_No>_Assign2`, with all your solutions.
- For late submissions, 10% is deducted for each day (including weekend) late after an assignment is due. Note that each student begins the course with 5 grace days for late submission of assignments, of which upto 3 grace days can be used for a single assignment. Late submissions will automatically use your grace days balance, if you have any left. You can see your balance on the Marks and Grace Days document, soon to be shared under the course Google drive.
- Please read the department plagiarism policy. Do not engage in any form of cheating - strict penalties will be imposed for both givers and takers. Please talk to instructor or TA if you have concerns.

Questions

1. **(5 points)** Define and give examples for the following assumptions of causal inference.
   (a) SUTVA (2 points)
   (b) Large Sample Size (1 point)
   (c) No Measurement Error (1 point)
   (d) Double Blindedness (1 point)
2. **(8 points)** Consider the following DAG.



(a) List all of the sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y (2 points)

(b) List all of the minimal sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y (i.e., any set of variables such that, if you removed any one of the variables from the set, it would no longer meet the criterion) (2 points)

(c) List all the minimal sets of variables that need to be measured in order to identify the effect of D on Y (2 points)

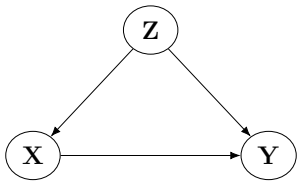
(d) Now suppose we want to know the causal effect of intervening on 2 variables. List all the minimal sets of variables that need to be measured in order to identify the effect of set $\{D, W\}$ on Y, i.e., $P(Y = y|do(D = d), do(W = w))$ (2 points)

3. **(11 Points)** Consider the SCM:

$$\begin{aligned} X &:= N_X \\ Y &:= 4X + N_Y \\ N_X, N_Y &\overset{i.i.d.}{\sim} N(0, 1) \end{aligned}$$

- The following quantities are normal distributions. What is the mean and variance of each of these quantities? (5 Points)
  (a) P_Y
  (b) $P_{Y|X=k}$
  (c) $P_{Y|do(X=k)}$
  (d) $P_{X|Y=k}$
  (e) $P_{X|do(Y=k)}$
- Write Python code to generate 100 samples from each of below distributions and visualize the results using appropriate plots (histogram, scatter plot etc). (6 Points)
  (a) $P_{X,Y}$
  (b) $P_{Y|X=2}$
  (c) $P_{Y|do(X=2)}$

4. **(10 Points)** Assume that we are studying a population of patients with a particular disease for which a treatment X may be helpful. It turns out that some of these patients are impacted by a syndrome Z that impacts survival Y but also impacts the patients ability to tolerate the treatment X.



The graph above describes the relationship between the syndrome Z, the treatment (or drug) X and the outcome Y (death or survival) for the population of patients. Suppose that a fraction of the population r suffers from the syndrome (i.e., have $Z = 1$) with the remaining proportion

$1-r$ not having the syndrome (i.e., $Z=0$). Let $X=1$ represent a patient taking the drug and $X=0$ represent a patient not taking the drug. And let $Y=1$ indicate that a patient dies and $Y=0$ indicate that a patient survives. Assume that patients without the syndrome die with probability p_1 if they don't take the drug and die with probability p_2 if they do take the drug. Patients with the syndrome die with probability p_3 if they don't take the drug and die with probability p_4 if they do take the drug. The complication is that having the syndrome makes it uncomfortable to take the potentially life saving drug. Assume that patients with the syndrome take the drug with probability q_2 and patients without the syndrome take the drug with probability q_1.

(a) Find the joint distribution $P(x,y,z)$ for all x,y,z (8 values) in terms of the parameters $r,p_1,p_2,p_3,p_4,q_1,q_2$.(1 point)

(b) Calculate the difference in death probabilities between takers and non takers of the drug, $P(y=1|x=1)-P(y=1|x=0)$(1 point)

(c) Calculate the difference in death probabilities between takers and non takers of the drug for those with $z=1$ (having the syndrome), $P(y=1|x=1,z=1)-P(y=1|x=0,z=1)$(1 point)

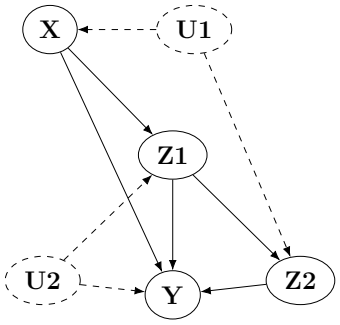
(d) Calculate the difference in death probabilities between takers and nontakers of the drug for those with $z=0$ (not having the syndrome)(1 point)

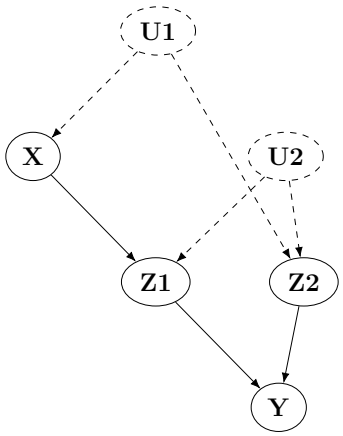
(e) Find a combination of parameter values that exhibit Simpson's paradox (i.e., where (c) and (d) show negative effect (lower death rate) but (b) doesn't(2 points)

(f) Compute $P(y|do(x))$ for all values of x and y(2 points)

(g) Compute the average treatment effect $P(y=1|do(x=1))-P(y=1|do(x=0))$. How does this quantity differ from the quantity computed in (b) above? Which is more relevant in assessing the effectiveness of the treatment? Explain(2 points)

5. **(6 points(3+3))** For each of the two graphs below, using the rules of *do*-calculus, explain whether $p(y|do(X=x))$ is identifiable or not?dashed nodes represents unobserved variables.




U1
X
U2
Z1
Z2
Y