BIOST/EPI 536 Homework 2

Ivy Zhang

10/12/2021

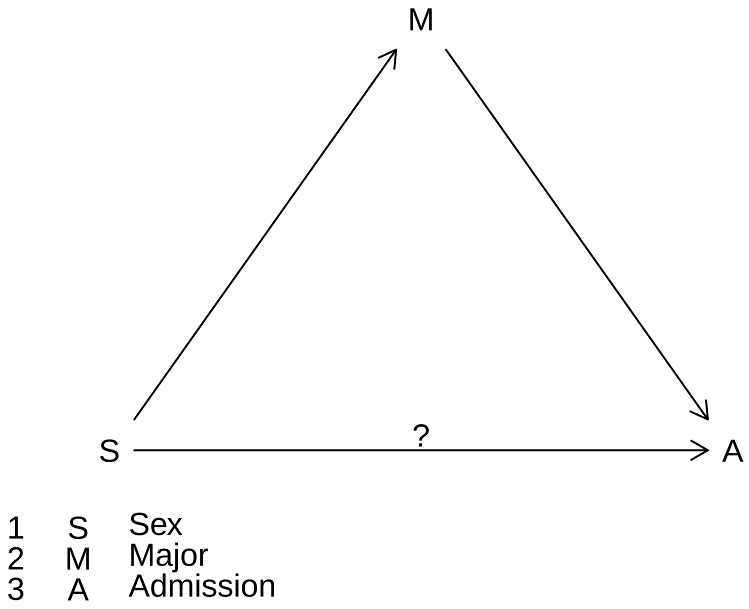
Question 1

(a) If the OR is not equal to 1, then the company would prefer the sex-adjusted OR. Because the treatment was randomly assigned and have no association with the sex, therefore, the OR for pooled is attenuated compared to the sex-adjusted OR.

(b)I will choose OR to summarizing the effect of the treatment of the data. Because RR depends on the baseline risks, and groups with different sex have different baseline risk. However, odds ratio haave no such limitation.

Question 2

(a)



1	S	Sex
2	M	Major
3	A	Admission

(b) From the logistic regression model, we estiamted that for female, the odds of being admiitted to the graduate school increase by a factor of 1.84 compared to male. Sex is a statisticall significant variable at the level of 0.05.

(c) From the logistic regression model, we estimated that the odds of the female being admitted to the graduate school decrease by a factor of 0.905 compared to the male applying to the same major. Sex is not a statistical significant variable at the level of 0.05.

(d) Yes. It is very different. If we do not adjust for major, it seems like female have an advantange in applying to the graduate school comared to male. However, if we adjust for major, female seems to have disadvantage of getting admitted by the graduate school to male. They are very different results from part b and part c.

(e)I think we should use the adjusted analysis to address the question onf whether the University discriminates against women in graduate school admissions. Although major looks like a mediator instead of a confounder in the previous dag, what we focus is whether there is another path that is directly related to the sex and admission rate. This analysis can help us to research for the question of whethere there is a discrimination against women in population who applied to the same major, therefore eliminating the effect caused by sex proportion differences within difference major applications.

(f) I think we also should consider about the year. The acceptance rate may vary from year to year because of different reasons such as COVID19.

Question 3

(a) We are fitting logistic model as

$$logit(P[D|asbestos, smoke]) = \beta_0 + \beta_1 I_{asbestos} + \beta_2 I_{smoke} + \beta_3 I_{asbestos} I_{smoke}$$

where p is the probabillity of getting lung-cancer at given asbesto exposure status and smoke status, and $I_{asbesto} = 1$ when the participant is exposured by asbesto and otherwise equals to 0. $I_{smoke} = 1$ when the participant smokes, otherwise equals to 0.

β_0 estimates the log-odds of the probability of getting lung-cancer when the participant is not exposed by the asbestos and do not smoke. This parameter does not estimate a population quantity because in the case-control study, the numbers of cases and controls have been fixed. The ratio of cases and controls in the study may not be exactly the same as the population is.Therefore, the distribution of the cases and controls is different in the study compared to the real-world population. Therefore this parapeter is not estimating a population quantity.

