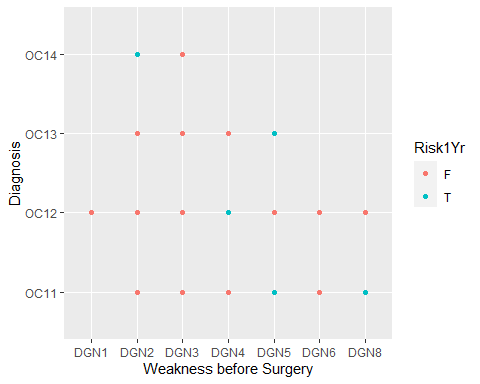
ASSIGNMENT 8 - Exercise 13: Fit a Logistic Regression Model to the Thoracic Surgery Binary Dataset

Ragunath Gunasekaran

2020-10-26

## Data Analysis Analysis

## Warning: package 'ggplot2' was built under R version 4.0.2



## Binary logistic regression model to the data set

1. 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.

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial, data = train.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4897 -0.5039 -0.3972 -0.2589 2.5788   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -15.87043 2399.54543 -0.007 0.99472   
## DGNDGN2 14.17874 2399.54480 0.006 0.99529   
## DGNDGN3 13.74844 2399.54478 0.006 0.99543   
## DGNDGN4 13.97917 2399.54483 0.006 0.99535   
## DGNDGN5 15.89783 2399.54486 0.007 0.99471   
## DGNDGN6 -0.03078 2768.68420 0.000 0.99999   
## DGNDGN8 34.43607 3393.46877 0.010 0.99190   
## PRE4 -0.15349 0.22220 -0.691 0.48969   
## PRE5 -0.02814 0.01873 -1.502 0.13313   
## PRE6PRZ1 -0.14579 0.64869 -0.225 0.82218   
## PRE6PRZ2 -0.03503 0.91426 -0.038 0.96943   
## PRE7T 0.84020 0.62418 1.346 0.17827   
## PRE8T 0.21402 0.46682 0.458 0.64662   
## PRE9T 1.91164 0.55991 3.414 0.00064 \*\*\*  
## PRE10T 0.62541 0.59916 1.044 0.29657   
## PRE11T 0.43165 0.47099 0.916 0.35942   
## PRE14OC12 0.24717 0.39834 0.621 0.53493   
## PRE14OC13 0.88279 0.74835 1.180 0.23814   
## PRE14OC14 1.62160 0.65809 2.464 0.01374 \*   
## PRE17T 0.88143 0.52079 1.692 0.09055 .   
## PRE19T -14.12968 2399.54475 -0.006 0.99530   
## PRE25T -0.87700 1.33369 -0.658 0.51081   
## PRE30T 1.05166 0.61645 1.706 0.08801 .   
## PRE32T -13.66563 1679.61076 -0.008 0.99351   
## AGE -0.02149 0.02169 -0.991 0.32182   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 284.38 on 352 degrees of freedom  
## Residual deviance: 238.68 on 328 degrees of freedom  
## AIC: 288.68  
##   
## Number of Fisher Scoring iterations: 15

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial, data = train.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4897 -0.5039 -0.3972 -0.2589 2.5788   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -15.87043 2399.54543 -0.007 0.99472   
## DGNDGN2 14.17874 2399.54480 0.006 0.99529   
## DGNDGN3 13.74844 2399.54478 0.006 0.99543   
## DGNDGN4 13.97917 2399.54483 0.006 0.99535   
## DGNDGN5 15.89783 2399.54486 0.007 0.99471   
## DGNDGN6 -0.03078 2768.68420 0.000 0.99999   
## DGNDGN8 34.43607 3393.46877 0.010 0.99190   
## PRE4 -0.15349 0.22220 -0.691 0.48969   
## PRE5 -0.02814 0.01873 -1.502 0.13313   
## PRE6PRZ1 -0.14579 0.64869 -0.225 0.82218   
## PRE6PRZ2 -0.03503 0.91426 -0.038 0.96943   
## PRE7T 0.84020 0.62418 1.346 0.17827   
## PRE8T 0.21402 0.46682 0.458 0.64662   
## PRE9T 1.91164 0.55991 3.414 0.00064 \*\*\*  
## PRE10T 0.62541 0.59916 1.044 0.29657   
## PRE11T 0.43165 0.47099 0.916 0.35942   
## PRE14OC12 0.24717 0.39834 0.621 0.53493   
## PRE14OC13 0.88279 0.74835 1.180 0.23814   
## PRE14OC14 1.62160 0.65809 2.464 0.01374 \*   
## PRE17T 0.88143 0.52079 1.692 0.09055 .   
## PRE19T -14.12968 2399.54475 -0.006 0.99530   
## PRE25T -0.87700 1.33369 -0.658 0.51081   
## PRE30T 1.05166 0.61645 1.706 0.08801 .   
## PRE32T -13.66563 1679.61076 -0.008 0.99351   
## AGE -0.02149 0.02169 -0.991 0.32182   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 284.38 on 352 degrees of freedom  
## Residual deviance: 238.68 on 328 degrees of freedom  
## AIC: 288.68  
##   
## Number of Fisher Scoring iterations: 15

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

DGNDGN8 had the greatest effect.

## Warning: package 'MASS' was built under R version 4.0.2

## [1] 0.6239316 0.7948718 0.7948718 0.8034188 0.8034188 0.7948718 0.7863248  
## [8] 0.7948718 0.7948718

## c. 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?

The highest accuracy of the model which we achieved was arround ~84%.

# References

1. Generalized Linear Models, Quick R by Datacamp - <https://www.statmethods.net/advstats/glm.html>
2. Thoracic Surgery by Meagan Londa - <https://rpubs.com/melonda/190569>