# Rapport du TP3 de Régression Linéaire

*Rémy Gaudré Baptiste Boisson*

*7 juin 2019*

## Introduction :

Dans ce rapport de TP, nous allons étudier un exemple de séléction de modèles de régression linéaire avec R.

Le jeu de données est tiré de Cornell et concerne des proportions de sept composants sur l'indice d'octane moteur de douze différents mélanges d'essences.

X1 : Distillation directe (entre 0 et 0.21) X2 : Reformat (entre 0 et 0.62) X3 : Naphta de craquage thermique (entre 0 et 0.12) X4 : Naphta de craquage catalytique (entre 0 et 0.62) X5 : Polymère (entre 0 et 0.12) X6 : Alkylat (entre 0 et 0.74) X7 : Essence naturelle (entre 0 et 0.08) Y : Indice d'octane moteur

```
     X1   X2   X3   X4   X5   X6   X7    Y
1  0.00 0.23 0.00 0.00 0.00 0.74 0.03 98.7
2  0.00 0.10 0.00 0.00 0.12 0.74 0.04 97.8
3  0.00 0.00 0.00 0.10 0.12 0.74 0.04 96.6
4  0.00 0.49 0.00 0.00 0.12 0.37 0.02 92.0
5  0.00 0.00 0.00 0.62 0.12 0.18 0.08 86.6
6  0.00 0.62 0.00 0.00 0.00 0.37 0.01 91.2
7  0.17 0.27 0.10 0.38 0.00 0.00 0.08 81.9
8  0.17 0.19 0.10 0.38 0.02 0.06 0.08 83.1
9  0.17 0.21 0.10 0.38 0.00 0.06 0.08 82.4
10 0.17 0.15 0.10 0.38 0.02 0.10 0.08 83.2
11 0.21 0.36 0.12 0.25 0.00 0.00 0.06 81.4
12 0.00 0.00 0.00 0.55 0.00 0.37 0.08 88.1
```

## 1.

Réaliser les statistiques descriptives univariées et bivariées (y versus les autres variables)

Pour cela, nous créont les fonction respectives "statistiques_descriptives" et "statistiques_bivariees". La fonction statistiques_descriptives nous donne le minimum, Q1(Xi), Q2(Xi), E(Xi), Q3(Xi), Max(Xi). La fonction statistiques_bivariees nous donne $\rho$(Y,Xi), et Cov(Y,Xi).

```
statistiques_bivariees = function (x) {
  return (c(cor(donnees$Y,x,use="pairwise.complete.obs"),cov(donnees$Y,x,use="pairwise.complete.obs")))
}

statistiques_descriptives = function (x) {
  return (summary(x))
}
```

Les statistiques descriptives univariées des Xi sont données ci-dessous.

```
       X1                X2               X3                X4
 Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.0750   1st Qu.:0.00000   1st Qu.:0.0000
 Median :0.00000   Median :0.2000   Median :0.00000   Median :0.3150
```

```
 Mean   :0.07417   Mean   :0.2183   Mean   :0.04333   Mean   :0.2533
 3rd Qu.:0.17000   3rd Qu.:0.2925   3rd Qu.:0.10000   3rd Qu.:0.3800
 Max.   :0.21000   Max.   :0.6200   Max.   :0.12000   Max.   :0.6200
       X5                X6                X7                Y
 Min.   :0.00000   Min.   :0.0000   Min.   :0.01000   Min.   :81.40
 1st Qu.:0.00000   1st Qu.:0.0600   1st Qu.:0.03750   1st Qu.:82.92
 Median :0.01000   Median :0.2750   Median :0.07000   Median :87.35
 Mean   :0.04333   Mean   :0.3108   Mean   :0.05667   Mean   :88.58
 3rd Qu.:0.12000   3rd Qu.:0.4625   3rd Qu.:0.08000   3rd Qu.:93.15
 Max.   :0.12000   Max.   :0.7400   Max.   :0.08000   Max.   :98.70
```

On s'intéresse maintenant aux statistiques bivariées de Y par rapport aux Xi comme indiqué ci-dessous :

