Section 2 Assignment

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."

Other than the fact that they mix up singular and plural by stating that we should let the "data" (plural) speak for "itself" (singular) :-), describe in one paragraph (no longer than one half page) why this statement is problematic. Provide an example illustrating how their interpretation of the scenario may be erroneous.

Every stats instructor at one point or another in a semester says "correlation does not mean causation." To claim causation, the data and models must meet a certain combination of assumptions and must uphold a defined standard. The question being asked must be highly specified instead of a vauge "some amount of aspirin and OJ were associated with cancer remission." While it is good to be alerted that something is happening, with almost every data set if you slice and dice it enough you can find some sort of correlation. It breaks assumptions such as counter factual consistency which states that the potential outcome that would be observed if we *set* the exposure to the *observed* value is the observed outcome. If we set a specific level of daily aspirin and OJ consumption, we would most likely see no change in cancer remission. There is also most likely exchangability problems. It could be that cancer patients that are able to eat breakfast every morning are in general doing better than those who cannot consistently eat food. If we exchanged the OJ exposure with the group that did not drink OJ, because originally unexposed group's disease progression is worse they will most likely not see remission the same way that originally exposed group's disease progression did. These assumptions are important to claim causality, otherwise you may come to some erroneous conclusions.

Question 2) In randomized controlled trial settings, researchers are often interested in estimating *per protocol effects*. Consider a simple scenario with a randomization indicator R, with $R = 0$ denoting "assigned

to placebo" and $R = 1$ denoting "assigned to treated", an adherence indicator A, with $A = 0$ denoting "did not adhere" and $A = 1$ denoting "adhered by taking treatment on the day randomized", and an outcome variable Y, with $Y = 1$ denoting "event", and $Y = 0$ denoting "no event". Can you write the per protocol effect, defined as being assigned to treatment and adhering relative to being assigned to placebo and adhering, using potential outcomes notation? Write these effects on the risk difference, risk ratio, and odds ratio scales.

Rd

$$E(Y^{a=1,r=1} - Y^{a=1,r=0})$$

RR

$$\frac{E(Y^{a=1,r=1})}{E(Y^{a=1,r=0})}$$

OR

$$\frac{Odds(Y^{a=1,r=1})}{Odds(Y^{a=1,r=0})}$$

Question 3) 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:

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

The effect of 6mg of Dexamethasone daily for four days post admission and then three days placebo on lung function for the week after admission to the hospital due to respiratory symptoms resulting from infection with SARS-CoV-2 compared to 6mg of Dexamethasone daily for three days post admission and then four days of placebo

---

Question 4) Please re-write the right-hand side of the equation in Question 3 more compactly (instead of writing out the exposure value on each of the seven days).

$$\psi = E(Y^{\bar{d}_4=1,\underline{d}_5=0}) - E(Y^{\bar{d}_3=1,\underline{d}_4=0})$$

Table 1: Table Under SUTVA (n=6)

ID	Exposure (A)	Outcome (Y)	Y(a=1)	Y(a=0)
1	1	1	1	na
2	1	1	1	na
3	0	1	na	1
4	1	0	0	na
5	0	0	na	0
6	0	1	na	1

Question 5: Please complete the Table under SUTVA:

Question 6: Suppose the average treatment effect on the risk difference, risk ratio, and odds ratio scales for the relation between quitting smoking and high blood pressure is -0.129, 0.53, and 0.45, respectively. Suppose further that the identification assumptions required for interpreting these associations as causal effects holds. Please interpret these effect measures.

RD: There will be 13 fewer cases of high blood pressure out of every 100 people in our sample if everyone quit smoking relative to if everyone continued to smoke.

RR: There will be a 53% decrease in high blood pressure in our sample if everyone quit smoking compared to if everyone continued to smoke.

OR: The odds of high blood pressure is 55% lower in our sample if everyone quit smoking compared to the outcome if everyone continued to smoke.

Question 7: For the example of the relation between quitting smoking and high blood pressure, do you think the average treatment effect or the effect of treatment on the treated is more relevant? Explain why or why not.

The effect of treatment on treatment is more relevant because we can isolate the effect among those who do receive intervention thereby getting a more precise understanding of how the treatment works among those who actually recieved treatment. It is able to isolate the beneficial effect.

The average treatment effect is less relevant because it averages out both of the beneficial and non beneficial effects in the entire population. Overall, that will yield a non-beneficial effect. Among the target population

of those who smoke, it could be that the treatment was overall beneficial for their specific group, whereas it may not be for the whole population. It would be beneficial to use the ETT as it would condition on the treated population vs the whole population.

Question 8: Again, for the example of the relation between quitting smoking and high blood pressure, can you describe a scenario where we may collect some data and where the no interference assumption would be violated?

In a study where people quit smoking, if someone was assigned to quit smoking but someone in their household was assigned to stay smoking, there could be an on blood pressure due to second hand smoke inhaled by the person assigned to quit smoking and the no inference assumption would be violated. The potential outcome would be different based on the exposure status of another subject.

Question 9: Consider the following statement from a paper by Athey et al (2020)[https://arxiv.org/pdf/1909.02210.pdf], 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?

The authors rely on counter factual consistency and no interference assumption (together = SUTVA!). In order to have the potential outcome equal to observed outcome (as stated in that phrase), both of these assumptions (counter factual consistency and no interference) must be met. Counter factual consistency assumption states that the potential outcome that would be observed the exposure was set to the observed value is the observed outcome. No interference assumption is when the potential outcome for any given individual does not depend on the exposure status of another individual. Both of the definitions of the assumptions allow it so that for each unit in the population, there are two potential outcomes and that the observed outcome corresponds to the potential outcome for the received treatment.

