Supervised project: Cardiovascular disease

Tommaso Locatelli

11/11/2021

Abstract

In this report I will analyze the risk factors of cardiovascular disease starting from a dataset that includes different types of features. Using logistic regressions in different ways, I will have to deal with various problems related to confounding effects which, once identified, I will show that they are sensitive to being eliminated by a penalized logistic regression.

Research question and dataset

The aim of this project is to analyze risk factors for heart disease starting from the Cardiovascular Disease dataset available on Kaggle (linked phrase). According to the World Health Organization the most important behavioural risk factors of heart disease are unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol. The effects of behavioural risk factors may show up in individuals as raised blood pressure, raised blood glucose, raised blood lipids, and overweight and obesity. This work will focus on verifying these risk factors and evaluating others according to the variables available in the dataset. In particular, it will be of interest not the accuracy of the prediction but the inference of a relationship between disease and factors.

First we load the dataset, drop the id and start loocking at it.

```
##     age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active
## 1 18393      2    168     62   110    80           1    1     0    0      1
## 2 20228      1    156     85   140    90           3    1     0    0      1
## 3 18857      1    165     64   130    70           3    1     0    0      0
##   cardio
## 1      0
## 2      1
## 3      1
```

Legend of the variables.

There are 3 types of input features:

Objective: factual information;

Examination: results of medical examination;

Subjective: information given by the patient.

Features:

Age | Objective Feature | age | int (days)

Height | Objective Feature | height | int (cm) |

Weight | Objective Feature | weight | float (kg) |

Gender | Objective Feature | gender | categorical code | 1:Female, 2:Male

Systolic blood pressure | Examination Feature | ap_hi | int |

Diastolic blood pressure | Examination Feature | ap_lo | int |

Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |

Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |

Smoking | Subjective Feature | smoke | binary |

Alcohol intake | Subjective Feature | alco | binary |

Physical activity | Subjective Feature | active | binary |

Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

All of the dataset values were collected at the moment of medical examination.

Data cleaning and preparation

Starting from the height and weight columns we obtain the body mass index column using the formula $bmi = weight/height^2$ expressed in kg and m; we report the age in years for convenience and recode 0 as Male instead of 2 for convenience too.

We check the dataset for abnormal values or outliers.

```
##       age            gender           height          weight
##  Min.   :29.00   Min.   :0.0000   Min.   : 55.0   Min.   : 10.00
##  1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:159.0   1st Qu.: 65.00
##  Median :53.00   Median :1.0000   Median :165.0   Median : 72.00
##  Mean   :52.84   Mean   :0.6504   Mean   :164.4   Mean   : 74.21
##  3rd Qu.:58.00   3rd Qu.:1.0000   3rd Qu.:170.0   3rd Qu.: 82.00
##  Max.   :64.00   Max.   :1.0000   Max.   :250.0   Max.   :200.00
##      ap_hi             ap_lo           cholesterol         gluc
##  Min.   : -150.0   Min.   :  -70.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:  120.0   1st Qu.:   80.00   1st Qu.:1.000   1st Qu.:1.000
##  Median :  120.0   Median :   80.00   Median :1.000   Median :1.000
##  Mean   :  128.8   Mean   :   96.63   Mean   :1.367   Mean   :1.226
##  3rd Qu.:  140.0   3rd Qu.:   90.00   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :16020.0   Max.   :11000.00   Max.   :3.000   Max.   :3.000
##      smoke              alco             active           cardio
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.00000   Median :1.0000   Median :0.0000
##  Mean   :0.08813   Mean   :0.05377   Mean   :0.8037   Mean   :0.4997
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##       bmi
##  Min.   :  3.472
##  1st Qu.: 23.875
##  Median : 26.374
##  Mean   : 27.557
##  3rd Qu.: 30.222
##  Max.   :298.667
```

Extreme not acceptable values are present in the columns: ap_hi, ap_lo. Looking at some of the abnormal registration show that there must have been some typo.

```
##   ap_hi ap_lo
## 1   902    60
## 2   906     0
## 3   909    60
```

```
## 4 11500    90
## 5  1420    80
## 6   701   110
```

Due to medical reasons we use as benchmark a minimum pressure of 40 and a maximum of 250.

It seems that there are also anomalous values related to height, weight and bmi.

```
##   height weight       bmi
## 1     76     55  95.22161
## 2     97    170 180.67807
## 3     75    168 298.66667
## 4     71     68 134.89387
## 5     67     57 126.97706
## 6     70     68 138.77551
```

Some lines may have inverted weight and height due to input errors and that cause extreme bmi values, so we cut any record with bmi bigger than 50 and lower than 10, since values lower than 16 and greater than 40 are already considered extreme.

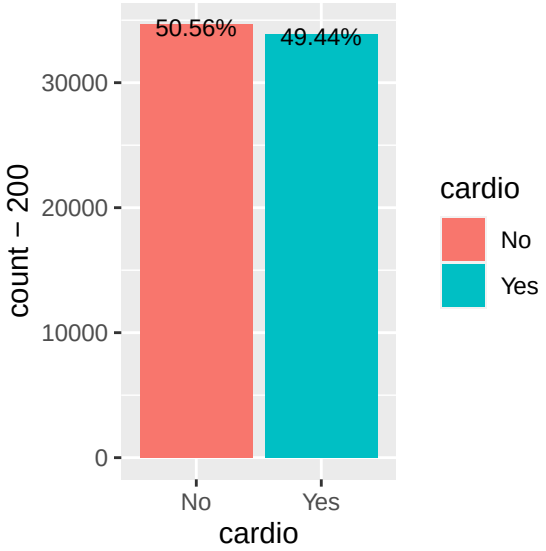
So we get our final dataset with 68519 observation and 13 variables.

```
##      height          weight           ap_hi           ap_lo
##  Min.   :120.0   Min.   : 28.00   Min.   : 70.0   Min.   : 45.00
##  1st Qu.:159.0   1st Qu.: 65.00   1st Qu.:120.0   1st Qu.: 80.00
##  Median :165.0   Median : 72.00   Median :120.0   Median : 80.00
##  Mean   :164.4   Mean   : 73.97   Mean   :126.6   Mean   : 81.38
##  3rd Qu.:170.0   3rd Qu.: 82.00   3rd Qu.:140.0   3rd Qu.: 90.00
##  Max.   :207.0   Max.   :180.00   Max.   :240.0   Max.   :190.00
##       bmi
##  Min.   :10.73
##  1st Qu.:23.88
##  Median :26.31
##  Mean   :27.38
##  3rd Qu.:30.11
##  Max.   :50.00
```

