P8122 Homework 3

Due: October 14 2022tober 18 20 1 at 5:00pm

Roxy Zhang rz2570

Instructions

- Upload a single pdf file for your homework on Canvas.
- The file should include the code in the appendix. Comment your code.
- You may discuss these problems with each other verbally, but must write up the answers on your own, and may not share or show your answers to anyone else.
- Short and clear answers, please.
- No late homework are allowed.

Inspired by the study on the effect that light at night has on weight gain and other variables in mice, we created data by simulating new attributes and exposure assignment for the mice, representing what we would expect to see in an observational study of this relationship. The code to simulate the data is provided. We will proceed using the g-formula to estimate the sample average causal effect.

a) (30 points) We have provided R-code to simulate data of 16 mice. Obesity, a binary covariate (C), is simulated using a Bernoulli distribution by assuming that 1/5 (p) of the mice are obese at baseline. Exposure to light variable (A) is simulated as a Bernoulli distribution conditional on the obesity covariate C. Here we assume that obese mice are less likely to get exposed to light because they are more sedentary (theta1 = -1/5). Finally, we generate a glucose outcome variable (Y), a normally distributed variable whose mean (mu_Y) is both a function of obesity at baseline and light (beta0+beta1*obesity+beta2*light, beta0=110, beta1=20, beta2=-5). Interpret all parameters (p, theta0, theta1, beta0, beta1, beta2).

P: baseline covariate, the proportion of obese mice at baseline is 1/5.

Theta0: the probability that non-obese mice get exposed to the light is 1/2.

Theta1: the probability that obese mice are less likely to get exposed to light is 1/5 lower compared to non-obese mice.

Beta0: the mean of the glucose outcome when the mice are not obese and not exposed to light is 110.

Beta1: holding light exposure the same, the difference in the glucose outcome between the obese mice and non-obese mice is 20.

Beta2: holding obesity status the same, the difference in the glucose outcome between the mice exposed to light and the mice not exposed to light is -5.

b) (10 points) Write the marginal and conditional PACE. Under which assumptions marginal and conditional PACE are identified?

Marginal PACE: $E[Y_1 - Y_0]$

Conditional PACE: $E[Y_1|C=c] - E[Y_0|C=c] = E[Y|A=1, C=c] - E[Y|A=0, C=c]$

No unobserved confounding assumption (NUCA): data is collected on as many "pre-treatment" variables as possible that affect both the treatment/exposure under consideration and the outcome. Thus, conditional on C, A should be independent of all unobserved covariates U. And A is independent of Y_0 and Y_1 given C.

Marginal and conditional PACE are also identified under consistency, SUTVA, exchangeability and positivity assumptions.

c) (10 points) Show the g-formula for the randomized study we considered in the previous homework in which we studied the effect of light (dark=DL Vs bright=LL) on obesity. Compare the g-formula of the current simulated observational study.

The g-formula for randomized trials: $E(Y_a) = \sum_c E(Y|A = a, C = c)\Pr(C = c) = E(Y|A=a)$

The g-formula for observational study: $E(Y_a) = E\{E(Y_a|C)\} = \sum_c E(Y|A = a, C = c)\Pr(C = c)$

G-formula unifies randomized trials and observational studies. However, in observational study, the g-formula no longer has a simple form as a conditional mean, i.e. $E(Y|A=a)$ is no longer equal to $E(Y_a)$.

d) (20 points) Provide the estimate of $E[Y|A = 1] - E[Y|A = 0]$ in your simulated data from part (a). Interpret the results.

```
> mean(Y[A == 1]) - mean(Y[A == 0])
[1] -2.566766
```

The average glucose outcome in mice exposed to light is 2.567 lower that the average glucose outcome in mice not exposed to light. Light has a negative causal relationship with weight gain.

e) (20 points) Provide the estimate of $E[Y_1] - E[Y_0]$ in your simulated data from part (a) using the gformula. Interpret the results. Explain the differences between the inferences obtained in (d) and (e).

```
knitr::kable(bootstrap, digits = 3)
```

V1	mean	se	ll	ul
Observed	115.831804247806	1.01744285848755	113.837652888843	117.825955606769
No Treatment	113.417282662121	1.28581362142713	110.897134273293	115.937431050949
Treatment	118.24632583349	1.15864358192083	115.975426142007	120.517225524974
Treatment - No Treatment	4.82904317136914	0.530261148976293	3.78975041697477	5.8683359257635

