MA592 Introduction to Causal Inference

Homework 1: Simpson's paradox

Zhiwei Liang

9/10/2021

1.Estimate the risk of Myocardial Infarction without statins in high risk men, based on the data above

$P(MI|highrisk, nostatins) = 400 \div 50001 = 0.0079998$

2.Estimate the risk of Myocardial Infarction with statins in high risk men, based on the data above.

$P(MI|highrisk, statins) = 240 \div 50002 = 0.0047998$

3.Estimate the risk difference in MI risk due to statins in high risk men, assuming that given they are high risk, the men not on statins and the men on statins are comparable (except for their statin status). The risk difference is the probability of MI under no statins minus the probability of MI under statins.

$Diff = P(MI|highrisk, nostatins) - P(MI|highrisk, statins) = 0.0032$

4.Based on your results, and assuming your results are not due to random variation, do statins prevent MI in high risk men?

In high risk men, the risk without statins is **0.32% higher** than with statins. **So I think statins prevent MI in this population.**

5.Repeat the above for the lower risk men.

$P(MI|lowrisk, nostatins) = 360 \div 180002 = 0.002$

$P(MI|lowrisk, statins) = 30 \div 20001 = 0.0014999$

$Diff = P(MI|lowrisk, nostatins) - P(MI|lowrisk, statins) = 5.0005277 \times 10^{-4}$

In low risk men, the risk without statins is **0.05% higher** than with statins. **So I think statins prevent MI in this population.**

6. Based on these results, do statins prevent Myocardial Infarction?

Based on the results above, I think statins do prevent Myocardial Infarction.

7. Ignoring the risk status, estimate the overall risk of Myocardial Infarction in all men who are not treated with statins in the above population.

$P(MI|nostatins) = (360 + 400) \div (50001 + 180002) = 0.0033043$

8.Ignoring the risk status, estimate the overall risk of Myocardial Infarction in all men who are teated with statins in the above population.

$P(MI|statins) = (240 + 30) \div (50002 + 20001) = 0.003857$

9.If we would not know about the risk status of these men, or chose to ignore it, and then calculated the risk difference between men not on statins and men on statins, what would we conclude?

$Diff = P(MI|nostatins) - P(MI|statins) = -5.5267283 \times 10^{-4}$

In the whole men population, the risk without statins is **0.06% lower** than with statins. **So I think statins can't prevent MI for these men.**

10.Would you please explain the difference between the conclusions in 6. and 9.? Which of these conclusions has a causal interpretation? Why? (Tips: Regarding question 10, the question can be interpreted as: Among the numbers you calculated, which are the numbers of interest for patients/doctors to decide on whether or not to take statins for its effect on MI risk? And which are the numbers that are better ignored when making this decision?)

1)The two conclusions

In question 6, it seems like statins do prevent Myocardial Infarction in either high risk and low risk men groups.

In question 9, it seems like statins can't prevent MI in the population.

2)The explanation

There may be a problem with the selection of men for the two experimental groups, neither of which is sufficiently representative. Physicians seems to feel that high risk men need statins more whereas low risk men need statins less. Therefore, there would be more high risk men in statins group and more low risk men in no-statins group.

The proportion of the high risk men and low risk men between statins and no-statins group is different. I think this gap leads to the paradox in Q6 and Q9.

The *risk status* may also influences the MI, even more than statins. The main reason why statins can't prevent MI well because there are more high risk men in no-statins group, not because of the usage of statins.

3)Which of these conclusions has a causal interpretation?

In my opinion, the **separate results** has a causal interpretation.

The patients were grouped by *risk status* in this case. But is it a real grouping or fake grouping? **It depends on whether the risk_status is the real influence factor for MI.** In other words, the grouping is significant if there are difference between high risk and low risk groups.(Like *age sex race* those are groups which often show difference). We can test to see if there is group difference but we need more data.

So if the *risk status* is the real grouping factor, the risks of MI in separate case are the interests of patients and doctors. For example, for low risk patients, $P(MI|lowrisk, nostatins), P(MI|lowrisk, statins), Diff = P(MI|lowrisk, nostatins) - P(MI|lowrisk, statins)$ are important.

11.Please describe another example or general situation where such a phenomenon could be expected.

In the Internet company, data analysts always do A/B Test to test certain function.

For example, data analysts want to test whether the new function leads to more purchase behaviors. They select several users and divide them into 2 groups. The A group users use the original version function while the B group users use the new version. Among these users, some shopped more while others shopped less before.

The data analysts may find that the new version works in either more-shopping and less-shopping group. But when they consider the whole samples, the new version doesn't work.

This is a usual situation for the data analysts.

12.Introduce some notation and write out what is estimated above using this notation.

Notation: P(A) means the probability or rate of event A, | means the condition.

