MET CS 555 Assignment 6 – 20 points Fall 2, 2019

**SUBMISSION REQUIREMENTS: Please submit a single document (word or PDF) for submission.  Your submission should contain a summary of your results (and answers to questions asked on the homework) as well as your R code used to generate your results (please append to the end of your submission). Please use R for the calculations whenever possible. You will lose points if you are not utilizing R. You will also lose 10 points per day for late submissions unless prior arrangements are made with your facilitator.**

**The data in this document consists of body temperature measurements and heart rate measurements for 65 men and 65 women.**  **Save the data to excel and read the data into R. Use this data to address the following questions.**

(1) We are interested in whether the proportion of men and women with body temperatures greater than or equal to 98.6 degrees Fahrenheit are equal. Therefore, we need to dichotomize the body temperature variable. Create a new variable, called “temp\_level” in which temp\_level = 1 if body temperature >= 98.6 and temp\_level=0 if body temperature < 98.6. (1 point)

**My answer:**

See my R script.

(2) Summarize the data relating to body temperature level by sex. (2 points)

**My answer:**

Using R, we can calculate the numbers of men and women of different levels of body temperatures. There are 51 men and 30 women having body temperature level = 0, which mean the body temperature of these people are lower than 98.6, and 14 men and 35 women have a greater temperature level. Moreover, we get sample population proportion from men and women who has lower temperature level are 51/65 = 0.78 and 30/65 = 0.46, which means the proportion of men who has lower temperature level is greater than women.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Population | Sex | sample size | lower temp level | higher temp level | sample proportion(lower) |
| 1 | Men | 65 | 51 | 14 | 0.78 |
| 2 | Women | 65 | 30 | 35 | 0.46 |

(3) Calculate the risk difference. Formally test (at the α=.05 level) whether the proportion of people with higher body temperatures (greater than or equal to 98.6) is the same across men and women, based on this effect measure (go through the 5-step recipe for testing). Do females have higher body temperatures than males? (4.5 points)

**My answer:**

Denote that p1 = 35 / 65 = 0.538 is the proportion of women with higher body temperature, and p2 = 14 / 65 = 0.215 is the proportion of men with higher body temperature, so we get the risk difference is (p1 - p2) = 0.323. Here is the 5-step recipe for testing whether the proportion of people with higher body temperatures is the same across men and women.

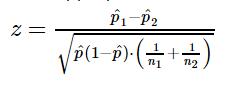
1. Set up the hypotheses and select the alpha level

*H*0 : p1 = p2 (the proportion of people with higher body temperatures for males and females are the same)

*H*1 : p1 ≠ p2 (the proportion of people with higher body temperatures for males and females are different)

α = 0.05

1. Select the appropriate test statistic



1. State the decision rule

Decision rule: Reject H0 if the p-value p ≤ α = 0.05

Otherwise, do not reject H0

1. Compute the test statistic and the associated p-value

prop.test(c(35,14), c(65,65), alternative = "two.sided",

conf.level = 0.95, correct = FALSE)

X-squared = 14.444, df = 1, p-value = 0.0001444

Z = sqrt(X-squared) = 3.80

1. Conclusion

95 percent confidence interval:

0.1659946 0.4801593

Reject H0 since the p-value is less than 0.05. We have significant evidence at α = 0.05 level that the proportion of people with higher body temperatures for men and women are different. Yes, we are 95% confident that percentage of females with higher body temperature is between 16.6% and 48.0% greater than males.

(4) Perform a logistic regression with sex as the only explanatory variable. Formally test (at the α=.05 level) if the odds of having a temperature greater than or equal to 98.6 is the same between males and females (go through the 5-step recipe for testing). Include the odds ratio for sex and the associated 95% confidence interval based on the model in your summary and interpret this value. What is the c-statistic for this model? (5.5 points)

**My answer:**

data$sex2 <- ifelse(data$sex == 1,"M","F")

data$sex2 <- factor(data$sex2, levels = c("M","F"))

data$female <- ifelse(data$sex2 == "F", 2, 1)

m <- glm(data$temp\_level ~ data$female, family = binomial)

summary(m)

First, we create a dummy variable “data$female” (Female versus male). Here’s formally test for simple logistic regression.

1. Set up the hypotheses and select the alpha level

H0 : βsex = 0 or ORsex = 1(there is no association between sex and higher body temperature level)

H1 : βsex ≠ 0 or ORsex ≠ 1(there is an association between sex and higher body temperature level)

α=0.05

1. Select the appropriate test statistic



1. State the decision rule

Decision rule: Reject H0 if p ≤ α = 0.05

Otherwise, do not reject H0

1. Compute the test statistic

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.7397 0.6527 -4.197 2.7e-05 \*\*\*

data$female 1.4469 0.3911 3.700 0.000216 \*\*\*

OR 2.5 % 97.5 %

(Intercept) 0.06459054 0.01797084 0.2321505

data$female 4.25000000 1.97471185 9.1469041

1. Conclusion

Reject H0 since p-value = 0.000216 < α. We have significant evidence at the α = 0.05 level that βsex ≠ 0. That is, there is evidence of an association between sex and having a temperature greater than or equal to 98.6. The odd ratio for sex is 4.25, and the associated 95% confidence interval is between 1.97 and 9.14, that is, we are 95% confident that the odds of having a temperature greater than or equal to 98.6 in women are between 1.97 and 9.14 times higher than men.

R result:

Data: data$prob1 in 81 controls (data$temp\_level 0) < 49 cases (data$temp\_level 1).

Area under the curve: 0.672

The c-statistic is the area under the curve, which is 0.672.

