Pset 3

Due 09/28/2023

Problem Set 3. DAGs.

Relevant material will be covered by **Sep 21**. Problem set is due **Sep 28**.

To complete the problem set, **copy this code** into a .Rmd and complete the homework. Omit your name so we can have anonymous peer feedback. Compile to a PDF and submit the PDF on Canvas.

Note: For this problem set only, you alternatively may complete the homework by hand. This is because you are welcome to draw DAGs by hand instead of producing them by code. If you do this, scan or take a picture of your document.

1. True or False

For 1.1–1.5, answer True or False: X is a sufficient adjustment set to identify the causal effect of A on Y. Explain in one sentence. If False, state the backdoor path that is unblocked conditional on X. A path is a linear series of nodes connected by arrows; see examples in 1.6 and 1.7.

1.1 Answer.

True as all the paths to Y from A are casual and open, and when we condition on X the path is blocked leaving no other paths to Y

1.2

False as even if we condition on X and block that path, another backdoor path can be taken. This other path goes A <-U->Y

1.3 Answer.

True. This is because even though there is a path going from U1 it has to pass through X, so if we condition on X it blocks all possible backdoor paths

1.4 Answer.

True. X is not a collider in this scenario, also all backdoor paths go through it, so by conditioning on X it blocks all other available back paths.

1.5 Answer.

True, because all of the paths that get to Y require passing through X so conditioning on it it blocks all the other paths making it enough for the situation.

1.6 (3 points)

True or False? Conditioning on X blocks this path: $A \leftarrow B \leftarrow X \rightarrow C \rightarrow Y$

Answer.

TRUE, as X is a non-collider, by conditioning it, it will become a collider and block this path

1.7 (3 points)

True or False? Conditioning on X blocks this path: $A \leftarrow B \rightarrow X \leftarrow C \rightarrow Y$

Answer.

FALSE as X is already a collider in this path so by conditioning on it, makes the path open by turning it into a non collider.

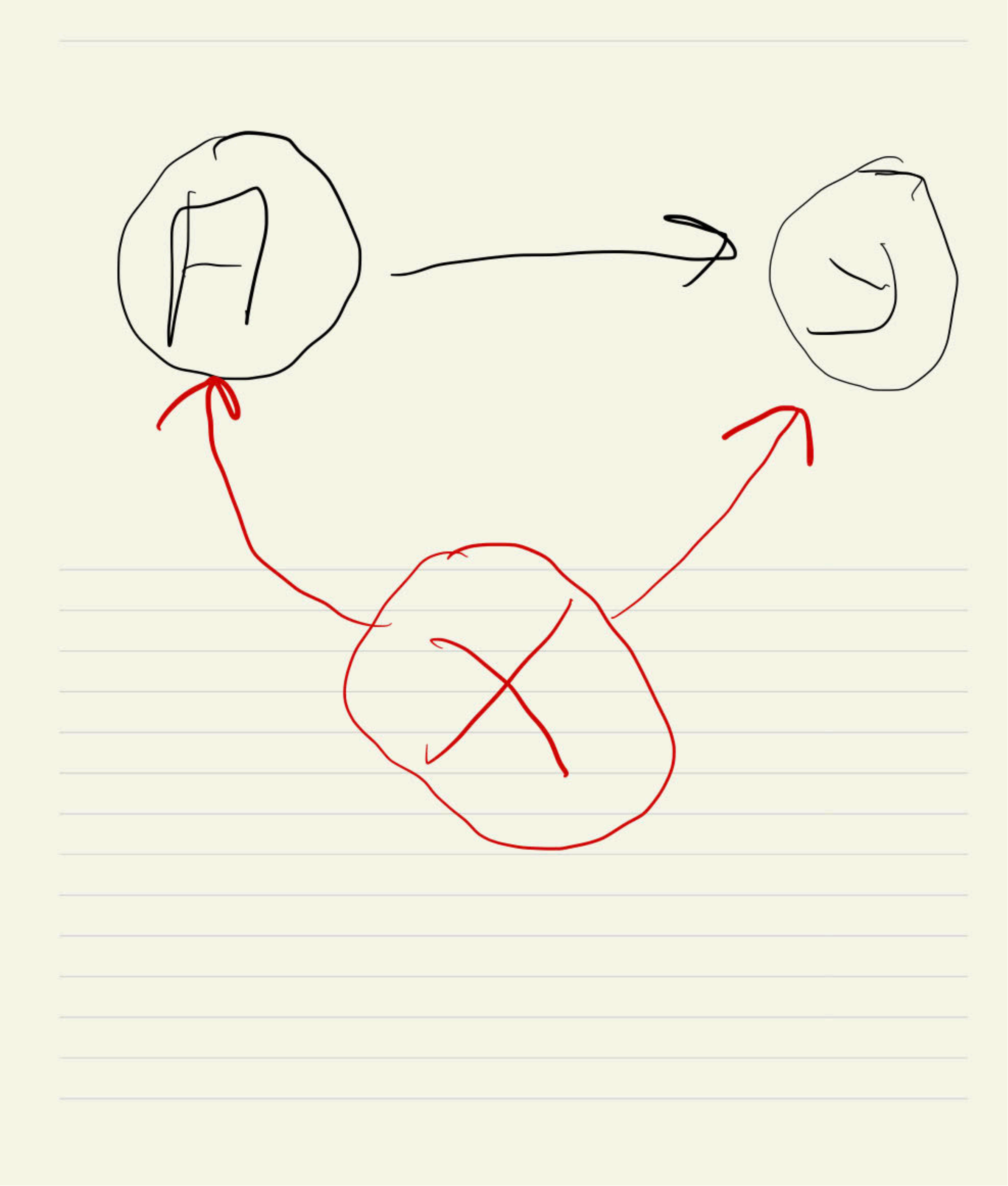
2. Draw a DAG (10 points)

