Logistic Regression

Oliver Imhans

9/25/2021

```
library(tidyverse)
library(dplyr)
library(corrgram)
library(DataExplorer)
library(ppcor)
library(caTools)
library(ggplot2)
library(corrplot)
library(data.table)
library(plotly)
library(modelr)
library(arm)
library(cowplot)
```

Logistic regression which is also known as the logit model, predicts the probability of an event occurring. It is used to model dichotomous outcome variables. Logistic regression implies that the possible outcomes are not numerical but rather categorical. Examples of such categories are: Yes / No or 1 / 0.

Logistic regression in R Programming is a classification algorithm used to find the probability of event success and event failure.

Logistic sample: model <- glm(Y ~ x, binomial(), data)

The *logit function* is shown below

$$logit(p) = log(\frac{p}{1-p})$$

To get values of x between 0 and 1, take the inverse.

$$logit^{-1}(a) = \frac{1}{1+e^{-a}}$$

$$Deviance = -2 * log - likelihood = -2LL$$

$$AIC = -2LL + 2k$$

Using AIC, we can compare multiple models, but as we increase the number of predictors the AIC k penalty will increase.

Below is an example of how to implement the Logistic Regression.

The dataset used for this project can be downloaded at this link: https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression?resource=download

The variables in the dataset are:

Male: Gender (1 = Male, 0 = Female) Age: Patient age Education: Education level (1 = some high school, 2 = high school/GED, 3 = some college, 4 = college) CurrentSmoker: 1 = patient is smoker CigsPerDay: Number of cigarettes patient smokes per day BPMeds: 1 = patient is on blood pressure medication PrevalentStroke: 1 = patient has previously had a stroke PrevalentHyp: 1 = patient has hypertension Diabetes: 1 = patient has diabetes Chol: total cholesterol (mg/dL) SysBP: systolic blood pressure (mmHg) DiaBP: diastolic blood pressure (mmHg) BMI: body mass index (weight / $height^2$) HeartRate: Heart rate (beats per minute) Glucose: blood glucose level (mg/dL) TenYearCHD: 1 = patient developed coronary heart disease within 10 years of exam

```
# Importing the dataset
data = read.csv('framingham.csv')
```

```
head(data)
```

```
##   male age education currentSmoker cigsPerDay BPMeds prevalentStroke
## 1    1  39         4             0          0      0               0
## 2    0  46         2             0          0      0               0
## 3    1  48         1             1         20      0               0
## 4    0  61         3             1         30      0               0
## 5    0  46         3             1         23      0               0
## 6    0  43         2             0          0      0               0
##   prevalentHyp diabetes totChol sysBP diaBP   BMI heartRate glucose TenYearCHD
## 1            0        0     195 106.0    70 26.97        80      77          0
## 2            0        0     250 121.0    81 28.73        95      76          0
## 3            0        0     245 127.5    80 25.34        75      70          0
## 4            1        0     225 150.0    95 28.58        65     103          1
## 5            0        0     285 130.0    84 23.10        85      85          0
## 6            1        0     228 180.0   110 30.30        77      99          0
```

```
dim(data)
```

```
## [1] 4238   16
```

```
summary(data)
```

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.003   Mean   :0.02963   Mean   :0.005899   Mean   :0.3105
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.00
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.00
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.00
##  Mean   :0.02572   Mean   :236.7   Mean   :132.4   Mean   : 82.89
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 89.88
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.50  
##                    NA's   :50                                       
##       BMI          heartRate         glucose         TenYearCHD   
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.000  
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.000  
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.000  
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.97   Mean   :0.152  
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.000  
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.000  
##  NA's   :19      NA's   :1        NA's   :388                     
```

