Homework 3

Advanced Methods for Data Analysis (36-402)

Due Friday February 11, 2022 *at 3:00 pm*

Solutions – not to be posted online or shared, even after the end of the semester.

You should **always show all your work** and submit both a writeup and *R* code.

- Assignments must be submitted through Gradescope as a PDF. Follow the instructions here: `https://www.cmu.edu/teaching/gradescope/`
- Gradescope will ask you to mark which parts of your submission correspond to each homework problem. This is mandatory; if you do not, grading will be slowed down, and your assignment will be penalized.
- Make sure your work is legible in Gradescope. You may not receive credit for work the TAs cannot read. **Note:** If you submit a PDF with pages much larger than 8.5×11", they will be blurry and unreadable in Gradescope.
- For questions involving R code, we strongly recommend using R Markdown. The relevant code should be included with each question, rather than in an appendix. A template Rmd file is provided on Canvas.

1. **Estimating Causal Effect by Random Assignment to Treatment.** Suppose that $X \in \mathbb{R}$ is the treatment variable, and $Y \in \mathbb{R}$ is the outcome variable. The regression function, which measures association, is $r(x) = \mathbb{E}(Y|X = x)$. Prove that when treatment X is randomly assigned (i.e., X is independent of $C(x)$ for all x), then
$$\theta(x) = r(x),$$
where $\theta(x)$ denotes the **causal regression function** (defined in class and in Chapter 16 of *All of Statistics* by Wasserman.)

Hint: During lecture, we went through a similar proof but for binary treatments (Theorem 16.3 in All of Statistics).

Solution: Assume that X is independent of $C(x)$ for all x. Then

$$\begin{aligned}
r(x) &= \mathbb{E}[Y \mid X = x] && \text{by definition} \\
&= \mathbb{E}[C(X) \mid X = x] && \text{because } Y = C(X) \\
&= \mathbb{E}[C(x) \mid X = x] && \text{by conditioning} \\
&= \mathbb{E}[C(x)] && \text{because } C(x) \text{ and } X \text{ are independent} \\
&= \theta(x) && \text{by definition of } \theta(x).
\end{aligned}$$

2. **Association and Causation.** In Lecture 2B, we used Toy Example 2 to demonstrate that the association α and average treatment effect θ need not be equal in the population. In the example, if the table represented the entire population, $\theta \neq \alpha$.

   Create a similar example relating a binary treatment X, binary outcome Y, and counterfactuals $C(0)$ and $C(1)$. Make a table with numbers for a population of 10 individuals so that $\alpha > 0$ **and** $\theta < 0$. Include a calculation of α and θ using the data in your table.

   Hint: See the R Markdown cheat sheet (right column) for how to format a table in R Markdown.

Solution: Here's one possible table, though there can be many:

X	$C(0)$	$C(1)$	Y
0	0	0	0
0	1	0	1
0	1	0	1
0	1	0	1
0	1	0	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

In this example,

$$\begin{aligned}
\alpha &= \mathbb{E}[Y \mid X = 1] - \mathbb{E}[Y \mid X = 0] \\
&= 1 - \frac{4}{5} \\
&= \frac{1}{5} \\
\theta &= \mathbb{E}[C(1)] - \mathbb{E}[C(0)] \\
&= \frac{1}{2} - \frac{9}{10} \\
&= -\frac{2}{5}.
\end{aligned}$$

3. **Review of Conditional Expectation.** Suppose Z is a random variable that is either 0 or 1, with probability 0.5 of either outcome. If $Z = 0$, then $X \sim \text{Uniform}(0, 1)$; if $Z = 1$, then $X \sim \text{Uniform}(3, 4)$.

   (a) Find $\mathbb{E}[X]$.

Solution: By the law of total expectation,

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \mid Z]] \\
&= \mathbb{E}[X \mid Z=0]\Pr(Z=0) + \mathbb{E}[X \mid Z=1]\Pr(Z=1) \\
&= \frac{1}{2} \times \frac{1}{2} + \frac{7}{2} \times \frac{1}{2} \\
&= \frac{1}{4} + \frac{7}{4} \\
&= 2.
\end{aligned}$$

(b) Find $\mathrm{Var}(X)$.

Solution: By the law of total variance,

$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Z)] + \mathrm{Var}(\mathbb{E}[X \mid Z]).$$

First, it's useful to remember that the mean of a Uniform(a, b) is $(a+b)/2$ and the variance is $(b-a)^2/12$. Let's calculate each term of the sum using these results. The first term uses the law of total expectation:

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}(X \mid Z)] &= \mathrm{Var}(X \mid Z=0)\Pr(Z=0) + \mathrm{Var}(X \mid Z=1)\Pr(X=1) \\
&= \frac{1}{12} \times \frac{1}{2} + \frac{1}{12} \times \frac{1}{2} \\
&= \frac{1}{12}.
\end{aligned}$$

The second term:

