# Analyse statistique avec R

*Arthur Tenenhaus*

*Monday, September 07, 2015*

## La R?gression multiple

La r?gression mutiple est une m?thode statistique adapt?e ? l'?tude de la liaison entre une variable quantitative $Y$ et un ensemble de $p$ variables explicatives $X_1, X_2, \ldots, X_p$ quantitatives ou qualitatives. L'exemple compagnon de cette s?ance, **pr?vision du prix d'une automobile**, servira ? illustrer cette m?thode.

Les commandes suivantes permettent de charger et mettre en forme le jeu de donn?es AUTO.

```
library(gdata)
library(pheatmap)
A = read.table("AUTO.csv", header=TRUE, sep="\t")
rownames(A) = A[, 1]
A = A[, -1]
head(A)
```

```
##                    CYL PUI LON LAR POIDS VITESSE  PRIX
## ALFASUD-TI-1350   1350  79 393 161   870     165 30579
## AUDI-100-L        1588  85 468 177  1110     160 39990
## SIMCA-1307-GLS    1294  68 424 168  1050     152 29600
## CITROEN-GS-CLUB   1222  59 412 161   930     151 28250
## FIAT-132-1600GLS 1585  98 439 164  1105     165 34900
## LANCIA-BETA-1300 1297  82 429 169  1080     160 35480
```

On se propose de construire un mod?le de pr?diction du prix d'une automobile ? partir des 6 variables caract?ristiques suivantes.

```
colnames(A)[-7]
```

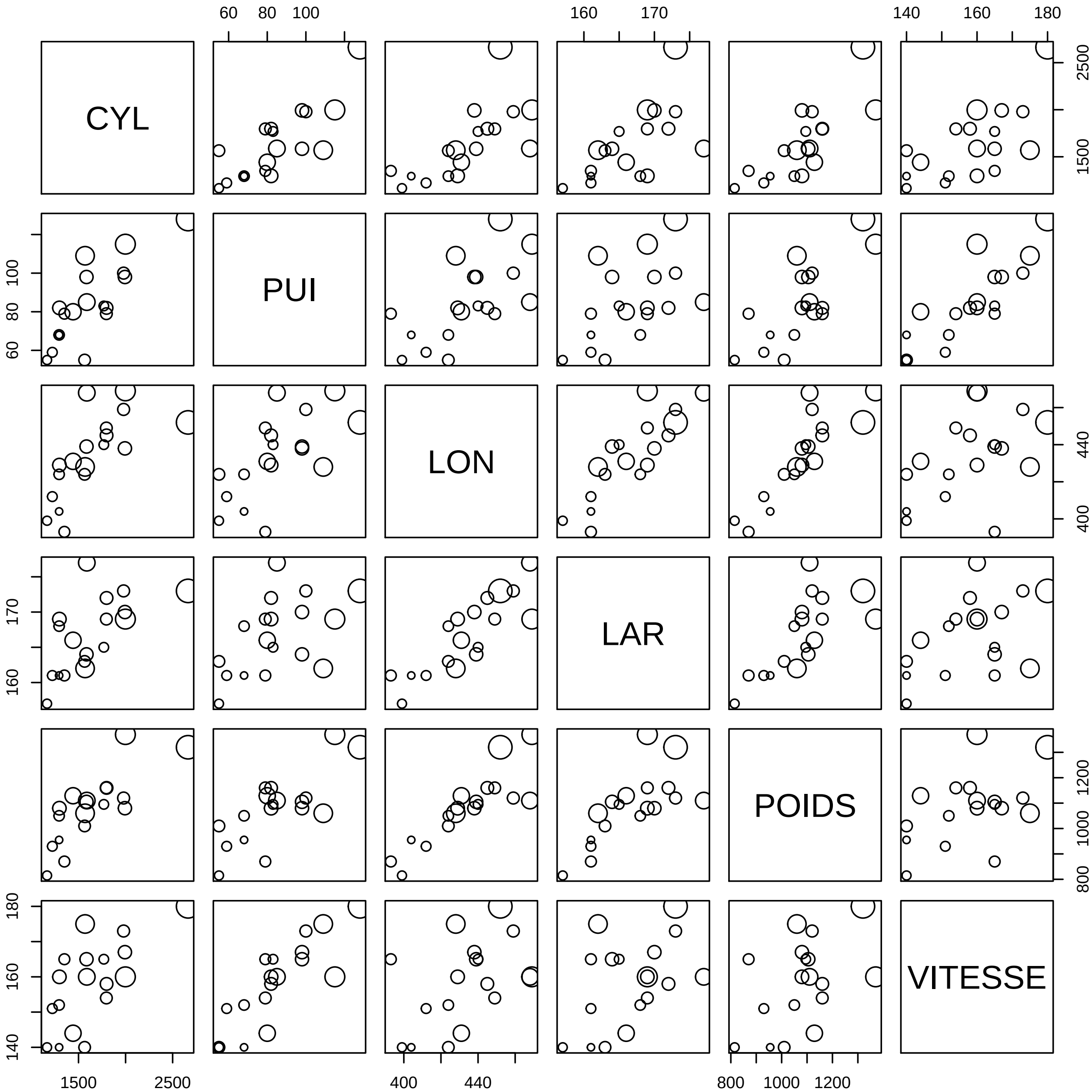
```
## [1] "CYL"     "PUI"     "LON"     "LAR"     "POIDS"   "VITESSE"
```

Ces variables ont ?t? mesur?es sur 18 automobiles.

