

Thoracic Surgery and Binary Classifier Assignment

Theodore Koby-Hercsky

5/20/2021

html_document: <https://rpubs.com/theoKoby/772128>

Set the working directory to the root of my week 10 Assignment directory

setwd("~/Documents/Bellevue University Classes/DSC520/Week 10 Assignment")

The ThoracicSurgery.csv file and a summary of the data

I am importing readr from the library so I can use the read_csv function to create my Thoracic Surgery data frame.

```
library(readr)
```

Creating the Thoracic Surgery data frame by using the read_csv function to pull my Thoracic Surgery data.

```
ThoracicSurgery <- read_csv("ThoracicSurgery.csv")
```

Creating the Thoracic Surgery data frame by using the read_csv function to pull my Thoracic Surgery data.

```
View(ThoracicSurgery)
```

```
head(ThoracicSurgery)
```

```
## # A tibble: 6 x 17
```

```
##   Diagnosis    FVC    FEV1 Performance Pain  Haemoptysis Dyspnoea Cough Weakness
```

```
##   <dbl> <dbl> <dbl>          <dbl> <lgl> <lgl>          <lgl> <lgl> <lgl>
```

```
## 1      2  2.88  2.16          1 FALSE FALSE          FALSE TRUE  TRUE
```

```
## 2      3  3.4   1.88          0 FALSE FALSE          FALSE FALSE FALSE
```

```
## 3      3  2.76  2.08          1 FALSE FALSE          FALSE TRUE  FALSE
```

```
## 4      3  3.68  3.04          0 FALSE FALSE          FALSE FALSE FALSE
```

```
## 5      3  2.44  0.96          2 FALSE TRUE          FALSE TRUE  TRUE
```

```
## 6      3  2.48  1.88          1 FALSE FALSE          FALSE TRUE  FALSE
```

```
## # ... with 8 more variables: Tumor_Size <dbl>, Diabetes_Mellitus <lgl>,
```

```
## #   MI_6mo <lgl>, PAD <lgl>, Smoking <lgl>, Asthma <lgl>, Age <dbl>,
```

```
## #   Risk1Y <dbl>
```

As seen below we can use the summary function to analyze the descriptive statistics for this data set. As I have updated the names of all the variables and taken out unneeded characters in the csv file before uploading the file to my Rmarkdown report.

```
summary(ThoracicSurgery)
```

```
##   Diagnosis          FVC          FEV1          Performance
##   Min.   :1.000    Min.   :1.440    Min.    : 0.960    Min.     :0.0000
##   1st Qu.:3.000    1st Qu.:2.600    1st Qu.: 1.960    1st Qu.:0.0000
##   Median :3.000    Median :3.160    Median : 2.400    Median :1.0000
```

```
## Mean :3.096 Mean :3.282 Mean : 4.569 Mean :0.7809
## 3rd Qu.:3.000 3rd Qu.:3.808 3rd Qu.: 3.080 3rd Qu.:1.0000
## Max. :8.000 Max. :6.300 Max. :86.300 Max. :2.0000
## Pain Haemoptysis Dyspnoea Cough
## Mode :logical Mode :logical Mode :logical Mode :logical
## FALSE:439 FALSE:402 FALSE:439 FALSE:147
## TRUE :31 TRUE :68 TRUE :31 TRUE :323
##
##
## Weakness Tumor_Size Diabetes_Mellitus MI_6mo
## Mode :logical Min. :11.00 Mode :logical Mode :logical
## FALSE:392 1st Qu.:11.00 FALSE:435 FALSE:468
## TRUE :78 Median :12.00 TRUE :35 TRUE :2
## Mean :11.74
## 3rd Qu.:12.00
## Max. :14.00
## PAD Smoking Asthma Age
## Mode :logical Mode :logical Mode :logical Min. :21.00
## FALSE:462 FALSE:84 FALSE:468 1st Qu.:57.00
## TRUE :8 TRUE :386 TRUE :2 Median :62.00
## Mean :62.53
## 3rd Qu.:69.00
## Max. :87.00
## Risk1Y
## Min. :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean :0.1489
## 3rd Qu.:0.0000
## Max. :1.0000
```

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.

