BIOST 536 Homework 2

Benjamin Stan

October 16, 2021

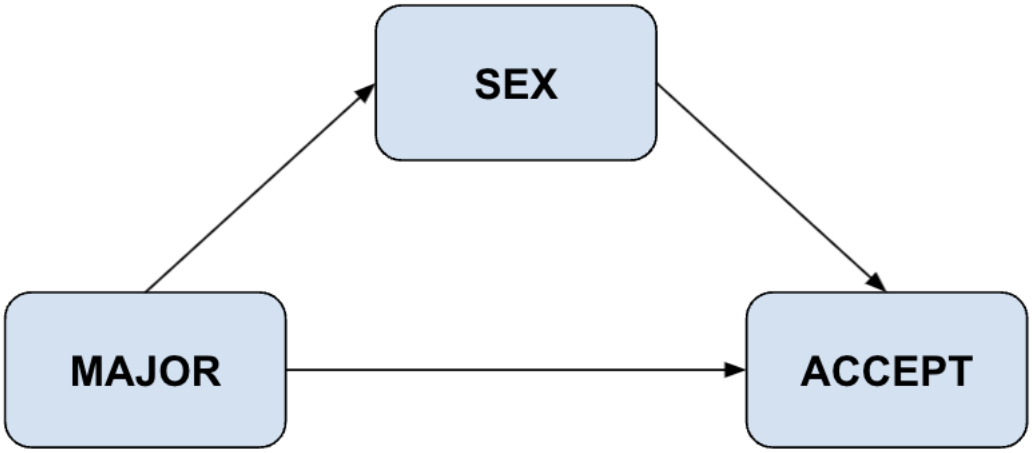
Question 1

a.) Given that the drug is effective at reducing the chances of severe disease, it would be expected that the crude, or pooled, OR would be attenuated (closer to 1) relative to the OR for each of the sexes individually. Thus, the sex-adjusted OR would be preferred for the marketing department, as this value would be more extreme (further from 1). Graphically, the point on the L'Abbe plot for males will have larger values on the x and y axis than for females because they are more likely to develop the disease in both the exposed and unexposed cases.

b.) In this trial, to summarize the effect of the treatment, I would use the relative risk (RR) because it provides a measure for the efficacy of the treatment relative to the baseline values. In this trial, the baseline values are interpretable, as the approach is a randomized control trial. We also know that the baseline values differ between sexes, so the RR will provide a means to compare relative to respective baselines.

Question 2

a.) This DAG shows major as a confounder between sex and accept because the choice of major determines the breakdown of sex that will be considered.



b.) Based on the results of a logistic regression, the odds of acceptance to graduate school are 1.84 times greater (95% robust CI: 1.62-2.09) for a male population compared to a female population.

c.) Based on the results of a logistic regression, the odds of acceptance to graduate school for a male population are 0.905 times that of a female population of the same major (95% robust CI: 0.773-1.06).

d.) The results from problems b and c are very different in that they lead to competing conclusions. The results from b suggest that men are substantially more likely to be admitted, while the results from c suggest that they are less likely to be admitted. The 95% confidence interval for problem b does not include 1, while that from problem c does include 1; however, the point estimates still show opposing conclusions.

e.) The results from the analysis in question c better address the question of University discrimination because the selected majors have an impact on the admission rate and will affect the sex breakdown of applicants. As such, it needs to be adjusted for in the logistic regression analysis.

f.) It would have been beneficial to have additional information regarding the difference in acceptance rate by major. For instance, information regarding the amount of funding for the department might be a potential confounder in that it affects the acceptance rate.

Question 3

a.) We will use the following model for reference:

$$logit(P(LungCancer)) = \beta_0 + \beta_1 * Smoke + \beta_2 * Asbestos + \beta_3 * Smoke * Asbestos$$

In this model

- β_0 does not measure any population parameter, as it includes the term for the likelihood of being sampled from the population, a quantity that we do not have available
- β_1 represents the log-odds ratio of lung cancer for the sub-population that smokes and does not have asbestos exposure compared to the sub-population that neither smokes nor has asbestos exposure
- β_2 represents the log-odds ratio of lung cancer for the sub-population that has asbestos exposure but does not smoke
- β_3 represents the log-odds ratio of lung cancer for the sub-population that has asbestos exposure and smokes compared to the sub-population that has asbestos exposure but does not smoke.

b.) According to the model, the OR for lung cancer for a sub-population with asbestos exposure and no smoking behavior is 2.00 (95% robust CI: 0.61-6.58). This is based on the coefficients of the logistic regression in part a.

c.) According to the model, the OR for lung cancer for a sub-population with asbestos exposure and smoking behavior is 60.0 (95% CI: 21.4-168.0).

d.) The odds ratio between two sub-populations, one that has asbestos exposure and smokes and another that has asbestos exposure and does not smoke, is 30.0 (95% robust CI: 6.2-144.8). There is strong evidence to reject the null hypothesis that the odds ratio between these two groups is one (p-value: $2.3 * 10^{-5}$).

