8.1 Thorasic Surgery Binary Data Set

Arindam Samanta

February 02 2020

### Source for the Data

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007â€“2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

#### Loading the required libraries for our analysis

library(foreign)  
library(ggplot2)  
library(dplyr)

#populating the housing\_data dataframe  
wd <- getwd()  
fname <- "ThoraricSurgery.arff"  
path\_to\_file <- paste(wd,'/dataset/',fname, sep = "")  
path\_to\_file  
  
my\_init\_df <- read.arff(path\_to\_file)  
glimpse(my\_init\_df)  
  
summary(my\_init\_df)  
  
head(my\_init\_df)

The data set has 470 observations and 17 predictors out of which the binary response variable Risk1Yr, shows 1 year survival period(T if died). The predictor variables constitutes both continuous and binary.

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. The code below estimates a logistic regression model using the glm(generalized linear model) function. The model uses all the predictors for the model.

## 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 0:400   
## 1st Qu.:57.00 1: 70   
## Median :62.00   
## Mean :62.53   
## 3rd Qu.:69.00   
## Max. :87.00   
##

##   
## Call:  
## glm(formula = Risk1Yr ~ ., family = binomial(link = "logit"),   
## data = survive)  
##   
## 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

## # A tibble: 25 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 DGNDGN8 18.0 2400. 0.00752 0.994   
## 2 DGNDGN5 16.4 2400. 0.00683 0.995   
## 3 DGNDGN2 14.7 2400. 0.00614 0.995   
## 4 DGNDGN4 14.6 2400. 0.00609 0.995   
## 5 DGNDGN3 14.2 2400. 0.00591 0.995   
## 6 PRE14OC14 1.65 0.609 2.71 0.00668  
## 7 PRE9T 1.37 0.487 2.81 0.00494  
## 8 PRE14OC13 1.18 0.617 1.91 0.0558   
## 9 PRE30T 1.08 0.499 2.17 0.0298   
## 10 PRE17T 0.927 0.444 2.08 0.0371   
## # … with 15 more rows

1. According to the summary, which variables had the greatest effect on the survival rate? Both summary and tidy function have been applied to the model. As per the output it seems Diagnosis 8 has the greatest effect on survival as it shows from the estimate the log odds of survival than any other diagnosis by 18.032.
2. 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?

Now using the regression model survive\_m we are trying to predict the outcome from the data set.

##   
## FALSE TRUE  
## 0 390 10  
## 1 67 3

## [1] 0.8361702

In the above prediction type = “response”, indicates to compute the response probability. In order to compute the confusion matrix, predict>0.5 means it returns 1 if the predicted probabilities are above 0.5, else 0. Now each row in a confusion matrix represents an actual target, while each column represents a predicted target. The model accuracy can be calculated by summing the TP+TN over the total observations.

The accuracy of the model shows here as 83%.