Dhakal_Module6

Kokil Dhakal

2023-12-10

Data loading

```r
data.module6 <- read_excel("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module6/modudle6_data.xlsx")
head(data.module6)
```

```
## # A tibble: 6 x 3
##    temp   sex `Heart rate`
##   <dbl> <dbl>        <dbl>
## 1  96.3     1           70
## 2  96.7     1           71
## 3  96.9     1           74
## 4  97       1           80
## 5  97.1     1           73
## 6  97.1     1           75
```

```r
attach(data.module6)
```

Question 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. (2 points)

Answer:

```r
data.module6$temp_level <- ifelse(temp >= 98.6,1,0)
#Also creating dummy variable(binary) for sex to be used in this module here.
data.module6$yes_female <- ifelse(sex== 1,0,1)
high.temp <- data.module6 %>% filter(temp_level==1)
high.temp <- table(high.temp$yes_female)
high.temp
```

```
##
##  0  1
## 14 35
```

```r
head(data.module6)
```

```
## # A tibble: 6 x 5
##    temp   sex `Heart rate` temp_level yes_female
##   <dbl> <dbl>        <dbl>      <dbl>      <dbl>
## 1  96.3     1           70          0          0
## 2  96.7     1           71          0          0
## 3  96.9     1           74          0          0
## 4  97       1           80          0          0
## 5  97.1     1           73          0          0
## 6  97.1     1           75          0          0
```

Here, I create a variable named temp_level from temp variable where > =98.6 assigned to 1 and other assigned to 0 using ifelse() method.Also, new dummy variable yes_female was made where 1 represent female and 0 represent male.After that I used table() method to find the number of male and female in this high body temperature group (group 1).In this high body temperature group, there are 35 females and 14 males.The reason of making male as reference(0) is to make more convenience to interpret the result as females high temp is higher than males high temp.

Question 2

Summarize the data relating to body temperature level (i.e., the variable you created above) by sex. (2 points)

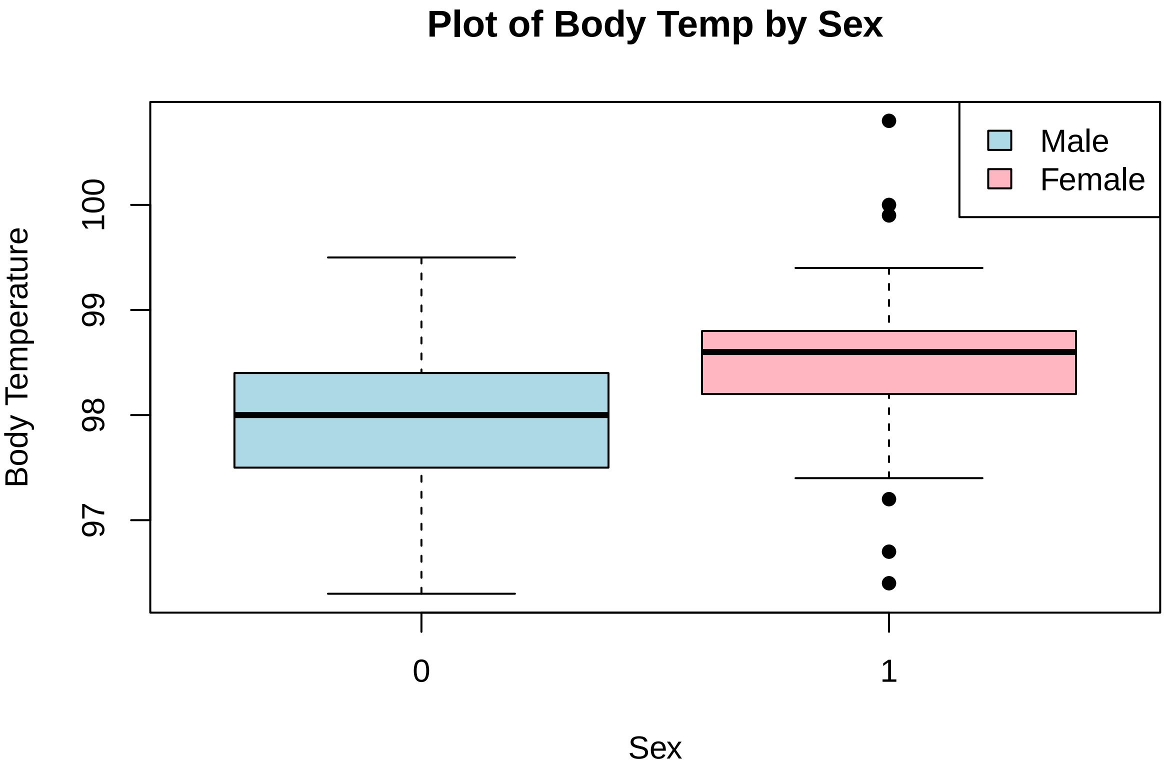
Answer:

```r
temp.level.summary <-data.frame(as.matrix(aggregate(temp, by=list(data.module6$yes_female), summary)))
colnames(temp.level.summary) <- c("group","min","Q1","median","mean","Q3","max")
kable(temp.level.summary,align = "c")
```

