Exercise_10.2_FigueroaHolly

Holly Figueroa

5/19/2021

```
knitr::opts_chunk$set(echo = TRUE)
```

Question 1b.i

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(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.5

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

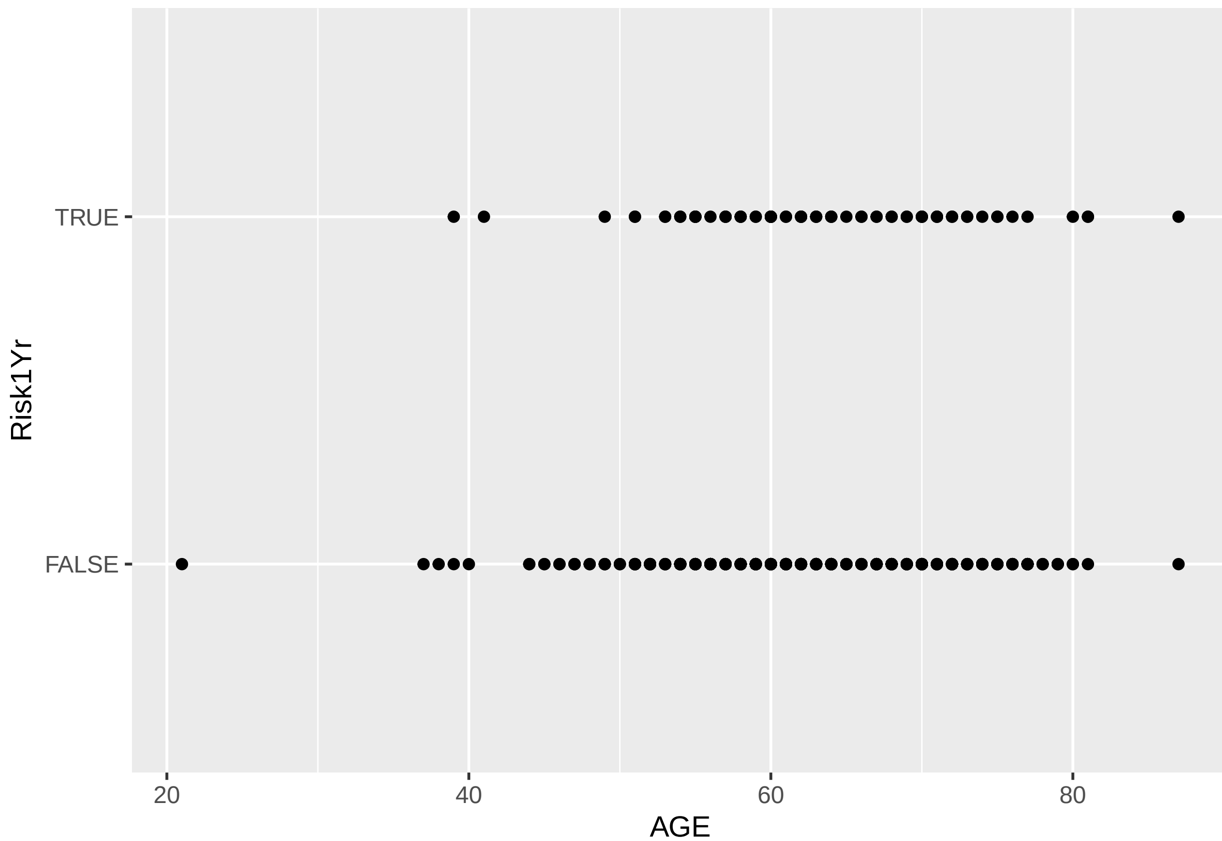
```
library(stats)
library(ggplot2)
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.0.5
```

