Exercise 13

Daniel Angel

February 7, 2020

Fit a Logistic Regression Model to the Thoracic Surgery Binary Dataset

Problem Statement : 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.

Data Set Information (From UCI website)

“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.”

```r
## Set working directory to root of your directory
setwd("C:/Users/Danny/Documents")

## Load the ‘foreign‘ library
library(foreign)

## Load the UC Irvine data set
thor_surg_df <- read.arff('data/ThoraricSurgery.arff')
head(thor_surg_df)
```

```
##    DGN PRE4 PRE5 PRE6 PRE7 PRE8 PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25 PRE30
## 1 DGN2 2.88 2.16 PRZ1    F    F    F     T     T  OC14     F     F     F     T
## 2 DGN3 3.40 1.88 PRZ0    F    F    F     F     F  OC12     F     F     F     T
## 3 DGN3 2.76 2.08 PRZ1    F    F    F     T     F  OC11     F     F     F     T
## 4 DGN3 3.68 3.04 PRZ0    F    F    F     F     F  OC11     F     F     F     F
## 5 DGN3 2.44 0.96 PRZ2    F    T    F     T     T  OC11     F     F     F     T
## 6 DGN3 2.48 1.88 PRZ1    F    F    F     T     F  OC11     F     F     F     F
##   PRE32 AGE Risk1Yr
## 1     F  60       F
## 2     F  51       F
## 3     F  59       F
## 4     F  54        F
## 5     F  73        T
## 6     F  51        F
```

Variable/Contributing Factors Details:

1. DGN: Diagnoses - combination of ICD-10 codes for primary and secondary and multiple tumors if applicable (DGN1-DGN6, DGN8)
2. PRE4: Forced vital capacity - FVC (Range is 1-6.5)
3. PRE5: Volume at end of the 1st second of forced exhalation - FEV1 (Range 0-90)
4. PRE6: Performance on Zubrod scale (PRZ2,PRZ1,PRZ0)
5. PRE7: Pain prior to surgery (True,False)
6. PRE8: Haemoptysis (Coughing up blood) prior to surgery (True,False)
7. PRE9: Dyspnoea (Difficulty Breathing / Shortness of Breath) prior to surgery (True,False)
8. PRE10: Cough prior to surgery (True,False)
9. PRE11: Weakness prior to surgery (True,False)
10. PRE14: Size of original tumor, from OC11 (smallest) to OC14 (largest)
11. PRE17: Type 2 Diabetes - diabetes mellitus (True,False)
12. PRE19: MI up to 6 months (True,False)
13. PRE25: PAD - Peripheral Arterial Diseases (True,False)
14. PRE30: Smoker (True,False)
15. PRE32: Asthma (True,False)
16. AGE: Age at surgery (In years, Range 20-90)
17. Risk1Y: 1 year survival period - (T)rue value if patient passed, (F)alse if still alive

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.

```
# This model includes all other parameters as dependent
lm1 <- glm(Risk1Yr ~ . , family ='binomial' , data = thor_surg_df)
summary(lm1)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ ., family = "binomial", data = thor_surg_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
```

```
# Trying out another model
# This model includes only DGN, Smoking, and Forced Exhale Volume as dependent
lm2 <- glm(Risk1Yr ~ DGN + PRE5 + PRE30, family =’binomial’ , data = thor_surg_df)
summary(lm2)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ DGN + PRE5 + PRE30, family = "binomial",
##     data = thor_surg_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2957  -0.5495  -0.5446  -0.3462   2.3923
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -16.47841 1455.39760  -0.011   0.9910
## DGNDGN2       14.51866 1455.39757   0.010   0.9920
## DGNDGN3       13.73306 1455.39754   0.009   0.9925
## DGNDGN4       13.99432 1455.39759   0.010   0.9923
## DGNDGN5       15.82019 1455.39764   0.011   0.9913
## DGNDGN6        0.19571 1623.14502   0.000   0.9999
## DGNDGN8       16.54764 1455.39829   0.011   0.9909
## PRE5           -0.01897    0.01712  -1.108   0.2679  
## PRE30T          0.96545    0.45948   2.101   0.0356 *
## ---
## 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: 372.03  on 461  degrees of freedom
## AIC: 390.03
##
## Number of Fisher Scoring iterations: 14
```

```
#Using info from the first linear model summary, focus this model to be dependent on only those factors
lm3 <- glm(Risk1Yr ~ PRE9 + PRE14 + PRE30 + PRE17  , family =’binomial’ , data = thor_surg_df)
summary(lm3)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE9 + PRE14 + PRE30 + PRE17, family = "binomial",
##     data = thor_surg_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max  
## -1.4318  -0.5496  -0.4601  -0.3614   2.4980  
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -3.0748     0.4732  -6.498 8.14e-11 ***
## PRE9T         1.0384     0.4434   2.342  0.01919 *  
## PRE14OC12     0.3790     0.3090   1.226  0.22004    
## PRE14OC13     1.2999     0.5735   2.267  0.02341 *  
## PRE14OC14     1.7493     0.5625   3.110  0.00187 ** 
## PRE30T        0.8821     0.4362   2.022  0.04316 *  
## PRE17T        1.0239     0.4174   2.453  0.01418 *  
## ---
## 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: 368.78  on 463  degrees of freedom
## AIC: 382.78
##
## Number of Fisher Scoring iterations: 5
```