A researcher comes to you with a proposal: identify the causal effect of A on Y by adjusting for any variable X that predicts A and also predicts Y. They propose that machine learning can thus solve causal identification for us.

The researcher is wrong. Show them why. Draw a DAG in which

- the effect of A on Y is unconfounded
- a variable X is statistically associated with A
- a variable X is statistically associated with Y
- but one does not need to adjust for X to identify the causal effect

Answer.



There is an edge from A to Y with no hidden variables, which shows that it is indeed unconfounded. Next, we have an edge going from X into A like this -> to show that X is associated with A in some form, and I did the exact same thing with X->Y when it comes to showing association between the two variables. Finally I made it so there lies a path A->Y to show that X does not need to be adjusted in order to show a causal effect.

3. Using DAGs in a new context

DAGs are not just useful for causal inference: they can be useful whenever we need to know whether one variable is statistically independent of another. This is true, for example, when drawing inference about a population from a sample.

A researcher uses an opt-in online web survey to draw inference about support for President Biden. They ask respondents: “Do you approve of President Biden’s performance in office?’ ’ with the answer choices Yes/No. The researcher also gathers data on two demographic characteristics: whether the respondent completed college and current employment. They write:

> My sample is not representative. Suppose for every person in the population, S denotes whether they are included in my sample. Then S is related to their approval of President Biden (Y).
>
> However, I believe my sample *is* representative when I look at a set of people who all take the same value along college completion and employment, such as those who finished college and are currently employed. If these variables are X_1, X_2, I believe this independence statement: $S \perp\!\!\!\perp Y \mid X_1, X_2$. I will therefore get population estimates by a procedure with several steps: use my sample to estimate the mean outcome $E(Y \mid \vec{X} = \vec{x})$ in each stratum, then use Census data to estimate the size of each stratum $P(\vec{X} = \vec{x})$ in the population, then estimate $E(Y) = \sum_{\vec{x}} E(Y \mid \vec{X} = \vec{x}) P(\vec{X} = \vec{x})$.