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | Y |
|---|---|---|---|---|---|---|---|
| -0.8372958 | -0.07081888 | -0.8379578 | -0.7067135 | 0.4937991 | 0.9850704 | -0.7411162 | 1.00000 |
| -0.5039242 | -0.09030303 | -0.2941212 | -1.0462121 | 0.1838788 | 1.8815606 | -0.1308788 | 42.52697 |

## 2.

Réaliser le modèle de régression linéaire entre y et toutes les autres variables (fonction R : lm). Que constatez vous ?

```
Call:
lm(formula = Y ~ ., data = donnees)

Residuals:
         1          2          3          4          5          6 
 1.207e+00 -2.218e-01 -5.475e-01 -8.195e-02  8.105e-01 -3.527e-01 
         7          8          9         10         11         12 
 5.906e-02  3.598e-01 -3.036e-01 -1.153e-01  1.055e-15 -8.141e-01 

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   116.92      81.10   1.442    0.209
X1            -82.60     173.22  -0.477    0.654
X2            -31.00      80.84  -0.383    0.717
X3             24.33     431.97   0.056    0.957
X4            -39.74      90.25  -0.440    0.678
X5            -29.17      84.02  -0.347    0.743
X6            -16.62      84.45  -0.197    0.852
X7                NA         NA      NA       NA

Residual standard error: 0.8362 on 5 degrees of freedom
Multiple R-squared:  0.9925,  Adjusted R-squared:  0.9836 
F-statistic: 110.7 on 6 and 5 DF,  p-value: 3.762e-05
```

On s'apperçoit déjà qu'il y a un problème de singularité. Cela signifie qu'il y a probablement 2 variables parfaitement colinéaire. Le $R^2$ est très bon ($R^2 = 99.25\%$), il explique donc une grande partie de l'information. Et selon la p-value, le modèle est significatif.

## 3.

Puisque $n = 12 > p = 7$, il ne reste qu'à vérifier qu'il n'y a pas une relation entre les variables explicatives (multi-collinéarité). En effet, les variables X's représentent les taux de chaque composante dans l'essence. Du coup, la somme sur ligne doit faire 100%. Vérifier (fonction apply). Donc on n'a pas besoin de toutes les 7 variables puisque 6 suffisent ! On calculera aussi le déterminant de la matrice XTX (voir cours). Utiliser la foction det en R.

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1
```

```
[1] 2.510764e-12
```

La somme des lignes est bien égale à 100%. Le déterminant de la matrice est très proche de 0, il y a donc bien un problème de colinéarité entre les variables.

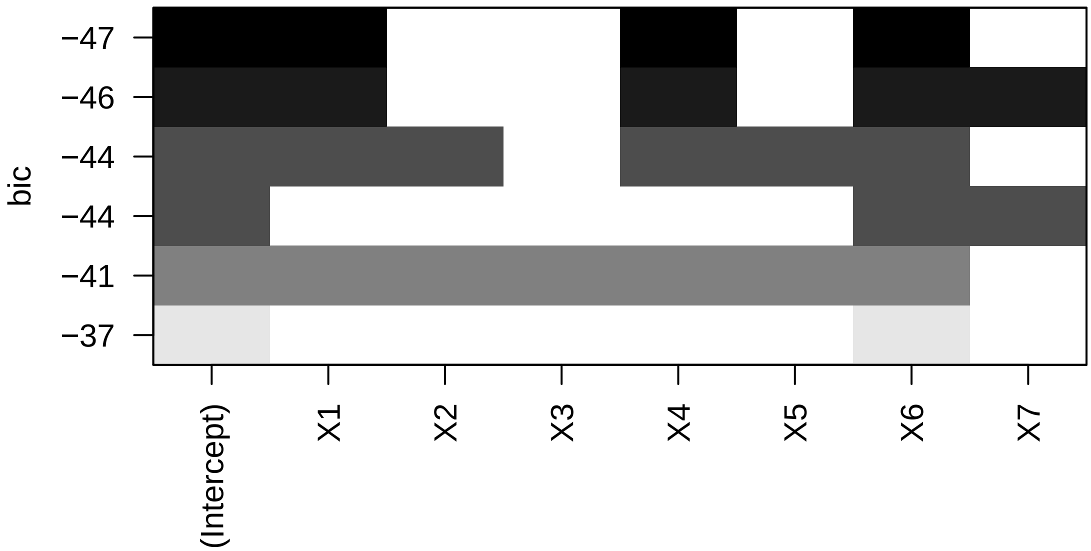
## 4.

One ne peut pas donc faire un modèle avec toutes les variables. Mais lesquelles éliminer ? On procédera à une sélection des variables. Explorez la fonction regsubsets du package « leaps ».

