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

```
setwd("/Users/alissa/Documents/Grad/DSC 520/dsc520")
library(foreign)
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.

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

Survival <- thoracicSurgery$Survival
Risk1Y <- thoracicSurgery$Risk1Yr
Pain <- thoracicSurgery$PRE7
Cough <- thoracicSurgery$PRE10
Weak <- thoracicSurgery$PRE11
Size <- thoracicSurgery$PRE14
Smoke <- thoracicSurgery$PRE30
Age <- thoracicSurgery$AGE

thoracicSurgeryX <- data.frame(Risk1Y, Survival, Pain, Cough,
    Weak, Size, Smoke, Age)
```

b. Analysis

i. Logistic Regression The variables I chose were pain before surgery, cough, weakness, size of the tumor (with a baseline of OC11, the smallest and most common), whether the patient smokes, and age.

```
thoracic_Model.1 <- glm(Survival ~ Pain + Cough + Weak + Size +
    Smoke + Age, data = thoracicSurgeryX, family = binomial())

summary(thoracic_Model.1)
```

```
##
## Call:
## glm(formula = Survival ~ Pain + Cough + Weak + Size + Smoke +
##     Age, family = binomial(), data = thoracicSurgeryX)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5054   0.3936   0.4928   0.5852   1.2401
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.463965   1.094318   3.165  0.00155 **
## PainT       -0.548374   0.479563  -1.143  0.25284
## CoughT      -0.273954   0.331171  -0.827  0.40811
## WeakT       -0.492923   0.339637  -1.451  0.14669
## SizeOC12    -0.451806   0.310199  -1.457  0.14525
## SizeOC13    -1.265129   0.575681  -2.198  0.02798 *
## SizeOC14    -1.744860   0.566464  -3.080  0.00207 **
## SmokeT      -0.652147   0.434374  -1.501  0.13327
## Age         -0.006375   0.016114  -0.396  0.69238
## ---
## 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: 373.60  on 461  degrees of freedom
## AIC: 391.6
##
## Number of Fisher Scoring iterations: 5
```

ii. Variable Analysis We only had a few variables that had a significant effect on whether a patient survived the first year after surgery. They both 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. OC13 was significant at a $p = 0.05$ level while OC14 was significant at a $p = 0.01$ level. A couple of other variables (smoking, pain before surgery) have sizable effect sizes, but also have high standard errors so they are not reliable predictors.

```
thoracicTest <- thoracicSurgeryX
thoracicTest$model_prob <- predict(thoracic_Model.1, thoracicTest,
    type = "response")
thoracicTest$model_predict <- thoracicTest$model_prob >= 0.5
thoracicTest$model_correct <- thoracicTest$model_predict ==
    thoracicTest$Survival
model_accuracy <- sum(thoracicTest$model_correct)/nrow(thoracicTest)

model_accuracy
```

iii. Model Accuracy

```
## [1] 0.8468085
```

I used the predict function to predict the survival of each patient after the first year after surgery based on the model I designed. If the model predicted the patient was 50% or more likely to survive, I indicated that the model predicted survival. Then I compared the predicted survival to the patient's actual survival after a year. This is represented in the model_correct variable. I divided the model_correct by the number of total observations to discover that the model has an 84.7% accuracy rate.

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

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

summary(binary_Model.1)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial(), data = binary_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

b. Accuracy

```
binaryTest <- binary_df

binaryTest$model_prob <- predict(binary_Model.1, binaryTest,
    type = "response")
binaryTest$model_predict <- binaryTest$model_prob >= 0.5
binaryTest$model_correct <- binaryTest$model_predict == binaryTest$label
b_model_accuracy <- sum(binaryTest$model_correct)/nrow(binaryTest)

b_model_accuracy
```

```
## [1] 0.5834446
```

The accuracy of the logistic regression classifier is 58.3%.