BIOST 536: Homework 2

Department of Biostatistics @ University of Washington

Alejandro Hernandez

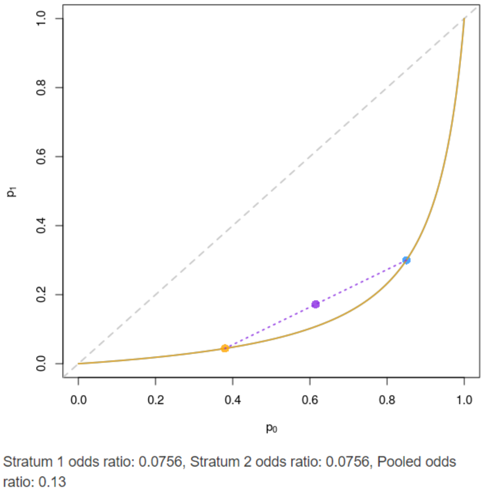
October 12, 2024

Hypothetical Pharmacuetical Trial

1. A pharmaceutical company has run a randomized controlled trial of a drug intended for patients with symptomatic infection to prevent progression to severe disease. Assume that males with symptomatic infection are substantially more likely to progress to severe disease than females. In the trial, treatment vs. placebo was randomly assigned, with males and females each evenly divided between the treatment vs. placebo groups. You can also assume the treatment is effective – it does reduce the chances of severe disease.

a. Suppose the company chooses to summarize results with OR, and that the OR in males is the same as the OR in females. If the marketing department wants to make their drug look as impressive as possible, would they prefer the crude OR or the sex-adjusted OR? Why?

They would prefer the sex-adjusted OR, because the crude/pooled OR is closer to 1, suggesting a weaker treatment effect than the adjusted OR. Below is an illustration of this example, where strata 1 and 2 have the same OR and the pooled OR is *attenuated* (reduced in effect) in comparison. See a comparable example in Figure 1, where risk of disease is higher in one strata than the other.



Stratum 1 odds ratio: 0.0756, Stratum 2 odds ratio: 0.0756, Pooled odds ratio: 0.13

Figure 1: Example of stratified and attenuated pooled odds ratio.

b. Which measure would you prefer for summarizing the effect of the treatment? RR, RD, or OR? Explain briefly.

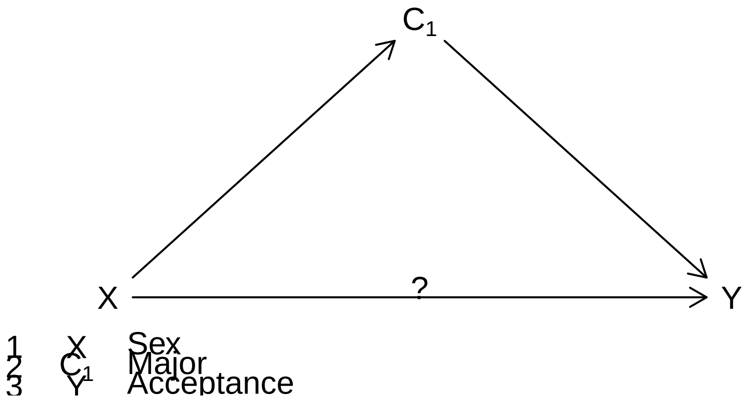
I like relative risk because it's a clear metric how much more the risk of disease is between the treated and untreated groups. I also like the language "reduce your risk by #%" that a risk difference allows.

Parameters of Logistic Regression

2. A major public research university has been accused of discriminating against women in admission to its graduate programs. A task force randomly selects 6 graduate programs ("majors") from across the university to investigate the question. Use the dataset `sexbias` to investigate the following.

Characteristic	**Accepted**, N = 1,755	**Not accepted**, N = 2,771
Sex		
female	557 (32%)	1,278 (46%)
male	1,198 (68%)	1,493 (54%)
Major		
A	601 (34%)	332 (12%)
B	370 (21%)	215 (7.8%)
C	322 (18%)	596 (22%)
D	269 (15%)	523 (19%)
E	147 (8.4%)	437 (16%)
F	46 (2.6%)	668 (24%)

a. For the three variables in the dataset, draw a DAG representing the most appropriate scientific model to approach the question.



b. Use logistic regression to examine the unadjusted association between sex and acceptance to graduate school. Summarize the results in language suitable for the task force's report.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.436	0.051	-16.36	0	0.394	0.481
SEXmale	1.841	0.064	9.55	0	1.625	2.087

Sex = {0:male, 1:female} logodds(acceptance|sex) = B0 + B1*sex

Then exp(B1) estimates the odds ratio of males to females.

c. Use logistic regression to examine the association between sex and acceptance to graduate school adjusted for "major". Summarize the results in language suitable for the task force's report.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.978	0.099	6.880	0.000	1.630	2.404
SEXmale	0.905	0.081	-1.235	0.217	0.772	1.060
MAJORB	0.958	0.110	-0.395	0.693	0.772	1.188
MAJORC	0.283	0.107	-11.841	0.000	0.229	0.348
MAJORD	0.274	0.106	-12.234	0.000	0.222	0.337
MAJORE	0.176	0.126	-13.792	0.000	0.137	0.224
MAJORF	0.037	0.170	-19.452	0.000	0.026	0.051

d. Are the results from b and c very different? Why or why not? (Don't answer in general terms, answer in terms of this dataset.)

e. Which analysis best addresses the question of whether the University discriminates against women in graduate school admissions?

f. Is there any other information you would have liked to have had for this analysis (e.g. any unmeasured potential confounders)?