group	min	Q1	median	mean	Q3	max
0	96.3	97.5	98.0	97.98923	98.4	99.5
1	96.4	98.2	98.6	98.50923	98.8	100.8

```r
boxplot(temp~yes_female,data = data.module6,
        pch=16,
        col=c("lightblue","lightpink"),
        ylab = "Body Temperature",
        xlab = "Sex",
        main="Plot of Body Temp by Sex")

legend("topright", legend = c("Male","Female"),
       fill = c("lightblue","lightpink"))
```



In this section, I use aggregate method to summarize body tempt by the gender.Here 0 represent male and 1 represent female.Mean body temperature for male is 97.989 and mean body temp for female is 98.509.From box plot it is found that there are some outlier data points in case of female group while there is non for male.From plot and summary table, we can say that the min value,mean, median, Q3,max value and Q1 of female group body temperature are higher than values of body temperature in male group.

Question3

Calculate the risk difference for high body temperature level between men and women. Formally test (at the alpha=0.05 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. You should be showing all 5 steps in the 5-step recipe for testing.

Answer:

```
#calculating proportion of male high temp among males
p_male <- high.temp[1]/65 #total number of male =65
#calculating proportion of female high temp among females
p_female <- high.temp[2]/65 #total number of female =65

#risk difference
risk.difference <-p_female- p_male # considering male group as reference group
risk.difference
```

```
##         1
## 0.3230769
```

Testing hypothesis: 1.Setting hypothesis and alpha level

- H0: p_male=p_female (proportion of people with higher body temperatures (greater than or equal to 98.6) is the same across men and women)
- H1: p_male $\neq$ p_female (proportion of people with higher body temperatures (greater than or equal to 98.6) is not the same across men and women)
- $\alpha = 0.05$

2. Determining test statstics

- z = (p1-p2)/sqrt(p(1-p)*(1/n1+1/n2))
  Where:
  p1= proportion of male with high temp across male
  p2= proportion of female with high tempt accros female
  p= proportion of male +female with high temp
  n1= total number of male
  n2 =total number of female

3. State the decision rule

- Determine the appropriate critical value from the standard normal distribution associated with a right hand tail probability of alpha/2=0.05/2=0.025

```
qnorm(0.025,lower.tail = FALSE)
```

```
## [1] 1.959964
```

- Decision Rule: Reject H0 if |z| $\geq$ 1.960
- Otherwise, do not reject H0

4. Calculate the z value from the provided data

```
# putting values in above formula
z= (p_female-p_male)/sqrt(((14+35)/(65+65)*(1-(14+35)/(65+65)))*(1/65+1/65))
z
```

```
##        1
## 3.800585
```

```
#calculating p-value
pnorm(z,lower.tail = FALSE)
```

```
##            1
## 7.217752e-05
```

5.Conclusion:

Since |z| > critical value(1.96), we reject null hypothesis which is interpreted as proportion of people with higher body temperatures (greater than or equal to 98.6) is not the same across men and women(p <0.01). it can be also say that risk of a high body temp is around 32% higher among female as among males, when male is considering reference group(i.e.risk difference is 0.323).