I fit a binary logistic regression model to determine if the tumor size was 12 or above which will indicate a True.

```
ThoracicSurgery$Patiet_Predicts_Survival <- with(ThoracicSurgery, Tumor_Size >= 12 & Risk1Y >= 1)
```

to the data from the ThoracicSurgery data frame by using an glm() function to perform a logistic regression.

```
patient_survival_regression <- glm(Patiet_Predicts_Survival ~ Age + Asthma + Smoking + PAD + MI_6mo + Diabetes_Mellitus + Tumor_Size + Weakness + Cough + Dyspnoea + Haemoptysis + Pain + Performance + FEV1 + FVC + Diagnosis + Risk1Y, data = ThoracicSurgery, family = binomial(link = "logit"))
```

As seen below in the summary we see the Number of Fisher Scoring iterations being 25. While the Null deviance: 3.2698e+02 on 469 degrees of freedom which shows how well the response variable is predicted by a model that includes only the intercept.

```
summary(patient_survival_regression)
```

```
##
## Call:
## glm(formula = Patient_Predicts_Survival ~ Age + Asthma + Smoking +
##      PAD + MI_6mo + Diabetes_Mellitus + Tumor_Size + Weakness +
##      Cough + Dyspnoea + Haemoptysis + Pain + Performance + FEV1 +
##      FVC + Diagnosis + Risk1Y, family = binomial(link = "logit"),
##      data = ThoracicSurgery)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.503e-05 -2.100e-08 -2.100e-08 -2.100e-08  4.573e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.322e+02  1.744e+05  -0.004    0.997
## Age             4.869e-02  8.099e+02   0.000    1.000
## AsthmaTRUE      8.507e+01  2.237e+05   0.000    1.000
## SmokingTRUE    -4.307e+00  5.247e+04   0.000    1.000
## PADTRUE        -1.488e+00  2.667e+04   0.000    1.000
## MI_6moTRUE      8.563e+01  2.218e+05   0.000    1.000
## Diabetes_MellitusTRUE 4.437e-01  1.378e+04   0.000    1.000
## Tumor_Size      4.387e+01  1.112e+04   0.004    0.997
## WeaknessTRUE   -1.170e+00  1.301e+04   0.000    1.000
## CoughTRUE       1.084e+00  2.254e+04   0.000    1.000
## DyspnoeaTRUE    1.588e+00  2.373e+04   0.000    1.000
## HaemoptysisTRUE -4.300e-01  1.265e+04   0.000    1.000
## PainTRUE        2.566e-01  2.172e+04   0.000    1.000
## Performance    -2.475e-01  1.533e+04   0.000    1.000
## FEV1            9.586e-01  1.556e+03   0.001    1.000
## FVC            -1.274e+00  8.571e+03   0.000    1.000
## Diagnosis       -5.902e-01  7.682e+03   0.000    1.000
## Risk1Y          1.325e+02  3.052e+04   0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3.2698e+02  on 469  degrees of freedom
## Residual deviance: 3.2981e-08  on 452  degrees of freedom
## AIC: 36
##
## Number of Fisher Scoring iterations: 25
```

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

As seen below we use the summary function to show a summary of our logistic regression.

