Marie-Laure Charpignon
mcharpig
Problem Set 5
Slack days: I am using my last slack day of the semester.

Question 1:

Proposed table illustrating the presence of a hidden confounder (H).
Note that Y = t*Y1 + (1-t)*Y0.

T	Y1	Y0	Y	H
1	1	0	1	1
0	1	0	0	0
1	0	1	0	0
1	1	0	1	1
0	0	1	1	0
1	1	0	1	0

- We have:
  - $E(Y|t=1) = 3/4$
  - $E(Y|t=0) = 1/2$

  Which leads to:

  - $E(Y|t=1) - E(Y|t=0) = 9/12 - 6/12 = 3/12 = 1/4$
  - $E(Y1-Y0) = 2/6 = 1/3$

  The quantity $[E(Y|t=1) - E(Y|t=0)]$ differs from the quantity $[E(Y1-Y0)]$.
  Thus, the equality is not satisfied (second criterion).

- Additionally:
  - $P(H=1) = 2/6 = 1/3$
  - $P(H=0) = 4/6 = 2/3$

  - $P(T=1|H=1) = 2/2 = 1$
  - $P(T=0|H=1) = 0/2 = 0$
  - $P(T=1|H=0) = 2/4 = 1/2$
  - $P(T=0|H=0) = 2/4 = 1/2$

  - $P(Y1=1|H=1) = 2/2 = 1$
  - $P(Y1=0|H=1) = 0/2 = 0$

  - $P(Y0=1|H=0) = 1/2$
  - $P(Y0=0|H=0) = 1/2$

  - $P(Y1=1|H=0) = 1/2$
  - $P(Y1=0|H=0) = 1/2$

- We have: **P(Y0 = 1| T=1) = 1/4, while P(Y0 = 1 | T=0) = 1/2. Hence Y0 is dependent on T. Thus, ignorability does not hold (first criterion).**
- Also, note that P((Y0,Y1) = (0,1)) = 4/6 = 2/3,
  but P((Y0,Y1) = (0,1) | T=0) = 1/2 and P((Y0,Y1) = (0,1) | T=1) = 3/4.
  So the joint distribution (Y0,Y1) is not independent of T.
- But let us prove that: conditioned on H, (Y0,Y1) is independent of T.

P((Y0,Y1) = (0,1), T=1 | H=1) = 2/2 = 2/2 * 2/2 = P((Y0,Y1) = (0,1) | H=1) * P(T=1 | H=1)
P((Y0,Y1) = (1,1), T=1 | H=1) = 0/2 = 0/2 * 2/2 = P((Y0,Y1) = (1,1) | H=1) * P(T=1 | H=1)
P((Y0,Y1) = (1,1), T=0 | H=1) = 0/2 = 0/2 * 0/2 = P((Y0,Y1) = (1,1) | H=1) * P(T=0 | H=1)
P((Y0,Y1) = (0,1), T=0 | H=1) = 0/2 = 2/2 * 0/2 = P((Y0,Y1) = (0,1) | H=1) * P(T=0 | H=1)
P((Y0,Y1) = (1,0), T=0 | H=1) = 0/2 = 0/2 * 0/2 = P((Y0,Y1) = (1,0) | H=1) * P(T=0 | H=1)
P((Y0,Y1) = (1,0), T=1 | H=1) = 0/2 = 0/2 * 2/2 = P((Y0,Y1) = (1,0) | H=1) * P(T=1 | H=1)
P((Y0,Y1) = (0,0), T=1 | H=1) = 0/2 = 0/2 * 2/2 = P((Y0,Y1) = (0,0) | H=1) * P(T=1 | H=1)
P((Y0,Y1) = (0,0), T=0 | H=1) = 0/2 = 0/2 * 0/2 = P((Y0,Y1) = (0,0) | H=1) * P(T=0 | H=1)

P((Y0,Y1) = (0,1), T=1 | H=0) = 1/4 = 1/2 * 1/2 = P((Y0,Y1) = (0,1) | H=0) * P(T=1 | H=0)
P((Y0,Y1) = (1,1), T=1 | H=0) = 0/4 = 0/4 * 1/2 = P((Y0,Y1) = (1,1) | H=0) * P(T=1 | H=0)
P((Y0,Y1) = (1,1), T=0 | H=0) = 0/4 = 0/4 * 1/2 = P((Y0,Y1) = (1,1) | H=0) * P(T=0 | H=0)
P((Y0,Y1) = (0,1), T=0 | H=0) = 1/4 = 1/2 * 1/2 = P((Y0,Y1) = (0,1) | H=0) * P(T=0 | H=0)
P((Y0,Y1) = (1,0), T=0 | H=0) = 1/4 = 1/2 * 1/2 = P((Y0,Y1) = (1,0) | H=0) * P(T=0 | H=0)
P((Y0,Y1) = (1,0), T=1 | H=0) = 1/4 = 1/2 * 1/2 = P((Y0,Y1) = (1,0) | H=0) * P(T=1 | H=0)
P((Y0,Y1) = (0,0), T=1 | H=0) = 0/4 = 0/4 * 1/2 = P((Y0,Y1) = (0,0), T=1 | H=0) * P(T=1 | H=0)
P((Y0,Y1) = (0,0), T=0 | H=0) = 0/4 = 0/4 * 1/2 = P((Y0,Y1) = (0,0), T=0 | H=0) * P(T=0 | H=0)