```
Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
force.in = force.in, : 1 linear dependencies found
```

```
Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
force.in = force.in, : nvmax reduced to 6
```

```
Subset selection object
Call: regsubsets.formula(Y ~ ., int = T, nbest = 1, nvmax = 7, method = "exh",
    data = donnees)
7 Variables  (and intercept)
   Forced in Forced out
X1     FALSE      FALSE
X2     FALSE      FALSE
X3     FALSE      FALSE
X4     FALSE      FALSE
X5     FALSE      FALSE
X6     FALSE      FALSE
X7     FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         X1  X2  X3  X4  X5  X6  X7 
1  ( 1 ) " " " " " " " " " " "*" " "
2  ( 1 ) " " " " " " " " " " "*" "*"
3  ( 1 ) "*" " " " " "*" " " "*" " "
4  ( 1 ) "*" " " " " "*" " " "*" "*"
5  ( 1 ) "*" "*" " " "*" "*" "*" " "
6  ( 1 ) "*" "*" "*" "*" "*" "*" " "
```




### 4.a.

Quel est le meilleur modèle ? Combien de variables fait-il rentrer dans la régression ? Estimer ce modèle, analyser la validité et les performances du modèle complet (R2, significativité coefficients). Selon le critère BIC, le meilleur modèle expliquant Y est $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_6$.

```
Call:
lm(formula = Y ~ X1 + X4 + X6, data = donnees)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00154 -0.41198  0.02205  0.29286  1.00148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  85.9435     0.9964  86.255 3.64e-13 ***
X1          -14.0924     4.1175  -3.423  0.00905 **
X4           -4.9445     1.3018  -3.798  0.00525 **
X6           15.8852     1.5779  10.067 8.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.707 on 8 degrees of freedom
Multiple R-squared:  0.9915,  Adjusted R-squared:  0.9882
F-statistic: 309.3 on 3 and 8 DF,  p-value: 1.31e-08
```

Le modèle est le suivant : $Y = 85.9435 - 14.0924X_1 - 4.9445X_4 + 15.8852X_6$ Selon la p-value du modèle (p-value = 1.31e-08), le modèle est significatif. Le $R^2 = 99.15\%$, donc le modèle explique 99.15% de l'information de Y.

### 4.b.

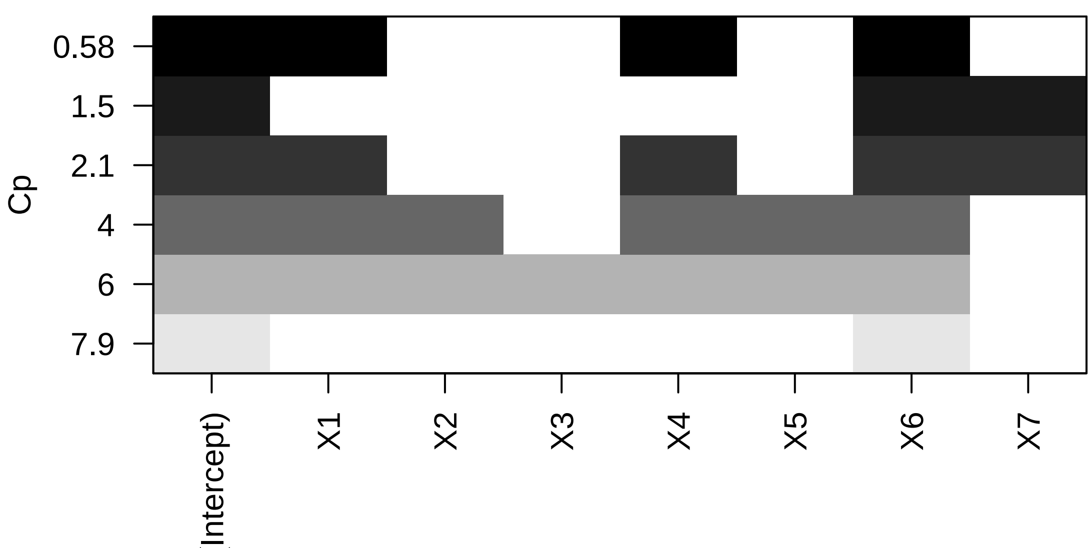
Le meilleur modèle est le modèle avec X6 et X7.