β_1 estimates the difference between the log-odds of of the probability of getting lung-cancer when the participant is not exposed by the asbestos and smoke and the the log-odds of the probability of getting lung-cancer when the participant is not exposed by the asbestos and do not smoke, which is the log-odds ratio for asbestos exposure among non-smokers. It is estimating a population quantity.

β_2 estimates difference between the log-odds of of the probability of getting lung-cancer when the participant is exposed by the asbestos and smoke and the the log-odds of the probability of getting lung-cancer when the participant is exposed by the asbestos and does not smoke, which is the log-odds ratio for smoking among participants are not exposed by the asbestos. It is estimating a population quantity.

β_3 estimates the difference between the log-odds ratio for smoking when the participant is exposed to the asbestos and the log-odds ratio for smoking among participants are not exposed by the asbestos. It is estimating a population quantity.

(b) Based on the logistic regression result, we estimated that the OR for asbestos among non-smokers is 0.5.

(c) Based on the logistic regression result, we estimated that the OR for asbestos among smokers is 0.0167.

(d) We applied a logistic regression with lung-cancer as the response variable, smoking and asbestos as the independent variables and with interaction term for asbestos exposure and smoking. Based on the model, we estiamted that the interaction term between asbestos exposure and smoking is statistically significant at the level of 0.05. Therefore, we estimated that the odds of asbestos exposure among smokers is smaller at a facor of 0.0333 than the odds of asbestos exposure among non-smokers.

(e) Based on the simple logistic regression model using the subset of the data on smokers, we estimated the OR for asbestos among smokers is 0.0167.We have 95% confidence the real odds ratio is between 0.0053 and 0.0469. The result is exactly same as the result from c as our expectation. Because it is the odds-ratio for asbestos among smokes in this dataset.

(f) Based on the logistic regression model, we estimated that the common smoking-adjusted OR for asbestos for smokers and non-smokers is 0.056.

(g) Based on the wald test and a significance level of 0.05, we will reject the null hypothesis that the smoking-adjusted odds ratio is 1. It is from the fact that p-value is smaller than 0.05 from the summary of the model in 3f.

Code Appendix

```
#----------------Set Up--------------------
library(dplyr)
options(digits = 3)
#-------------------Q2 Draw Dag--------------------
library(dagR)
library(readr)
sexbias = read_csv("~/Desktop/R hw/sexbias.csv")
q2dag = dag.init(outcome = NULL, exposure = NULL,
                 covs = c(1),arcs = c(0,1,
                                          1,-1),
                 x.name = "Sex", y.name = "Admission",
                 cov.names = "Major",
                 symbols = c("S","M","A"))
dag.draw(q2dag)
#-----------------Q2 Logistic regression------------------------
q2b = glm(ACCEPT~SEX, data = sexbias, family = binomial)
summary(q2b)
exp(coef(q2b))
q2c = glm(ACCEPT~SEX+MAJOR, data = sexbias, family = binomial)
summary(q2c)
exp(coef(q2c))
#--------------------Q3 Logistic Regression--------------------------------
load("/Users/ivyyuezhang/Desktop/R hw/asbestos.Rdata")

q3bc = glm(LUNGCA~relevel(SMOKE, ref= "No")*relevel(ASBESTOS,ref = "No"),
           data = asbestos_data, family = binomial)
summary(q3bc)
exp(coefficients(q3bc))
smokers = asbestos_data[which(asbestos_data$SMOKE=="Yes"),]
q3e = glm(LUNGCA~relevel(ASBESTOS,ref = "No"), data = smokers,
          family = binomial)
summary(q3e)
confint(q3e)
q3f = glm(LUNGCA~relevel(SMOKE, ref= "No")+relevel(ASBESTOS,ref = "No"),
          data = asbestos_data, family = binomial)
summary(q3f)
```