Question 2

1) **Causal graph describing the experiment.**

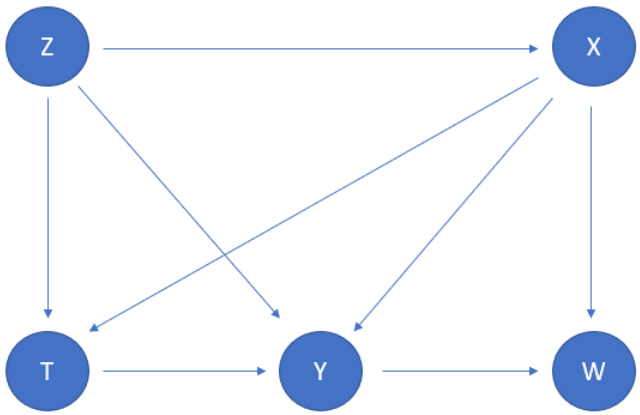

Table representing the experiment.

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

- P(Z=0) = 7/18
- P(Z=1) = 11/18

- P(X=0|Z=0) = 5/7
- P(X=1|Z=0) = 2/7
- P(X=0|Z=1) = 6/11
- P(X=1|Z=1) = 5/11
- E(Y|X=0, Z=0, T=0) = 0/2 = 0
- E(Y|X=0, Z=0, T=1) = 2/3
- E(Y|X=0, Z=1, T=0) = 2/3
- E(Y|X=0, Z=1, T=1) = 2/3
- E(Y|X=1, Z=1, T=1) = 1/2
- E(Y|X=1, Z=0, T=1) = 1/1 = 1
- E(Y|X=1, Z=0, T=0) = 0/1 = 0
- E(Y|X=1, Z=1, T=0) = 1/3

- **For Z=0 and X=0**
  - Contribution to E(Y1)
    - Formula: P(Z=0) * P(X=0|Z=0) * E(Y|X=0, Z=0, T=1)
    - Result: 7/18 * 5/7 * 2/3 = 5/18 * 2/3 = 5/27
  - Contribution to E(Y0)
    - Formula: P(Z=0) * P(X=0|Z=0) * E(Y|X=0, Z=0, T=0)
    - Result: 7/18 * 5/7 * 0 = 0

- **For Z=0 and X=1**
  - Contribution to E(Y1)
    - Formula: P(Z=0) * P(X=1|Z=0) * E(Y|X=1, Z=0, T=1)
    - Result: 7/18 * 2/7 * 1 = 1/9
  - Contribution to E(Y0)
    - Formula: P(Z=0) * P(X=1|Z=0) * E(Y|X=1, Z=0, T=0)
    - Result: 7/18 * 2/7 * 0 = 0
- **For Z=1 and X=0**
  - Contribution to E(Y1)
    - Formula: P(Z=1) * P(X=0|Z=1) * E(Y|X=0, Z=1, T=1)
    - Result: 11/18 * 6/11 * 2/3 = 2/9
  - Contribution to E(Y0)
    - Formula: P(Z=1) * P(X=0|Z=1) * E(Y|X=0, Z=1, T=0)
    - Result: 11/18 * 6/11 * 2/3 = 2/9
- **For Z=1 and X=1**
  - Contribution to E(Y1)
    - Formula: P(Z=1) * P(X=1|Z=1) * E(Y|X=1, Z=1, T=1)
    - Result: 11/18 * 5/11 * 1/2 = 5/36
  - Contribution to E(Y0)
    - Formula: P(Z=1) * P(X=1|Z=1) * E(Y|X=1, Z=1, T=0)
    - Result: 11/18 * 5/11 * 1/3 = 5/54

2) Average Treatment Effect (ATE)

- Hence: E(Y1) = 5/27 + 1/9 + 2/9 + 5/36 = 71/108
- Thus: E(Y0) = 0 + 0 + 2/9 + 5/54 = 17/54
- **Conclusion: ATE** = E(Y1) – E(Y0) = 71/108 – 17/54 = (71 – 34)/108 **= 37/108**

3) Conditional Average Treatment Effect (CATE)

Condition: filtering for patients without prior education (Z=0)

- **For X=1**
  - Contribution to E(Y1|Z=0)
    - Formula: P(X=1|Z=0) * E(Y|X=1, Z=0, T=1)
    - Result: 2/7 * 1 = 2/7
  - Contribution to E(Y0|Z=0)
    - Formula: P(X=1|Z=0) * E(Y|X=1, Z=0, T=0)
    - Result: 2/7 * 0 = 0
- **For X=0**
  - Contribution to E(Y1|Z=0)
    - Formula: P(X=0|Z=0) * E(Y|X=0, Z=0, T=1)
    - Result: 5/7 * 2/3 = 10/21