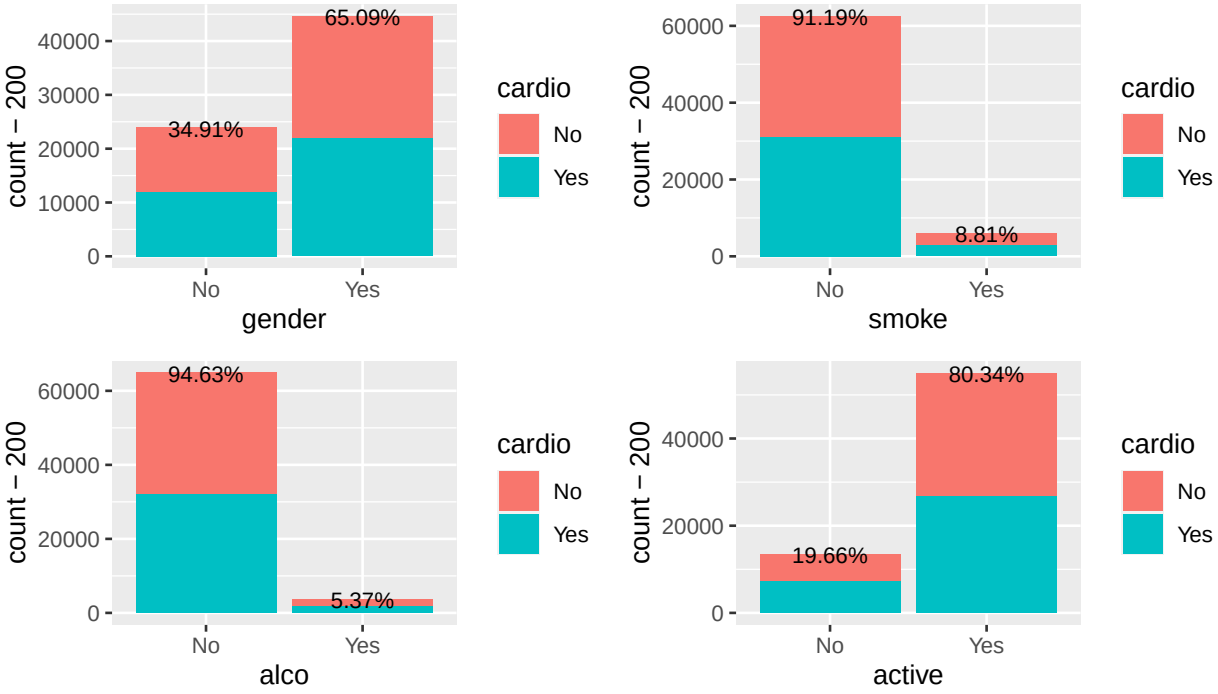
```
## [1] 68518    13
```

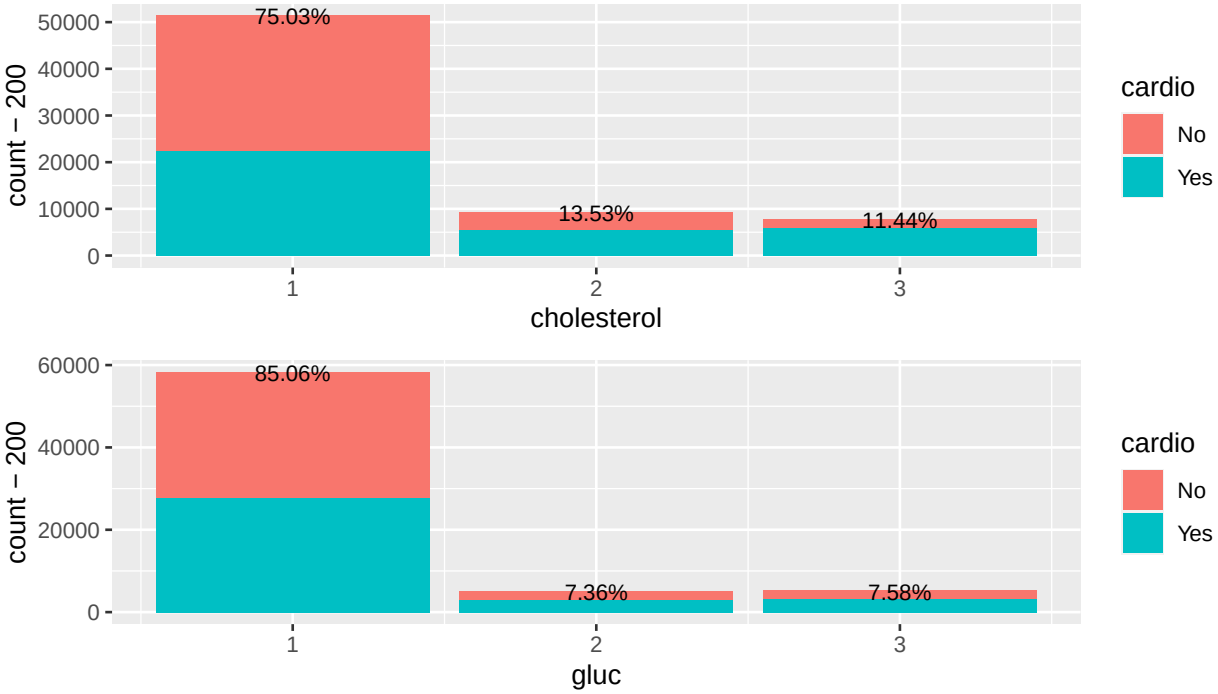
Data description

Let's start by seeing how our variable of interest is distributed within the database and in the various features to evaluate the balance of the sample.