In the summary function we see that the variables SmokingTRUE, HaemoptysisTRUE, and Diagnosis had the greatest negative effect while the variables Age, AsthmaTRUE, MI_6moTRUE, Diabetes_MellitusTRUE, Tumor_Size, and FEV1 greatest positive effect on the survival rate as seen in the estimate.

```
summary(patient_survival_regression)

##
## Call:
## glm(formula = Patient_Predicts_Survival ~ Age + Asthma + Smoking +
##      PAD + MI_6mo + Diabetes_Mellitus + Tumor_Size + Weakness +
##      Cough + Dyspnoea + Haemoptysis + Pain + Performance + FEV1 +
##      FVC + Diagnosis + Risk1Y, family = binomial(link = "logit"),
##      data = ThoracicSurgery)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -4.503e-05 -2.100e-08 -2.100e-08 -2.100e-08  4.573e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.322e+02  1.744e+05  -0.004    0.997
## Age             4.869e-02  8.099e+02   0.000    1.000
## AsthmaTRUE      8.507e+01  2.237e+05   0.000    1.000
## SmokingTRUE    -4.307e+00  5.247e+04   0.000    1.000
## PADTRUE        -1.488e+00  2.667e+04   0.000    1.000
## MI_6moTRUE      8.563e+01  2.218e+05   0.000    1.000
## Diabetes_MellitusTRUE 4.437e-01  1.378e+04   0.000    1.000
## Tumor_Size      4.387e+01  1.112e+04   0.004    0.997
## WeaknessTRUE   -1.170e+00  1.301e+04   0.000    1.000
## CoughTRUE       1.084e+00  2.254e+04   0.000    1.000
## DyspnoeaTRUE    1.588e+00  2.373e+04   0.000    1.000
## HaemoptysisTRUE -4.300e-01  1.265e+04   0.000    1.000
## PainTRUE        2.566e-01  2.172e+04   0.000    1.000
## Performance    -2.475e-01  1.533e+04   0.000    1.000
## FEV1            9.586e-01  1.556e+03   0.001    1.000
## FVC            -1.274e+00  8.571e+03   0.000    1.000
## Diagnosis       -5.902e-01  7.682e+03   0.000    1.000
## Risk1Y          1.325e+02  3.052e+04   0.004    0.997
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 3.2698e+02  on 469  degrees of freedom
## Residual deviance: 3.2981e-08  on 452  degrees of freedom
```

```
## AIC: 36
##
## Number of Fisher Scoring iterations: 25
```

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?

First I will pull up the summary of my updated ThoracicSurgery
summary(ThoracicSurgery)

```
##      Diagnosis      FVC      FEV1      Performance
## Min.   :1.000   Min.   :1.440   Min.    : 0.960   Min.     :0.0000
## 1st Qu.:3.000   1st Qu.:2.600   1st Qu.: 1.960   1st Qu.:0.0000
## Median :3.000   Median :3.160   Median : 2.400   Median :1.0000
## Mean   :3.096   Mean   :3.282   Mean    : 4.569   Mean    :0.7809
## 3rd Qu.:3.000   3rd Qu.:3.808   3rd Qu.: 3.080   3rd Qu.:1.0000
## Max.   :8.000   Max.   :6.300   Max.    :86.300   Max.    :2.0000
##      Pain      Haemoptysis      Dyspnoea      Cough
## Mode :logical   Mode :logical   Mode :logical   Mode :logical
## FALSE:439      FALSE:402      FALSE:439      FALSE:147
## TRUE :31        TRUE :68        TRUE :31        TRUE :323
##
##
##      Weakness      Tumor_Size      Diabetes_Mellitus      MI_6mo
## Mode :logical   Min.    :11.00   Mode :logical   Mode :logical
## FALSE:392      1st Qu.:11.00   FALSE:435      FALSE:468
## TRUE :78        Median :12.00   TRUE :35        TRUE :2
##                  Mean     :11.74
##                  3rd Qu.:12.00
##                  Max.     :14.00
##      PAD      Smoking      Asthma      Age
## Mode :logical   Mode :logical   Mode :logical   Min.    :21.00
## FALSE:462      FALSE:84        FALSE:468      1st Qu.:57.00
## TRUE :8         TRUE :386       TRUE :2         Median :62.00
##                  Mean     :62.53
##                  3rd Qu.:69.00
##                  Max.     :87.00
##      Risk1Y      Patient_Predicts_Survival
## Min.    :0.0000   Mode :logical
## 1st Qu.:0.0000   FALSE:418
## Median :0.0000   TRUE :52
## Mean    :0.1489
## 3rd Qu.:0.0000
## Max.    :1.0000
```

