Linear Classification: Heart Disease

Data Mining

Data

In this exercise we are going to analyze coronary heart disease determinants. We use data taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal. It contains measurements for several variables of males in a heart-disease high-risk region of the Western Cape, South Africa. These data are, amomg others,

Variable	Description
sbp	systolic blood pressure
tobacco	cumulative tobacco (kg)
ldl	low density lipoprotein cholesterol
famhist	family history of heart disease (Present, Absent)
alcohol	current alcohol consumption
age	age at onset
chd	response, coronary heart disease

The data is available online and can be loaded directly into R.

```
heart = read.table("https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.data",
                   sep=",",head=T,row.names=1)
```

We use *attach()* so that we can refer to the names of the variables directly.

```
attach(heart)
```

The response variable is the existence of myocardial infarctions, heart attacks. As explanatory variables, we will use measurements of health related variables. We are interested in fitting a linear model like

$$cdh = \beta_0 + \beta_1 tobacco + \beta_2 ldl + \beta_3 famhist + \beta_4 age + beta_5 sbp + \beta_6 alcohol.$$

As such, we are in a classification setting.

Logistic Regression

We start by fitting a logistic regression to the data. Logistic regression can be called by the generalized linear models function, *glm()*, by using the additional option *family=binomial*. We then print the results of the estimation by the *summary()* command.

```
glm.fit = glm(chd~tobacco+ldl+famhist+age+sbp+alcohol,family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + famhist + age + sbp + alcohol,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7781  -0.8517  -0.4601   0.9250   2.5059
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.7732879  0.8051050  -5.929 3.05e-09 ***
## tobacco         0.0796772  0.0260951   3.053  0.00226 **
## ldl             0.1654703  0.0544310   3.040  0.00237 **
## famhistPresent  0.9280401  0.2241231   4.141 3.46e-05 ***
## age             0.0416009  0.0101936   4.081 4.48e-05 ***
## sbp             0.0048733  0.0055584   0.877  0.38062
## alcohol         0.0006143  0.0044321   0.139  0.88976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 484.61  on 455  degrees of freedom
## AIC: 498.61
##
## Number of Fisher Scoring iterations: 4
```

We notice that the standard errors of the estimators associated to *sbp* and *alcohol* are large relative to the values of the estimators, which points to them not beign different than zero. This notion can be statistically shown by looking at their z-values, which are quite close to zero, with associated p-values quite large.

We may be interested in removing these non-significant variables from the specification; thus, we estimate the following model

$$cdh = \beta_0 + \beta_1 tobacco + \beta_2 ldl + \beta_3 famhist + \beta_4 age + beta_5 sbp + \beta_6 alcohol.$$

by logistic regression.

```
glm.fit = glm(chd~tobacco+ldl+famhist+age,family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = chd ~ tobacco + ldl + famhist + age, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7559  -0.8632  -0.4545   0.9457   2.4904
```

```
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.204275   0.498315  -8.437  < 2e-16 ***
## tobacco         0.080701   0.025514   3.163  0.00156 **
## ldl             0.167584   0.054189   3.093  0.00198 **
## famhistPresent  0.924117   0.223178   4.141 3.46e-05 ***
## age             0.044042   0.009743   4.521 6.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 461  degrees of freedom
## Residual deviance: 485.44  on 457  degrees of freedom
## AIC: 495.44
##
## Number of Fisher Scoring iterations: 4
```

As we will see in future lectures, we can use the Akaike Criterion AIC to choose between competing models. In this example, we notice that the both AIC are close although it is smaller for the reduced model. This may be used as a determinant as to prefer the reduced model given its reduced complexity over the full one above.

Moreover, in the reduced model all variables are statistically significant and have the expected sign. In particular, smoking and higher levels of cholesterol increase the probability of developing a myocardial infarction, a heart attack.

Linear Discriminant Analysis

We fit now LDA to the specification. We use the *lda()* function which is part of the "MASS" package.

```
library("MASS")
lda.fit = lda(chd~tobacco+ldl+famhist+age)
lda.fit
```

```
## Call:
## lda(chd ~ tobacco + ldl + famhist + age)
##
## Prior probabilities of groups:
##         0         1
## 0.6536797 0.3463203
##
## Group means:
##    tobacco      ldl famhistPresent      age
## 0 2.634735 4.344238      0.3178808 38.85430
## 1 5.524875 5.487938      0.6000000 50.29375
##
## Coefficients of linear discriminants:
##                        LD1
## tobacco        0.08804111
## ldl            0.16591563
## famhistPresent 0.90867059
## age            0.03486466
```

We note that the estimates are quite close to the ones from logistic regression. This due to the fact that the two algorithms assume a linear model, differing only on the estimation technique.

Prediction

Once we have fitted all the models, we can use the estimates to assess the probabilities of developing myocardial infarctions at different levels of smoking.

To do so, we construct a new variable with the rest of the other variables fixed at their mean, while we fix tobacco to three different levels: 0, its mean, and the third quartile.

```
new = data.frame(tobacco=c(0,mean(tobacco),13.60),ldl=rep(mean(ldl),3),
                 famhist=c("Present","Present","Present"),age=rep(mean(age),3))
```

Once we have defined the new data frame we can make predictions by the *predict()* function.

```
lda.pred = predict(lda.fit,newdata = new)
glm.pred = predict(glm.fit,newdata = new,type="response")
```

We can then note the probabilityes by

```
glm.pred
```

```
##         1         2         3
## 0.3543363 0.4239367 0.6218706
```

```
lda.pred$posterior[,2]
```

```
##         1         2         3
## 0.3568463 0.4409821 0.6742106
```

Note the increase in the probabilities increase as the variable tobacco increases.

Furthermore, note that the effect depends on the value of the other variables. To show this, we evaluate the effect of an increase in tobacco consumption for individuals without famility history of heart disease.

```
new2 = data.frame(tobacco=c(0,mean(tobacco),13.60),ldl=rep(mean(ldl),3),
                 famhist=c("Absent","Absent","Absent"),age=rep(mean(age),3))
lda.pred = predict(lda.fit,newdata = new2)
glm.pred = predict(glm.fit,newdata = new2,type="response")
glm.pred
```

```
##         1         2         3
## 0.1788514 0.2260502 0.3949335
```

```
lda.pred$posterior[,2]
```

```
##         1         2         3
## 0.1696562 0.2251043 0.4324908
```

Note that, as before, the probabilities increase as the variable tobacco increases. Nonetheless, the probabilities are lower than for individuals with family history of heart disease.