### Analyse exploratoire des donn?es

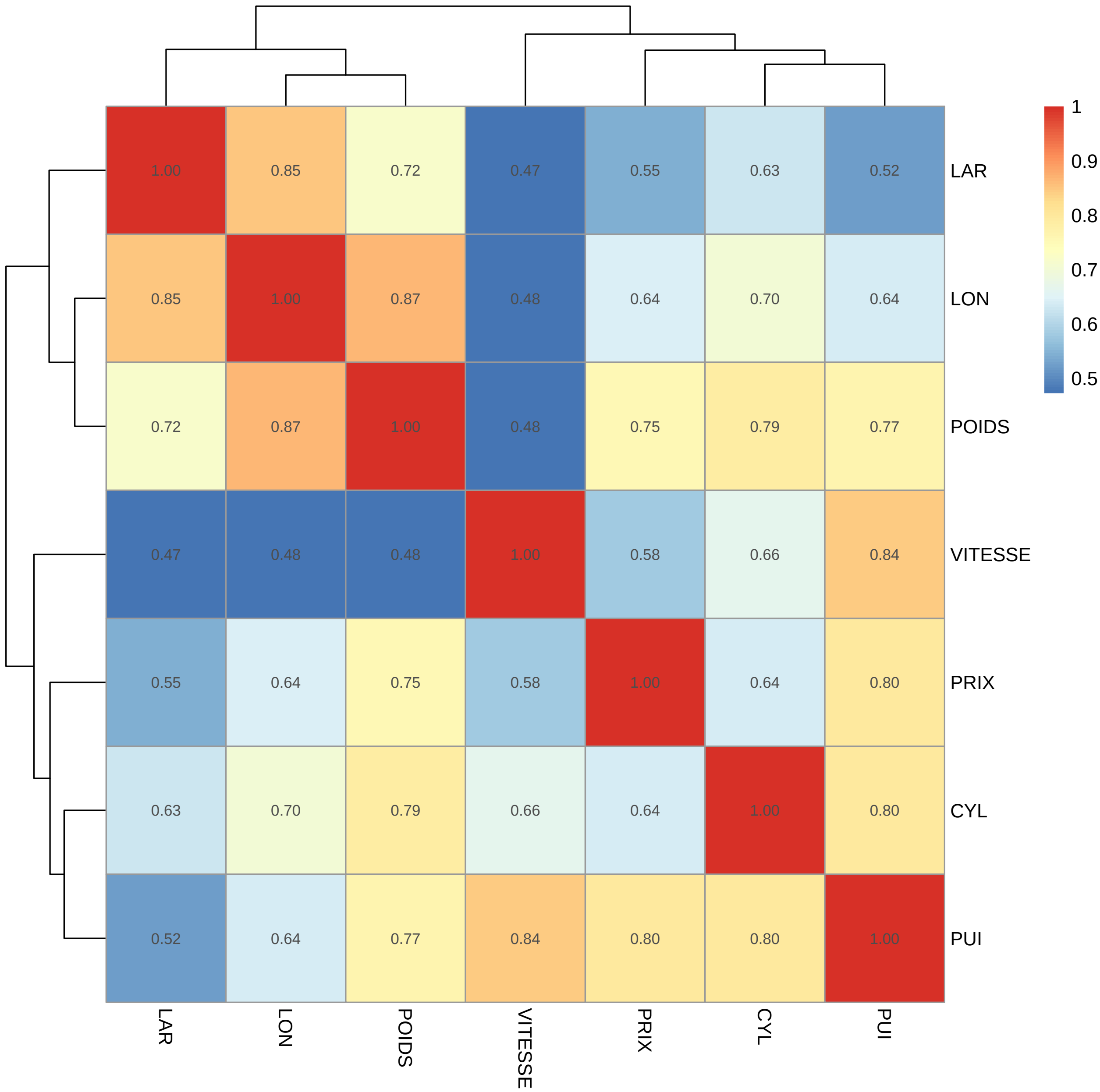
Toute bonne mod?lisation doit ?tre pr?c?d?e d'une ?tape d'analyse exploratoire des donn?es. L'objectif de cette analyse exploratoire est de mat?rialiser au travers de figures de m?rite et d'indicateurs le contenu des donn?es. La figure suivante renvoie l'ensemble des graphes bivari?s, la taille des points refl?tant la valeur de la variable ? expliquer.

```
pairs(A[, -7], cex = exp(A$PRIX/min(A$PRIX)-1), )
```




La figure suivante permet de visualiser la structure de corr?lation et les relations entre variables.

```
pheatmap(cor(A), display_numbers = TRUE)
```




On constate que les variables CYL, PUI et VITESSE sont corr?l?es entre elles tout comme LAR, LON et POIDS.

