Exercise 10

Alissa Trujillo

2022-05-17

Exercise 10

Rmd file: https://github.com/alisssaa/dsc520/blob/master/completed/Exercise10/Exercise10.Rmd

1. Thoracic Surgery

a. Importing the Data

```r
setwd("/Users/alissa/Documents/Grad/DSC 520/dsc520")
library(foreign)
library(caTools)
thoracicSurgery <- read.arff("data/ThoracicSurgery.arff")
```

In order to make the data a bit easier to read, I am going to create a new data table including the pertinent information I will be using in my model. I will be assigning the variables more descriptive names so I am able to better do analysis.

I will also be converting the Risk1Y variable into a Survival variable. The Risk1Y determines whether a patient died within the first year after surgery. The Survival variable, on the other hand, will measure whether the patient survived the first year. I feel like this is a better baseline for my model, rather than having the baseline be measuring death. To do so, I must first convert the Risk1Y into a logical operator, as it was imported as a factor.

```r
thoracicSurgery$Risk1Yr <- as.logical(thoracicSurgery$Risk1Yr)
thoracicSurgery$Survival <- !thoracicSurgery$Risk1Yr

Survival <- thoracicSurgery$Survival
Risk1Y <- thoracicSurgery$Risk1Yr
FVC <- thoracicSurgery$PRE4
Perf <- thoracicSurgery$PRE6
Pain <- thoracicSurgery$PRE7
Haem <- thoracicSurgery$PRE8
Dysp <- thoracicSurgery$PRE9
Cough <- thoracicSurgery$PRE10
Weak <- thoracicSurgery$PRE11
Size <- thoracicSurgery$PRE14
Diab <- thoracicSurgery$PRE17
MI <- thoracicSurgery$PRE19
PAD <- thoracicSurgery$PRE25
```

```
Smoke <- thoracicSurgery$PRE30
Asthma <- thoracicSurgery$PRE32
Age <- thoracicSurgery$AGE

thoracicSurgeryX <- data.frame(Risk1Y, Survival, FVC, Perf,
    Pain, Haem, Dysp, Cough, Weak, Size, Diab, MI, PAD, Smoke,
    Asthma, Age)

thoracicSplit <- sample.split(thoracicSurgeryX, SplitRatio = 0.75)
thoracicTrain <- subset(thoracicSurgeryX, thoracicSplit ==
    "TRUE")
thoracicValidate <- subset(thoracicSurgeryX, thoracicSplit ==
    "FALSE")
```

b. Analysis

```
thoracic_Model.1 <- glm(Survival ~ FVC + Perf + Pain + Haem +
    Dysp + Cough + Weak + Size + Diab + MI + PAD + Smoke +
    Asthma + Age, data = thoracicTrain, family = binomial())

summary(thoracic_Model.1)
```

i. Logistic Regression

```
##
## Call:
## glm(formula = Survival ~ FVC + Perf + Pain + Haem + Dysp + Cough +
##     Weak + Size + Diab + MI + PAD + Smoke + Asthma + Age, family = binomial(),
##     data = thoracicTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6075   0.3282   0.4688   0.5497   2.0926
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.90715    1.61535   1.181  0.23775
## FVC           0.16261    0.21179   0.768  0.44261
## PerfPRZ1      0.59289    0.55733   1.064  0.28741
## PerfPRZ2     -0.11376    0.92410  -0.123  0.90202
## PainT        -0.27551    0.72841  -0.378  0.70526
## HaemT        -0.19915    0.44343  -0.449  0.65336
## DyspT        -0.54195    0.57695  -0.939  0.34755
## CoughT       -0.37790    0.54195  -0.697  0.48562
## WeakT        -0.42201    0.48159  -0.876  0.38087
## SizeOC12     -0.24460    0.35602  -0.687  0.49206
## SizeOC13     -1.03299    0.69623  -1.484  0.13789
## SizeOC14     -2.14219    0.79423  -2.697  0.00699 **
## DiabT        -1.20576    0.49171  -2.452  0.01420 *
## MIT          13.70576 1016.46792   0.013  0.98924
## PADT          -0.79550    0.90707  -0.877  0.38049
## SmokeT        -1.12277    0.57814  -1.942  0.05213 .
## AsthmaT       12.32950 1455.39769   0.008  0.99324
## Age            0.01241    0.02027   0.612  0.54037
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 294.80  on 351  degrees of freedom
## Residual deviance: 266.68  on 334  degrees of freedom
## AIC: 302.68
##
## Number of Fisher Scoring iterations: 14
```

