MA592 Introduction to Causal Inference

Homework 2: Estimating the effect of a treatment

Zhiwei Liang

10/5/2021

1. The effect of statins on MI

(a) Estimate the risk of Myocardial Infarction in the entire group of 300,000 men, had none of them used statins, based on the data above, using a conditioning argument.

$$P(Y^{(0)}=1)=E(Y^{(0)})=E[E(Y^{(0)}|X,A=0)]$$
$$=P(X=1)E(Y^{(0)}|X=1,A=0)+P(X=0)E(Y^{(0)}|X=0,A=0)$$
$$=\frac{100003}{300006}\times\frac{400}{50001}+\frac{200003}{300006}\times\frac{360}{180002}$$

So the result is 0.004

(b) Estimate the risk of Myocardial Infarction in the entire group of 300,000 men, had they all used statins, based on the data above, using a conditioning argument.

$$P(Y^{(1)}=1)=E(Y^{(1)})=E[E(Y^{(1)}|X,A=1)]$$
$$=P(X=1)E(Y^{(1)}|X=1,A=1)+P(X=0)E(Y^{(1)}|X=0,A=1)$$
$$=\frac{100003}{300006}\times\frac{240}{50002}+\frac{200003}{300006}\times\frac{30}{20001}$$

So the result is 0.0025999

(c) Estimate the average effect of statins on Myocardial Infarction in the entire group of 300,000 men, based on the data above.

$ATE=E(Y^{(1)}-Y^{(0)})=E(Y^{(1)})-E(Y^{(0)})$ = -0.0014001

(d) Compare your results with the results you obtained in Questions 3 and 5 of Homework 1.

In HW1(3): the $Diff=P(MI|highrisk,statins)-P(MI|highrisk,nostatins)$ = -0.0032

In HW1(5): $Diff=P(MI|lowrisk,statins)-P(MI|lowrisk,nostatins)=-5.0005277\times10^{-4}$

In this case, the ATE is -0.0014001 in the entire group of men.

(e) Compare your results with the results you obtained in Question 9 of Homework 1. Can you say a few words about bias, and in this case, provide the numeric value of the bias?For now, just ignore random variation.

In HW1(9): $Diff=P(MI|statins)-P(MI|nostatins)=5.5267283\times10^{-4}$

The result indicates that the risk with statins is 0.06% higher than without statins.

In 1(c): The result is -0.0014 in the entire group of men, and it indicates that the MI risk with statins is 0.14% lower than without statins.

The conclusions in HW1 and HW2 are totally contrast, the former indicates statins can prevent MI while the latter indicates the statins cannot. The reason of the bias is the given condition of X. In HW1, the probabilities of different risk status were ignored, we just calculate the $E(Y^{(i)}|A=i)$; In HW2, considering $P(X)$, we calculate $E(Y^{(i)}|X, A=i)$.

Bias $=$ Q9 - 1(e) $= 5.5267283 * 10^{-4} - (-0.0014) = 0.002$

(f) Do you think your results are very sentitive to model specification? Why/why not?

No, since I don't use any model.

2.The effect of antibiotic medications on hospital death

The **notation** used in this question:

- A: treatment=1 for caz-avi, treatment=0 for colistin
- X_1: creatininehigh = 1 for patients with high creatinine levels, 0 otherwise
- X_2: infectiontype = 1 for bloodstream infections, 2 for urinary tract infections, and 3 for other types of infection
- X_3: pitt score is a measure of disease severity
- Y: hospitaldeath=1 for patients who died in the hospital and hospitaldeath=0 otherwis

```r
mydata = readxl::read_xlsx("hw2_1000.xlsx")

colnames(mydata)[2] = "x3"
colnames(mydata)[3] = "x2"
colnames(mydata)[4] = "x1"
colnames(mydata)[5] = "a"
colnames(mydata)[6] = "y"

# make the Infection type as factor
mydata$x2 = as.factor(mydata$x2)
```

(a) Estimate the probability of hospital death in the entire simulated population, had everyone been treated with caz-avi, using a conditioning argument. Please include all potential confounders in your analysis; "pre-testing" is often not recommended, we often try to decide which confounders to include based on subject matter knowledge.