```
Call:
lm(formula = Y ~ X6 + X7, data = donnees)

Residuals:
     Min       1Q   Median       3Q      Max
-1.02075 -0.47918 -0.07922  0.42667  1.50196

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   84.788      1.017  83.388  2.6e-14 ***
X6            19.504      1.161  16.801  4.2e-08 ***
X7           -40.006     12.556  -3.186   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8508 on 9 degrees of freedom
Multiple R-squared:  0.9861,  Adjusted R-squared:  0.983
F-statistic: 318.6 on 2 and 9 DF,  p-value: 4.44e-09
```

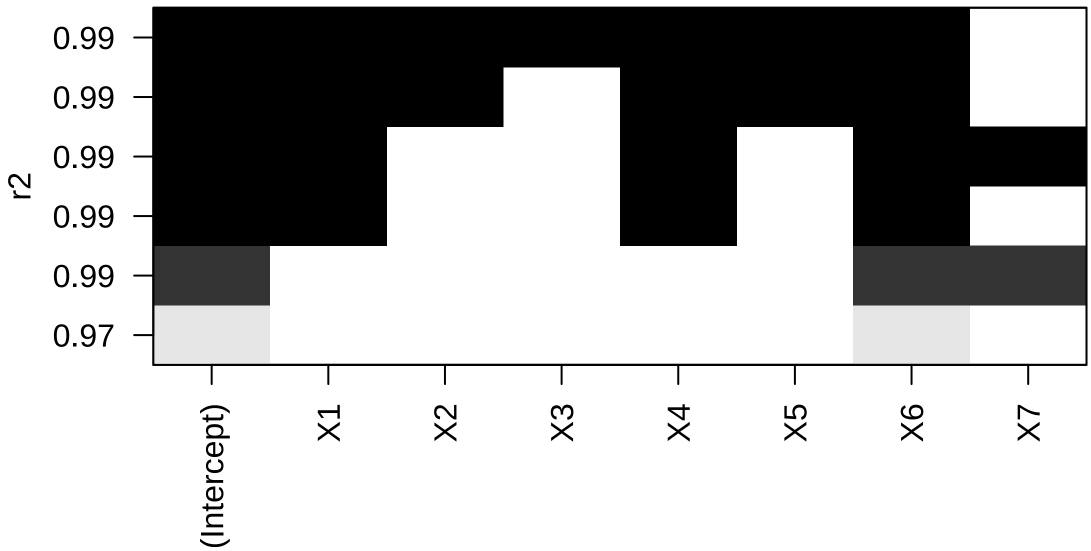
Le meilleur modèle à 2 variables est le suivant : $Y = 84.788 + 19.504X_6 - 40.006X_7$ Selon la p-value du modèle (p-value = 4.44e-09), le modèle est significatif. Le $R^2 = 0.98.61\%$, donc le modèle explique 98.61% de l'information de Y.

## 5

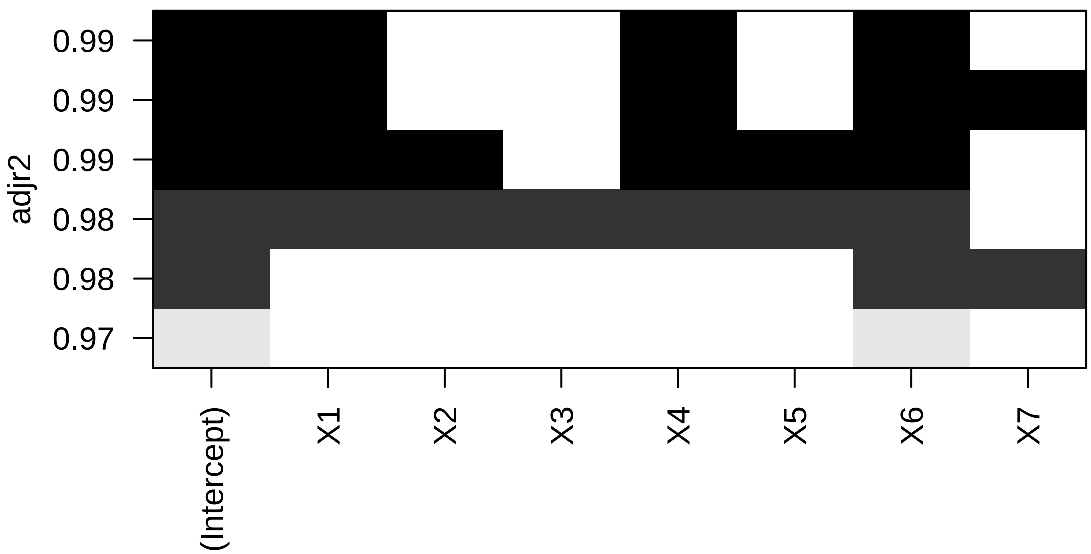
Remplacer précédemment le critère BIC par le critère Cp, R2 ajusté (adjr2) ou encore R2. (évidemment AIC donne le mêmes résultats que BIC à cause du lien entre les deux critères). Préciser pour chaque critère le meilleur modèle.




Selon le critère Cp, le meilleur modèle est le suivant : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_6$.




Selon le critère du $R^2$, le meilleur modèle est le suivant : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$




Selon le critère du $R^2_{ajusté}$, le meilleur modèle est le suivant : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_6$.

## 6.

Les recherches précédentes étaient exhaustives. Cela pose un problème lorsque le nombre de variables est grand. Faisons une sélection de variables pas-à-pas.