Question4

Perform a logistic regression with sex as the only explanatory variable. Formally test (at the alpha=0.05 level) if the odds of having a temperature greater than or equal to 98.6 is the same between males and females. Again, please show all 5 steps. Additionally, include the odds ratio for sex and the associated 95% confidence interval in your summary, and interpret the value of the odds ratio. Lastly, what is the c-statistic for this model?

```r
#building the simple logistic regression
set.seed(12345)
log.model <- glm(temp_level~yes_female,data = data.module6,family = "binomial")

summary(log.model)
```

```
##
## Call:
## glm(formula = temp_level ~ yes_female, family = "binomial", data = data.module6)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.2928     0.3017  -4.285 1.83e-05 ***
## yes_female    1.4469     0.3911   3.700 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 172.26  on 129  degrees of freedom
## Residual deviance: 157.45  on 128  degrees of freedom
## AIC: 161.45
##
## Number of Fisher Scoring iterations: 4
```

Hypothesis Testing:

1. setting hypothesis and alpha value

- H0: beta=0 or R=1(There is no association between high body temp and sex(yes_female))

-H1: beta ( )0 or R ( )1 (There is association between high body temp and sex(yes_female))

- $\alpha = 0.05$

2. Determining appropriate test statistics

- z= B1/S.E.(B1)

3. Decision rule.

- Determine the appropriate value from the standard normal distribution associated with a right hand tail probability of alpha/2=0.05/2=0.025 Using the table, z =1.960

-Decision Rule: Reject H0 if $|z| \geq 1.960$ or Reject H0 if $p \leq \alpha$

-Otherwise, do not reject H0

4.Compute the test statistic

```
# calculate value of z and p or just get value of z from above summary table
z <- 1.4469/0.3911 # just get it from above summary table i.e. 3.7
z
```

```
## [1] 3.699565
```

5.Conclusion:

Since, absolute value of |z| (3.700) is greater than critical value(1.96), we reject null hypothesis which can be interpreted as there is association between temp_level(high body temp) and gender variables(yes_female).

For odds ratio and confidence intervals

```
#calculating odds ratio and confidence intervals(95%)
exp(cbind(Odds_ratio = log.model$coefficients, confint.default(log.model)))
```

```
##             Odds_ratio     2.5 %    97.5 %
## (Intercept)  0.2745098 0.1519607 0.4958888
## yes_female   4.2500000 1.9747119 9.1469041
```

```
#another way to calculate odds ratio
odd_ratio <- ((p_female/(1-p_female))/(p_male/(1-p_male)))
odd_ratio
```

```
##    1
## 4.25
```

Now calculating c-statstic for the model

```
roc_obj <- roc(data.module6$temp_level, predict(log.model, type = "response"))
c_statistic <- auc(roc_obj)
c_statistic
```

```
## Area under the curve: 0.672
```

The odd ratio is 4.25 for each unit increase.This can be interpreted in case of categorical independent variable(i.e.yes_female variable) as holding all other variables constant, the odds of having a temperature level of 1 (greater or equal than 98.6) for females are 4.25 times higher than the for males. Also,need to remember that the odds ratio represents the multiplicative change in odds associated with a one-unit increase in the independent variable.The above table also gives 95% confidence interval (lower value is1.9747119 and upper value is 9.1469041). And, lastly c-statstic is 0.67 which means this model has some discriminatory power.

Question 5

Perform multiple logistic regression predicting body temperature level from sex and heart rate. Briefly summarize the output from this model (no need to go through all 5 steps). Give the odds ratio for sex. Also, report the odds ratio for heart rate (for a 10-beat increase). What is the c-statistic of this model?

performing multiple logistic regression:

```
set.seed(1234)
log.multi.model <- glm(temp_level~ yes_female+`Heart rate`,data = data.module6)
summary(log.multi.model)
```

```
##
## Call:
## glm(formula = temp_level ~ yes_female + `Heart rate`, data = data.module6)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.747932   0.418376  -1.788 0.076209 .
## yes_female    0.300513   0.080008   3.756 0.000262 ***
## `Heart rate`  0.013213   0.005687   2.324 0.021739 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2049751)
##
##     Null deviance: 30.531  on 129  degrees of freedom
## Residual deviance: 26.032  on 127  degrees of freedom
## AIC: 167.86
##
## Number of Fisher Scoring iterations: 2
```