e.) The logistic regression of the log-odds of lung cancer in the smoking population with a covariate for asbestos exposure has the following model:

$$logit(P(LungCancer)) = \beta_0 + \beta_1 * Asbestos$$

Based on this model, the odds ratio for a population with asbestos exposure and smoking behavior is 60.0 (95% robust CI: 21.4-168.0). The point estimate and confidence interval bounds all match the results of part c. This is to be expected because both regression estimates are referring to a population that is both asbestos-exposed and smokes. Part c isolates this population with the effect modifier parameter, while this model isolates it via filtering.

f.) The logistic regression of the log-odds of lung cancer with covariates for smoking status and asbestos exposure has the following model:

$$logit(P(LungCancer)) = \beta_0 + \beta_1 * Smoke + \beta_2 * Asbestos$$

Based on this model, the smoking-adjusted odds ratio for asbestos exposure is 17.9 (95% robust CI: 9.3-34.3). When comparing this result to those in parts b and c, the estimated OR in this case is between the two estimates for asbestos-exposed populations in the model that includes an effect modifier. This difference is expected because the interpretation of the parameters differs. The model referenced in parts b and c seeks to identify the difference in OR for those who were exposed to asbestos, distinguishing between those who smoke and those who do not. By comparison, this model identifies the average odds ratio for those with asbestos exposure and any given smoking status. It can be thought of as averaging over the effect modification in the previous model.

g.) The results of the logistic regression provide strong evidence to reject the null hypothesis that the smoking-adjusted odds ratio for a population exposed to asbestos is 1 (p-value $< 2.0 * 10^{-16}$).

Appendix:

```
## Question 2a
setwd("/Users/bstan/Documents/UW/Courses/BIOST 536")
rm(list = ls())
knitr::include_graphics("figs/hw2_q2_dag2.png")

## Question 2b
library(tidyverse)
library(tidyr)
library(tinytex)
library(sandwich)
sexbias <- read_csv("data/sexbias.csv")
table(sexbias$SEX,sexbias$MAJOR)
table(sexbias$ACCEPT,sexbias$MAJOR)

### Fit unadjusted logistic regression
sexbias <- sexbias %>% mutate(ACCEPT_BIN = ifelse(ACCEPT=="yes",1,0))
glm1 <- glm(ACCEPT_BIN ~ SEX, data = sexbias, family = "binomial")
glm1$coef %>% exp

coef1 <- glm1$coef
normal_se1 <- summary(glm1)$coefficients[, 2]
rob_se1 <- sqrt(diag(vcovHC(glm1, type = "HC0")))
conf_int_tx1 <- coef1[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se1[2]
conf_int_tx1 %>% exp

## Question 2c
glm <- glm(ACCEPT_BIN ~ SEX + MAJOR, data = sexbias, family = "binomial")
glm$coef %>% exp

coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int_tx <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
```

```r
conf_int_tx %>% exp

## Question 3a
asbestos <- read_csv("data/asbestos.csv")
asbestos <- asbestos %>% mutate(SMOKE_BIN = ifelse(SMOKE=="Yes",1,0)) %>%
  mutate(ASBESTOS_BIN = ifelse(ASBESTOS=="Yes",1,0)) %>%
  mutate(LUNGCA_BIN = ifelse(LUNGCA=="Yes",1,0))
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN*ASBESTOS_BIN, data = asbestos, family = "binomial")

## Question 3b
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN*ASBESTOS_BIN, data = asbestos, family = "binomial")
glm$coef %>% exp
coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int_tx <- coef[3] + c(0, qnorm(c(0.025, 0.975))) * rob_se[3]
conf_int_tx %>% exp

## Question 3c
library(multcomp)
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN*ASBESTOS_BIN, data = asbestos, family = "binomial")
coef <- glm$coef
coef
### Use glht function to produce confidence intervals for combination of coefficients
t_complete <- glht(glm, "ASBESTOS_BIN + SMOKE_BIN:ASBESTOS_BIN = 0")
summary(t_complete)
confint(t_complete)
print(c(exp(4.094),exp(3.065),exp(5.124)))

## Question 3d
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN*ASBESTOS_BIN, data = asbestos, family = "binomial")
coef <- glm$coef
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int_tx <- coef[4] + c(0, qnorm(c(0.025, 0.975))) * rob_se[4]
conf_int_tx %>% exp
summary(glm)

## Question 3e
asbestos_smoke <- asbestos %>% filter(SMOKE_BIN==1)
glm <- glm(LUNGCA_BIN ~ ASBESTOS_BIN, data = asbestos_smoke, family = "binomial")
coef <- glm$coef
coef %>% exp

normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int_tx <- coef[2] + c(0, qnorm(c(0.025, 0.975))) * rob_se[2]
conf_int_tx %>% exp

## Question 3f
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN+ASBESTOS_BIN, data = asbestos, family = "binomial")
coef <- glm$coef
coef %>% exp
```

```r
normal_se <- summary(glm)$coefficients[, 2]
rob_se <- sqrt(diag(vcovHC(glm, type = "HC0")))
conf_int_tx <- coef[3] + c(0, qnorm(c(0.025, 0.975))) * rob_se[3]
conf_int_tx %>% exp

## Question 3g
glm <- glm(LUNGCA_BIN ~ SMOKE_BIN+ASBESTOS_BIN, data = asbestos, family = "binomial")
summary(glm)
```