# Week 10 - Exercise 10.2

Ganesh Kale

May 22, 2021

## *Exercise 10.2.1 - Thoracic Surgery Analysis*

### Load the required packages

```
library(readxl)
library(dplyr)
library(QuantPsyc)
library(car)
library(foreign)
```

#### Load the file to dataframe

First 5 rows of the data frame:

```
##    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
##   PRE32 AGE Risk1Yr
## 1     F  60       F
## 2     F  51       F
## 3     F  59       F
## 4     F  54       F
## 5     F  73       T
```

***b.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. Include a summary using the summary() function in your results.***

Changed the baseline of all the binary predictors (T,F) using relevel() function because here we need to predict the survival that means Risky1Yr value as F.

```
thor_formula <- 'Risk1Yr ~ DGN + PRE4 + PRE5 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE
thor$PRE7 <- relevel(thor$PRE7, "T")
thor$PRE8 <- relevel(thor$PRE8, "T")
```

```r
thor$PRE9 <- relevel(thor$PRE9, "T")
thor$PRE10 <- relevel(thor$PRE10, "T")
thor$PRE11 <- relevel(thor$PRE11, "T")
thor$PRE17 <- relevel(thor$PRE17, "T")
thor$PRE19 <- relevel(thor$PRE19, "T")
thor$PRE25 <- relevel(thor$PRE25, "T")
thor$PRE30 <- relevel(thor$PRE30, "T")
thor$PRE32 <- relevel(thor$PRE32, "T")
thor$Risk1Yr <- relevel(thor$Risk1Yr, "T")

patientModel.1 <-  glm(thor_formula, data = thor, family = binomial())
summary(patientModel.1)
```

```
##
## Call:
## glm(formula = thor_formula, family = binomial(), data = thor)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4682   0.2716   0.4275   0.5483   1.4319
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.060e+01  3.354e+03   0.012  0.99034
## DGNDGN2     -1.461e+01  2.400e+03  -0.006  0.99514
## DGNDGN3     -1.407e+01  2.400e+03  -0.006  0.99532
## DGNDGN4     -1.447e+01  2.400e+03  -0.006  0.99519
## DGNDGN5     -1.628e+01  2.400e+03  -0.007  0.99459
## DGNDGN6     -2.927e-01  2.674e+03   0.000  0.99991
## DGNDGN8     -1.801e+01  2.400e+03  -0.008  0.99401
## PRE4         2.116e-01  1.832e-01   1.155  0.24805
## PRE5         2.679e-02  1.704e-02   1.572  0.11590
## PRE7F        5.911e-01  5.288e-01   1.118  0.26362
## PRE8F        1.859e-01  3.857e-01   0.482  0.62985
## PRE9F        1.342e+00  4.795e-01   2.799  0.00512 **
## PRE10F       3.141e-01  3.587e-01   0.876  0.38123
## PRE11F       4.855e-01  3.538e-01   1.372  0.16998
## PRE14OC12   -4.422e-01  3.296e-01  -1.342  0.17972
## PRE14OC13   -1.169e+00  6.173e-01  -1.894  0.05824 .
## PRE14OC14   -1.692e+00  6.015e-01  -2.812  0.00492 **
## PRE17F       8.978e-01  4.398e-01   2.042  0.04119 *
## PRE19F      -1.468e+01  1.656e+03  -0.009  0.99293
## PRE25F      -1.758e-01  9.999e-01  -0.176  0.86042
## PRE30F       1.067e+00  5.015e-01   2.127  0.03342 *
## PRE32F      -1.390e+01  1.658e+03  -0.008  0.99331
## AGE          7.641e-03  1.770e-02   0.432  0.66602
## ---
## 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: 342.00  on 447  degrees of freedom
```

```
## AIC: 388
##
## Number of Fisher Scoring iterations: 15
```

***b.ii] According to the summary, which variables had the greatest effect on the survival rate?***

```
##  (Intercept)      DGNDGN2      DGNDGN3      DGNDGN4      DGNDGN5      DGNDGN6
## 4.292641e+17 4.508693e-07 7.759461e-07 5.171320e-07 8.546863e-08 7.462814e-01
##      DGNDGN8         PRE4         PRE5        PRE7F        PRE8F        PRE9F
## 1.500581e-08 1.235629e+00 1.027156e+00 1.805979e+00 1.204264e+00 3.827055e+00
##       PRE10F       PRE11F    PRE14OC12    PRE14OC13    PRE14OC14       PRE17F
## 1.368961e+00 1.625042e+00 6.426425e-01 3.106154e-01 1.842083e-01 2.454314e+00
##       PRE19F       PRE25F       PRE30F       PRE32F          AGE
## 4.207677e-07 8.387722e-01 2.905861e+00 9.145265e-07 1.007670e+00
```

Based on the Summary result and odds ratio result above, we can say -

1. PRE9F - The z-value = 2.799 and p < 0.00512, odds ratio = 3.8270 which is greater than 1 that means as the predictor increases the odds of outcome occurring increases,False value of this predictor is significant.
2. PRE14OC14 - The z-value = -2.812 and p < 0.00492, odds ratio = 1.8420 which is greater than 1 that means as the predictor increases the odds of outcome occurring increases,False value of this predictor is significant.
3. PRE17F - the z-value = .042, p < 0.04119, odds ratio = 2.4543 which is greater than 1 that means as the predictor increases the odds of outcome occurring increases,False value of this predictor is significant.
4. PRE30F - the z-value = 2.127, p < 0.03342, odds ratio = 2.9058 which is greater than 1 that means as the predictor increases the odds of outcome occurring increases,False value of this predictor is significant.