This researcher’s reasoning is a common strategy known as **post-stratification**. This question is about formalizing a set of conditions under which the researcher is right and wrong.

Before you begin, we want to emphasize one aspect of the researcher’s assumption that is different from the exchangeability assumption for causal inference.

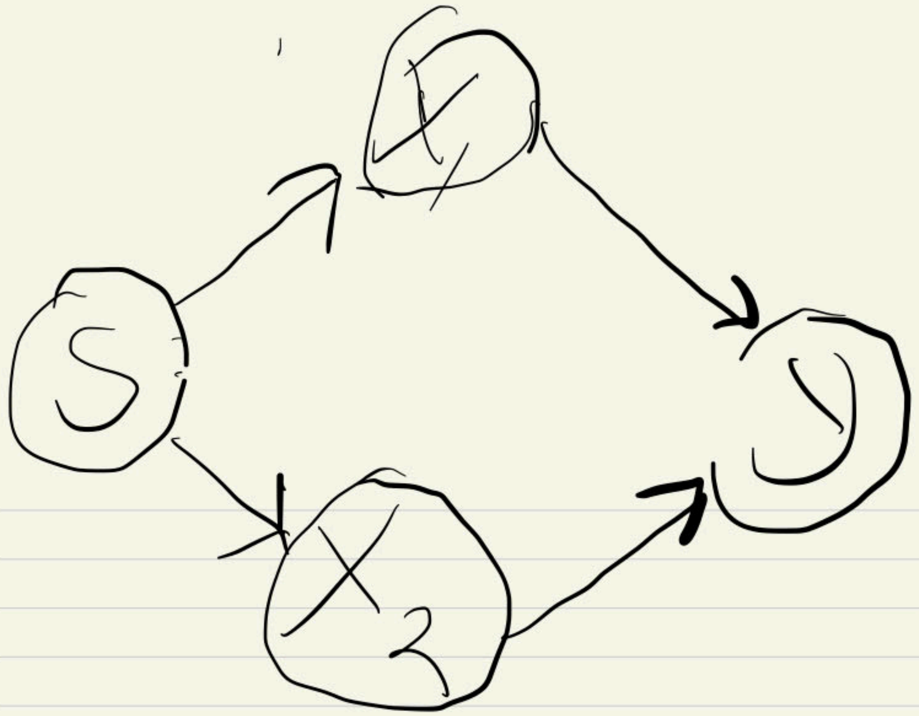
- for causal claims, we assume conditional exchangeability $A \perp\!\!\!\perp Y^a \mid \vec{X}$
  - involves the potential outcome Y^a
  - holds if the only unblocked paths between A and Y are causal paths
- for sample-to-population inference, we assume conditionally independent sampling $S \perp\!\!\!\perp Y \mid \vec{X}$
  - involves the factual outcome Y; there is no intervention here
  - holds if there are no unblocked paths between S and Y

Although the assumption is different, the principles of DAGs are still relevant.

3.1. (5 points)

Draw a DAG under which the researcher's claim is valid. Use S, Y, X_1, X_2.

Answer.



3.2. (2 points)

In a sentence or two, explain your DAG from 3.1 to the researcher. Tell us in words what is meant by each edge in your DAG.

