6.S897/HST.S53
Problem Set 5: Causal Inference 1
Repo: https://github.com/mlhc19mit/psets/tree/master/pset5
Due: April 9, 2019 at 11:59pm by midnight through Stellar

Section 0: Terminology

For every question, we assume the intervention is a binary treatment t where t=0 indicates no treatment and t=1 indicates treatment.

Helpful reminders:

- $\mathbf{Y_1}$ the value of Y in a world where the treatment is given.
- $\mathbf{E[Y_0]}$ the average value of Y in a world where the treatment is never given (over *all* patients, including to those who were actually treated).
- **E[Y|t=0]** the average value of Y among patients who were not treated *in our observed dataset*.
- $\mathbf{Y_1 - Y_0}$ the "value-added" of the intervention (i.e. how much giving the treatment increased Y above however much it would have increased without getting treated).

Q1. Do You Like Apples? *(5 points)*
(adapted from problem by Uri Shalit)

You are walking through the halls of Building 2, and you see some confused mathematicians staring at a chalkboard. After minutes of scratching their head, they give up and walk away dejectedly. You decide to wander over to the board to see for yourself what mysterious problem is so challenging. It reads:

> Give an example of a joint distribution (in the form of a table) which includes a hidden confounder H, a binary treatment t, and two potential outcomes Y_0 and Y_1 such that
>
> 1. Ignorability does *not* hold and
> 2. $E[Y_1 - Y_0] \neq E[Y|t=1] - E[Y|t=0]$,
>   where $Y = t * Y_1 + (1-t) * Y_0$.

Specify probability distributions $P(H)$, $P(T|H)$, $P(Y_0|H)$, and $P(Y_1|H)$ which satisfy the criteria above. Show that it satisfies each criteria.

Q2: Who Says You Can't Teach an Old Doc New Tricks? *(8 points)*
(adapted from problem by Uri Shalit and Rom Gutman)

As you scribble the answer on the board, Professor Sontag notices you and pulls you aside. He is very impressed that you solved that problem, and he invites you to begin collaborating with him to solve as many causal inference problems as you can. You begin with data about the effectiveness of a surgical fellowship from Sunnydale Hospital.

The data comes from an experiment run from 2004-2008 at Sunnydale Hospital. In the data (See Table 1), you have binary indicators for: prior education (Z), whether the surgeon had ≥ 100 successful surgeries in 2004 (X), whether the surgeon enrolled in the fellowship from 2005-2006 (T), whether the surgeon had ≥ 150 successful surgeries in 2007 (Y), and whether the surgeon cumulatively had ≥ 500 successful surgeries in their lifetime by 2008 (W).

We know the following:

(i) The number of successful surgeries in 2004 depends solely on the prior education.

(ii) Whether a doctor is selected to the surgery fellowship program depends on their prior education and number of successful surgeries in 2004.

(iii) The number of successful surgeries in 2007 depends on the fellowship, prior education, and number of successful surgeries in 2004.

(iv) The cumulative number of successful surgeries by 2008 is directly based on how many successful surgeries a doctor performs in 2004 and 2007.

Your task is as follows:

1. Draw the causal graph that describes the above experiment.
2. Calculate the Average Treatment Effect (ATE) of the fellowship (T) on 2007 success (Y). Use covariate adjustment and empirically estimate the probabilities/expectations from the observed data. Recall that

   $$\text{ATE} = E[Y_1 - Y_0] = E[Y_1] - E[Y_0]$$

   where (for this causal graph)

   $$E[Y_t] = \sum_z \sum_x P(Z=z) * P(X=x \mid Z=z) * E[Y \mid X=x, Z=z, T=t]$$

3. Calculate the Conditional Average Treatment Effect (CATE) of the fellowship (T) on 2007 success (Y) for patients without prior education (Z=0). Recall that

$$\text{CATE} = E[Y_1 - Y_0 \mid Z=0] = E[Y_1 \mid Z=0] - E[Y_0 \mid Z=0]$$

where (for this causal graph)

$$E[Y_t \mid Z=0] = \sum_x P(X=x \mid Z=0) * E[Y \mid X=x, Z=0, T=t]$$

Z	X	T	Y	W
0	1	1	1	0
0	0	1	1	0
0	0	1	1	0
1	1	1	1	1
1	0	1	1	1
1	0	1	1	1
1	1	1	0	0
1	0	1	0	0
0	0	1	0	0
1	1	0	1	1
1	0	0	1	0
1	0	0	1	0
1	1	0	0	1
1	1	0	0	1
0	1	0	0	0
1	0	0	0	0
0	0	0	0	0
0	0	0	0	0

Table 1: Table1

For the rest of the pset, you will use the low birth weight (lbw) data. This is publicly available data of linked birth and infant death data collected and maintained by the CDC (see repo for csv files to use for this pset). Each line represents an infant, and has data about the infant's date of birth and death if he/she died within a year. It also has information about the mother, her education, race, smoking habits, alcohol consumption and other as well as some information about the father.