Note: The authors do not rely on exchangability assumption in that statement because they do not mention independence of potential outcome from observed exposure explicitly.

Question 10: Consider the exchangeability assumption. Why is the word "exchangeable" used to describe this concept? What, precisely, is being exchanged?

Exchangeability implies that the potential outcome of the exposure Yˆx is independent of the observed exposure. Essentially, if you exchange the exposure between the two groups, the potential outcomes will still be the same if the exchangeability assumption is met. The different groups are a good representation and are able to be exchanged for one another and get the same potential outcome if exposed.

Question 11: Consider a regression model with an exposure and 11 confounders, for a total of 12 variables:

$$E(Y \mid X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \ldots + \beta_{12} C_{11}$$

What is the total number of possible interactions in this model? What are the total number of 2-way interactions? Show your reasoning.

We can have up to 4083 k-way interactions.

$$2^{12} - 12 - 1 = 4083$$

Mathematically, there are 66 two way interactions based on the binomial coefficient.

$$\binom{12}{2} = \frac{12!}{2!(12-2)!} = 66$$

Question 12: Suppose you had superpowers and were able to measure 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), and that there was no measured confounding, selection bias, and information bias (i.e., exchageability was met). If you included the potential outcomes in the regression model:

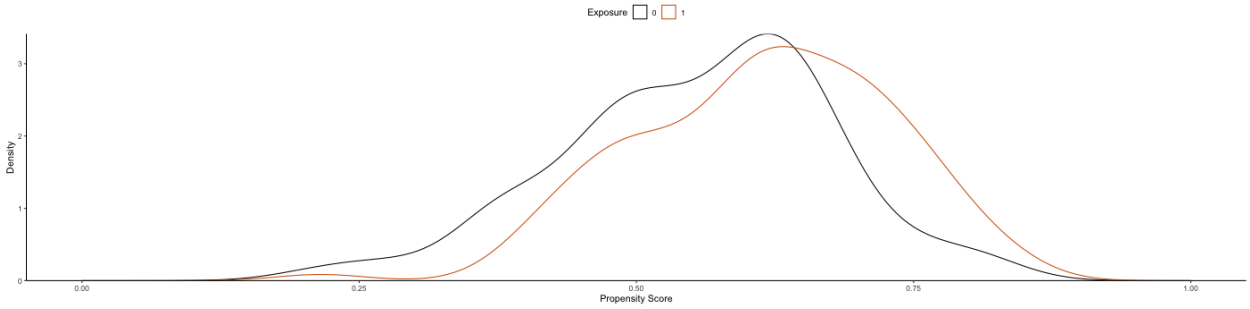
$$logit\{P(A = 1 \mid C, Y^a)\} = \beta_0 + \beta_1 C_1 + \ldots + \beta_p C_p + \theta Y^a$$

Can you determine from this information alone what the value of θ is if exchangeability holds? Can you determine what the value of θ is if exchangeability doesn't hold?

Exchangability implies that under a specific exposure, the independent outcomes are independent of the observed exposure. In this equation, if exchangability holds theta = 0 wont be associated with the observed treatment (and is thereby independent of the observed exposure A). If exchangability does not hold, then theta is a non-zero value because there will be a relationship between the observed exposure and the potential outcome.

Question 13: Install the AIPW package from CRAN and load the library. Then load the "eager_sim_obs" dataset using the `data("eager_sim_obs")` command. Using these data, generate (i) a propensity score over lap plot, (ii) a list of the five largest and smallest propensity scores, and (iii) the summary distribution of the stabilized inverse probability weights using a propensity score model that adjusts for `eligibility`, `loss_num`, `age`, `time_try_pregnant`, `BMI`, and `meanAP`. Is positivity violated in these data? Why or why not?

Positivity is when there are exposed and unexposed individuals at all confounding levels. Nonpositivity is not good because those who were unexposed in the sample are unlikely to be exposed and vise versa. When this occurs, it doesn't make sense in most cases to estimate the average treatment effect, since a subset of the population who may never realistically be exposed (or unexposed).



When looking at the graph, positivity is not violated. This is because the mass of density for the exposed (orange) occurs in the same place as the density mass for the unexposed (black).

The top 5 propensity scores are 0.8417561 0.8305905 0.8250527 0.8147493 0.8094557.

The bottom 5 scores are 0.2136926 0.2155983 0.2607334 0.2923699 0.3579733

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.5214 0.8209 0.9507 1.0003 1.1231 2.7366

The mean of the stabilized weights is essentially 1. The max weight is not large relative to the mean and min. This suggests that the weights are "well-behaved." Thus, in this particular case, we are not concerned with violations of the positivity assumption.

Question 14 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 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 $S = 4$. Can you determine all of the propensity score values in the sample of individuals in the trial?

Propensity scores are defined as the probability of receiving treatment, usually used for specific strata characteristics. Using these probabilities given in the question for the four mutually exclusive strata (S=1 20%, S=2 30%, S=3 15%, & S=4 10%) in the two-arm placebo controlled randomized trial, researchers can determine all of the propensity score values in the sample of individuals.

(Bonus?) Question 15 Using the information provided in Question 14, please write a logistic regression equation that the defines the propensity score for this randomized trial. What are the parameter values in this logistic regression model?

P(X|S1,S2,S3,S4) = B0 + (B1*S1)* + *(B2*S2) + (B3*S3)* + *(B4*S4)

where S1= .20 S2=.30 S3=.15 S4=.10