(5) Perform a multiple logistic regression predicting body temperature level from sex and heart rate. Summarize briefly the output from this model – no need to go through the formal 5 step testing procedure. Give the odds ratio for sex and heart rate (for a 10 beat increase). What is the c-statistic of this model? (5 points)

**My answer:**

R result:

Wald test:

----------

Chi-squared test:

X2 = 21.9, df = 2, P(> X2) = 1.7e-05

----------

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.048445 0.423904 -2.473 0.014708 \*

data$heart\_rate 0.013213 0.005687 2.324 0.021739 \*

data$female 0.300513 0.080008 3.756 0.000262 \*\*\*

As given above, both the p-value for global test is 1.7e-05, so we reject the null hypothesis and conclude that there is at least one βi ≠ 0. Also, each p-value of heart rate and gender are less than 0.05. In summary, first, reject the null hypothesis that βheart\_rate = 0 after adjusting for sex. That is, there is evidence of an association between heart rate and body temperature level after adjusting for gender. For gender, likewise, there is evidence of an association between gender and body temperature level after adjusting for heart rate.

The odd ratio for sex is 1.35, and for a 10-beat increase, the odd ratio for heart rate is 1.14, which means the odds ratio for body temperature level is 1.14 for every 10-beats increase in heart rate.

Data: data$prob2 in 81 controls (data$temp\_level 0) < 49 cases (data$temp\_level 1).

Area under the curve: 0.7297

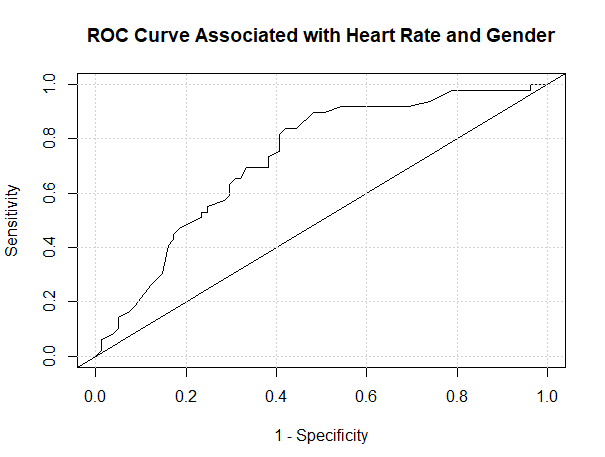
Thus, the c-statistic of this model is 0.7297.

(6) Which model fit the data better? Support your response with evidence from your output. Present the ROC curve for the model you choose. (2 points)

**My answer:**

I think the second model that multiple logistic regression taking heart rate and gender into consideration is better, since we know that the range of area under the ROC curve is between 0.5 and 1, with larger values indicating better fit. Compared with the first model that it’s c-statistic is 0.672, the c-statistic of the second model is 0.7297 and it performs much closer to 1

ROC curve:



**R script:**

setwd("C:\\Users\\Lin\\Desktop\\2019 Fall\\555 DAV\_ R\\Homework\\hw6")

# read the human.csv

data = read.csv("human.csv")

data

data$temp\_level = ifelse(data$temp >= 98.6, 1,0)

# get numbers of each group

# males = 1, females = 2

length(which(data$sex==1 & data$temp\_level == 0)) # man lower

length(which(data$sex==1 & data$temp\_level == 1)) # man higher

length(which(data$sex==2 & data$temp\_level == 0)) # women lower

length(which(data$sex==2 & data$temp\_level == 1)) # women higher

# or using table()

table(data$sex, data$temp\_level)

# higher in women and men

p1 <- length(which(data$sex==2 & data$temp\_level == 1)) / 65

p2 <- length(which(data$sex==1 & data$temp\_level == 1)) / 65

p1 - p2 # risk difference

# two-sample tests for proportion

prop.test(c(35,14), c(65,65), alternative = "two.sided",

conf.level = 0.95, correct = FALSE)

# simple logistic regression model

data$sex2 <- ifelse(data$sex == 1,"M","F")

data$sex2 <- factor(data$sex2, levels = c("M","F"))

data$female <- ifelse(data$sex2 == "F", 2, 1)

m <- glm(data$temp\_level ~ data$female, family = binomial)

summary(m)

#m1 <- glm(data$temp\_level ~ data$sex,family = binomial)

#summary(m1)

# conf int

library(aod)

wald.test(b = coef(m), Sigma = vcov(m), Term = 2:2)

exp(cbind(OR = coef(m), confint.default(m)))

# plotting ROC Curve

library(pROC)

data$prob1 <- predict(m, type = c("response"))

g <- roc(data$temp\_level ~ data$prob1)

plot(1-g$specificities, g$sensitivities, type = "l",

xlab = "1 - Specificity", ylab = "Sensitivity", main = "ROC Curve Associated with Gender")

abline(a=0,b=1)

grid()

# Multiple logistic regression (sex, heart rate)

m2 <- glm(data$temp\_level ~ data$heart\_rate+data$female)

summary(m2)

wald.test(b = coef(m2), Sigma = vcov(m2), Term = 2:3)

exp(cbind(OR = coef(m2), confint.default(m2)))

# for one unit

exp(m2$coefficients[2])

# for 10 unit

exp(m2$coefficients[2]\*10)

# model2 curve

data$prob2 <- predict(m2, type = c("response"))

g2 <- roc(data$temp\_level ~ data$prob2)

plot(1-g2$specificities, g2$sensitivities, type = "l",

xlab = "1 - Specificity", ylab = "Sensitivity", main = "ROC Curve Associated with Heart Rate and Gender")

abline(a=0,b=1)

grid()