Cette structure de corr?lation entre variables est ?galement exhib?? gr?ce ? l'analyse en composante principale (ACP). La commande suivante permet de r?aliser une ACP

```
res.pca = princomp(A, cor = TRUE)
summary(res.pca)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
## Standard deviation     2.2516392 0.9312795 0.67398876 0.56826510
## Proportion of Variance 0.7242684 0.1238973 0.06489441 0.04613217
```

```
## Cumulative Proportion  0.7242684 0.8481658 0.91306016 0.95919234
##                            Comp.5     Comp.6      Comp.7
## Standard deviation     0.40365770 0.28547690 0.203019773
## Proportion of Variance 0.02327708 0.01164244 0.005888147
## Cumulative Proportion  0.98246942 0.99411185 1.000000000
```

La variance de la composante s'obtient ?galement par la commande suivante

```
apply(res.pca$scores, 2, function(x) sd(x)*sqrt((NROW(A)-1)/(NROW(A))))
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 2.2516392 0.9312795 0.6739888 0.5682651 0.4036577 0.2854769 0.2030198
```

La variance cumul?e s'obtient comme suit:

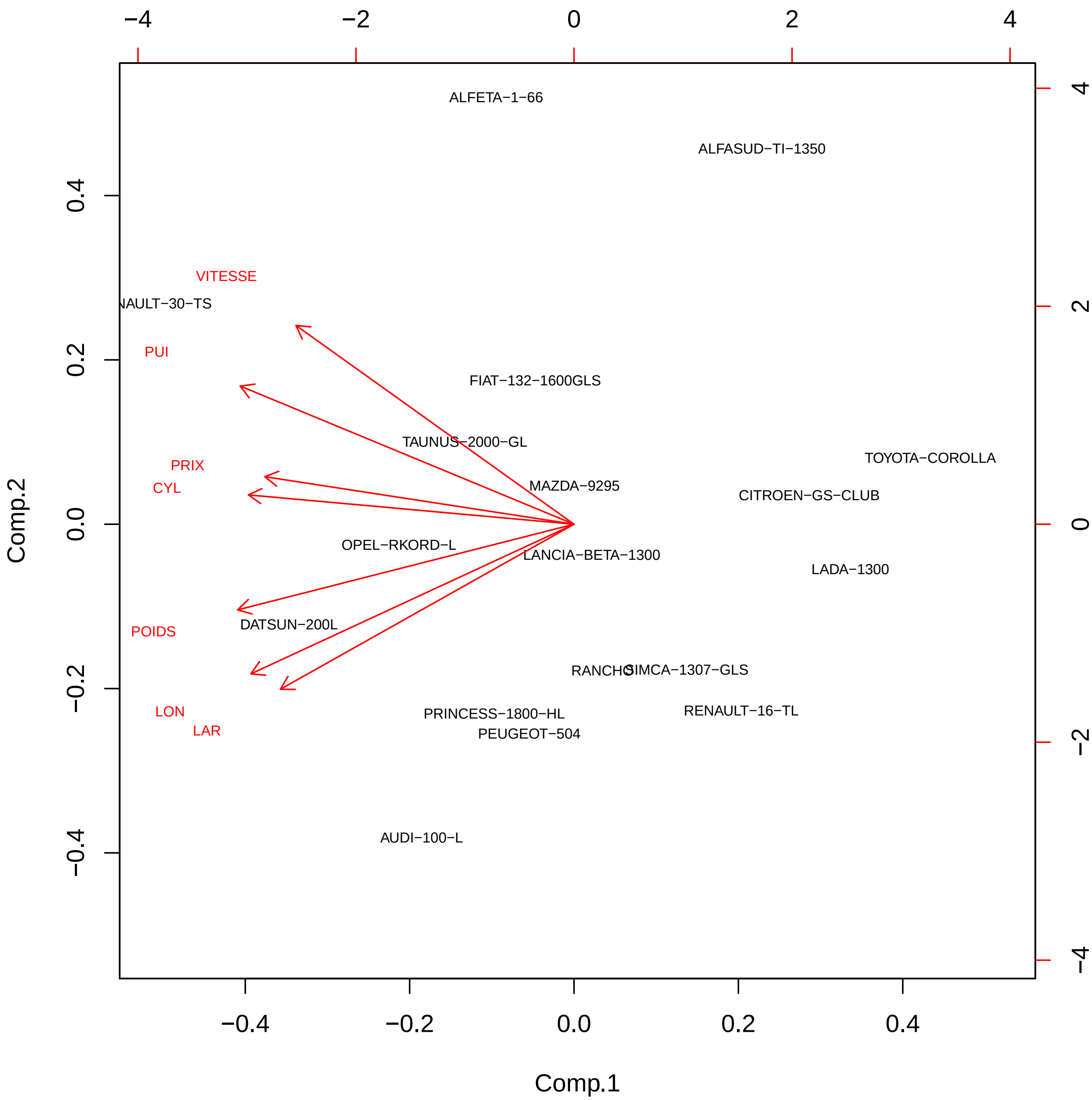
```
variance = apply(res.pca$scores, 2, function(x) var(x)*(NROW(A)-1)/(NROW(A)))
cumsum(variance)/sum(variance)
```

```
##    Comp.1    Comp.2    Comp.3    Comp.4    Comp.5    Comp.6    Comp.7
## 0.7242684 0.8481658 0.9130602 0.9591923 0.9824694 0.9941119 1.0000000
```

On constate que les deux premi?res composantes capturent pr?s de 85% de l'information pr?sente dans les donn?es. On en d?duit qu'une repr?sentation des donn?es sur le premier plan principal fournit une bonne approximation des relations entre individus.