ii. Variable Analysis We only had a few variables that had a significant effect on whether a patient survived the first year after surgery. The first two were in regards to the size of the tumor, OC13 and OC14. OC11 is the smallest tumor, which we used as the baseline factor for this variable. A size of OC12, the next largest tumor, did not seem to have a significant effect on survival. However, the two larger sizes of tumors OC13 and OC14 did. A tumor of size OC13 or OC14 was unfortunately significantly correlated with a patient's failure to survive the first year after their surgery, all other variables held constant. Both OC13 and OC14 were significant at a $p = 0.001$ level. Dyspnoea before surgery was also indicative of a patient not surviving the first year after surgery, statistically significant at a $p = 0.01$ level. Weakness before surgery was statistically significant as well, at a $p = 0.05$ level.

```r
thoracicTestV <- predict(thoracic_Model.1, thoracicValidate,
    type = "response")
thoracicTestT <- predict(thoracic_Model.1, thoracicTrain, type = "response")

confmatrix <- table(Actual_Value = thoracicTrain$Survival,
    Predicted_Value = thoracicTestT > 0.5)

confmatrix
```

iii. Model Accuracy

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##        FALSE     3   49
##        TRUE      2  298
```

After using my training data to train my model, I am able to plug in my validation data into my model to see how accurate it is. I split it 75/25 for the purposes of this assignment. Looking at the confusion matrix, we can see that 302 of the values were predicted correctly and 48 were predicted incorrectly. We can calculate the accuracy by dividing the correct responses by the total number of data points:

```r
(confmatrix[[1, 1]] + confmatrix[[2, 2]])/sum(confmatrix)
```

```
## [1] 0.8551136
```

The accuracy of the model is 85.6%. If we had guessed “true” for the one year survival rate for all patients, we would have a model that is accurate that is accurate for 86.4% of the data. Our model is unfortunately less accurate than simply assuming every patient survives the year.

2. Logistic Regression Model

a. Binary Classifier Data

Importing the Data

```
setwd("/Users/alissa/Documents/Grad/DSC 520/dsc520")
binary_df <- read.csv("data/binary-classifier-data.csv", header = TRUE)
```

Creating the Model

```
binarySplit <- sample.split(binary_df, SplitRatio = 0.75)
binaryTrain <- subset(binary_df, binarySplit == "TRUE")
binaryValidate <- subset(binary_df, binarySplit == "FALSE")

binary_Model.1 <- glm(label ~ x + y, data = binaryTrain, family = binomial())

summary(binary_Model.1)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial(), data = binaryTrain)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3766  -1.1693  -0.9522   1.1648   1.3896
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.433172   0.143853   3.011 0.002602 **
## x           -0.002722   0.002231  -1.220 0.222475
## y           -0.008017   0.002286  -3.507 0.000453 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1384.3  on 998  degrees of freedom
## Residual deviance: 1368.0  on 996  degrees of freedom
## AIC: 1374
##
## Number of Fisher Scoring iterations: 4
```

b. Accuracy

```
binaryTestV <- predict(binary_Model.1, binaryValidate, type = "response")
binaryTestT <- predict(binary_Model.1, binaryTrain, type = "response")

confmatrixB <- table(Actual_Value = binaryTrain$label, Predicted_Value = binaryTestT >
    0.5)

confmatrixB
```

```
##              Predicted_Value
## Actual_Value FALSE TRUE
##            0   283  229
##            1   190  297
```

```
(confmatrixB[[1, 1]] + confmatrixB[[2, 2]])/sum(confmatrixB)
```

```
## [1] 0.5805806
```

The accuracy of the logistic regression classifier is 58.5%. Simply guessing true or false would result in an accuracy of roughly 50%, so this is a slight improvement on the base model.