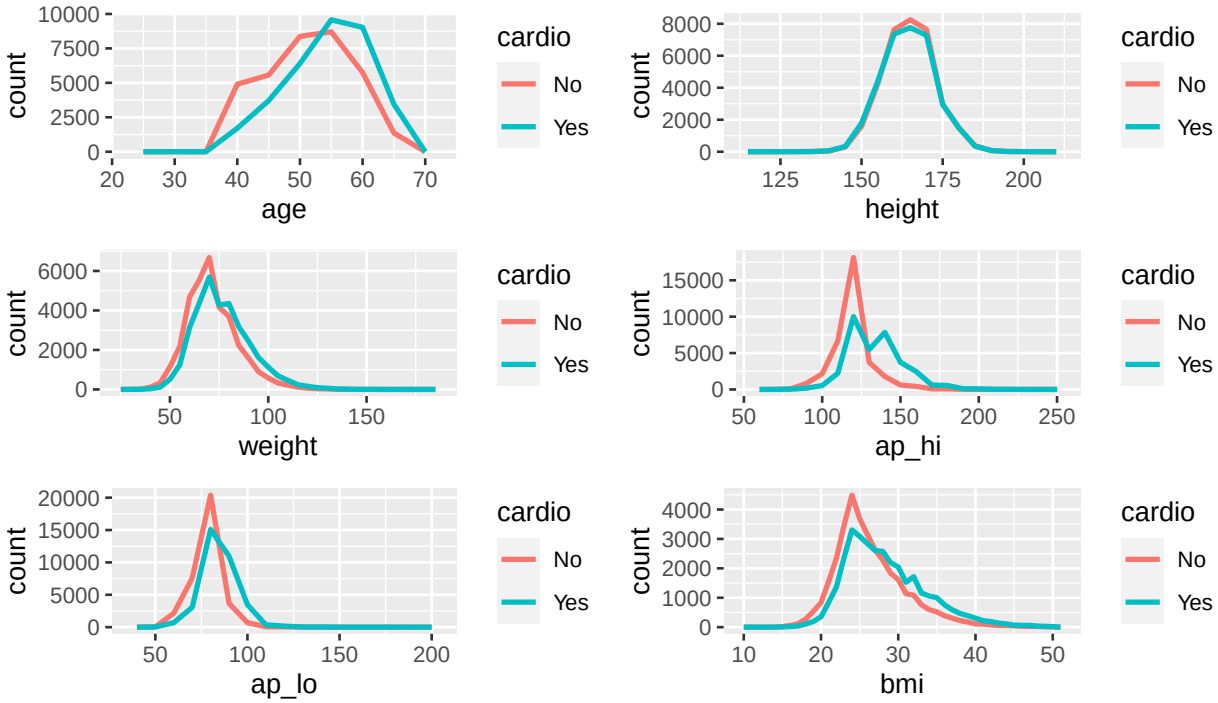


The variable of interest is well balanced within the overall database. As regards the discrete variables, some classes are underrepresented, but the proportion of ill patients seems to grow among those with high levels of cholesterol and glucose.





While for the continuous variables we can see how the share of sick people exceeds that of healthy ones with the increase of: age, weight, bmi and blood pressure.



Before proceeding with the analysis it is good to check the correlation between variables through the correlation matrix and the relative p-values.

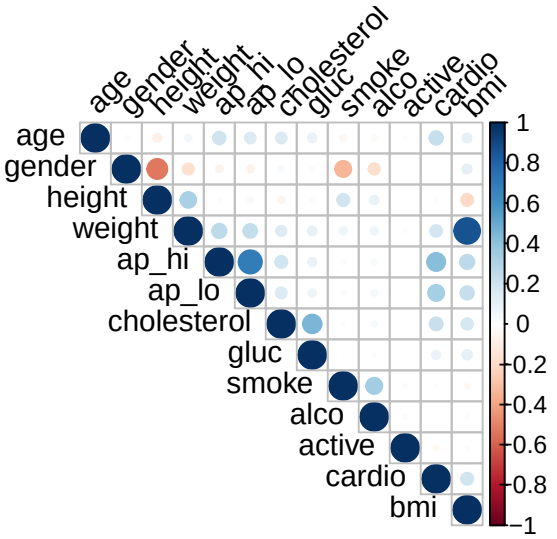
```
##                age gender height weight ap_hi ap_lo cholesterol  gluc smoke
## age           1.00   0.02  -0.09   0.06  0.21  0.15        0.15  0.10 -0.05
## gender        0.02   1.00  -0.52  -0.16 -0.06 -0.07        0.04  0.02 -0.34
```

```
## height      -0.09  -0.52   1.00   0.32  0.02  0.04       -0.05 -0.02  0.20
## weight       0.06  -0.16   0.32   1.00  0.27  0.25        0.14  0.10  0.07
## ap_hi        0.21  -0.06   0.02   0.27  1.00  0.70        0.19  0.09  0.03
## ap_lo        0.15  -0.07   0.04   0.25  0.70  1.00        0.16  0.08  0.02
## cholesterol  0.15   0.04  -0.05   0.14  0.19  0.16        1.00  0.45  0.01
## gluc         0.10   0.02  -0.02   0.10  0.09  0.08        0.45  1.00 -0.01
## smoke       -0.05  -0.34   0.20   0.07  0.03  0.02        0.01 -0.01  1.00
## alco        -0.03  -0.17   0.10   0.07  0.03  0.04        0.04  0.01  0.34
## active      -0.01  -0.01  -0.01  -0.02  0.00  0.00        0.01 -0.01  0.03
## cardio       0.24  -0.01  -0.01   0.18  0.43  0.34        0.22  0.09 -0.02
## bmi          0.10   0.11  -0.19   0.86  0.27  0.24        0.17  0.12 -0.03
##              alco active cardio   bmi
## age         -0.03  -0.01   0.24  0.10
## gender      -0.17  -0.01  -0.01  0.11
## height       0.10  -0.01  -0.01 -0.19
## weight       0.07  -0.02   0.18  0.86
## ap_hi        0.03   0.00   0.43  0.27
## ap_lo        0.04   0.00   0.34  0.24
## cholesterol  0.04   0.01   0.22  0.17
## gluc         0.01  -0.01   0.09  0.12
## smoke        0.34   0.03  -0.02 -0.03
## alco         1.00   0.03  -0.01  0.02
## active       0.03   1.00  -0.04 -0.02
## cardio      -0.01  -0.04   1.00  0.19
## bmi          0.02  -0.02   0.19  1.00
```

```
##             age gender height weight ap_hi ap_lo cholesterol gluc smoke alco
## age         1
## gender          1
## height          .      1
## weight                 .      1
## ap_hi                                 1
## ap_lo                                 ,     1
## cholesterol                                       1
## gluc                                              .           1
## smoke           .                                                   1
## alco                                                                .     1
## active
## cardio                                .     .
## bmi                           +
##             active cardio bmi
## age
## gender
## height
## weight
## ap_hi
## ap_lo
## cholesterol
## gluc
## smoke
## alco
## active      1
## cardio             1
## bmi                       1
```

```
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

We note how in addition to trivial relationships, such as: weight and height, systolic pressure and diastolic pressure, gender and height, bmi and weight; that there are a positive and significant correlations between glucose and cholesterol, smoking and alcohol, pressures and diseases and a negative correlation between smoking and gender. Let's summarize these observations with the following plot.



Before proceeding with the analysis it is good to point out two major limitations of this dataset: the absence of many variables related to the risk of suffering from cardiovascular diseases, such as genetic factors, diet, celiac disease, air pollution, lipid concentration in the blood and more; finally, the fact that many variables have a very small range of values, especially the subjective ones.

Data analysis

Before studying heart disease starting from all the features at the same time, let's take a closer look at the behavioral factors and their expressions recognized by the WHO as risky.

Behavioral risks

Due to the absence of a variable linked to the type of diet followed we will limit ourselves to assessing smoking, alcohol consumption and inactivity. To do that we fit a logistic regression for each feature also reporting the odds ratio.

Smoke

```
##
## Call:
## glm(formula = cardio ~ smoke, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.172  -1.172  -1.124   1.183   1.232
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.012420   0.008001  -1.552    0.121
## smokeYes    -0.114281   0.027003  -4.232 2.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 94960  on 68516  degrees of freedom
## AIC: 94964
##
## Number of Fisher Scoring iterations: 3
```

```
## Waiting for profiling to be done...
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 0.9876567 0.9722882 1.0032675
## smokeYes    0.8920072 0.8460027 0.9404671
```

Alcohol