Le biplot associ? est repr?sent? ci-dessous

```
biplot(res.pca, cex = .6)
```




## Calcul manuel des coefficients de r?gression et via la fonction lm()

```
y = A$PRIX
X = cbind(1, as.matrix(A[, -7]))
colnames(X) = c("Intercept", "CYL", "PUI", "LON", "LAR", "POIDS", "VITESSE")
beta_hat = solve(t(X)%*%X)%*%t(X)%*%y
beta_hat
```

```
##                    [,1]
## Intercept -8234.624688
```

```
## CYL          -3.505583
## PUI         282.174039
## LON         -15.126858
## LAR         208.831802
## POIDS        12.574568
## VITESSE    -111.039281
```

```r
res.lm = lm(PRIX ~ ., data = A)
summary(res.lm)
```

```
##
## Call:
## lm(formula = PRIX ~ ., data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8290.9 -1721.2  -167.4  2912.2  5420.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8234.625  42720.145  -0.193    0.851
## CYL            -3.506      5.551  -0.632    0.541
## PUI           282.174    174.890   1.613    0.135
## LON           -15.127    129.753  -0.117    0.909
## LAR           208.832    412.064   0.507    0.622
## POIDS          12.575     24.623   0.511    0.620
## VITESSE      -111.039    222.266  -0.500    0.627
##
## Residual standard error: 4406 on 11 degrees of freedom
## Multiple R-squared:  0.7091, Adjusted R-squared:  0.5504
## F-statistic: 4.468 on 6 and 11 DF,  p-value: 0.01562
```

On constate que les deux approches conduisent aux m?mes coefficients de r?gression.

**Il faut en permanence conserver un esprit critique sur les mod?les g?n?r?s** : L'examen des coefficients de r?gression et leur niveau de significativit? nous conduit ? rejeter le mod?le. En effet, contrairement ? ce que nous laissait conclure l'analyse exploratoire des donn?es, aucune des variables n'est significative. De plus, les signes des coefficients de r?gression associ?s ? CYL, LON et VITESSE ne sont pas en coh?rence avec l'intuition. D'o? peut provenir le probl?me ?

Comme discut? ci-dessous, la matrice de corr?lation montre une forte corr?lation entre variables. Or, nous savons que $\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^t\mathbf{X}\right)^{-1}$, et donc la variance de l'estimateur des moindres carr?es peut exploser en pr?sence de fortes multicolin?arit?s entre variables. Il faut donc, parmi les paquets de variables corr?l?es, s?lectionner un repr?sentant de chaque. Pour ce faire, nous proposons l'utilisation d'une approche de type stepwise.

```r
step(res.lm, direction = "backward")
```

```
## Start:  AIC=307.2
## PRIX ~ CYL + PUI + LON + LAR + POIDS + VITESSE
##
##           Df Sum of Sq       RSS    AIC
## - LON      1    263897 213844971 305.23
```

```
## - VITESSE  1   4845952 218427026 305.61
## - LAR      1   4986927 218568000 305.62
## - POIDS    1   5063698 218644772 305.63
## - CYL      1   7744198 221325272 305.85
## <none>                 213581074 307.20
## - PUI      1  50544530 264125604 309.03
##
## Step:  AIC=305.23
## PRIX ~ CYL + PUI + LAR + POIDS + VITESSE
##
##           Df Sum of Sq       RSS    AIC
## - VITESSE  1   5092979 218937950 303.65
## - POIDS    1   5439140 219284111 303.68
## - LAR      1   5654618 219499589 303.70
## - CYL      1   7601072 221446042 303.86
## <none>                 213844971 305.23
## - PUI      1  51510943 265355914 307.11
##
## Step:  AIC=303.65
## PRIX ~ CYL + PUI + LAR + POIDS
##
##         Df Sum of Sq       RSS    AIC
## - LAR    1   2214253 221152204 301.83
## - CYL    1  10437927 229375877 302.49
## <none>               218937950 303.65
## - POIDS  1  26475045 245412996 303.70
## - PUI    1  94872221 313810171 308.13
##
## Step:  AIC=301.83
## PRIX ~ CYL + PUI + POIDS
##
##         Df Sum of Sq       RSS    AIC
## - CYL    1   8927204 230079407 300.54
## <none>               221152204 301.83
## - POIDS  1  44620350 265772554 303.14
## - PUI    1  92659210 313811414 306.13
##
## Step:  AIC=300.54
## PRIX ~ PUI + POIDS
##
##         Df Sum of Sq       RSS    AIC
## <none>               230079407 300.54
## - POIDS  1  35703512 265782919 301.14
## - PUI    1  87548745 317628152 304.35
```