Calculating odds ratio

```
odds_ratio_sex <- exp(log.multi.model$coefficients[2]) #for a unit increase

odds_ratio_heartrate <- exp(log.multi.model$coefficients[3]*10) #for a 10 unit increase

odds_ratio_sex
```

```
## yes_female
##   1.350552
```

```
odds_ratio_heartrate
```

```
## `Heart rate`
##     1.141255
```

Calculating c_statstic

```
roc_multi.model <- roc(data.module6$temp_level, predict(log.multi.model,
                                                        type = "response"))

roc_multi.model$auc
```

```
## Area under the curve: 0.7297
```

For this question, I perform multiple logistic regression with temp_level as dependent variable and gender(yes_female) and heart_rate are dependent variables.After printing out the summary of the model,it is found that p-value for female variable is less than 0.001 while p value for heart rate is 0.021739.This shows null hypothesis at level of 0.05 is rejected for favoring alternative hypothesis.We can interpret at level of 0.05 that there is association between sex(yes_female) and high body temperature(temp_level).Similarly, there is association between heart rate and high body temperature(temp_level). We have odds ratio 1.35 for yes_female variable for unit change and 1.14 for heart rate variable for 10 unit change.Odds ratio of 1.35 can be interpreted as odds of a high body temp is 1.35 times higher among female as among males keeping other variables constant.Similarly, Keeping other variables constant, odds of a high body temperature (temp_level) is 1.14 times higher per 10 beat increase in heart rate. From c_stat value which is 0.7297, it can be said that our multiple logistic regression model has moderate level of discriminatory power.

Question 6

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

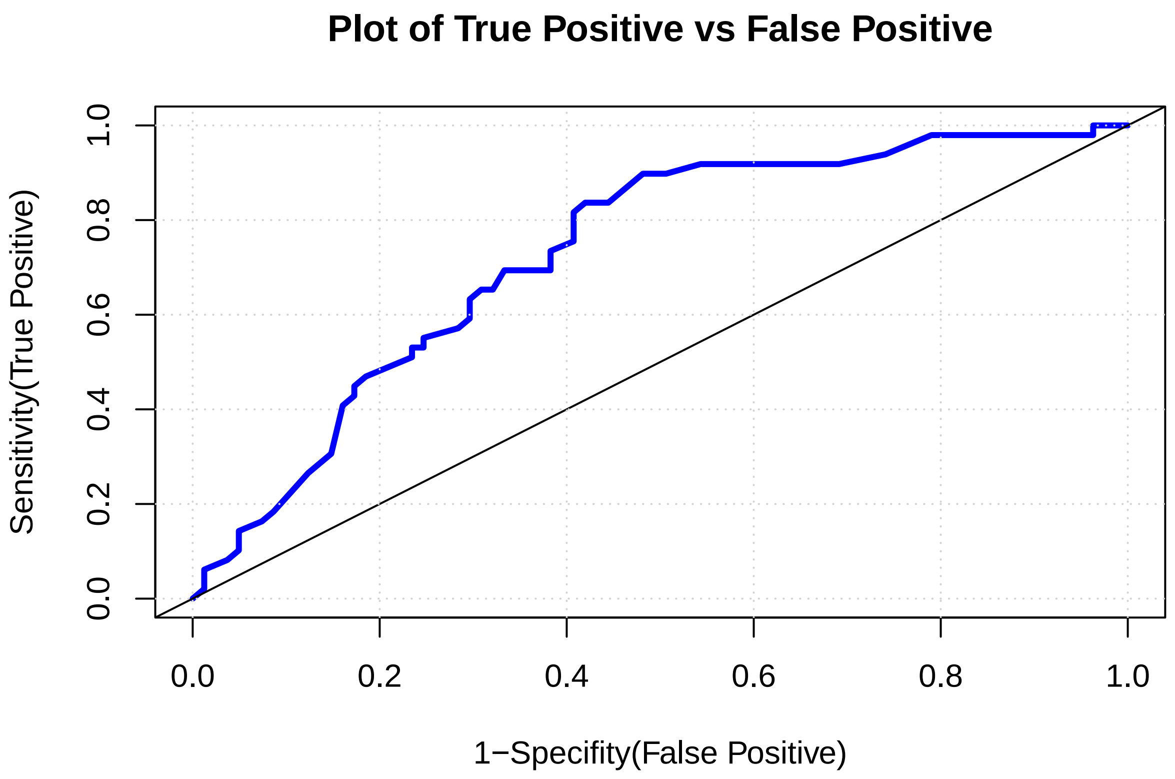
Answer:

Area under curve(AUC) is used to compare between models. The model with high AUC value consider to be better than one with low AUC value. In our models. AUC value for simple logistic regression from question no. 4 has 0.672 while AUC value for multiple regression from question no. 5 has 0.7297.Since, Area under curve for multiple regression has high value,that is why this model is better than simple logistic regression from 4.