```
##
## Call:
## glm(formula = cardio ~ alco, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.169  -1.169  -1.140   1.185   1.215
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -0.018754   0.007855  -2.388   0.0170 *
## alcoYes     -0.069443   0.033940  -2.046   0.0408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 94973  on 68516  degrees of freedom
## AIC: 94977
##
## Number of Fisher Scoring iterations: 3
```

```
## Waiting for profiling to be done...
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 0.9814209 0.9664276 0.9966461
## alcoYes     0.9329137 0.8728436 0.9970598
```

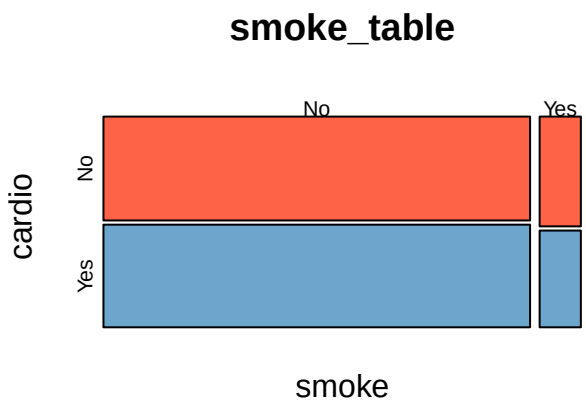
Physical activity

```
##
## Call:
## glm(formula = cardio ~ active, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.233  -1.152  -1.152   1.203   1.203
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.12936    0.01727   7.491 6.84e-14 ***
## activeYes   -0.18896    0.01926  -9.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 94881  on 68516  degrees of freedom
## AIC: 94885
##
## Number of Fisher Scoring iterations: 3
```

```
## Waiting for profiling to be done...
```

```
##                    OR     2.5 %    97.5 %
## (Intercept) 1.1380952 1.1002330 1.1772895
## activeYes   0.8278213 0.7971465 0.8596588
```

Contrary to our expectations, logistic regression seems to point to smoke and alco as protective factors of the disease. In the case of smoking even with a good significance.



We note that the proportion of sick people is actually slightly lower among smokers, a similar situation is also found in alcohol users since the two variables are positively correlated. We can also check it numerically with the proportional crosstabs that show how the percentage of ill patients decrease in the smoking class.

```
##      cardio
## smoke       No      Yes
##   No  0.503105 0.496895
##   Yes 0.531633 0.468367
```

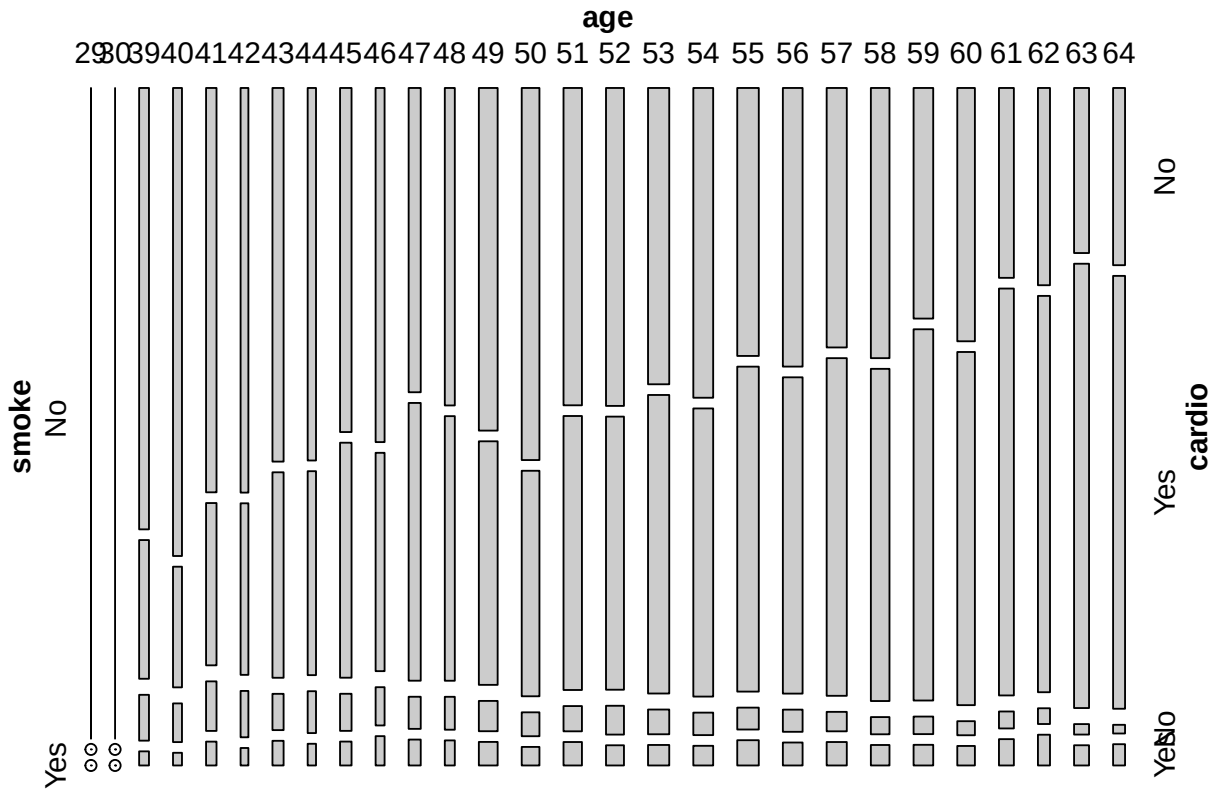
```
##      cardio
## alco         No       Yes
##   No  0.5046883 0.4953117
##   Yes 0.5220348 0.4779652
```

Finally we can check the indipendence of the two classes with the Chi-squared test.

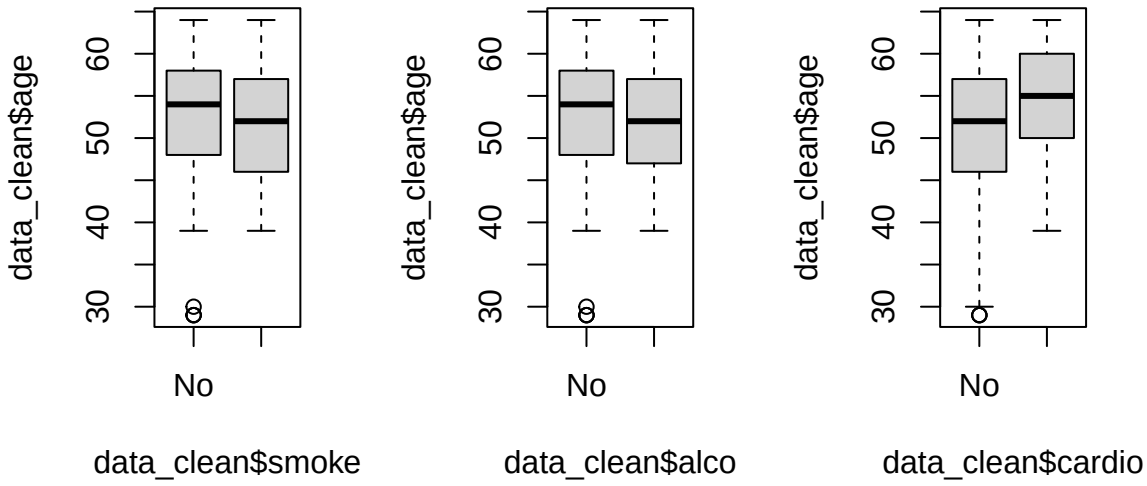
```
##
##  Pearson's Chi-squared test
##
## data:  smoke_table
## X-squared = 17.926, df = 1, p-value = 2.296e-05
```

