Week 10_1

Janine Par

2022-05-20

```
# a.For this problem, you will be working with the thoracic surgery data set from the University of Cal

## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/janin/OneDrive/Documents/R_repo/dsc520/")

## Load the `data/ThoraricSugery` to
thoraricsurgery_df <- read.csv("data/ThoraricSurgery.arff", header=FALSE, comment.char = "@")

names(thoraricsurgery_df) <- c("DGN","PRE4","PRE5","PRE6","PRE7","PRE8","PRE9","PRE10","PRE11","PRE14",

str(thoraricsurgery_df)
```

```
## 'data.frame':    470 obs. of  17 variables:
##  $ DGN  : chr  "DGN2" "DGN3" "DGN3" "DGN3" ...
##  $ PRE4 : num  2.88 3.4 2.76 3.68 2.44 2.48 4.36 3.19 3.16 2.32 ...
##  $ PRE5 : num  2.16 1.88 2.08 3.04 0.96 1.88 3.28 2.5 2.64 2.16 ...
##  $ PRE6 : chr  "PRZ1" "PRZ0" "PRZ1" "PRZ0" ...
##  $ PRE7 : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PRE8 : logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
##  $ PRE9 : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PRE10: logi  TRUE FALSE TRUE FALSE TRUE TRUE ...
##  $ PRE11: logi  TRUE FALSE FALSE FALSE TRUE FALSE ...
##  $ PRE14: chr  "OC14" "OC12" "OC11" "OC11" ...
##  $ PRE17: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PRE19: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PRE25: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ PRE30: logi  TRUE TRUE TRUE FALSE TRUE FALSE ...
##  $ PRE32: logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ AGE  : int  60 51 59 54 73 51 59 66 68 54 ...
##  $ Risk1: logi  FALSE FALSE FALSE FALSE TRUE FALSE ...
```

```
head (thoraricsurgery_df)
```

```
##    DGN PRE4 PRE5 PRE6  PRE7  PRE8  PRE9 PRE10 PRE11 PRE14 PRE17 PRE19 PRE25
## 1 DGN2 2.88 2.16 PRZ1 FALSE FALSE FALSE  TRUE  TRUE  OC14 FALSE FALSE FALSE
## 2 DGN3 3.40 1.88 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC12 FALSE FALSE FALSE
## 3 DGN3 2.76 2.08 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
## 4 DGN3 3.68 3.04 PRZ0 FALSE FALSE FALSE FALSE FALSE  OC11 FALSE FALSE FALSE
## 5 DGN3 2.44 0.96 PRZ2 FALSE  TRUE FALSE  TRUE  TRUE  OC11 FALSE FALSE FALSE
## 6 DGN3 2.48 1.88 PRZ1 FALSE FALSE FALSE  TRUE FALSE  OC11 FALSE FALSE FALSE
##   PRE30 PRE32 AGE Risk1
## 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
```

```
#b.Assignment Instructions:

#i. Fit a binary logistic regression model to the data set that predicts whether or not the patient sur
#Use the glm() function to perform the logistic regression. See Generalized Linear Models for an exampl


thoraricsurvice.model <- glm(Risk1~DGN+PRE4+PRE5+PRE6+PRE7+PRE8+PRE9+PRE10+PRE11+PRE14+PRE17+PRE19+PRE25

summary (thoraricsurvice.model)
```

```
##
## Call:
## glm(formula = Risk1 ~ DGN + PRE4 + PRE5 + PRE6 + PRE7 + PRE8 +
##     PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 +
##     PRE32 + AGE, family = binomial(), data = thoraricsurgery_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
## DGNDGN2      1.474e+01  2.400e+03   0.006  0.99510
## DGNDGN3      1.418e+01  2.400e+03   0.006  0.99528
## DGNDGN4      1.461e+01  2.400e+03   0.006  0.99514
## DGNDGN5      1.638e+01  2.400e+03   0.007  0.99455
## DGNDGN6      4.089e-01  2.673e+03   0.000  0.99988
## DGNDGN8      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 .
## PRE6PRZ1    -4.427e-01  5.199e-01  -0.852  0.39448
## PRE6PRZ2    -2.937e-01  7.907e-01  -0.371  0.71030
## PRE7TRUE     7.153e-01  5.556e-01   1.288  0.19788
## PRE8TRUE     1.743e-01  3.892e-01   0.448  0.65419
## PRE9TRUE     1.368e+00  4.868e-01   2.811  0.00494 **
## PRE10TRUE    5.770e-01  4.826e-01   1.196  0.23185
## PRE11TRUE    5.162e-01  3.965e-01   1.302  0.19295
## PRE14OC12    4.394e-01  3.301e-01   1.331  0.18318
## PRE14OC13    1.179e+00  6.165e-01   1.913  0.05580 .
## PRE14OC14    1.653e+00  6.094e-01   2.713  0.00668 **
## PRE17TRUE    9.266e-01  4.445e-01   2.085  0.03709 *
## PRE19TRUE   -1.466e+01  1.654e+03  -0.009  0.99293
## PRE25TRUE   -9.789e-02  1.003e+00  -0.098  0.92227
## PRE30TRUE    1.084e+00  4.990e-01   2.172  0.02984 *
## PRE32TRUE   -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
```

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

According to the summary, these variables have P value < .05 indicating that they are statistically significant

1. PRE9TRUE
2. PRE14OC14
3. PRE17TRUE
4. PRE30TRUE

```
#iii To compute the accuracy of your model, use the dataset to predict the outcome variable. #The perce

#Split data
tssplit <- sample.split(thoraricsurgery_df, SplitRatio = 0.8)

tssplit_train <- subset(thoraricsurgery_df,tssplit='True')
tssplit_train

tssplit_test <- subset(thoraricsurgery_df,tssplit='False')
tssplit_test

#Predict
res.train <- predict(thoraricsurvice.model,tssplit_train,type ="response")
res.train

res.test <- predict(thoraricsurvice.model,tssplit_test,type ="response")
res.test
```

```
confmatrix <- table(Actual_value=tssplit_train$Risk1, Predicted_Value= res.train > 0.5)

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

```
## [1] 0.8361702
```

Accuracy of the model is 83.6%