```
#checking for missing values
sum(is.na(data))
```

```
## [1] 645
```

```
#removes cases with missing data
data <- data %>% drop_na
```

```
#checking again for missing values
sum(is.na(data))
```

```
## [1] 0
```

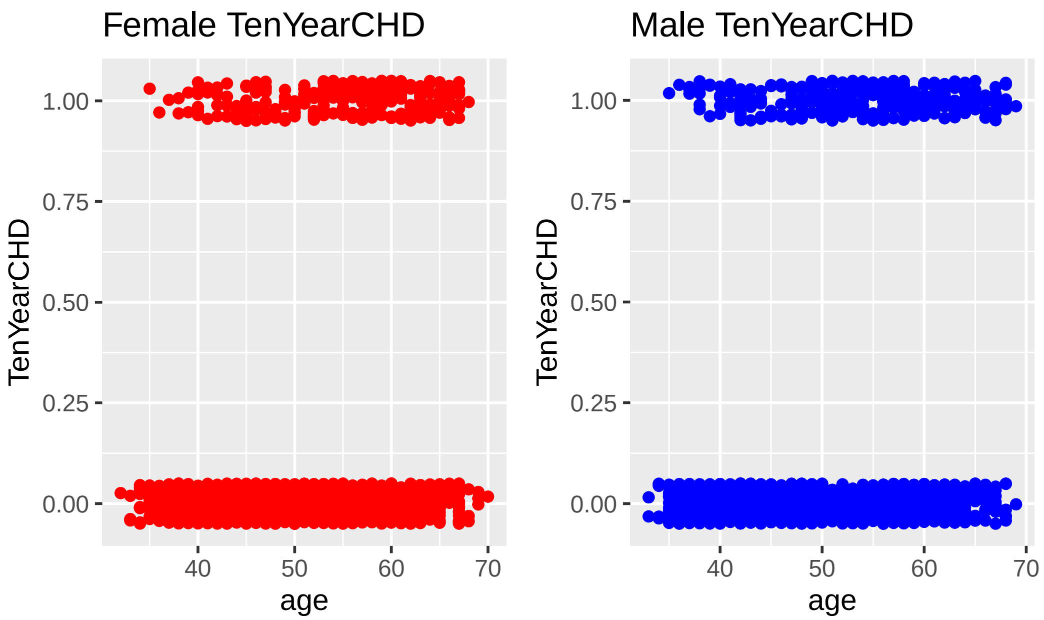
```
# Sub-setting the data
male <- filter(data, male == 1)
dim(male)
```

```
## [1] 1622   16
```

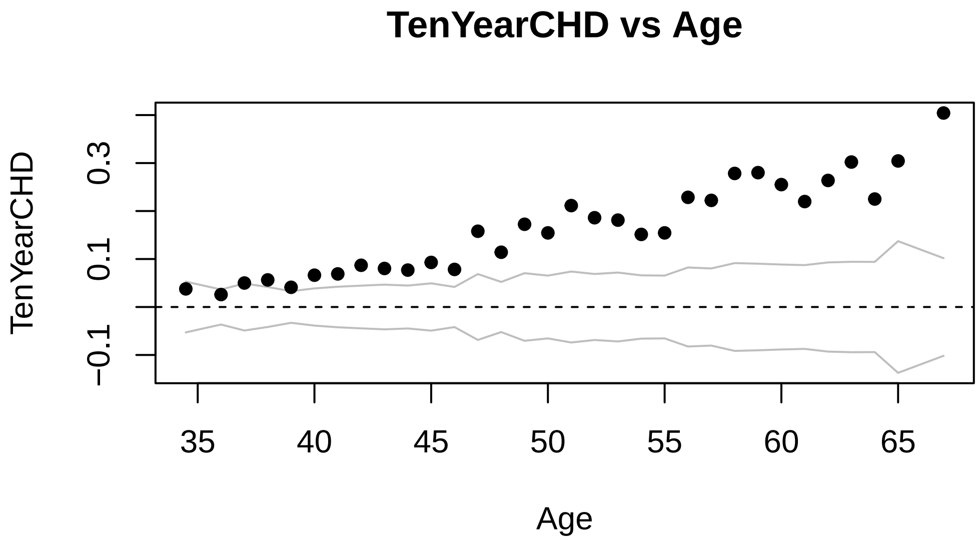
```
female <- filter(data, male == 0)
dim(female)
```