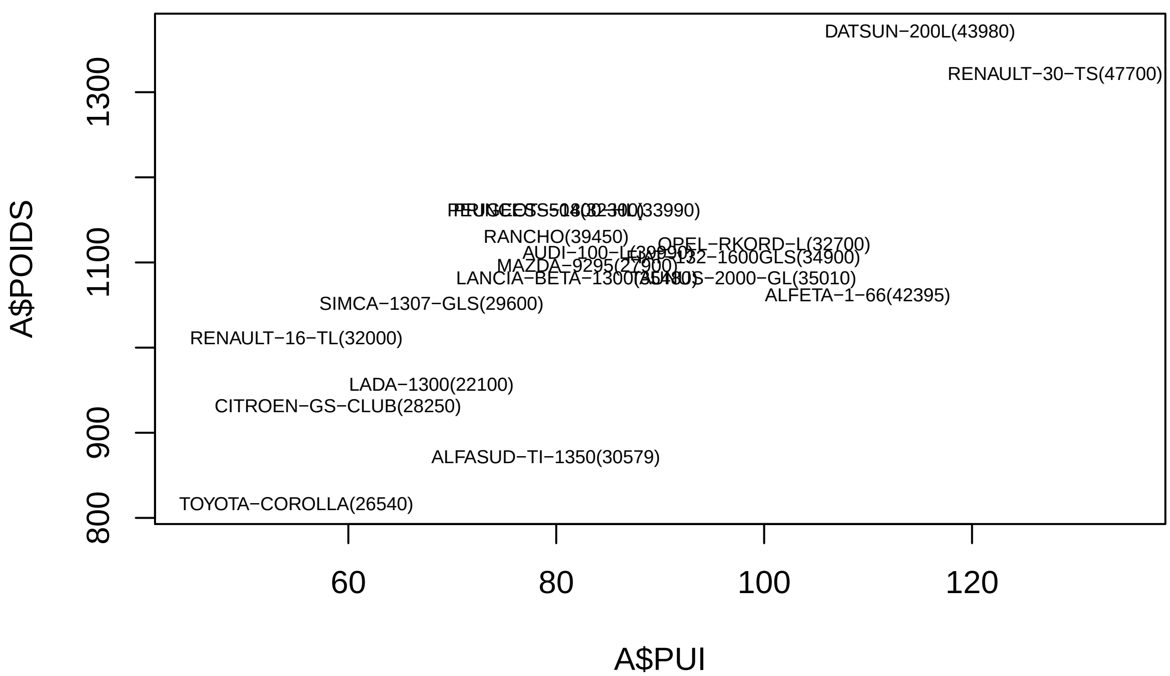
```
##
## Call:
## lm(formula = PRIX ~ PUI + POIDS, data = A)
##
## Coefficients:
## (Intercept)          PUI        POIDS
##     1784.60       173.02        16.44
```

Les variables PUI et POIDS sont conserv?es dans le mod?le final.

```r
res.final = lm(PRIX~PUI+POIDS, data = A)
summary(res.final)
```

```
##
## Call:
## lm(formula = PRIX ~ PUI + POIDS, data = A)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7149.3 -1875.6   300.1  2012.2  5264.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1784.60    8031.21   0.222   0.8271
## PUI           173.02      72.42   2.389   0.0305 *
## POIDS          16.44      10.77   1.526   0.1479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3916 on 15 degrees of freedom
## Multiple R-squared:  0.6866, Adjusted R-squared:  0.6448
## F-statistic: 16.43 on 2 and 15 DF,  p-value: 0.0001663
```

```r
plot(A$PUI, A$POIDS, col = "white", cex = A$PRIX, xlim = c(45, 135))
text(A$PUI, A$POIDS, paste(rownames(A), "(", A$PRIX, ")", sep = ""), cex = .6)
```




## La r?gression logistique

La r?gression logistique est une m?thode statistique adapt?e ? l'?tude de la liaison entre une variable qualitative $Y$ et un ensemble de $p$ variables explicatives $X_1, X_2, \ldots, X_p$ quantitatives ou qualitatives. L'exemple compagnon de cette s?ance, **pr?vision de la faillite d'entreprises**, servira ? illustrer cette m?thode.

### Pr?sentation des donn?es et notations

On se propose de construire un mod?le de pr?vision de la faillite d'une entreprise ? partir de donn?es financi?re r?colt?es par R.A. Johnson et D.W. Wichern en 1982. Ces donn?es financi?res annuelles ont ?t? recueillies sur 21 entreprises approximativement deux ans avant leur faillite, et, ? peu pr?s ? la m?me ?poque, sur 25 soci?t?s financi?rement solides. Les donn?es r?unissent, pour chaque entreprise, deux ratios financiers et leur situation deux ans plus tard :

| variable | Signification |
|---|---|
| $\mathbf{x}_1$ | Flux de tr?sorerie / Dette totale |
| $\mathbf{x}_2$ | Actif ? court terme / Dette ? court terme |
| $\mathbf{y}$ | Faillite (1), non faillite (0) |

