Week 10 Assignment

Ruben Brionez Jr

February 18th 2024

Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset

```
# Loading required libraries
library(dplyr)
library(ggplot2)
```

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.

```
# Loading data
df <- read.csv('week-10-data.csv')

# Reviewing data
head(df)
```

```
##   Column1 Column2 Column3 Column4 Column5 Column6 Column7 Column8 Column9
## 1              NA      NA              NA      NA      NA      NA      NA
## 2    DGN2    2.88    2.16    PRZ1   FALSE   FALSE   FALSE    TRUE    TRUE
## 3    DGN3    3.40    1.88    PRZ0   FALSE   FALSE   FALSE   FALSE   FALSE
## 4    DGN3    2.76    2.08    PRZ1   FALSE   FALSE   FALSE    TRUE   FALSE
## 5    DGN3    3.68    3.04    PRZ0   FALSE   FALSE   FALSE   FALSE   FALSE
## 6    DGN3    2.44    0.96    PRZ2   FALSE    TRUE   FALSE    TRUE    TRUE
##   Column10 Column11 Column12 Column13 Column14 Column15 Column16 Column17
## 1               NA       NA       NA       NA       NA       NA       NA
## 2     OC14    FALSE    FALSE    FALSE     TRUE    FALSE       60    FALSE
## 3     OC12    FALSE    FALSE    FALSE     TRUE    FALSE       51    FALSE
## 4     OC11    FALSE    FALSE    FALSE     TRUE    FALSE       59    FALSE
## 5     OC11    FALSE    FALSE    FALSE    FALSE    FALSE       54    FALSE
## 6     OC11    FALSE    FALSE    FALSE     TRUE    FALSE       73     TRUE
```

```
# Renaming Columns
df <- df %>% rename('DGN'=Column1, 'PRE4'=Column2,'PRE5'=Column3,
                    'PRE6'=Column4,'PRE7'=Column5,'PRE8'=Column6,
                    'PRE9'=Column7,'PRE10'=Column8,'PRE11'=Column9,
                    'PRE14'=Column10,'PRE17'=Column11,'PRE19'=Column12,
                    'PRE25'=Column13,'PRE30'=Column14,
```

```
                    'PRE32'=Column15,'Age'=Column16,'Risk1Yr'=Column17)

# Checking for NA Values
sum(is.na(df))
```

```
## [1] 14
```

```
# Removing NA Values
dataset <- df[complete.cases(df), ]

# Using glm() function to fit a model
risk_model <- glm(Risk1Yr ~ PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 +
                   PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32
                  + Age, data = df, family = binomial)
```

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

```
summary(risk_model)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE4 + PRE5 + PRE6 + PRE7 + PRE8 + PRE9 +
##     PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30 + PRE32 +
##     Age, family = binomial, data = df)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.262e+00  1.393e+00  -1.624  0.10430
## PRE4        -1.581e-01  1.758e-01  -0.899  0.36850
## PRE5        -2.246e-02  1.697e-02  -1.324  0.18554
## PRE6PRZ1    -4.713e-01  5.073e-01  -0.929  0.35294
## PRE6PRZ2    -2.844e-01  7.596e-01  -0.374  0.70809
## PRE7TRUE     6.323e-01  5.339e-01   1.184  0.23631
## PRE8TRUE     2.593e-01  3.723e-01   0.696  0.48613
## PRE9TRUE     1.185e+00  4.771e-01   2.483  0.01301 *
## PRE10TRUE    4.836e-01  4.726e-01   1.023  0.30628
## PRE11TRUE    5.423e-01  3.866e-01   1.403  0.16063
## PRE14OC12    4.387e-01  3.195e-01   1.373  0.16974
## PRE14OC13    1.281e+00  5.904e-01   2.170  0.03000 *
## PRE14OC14    1.674e+00  5.804e-01   2.884  0.00392 **
## PRE17TRUE    9.511e-01  4.307e-01   2.208  0.02723 *
## PRE19TRUE   -1.380e+01  1.003e+03  -0.014  0.98902
## PRE25TRUE    3.013e-01  8.910e-01   0.338  0.73526
## PRE30TRUE    7.976e-01  4.486e-01   1.778  0.07541 .
## PRE32TRUE   -1.325e+01  1.002e+03  -0.013  0.98945
## Age         -6.039e-03  1.722e-02  -0.351  0.72588
## ---
## 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: 359.28  on 451  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 397.28
##
## Number of Fisher Scoring iterations: 14
```

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

The closer to zero the p-value is the more statistical significance

The summary provides an easy to see significance code for variables

In this model, the variable PRE9, which is described as "Dyspnoea before surgery" is the most statistically significant variable

Following the variable PRE9, PRE14 which is described as "T in clinical TNM - size of the original tumor, from OC11 (smallest) to OC14 (largest)"

The variable PRE30 and PRE17 are the next most significant variables, PRE30 being described as "Smoking" the variable PRE17 being described as Type 2 DM - diabetes mellitus

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?

```
# Running data through the model
risk_predict <- predict(risk_model, type = 'response')

# Reviewing
head(risk_predict)
```

```
##         2         3         4         5         6         7
## 0.47443727 0.12854758 0.09180030 0.03774021 0.21212385 0.04777117
```

```
# Validate the model - Confusion Matrix
confmatrix <- table(Actual_Value=dataset$Risk1Yr,
                    Predicted_Value=risk_predict >0.5)

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

```
## [1] 0.8425532
```

The accuracy appears to be about 84%

Part 2 of Assignment

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

```
#Load data
binary_df <- read.csv("binary-classifier-data.csv")

head(binary_df)
```

```
##   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
```

```
# Creating the model
binary_model <- glm(label ~ x + y, data = binary_df, family = "binomial")

# Summary of model
summary(binary_model)
```

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

```
# Run Data through the model
binary_predict <- predict(binary_model, binary_df, type = 'response')

# Validate the model - Confusion Matrix
confmatrix <- table(Actual_Value=binary_df$label,
                    Predicted_Value=binary_predict >0.5)
head(confmatrix)
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0   429  338
##            1   286  445
```

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

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

The accuracy of the model appears to be about 58%