Exercise 13 - Thoraricsurgery

Kolekar, Shilpa

October 25th, 2020

## install.packages("foreign")  
library("foreign")

## Warning: package 'foreign' was built under R version 4.0.3

## Set the working directory   
setwd("C:/Users/shilp/Documents/GitHub/dsc520/data")  
  
## Load the `data/ThoraricSurgery` to  
surgery\_df <- read.arff("ThoraricSurgery.arff")  
summary(surgery\_df)

## DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9   
## DGN1: 1 Min. :1.440 Min. : 0.960 PRZ0:130 F:439 F:402 F:439   
## DGN2: 52 1st Qu.:2.600 1st Qu.: 1.960 PRZ1:313 T: 31 T: 68 T: 31   
## DGN3:349 Median :3.160 Median : 2.400 PRZ2: 27   
## DGN4: 47 Mean :3.282 Mean : 4.569   
## DGN5: 15 3rd Qu.:3.808 3rd Qu.: 3.080   
## DGN6: 4 Max. :6.300 Max. :86.300   
## DGN8: 2   
## PRE10 PRE11 PRE14 PRE17 PRE19 PRE25 PRE30 PRE32   
## F:147 F:392 OC11:177 F:435 F:468 F:462 F: 84 F:468   
## T:323 T: 78 OC12:257 T: 35 T: 2 T: 8 T:386 T: 2   
## OC13: 19   
## OC14: 17   
##   
##   
##   
## AGE Risk1Yr  
## Min. :21.00 F:400   
## 1st Qu.:57.00 T: 70   
## Median :62.00   
## Mean :62.53   
## 3rd Qu.:69.00   
## Max. :87.00   
##

## a. Fit a binary logistic regression model to the data set that predicts whether or not the patient survived for one year (the Risk1Y variable) after the surgery. Use the glm() function to perform the logistic regression. See Generalized Linear Models for an example. Include a summary using the summary() function in your results.

newModel <- glm(Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 + AGE, data = surgery\_df, family = binomial)  
summary(newModel)

##   
## Call:  
## glm(formula = Risk1Yr ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +   
## PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +   
## PRE32 + AGE, family = binomial, data = surgery\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6084 -0.5439 -0.4199 -0.2762 2.4929   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.655e+01 2.400e+03 -0.007 0.99450   
## DGNDGN2 1.474e+01 2.400e+03 0.006 0.99510   
## DGNDGN3 1.418e+01 2.400e+03 0.006 0.99528   
## DGNDGN4 1.461e+01 2.400e+03 0.006 0.99514   
## DGNDGN5 1.638e+01 2.400e+03 0.007 0.99455   
## DGNDGN6 4.089e-01 2.673e+03 0.000 0.99988   
## DGNDGN8 1.803e+01 2.400e+03 0.008 0.99400   
## PRE4 -2.272e-01 1.849e-01 -1.229 0.21909   
## PRE5 -3.030e-02 1.786e-02 -1.697 0.08971 .   
## PRE6PRZ1 -4.427e-01 5.199e-01 -0.852 0.39448   
## PRE6PRZ2 -2.937e-01 7.907e-01 -0.371 0.71030   
## PRE7T 7.153e-01 5.556e-01 1.288 0.19788   
## PRE8T 1.743e-01 3.892e-01 0.448 0.65419   
## PRE9T 1.368e+00 4.868e-01 2.811 0.00494 \*\*  
## PRE10T 5.770e-01 4.826e-01 1.196 0.23185   
## PRE11T 5.162e-01 3.965e-01 1.302 0.19295   
## PRE14OC12 4.394e-01 3.301e-01 1.331 0.18318   
## PRE14OC13 1.179e+00 6.165e-01 1.913 0.05580 .   
## PRE14OC14 1.653e+00 6.094e-01 2.713 0.00668 \*\*  
## PRE17T 9.266e-01 4.445e-01 2.085 0.03709 \*   
## PRE19T -1.466e+01 1.654e+03 -0.009 0.99293   
## PRE25T -9.789e-02 1.003e+00 -0.098 0.92227   
## PRE30T 1.084e+00 4.990e-01 2.172 0.02984 \*   
## PRE32T -1.398e+01 1.645e+03 -0.008 0.99322   
## AGE -9.506e-03 1.810e-02 -0.525 0.59944   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 395.61 on 469 degrees of freedom  
## Residual deviance: 341.19 on 445 degrees of freedom  
## AIC: 391.19  
##   
## Number of Fisher Scoring iterations: 15

1. According to the summary, which variables had the greatest effect on the survival rate?

By looking at the summary, PRE9T, PRE14OC14, PRE17T, and PRE30T variables have the greatest effect on the survival rate because their P value is 0.05 or below which means these variable are statistically significant.

1. To compute the accuracy of your model, use the dataset to predict the outcome variable. The percent of correct predictions is the accuracy of your model. What is the accuracy of your model?

## install.packages("caTools")  
library("caTools")

## Warning: package 'caTools' was built under R version 4.0.3

split <- sample.split(surgery\_df, SplitRatio = 0.8)  
train <- subset(surgery\_df, split == "TRUE")  
test <- subset(surgery\_df, split == "FALSE")  
# run the test data through model  
res <- predict(newModel, test, type= "response")  
res <- predict(newModel, train, type = "response")  
##validate the model confusion matrix  
  
confmatrix <- table(Actual\_Value = train$Risk1Yr, Predicted\_Value = res > 0.5)  
confmatrix

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 306 7  
## T 43 3

# Accuracy  
(confmatrix[[1,1]]+ confmatrix[[2,2]]) / sum(confmatrix)

## [1] 0.8607242

Accuracy of model is 86.07%.