Logistic Regression model for AHD

Huy N. Pham

4/11/2020

Introduction

The data table is the result from an cardiovascular (CV) study in Cleveland Clinic between May 1981 and September 1984. No patients had a history of CV disease. After providing the historical information, all patients performed a number of clinical tests. A part of features from these results was collected in the studying data table.

Data loading

The data table was loaded and assigned to a variable call `data`.

```
# Install the required library
list.of.packages <- c("dplyr", "qwraps2", "ggplot2", "gridExtra", "MASS")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
library(qwraps2)
options(qwraps2_markup = "markdown")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
```

```
##     select
```

```r
# Load the data table
data_link <- "https://github.com/pnhuy/datasets/raw/master/heart_uci/heart.csv"
heart <- read.csv(data_link)
```

Exploratory Data analysis

The data have 303 rows or patients, and 15 fields: X, Age, Sex, ChestPain, RestBP, Chol, Fbs, RestECG, MaxHR, ExAng, Oldpeak, Slope, Ca, Thal, AHD.

```r
# Remove the first column
data_col <- colnames(heart)
data_col <- data_col[data_col != "X"]
heart <- heart[data_col]

# Process the categorical variables
heart$Sex <- factor(heart$Sex, levels=c(0,1), labels=c('female', 'male'))
#heart$ChestPain <- factor(heart$ChestPain)
heart$Fbs <- factor(heart$Fbs)
heart$RestECG <- factor(heart$RestECG)
heart$ExAng <- factor(heart$ExAng)
heart$Slope <- factor(heart$Slope)
heart$Ca <- factor(heart$Ca)
heart$Thal <- factor(heart$Thal)
heart$AHD <- factor(heart$AHD)

factor_feature <- c("Sex", "Fbs", "RestECG", "ExAng", "Slope", "Ca", "Thal")
numberic_feature <- c("Age", "RestBP", "Chol", "MaxHR", "Oldpeak")
```

The basic statistics of data was below:

```r
summary_table(heart)
```

	heart (N = 303)
Age	
minimum	29
median (IQR)	56 (48.00, 61.00)
mean (sd)	54.44 ± 9.04
maximum	77
Sex	
female	97 (32)
male	206 (68)
ChestPain	
asymptomatic	144 (48)
nonanginal	86 (28)
nontypical	50 (17)
typical	23 (8)
RestBP	
minimum	94
median (IQR)	130 (120.00, 140.00)
mean (sd)	131.69 ± 17.60
maximum	200
Chol	
minimum	126

	heart (N = 303)
median (IQR)	241 (211.00, 275.00)
mean (sd)	246.69 ± 51.78
maximum	564
Fbs	
0	258 (85)
1	45 (15)
RestECG	
0	151 (50)
1	4 (1)
2	148 (49)
MaxHR	
minimum	71
median (IQR)	153 (133.50, 166.00)
mean (sd)	149.61 ± 22.88
maximum	202
ExAng	
0	204 (67)
1	99 (33)
Oldpeak	
minimum	0.00
median (IQR)	0.80 (0.00, 1.60)
mean (sd)	1.04 ± 1.16
maximum	6.20
Slope	
1	142 (47)
2	140 (46)
3	21 (7)
Ca	
0	176 (59)
1	65 (22)
2	38 (13)
3	20 (7)
Unknown	4/303 (1)
Thal	
fixed	18 (6)
normal	166 (55)
reversable	117 (39)
Unknown	2/303 (1)
AHD	
No	164 (54)
Yes	139 (46)

The data might contain missing value and they would be removed before building the model.