The following table is intended to clarify the differences between the following two exercises.

Question	Treatment	Outcome	Dataset
Q3	Birth weight under 2700g? (*binary*)	1-Year Mortality? (*binary*)	*twins.gz*
Q4	Smoked during pregnancy? (*binary*)	Low birth weight? (*real*)	*singletons.gz*

Q3. Low Birth Weight Causes Infant Mortality? *(10 points)*
(adapted from problem by Maggie Makar)

You are doing so well at causal inference that Prof Sontag decides you are ready to meet some of his clinical collaborators. He introduces you to Dr. Buffy Summers at Sunnydale Hospital. She wants to know whether low birth weight (i.e. < 2700g) causes infant mortality, but she really can't figure out all of the jargon like "propensity" or "covariate." She hopes that you will be able to explain things in a more intuitive way.

Fortunately for you, you notice that the data has a section for twin babies (*see twins.gz*). Sometimes for a given pair of twins (where the environmental factors are the same), one baby is born below 2700g while the other is born above. "This will be great for explaining it to Dr. Summers without any fancy math!", you excitedly say to yourself.

a) Why would this dataset be so well-suited for to studying the counterfactual effect of low birth weight on infant mortality? Explain in 1-3 sentences.

b) Filter the cohort so that you only have pairs where exactly one of them is below 2700g. Use these "counterfactual" pairs to estimate the ATE of low birth weight on the one year mortality rate for twins. Recall that

$$CATE(x) = y_1 - y_0$$

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} CATE(x_i)$$

Where t=0 indicates birth weight >= 2700g and t=1 indicates birthweight < 2700g.

c) Dr. Summers likes your explanation. She then asks whether this ATE generalizes to the whole population, including singletons? In other words, can we assume that the ATE of lbw in the singletons population is roughly the same? Justify why or why not. Hint: compute the mortality rates among the "counterfactual" twin pairs and among the singleton population.

Q4: Smoking During Pregnancy Causes Low Birth Weight? *(15 points)*
(adapted from problem by Maggie Makar)

Dr. Summers appreciates the simplified description, but she wants to focus on methods that can be applied to the full population of babies. She suggests that you shift focus and instead look at whether smoking during pregnancy causes low birth weight. She provides a large dataset with many factors about the mother and father (*see singletons.gz*).

Covariates: ['dmage', 'dmar', 'dlivord', 'anemia', 'cardiac', 'lung', 'diabetes', 'herpes', 'hydra', 'hemo', 'chyper', 'phyper', 'eclamp', 'incervix', 'pre4000', 'preterm', 'renal', 'rh', 'uterine', 'othermr', 'alcohol', 'm_race_black', 'm_race_other', 'm_race_white', 'm_edu_college', 'm_edu_elementary', 'm_edu_highschool', 'm_edu_morethancollege', 'm_edu_noedu']
Treatment: 'tobacco'
Outcome: 'dbirwt'

a) Dr. Summers asks you to start simple. Split the dataset into those who were treated (i.e. babies whose mothers smoked during pregnancy) and those who weren't. Compute the average birth weight of each cohort, and report the difference between these two average birth weights. Justify why, in general, this naive approach won't allow you to reason about the causal effects of the treatment.

b) Do a covariate adjustment to estimate the effect of mother's smoking habits on the baby's birth weight. Use sklearn's LinearRegression (with default hyperparameters) to fit a linear model, f, to predict the potential outcome given the covariates and treatment:

$$f(x,t) \approx \mathbb{E}[Y_t | T = t, x]$$

Report the estimated ATE using this method. In 1 sentence, describe the relationship between the ATE (as calculated using this method) and the coefficient for the treatment variable that is learned by the linear model. Recall that

$$\widehat{ATE} = \frac{1}{n}\sum_{i=1}^{n} f(x_i, 1) - f(x_i, 0)$$

c) An alternative to covariate adjustment is propensity score weighting. Use sklearn's LogisticRegression (with default hyperparameters) to fit a linear model to predict the the treatment given the covariates. The probability of getting treatment t is known as the propensity score:

$$\hat{p}(T = t|x)$$

Report the estimated ATE using this method. State in 1 sentence what this ATE means to someone considering smoking during pregnancy. Recall that

$$A\hat{T}E = \frac{1}{n_1} \sum_{i \text{ s.t. } t_i = 1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n_0} \sum_{i \text{ s.t. } t_i = 0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

where n_T are the effective counts of the re-weighted cohorts:

$$n_T = \sum_{i \text{ s.t. } t_i = T} \frac{1}{\hat{p}(t_i = T|x_i)}$$

Dr. Summers thanks you for your hard work. She loves the story of how you and Professor Sontag began working together, so she asks you to write up the full story (i.e. the answers to these four questions) in one single writeup so that she can write a screenplay about it.

Q5: Not graded by please answer *(0 points)*

a) How many hours did you spend on this problem set? This answer is very useful feedback for whether psets were too short, too long, or just right. Future students of 6.S897/HST.S53 will be grateful for your response.