Week 8: Assignment 13

Megan Holford

2020-10-24

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

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.

```
library(foreign)

thoracic_df <- read.arff(’data/ThoraricSurgery.arff’)

head(thoracic_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
```

```
thoracic_df$Risk1Yr<-relevel(thoracic_df$Risk1Yr, "T")

thoraciclm <- glm(Risk1Yr ~ . , family =’binomial’ , data = thoracic_df)

summary(thoraciclm)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ ., family = "binomial", data = thoracic_df)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.4929   0.2762   0.4199   0.5439   1.6084
##
## 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
```

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

1- PRE5
2- PRE9T
3- PRE14OC13
4- PRE14OC14
5- PRE17T 6- 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?

```
thoracic_predict <- predict(thoraciclm,thoracic_df,type = "response")

thoracic_cm <- table(ActualValue=thoracic_df$Risk1Yr, PredictedValue = thoracic_predict < 0.5)

thoracic_cm
```

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

The accuracy of this model is good for predicting false values. So this model is good for estimating when someone would live, but not for when someone would die (true values)