Data 102 (Fall 2022) **Names:**
Discussion Worksheet 9
10/28/2022

1. Recall the kidney stone problem from last week, where treatment A was given to one half of a group of kidney stone patients and treatment B to the other half. The results of the study were

	Treatment A	Treatment B
# cases given treatment	100	100
# cases after treatment	25	20

When taking into account size of the kidney stones, we saw that

	Treatment A	Treatment B
# cases with small kidney stones given treatment	20	90
# cases with small kidney stones after treatment	2	15
# cases with large kidney stones given treatment	80	10
# cases with large kidney stones after treatment	23	5

(a) What are the confounder X, the treatment Z, the outcome Y, and the propensity score $e(X)$ in this problem?

Solution: X: size of kidney stone. Z: treatment ($Z = 1$ if treatment A, $Z = 0$ if treatment B). Y: outcome ($Y(1) = 1$ if treatment A was successful, 0 if not; $Y(0) = 1$ if treatment B was successful, 0 if not). $e(X) = \mathbb{P}[Z = 1|X]$.

(b) Assume that Y is unconfounded and the probability of treatment depends only on whether the kidney stone is large or small. Write down and compute the IPW estimate for the difference in treatment effect from the two treatments.

Solution:

$$\begin{aligned}
\hat{\tau}_{\text{IPW}} &= \frac{1}{200}\sum_{i:Z_i=1}\frac{Y_i}{e(X_i)} - \frac{1}{200}\sum_{i:Z_i=0}\frac{Y_i}{e(X_i)} \\
&= \frac{1}{200}\left(\frac{18}{\hat{e}_s}+\frac{57}{\hat{e}_l}\right) - \frac{1}{200}\left(\frac{75}{1-\hat{e}_s}+\frac{5}{1-\hat{e}_l}\right) \\
&= \frac{1}{200}\left(99+\frac{513}{8}\right) - \frac{1}{200}\left(\frac{825}{9}+45\right) \\
&= \frac{1}{200}\left(99+64.125\right) - \frac{1}{200}\left(91\frac{2}{3}+45\right) \\
&\approx \frac{26.5}{200} = 0.1325,
\end{aligned}$$

where

$$\hat{e}_s = \frac{20}{20+90} = \frac{20}{110}$$

$$\hat{e}_l = \frac{80}{80+10} = \frac{80}{90}.$$

(c) Compute the difference-in-means estimate and compare it to the IPW estimate. How does the fact that these differ relate to Simpson's Paradox?

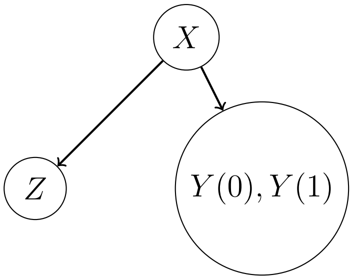
Solution:

$$\hat{\tau}_{\text{SDO}} = \frac{75}{100} - \frac{80}{100} = -0.05.$$

(d) Write down the unconfoundedness assumption as a conditional independence condition. Express this same assumption using a causal DAG.

Solution: The unconfoundedness assumption is

$$Z \perp\!\!\!\perp \{Y(1), Y(0)\} | X.$$



2. (a) Consider the linear structural model.

$$Y = \alpha + \tau Z + \beta X + \epsilon$$

$$Z = \alpha' + \gamma W + \eta X + \delta$$

We want to estimate the treatment effect τ of Z on Y using W as an instrumental variable. In order for W to be a valid instrumental variable, we need some assumptions on Y, Z, W, and X. Specify whether each of the quantities below **must be zero**($= 0$), **must be non-zero**($\neq 0$), or **does not matter**.

- Cov(W, Y)
- Cov(W, X)
- Cov(W, Z)
- Cov(W, ϵ)
- Cov(W, δ)

Solution:

- Does not matter.
- $= 0$. The instrumental variable W must be independent of the confounder X.
- $\neq 0$. The treatment Z must depend on the instrumental variable W.
- $= 0$. ϵ must be independent of W so that $\frac{\text{Cov}(Y,W)}{\text{Var}(W)}$ simplifies to $\tau\gamma$, i.e. we want Cov(W, ϵ) term when expanding Cov(W, ϵ) to be 0.
- $= 0$. δ must be independent of Z so that $\frac{\text{Cov}(Z,W)}{\text{Var}(W)}$ simplifies to γ, i.e. we want Cov(W, δ) term when expanding Cov(W, δ) to be 0.