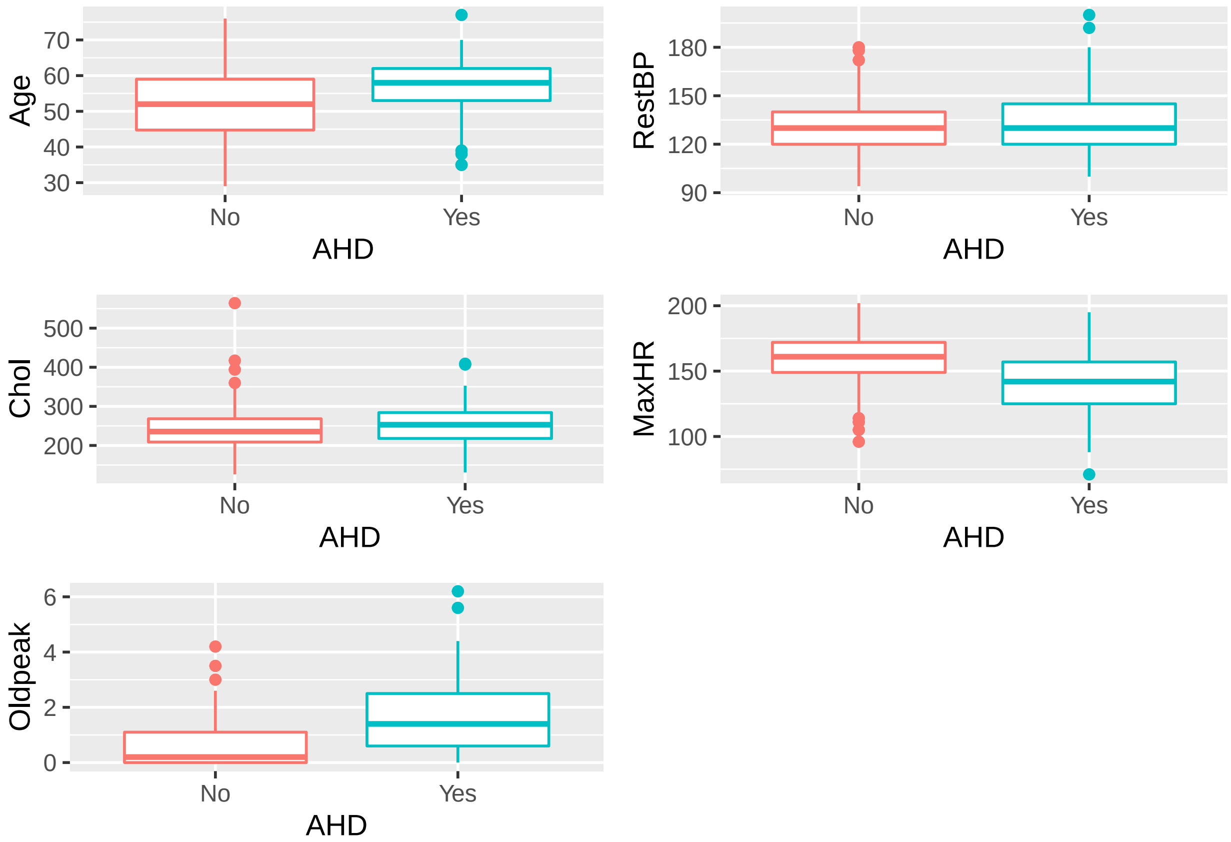
```
heart <- na.omit(heart)
```

After removing the NA, the table consisted of 297 records.

The distibution of target (`AHD`) which was grouped by numerical variables was illustrated by using boxplot.

```
p1 <- ggplot(heart, aes(x=AHD, y=Age, color=AHD)) + geom_boxplot() + theme(legend.position = "none")
p2 <- ggplot(heart, aes(x=AHD, y=RestBP, color=AHD)) + geom_boxplot() + theme(legend.position = "none")
p3 <- ggplot(heart, aes(x=AHD, y=Chol, color=AHD)) + geom_boxplot() + theme(legend.position = "none")
```

```r
p4 <- ggplot(heart, aes(x=AHD, y=MaxHR, color=AHD)) + geom_boxplot() + theme(legend.position = "none")
p5 <- ggplot(heart, aes(x=AHD, y=Oldpeak, color=AHD)) + geom_boxplot() + theme(legend.position = "none"
grid.arrange(p1, p2, p3, p4, p5)
```



The boxplot show the difference in distribution of `AHD` by `Age`, `MaxHR` and `Oldpeak`. These hypothesis would be tested by t-test.

Hypothesis testing

```r
t.test(Age ~ AHD, data=heart)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age by AHD
## t = -4.0636, df = 294.66, p-value = 6.204e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.108514 -2.122234
## sample estimates:
##  mean in group No mean in group Yes
##          52.64375          56.75912
```

```r
t.test(MaxHR ~ AHD, data=heart)
```

```
##
##  Welch Two Sample t-test
```

```
##
## data:  MaxHR by AHD
## t = 7.9286, df = 266.44, p-value = 6.108e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  14.63637 24.30715
## sample estimates:
##  mean in group No mean in group Yes
##          158.5813          139.1095
```

```
t.test(Oldpeak ~ AHD, data=heart)
```

```
##
##  Welch Two Sample t-test
##
## data:  Oldpeak by AHD
## t = -7.7558, df = 216, p-value = 3.429e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.241970 -0.738632
## sample estimates:
##  mean in group No mean in group Yes
##          0.598750          1.589051
```

The `t.test` show there was a significant difference in mean of `Age`, `MaxHR`, `Oldpeak` between group of patiend who had AHD or no AHD.

Logistic regression model

A logistic model would be built to predict the probability of AHD by the remaining variables. Stepwise algorithm was used to select the best model.

```
full.model <- glm(AHD ~ ., data=heart, family = "binomial")
step.model <- stepAIC(full.model, direction = "both",
                      trace = FALSE)
summary(step.model)
```

```
##
## Call:
## glm(formula = AHD ~ Sex + ChestPain + RestBP + MaxHR + ExAng +
##     Oldpeak + Slope + Ca + Thal, family = "binomial", data = heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0614  -0.4830  -0.1201   0.3257   2.9129
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -4.44644    2.30062  -1.933 0.053271 .
## Sexmale              1.55489    0.52222   2.977 0.002907 **
## ChestPainnonanginal -2.13460    0.51462  -4.148 3.36e-05 ***
## ChestPainnontypical -0.97816    0.56426  -1.734 0.083000 .
## ChestPaintypical    -2.44156    0.69208  -3.528 0.000419 ***
## RestBP               0.02493    0.01070   2.329 0.019861 *
## MaxHR               -0.01493    0.01071  -1.394 0.163450
## ExAng1               0.63056    0.43943   1.435 0.151300
## Oldpeak             0.44279    0.22944   1.930 0.053623 .
## Slope2              1.31410    0.47721   2.754 0.005893 **
## Slope3              0.56249    0.91149   0.617 0.537160
## Ca1                 2.16201    0.49319   4.384 1.17e-05 ***
## Ca2                 2.94081    0.72291   4.068 4.74e-05 ***
## Ca3                 2.00426    0.89728   2.234 0.025502 *
## Thalnormal          0.37503    0.78094   0.480 0.631064
## Thalreversable      1.74964    0.76698   2.281 0.022536 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 409.95  on 296  degrees of freedom
## Residual deviance: 187.74  on 281  degrees of freedom
## AIC: 219.74
##
## Number of Fisher Scoring iterations: 6
```

Interpret the Model

Conclusion