```
##
##  Pearson's Chi-squared test
##
## data:  alco_table
## X-squared = 4.1876, df = 1, p-value = 0.04072
```

In both cases we find a significant difference from the independence, especially for the smoke variable. All these considerations leads us to consider the possibility of finding ourselves in front of a confounding effect. Perhaps a third variable influences both the presence of heart disease and our explanatory variables causing a spurious association.



Through a mosaic graph that shows the distribution of patients and smokers according to age, it is possible to hypothesize that age is a determining factor for heart problems while the proportion of smokers decreased, this could induce the observed spurious relationship.



Coherently we find that smokers and alcohol users are on average younger while the sick are often the oldest.

Let's check the hypothesis that probability of cardiovascualar disease increase due to the age and also that probability of to be a smoker decrease with increasing age through two other logistic regressions.

Age

```
##
## Call:
## glm(formula = cardio ~ age, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5305  -1.1402  -0.7746   1.1211   1.6434
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.937232   0.064384  -61.15   <2e-16 ***
## age          0.074023   0.001206   61.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 90955  on 68516  degrees of freedom
## AIC: 90959
##
## Number of Fisher Scoring iterations: 4
```

Smoke prediction with respect to age

```
##
## Call:
## glm(formula = smoke ~ age, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5626  -0.4468  -0.4214  -0.3926   2.3206
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.051750   0.102769  -10.23   <2e-16 ***
## age         -0.024542   0.001963  -12.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 40860  on 68517  degrees of freedom
## Residual deviance: 40705  on 68516  degrees of freedom
## AIC: 40709
##
```

```
## Number of Fisher Scoring iterations: 5
```

Expression of risks in medical examination

Let's look at the examination features and bmi as WHO considers them symptoms of a high risk of heart disease. We note, however, that the concentration of lipids in the blood is not present.

Blood pressure

```
##
## Call:
## glm(formula = cardio ~ ap_hi + ap_lo, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8608  -1.0136  -0.3785   1.0132   2.6193
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.2047222  0.0959978  -95.89   <2e-16 ***
## ap_hi        0.0606420  0.0008645   70.14   <2e-16 ***
## ap_lo        0.0191168  0.0013428   14.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 80414  on 68515  degrees of freedom
## AIC: 80420
##
## Number of Fisher Scoring iterations: 4
```

Cholesterol

```
##
## Call:
## glm(formula = cardio ~ cholesterol, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.681  -1.067  -1.067   1.292   1.292
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.96413    0.01820  -52.98   <2e-16 ***
## cholesterol  0.69913    0.01251   55.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
```

```
## Residual deviance: 91501  on 68516  degrees of freedom
## AIC: 91505
##
## Number of Fisher Scoring iterations: 4
```

Glucose

```
##
## Call:
## glm(formula = cardio ~ gluc, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.416  -1.138  -1.138   1.217   1.217
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.41214    0.01839  -22.41   <2e-16 ***
## gluc         0.31885    0.01375   23.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 94425  on 68516  degrees of freedom
## AIC: 94429
##
## Number of Fisher Scoring iterations: 4
```

Bmi

```
##
## Call:
## glm(formula = cardio ~ bmi, family = binomial(link = "logit"),
##     data = data_clean)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9669  -1.1065  -0.8808   1.1821   1.7814
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.220308   0.045166  -49.16   <2e-16 ***
## bmi          0.080409   0.001634   49.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 94977  on 68517  degrees of freedom
## Residual deviance: 92369  on 68516  degrees of freedom
## AIC: 92373
##
```

```
## Number of Fisher Scoring iterations: 4
```

This time we get the coefficients we expected and in all cases they turn out to be significant.

Attempts to build a multiple model

At this point it would be interesting to build a model that takes into account several variables at the same time and if possible to avoid problems of confounding effect.

Total multiple model

As a first approach we try to construct a multiple logistic model that takes into account all the available variables parameterizing it starting from a training set and judging its performance in the test set.

```
##
## Call:
## glm(formula = cardio ~ ., family = binomial(link = "logit"),
##     data = dtrain[-c(3, 4)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7645  -0.9227  -0.3430   0.9365   2.5644
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.850873   0.147322 -80.442  < 2e-16 ***
## age           0.050674   0.001610  31.472  < 2e-16 ***
## genderYes    -0.043695   0.023475  -1.861 0.062698 .
## ap_hi         0.053824   0.001052  51.158  < 2e-16 ***
## ap_lo         0.015484   0.001643   9.426  < 2e-16 ***
## cholesterol   0.487317   0.018632  26.156  < 2e-16 ***
## gluc         -0.109570   0.021107  -5.191 2.09e-07 ***
## smokeYes     -0.185991   0.041703  -4.460 8.20e-06 ***
## alcoYes      -0.182201   0.051137  -3.563 0.000367 ***
## activeYes    -0.231732   0.026223  -8.837  < 2e-16 ***
## bmi           0.030396   0.002232  13.617  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66485  on 47962  degrees of freedom
## Residual deviance: 53861  on 47952  degrees of freedom
## AIC: 53883
##
## Number of Fisher Scoring iterations: 4
```

Regarding the fit of the model several things should be noted:

1. almost any coefficient is significant
2. the glucose is now considered a protective factor
3. alcohol and smoke are considered as risk factors anyway
4. gender is correctly recognized as a protective factor as we know that incidence of CVD in women is usually lower than in men (despite the low significance)

5. we excluded height and weight as it would have suffered from high collinearity with the bmi

Performance in the training set:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    No   Yes
##        No  19057  7916
##        Yes  5194 15796
##
##                Accuracy : 0.7267
##                  95% CI : (0.7227, 0.7306)
##     No Information Rate : 0.5056
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4526
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6662
##             Specificity : 0.7858
##          Pos Pred Value : 0.7525
##          Neg Pred Value : 0.7065
##              Prevalence : 0.4944
##          Detection Rate : 0.3293
##    Detection Prevalence : 0.4376
##       Balanced Accuracy : 0.7260
##
##        'Positive' Class : Yes
##
```