*## Next we can calculate the amount of Risk1Y that was a 1 for died within a year. This is calculated by taking the 470 amount of lines and multiply it by the mean 0.1489 which is seen as 470 * 0.1489 = 69.983. Which we see 70.*

```

data_set_deaths <- 470*0.1489
data_set_deaths

## [1] 69.983

## While the predicted amount from our Patiet_Predicts_Survival shows 52 deaths seen as the TRUE amount from our summary.
Patiet_Predicts_Survival_amount <- 52
Patiet_Predicts_Survival_amount

## [1] 52

## Finally we will take the Patiet_Predicts_Survival_amount and divide by the data_set_deaths which will give us the percent of accuracy of the model.
percent_of_accuracy <- Patiet_Predicts_Survival_amount/data_set_deaths
percent_of_accuracy

## [1] 0.7430376

```

The binary-classifier-data.csv file and a summary of the data

```

## I am importing readr from the library so I can use the read_csv function to create my binary-classifier data frame.
library(readr)

## Creating the binary-classifier data frame by using the read_csv function to pull my binary-classifier data.
binary_classifier_data <- read_csv("data/binary-classifier-data.csv")
## Creating the binary-classifier data frame by using the read_csv function to pull my binary-classifier data.
View(binary_classifier_data)
head(binary_classifier_data)

## # A tibble: 6 x 3
##   label     x     y
##   <dbl> <dbl> <dbl>
## 1     0  70.9  83.2
## 2     0  75.0  87.9
## 3     0  73.8  92.2
## 4     0  66.4  81.1
## 5     0  69.1  84.5
## 6     0  72.2  86.4

```

Fit a logistic regression model to the binary-classifier-data.csv dataset

```

## I fit a logistic regression model to determine if the x variable is greater than or equal to 32 and y variable is greater than or equal to 45 which will show as true.
binary_classifier_data$label_regression <- with(binary_classifier_data, x >= 32 & y >= 45)

## I am going to use the glm() function to fit a logistic regression model with my new label regression variable.
binary_classifier_regression <- glm(label_regression ~ label + x + y, data =

```

```
binary_classifier_data, family = binomial())

## As seen in our logistic regression model we see that the AIC is 292.94 and
## we have a Null Deviance of 1913.40 on 1497 degrees of freedom.
summary(binary_classifier_regression)

##
## Call:
## glm(formula = label_regression ~ label + x + y, family = binomial(),
##      data = binary_classifier_data)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -3.6378  -0.0647  -0.0237   0.0159   2.1631
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -17.01254    1.29728  -13.114  <2e-16 ***
## label        -0.60551    0.39479   -1.534    0.125
## x             0.14295    0.01303   10.970  <2e-16 ***
## y             0.18431    0.01364   13.513  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1913.40  on 1497  degrees of freedom
## Residual deviance:  284.94  on 1494  degrees of freedom
## AIC: 292.94
##
## Number of Fisher Scoring iterations: 8
```

What is the accuracy of the logistic regression classifier?

```
summary(binary_classifier_data)

##      label      x      y      label_regression
## Min.   :0.000 Min.   : -5.20 Min.   : -4.019 Mode :logical
## 1st Qu.:0.000 1st Qu.: 19.77 1st Qu.: 21.207 FALSE:994
## Median :0.000 Median : 41.76 Median : 44.632 TRUE :504
## Mean   :0.488 Mean   : 45.07 Mean   : 45.011
## 3rd Qu.:1.000 3rd Qu.: 66.39 3rd Qu.: 68.698
## Max.   :1.000 Max.   :104.58 Max.   :106.896

## I am going to view the summary of the binary_classifier_data and compare t
## he mean of our labels to the percentage of true in the label regression.
number_of_values_in_label_regression <- 504 + 994
percentage_of_true_label_regression <- 504/1498
percentage_of_true_label_regression

## [1] 0.3364486
```

This shows us that our accuracy in comparison to our label variable is less due to the label being 0.488 while the label_regression percent is 0.3364.