### Chargement des donn?es

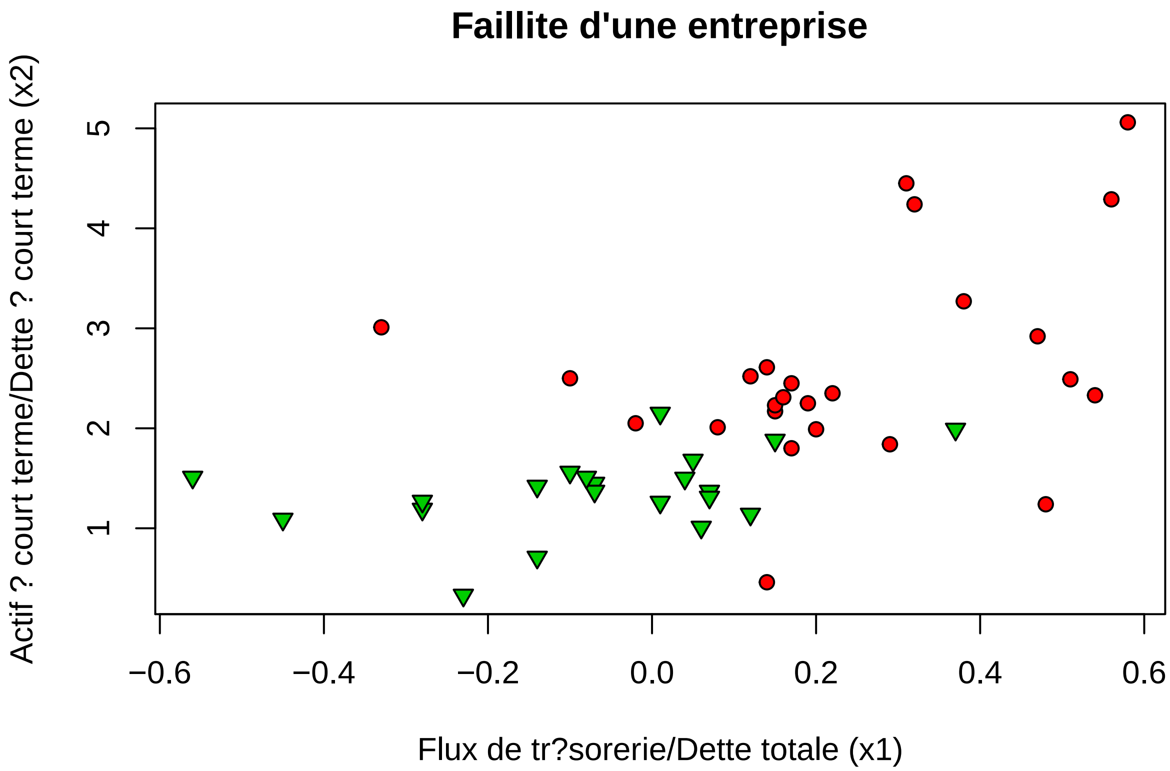
```
A = read.table("faillite.txt",
               header = TRUE)
X = as.matrix(A[, c(2,4)])
y = A[, 6]
head(cbind(X, y))
```

```
##         x1   x3 y
## [1,] -0.45 1.09 1
## [2,] -0.56 1.51 1
## [3,]  0.06 1.01 1
## [4,] -0.07 1.45 1
## [5,] -0.10 1.56 1
## [6,] -0.14 0.71 1
```

### Visualisation de donn?es

Toute bonne mod?lisation doit ?tre pr?c?d?e d'une ?tape d'analyse exploratoire des donn?es. L'objectif de cette analyse exploratoire est de mat?rialiser au travers de figures de m?rite et d'indicateurs le contenu des donn?es. La figure suivante renvoie le graphe bivari? $(\mathbf{x}_1, \mathbf{x}_2)$, la couleur/forme des points refl?tant la valeur de la variable ? expliquer.

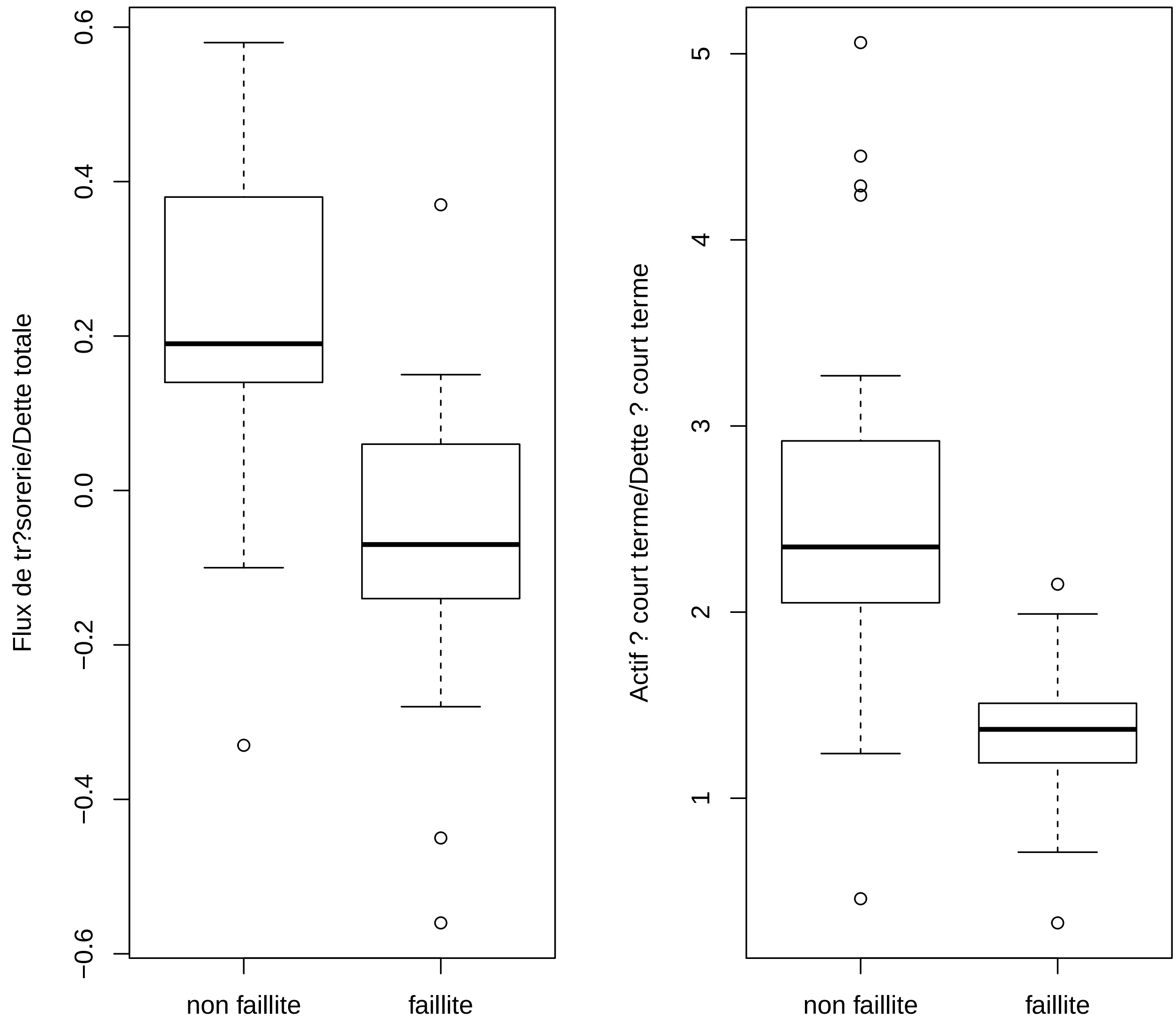
```
plot(X,
     bg = c("red", "green3")[as.factor(y)],
     pch = c(21, 25)[as.factor(y)],
     main = "Faillite d'une entreprise",
     xlab = "Flux de tr?sorerie/Dette totale (x1)",
     ylab = "Actif ? court terme/Dette ? court terme (x2)"
)
```