Performance in the test set:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  8196 3411
##        Yes 2197 6751
##
##                Accuracy : 0.7272
##                  95% CI : (0.721, 0.7333)
##     No Information Rate : 0.5056
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4535
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.6643
##             Specificity : 0.7886
##          Pos Pred Value : 0.7545
##          Neg Pred Value : 0.7061
##              Prevalence : 0.4944
##          Detection Rate : 0.3284
```

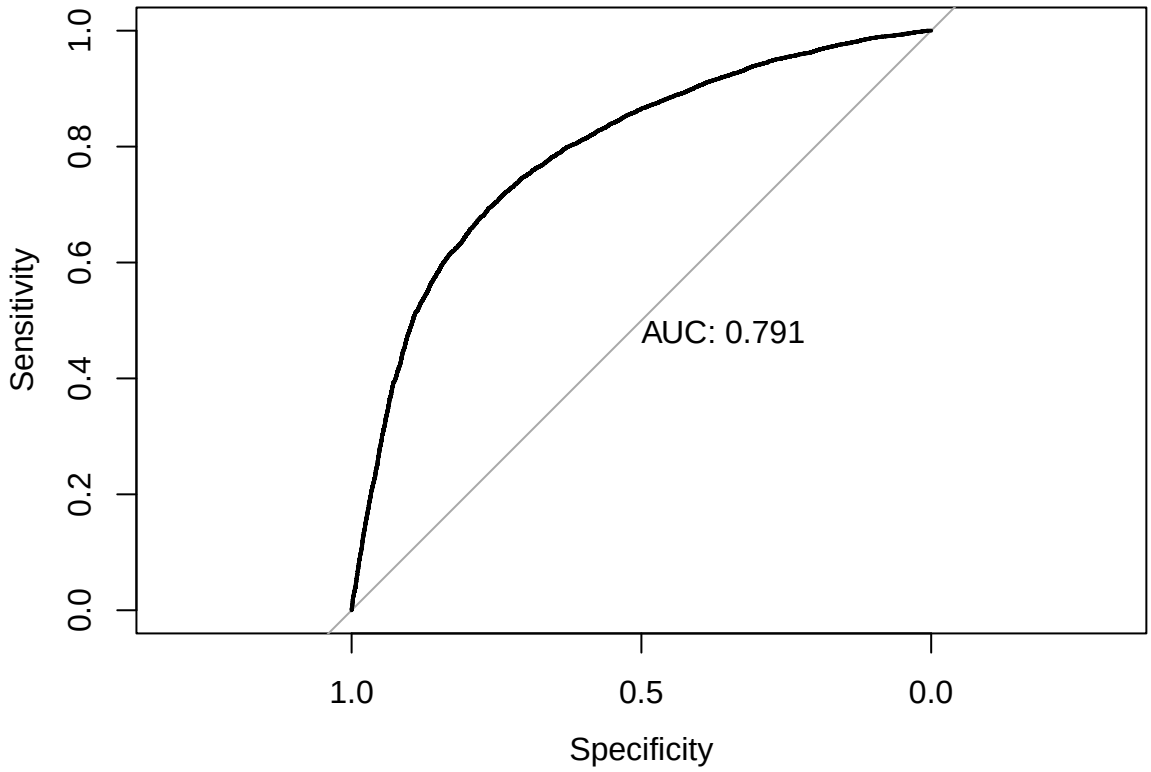
```
##    Detection Prevalence : 0.4353
##       Balanced Accuracy : 0.7265
##
##        'Positive' Class : Yes
##
```

It is important to note that the sensitivity is quite low in both cases

ROC curve:

```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

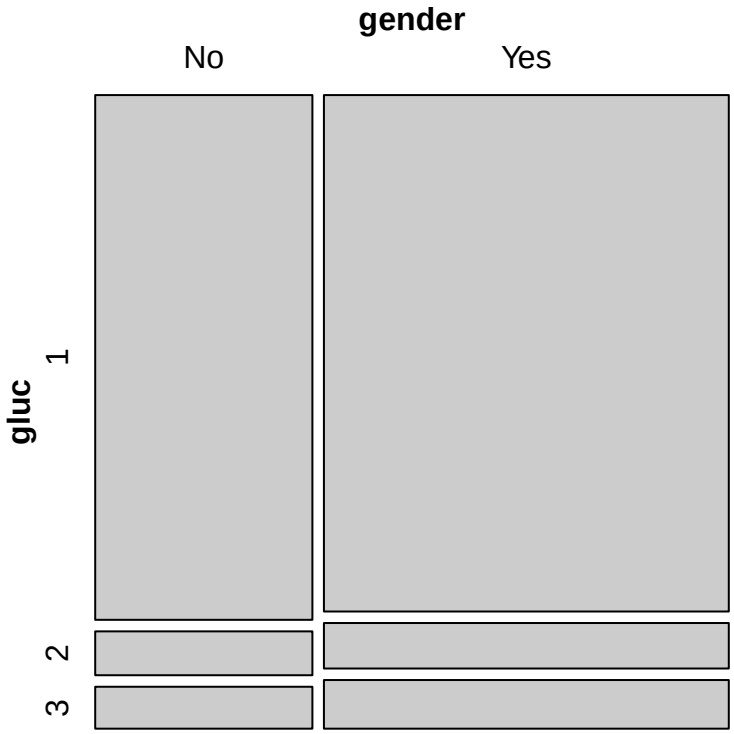


For what concern the change in glucose sign is significant as it depends on the insertion of cholesterol, in fact, excluding the latter from the regression, the glucose coefficient returns to being positive.

```
##
## Call:
## glm(formula = cardio ~ ., family = binomial(link = "logit"),
##     data = dtrain[-c(3, 4, 7)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8638  -0.9472  -0.3388   0.9631   2.6673
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.082637   0.147002 -82.194  < 2e-16 ***
```

```
## age           0.053587   0.001596  33.574  < 2e-16 ***
## genderYes    -0.021123   0.023276  -0.908 0.364141
## ap_hi         0.055565   0.001050  52.922  < 2e-16 ***
## ap_lo         0.016139   0.001634   9.879  < 2e-16 ***
## gluc          0.136161   0.018474   7.370  1.7e-13 ***
## smokeYes     -0.159128   0.041287  -3.854 0.000116 ***
## alcoYes      -0.143501   0.050409  -2.847 0.004417 **
## activeYes    -0.218123   0.026037  -8.377  < 2e-16 ***
## bmi           0.035084   0.002209  15.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 66485  on 47962  degrees of freedom
## Residual deviance: 54578  on 47953  degrees of freedom
## AIC: 54598
##
## Number of Fisher Scoring iterations: 4
```

This could be linked to the fact that both glucose and cholesterol, in addition to being important risk factors, are also generally higher among women which instead is a protective factor.



Finally, it is worth to evaluate the collinearity between the latter.

