McKibben DSC520 Ex. 9.2

Makayla McKibben

2024-08-03

```
knitr::opts_chunk$set(echo = TRUE)

# Install and call relevant packages
#install.packages("foreign")
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 4.4.1
```

```
#install.packages("tidyverse")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.1

## Warning: package 'ggplot2' was built under R version 4.4.1

## Warning: package 'tibble' was built under R version 4.4.1

## Warning: package 'tidyr' was built under R version 4.4.1

## Warning: package 'readr' was built under R version 4.4.1

## Warning: package 'purrr' was built under R version 4.4.1

## Warning: package 'dplyr' was built under R version 4.4.1

## Warning: package 'forcats' was built under R version 4.4.1

## Warning: package 'lubridate' was built under R version 4.4.1

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
#install.packages("mlogit")
library(mlogit)
```

```
## Warning: package 'mlogit' was built under R version 4.4.1
```

```
## Loading required package: dfidx
```

```
## Warning: package 'dfidx' was built under R version 4.4.1
```

```
##
## Attaching package: 'dfidx'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```r
# Ex. 9.2 problem 1
# Import data from file
surgery <- read.arff("ThoracicSurgery.arff")

# Get a sense of the data's structure
head(surgery)
```

```
##    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
```

```r
# Rename columns in dataset
surgery <- surgery %>%
  rename(Diagnosis = DGN, FVC = PRE4, FEV1 = PRE5, Zubrod = PRE6,
         Pain_before = PRE7, Haemoptysis_before = PRE8,
         Dyspnoea_before = PRE9, Cough_before = PRE10,
         Weakness_before = PRE11, Size_of_tumour = PRE14, T2_Diabetes = PRE17,
         MI = PRE19, PAD = PRE25, Smoker = PRE30, Asthma = PRE32)

# Check for rename
head(surgery)
```

```
##   Diagnosis  FVC FEV1 Zubrod Pain_before Haemoptysis_before Dyspnoea_before
## 1      DGN2 2.88 2.16   PRZ1           F                  F               F
```

```
## 2      DGN3 3.40 1.88   PRZ0           F                   F               F
## 3      DGN3 2.76 2.08   PRZ1           F                   F               F
## 4      DGN3 3.68 3.04   PRZ0           F                   F               F
## 5      DGN3 2.44 0.96   PRZ2           F                   T               F
## 6      DGN3 2.48 1.88   PRZ1           F                   F               F
##   Cough_before Weakness_before Size_of_tumour T2_Diabetes MI PAD Smoker Asthma
## 1            T               T           OC14           F  F   F      T      F
## 2            F               F           OC12           F  F   F      T      F
## 3            T               F           OC11           F  F   F      T      F
## 4            F               F           OC11           F  F   F      F      F
## 5            T               T           OC11           F  F   F      T      F
## 6            T               F           OC11           F  F   F      F      F
##   AGE Risk1Yr
## 1  60       F
## 2  51       F
## 3  59       F
## 4  54       F
## 5  73       T
## 6  51       F
```

```r
# Remove any missing entries/rows
surgery <- surgery[complete.cases(surgery),]

# Create model
surg_model <- glm(Risk1Yr ~ AGE + Diagnosis + FVC +
                    FEV1 + Zubrod + Pain_before + Haemoptysis_before +
                    Dyspnoea_before  + Cough_before + Weakness_before +
                    Size_of_tumour + T2_Diabetes + MI + PAD + Smoker +
                    Asthma, data = surgery, family = binomial)

# Check probabilities
head(surgery, 10)
```

```
##    Diagnosis  FVC FEV1 Zubrod Pain_before Haemoptysis_before Dyspnoea_before
## 1       DGN2 2.88 2.16   PRZ1           F                   F               F
## 2       DGN3 3.40 1.88   PRZ0           F                   F               F
## 3       DGN3 2.76 2.08   PRZ1           F                   F               F
## 4       DGN3 3.68 3.04   PRZ0           F                   F               F
## 5       DGN3 2.44 0.96   PRZ2           F                   T               F
## 6       DGN3 2.48 1.88   PRZ1           F                   F               F
## 7       DGN3 4.36 3.28   PRZ1           F                   F               F
## 8       DGN2 3.19 2.50   PRZ1           F                   F               F
## 9       DGN3 3.16 2.64   PRZ2           F                   F               F
## 10      DGN3 2.32 2.16   PRZ1           F                   F               F
##    Cough_before Weakness_before Size_of_tumour T2_Diabetes MI PAD Smoker Asthma
## 1             T               T           OC14           F  F   F      T      F
## 2             F               F           OC12           F  F   F      T      F
## 3             T               F           OC11           F  F   F      T      F
## 4             F               F           OC11           F  F   F      F      F
## 5             T               T           OC11           F  F   F      T      F
## 6             T               F           OC11           F  F   F      F      F
## 7             T               F           OC12           T  F   F      T      F
## 8             T               F           OC11           F  F   T      T      F
## 9             T               T           OC11           F  F   F      T      F
```

```
## 10            T               F           OC11           F  F   F      T      F
##    AGE Risk1Yr
## 1   60       F
## 2   51       F
## 3   59       F
## 4   54       F
## 5   73       T
## 6   51       F
## 7   59       T
## 8   66       T
## 9   68       F
## 10  54       F
```

```
# Get summary of model
summary(surg_model)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ AGE + Diagnosis + FVC + FEV1 + Zubrod +
##     Pain_before + Haemoptysis_before + Dyspnoea_before + Cough_before +
##     Weakness_before + Size_of_tumour + T2_Diabetes + MI + PAD +
##     Smoker + Asthma, family = binomial, data = surgery)
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.655e+01  2.400e+03  -0.007  0.99450