- Contribution to E(Y0|Z=0)
  - Formula: P(X=0|Z=0) * E(Y|X=0, Z=0, T=0)
  - Result: 5/7 * 0 = 0

- Hence: E(Y1|Z=0) = 2/7 + 10/21 = 16/21
- Hence: E(Y0|Z=0) = 0 + 0 = 0
- **Conclusion: CATE** = E(Y1|Z=0) – E(Y0|Z=0) = 16/21 – 0 **= 16/21**

Question 3a:

Propensity score matching is about finding out "long last twins" that just differ by one characteristic or parameter (e.g. the treatment or the condition), and to compare their outcomes (e.g. mortality or a specific complication). Identical (monozygotic or developing from one embryo) or fraternal (dizygotic, developing from two embryos) twins are well-suited since they share DNA and development conditions in the mother's womb. As they were born, the babies also get raised in the same environment and receive similar care. A statistician would say that they share the same set of covariates. If it happens that a pair of twins has one baby born below 2700g while the other is born above, then it is an ideal scenario to study the counterfactual effect of low birth weight on infant mortality because the two cases have everything identical but birth weight. I would like to think of this experiment as a "natural experiment" that meets the criteria of Randomized Clinical Trials (RCTs).

Question 3b:

Using the twin counterfactual pairs, the ATE of low birth weight on one-year mortality rate is estimate to **0.03272** i.e. that babies with birth weight lower than 2700g are **3.3%** more likely to die within a year than their twin born with a normal weight – all other things being equal. This confirms the increased possibility of potential complications suggested by the literature, as the birth weight of the baby decreases. Premature birth and low birth weights, especially at levels twice as low as 2700g, can indeed increase the risk of several health-related issues, such as vision, hearing loss as well as mental disabilities.

Question 3c:

- The mortality rate among eligible counterfactual twin pairs is quite different from the one in the singleton population (**0.0346 vs. 0.00505, i.e. more than 6 times larger**). More specifically, the mortality rate among counterfactual twin babies whose birth weight is below threshold is **0.0380** vs. **0.00525** for the twin babies with normal birth weight, which is closer to the ones of singletons. Because of this discrepancy between these two subsets of the larger population (twin babies have higher mortality rates than singletons), it may be difficult to generalize the results of the causal inference analysis we have just conducted on twins to the broader population. While it might be true that lower birth weight induces higher infant mortality rate, we would not be able to make any firm statement on the magnitude of this effect in the general population.

  On the next page, you will find a summary table of mortality rate per group.

Group	Mortality rate
Counterfactual twin pairs	3.46%
Counterfactual twin babies (normal weight)	0.525%
Counterfactual twin babies (low weight)	3.80%
Singletons	0.505%

- A potential explanation for the discrepancy (and a valid clinical reason not to draw a general statement) is that it is very common for twins to be born at a low birth weight. More than half of twins are born weighing less than 2700g, while the average birth weight of a healthy baby should be around 3000-4000g. This is largely due to the fact that twins are typically born premature (at a higher rate than singletons). Being born premature affects both singleton and twin babies' birth weights, but the prevalence of prematurity is higher among twins than among singletons.

Question 4a:

The average birth weight among non-smoking mothers is **3500.2g**, while the average birth weight among babies born from smoking mothers is **3335.3g**. This, the naïve approach consisting in comparing the two averages leads to a birth weight difference of **164.8g**. Without any adjustments, babies born from smoking mothers would be said to weigh approximately **164g** less than babies born from non-smoking mothers, on average. But the reality is that many other factors may influence a baby's birth weight, beyond the fact that the mother may be smoking. There are many other environmental factors, including other family conditions, that may influence the development of the fetus in the mother's womb. These factors may positively or negatively correlate with tobacco consumption, as well as interact. In order to get a more accurate estimate of the difference that can be linked to tobacco consumption, one needs to consider such factors or variables called confounders. This naïve approach consisting in computing the difference between these two average birth weights, in the presence of confounding variables that were not accounted for, would certainly not allow us to reason about the causal effects of the treatment.

Question 4b:

The estimated ATE using covariate adjustment is: **-188.4**. An appealing characteristic of covariate adjustment via linear regression is that the Average Treatment Effect (ATE) matches exactly with the coefficient for the treatment variable ('tobacco' in our case) learned by the linear model. With covariate adjustment, we found that babies born from smoking women weigh approximately **188g** less at birth than babies born from non-smoking mothers, on average.

Question 4c:

Using propensity score weighting, the estimated ATE is: **-407.11**. This estimate is actually more than twice lower than the values obtained with both the naïve approach and the covariate adjustment approach using linear regression. Adjusting for other confounders, tobacco has a negative impact on birth weight. The causal effect of tobacco itself is to reduce the baby's birth weight by more than **400g**. Lower birth weight may lead to health and mental complications, sometimes leading to infant mortality. This analysis highlights the dangers of considering smoking during pregnancy or provides compelling arguments to encourage mothers to avoid doing so.

Question 5: I spent approximately 12 hours on this problem set.