VIF for collinearity:

```
##       age genderYes     ap_hi     ap_lo      gluc  smokeYes   alcoYes activeYes
```

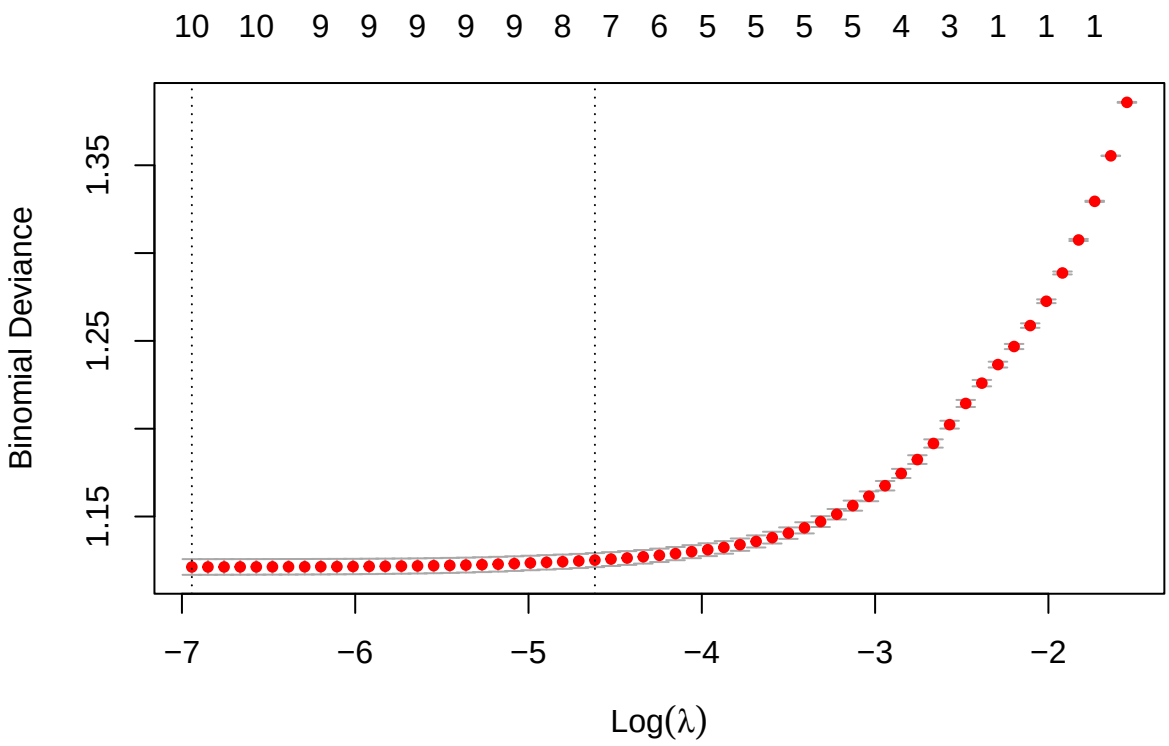
```
##  1.011684  1.152464  1.631240  1.616864  1.012928  1.246030  1.139376  1.002020
##       bmi
##  1.059126
```

But as can be seen there are no significantly high values and the two maxima related to pressure are clearly induced by the deep link between these two measures.

Lasso logistic regression

In light of the various problems that have emerged regarding the inclusion of multiple features at the same time, we apply a penalized logistic regression of the lasso type.

Find the optimal value of lambda that minimizes the cross-validation error:



As can be seen from the plot, the optimal value of lambda which corresponds to the vertical dashed line on the left tends to zero and its value is:

```
## [1] 0.0009656114
```

In fact it is not surprising that no features have been excluded:

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -11.85653902
## age           0.05104059
## genderYes    -0.01421430
## ap_hi         0.05290619
## ap_lo         0.01694021
## cholesterol   0.48253254
```

```
## gluc         -0.11551060
## smokeYes     -0.10501931
## alcoYes      -0.22068932
## activeYes    -0.20486310
## bmi           0.02879069
```

Since the interest was to simplify the model rather than improve its performance let's look also the value of lambda that gives the simplest model but also lies within one standard error of the optimal value of lambda and relatives coeficients.

```
## [1] 0.009883332
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -10.88233564
## age           0.04519213
## genderYes     .
## ap_hi         0.05021304
## ap_lo         0.01374941
## cholesterol   0.37502123
## gluc          .
## smokeYes      .
## alcoYes      -0.03781811
## activeYes    -0.08976485
## bmi           0.02191688
```

Glucose, gender and smoke were excluded probably because the weakness of the spurious relationship is somehow identified and therefore excluded by the introduction of bias that involves the increase of lambda.

Conclusion

Several problems have arisen in applying logistic regressions to this database, often leading to a not simple interpretation of the coefficients obtained. This could be linked to the fact that even if the database is well balanced as regards the target variable, other variables are poorly balanced between them, age and gender seem to make the proportion of other variables vary a lot, which is a fairly typical situation in observational studies like this one. In addition to this, however, it was possible to bring out the strong relationship that heart disease has with various variables, such as: cholesterol, high blood pressure and age. An interesting final conclusion could be that the penalty due to the application of Lasso affected the coefficients that were not in line with expectations, this could be interpreted as the fact that in this case the spurious relationship induced by the confounders is less supported by the data than other risk factors.

Appendix

```
cardio_train <- read.csv("C:/Users/tomma/Desktop/Salini projects/Supervised/cardio_train.csv", sep=";")
cardio_train = subset(cardio_train, select = -c(id) )
head(cardio_train, n=3L)
library(plyr)
library(dplyr)
library(epitools)
library(caret)
library(gridExtra)
library(tidyverse)
library(rsample)
library(e1071)
library(GGally)
library(data.table)
library(DT)
library(readr)
library(ggplot2)
library(dplyr)
library(tidyr)
library(corrplot)
library(rms)
library(MASS)
library(ROCR)
library(gplots)
library(pROC)
library(grid)
library(vcd)
library(ggpubr)
cardio_train$age=cardio_train$age%/%365
cardio_train = cardio_train %>%
  mutate(bmi = weight/((height/100)^2))
cardio_train$gender[cardio_train$gender==2]=0
summary(cardio_train)
head(cardio_train[c(5,6)] %>% filter(ap_hi > 250))
cardio_train=cardio_train[ which(cardio_train$ap_hi>40
              & cardio_train$ap_hi<250
              & cardio_train$ap_lo<250
              & cardio_train$ap_lo>40), ]
head(cardio_train[c(3,4,13)] %>% filter(height < 100))
cardio_train=cardio_train[ which(cardio_train$bmi>10
              & cardio_train$height<240
              & cardio_train$bmi<50), ]
data_clean=cardio_train
data_clean$gender <- as.factor(mapvalues(data_clean$gender,
                                         from=c("0","1"),
                                         to=c("No", "Yes")))
data_clean$smoke <- as.factor(mapvalues(data_clean$smoke,
                                        from=c("0","1"),
                                        to=c("No", "Yes")))
data_clean$alco <- as.factor(mapvalues(data_clean$alco,
                                       from=c("0","1"),
                                       to=c("No", "Yes")))
```

