4330 Assignment 2

Ravish Kamath: 213893664

04 December, 2022

Question 1

This data set is from the Duke University Cardiovascular Disease Databank and consists of 2258 patients and 6 variables. The patients were referred to Duke University Medical Center for chest pain. The variables included in the data set **acath2.csv** are the following:

- sex: sex of the patient (0=male, 1=female)
- age: age of the patient
- cad.dur: duration of symptoms of coronary artery disease
- cholest: cholesterol (in mg)
- sigdz: significant coronary disease by cardiac catheterization (definied as $\geq$ 75$ diameter narrowing in at least one important coronry artery - 1 = yes, 0 = no)
- tvdm: severe coronary disease (definied as three vessel or left main disease by cardiac catheterization - 1 = yes, 0 = no))

(a) **[3 points]** In R create a new vector that dichotomizes cholest into high and low, where the cutoff is the median of cholest. Calculate the odds ratio for significant coronary disease based on high/low cholesterol. Interpret the odds ratio.

(b) **[3 points]** Do the same as part (a), but use severe coronary disease instead of significant coronary disease.

(c) **[2 points]** Do you think you could estimate the risk ratio for significant or severe coronary disease in this example? If yes, then estimate the risk ratios for the relationships investigated in part (a) and (b). If not, say why. In either case justify your answer.

Solution

Part A

```
summary(A2df)
```

```
##       sex              age           cad.dur          cholest
##  Min.   :0.0000   Min.   :17.00   Min.   :  0.00   Min.   : 29.0
##  1st Qu.:0.0000   1st Qu.:45.00   1st Qu.:  6.00   1st Qu.:196.0
##  Median :0.0000   Median :51.00   Median : 19.00   Median :224.5
##  Mean   :0.3051   Mean   :50.82   Mean   : 41.91   Mean   :229.9
##  3rd Qu.:1.0000   3rd Qu.:57.00   3rd Qu.: 58.00   3rd Qu.:259.0
##  Max.   :1.0000   Max.   :81.00   Max.   :416.00   Max.   :576.0
##      sigdz            tvdlm
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
```

```
##  Mean   :0.6599   Mean   :0.3202
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
```

Observing our summary of the data set, clearly the median of the variable, cholest, would be 224.5.

```
A2df$cholest = dicho(A2df$cholest, dich.by = 'median')
summary(A2df$cholest)
```

```
##    0    1
## 1129 1129
```

Since we are splitting it based of its median, we can see based of the summary there are 1129 in both levels. 0 = cholesterol level below 224.5, and 1 = cholesterol level is above 224.5.

Now let us proceed with the calculation of the odds ratio for significant coronary disease based on high/low cholesterol.

```
odds_tb = table(A2df$cholest, A2df$sigdz)
colnames(odds_tb) = c("non-signficant coronary", "significant coronary")
rownames(odds_tb) = c('chol < 224.5', 'chol > 224.5')
odds_tb
```

```
##
##                non-signficant coronary significant coronary
##   chol < 224.5                     455                   674
##   chol > 224.5                     313                   816
```

```
n = sum(odds_tb)
chol_great= sum(odds_tb[2,])
chol_low = sum(odds_tb[1,])
#P(significant coronary disease | high cholesterol)
sigcoron_high_prob = odds_tb[2,2]/ chol_great

#P(significant coronary disease | low cholesterol)
sigcoron_low_prob = odds_tb[1,2]/ chol_low


odds_ratio = (sigcoron_high_prob/(1 - sigcoron_high_prob))/
             (sigcoron_low_prob/(1 - sigcoron_low_prob))
odds_ratio
```

```
## [1] 1.759938
```

With the odds ratio being greater than one, the individuals with higher cholesterol will have a higher odds of having significant coronary disease.