## AGE                 -9.506e-03  1.810e-02  -0.525  0.59944
## DiagnosisDGN2        1.474e+01  2.400e+03   0.006  0.99510
## DiagnosisDGN3        1.418e+01  2.400e+03   0.006  0.99528
## DiagnosisDGN4        1.461e+01  2.400e+03   0.006  0.99514
## DiagnosisDGN5        1.638e+01  2.400e+03   0.007  0.99455
## DiagnosisDGN6        4.089e-01  2.673e+03   0.000  0.99988
## DiagnosisDGN8        1.803e+01  2.400e+03   0.008  0.99400
## FVC                 -2.272e-01  1.849e-01  -1.229  0.21909
## FEV1                -3.030e-02  1.786e-02  -1.697  0.08971 .
## ZubrodPRZ1          -4.427e-01  5.199e-01  -0.852  0.39448
## ZubrodPRZ2          -2.937e-01  7.907e-01  -0.371  0.71030
## Pain_beforeT         7.153e-01  5.556e-01   1.288  0.19788
## Haemoptysis_beforeT  1.743e-01  3.892e-01   0.448  0.65419
## Dyspnoea_beforeT     1.368e+00  4.868e-01   2.811  0.00494 **
## Cough_beforeT        5.770e-01  4.826e-01   1.196  0.23185
## Weakness_beforeT     5.162e-01  3.965e-01   1.302  0.19295
## Size_of_tumourOC12   4.394e-01  3.301e-01   1.331  0.18318
## Size_of_tumourOC13   1.179e+00  6.165e-01   1.913  0.05580 .
## Size_of_tumourOC14   1.653e+00  6.094e-01   2.713  0.00668 **
## T2_DiabetesT         9.266e-01  4.445e-01   2.085  0.03709 *
## MIT                 -1.466e+01  1.654e+03  -0.009  0.99293
## PADT                -9.789e-02  1.003e+00  -0.098  0.92227
## SmokerT              1.084e+00  4.990e-01   2.172  0.02984 *
## AsthmaT             -1.398e+01  1.645e+03  -0.008  0.99322
## ---
## 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 if we look at Pr(>|z|) values we can see
# that Dyspnoea_before, Size of tumor, having type 2 diabetes,
# and smoking had a significant effect on 1yr survival.

# Predicting the results from the model
model_acc <- predict.glm(surg_model, newdata = surgery)
head(model_acc)
```

```
##          1          2          3          4          5          6
##  0.2817109 -2.1621770 -2.4039672 -3.8128355 -1.5908565 -3.3422297
```

```
# Create the base for finding percentage of correct predictions
accuracy <- cbind(surgery$Risk1Yr, model_acc)
colnames(accuracy) <- c("1 Year Survival", "Model")
accuracy <- data.frame(accuracy)
head(accuracy)
```

```
##   X1.Year.Survival      Model
## 1                1  0.2817109
## 2                1 -2.1621770
## 3                1 -2.4039672
## 4                1 -3.8128355
## 5                2 -1.5908565
## 6                1 -3.3422297
```

```
results_model <- ifelse(accuracy$Model >= 0.5, "Positive", "Negative")
results_data <- ifelse(surgery$Risk1Yr == 2, "Positive", "Negative")
head(results_data)
```

```
## [1] "Negative" "Negative" "Negative" "Negative" "Negative" "Negative"
```

```
results_comb <- cbind(results_data, results_model)
head(results_comb)
```

```
##      results_data results_model
## [1,] "Negative"   "Negative"
## [2,] "Negative"   "Negative"
## [3,] "Negative"   "Negative"
## [4,] "Negative"   "Negative"
## [5,] "Negative"   "Negative"
## [6,] "Negative"   "Negative"
```

```r
colnames(results_comb) <- c("Data", "Model")
results_comb <- data.frame(results_comb)
num_correct <- length(which(results_comb$Data == results_comb$Model))

percent_correct <- (num_correct/length(results_comb$Data))*100
percent_correct
```

```
## [1] 98.7234
```

```r
# Ex. 9.2 problem 2
# Import data from file binary classifier data
binary <- read.csv(file = 'binary-classifier-data.csv', header = TRUE,
                   sep =",", stringsAsFactors = FALSE)

# Check data structure
head(binary)
```

```
##   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
# Remove rows missing data
binary <- binary[complete.cases(binary),]

# Create model
binary_model <- glm(label ~ x + y, data = binary, family = binomial)

# Check model
summary(binary_model)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = binomial, data = binary)
##
## 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
# Predict results using model
pred_binary <- predict.glm(binary_model)

# Bind results to single dataframe
binary_acc <- cbind(binary$label, pred_binary)
colnames(binary_acc) <- c("Data", "Model")
binary_acc <- data.frame(binary_acc)

# Check results
head(binary_acc)
```

```
##   Data      Model
## 1    0 -0.4191462
## 2    0 -0.4674600
## 3    0 -0.4984068
## 4    0 -0.3911610
## 5    0 -0.4253135
## 6    0 -0.4481342
```

```r
# Make necessary transformations
results_model_bin <- ifelse(binary_acc$Model >= 0.5, "Positive", "Negative")
results_data_bin <- ifelse(binary_acc$Data == 1, "Positive", "Negative")
results_bin <- cbind(results_data_bin, results_model_bin)
colnames(results_bin) <- c("Data", "Model")
results_bin <- data.frame(results_bin)

# Find percent accuracy
num_correct_binary <- length(which(results_bin$Data == results_bin$Model))
percent_correct_binary <- (num_correct_binary/length(results_bin$Data))*100
percent_correct_binary
```

```
## [1] 51.2016
```