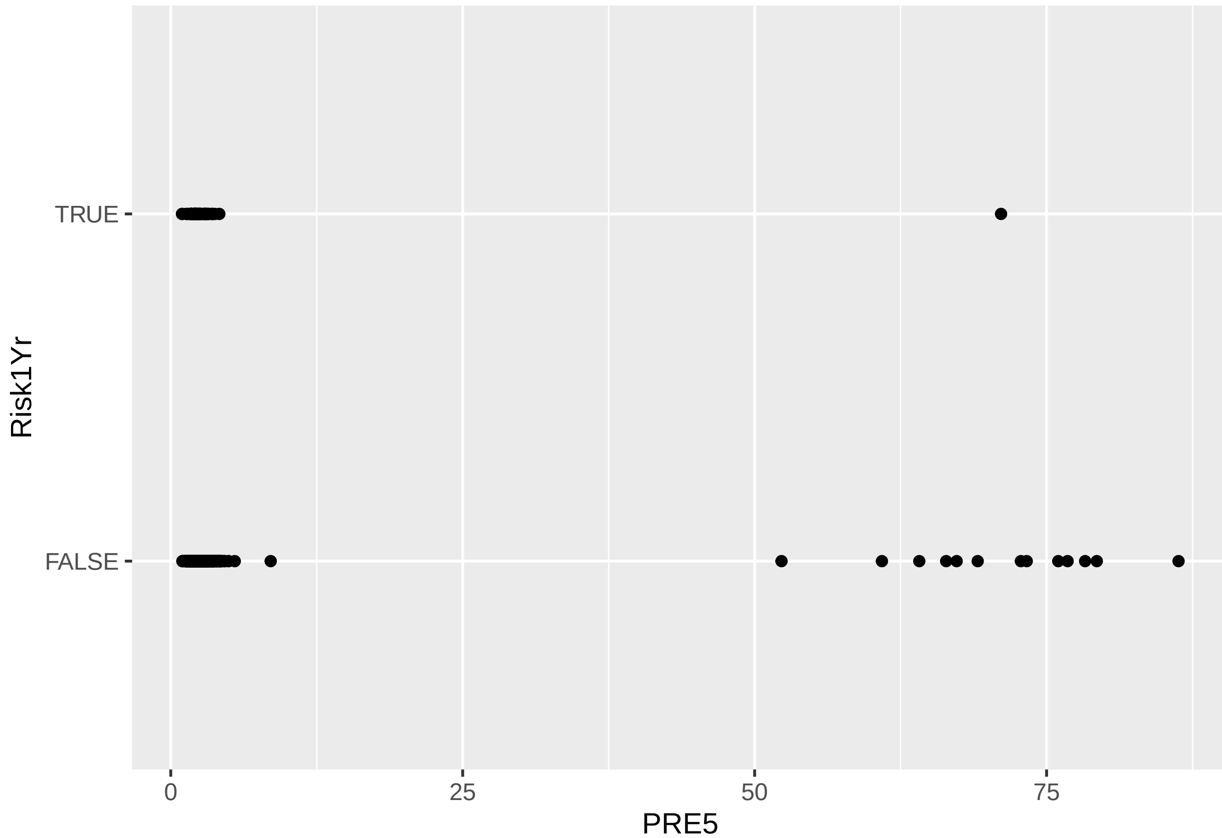
```
# Load/view csv file
surgery_data_orig=read.csv('thorasic_surgery.csv')

# Explore Data
head(surgery_data_orig)
```

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

```r
ggplot(surgery_data_orig, aes(AGE, Risk1Yr)) + geom_point()
```




```r
#PRE5 nearly all cases over the value 50 have FALSE Risk1Yr
ggplot(surgery_data_orig, aes(PRE5, Risk1Yr)) + geom_point()
```




```
surgery_data_orig$PRE5_group<-as.numeric(surgery_data_orig$PRE5 >= 50)
View(surgery_data_orig)

# Make model
surgery5_glm <- glm(Risk1Yr ~ PRE5_group + PRE6 + PRE9 + PRE17 + PRE30, data = surgery_data_orig, famil
summary(surgery5_glm)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE5_group + PRE6 + PRE9 + PRE17 + PRE30,
##     family = binomial(), data = surgery_data_orig)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1780  -0.5502  -0.5502  -0.3738   2.3933
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8052     0.4642  -6.043 1.51e-09 ***
## PRE5_group   -1.1283     1.0947  -1.031  0.30269
## PRE6PRZ1      0.1791     0.3344   0.536  0.59221
## PRE6PRZ2      0.8030     0.5426   1.480  0.13887
## PRE9TRUE      1.1889     0.4529   2.625  0.00866 **
## PRE17TRUE     1.0583     0.4100   2.581  0.00985 **
## PRE30TRUE     0.8148     0.4352   1.872  0.06116 .
```

```
## ---
## 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: 376.80  on 463  degrees of freedom
## AIC: 390.8
##
## Number of Fisher Scoring iterations: 5
```

Question 1b.ii

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

> The variables found to be associated with the largest standard deviation change in survival after one year, are called "PRE9" and "PRE17". They are also the most statistically significant according to the summary analysis with p values of 0.01.

