10.2Exercises.R

Rahul Rajeev

2023-02-15

```
# Assignment: 10.2 Exercises - Thoracic Surgery Binary Logistic Model
# Name: Rajeev, Rahul
# Date: 2023-02-13

## Set the working directory to the root of your DSC 520 directory
library(ggplot2)
library(foreign)
library(dplyr)

##
## 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(stringr)
library(caTools)
theme_set(theme_minimal())

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/rahul/Documents/Bellevue/DSC 520")

## loading data
surgery_df<- read.arff("data/ThoracicSurgery.arff")
head(surgery_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

# adjusting the non-numerical values to numerical values except for the ones
# that are binary (diagnosis from DGN, zubrod scale from PRE6, and tumor size
# from PRE14). Everything else should be fine.

adjusted_surgery_df <- surgery_df %>%
  mutate(diagnosis = str_sub(DGN, -1), zubrod = str_sub(PRE6, -1),
         tumor_size = str_sub(PRE14, -2, -1)) %>%
  select(diagnosis, PRE4, PRE5, zubrod, PRE7, PRE8, PRE9, PRE10, PRE11,
tumor_size,
         PRE17, PRE19, PRE25, PRE30, PRE32, AGE, Risk1Yr)

head(adjusted_surgery_df)

##   diagnosis PRE4 PRE5 zubrod PRE7 PRE8 PRE9 PRE10 PRE11 tumor_size PRE17
PRE19
## 1         2 2.88 2.16      1    F    F    F     T     T         14     F
F
## 2         3 3.40 1.88      0    F    F    F     F     F         12     F
F
## 3         3 2.76 2.08      1    F    F    F     T     F         11     F
F
## 4         3 3.68 3.04      0    F    F    F     F     F         11     F
F
## 5         3 2.44 0.96      2    F    T    F     T     T         11     F
F
## 6         3 2.48 1.88      1    F    F    F     T     F         11     F
F
##   PRE25 PRE30 PRE32 AGE Risk1Yr
## 1     F     T     F  60       F
## 2     F     T     F  51       F
## 3     F     T     F  59       F
## 4     F     F     F  54       F
## 5     F     T     F  73       T
## 6     F     F     F  51       F

# general linear model for binary logistic
binary_log <- glm(Risk1Yr ~ diagnosis + PRE4 + PRE5 + zubrod + PRE7 + PRE8 +
                    PRE9 + PRE10 + PRE11 + tumor_size + PRE17 + PRE19 + PRE25
+
```

```
                    PRE30 + PRE32 + AGE, family = 'binomial',
data=adjusted_surgery_df)
summary(binary_log)

##
## Call:
## glm(formula = Risk1Yr ~ diagnosis + PRE4 + PRE5 + zubrod + PRE7 +
##     PRE8 + PRE9 + PRE10 + PRE11 + tumor_size + PRE17 + PRE19 +
##     PRE25 + PRE30 + PRE32 + AGE, family = "binomial", data =
adjusted_surgery_df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6084  -0.5439  -0.4199  -0.2762   2.4929
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.655e+01  2.400e+03  -0.007  0.99450
## diagnosis2    1.474e+01  2.400e+03   0.006  0.99510
## diagnosis3    1.418e+01  2.400e+03   0.006  0.99528
## diagnosis4    1.461e+01  2.400e+03   0.006  0.99514
## diagnosis5    1.638e+01  2.400e+03   0.007  0.99455
## diagnosis6    4.089e-01  2.673e+03   0.000  0.99988
## diagnosis8    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 .
## zubrod1      -4.427e-01  5.199e-01  -0.852  0.39448
## zubrod2      -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
## tumor_size12  4.394e-01  3.301e-01   1.331  0.18318
## tumor_size13  1.179e+00  6.165e-01   1.913  0.05580 .
## tumor_size14  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

# according to the summary, teh variables that had the greatest effect on
survival
# were PRE9 with dysnopea before surgery, PRE14 with tumor size of 14, PRE17
with
# type 2 diabetes mellitus, and PRE30 with smoking. With meaning that the
# condition was true.

# creating test dataset and testing against model
split <- sample.split(adjusted_surgery_df, SplitRatio = 0.65)
train <- subset(adjusted_surgery_df, split == TRUE)
test <- subset(adjusted_surgery_df, split == FALSE)
my_model <- glm(Risk1Yr ~ diagnosis + PRE4 + PRE5 + zubrod + PRE7 + PRE8 +
                PRE9 + PRE10 + PRE11 + tumor_size + PRE17 + PRE19 + PRE25 +
                PRE30 + PRE32 + AGE, family = 'binomial', data=train)
summary(my_model)