Logistic Regression with Interaction

3. The course CANVAS site has a file of (fictitious) data from a case-control study of lung-cancer examining two exposures, smoking and asbestos. Asbestos is the exposure of interest. Fit a logistic regression model with a main effect for asbestos exposure, a main effect for smoking, and an interaction term for asbestos exposure and smoking. (This is "model A" in lecture 4.)

Characteristic	**Lung cancer**, N = 95	**No lung cancer**, N = 190
Smoke	80 (84%)	100 (53%)
Asbestos	80 (84%)	38 (20%)

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.139	0.337	-5.85	0.000	0.067	0.256
ASBESTOSYes	2.000	0.608	1.14	0.254	0.565	6.406

term	estimate	std.error	statistic	p.value	conf.low	conf.high
SMOKEYes	0.450	0.571	-1.40	0.162	0.135	1.329
ASBESTOSYes:SMOKEYes	30.000	0.803	4.23	0.000	6.563	157.672

a. For each of the four regression parameters in the model: what population quantity does the parameter estimate? If the parameter does not estimate a population quantity, briefly explain why.

0.138888888888885 CI: 0.0673065162562178 - 0.256350656962234

2.00000000000005 CI: 0.565494879601021 - 6.40585727334265

0.450000000148663 CI: 0.134818589822024 - 1.32926473635846

29.9999999900892 CI: 6.56328550210987 - 157.672176468547

b. According to the fitted model, what is the OR for asbestos among non-smokers?

2.00000000000005 CI: 0.565494879601021 - 6.40585727334265

c. According to the fitted model, what is the OR for asbestos among smokers?

59.9999999801798 CI: 3.71150434480275 - 1010.0254584348

d. Summarize the evidence that smokers and non-smokers have different ORs for asbestos. Write your answer in a few sentences suitable for a scientific publication.
29.9999999900892 CI: 6.56328550210987 - 157.672176468547

e. One could instead estimate the OR for asbestos among smokers by fitting a simple logistic regression model using the subset of the data on smokers. Do this. Compare your point estimates and confidence intervals here and part c and comment on whether any similarities or differences are to be expected.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.063	0.461	-6.01	0	0.022	0.139
ASBESTOSYes	60.000	0.525	7.80	0	23.305	188.883

59.9999999801796 CI: 23.3048770880592 - 188.883152849611

f. Use an appropriate logistic regression model to estimate the smoking-adjusted OR for asbestos. Compare your results with b and c above.

g. For the model in part f, perform a test of the null hypothesis that the smoking-adjusted odds ratio is 1.

End of report. Code appendix begins on the next page.

Code Appendix

```r
# setup options
knitr::opts_chunk$set(echo = FALSE, message = FALSE)
options(knitr.kable.NA = '-')
options(digits = 3)
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]
# clear environment
rm(list=ls())

# load relevant packages
library(dplyr)      # data frame manipulation
library(gtsummary)  # table summaries
library(dagR)       # DAG simulation
library(broom)      # model coefficient table
library(knitr)      # pretty tables

# load data
sbias <- read.csv("../data/sexbias.csv")
sbias <- sbias %>% select(-1) %>% mutate_all(as.factor)

asbt <- read.csv("../data/asbestos.csv")
asbt <- asbt %>% select(-1) %>% mutate_all(as.factor)

## ===============
## Question 2
## ===============
sbias %>%
  dplyr::mutate(ACCEPT = ifelse(ACCEPT=="no", "Not accepted", "Accepted")) %>%
  gtsummary::tbl_summary(by = ACCEPT,
                         label = list(SEX="Sex", MAJOR="Major")) %>%
  bold_labels()
dag <- dagR::dag.init(covs = 1, arcs = c(0, 1, 1, 2),
                      y.name = "Acceptance", x.name = "Sex",
                      cov.names = "Major") %>%
  dag.draw()
mod.lr <- glm(data=sbias, ACCEPT ~ SEX, family=binomial()) %>%
  broom::tidy(exponentiate = TRUE, conf.int = TRUE)
knitr::kable(mod.lr)
mod.lr2 <- glm(data=sbias, ACCEPT ~ SEX + MAJOR, family=binomial()) %>%
  tidy(exponentiate = TRUE, conf.int = TRUE)
kable(mod.lr2)

## ===============
## Question 3
## ===============
asbt %>%
  mutate(LUNGCA = ifelse(LUNGCA=="No", "No lung cancer", "Lung cancer")) %>%
  tbl_summary(by = LUNGCA,
              label = list(ASBESTOS="Asbestos", SMOKE="Smoke")) %>%
  bold_labels()
```

```r
mod.lr3 <- glm(data=asbt, LUNGCA ~ ASBESTOS * SMOKE, family=binomial()) %>%
  tidy(exponentiate = TRUE, conf.int = TRUE)
kable(mod.lr3)
mod.lr4 <- glm(data = subset(asbt, SMOKE=="Yes"),
               LUNGCA ~ ASBESTOS, family=binomial()) %>%
  tidy(exponentiate = TRUE, conf.int = TRUE)
kable(mod.lr4)
```

End of document.