```
Start:  AIC=-0.8
donnees$Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7


Step:  AIC=-0.8
donnees$Y ~ X1 + X2 + X3 + X4 + X5 + X6

       Df Sum of Sq    RSS      AIC
```

```
- X3    1  0.002218 3.4983 -2.79158
- X6    1  0.027082 3.5232 -2.70659
- X5    1  0.084293 3.5804 -2.51329
- X2    1  0.102814 3.5989 -2.45138
- X4    1  0.135593 3.6317 -2.34258
- X1    1  0.159001 3.6551 -2.26548
<none>              3.4961 -0.79919

Step:  AIC=-2.79
donnees$Y ~ X1 + X2 + X4 + X5 + X6

       Df Sum of Sq    RSS     AIC
- X6    1   0.08997 3.5883 -4.4869
- X5    1   0.23972 3.7380 -3.9962
- X2    1   0.28702 3.7853 -3.8454
- X4    1   0.36541 3.8637 -3.5994
- X1    1   0.42297 3.9213 -3.4219
<none>              3.4983 -2.7916

Step:  AIC=-4.49
donnees$Y ~ X1 + X2 + X4 + X5

       Df Sum of Sq     RSS    AIC
<none>                3.588 -4.487
- X5    1     3.533   7.121  1.738
- X2    1    50.419  54.008 26.051
- X1    1    92.385  95.973 32.950
- X4    1   135.027 138.615 37.362
Start:  AIC=45.96
donnees$Y ~ 1

       Df Sum of Sq    RSS    AIC
+ X6    1    453.93  13.86  5.732
+ X3    1    328.47 139.32 33.423
+ X1    1    327.96 139.84 33.467
+ X7    1    256.94 210.86 38.395
+ X4    1    233.64 234.16 39.653
+ X5    1    114.07 353.73 44.604
<none>              467.80 45.958
+ X2    1      2.35 465.45 47.897

Step:  AIC=5.73
donnees$Y ~ X6

       Df Sum of Sq     RSS     AIC
+ X7    1    7.3485  6.5152 -1.3292
+ X2    1    6.7120  7.1517 -0.2106
+ X4    1    4.0095  9.8542  3.6359
+ X3    1    2.6671 11.1967  5.1685
+ X1    1    2.6534 11.2104  5.1832
<none>              13.8638  5.7325
+ X5    1    0.8501 13.0136  6.9731
```

```
Step:  AIC=-1.33
donnees$Y ~ X6 + X7

       Df Sum of Sq    RSS      AIC
+ X1    1   1.42728 5.0880 -2.29636
+ X3    1   1.38121 5.1340 -2.18821
+ X5    1   1.08694 5.4283 -1.51938
<none>              6.5152 -1.32916
+ X4    1   0.36245 6.1528 -0.01603
+ X2    1   0.01238 6.5029  0.64802