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

Based on the summary of the first linear model which viewed all variables, only four variables had a p-value less than .05 which are in order of significance - ***PRE9, PRE14, PRE30, and PRE17***. A fifth variable, PRE5 is close to the level of significance which I seek but is slightly above the .05 threshold.

Therefore, ***Dyspnoea, Tumor Size, Cigarette Smoking, and Diabetes*** are the best factors at predicting if a patient will or will not survive one year post-operation.

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?

```
result1 <- predict(lm1,thor_surg_df,type = "response")
# Zero is False and non-zero values True so as in the data set true corresponds with someone who died.
confmatrix1 <- table(ActualValue=thor_surg_df$Risk1Yr, PredictedValue = result1 > 0.5)
confmatrix1
```

```
##            PredictedValue
## ActualValue FALSE TRUE
##           F   390   10
##           T    67    3
```

```
acc1 <- (confmatrix1[1,1]+confmatrix1[2,2])/sum(confmatrix1)
acc1
```

```
## [1] 0.8361702
```

```
result2 <- predict(lm2,thor_surg_df,type = "response")

confmatrix2 <- table(ActualValue=thor_surg_df$Risk1Yr, PredictedValue = result2 > 0.5)
confmatrix2
```

```
##            PredictedValue
## ActualValue FALSE TRUE
##           F   396    4
##           T    63    7
```

```
acc2 <- (confmatrix2[1,1]+confmatrix2[2,2])/sum(confmatrix2)
acc2
```

```
## [1] 0.8574468
```

```
result3 <- predict(lm3,thor_surg_df,type = "response")

confmatrix3 <- table(ActualValue=thor_surg_df$Risk1Yr, PredictedValue = result3 > 0.5)
confmatrix3
```

```
##            PredictedValue
## ActualValue FALSE TRUE
##           F   396    4
##           T    69    1
```

```
acc3 <- (confmatrix3[1,1]+confmatrix3[2,2])/sum(confmatrix3)
acc3
```

```
## [1] 0.8446809
```

I tested all 3 models and all of them under-predicted deaths. The first used all variables and was 83.6% accurate, the second focused on 3 variables and was 85.7% accurate, and the third used those four variables which were most significant from part B and was 84.4% accurate.

The second model was the most accurate and the only one to have better than 50% odds of being correct when it predicts a death; although viewed another way it was only able to predict 10% of all deaths.

The third model was overly optimistic and so although it correctly predicted the same amount of survivors as the 2nd it vastly underestimated the number of patients who would perish. Sometimes you must weigh your options considering the preference for tending towards more false positives or more false negatives.

The accuracy data for all was skewed by the fact that not many patients in the data set died, around fifteen percent.

It is interesting that cherry picking your variables as was the case in the 3rd model did not actually yield the best results but a combination of the obvious and not created the superior model. All of the models would have been improved by a larger training data set or a data set which included more people who would end up passing away a year after the surgery.