```r
data_clean$active <- as.factor(mapvalues(data_clean$active,
                                         from=c("0","1"),
                                         to=c("No", "Yes")))
data_clean$cardio <- as.factor(mapvalues(data_clean$cardio,
                                         from=c("0","1"),
                                         to=c("No", "Yes")))
summary(cardio_train[c(3,4,5,6,13)])
dim(cardio_train)
# cario distribution
p=ggplot(data_clean, aes(x = cardio)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)
p
#Gender plot
p1 <- ggplot(data_clean, aes(x = gender)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)

#Smoke plot
p2 <- ggplot(data_clean, aes(x = smoke)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)

#Alco plot
p3 <- ggplot(data_clean, aes(x = alco)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)

#Active plot
p4 <- ggplot(data_clean, aes(x = active)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)
```

```
#Plot demographic data within a grid
grid.arrange(p1, p2, p3, p4, ncol=2)
#Cholesterol plot
p5 <- ggplot(data_clean, aes(x = cholesterol)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)

#Gluc billing plot
p6 <- ggplot(data_clean, aes(x = gluc)) +
  geom_bar(aes(fill = cardio)) +
  geom_text(aes(y = ..count.. -200,
                label = paste0(round(prop.table(..count..),4) * 100, '%')),
            stat = 'count',
            position = position_dodge(.1),
            size = 3)

#Plot contract data within a grid
grid.arrange(p5, p6, ncol=1)
#age histogram
p7 <- ggplot(data = data_clean, aes(age, color = cardio))+
  geom_freqpoly(binwidth = 5, size = 1)

#height histogram
p8 <- ggplot(data = data_clean, aes(height, color = cardio))+
  geom_freqpoly(binwidth = 5, size = 1)

#weight charges histogram
p9 <- ggplot(data = data_clean, aes(weight, color = cardio))+
  geom_freqpoly(binwidth = 5, size = 1)

p10 <- ggplot(data = data_clean, aes(ap_hi, color = cardio))+
  geom_freqpoly(binwidth = 10, size = 1)

#ap_lo charges histogram
p11 <- ggplot(data = data_clean, aes(ap_lo, color = cardio))+
  geom_freqpoly(binwidth = 10, size = 1)

#bmi histogram
p12 <- ggplot(data = data_clean, aes(bmi, color = cardio))+
  geom_freqpoly(binwidth = 1, size = 1)

#Plot quantitative data within a grid
grid.arrange(p7, p8, p9,p10, p11, p12, ncol=2)
res<-cor(cardio_train)
round(res, 2)
symnum(res, abbr.colnames = FALSE)
library(corrplot)
corrplot(res, type = "upper",
         tl.col = "black", tl.srt = 45)
```

```
lr_fits <- glm(cardio ~ smoke, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fits)
exp(cbind(OR = coef(lr_fits), confint(lr_fits)))
lr_fit <- glm(cardio ~ alco, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
exp(cbind(OR = coef(lr_fit), confint(lr_fit)))
lr_fit <- glm(cardio ~ active, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
exp(cbind(OR = coef(lr_fit), confint(lr_fit)))
smoke_table <- xtabs(~smoke+cardio, data=data_clean)
alco_table <- xtabs(~alco+cardio, data=data_clean)
plot(smoke_table , col=c("tomato","skyblue3"))
prop.table(smoke_table, 1)
prop.table(alco_table, 1)
Test <- chisq.test(smoke_table, correct=FALSE)
Test

Test <- chisq.test(alco_table, correct=FALSE)
Test
library(vcd)
hec2 <- structable(cardio ~  age + smoke , data = data_clean)

mosaic(hec2, split_vertical = c(TRUE, FALSE, FALSE),
       labeling_args = list(abbreviate = c(Eye = 3)))
boxplot(data_clean$age ~ data_clean$smoke)
boxplot(data_clean$age ~ data_clean$alco)
boxplot(data_clean$age ~ data_clean$cardio)
lr_fita <- glm(cardio ~ age, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fita)
lr_fit <- glm(smoke ~ age, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
lr_fit <- glm(cardio ~ ap_hi + ap_lo, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
lr_fit <- glm(cardio ~ cholesterol, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
lr_fit <- glm(cardio ~ gluc, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
lr_fit <- glm(cardio ~ bmi, data = data_clean,
          family=binomial(link='logit'))
summary(lr_fit)
#Train and test
set.seed(1) #1,40,41,42
split_train_test <- createDataPartition(data_clean$cardio,p=0.7,list=FALSE)
dtrain<- data_clean[split_train_test,]
dtest<-  data_clean[-split_train_test,]
```

```r
#logistic regression
lr_fit <- glm(cardio ~., data = dtrain[-c(3,4)],
              family=binomial(link='logit'))
summary(lr_fit)
lr_prob1 <- predict(lr_fit, dtest, type="response")
lr_pred1 <- ifelse(lr_prob1 > 0.5,"Yes","No")
lr_prob2 <- predict(lr_fit, dtrain, type="response")
lr_pred2 <- ifelse(lr_prob2 > 0.5,"Yes","No")
lr_tab1 <- table(Predicted = lr_pred2, Actual = dtrain$cardio)
lr_tab2 <- table(Predicted = lr_pred1, Actual = dtest$cardio)
#train
confusionMatrix(
  as.factor(lr_pred2),
  as.factor(dtrain$cardio),
  positive = "Yes"
)
#test
confusionMatrix(
  as.factor(lr_pred1),
  as.factor(dtest$cardio),
  positive = "Yes"
)
#ROC, sensitivity TP, 1-specificity FP
lr_prob2 <- predict(lr_fit, dtest, type="response")
test_roc = roc(dtest$cardio ~ lr_prob2, plot = TRUE, print.auc = TRUE)
#logistic regression
lr_fit <- glm(cardio ~., data = dtrain[-c(3,4,7)],
              family=binomial(link='logit'))
summary(lr_fit)
hec4 <- structable(gluc~ gender, data = data_clean)
mosaic(hec4, split_vertical = c(T,F),
       labeling_args = list(abbreviate = c(Eye = 3)))
#vif for collinearity
vif(lr_fit)
# Split the data into training and test set
set.seed(123)
training.samples <- data_clean$cardio %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- data_clean[training.samples,-c(3,4) ]
test.data <- data_clean[-training.samples, -c(3,4)]

# Dumy code categorical predictor variables
x <- model.matrix(cardio~., train.data)[,-1]
# Convert the outcome (class) to a numerical variable
y <- ifelse(train.data$cardio == "Yes", 1, 0)

library(glmnet)
set.seed(123)

#fit <- glmnet(x, y)
#plot(fit, xvar = "lambda", label = TRUE)
#for (i in fit$lambda){
 # if(){
```

```
  #  print(coef(fit, s = i))
  #}
#}
cv.lasso <- cv.glmnet(x, y, alpha = 1, family = "binomial")
plot(cv.lasso)
cv.lasso$lambda.min
coef(cv.lasso, cv.lasso$lambda.min)
cv.lasso$lambda.1se
coef(cv.lasso, cv.lasso$lambda.1se)
```