Exercise 13: Fit a Logistic Regression Model to the Thoracic Surgery Binary Data

Michael Hotaling

2020-10-20

Exercise 13: Fit a Logistic Regression Model to the Thoracic Surgery Binary Data

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.

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.

```r
library(foreign)
library(caTools)


df <- read.arff("ThoraricSurgery.arff")

set.seed(520)

sample <- sample.split(df$Risk1Yr, SplitRatio = 0.70)

train = subset(df, sample == TRUE)
test = subset(df, sample == FALSE)

model = glm(Risk1Yr ~ . -1 , family = binomial(logit), data = train)

model <- step(model, trace=FALSE);
summary(model)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE8 + PRE9 + PRE14 + PRE17 + PRE30 -
##     1, family = binomial(logit), data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6274  -0.5210  -0.4328  -0.2685   2.1975
```

```
## 
## Coefficients:
##            Estimate Std. Error z value Pr(>|z|)    
## PRE8F       -3.6972     0.6347  -5.825 5.71e-09 ***
## PRE8T       -3.1466     0.7139  -4.408 1.04e-05 ***
## PRE9T        1.3338     0.5522   2.415  0.01572 *  
## PRE14OC12    0.3925     0.3819   1.028  0.30402    
## PRE14OC13    1.1599     0.7001   1.657  0.09758 .  
## PRE14OC14    2.1385     0.6525   3.277  0.00105 ** 
## PRE17T       1.1974     0.4821   2.484  0.01299 *  
## PRE30T       1.3763     0.5774   2.383  0.01715 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 456.09  on 329  degrees of freedom
## Residual deviance: 247.36  on 321  degrees of freedom
## AIC: 263.36
## 
## Number of Fisher Scoring iterations: 5
```

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

1- PRE8F
2- PRE8T
3- PRE9T
4- PRE14OC13
5- PRE14OC14
6- PRE17T 7- PRE30T

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?

```r
library(pander)

test$predicted = predict(model, newdata=test, type="response")
pander(table(test$Risk1Yr, test$predicted> 0.5))
```

	FALSE	TRUE
F	118	2
T	20	1

The logistic model is good at predicting when values will be False, but bad at making True estimates. This is likely due to rank deficiency and, if provided more data with more T values, we may be able to create a model that is more robust.