The g-formula for observational study: $E(Y_a) = E\{E(Y_a|C)\} = \sum_c E(Y|A = a, C = c)\Pr(C = c)$

$E(Y_1) = E\{E(Y_1|C)\} = \sum_c E(Y|A = 1, C = c)\Pr(C = c)$
$E(Y_0) = E\{E(Y_0|C)\} = \sum_c E(Y|A = 0, C = c)\Pr(C = c)$
$E(Y_1) - E(Y_0) = 4.829$

The average glucose outcome in mice exposed to light is 4.829 higher that the average glucose outcome in mice not exposed to light. Light has a beneficial causal effect on weight gaining.

The difference between (d) and reflects the effect of adjustment for confounding bias. In (d), the effect of confounder obese status is not considered, whereas in (e) the confounder is considered and $E(Y_a)$ can no longer have a simple form as a conditional term.

f) (10 points) Consider now a hypothetical observational study with 10 continuous covariates. Under which assumptions can you estimate $E[Y_1] - E[Y_0]$ using linear regression.

- The 10 covariates don't share co-linearity with each other.
- The 10 covariates are pre-exposure variables.
- The 10 covariates are observed confounders, i.e. they are observed correlates of A and Y.
- Within levels of each 10 covariates, A is as if randomized, i.e. A is independent of Y_0 and Y_1 given C.

Appendix: R code

```{r}

library(tidyverse)

library(boot)

```

```{r}

#generate data

set.seed(124)

n <- 16

p_C <- 1/5

C <- rbinom(n,1,p_C)

theta0 <- 1/2

theta1 <- -1/5

p_A <- theta0+theta1*C

A <- rbinom(n,1,p_A)

beta0 <- 110

beta1 <- 20

beta2 <- 5

sigma_Y <- 1

mu_Y <- beta0+beta1*C+beta2*A

Y <- rnorm(n,mu_Y, sigma_Y)



# create dataframe

library(tidyverse)
```

```
df = cbind(A, C, Y) %>%
  as.data.frame()
```

```{r}
# Provide the estimate of E[Y |A = 1] − E[Y |A = 0] in simulated data
mean(Y[A == 1]) - mean(Y[A == 0])
```

```{r}
# Provide the estimate of E[Y1] − E[Y0] in simulated data using the gformula
standardization <- function(data, indices) {
  # create a dataset with 3 copies of each subject
  d <- data[indices, ] # 1st copy: equal to original one`
  d$interv <- -1
  d0 <- d # 2nd copy: treatment set to 0, outcome to missing
  d0$interv <- 0
  d0$A <- 0
  d0$Y <- NA
```

```
d1 <- d # 3rd copy: treatment set to 1, outcome to missing

d1$interv <- 1

d1$A <- 1

d1$Y <- NA

d.onesample <- rbind(d, d0, d1) # combining datasets



# linear model to estimate mean outcome conditional on treatment and confounders

# parameters are estimated using original observations only (interv= -1)

# parameter estimates are used to predict mean outcome for observations with set

# treatment (interv=0 and interv=1)

fit <- glm(

  Y ~ as.factor(A) + as.factor(C),

  data = d.onesample

)



d.onesample$predicted_meanY <- predict(fit, d.onesample)



# estimate mean outcome in each of the groups interv=-1, interv=0, and interv=1

return(c(

  mean(d.onesample$predicted_meanY[d.onesample$interv == -1]),
```

```
    mean(d.onesample$predicted_meanY[d.onesample$interv == 0]),
    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]),
    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]) -
      mean(d.onesample$predicted_meanY[d.onesample$interv == 0])
  ))
}
```

```{r}
# bootstrap
results <- boot(data = df,
                statistic = standardization,
                R = 5)
```

```{r}
# generating confidence intervals
se <- c(sd(results$t[, 1]),
```

```
      sd(results$t[, 2]),
      sd(results$t[, 3]),
      sd(results$t[, 4]))
mean <- results$t0
ll <- mean - qnorm(0.975) * se
ul <- mean + qnorm(0.975) * se

bootstrap <-
  data.frame(cbind(
    c(
      "Observed",
      "No Treatment",
      "Treatment",
      "Treatment - No Treatment"
    ),
    mean,
    se,
    ll,
    ul
  ))
```

```
knitr::kable(bootstrap, digits = 3)

```