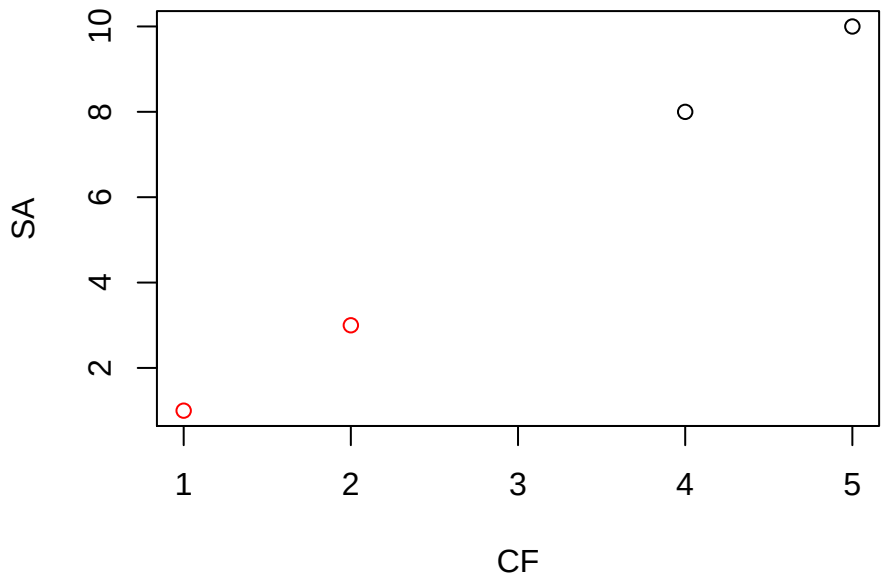
Corporate Risk - Scores & Rules with R

Introduction to Logistic Regression

We consider a sample of four companies.
The data frame: `mydata <- data.frame(CF=c(5,4,1,2),SA=c(10,8,1,3),risk=c('A','A','B','B'))` gives for each company: the values of two financial descriptors (CF, SA) as the class of risk.

1. Plot the data.

```
mydata <- data.frame(CF=c(5,4,1,2),SA=c(10,8,1,3),risk=c('A','A','B','B'))
attach(mydata)
plot(mydata[,1:2],col=risk)
```



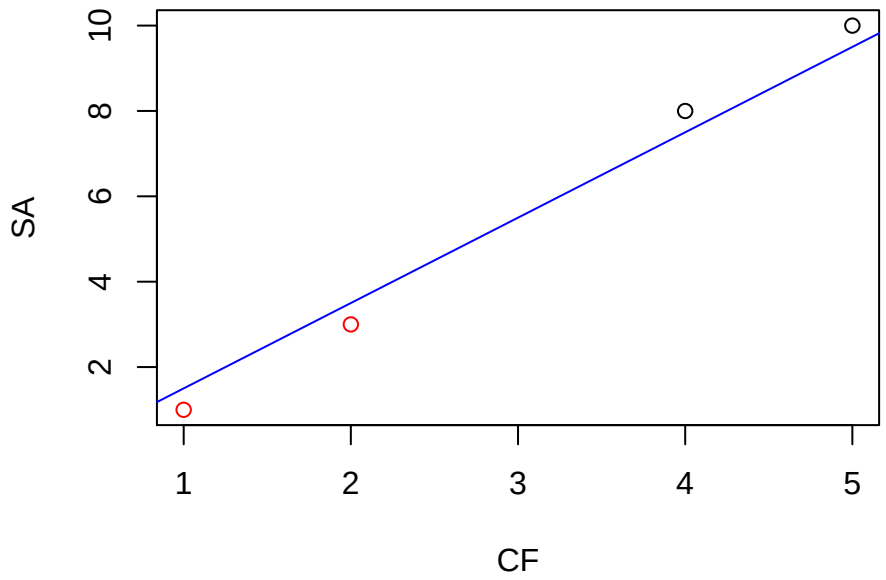
2. What is the logistic regression score?

```
LR <- glm(risk~CF+SA,data=mydata,family=binomial(lin='logit'))
summary(LR)

##
## Call:
## glm(formula = risk ~ CF + SA, family = binomial(lin = "logit"),
##     data = mydata)
##
## Deviance Residuals:
##          1           2           3           4
## -1.079e-05  -1.079e-05   1.079e-05   1.079e-05
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -23.57  361966.64       0        1
## CF              94.26  578492.10       0        1
```

```
## SA             -47.13  251280.91       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.5452e+00  on 3  degrees of freedom
## Residual deviance: 4.6610e-10  on 1  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 22
```

3. Plot the rule associated to the logistic score.

```
plot(mydata[,1:2],col=risk)
abline(-23.57/47.13,94.26/47.13,col='blue')
```



For the company called EcoNet, one observes: $CF = 4$ and $SA = 7.4$.