Il semble que ces deux variables soient porteuses d'information discriminante.

on s'int?resse maintenant aux deux variables s?par?ment. Les bo?tes ? moustaches des deux ratios financiers selon le crit?re de faillite sont pr?sent?s sur la figure suivante. La bo?te ? moustache permet de visualiser de mani?re tr?s compacte la dispersion des donn?es. La bo?te centrale est construite ? partir du premier et du troisi?me quartile et est partag?e par la m?diane. Les "moustaches" vont du premier quartile au minimum et du troisi?me quartile au maximum. Par convention, les moustaches ne doivent pas d?passer une fois et demi la distance interquartile. Si les points extr?mes sont trop loin des quartiles, ils appara?tront comme isol?s (outliers) sur le graphique.

```
layout(matrix(t(1:2), 1, 2))
boxplot(X[, 1]~factor(y,
                      labels = c("non faillite", "faillite")),
                      ylab = "Flux de tr?sorerie/Dette totale")
boxplot(X[, 2]~factor(y,
                      labels = c("non faillite", "faillite")),
                      ylab = "Actif ? court terme/Dette ? court terme")
```




## L'algorithme de Newton-Raphson pour la r?gression logistique

Les param?tres du mod?le sont estim?s par maximum de vraisemblance. L'algorithme utilis? pour trouver l'estimateur du maximum de vraisemblance est l'algorithme de Newton-Raphson.

Posons $\pi_i = P(Y = 1|X = \boldsymbol{x}_i)$, $\mathbf{X}$ la matrice form?e d'une premi?re colonne de coordonn?es constantes ?gales ? 1 et des 2 colonnes correspondant aux variables $\mathbf{x}_1$, $\mathbf{x}_2$ observ?es sur les $n$ individus et $\mathbf{V}$ la matrice diagonale form?e des $\pi_i(1-\pi_i)$. Enfin, notons $\boldsymbol{\pi}$ le vecteur de probabilit?s tel que le i?me ?l?ment ?gal $\pi_i$, l'?tape courante de l'algorithme de Newton-Raphson peut s'?crire comme suit :

$$\boldsymbol{\beta}^{(s+1)} = \boldsymbol{\beta}^{(s)} + \left(\mathbf{X}^t\mathbf{V}\mathbf{X}\right)^{-1}\mathbf{X}^t(\mathbf{y}-\boldsymbol{\pi}) \tag{1}$$

Le code ci-dessous impl?mente l'algorithme de Newton-Raphson pour la r?gression logistique

```
my_lr = function(X, y, tolerance = 1e-6, max.iter=200){
  X = cbind(1, X)
  beta_s = rep(0, NCOL(X))
  pi = runif(NROW(X), 0, 1)
```

```r
  V = diag(pi*(1-pi))
  iter = 1

  made.changes = TRUE

  while (made.changes & (iter < max.iter))
  {
    iter = iter + 1
    made.changes <- FALSE
    beta_s_plus_1 = beta_s + solve(t(X)%*%V%*%X)%*%t(X)%*%(y-pi)

    pi = drop(1/(1+exp(-X%*%beta_s_plus_1)))
    V = diag(pi*(1-pi))

    relative.change = drop(crossprod(beta_s_plus_1 - beta_s))/drop(crossprod(beta_s))
    made.changes = (relative.change > tolerance)

    beta_s = beta_s_plus_1

    if (iter == 200)
      warning("The Newton-Raphson algorithm did not converge after 200 iterations.")
  }

  return(list(beta = beta_s , proba = pi))
}
```

### Comparison between my_lr and glm

```r
res.mylr = my_lr(X, y)
res.mylr
```

```
## $beta
##          [,1]
##      5.940161
## x1 -6.556415
## x3 -3.019117
##
## $proba
##  [1] 9.963147e-01 9.936497e-01 9.239641e-01 8.830247e-01 8.682965e-01
##  [6] 9.911152e-01 7.593700e-01 9.057591e-01 7.933145e-01 9.289660e-01
## [11] 9.984248e-01 8.214450e-01 3.505767e-01 9.849795e-01 3.275600e-01
## [16] 7.629709e-02 8.705474e-01 6.318644e-01 8.880036e-01 8.470434e-01
## [21] 9.809542e-01 7.237522e-03 3.423639e-01 1.619867e-03 1.092357e-01
## [26] 1.285648e-04 7.282346e-05 7.908940e-02 4.705589e-01 6.932041e-02
## [31] 3.523238e-01 1.687117e-01 2.784847e-01 9.742559e-01 5.428809e-02
## [36] 1.448061e-01 1.107576e-01 1.799959e-01 9.614268e-03 2.721454e-01
## [41] 2.787595e-01 2.292076e-05 2.011431e-01 2.580531e-03 7.101070e-02
## [46] 1.966421e-06
```