$$P(Y^{(1)}=1) = E(Y^{(1)}) = E[E(Y^{(1)}|X_1, X_2, X_3, A=1)]$$

$$= \sum_{i_1=0}^{1}\sum_{i_2=1}^{3}\sum_{i_3=0}^{1} P(X_1=i_1, X_2=i_2, X_3=i_3)E(Y|X_1=i_1, X_2=i_2, X_3=i_3, A=1)$$

```r
library(dplyr)

# function to calculate P(X1,X2,X3)*E(Y^1|X1,X2,X3,A=1)
my_sum_1 = function(i1, i2, i3) {
    n = 1000
    n123 = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3))
    n123a = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3, a == 1))
    n123ay = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3, a == 1, y == 1))
    p = (n123/n) * (n123ay/n123a)
    return(p)
}

# iteration to sum all the combinations of X1-X3
p1 = 0
for (i1 in 0:1) {
    for (i2 in 1:3) {
        for (i3 in 0:1) {
            p1 = p1 + sum(my_sum_1(i1, i2, i3))
        }
    }
}
p1
## [1] 0.08729967
```

The probability is 0.0872997.

(b) Estimate the probability of hospital death in the entire simulated population, had everyone been treated with colistin, using a conditioning argument.

$$P(Y^{(0)}=1)=E(Y^{(0)})=E[E(Y^{(0)}\mid X_1,X_2,X_3,A=0)]$$

$$=\sum_{i_1=0}^{1}\sum_{i_2=1}^{3}\sum_{i_3=0}^{1}P(X_1=i_1,X_2=i_2,X_3=i_3)E(Y|X_1=i_1,X_2=i_2,X_3=i_3,A=0)$$

```
# function to calculate P(X1,X2,X3)*E(Y^1|X1,X2,X3,A=1)
my_sum_0 = function(i1, i2, i3) {
    n = 1000
    n123 = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3))
    n123a = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3, a == 0))
    n123ay = nrow(mydata %>%
        filter(x1 == i1, x2 == i2, x3 == i3, a == 0, y == 1))
    p = (n123/n) * (n123ay/n123a)
    return(p)
}

# iteration to sum all the combinations of X1-X3
p2 = 0
for (i1 in 0:1) {
    for (i2 in 1:3) {
        for (i3 in 0:1) {
            p2 = p2 + sum(my_sum_0(i1, i2, i3))
        }
    }
}
p2
## [1] 0.35744
```

The probability is 0.35744.

(c) What are the assumptions you made in 2(a) and 2(b)? Mention all assumptions, including potential modeling assumptions.

1. $(Y^{(0)},Y^{(1)})\ \ independent\ \ of A,\ given\ X1,X2,X3$

2. Consistency

- If $A = 0$, we observe $Y = Y^{(0)}$.
- If $A = 1$, we observe $Y = Y^{(1)}$.

(d) Estimate the average effect of caz-avi, as compared to colistin, on hospital death, in the entire simulated population.

$ATE = E(Y^{(1)} - Y^{(0)}) = E(Y^{(1)}) - E(Y^{(0)}) =$ -0.2701403

(e) Do you think your results are very sentitive to model specification? Why/why not?

No, since I don't use any model. But if we talk about the method specification, I think it may be sensitive. For example, maybe the results from method IPWT and Conditioning would be different, but it all depends.

(f) Is there confounding by indication in these data? Why do you think that?

one case: Ignoring the three potential confounders, we calculate the overall risk of hospital death: $Diff = P(hospitaldeath|cazavi) - P(hospitaldeath|colistin)$

```
caz = mydata %>%
    filter(a == 1)
coli = mydata %>%
    filter(a == 0)
p3 = mean(caz$y)
p4 = mean(coli$y)
p3 - p4
p1 - p2
## [1] -0.1887431
## [1] -0.2701403
```

The result is -0.1887431, which means the risk of death using cazavi is 18.87% lower than using colistin.

second case: In 2(d), the ATE is -0.2701403,which means the risk of death using cazavi is 27.01% lower than using colistin.

They both come to the conclusion that cazavi is better than colostin, but the numeric values are different. So I guess there's no confounding of indication.

(g) Estimate E[Y |A = a] for a caz-avi and for a colistin.

We use the $\overline{Y}$ average of Y given A = 1 to estimate $E(Y|A = 1)$, the expectation is 0.122807.

In the same way, we use the $\overline{Y}$ average of Y given A = 0 to estimate $E(Y|A = 0)$, the expectation is 0.3115502.

(h) Are your estimates in (g) biased or unbiased estimates for EY(caz—avi) and EY (colistin)? Estimate the biases.

$Bias(1) = E(Y|A = 1) - E(Y^{(1)}) =$ 0.0355073

$Bias(0) = E(Y|A = 0) - E(Y^{(0)}) =$ -0.0458898