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[X \mid Z]) &= \mathbb{E}\left[(\mathbb{E}[X \mid Z] - \mathbb{E}[\mathbb{E}[X \mid Z]])^2\right] && \text{def. of variance} \\
&= \mathbb{E}\left[(\mathbb{E}[X \mid Z] - 2)^2\right] && \text{from (a)} \\
&= (\mathbb{E}[X \mid Z=0] - 2)^2 \Pr(Z=0) \\
&\quad + (\mathbb{E}[X \mid Z=1] - 2)^2 \Pr(Z=1) && \text{law of total expectation} \\
&= \left(\frac{1}{2} - 2\right)^2 \times \frac{1}{2} + \left(\frac{7}{2} - 2\right)^2 \times \frac{1}{2} && \text{from (a)} \\
&= \frac{9}{4} \times \frac{1}{2} + \frac{9}{4} \times \frac{1}{2} \\
&= \frac{9}{4}.
\end{aligned}$$

Adding the two terms up, we get

$$\mathrm{Var}(X) = \frac{1}{12} + \frac{9}{4} = \frac{7}{3}.$$

4. **The Omitted Variables Effect for Categorical Data.** In 1973, the University of California at Berkeley feared that they would be sued for gender bias in their graduate school admissions.[1] Table 1

[1] For a bit more of this story, see `https://www.refsmmat.com/posts/2016-05-08-simpsons-paradox-berkeley.html`.

shows the numbers of applicants who were admitted and rejected (by the six largest departments), tabulated by sex.

Table 1: 1973 Berkeley graduate admissions for six largest departments by sex.

	Male	Female
Admitted	1198	557
Rejected	1493	1278

Interest lies in the effect (if any) of sex on admission status. Table 1 might seem to reveal gender bias.

The data for this problem are available in R as the object `UCBAdmissions` if you load the `graphics` library, e.g. `library(graphics)`. The `graphics` library is built into R, but not loaded by default.

(a) Show that the proportion of male applicants that were admitted is higher than the proportion of female applicants that were admitted.

Solution: We can get the marginal table that matches Table 1. The `margin.table` function is built in to R and calculates the sums of entries to make a table like this:

```
library(graphics)
mt <- margin.table(UCBAdmissions, c(1,2))
mt
##           Gender
## Admit      Male Female
##   Admitted 1198    557
##   Rejected 1493   1278
```

We can then calculate the proportions in each row:

```
props <- function(x) { x / sum(x) }
mta <- apply(mt, 2, props)[1,]

mta
##      Male    Female
## 0.4451877 0.3035422
```

The difference in admission rate appears to be fairly large.

(b) Table 2 gives the same data further tabulated according to the six different departments involved in Table 1.

Table 2: 1973 Berkeley graduate admissions by sex and department.

	Department A		Department B		Department C	
	Male	Female	Male	Female	Male	Female
Admitted	512	89	353	17	120	202
Rejected	313	19	207	8	205	391

	Department D		Department E		Department F	
	Male	Female	Male	Female	Male	Female
Admitted	138	131	53	94	22	24
Rejected	279	244	138	299	351	317

Show that women are admitted at a higher rate than men by most of the six departments, and say which departments they are. Based on those few departments where males are admitted at a higher rate, say why this might be called a "near example" of Simpson's paradox.

Solution:

```
dept.admits <- data.frame(dept = LETTERS[1:6],
                          male_admit = numeric(6),
                          female_admit = numeric(6))
for (j in 1:6) {
  dept.admits[j, 2:3] <- apply(UCBAdmissions[,,j], 2, props)[1,]
}

dept.admits
```

```
##   dept male_admit female_admit
## 1    A 0.62060606   0.82407407
## 2    B 0.63035714   0.68000000
## 3    C 0.36923077   0.34064081
## 4    D 0.33093525   0.34933333
## 5    E 0.27748691   0.23918575
## 6    F 0.05898123   0.07038123
```

We see that women are admitted in higher proportion than males in Departments A, B, D, and F. In the other two departments, C and E, the differences are small compared to those in Departments A and B, so this is might be called a "near example" of Simpson's paradox. ("Near" because the effect in groups is the opposite of the overall effect, but only for most groups, not all.)

(c) For the rest of this problem, let Y_j be the binary random variable taking the value 1 if the jth person was admitted and 0 if not. Let X_j be the binary random variable taking the value 1 if the jth person was female and 0 if male. Let Z_j be the categorical random variable taking the values A, B, C, D, E, F that indicate the department.

Compute an estimate of the *conditional regression*

$$r(x, z) \equiv \mathbb{P}(Y = 1|X = x, Z = z)$$

of Y on X given $Z = z$ for each of the six departments (values of z.)

Solution: For each department z, the estimate of the conditional regression function is $\hat{r}(x, z)$, equal to the proportion of those students of sex x in the applicant pool of department z who got admitted. These are the same as the proportions computed in part (b).

To be continued. We'll resume this problem in the next homework assignment.