La fonction **glm()** disponible dans le package **stats** impl?mente le mod?le lin?aire g?n?ralis?. La r?gression logistique binaire comme cas particulier du mod?le lin?aire g?n?ralis? est donc disponible via cette fonction.

```
res.glm = glm(y~X, family = binomial)
summary(res.glm)
```

```
##
## Call:
## glm(formula = y ~ X, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.70538  -0.48365  -0.00942   0.47678   2.26853
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.940      1.985   2.992  0.00277 **
## Xx1           -6.556      2.905  -2.257  0.02402 *
## Xx3           -3.019      1.002  -3.013  0.00259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 63.421  on 45  degrees of freedom
## Residual deviance: 28.636  on 43  degrees of freedom
## AIC: 34.636
##
## Number of Fisher Scoring iterations: 6
```

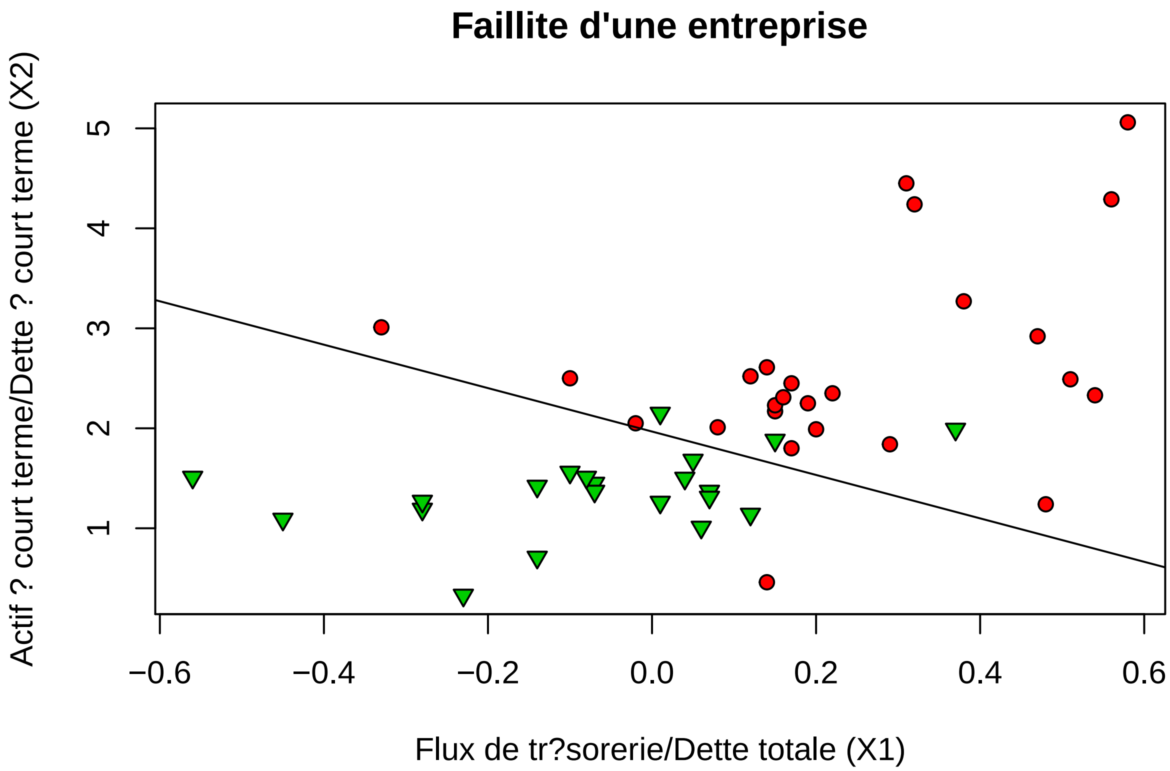
Reste alors ? comparer l'estimateur di maximum de vraisemblance issu de notre impl?mentation ? celui estim? par la fonction **glm()** ... verdict....

### Visualisation de la fronti?re de d?cision

La figure suivante repr?sente les entreprises dans le plan des variables $\mathbf{x}_1$ et $\mathbf{x}_2$ et s?parer les deux classes d'entreprise par la droite d'iso-probabilit? 0.5 d'?quation :

$$5.94 - 6.556X_1 - 3.019X_3 = 0 \iff X_3 = -\frac{6.556}{3.019}X_1 + \frac{5.94}{3.019}$$

```
plot(X,
     bg = c("red", "green3")[as.factor(y)],
     pch = c(21, 25)[as.factor(y)],
     main = "Faillite d'une entreprise",
     xlab = "Flux de tr?sorerie/Dette totale (X1)",
     ylab = "Actif ? court terme/Dette ? court terme (X2)"
)
b = res.mylr$beta[1]
a1 = res.mylr$beta[2]
a2 = res.mylr$beta[3]
abline(-b/a2, -a1/a2, col = "black")
```


Faillite d'une entreprise
Actif ? court terme/Dette ? court terme (X2)
1
2
3
4
5
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
Flux de tr?sorerie/Dette totale (X1)