ASSIGNMENT 10.2

Sharath Chandra Tummanapally

2021-11-06

#1. Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset # a.For this problem, you will be working with the thoracic surgery data set from the University of California Irvine machine learning repository. This dataset contains information on life expectancy in lung cancer patients after surgery. The underlying thoracic surgery data is in ARFF format. This is a text-based format with information on each of the attributes. You can load this data using a package such as foreign or by cutting and pasting the data section into a CSV file.

# b.Assignment Instructions:

## 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 = Thoracic\_Surgery\_Data\_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   
## PRE7TRUE 7.153e-01 5.556e-01 1.288 0.19788   
## PRE8TRUE 1.743e-01 3.892e-01 0.448 0.65419   
## PRE9TRUE 1.368e+00 4.868e-01 2.811 0.00494 \*\*  
## PRE10TRUE 5.770e-01 4.826e-01 1.196 0.23185   
## PRE11TRUE 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 \*\*  
## PRE17TRUE 9.266e-01 4.445e-01 2.085 0.03709 \*   
## PRE19TRUE -1.466e+01 1.654e+03 -0.009 0.99293   
## PRE25TRUE -9.789e-02 1.003e+00 -0.098 0.92227   
## PRE30TRUE 1.084e+00 4.990e-01 2.172 0.02984 \*   
## PRE32TRUE -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

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

PRE9, PRE14OC14 has the greatest effect on the survival rate.

## 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?

## Loading required package: ggplot2

## Loading required package: lattice

## Predicted\_value  
## Actual\_value FALSE TRUE  
## FALSE 311 8  
## TRUE 55 3

## [1] 0.8328912

Accuracy is 83%.

#2. Fit a Logistic Regression Model # a.Fit a logistic regression model to the binary-classifier-data.csv dataset

# b.The dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.

## What is the accuracy of the logistic regression classifier?

##   
## Call:  
## glm(formula = label ~ x + y, family = binomial(), data = binarydata\_df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3728 -1.1697 -0.9575 1.1646 1.3989   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 0.424809 0.117224 3.624 0.00029 \*\*\*  
## x -0.002571 0.001823 -1.411 0.15836   
## y -0.007956 0.001869 -4.257 2.07e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2075.8 on 1497 degrees of freedom  
## Residual deviance: 2052.1 on 1495 degrees of freedom  
## AIC: 2058.1  
##   
## Number of Fisher Scoring iterations: 4

## Predicted\_Label  
## Actual\_Label 0 1  
## 0 429 338  
## 1 286 445

## [1] 0.5834446

Accuracy of the logistic regression classifier is 58%.

## Keep this assignment handy, as you will be comparing your results from this week to next week.