Plotting ROC curve for multiple linear regression model:

```
roc_multi.model <- roc(data.module6$temp_level, predict(log.multi.model,
                                                        type = "response"))
plot(1-roc_multi.model$specificities,roc_multi.model$sensitivities,col="blue",
     main = "Plot of True Positive vs False Positive",type = "l",xlab = "1-Specifity(False Positive)",yl

abline(a=0,b=1)
grid()
```



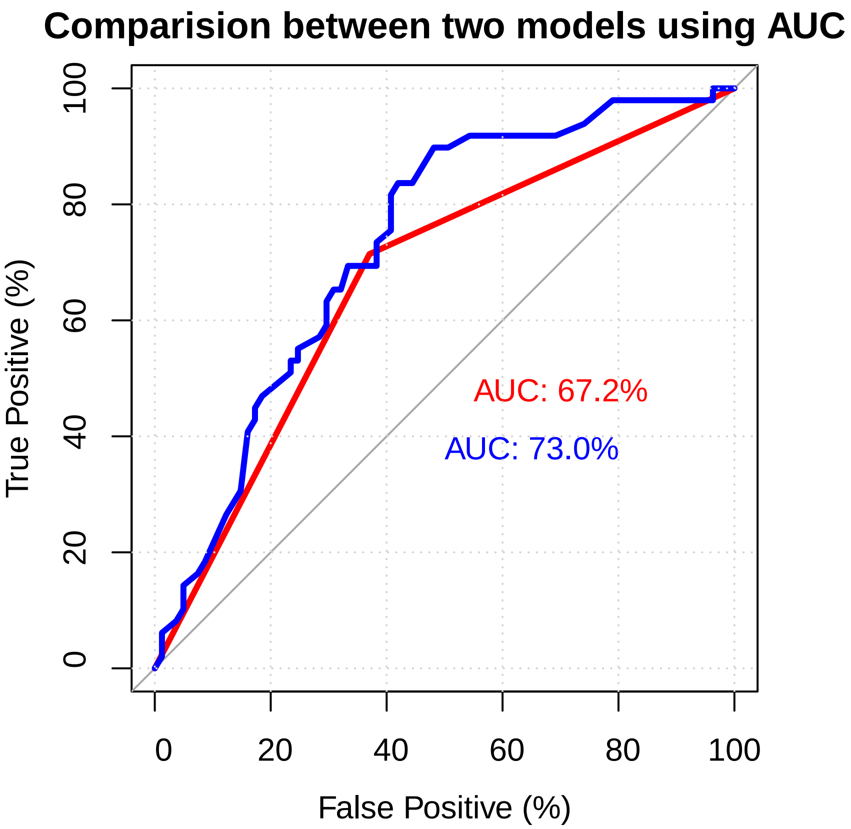
Alternatively,

we can plot AUC for both model to compare which one is bettwer and it seems that multiple logistic regression model has high AUC value and hence better model as shown below.

```
par(pty="s")
roc(data.module6$temp_level~predict(log.model,type = "response"),plot=TRUE,legacy.axes=T,percent=T,
    xlab="False Positive (%)",ylab="True Positive (%)",col="red",lwd=3,print.auc=T,print.auc.x=45,
    main="Comparision between two models using AUC")
```

```
##
## Call:
## roc.formula(formula = data.module6$temp_level ~ predict(log.model,     type = "response"), plot = TRU
##
## Data: predict(log.model, type = "response") in 81 controls (data.module6$temp_level 0) < 49 cases (da
## Area under the curve: 67.2%
```

```
plot.roc(data.module6$temp_level~predict(log.multi.model,type = "response"),col="blue",print.auc=T,lwd=
grid()
```


Comparision between two models using AUC
True Positive (%)
False Positive (%)
0
20
40
60
80
100
0
20
40
60
80
100
AUC: 67.2%
AUC: 73.0%

The End