***b.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?***

Based on the result, the percentage of correct predictions that means accuracy of model is - 0.8361702 or 83.6170%

Head of the final data frame with residuals and probabilities :

```
##    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
##   PRE32 AGE Risk1Yr pred.prob  std.resid stud.resid dfbeta.(Intercept)
## 1     F  60       F 0.4201373  1.3865719  1.3726405       1.078508e-01
## 2     F  51       F 0.9091711  0.4394349  0.4379941      -2.318582e-02
## 3     F  59       F 0.9139418  0.4262260  0.4252786       7.557292e-03
## 4     F  54       F 0.9806533  0.1983491  0.1980121      -5.112898e-03
## 5     F  73       T 0.8456124 -1.9653791 -1.9802712       2.884245e-02
```

```
##   dfbeta.DGNDGN2 dfbeta.DGNDGN3 dfbeta.DGNDGN4 dfbeta.DGNDGN5 dfbeta.DGNDGN6
## 1   4.362046e-02  -2.628836e-02  -1.001738e-02  -1.016901e-02   1.085944e-02
## 2   1.254537e-02   1.729310e-02   1.370523e-02   1.416859e-02   1.672384e-02
## 3  -3.389407e-03  -1.696571e-03  -4.518086e-03  -4.582569e-03  -7.126913e-03
## 4   9.409810e-04   1.842696e-03   1.103289e-03  -2.716791e-04   1.166918e-04
## 5   4.355516e-02   1.498183e-02   3.790779e-02   4.412107e-02   3.721550e-02
##   dfbeta.DGNDGN8   dfbeta.PRE4   dfbeta.PRE5  dfbeta.PRE7F  dfbeta.PRE8F
## 1   1.499242e-02 -1.282872e-02 -7.368483e-06 -2.122652e-02  3.388644e-02
## 2   4.583164e-03 -1.259495e-05 -4.157889e-05  4.415324e-03  8.634625e-04
## 3  -2.961018e-03 -1.826592e-03  1.198116e-05  2.208966e-04  2.591536e-03
## 4  -7.589477e-03  3.981492e-04  4.686899e-06  1.330670e-03  2.895880e-04
## 5   1.375606e-03  4.297488e-03  1.790986e-04 -3.057170e-02  8.606257e-02
##    dfbeta.PRE9F dfbeta.PRE10F dfbeta.PRE11F dfbeta.PRE14OC12 dfbeta.PRE14OC13
## 1 -1.793825e-02 -2.227940e-03 -6.517277e-02     7.839910e-03    -9.784288e-03
## 2  2.263211e-03  1.226363e-02 -5.521354e-04     5.835865e-03     3.439537e-03
## 3  1.790707e-03 -4.819423e-03  5.264588e-03    -8.864733e-03    -9.685223e-03
## 4  1.186119e-03  1.655397e-03  2.969029e-04    -1.856475e-03    -1.664841e-03
## 5 -1.454122e-02  3.539321e-03  4.216545e-02     4.295952e-02     2.785683e-02
##   dfbeta.PRE14OC14 dfbeta.PRE17F dfbeta.PRE19F dfbeta.PRE25F dfbeta.PRE30F
## 1     1.729563e-01  2.639995e-02  3.594423e-02  5.837049e-03 -5.711598e-03
## 2     2.117919e-03  1.765056e-03  5.004553e-03  9.574252e-04 -4.606584e-03
## 3    -9.613378e-03  2.766326e-03  3.218691e-03  3.039974e-04 -6.956682e-04
## 4    -1.757034e-03  9.985793e-04  1.889210e-04 -1.130179e-03  5.964002e-03
## 5     3.865440e-02 -2.160205e-02 -3.103414e-02 -4.097811e-02  4.483396e-03
##   dfbeta.PRE32F    dfbeta.AGE       dffit    leverage model_prob model_pred
## 1 -1.164916e-02 -8.189350e-04  0.52312425 0.097907039  0.4201373          0
## 2  1.233428e-02 -3.516221e-04  0.05931452 0.013770565  0.9091711          1
## 3  3.935924e-03 -1.138437e-04  0.04720497 0.009312954  0.9139418          1
## 4  2.895975e-03 -3.659178e-05  0.01882327 0.006857686  0.9806533          1
## 5  1.062652e-03 -1.334463e-03 -0.41471911 0.032655006  0.8456124          1
##   Risk1Yr_int
## 1           1
## 2           1
## 3           1
## 4           1
## 5           0

## [1] 0.8361702
```

## *Exercise 10.2.2 - Binary-Classifier_Data Analysis*

### Load the dataset

Head of the Data frame:

```
##   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
```

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

Summary of the logistic regression model:

```
binclaas.model <- glm(label ~ x + y, data = bin.class, family = binomial())

summary(binclaas.model)
```

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

***2.b.i] What is the accuracy of the logistic regression classifier?***

Accuracy of the logistic regression classifier is :0.512016

```
##   label        x        y model.prob model_pred accurate
## 1     0 70.88469 83.17702  0.3967211          0        1
## 2     0 74.97176 87.92922  0.3852176          0        1
## 3     0 73.78333 92.20325  0.3779152          0        1
## 4     0 66.40747 81.10617  0.4034378          0        1
## 5     0 69.07399 84.53739  0.3952460          0        1
## 6     0 72.23616 86.38403  0.3898045          0        1
```

```
## [1] 0.512016
```