1b.iii

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?

```
# Add column of probability of Risk1Yr based on model
surgery_data_orig$predicted_prob<-fitted(surgery5_glm)
head(surgery_data_orig)
```

```
##   id  DGN PRE4 PRE5 PRE6  PRE7  PRE8  PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25
## 1  1 DGN2 2.88 2.16 PRZ1 FALSE FALSE FALSE  TRUE  TRUE  OC14 FALSE FALSE FALSE
## 2  2 DGN3 3.40 1.88 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC12 FALSE FALSE FALSE
## 3  3 DGN3 2.76 2.08 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
## 4  4 DGN3 3.68 3.04 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC11 FALSE FALSE FALSE
## 5  5 DGN3 2.44 0.96 PRZ2 FALSE  TRUE FALSE  TRUE  TRUE  OC11 FALSE FALSE FALSE
## 6  6 DGN3 2.48 1.88 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
##   PRE30 PRE32 AGE Risk1Yr PRE5_group predicted_prob
## 1  TRUE FALSE  60   FALSE          0     0.14047903
## 2  TRUE FALSE  51   FALSE          0     0.12020985
## 3  TRUE FALSE  59   FALSE          0     0.14047903
## 4 FALSE FALSE  54   FALSE          0     0.05704306
## 5  TRUE FALSE  73    TRUE          0     0.23371271
## 6 FALSE FALSE  51   FALSE          0     0.06747828
```

```
# Add column of TRUE/FALSE predictions based on probability scores above .25
surgery_data_orig$predictionTF<-if_else(surgery_data_orig$predicted_prob > .25, TRUE, FALSE)
head(surgery_data_orig)
```

```
##   id  DGN PRE4 PRE5 PRE6  PRE7  PRE8  PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25
## 1  1 DGN2 2.88 2.16 PRZ1 FALSE FALSE FALSE  TRUE  TRUE  OC14 FALSE FALSE FALSE
## 2  2 DGN3 3.40 1.88 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC12 FALSE FALSE FALSE
```

```
## 3  3 DGN3 2.76 2.08 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
## 4  4 DGN3 3.68 3.04 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC11 FALSE FALSE FALSE
## 5  5 DGN3 2.44 0.96 PRZ2 FALSE  TRUE FALSE  TRUE  TRUE  OC11 FALSE FALSE FALSE
## 6  6 DGN3 2.48 1.88 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
##   PRE30 PRE32 AGE Risk1Yr PRE5_group predicted_prob predictionTF
## 1  TRUE FALSE  60   FALSE          0     0.14047903        FALSE
## 2  TRUE FALSE  51   FALSE          0     0.12020985        FALSE
## 3  TRUE FALSE  59   FALSE          0     0.14047903        FALSE
## 4 FALSE FALSE  54   FALSE          0     0.05704306        FALSE
## 5  TRUE FALSE  73    TRUE          0     0.23371271        FALSE
## 6 FALSE FALSE  51   FALSE          0     0.06747828        FALSE
```

```
# Choose probability threshold and compare model outcome with actual values
confmatrix <- table(actual_value = surgery_data_orig$Risk1Yr, Prediction = surgery_data_orig$predictionT
confmatrix
```

```
##             Prediction
## actual_value FALSE TRUE
##        FALSE   369   31
##        TRUE     55   15
```

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

```
## [1] 0.8170213
```

> After gaining probabilities based on our model, and choosing a threshold, testing shows that the model was approx 82% accurate.

2a.

Fit a logistic regression model to the binary-classifier-data.csv dataset.The dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables

```
binary_dataset <- read.csv('binary-classifier-data.csv')
head(binary_dataset)
```

```
##   label        x        y
## 1     0 70.88469 83.17702
## 2     0 74.97176 87.92922
## 3     0 73.78333 92.20325
## 4     0 66.40747 81.10617
## 5     0 69.07399 84.53739
## 6     0 72.23616 86.38403
```

```
library(mlogit)
```

```
## Warning: package 'mlogit' was built under R version 4.0.5
```

```
## Loading required package: dfidx
```

```
## Warning: package 'dfidx' was built under R version 4.0.5

##
## Attaching package: 'dfidx'

## The following object is masked from 'package:stats':
##
##     filter
```

```
binary_model <-glm(label ~ x + y, data = binary_dataset, family = binomial())
```

2b.i

What is the accuracy of the logistic regression classifier?

```
summary(binary_model)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial(), data = binary_dataset)
##
## 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
```

```
binary_dataset$pred_prob <-fitted(binary_model)
binary_dataset$pred_label<-if_else(binary_dataset$pred_prob >= .50
                                   , 1, 0)
confmatrix2 <- table(Actual_Label = binary_dataset$label, Predicted_Label = binary_dataset$pred_label)
confmatrix2
```

```
##             Predicted_Label
## Actual_Label   0   1
##            0 429 338
##            1 286 445
```

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

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

> Output for the accuracy went down when I adjusted the threshold below or above .50, leaving me to conclude the best threshold I could get was at .50 probability where that or over would be predicted as labeled 1 and under would be predicted as labeled 0. The accuracy for this model was only 58% suggesting the variables might not have a straight, linear relationship.

2b.ii

Keep this assignment handy, as you will be comparing your results from this week to next week.