Part B

Here is the calculation of the odds ratio for severe coronary disease based on high/low cholesterol.

```
odds_tb = table(A2df$cholest, A2df$tvdlm)
colnames(odds_tb) = c("non-severe coronary", "severe coronary")
rownames(odds_tb) = c('chol < 224.5', 'chol > 224.5')
odds_tb
```

```
##
##                non-severe coronary severe coronary
##   chol < 224.5                 810             319
##   chol > 224.5                 725             404
```

```
chol_great= sum(odds_tb[2,])
chol_low = sum(odds_tb[1,])

#P(severe coronary disease | high cholesterol)
sevcoron_high_prob = odds_tb[2,2]/ chol_great

#P(severe coronary disease | low cholesterol)
sevcoron_low_prob = odds_tb[1,2]/ chol_low


odds_ratio = (sevcoron_high_prob/(1 - sevcoron_high_prob))/
  (sevcoron_low_prob/(1 - sevcoron_low_prob))
odds_ratio
```

```
## [1] 1.414939
```

With the odds ratio being greater than one, the individuals with higher cholesterol will have a higher odds of having severe coronary disease.

Part C

We cannot estimate the risk ratio for either significant nor sever coronary disease. It is clear that this is a retrospective study/case control since we are already sampling relative to the outcome, which would be coronary heart disease. Furthermore, a quick Google search shows us that coronary heart disease is quite common, and hence the probability of coronary heart disease is greater than 5%. Hence we cannot use the odds ratio to estimate the risk ratio.

Question 2

Use the same data set as in Question 1. Run a linear regression model to investigate the joint effect of age and sex on cholesterol (mg).

(a) **[3 points]** Write out the fitted regression model based on the R output. Interpret all the estimated β parameters in the model.

(b) **[2 points]** Calculate and report 95% confidence intervals for the coefficients of age and sex. Interpret.

(c) **[2 points]** Predict the cholesterol of a 50-year-old female.

(d) **[2 points]** Predict the cholesterol of a 10-year-old male. Are you less confident in this prediction than the one you made in part (c)? Why?

Solution

Part A

```
fit <- lm(A2df$cholest~ age + sex, data=A2df)
s.fit <- summary(fit)
s.fit
```

```
##
## Call:
## lm(formula = A2df$cholest ~ age + sex, data = A2df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -208.28  -34.10   -4.92   28.47  339.19
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 224.96743    5.83456  38.558  < 2e-16 ***
## age           0.03886    0.11307   0.344    0.731
## sex           9.78549    2.31139   4.234 2.39e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.43 on 2255 degrees of freedom
## Multiple R-squared:  0.008077,   Adjusted R-squared:  0.007198
## F-statistic: 9.182 on 2 and 2255 DF,  p-value: 0.0001068
```

Our model will be:

$$cholesterol = 224.96743 + 0.03886Age + 9.78549Sex$$

.

β_0 interpretation: Given that the age of the individual is 0 year old and male,their cholesterol level will be 224.96743

β_1 interpretation: Given that sex is a constant, for every increase in unit of Age, the cholesterol level would increase by 0.03886mg.

β_2 interpretation: Given that age is constant, then cholesterol level would increase by 9.78549 if they are female.

Part B

```
confint(fit, level=0.95)
```

```
##                   2.5 %      97.5 %
## (Intercept) 213.5257660 236.4090892
## age          -0.1828807   0.2605961
## sex           5.2528151  14.3181654
```

Age C.I.
Our confidence interval for $\alpha = 0.05$ would be **(-0.1828807, 0.2605961)**.

Sex C.I.
Our confidence interval for $\alpha = 0.05$ would be **(5.258151, 14.3181654)**.

Part C

```
new.x <- data.frame(age= 50, sex= 1)
predict(fit, newdata = new.x)
```

```
##        1
## 236.6958
```

Based of the model we would predict that their cholesterol level would be 236.6958mg for a 50 year old female.

Part D

Based of the model we would predict that their cholesterol level would be **225.356** for a 10 year old male. **Yes I would be less confident** in this prediction. Firstly, most of these patients are experiencing chest pain, which would be common within older people, rather than someone who is 10 year's old. Younger people tend to have a much healthier heart do you to younger age. To have a 225mg cholesterol level, would seem too unrealistic for that individual, unless they have an obesity issue.