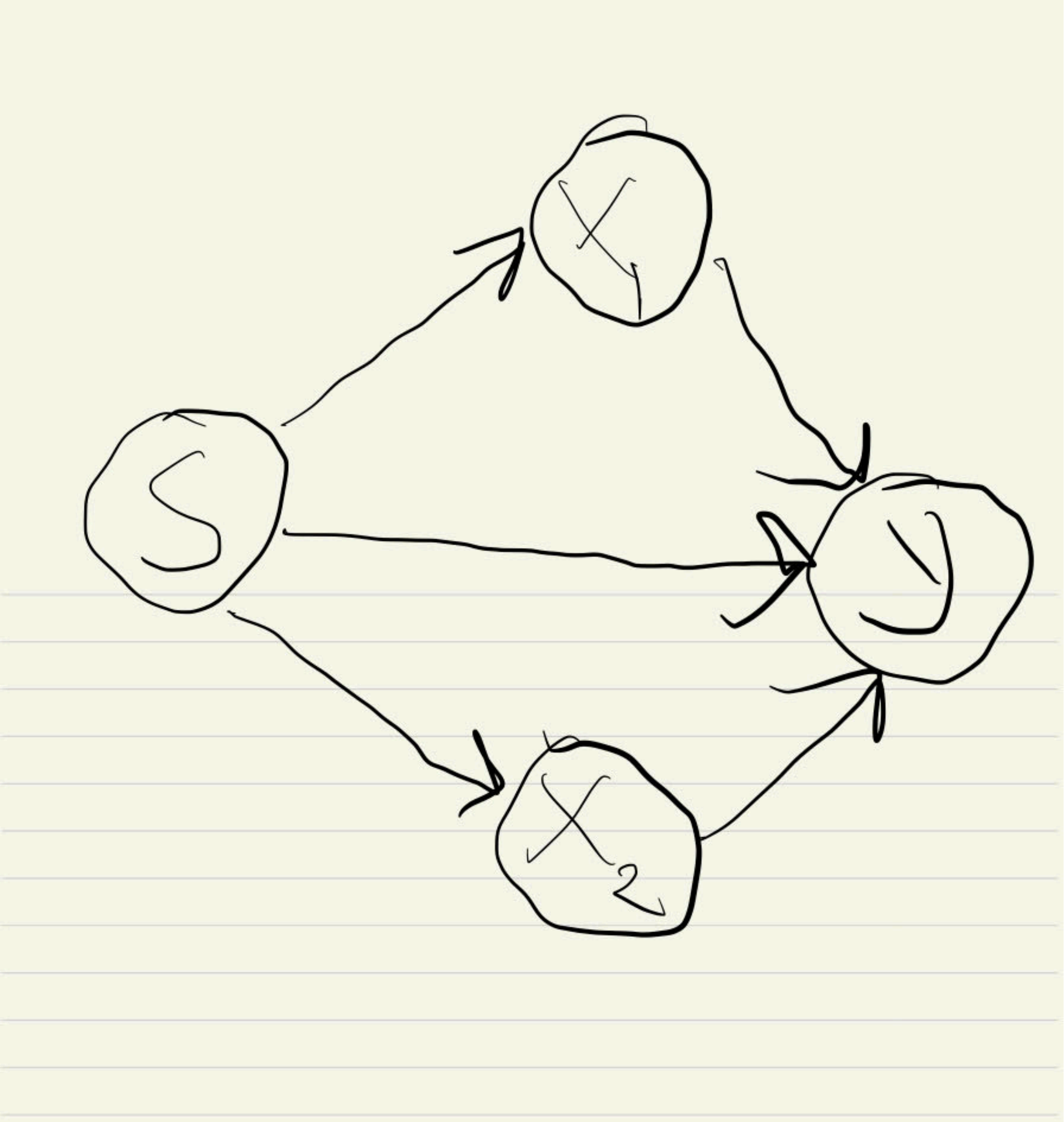
Answer.

So the edges S->X1 and S->X2 show that the number of persons who graduated college and are employed are affected by S . Finally the edges X1->Y and X2->Y show that Y is affected by both those variables. Now in order to show independence I made it so that S->Y does not exist directly

3.3. (5 points)

Draw a DAG showing a counterexample under which the researcher's claim is invalid.

Answer.



3.4 (2 points)

In a sentence or two, explain your DAG from 3.3 to the researcher. Tell us, particularly about the path that creates a statistical dependence between S and Y.

Answer.

This graph is quite similar to the first one in that S->X1 and S->X2 show that the number of persons who graduated college and are employed are affected by S . Finally the edges X1->Y and X2->Y show that Y is affected by both those variables. Now the part that shows dependence between S and Y is the edge S->Y. With that edge, S has a direct relationship on Y making it so that whatever S is has an effect on Y.