4. To which class of risk would you affect EcoNet?

```
score <- predict(LR,new=data.frame(CF=4,SA=7.4))
score

##        1
## 4.713214
```

So, which is the class of EcoNet ?

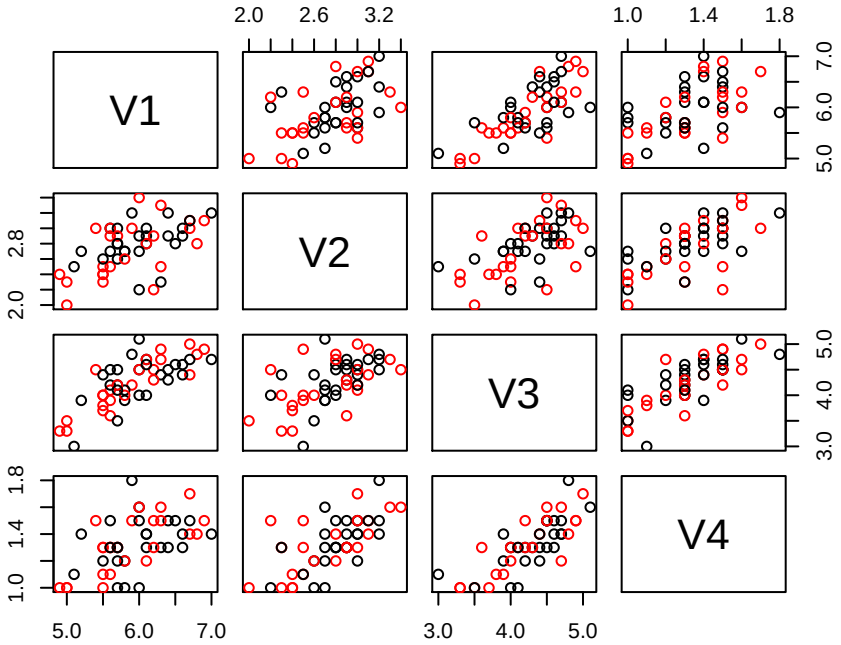
```
predict(LR,new=mydata[,c(1,2)])

##         1         2         3         4
## -23.56607 -23.56607  23.56607  23.56607
```

Logistic Regression on real data

We consider two samples: the first one is called `corpo_train`, it countains 25 firms described by four financial descriptors and the class of risk; the second one is called `corpo_test` it countains also 25 firms described by the same descriptors. Both datasets can be loaded thanks to:
`train <-read.table('http://alexandrelourme.free.fr/corpo_train',dec='.',sep=',',header=TRUE)`
`test <-read.table('http://alexandrelourme.free.fr/corpo_test',dec='.',sep=',',header=TRUE).`

5. Plot the train data.

```
train <- read.table('http://alexandrelourme.free.fr/corpo_train',dec='.',sep=',',header=TRUE)
plot(train[,1:4],col=risk)
```



6. What is the logistic regression score?

```
LR <- glm(risk~V1+V2+V3+V4,data=train,family=binomial(lin='logit'))
summary(LR)

##
## Call:
## glm(formula = risk ~ V1 + V2 + V3 + V4, family = binomial(lin = "logit"),
##     data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.676  -1.193   0.833   1.070   1.474
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.5379     3.5110   0.438    0.661
## V1           -0.7656     0.9080  -0.843    0.399
## V2            1.2875     1.3449   0.957    0.338
## V3            0.9541     1.3229   0.721    0.471
## V4           -3.3621     2.8507  -1.179    0.238
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.994  on 49  degrees of freedom
## Residual deviance: 66.819  on 45  degrees of freedom
## AIC: 76.819
##
## Number of Fisher Scoring iterations: 4
```

7. To which class would be allocated the firm: $V_1 = 6.9$, $V_2 = 4.2$, $V_3 = 4.5$, $V_4 = 2.1$?

```
predict(LR,new=data.frame(V1=6.9,V2=4.2,V3=4.5,V4=2.1))

##         1
## -1.104249
```

8. Estimate the scores of the firms into the test sample?

```
test <- read.table('http://alexandrelourme.free.fr/corpo_test',dec='.',sep=',',header=TRUE)
score <- predict(LR,new=test[,1:4])
```

9. Compare the estimated class and the true class of the firms within the sample test.

```
estimated_class <- (score>0)
true_class <- (test[,5])
table(true_class,estimated_class)

##           estimated_class
## true_class FALSE TRUE
##          n    24    0
##          y    23    3
```

10. Compute and interpret: (i) the misclassification error rate (ii) the false positive rate.
