Module 8 Assignment

Umakanth Sai Balguri - (U01403700)



Screenshot from Canvas

# Preprocessing

### Loading Library files

rm(list=ls())  
library(rio)  
library(car)

## Loading required package: carData

library(moments)  
library(robustHD)

## Loading required package: ggplot2

## Loading required package: perry

## Loading required package: parallel

## Loading required package: robustbase

### Loading the file into R and

patients=import("6304 Module 8 Assignment Data.xlsx")

### Fixing Coloumn names and creating a working Dataset with using U-number as seed

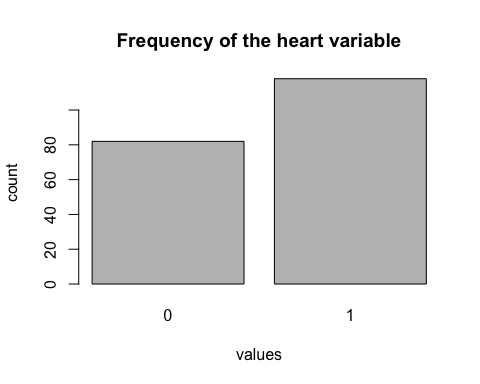
colnames(patients)=tolower(make.names(colnames(patients)))  
patients$chest.pain=as.factor(patients$chest.pain)  
patients$gender=as.factor(patients$gender)  
patients$exercise=as.factor(patients$exercise)  
patients$blood.sugar=as.factor(patients$blood.sugar)  
patients$heart=as.factor(patients$heart)  
set.seed(01403700)  
sample=patients[sample(1:nrow(patients),200,replace=FALSE),]  
attach(sample)

## The following object is masked from package:robustbase:  
##   
## heart

# Analysis

### Response to Q1

plot(sample$heart, main="Frequency of the heart variable", xlab="values", ylab="count")



### Response to Q2, Q3, Q4

sample.out=glm(heart~age+gender+chest.pain+blood.pressure+cholesterol+blood.sugar+max.heart.rate+exercise+diabetes,data=sample, family="binomial")  
summary(sample.out)

##   
## Call:  
## glm(formula = heart ~ age + gender + chest.pain + blood.pressure +   
## cholesterol + blood.sugar + max.heart.rate + exercise + diabetes,   
## family = "binomial", data = sample)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.30408 -0.05195 0.00848 0.15403 2.41673   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -36.601436 8.188564 -4.470 7.83e-06 \*\*\*  
## age 0.464298 0.100862 4.603 4.16e-06 \*\*\*  
## gendermale -1.426990 0.751785 -1.898 0.057678 .   
## chest.painmoderate -1.559125 0.960531 -1.623 0.104548   
## chest.painno\_pain -3.157196 1.052119 -3.001 0.002693 \*\*   
## chest.painsevere -1.746490 1.232415 -1.417 0.156446   
## blood.pressure 0.008396 0.017902 0.469 0.639075   
## cholesterol 0.016159 0.023498 0.688 0.491650   
## blood.sugar1 -0.901701 0.844420 -1.068 0.285595   
## max.heart.rate 0.089464 0.024473 3.656 0.000257 \*\*\*  
## exerciseyes -3.277553 1.051866 -3.116 0.001834 \*\*   
## diabetes NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 270.743 on 199 degrees of freedom  
## Residual deviance: 71.598 on 189 degrees of freedom  
## AIC: 93.598  
##   
## Number of Fisher Scoring iterations: 8

Yes, Residual Deviance is remarkably different from Null Deviance

### Response to Q5

Considering variables with significance level from Model in Part-2, “age” variable will have the greatest influence in increasing the modeled probability.

### Response to Q6

Considering variables with significance level from Model in Part-2, exercise= “Yes” variable will have the greatest influence in decreasing the modeled probability.

### Response to Q7

Yes, R hadn’t determines a beta coefficient in the model for the diabetes variable because of singularity i.e.diabetes variable has only one value for all the observations .

### Response to Q8,Q9

for(i in 1:nrow(sample)) {  
 if(sample$age[i]<40) {  
 sample$age.range[i]="Young"  
 }  
 else if (sample$age[i]>60) {  
 sample$age.range[i]="Older"  
 }  
 else {  
 sample$age.range[i]="Middle"  
   
 }  
 i=i+1  
 }  
sample$age.range=relevel(as.factor(sample$age.range), ref = "Young")

### Response to 10

sample.out.new=glm(heart~age.range+gender+chest.pain+blood.pressure+cholesterol+blood.sugar+max.heart.rate+exercise+diabetes,data=sample, family="binomial")  
summary(sample.out.new)

##   
## Call:  
## glm(formula = heart ~ age.range + gender + chest.pain + blood.pressure +   
## cholesterol + blood.sugar + max.heart.rate + exercise + diabetes,   
## family = "binomial", data = sample)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.08909 -0.21960 0.00001 0.00015 2.34638   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -16.89930 5.15973 -3.275 0.001056 \*\*   
## age.rangeMiddle 2.47807 1.91220 1.296 0.195002   
## age.rangeOlder 25.28694 1663.19838 0.015 0.987870   
## gendermale -1.54191 0.67197 -2.295 0.021755 \*   
## chest.painmoderate -0.41541 0.84078 -0.494 0.621256   
## chest.painno\_pain -2.17846 0.83207 -2.618 0.008842 \*\*   
## chest.painsevere -2.09129 1.17214 -1.784 0.074398 .   
## blood.pressure 0.02339 0.01928 1.213 0.225202   
## cholesterol 0.03211 0.02349 1.367 0.171655   
## blood.sugar1 -0.87749 0.79563 -1.103 0.270077   
## max.heart.rate 0.07799 0.02176 3.583 0.000339 \*\*\*  
## exerciseyes -2.11824 0.98099 -2.159 0.030828 \*   
## diabetes NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 270.743 on 199 degrees of freedom  
## Residual deviance: 81.103 on 188 degrees of freedom  
## AIC: 105.1  
##   
## Number of Fisher Scoring iterations: 19

Based on the new category of age variable, age middle has a positive influence on heart (i.e middle age is positively likely to have any kind of heart disease). This increases with age. For age range, >60 years i.e. old age the positive likely-hood is very high.