The separate case:

$P(purchase|newfunction, moreshoppingusers)$

$P(purchase|originalfunction, moreshoppingusers)$

$P(purchase|newfunction, lessshoppingusers)$

$P(purchase|originalfunction, lessshoppingusers)$

Diff in less shopping users =

$$P(purchase|newfunction, lessshoppingusers) - P(purchase|originalfunction, lessshoppingusers)$$

Diff in more shopping users =

$$P(purchase|newfunction, moreshoppingusers) - P(purchase|originalfunction, moreshoppingusers)$$

The aggregate case:

$P(purchase|newfunction)$

$P(purchase|originalfunction)$

Diff =

$$P(purchase|newfunction) - P(purchase|originalfunction)$$

13.Fit a logistic regression model for the probability of MI given statin treatment (yes/no) and risk status (high/lower risk). Do this first without an interaction term for treatment and risk status. Which coefficient do you find for the statin treatment variable? Is the treatment effect significant?

```r
# build the dataset
g1 = c(1, 0, 1)
g1 = rep(1, 400) %*% t.default(g1) %>%
    as.data.frame()

g2 = c(1, 0, 0)
g2 = rep(1, 49601) %*% t.default(g2) %>%
    as.data.frame()

g3 = c(1, 1, 1)
g3 = rep(1, 240) %*% t.default(g3) %>%
    as.data.frame()
```

```r
g4 = c(1, 1, 0)
g4 = rep(1, 49762) %*% t.default(g4) %>%
    as.data.frame()

g5 = c(0, 0, 1)
g5 = rep(1, 360) %*% t.default(g5) %>%
    as.data.frame()

g6 = c(0, 0, 0)
g6 = rep(1, 179642) %*% t.default(g6) %>%
    as.data.frame()

g7 = c(0, 1, 1)
g7 = rep(1, 30) %*% t.default(g7) %>%
    as.data.frame()

g8 = c(0, 1, 0)
g8 = rep(1, 19971) %*% t.default(g8) %>%
    as.data.frame()

mydata = rbind(g1, g2, g3, g4, g5, g6, g7, g8)
# head(mydata,10)

colnames(mydata) = c("risk_status", "treatment", "MI")
```

```r
# log reg model without interaction term
fit1 = glm(MI ~ risk_status + treatment, family = binomial(link = "logit"), data = mydata)

summary(fit1)
##
## Call:
## glm(formula = MI ~ risk_status + treatment, family = binomial(link = "logit"),
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.1260  -0.0991  -0.0637  -0.0637   3.6555
##
## Coefficients:
##             Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -6.19902    0.05092 -121.736  < 2e-16 ***
## risk_status  1.36624    0.06876   19.869  < 2e-16 ***
## treatment   -0.48120    0.07587   -6.342 2.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13745  on 300005  degrees of freedom
## Residual deviance: 13346  on 300003  degrees of freedom
## AIC: 13352
##
## Number of Fisher Scoring iterations: 9
```

The coefficient of statins treatment is -0.4811988.

It seems like the treatment effect is significant.

14.Fit a logistic regression model for the probability of MI given statin treatment (yes/no) and risk status (high/lower risk). Do this with an interaction term for treatment and risk status. Which coefficient(s) do you find for the statin treatment variable(s)? Is the effect of statins significant in any of the risk groups?

```
fit2 = glm(MI ~ risk_status + treatment + risk_status * treatment, family = binomial(link = "logit"),
    data = mydata)
summary(fit2)
##
## Call:
## glm(formula = MI ~ risk_status + treatment + risk_status * treatment,
##     family = binomial(link = "logit"), data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.1267  -0.0981  -0.0633  -0.0633   3.6062
##
## Coefficients:
##                       Estimate Std. Error  z value Pr(>|z|)
## (Intercept)           -6.21262    0.05276 -117.758   <2e-16 ***
## risk_status            1.39232    0.07283   19.119   <2e-16 ***
## treatment             -0.28822    0.19017   -1.516    0.130
## risk_status:treatment -0.22584    0.20706   -1.091    0.275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13745  on 300005  degrees of freedom
## Residual deviance: 13345  on 300002  degrees of freedom
## AIC: 13353
##
## Number of Fisher Scoring iterations: 9
```

In high risk group(risk_status = 1), the coefficient of statins treatment is -0.5140663;

In low risk group(risk_status = 0), the coefficient of statins treatment is -0.2882219;

When considering the interaction of statins_treatment and risk_status, it seems like the treatment effect is not significant in both groups since the p-values are all large.

15.Given the fit of the above 2 models, what are your conclusions about the effect of statins in this study?

The statins treatment is negative correlated with probability of MI. But maybe the risk_status is not a significant grouping factor.