Causal Inference Questions

Causal Inference

Question 1

Consider the following statement from Mayer-Schonberger and Cukier (2013) "Big Data: A Revolution That Will Transform How we Live, Work, and Think", page 14:

"Correlations may not tell us precisely *why* something is happening, but they alert us *that* it is happening. And in many situations this is good enough. If millions of electronic medical records reveal that cancer sufferers who take a certain combination of aspirin and orange juice see their disease go into remission, then the exact cause for the improvement in health may be less important than the fact that they lived. ... we can let the data speak for itself."

Do you agree with this statement? If so, why? If not, why not?

Question 2

Suppose we conduct a study of the the effect of 6 mg Dexamethasone daily versus placebo on a measure of lung function one week after admission to the hospital due to respiratory symptoms resulting from infection with SARS-CoV-2. Suppose we let Y denote lung function at the end of seven days, and D_j denote Dexamethasone treatment on day j of follow-up (e.g., $D_j = 1$ denotes treated with Dexamethasone on day j; $D_j = 0$ denotes not treated with Dexamethasone on day j). Please describe, in words, the effect that the following contrast of potential outcomes captures:

$$\psi = E(Y^{d_1=1,d_2=1,d_3=1,d_4=1,d_5=0,d_6=0,d_7=0}) - E(Y^{d_1=1,d_2=1,d_3=1,d_4=0,d_5=0,d_6=0,d_7=0})$$

Question 3

Please re-write the right-hand side of this causal effect:

$$\psi = E(Y^{d_1=1,d_2=1,d_3=1,d_4=1,d_5=0,d_6=0,d_7=0}) - E(Y^{d_1=1,d_2=1,d_3=1,d_4=0,d_5=0,d_6=0,d_7=0})$$

more compactly (instead of writing out the exposure value on each of the seven days).

Question 4

Consider the following statement from a paper by Athey et al 2020, page 14: "In the setting of interest we have data on an outcome Y_i, a set of pretreatment variables X_i and a binary treatment $W_i \in \{0, 1\}$. We postulate that there exists for each unit in the population two potential outcomes $Y_i(0)$ and $Y_i(1)$, with the observed outcome equal to corresponding to the potential outcome for the treatment received, $Y_i = Y_i(W_i)$."

What assumption(s) are the authors relying on when they say "We postulate that there exists ..."? Why?

Question 5

Suppose you had superpowers and were able to measure unobservable potential outcomes. Suppose you used these measures to fit a model that regresses the exposure A against all measured confounders C (i.e., propensity score model).

If you included the potential outcomes in a logistic regression model such as:

$$\text{logit}\{P(A = 1 \mid C, Y^a)\} = \beta_0 + \beta_1 C_1 + \ldots + \beta_p C_p + \theta Y^a$$

Can you determine the value of θ if exchangeability holds?

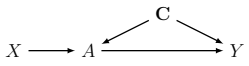
Can you determine the value of θ if exchangeability doesn't hold?

Question 6

Consider a two-arm placebo controlled randomized trial with four mutually exclusive strata labeled $S = 1, S = 2, S = 3$ and $S = 4$. Suppose that the treatment was not conditional on any other variables, and was assigned to: 20% of individuals in stratum $S = 1$; 30% of individuals in stratum $S = 2$; 15% of individuals in stratum $S = 3$; and 10% of individuals in stratum $S = 4$. If you assume no finite sample error, can you determine all of the propensity score values in the sample of individuals in the trial?

Question 7

Consider the setting in a double blind placebo controlled randomized trial with a treatment assignment indicator (X), a treatment complying variable (A), confounders (C) and an outcome Y



Using potential outcomes notation and clear notation for the values of X and A, write out a per protocol effect, defined as the effect that would be observed if everyone was assigned to treatment and took it, versus if everyone was assigned to placebo and took it.

Question 8

Consider the following regression model fit using the NHEFS data:

```
modelOR <- glm(wt_delta ~ qsmk + sex + age + income + sbp + dbp + price71 + tax71 + race,
               data=nhefs,
               family = binomial("logit"))
```

Supposed we used the S learner as outline in the notes on CATEs, which consists of predicting individual outcomes under `qsmk = 1` and `qsmk = 0` for each person, taking the difference between the predicted outcomes, and averaging these differences among subsets of the cohort defined by some of the variables in the model (e.g., race).

Would we expect to see that race is a modifier of the marginally standardized odds ratio? Why or why not?

Question 9

Write a contrast of potential outcomes for the following research questions. If it is not possible, explain why. If it is possible, explain potential threats to the validity of a causal inference :

(a) Should women > 50 years of age be screened regularly for breast cancer?

(b) Does air pollution kill citizens of Los Angeles?

(c) How much mortality was caused by the tobacco industry's misconduct?

(d) Does school type (private, public, charter) affect a child's later mortality risk?

(e) Does a depression treatment work better in individuals who exercise regularly, eat a diet rich in green leafy vegetables, and who have never smoked?

(f) Are parents more conservative than their children because they are older?

Question 10

Using the following table, compute:

a) The Average Treatment Effect

b) The Effect of Treatment on the Treated

c) The Effect of Treatment on the Untreated

d) The CATE for C1 = 0 and C2 = 1

e) The CATE for C1 = 1 and C2 = 1

ID	X	Y	Y1	Y0	C1	C2
1	1	10	10	15	1	1
2	1	5	5	9	1	0
3	1	16	16	18	1	1
4	1	3	3	5	1	0
5	1	6	6	9	0	1
6	1	8	8	8	0	0
7	1	12	12	11	0	1
8	1	8	8	10	0	0
9	0	5	8	5	1	1
10	0	7	12	7	1	0
11	0	8	8	8	1	1
12	0	10	12	10	1	0
13	0	3	7	3	0	1
14	0	5	9	5	0	0
15	0	5	5	5	0	1
16	0	2	5	2	0	0