(b) Assuming W is a valid instrumental variable, write down an estimate for τ. Your answer should involve the ratio of two covariances.

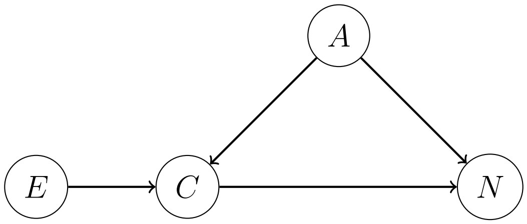
Solution:

$$\hat{\tau} = \frac{\text{Cov}(Y, W)}{\text{Cov}(Z, W)}$$

(c) If W was a vector instead of scalar, what could go wrong with the formula for estimating τ that you got in part b)? At a high-level, describe what you would do instead to estimate τ in this case.

Solution: The division in the scalar solution is ill-defined if W is a vector, since both numerator and denominator of that expression would be vectors. For this setting, we would have to employ 2SLS. 1) First get a prediction $\tilde{Z}$ from W and X using the second equation. 2) Use $\tilde{Z}$ in the first equation to get an estimate $\hat{\tau}$.

3. You are observing the Data 102 Halloween Party where trick-or-treaters are arriving and the GSIs are passing out candy to each one. Some of the trick-or-treaters are children and some of them are adults. Some are wearing more intricate (fancy) costumes than others. You notice that the GSIs give different trick-or-treaters different amounts of candy. You believe that the amount of candy given to a specific trick-or-treater is related to the intricacy of their costume, the effort put into their costume, and their age. You come up with the causal diagram below:



Age (A), Costume intricacy (C), and Artistic effort (E) are all binary random variables (0 =low, 1 =high). Number of candies received (N) is an integer. For this question, you should assume that the diagram above represents the true causal relationships, and that all relationships are linear.

Over the course of the party, you count 120 trick-or-treaters. You want to quantify the causal relationship between costume intricacy (C) and number of candies received (N).

(a) Can you use Artistic Effort as the instrumental variable when performing 2-stage least squares regression to predict the treatment effect of costume intricacy on Number of candies?

Solution: Yes, because it only affects the treatment (Costume intricacy) and is independent of the confounder (Age).

(b) Fill in the covariates for the Fill in the covariates for the linear structural model.

$$N = \alpha + \tau\_\_\_ + \beta\_\_\_ + \epsilon$$

$$C = \alpha' + \gamma\_\_\_\_ + \eta\_\_\_\_ + \delta.$$

If Cov$(C, E) = 5$, Cov$(N, E) = 10$, and Cov$(E, E) = 2$, what is the ATE?

Solution:

$$N = \alpha + \tau C + \beta A + \epsilon$$
$$C = \alpha' + \gamma E + \eta A + \delta,$$

and

$$\text{ATE} = \frac{\text{Cov}(N, E)}{\text{Cov}(C, E)} = 2. \tag{1}$$

(c) You decide to investigate the treatment effect of Costume intricacy on Number of candy pieces. Which of the following produce an unbiased estimate of the ATE?

i. $\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^{n} N_i C_i - \frac{1}{120-n_1}\sum_{i=1}^{n} N_i(1 - C_i)$, where n_1 is the number of students with $C_i = 1$.

ii. The slope of a line that uses simple linear regression to predict N from C.

iii. The coefficient β_C, where (β_A, β_C) are the result of running least squares on N with covariates A and C.

iv. $\mathbb{P}(N = 1|C)$.

Solution: Only option C provides an unbiased estimate of the ATE. Both options A and B fail to account for the confounding effect of A, while option D both fails to account for confounding and measure the target quantity. C works (assuming the true relationship is linear) because it controls for the confounder A.

Feedback Form

On a scale of 1-5, where 1 = much too slow and 5 = much too fast, how was the pace of the discussion section?

1 2 3 4 5

Which problem(s) did you find most useful?

Which were least useful?

Any other feedback?