```
## [1] 2034   16
```

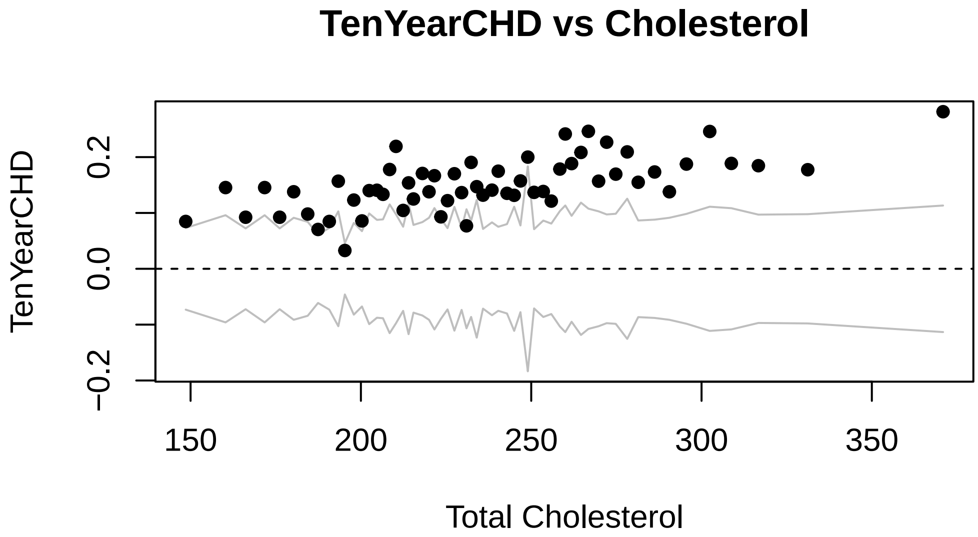
```
# Visualizing Data
fcolor = "red"
mcolor = "blue"

female.plot <- ggplot(female, aes(x = age, y = TenYearCHD)) +
  geom_jitter(width = 0, height = 0.05, color = fcolor) +
  labs(title = "Female TenYearCHD")

male.plot <- ggplot(male, aes(x = age, y = TenYearCHD)) +
  geom_jitter(width = 0, height = 0.05, color = mcolor) +
  labs(title = "Male TenYearCHD")

plot_grid(female.plot, male.plot)
```



```
# A quick plot of age against TenYearCHD
binnedplot(data$age, data$TenYearCHD,xlab="Age",ylab="TenYearCHD",main="TenYearCHD vs Age")
```



```
binnedplot(data$totChol, data$TenYearCHD, xlab="Total Cholesterol",ylab="TenYearCHD",main="TenYearCHD vs
```



Fitting the logistic regression model

```
# split the data into training and testing data:

set.seed(42)

split = sample.split(data, SplitRatio = 0.8)
train = subset(data, split == TRUE)
test = subset(data, split == FALSE)

row.names(train) <- NULL
row.names(test) <- NULL
```

```
# using the glm() command on the training data.
# summary to view the outcome
logit <- glm(TenYearCHD ~., family="binomial", data=train)
summary(logit)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0767  -0.5770  -0.4204  -0.2858   2.8426
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -7.944280   0.840272  -9.454  < 2e-16 ***
## male               0.600154   0.128972   4.653 3.27e-06 ***
## age                0.063844   0.007954   8.027 1.00e-15 ***
## education         -0.047638   0.059345  -0.803  0.42213
## currentSmoker      0.063430   0.184888   0.343  0.73154
## cigsPerDay         0.016502   0.007335   2.250  0.02445 *
## BPMeds             0.416300   0.276493   1.506  0.13216
## prevalentStroke  0.179263   0.566743   0.316  0.75177
## prevalentHyp     0.099443   0.165464   0.601  0.54784
## diabetes         0.145469   0.373150   0.390  0.69665
## totChol          0.002335   0.001320   1.768  0.07698 .
## sysBP            0.014477   0.004448   3.255  0.00113 **
## diaBP           -0.001067   0.007677  -0.139  0.88950
## BMI             -0.014027   0.015824  -0.886  0.37536
## heartRate       -0.005037   0.004997  -1.008  0.31340
## glucose          0.008371   0.002667   3.138  0.00170 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2239.3  on 2741  degrees of freedom
## Residual deviance: 1988.8  on 2726  degrees of freedom
## AIC: 2020.8
##
## Number of Fisher Scoring iterations: 5
```

Making Prediction based on the model

```
subject_1 = data.frame(male = 1, age = 50, education = 2, currentSmoker = 1,  cigsPerDay =  0, BPMeds =
predict(logit,subject_1,type='response')
```

```
##          1
## 0.06859464
```

According to the model, the probability of this subject(smoker with diabetes) developing heart disease within 10 years is 68.6%.

```
subject_2 = data.frame(male = 1, age = 50, education = 2, currentSmoker = 0,  cigsPerDay =  0, BPMeds =
predict(logit,subject_2,type='response')
```

```
##          1
## 0.05639219
```

According to the model, the probability of this subject(none smoker without diabetes) developing heart disease within 10 years is 56.39%.

It is important to note, that the probability of a subject developing heart disease within 10 years, varies for a smoker with diabetes and a none smoker without diabetes.

Other Predictions using classification is also possible. That discussion will be carried out in the python version of Logistic Regression.