##
## Call:
## glm(formula = Risk1Yr ~ diagnosis + PRE4 + PRE5 + zubrod + PRE7 +
##     PRE8 + PRE9 + PRE10 + PRE11 + tumor_size + PRE17 + PRE19 +
##     PRE25 + PRE30 + PRE32 + AGE, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0872  -0.5326  -0.3575  -0.2186   2.3631
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -13.98103 2399.54557  -0.006 0.995351
## diagnosis2     15.51206 2399.54482   0.006 0.994842
## diagnosis3     14.19978 2399.54478   0.006 0.995278
## diagnosis4     15.07615 2399.54481   0.006 0.994987
## diagnosis5     16.81641 2399.54514   0.007 0.994408
## diagnosis6      0.84408 2663.91199   0.000 0.999747
## diagnosis8     18.30895 2399.54528   0.008 0.993912
## PRE4           -0.46821    0.25193  -1.859 0.063097 .
## PRE5           -0.02799    0.02035  -1.375 0.169089
## zubrod1        -1.25901    0.70303  -1.791 0.073322 .
## zubrod2        -0.18657    1.05049  -0.178 0.859032
## PRE7T           0.95312    0.90086   1.058 0.290051
## PRE8T          -0.25330    0.58976  -0.430 0.667558
## PRE9T           1.20379    0.64853   1.856 0.063428 .
## PRE10T          1.10948    0.69395   1.599 0.109867
## PRE11T          0.28626    0.59118   0.484 0.628233
## tumor_size12    0.79439    0.45365   1.751 0.079926 .
## tumor_size13    1.60889    0.80617   1.996 0.045964 *
## tumor_size14    2.73471    0.77920   3.510 0.000449 ***
```

```
## PRE17T          0.78070    0.62359   1.252 0.210594
## PRE19T        -14.94655 2399.54479  -0.006 0.995030
## PRE25T         -0.12000    1.25896  -0.095 0.924062
## PRE30T          1.25281    0.72455   1.729 0.083793 .
## PRE32T        -14.15236 1595.19837  -0.009 0.992921
## AGE            -0.04475    0.02360  -1.897 0.057874 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 254.92  on 303  degrees of freedom
## Residual deviance: 201.24  on 279  degrees of freedom
## AIC: 251.24
##
## Number of Fisher Scoring iterations: 15

# running the test data through the model
res <- predict(my_model, test, type = 'response')
res <- predict(my_model, train, type='response')

# confusion matrix
length(train$Risk1Yr)

## [1] 304

length(res[res > 0.5])

## [1] 17

confmatrix <- table(Actual_Value = train$Risk1Yr, Predicted_Value = res >
0.5)
confmatrix

##             Predicted_Value
## Actual_Value FALSE TRUE
##            F   251    8
##            T    36    9

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

## [1] 0.8552632

# the accuracy of my model is 84.3%

# Assignment: 10.2 Exercises - Binary Classifier Dataset
# Name: Rajeev, Rahul
# Date: 2023-02-13

# additional libraries
library(tidyr)
```

```r
# loading data
classifier_data <- read.csv('data/binary-classifier-data.csv')
head(classifier_data)
```

```
##   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
```

```r
# general linear model for binary logistic
binary_log2 <- glm(label ~ x + y, family = 'binomial', data=classifier_data)
summary(binary_log2)
```

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

```r
# splitting data
split2 <- sample.split(classifier_data, SplitRatio = 0.65)
train2 <- subset(classifier_data, split2 == TRUE)
test2 <- subset(classifier_data, split2 == FALSE)
my_model2 <- glm(label ~ x + y, family = 'binomial', data=train2)
summary(my_model2)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = train2)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3626  -1.1678  -0.9679   1.1643   1.3961
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.408173   0.202259   2.018   0.0436 *
## x           -0.002267   0.003161  -0.717   0.4733
## y           -0.007832   0.003245  -2.413   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 691.52  on 498  degrees of freedom
## Residual deviance: 684.09  on 496  degrees of freedom
## AIC: 690.09
##
## Number of Fisher Scoring iterations: 4

# running the test data through the model
res2 <- predict(my_model2, test2, type = 'response')
res2 <- predict(my_model2, train2, type='response')

# confusion matrix
confmatrix2 <- table(Actual_Value = train2$label, Predicted_Value = res2 >
0.5)
confmatrix2

##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   144  111
##            1   102  142

# accuracy
(confmatrix2[[1,1]] + confmatrix2[[2,2]]) / sum(confmatrix2)

## [1] 0.5731463

# the accuracy of the logistic regression classifier is 54.3% which is very
low.
```