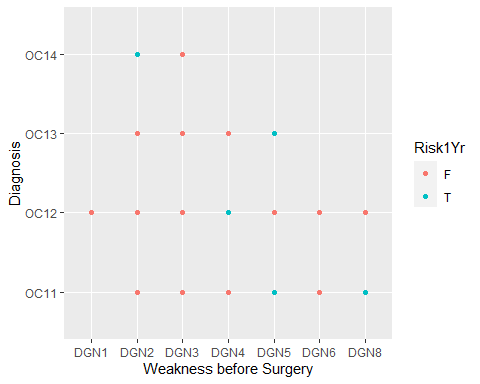
ASSIGNMENT 8 - Exercise 13: Fit a Logistic Regression Model to the Thoracic Surgery Binary Dataset

Ragunath Gunasekaran

2020-10-26

## Data Analysis Analysis

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



## 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.

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial, data = train.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5882 -0.5837 -0.4022 -0.2504 2.4831   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -15.87307 2399.54542 -0.007 0.99472   
## DGNDGN2 15.05462 2399.54479 0.006 0.99499   
## DGNDGN3 14.34478 2399.54477 0.006 0.99523   
## DGNDGN4 14.93272 2399.54480 0.006 0.99503   
## DGNDGN5 16.53422 2399.54487 0.007 0.99450   
## DGNDGN6 2.01894 3393.46880 0.001 0.99953   
## DGNDGN8 18.38268 2399.54524 0.008 0.99389   
## PRE4 -0.37213 0.21539 -1.728 0.08405 .   
## PRE5 -0.03792 0.01885 -2.012 0.04424 \*   
## PRE6PRZ1 -0.69628 0.56797 -1.226 0.22023   
## PRE6PRZ2 -0.64671 0.88525 -0.731 0.46506   
## PRE7T 1.48640 0.62104 2.393 0.01669 \*   
## PRE8T -0.03179 0.43738 -0.073 0.94206   
## PRE9T 1.58756 0.58735 2.703 0.00687 \*\*  
## PRE10T 0.53324 0.51547 1.034 0.30092   
## PRE11T 0.83255 0.44750 1.860 0.06282 .   
## PRE14OC12 0.69206 0.38856 1.781 0.07490 .   
## PRE14OC13 1.21290 0.68789 1.763 0.07786 .   
## PRE14OC14 1.41468 0.69221 2.044 0.04098 \*   
## PRE17T 1.05454 0.47604 2.215 0.02674 \*   
## PRE19T -15.48217 2399.54476 -0.006 0.99485   
## PRE25T 0.18444 1.11877 0.165 0.86905   
## PRE30T 1.23139 0.57487 2.142 0.03219 \*   
## PRE32T -14.16789 1592.62391 -0.009 0.99290   
## AGE -0.01768 0.02050 -0.863 0.38841   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 312.13 on 352 degrees of freedom  
## Residual deviance: 263.17 on 328 degrees of freedom  
## AIC: 313.17  
##   
## Number of Fisher Scoring iterations: 15

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial, data = train.data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5882 -0.5837 -0.4022 -0.2504 2.4831   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -15.87307 2399.54542 -0.007 0.99472   
## DGNDGN2 15.05462 2399.54479 0.006 0.99499   
## DGNDGN3 14.34478 2399.54477 0.006 0.99523   
## DGNDGN4 14.93272 2399.54480 0.006 0.99503   
## DGNDGN5 16.53422 2399.54487 0.007 0.99450   
## DGNDGN6 2.01894 3393.46880 0.001 0.99953   
## DGNDGN8 18.38268 2399.54524 0.008 0.99389   
## PRE4 -0.37213 0.21539 -1.728 0.08405 .   
## PRE5 -0.03792 0.01885 -2.012 0.04424 \*   
## PRE6PRZ1 -0.69628 0.56797 -1.226 0.22023   
## PRE6PRZ2 -0.64671 0.88525 -0.731 0.46506   
## PRE7T 1.48640 0.62104 2.393 0.01669 \*   
## PRE8T -0.03179 0.43738 -0.073 0.94206   
## PRE9T 1.58756 0.58735 2.703 0.00687 \*\*  
## PRE10T 0.53324 0.51547 1.034 0.30092   
## PRE11T 0.83255 0.44750 1.860 0.06282 .   
## PRE14OC12 0.69206 0.38856 1.781 0.07490 .   
## PRE14OC13 1.21290 0.68789 1.763 0.07786 .   
## PRE14OC14 1.41468 0.69221 2.044 0.04098 \*   
## PRE17T 1.05454 0.47604 2.215 0.02674 \*   
## PRE19T -15.48217 2399.54476 -0.006 0.99485   
## PRE25T 0.18444 1.11877 0.165 0.86905   
## PRE30T 1.23139 0.57487 2.142 0.03219 \*   
## PRE32T -14.16789 1592.62391 -0.009 0.99290   
## AGE -0.01768 0.02050 -0.863 0.38841   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 312.13 on 352 degrees of freedom  
## Residual deviance: 263.17 on 328 degrees of freedom  
## AIC: 313.17  
##   
## 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] 0.5641026 0.7008547 0.7692308 0.8119658 0.8205128 0.8461538 0.8547009  
## [8] 0.8547009 0.8632479

## 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 around ~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>