Step:  AIC=-2.3
donnees$Y ~ X6 + X7 + X1

       Df Sum of Sq    RSS     AIC
+ X4    1   1.52705 3.5609 -4.5787
<none>              5.0880 -2.2964
+ X3    1   0.49613 4.5918 -1.5275
+ X5    1   0.44979 4.6382 -1.4071
+ X2    1   0.06527 5.0227 -0.4513

Step:  AIC=-4.58
donnees$Y ~ X6 + X7 + X1 + X4

       Df Sum of Sq    RSS     AIC
<none>              3.5609 -4.5787
+ X2    1  0.050479 3.5104 -2.7500
+ X5    1  0.049644 3.5113 -2.7472
+ X3    1  0.001742 3.5592 -2.5846
Start:  AIC=45.96
donnees$Y ~ 1

       Df Sum of Sq    RSS    AIC
+ X6    1    453.93  13.86  5.732
+ X3    1    328.47 139.32 33.423
+ X1    1    327.96 139.84 33.467
+ X7    1    256.94 210.86 38.395
+ X4    1    233.64 234.16 39.653
+ X5    1    114.07 353.73 44.604
<none>              467.80 45.958
+ X2    1      2.35 465.45 47.897

Step:  AIC=5.73
donnees$Y ~ X6

       Df Sum of Sq    RSS    AIC
+ X7    1      7.35   6.52 -1.329
+ X2    1      6.71   7.15 -0.211
+ X4    1      4.01   9.85  3.636
+ X3    1      2.67  11.20  5.169
+ X1    1      2.65  11.21  5.183
<none>              13.86  5.732
+ X5    1      0.85  13.01  6.973
```

```
- X6    1    453.93 467.80 45.958

Step:  AIC=-1.33
donnees$Y ~ X6 + X7

       Df Sum of Sq     RSS    AIC
+ X1    1     1.427   5.088 -2.296
+ X3    1     1.381   5.134 -2.188
+ X5    1     1.087   5.428 -1.519
<none>                6.515 -1.329
+ X4    1     0.362   6.153 -0.016
+ X2    1     0.012   6.503  0.648
- X7    1     7.349  13.864  5.732
- X6    1   204.343 210.858 38.395

Step:  AIC=-2.3
donnees$Y ~ X6 + X7 + X1

       Df Sum of Sq    RSS     AIC
+ X4    1     1.527  3.561 -4.5787
<none>               5.088 -2.2964
+ X3    1     0.496  4.592 -1.5275
+ X5    1     0.450  4.638 -1.4071
- X1    1     1.427  6.515 -1.3292
+ X2    1     0.065  5.023 -0.4513
- X7    1     6.122 11.210  5.1832
- X6    1    93.651 98.739 31.2909

Step:  AIC=-4.58
donnees$Y ~ X6 + X7 + X1 + X4

       Df Sum of Sq     RSS     AIC
- X7    1    0.4379  3.9988 -5.1868
<none>               3.5609 -4.5787
+ X2    1    0.0505  3.5104 -2.7500
+ X5    1    0.0496  3.5113 -2.7472
+ X3    1    0.0017  3.5592 -2.5846
- X4    1    1.5271  5.0880 -2.2964
- X1    1    2.5919  6.1528 -0.0160
- X6    1   12.5943 16.1552 11.5680

Step:  AIC=-5.19
donnees$Y ~ X6 + X1 + X4

       Df Sum of Sq    RSS     AIC
<none>               3.999 -5.1868
+ X7    1     0.438  3.561 -4.5787
+ X2    1     0.261  3.738 -3.9962
+ X5    1     0.214  3.785 -3.8454
+ X3    1     0.177  3.821 -3.7316
- X1    1     5.855  9.854  3.6359
- X4    1     7.212 11.210  5.1832
- X6    1    50.660 54.659 24.1945
```

```
[1] 1.287098
[1] 1.012127
[1] 1.017572
```

Selon le critère du PRESS, le meilleur modèle est obtenu avec la méthode “forward” : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_6 + \beta_4 X_7$.