BE 1 - Régression Linéaire

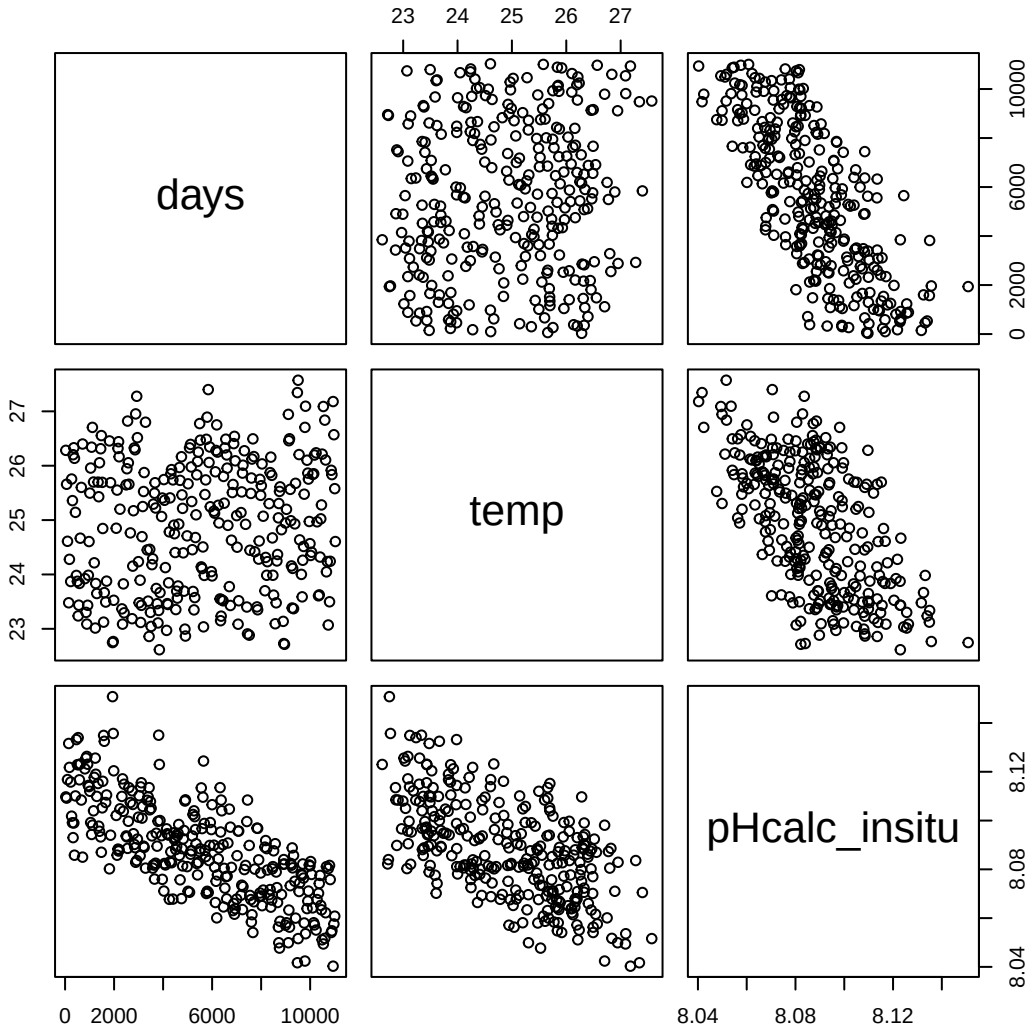
Exercice 1 - Acidification de l'océan

0) Chargement des données

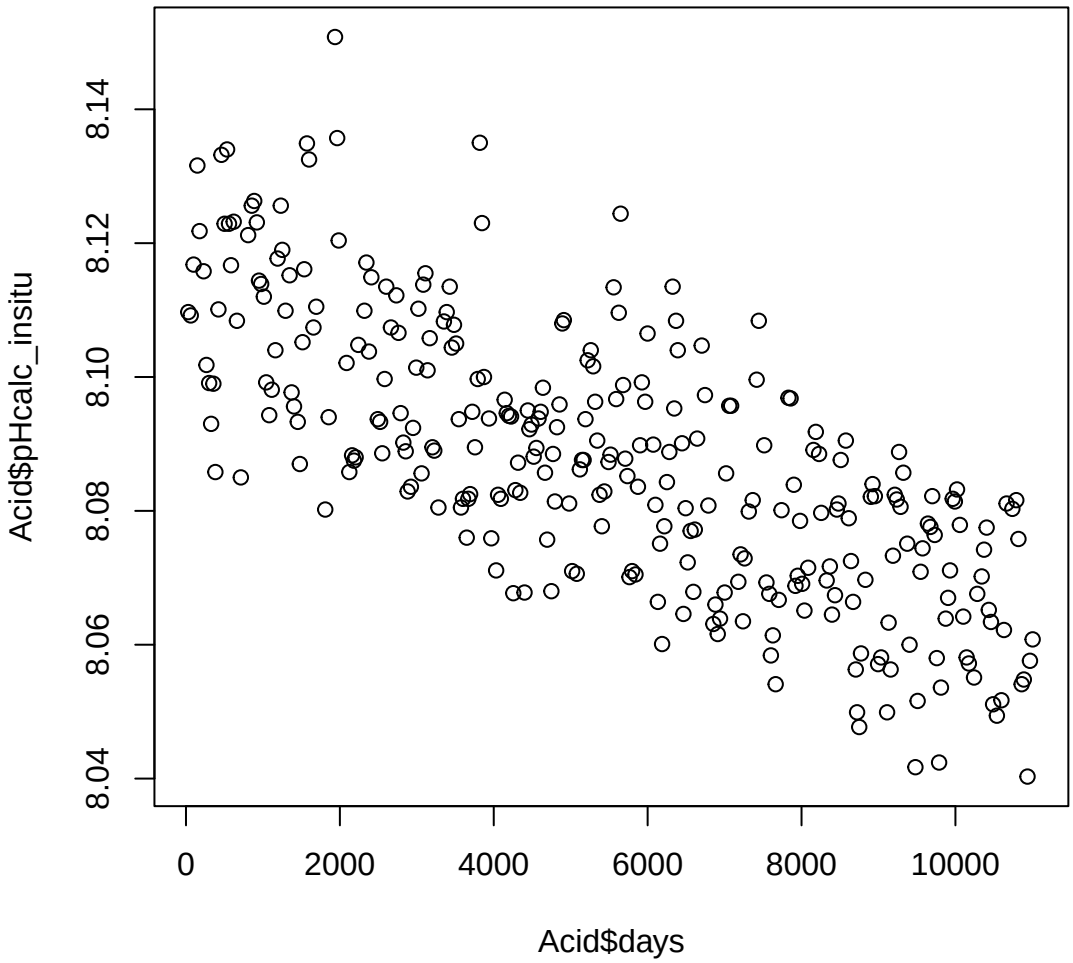
```
Acid = read.table(file = "Data/ph.txt", header = T, sep='\t', dec=',')
head(Acid)
```

```
##   days   temp pHcalc_insitu
## 1   30 26.283        8.1097
## 2   62 25.659        8.1092
## 3   99 24.610        8.1168
## 4  148 23.479        8.1316
## 5  177 24.278        8.1218
## 6  229 23.870        8.1158
```

```
pairs(Acid)
```




```
plot(Acid$days,Acid$pHcalc_insitu)
```



1-0) Modèle de régression linéaire

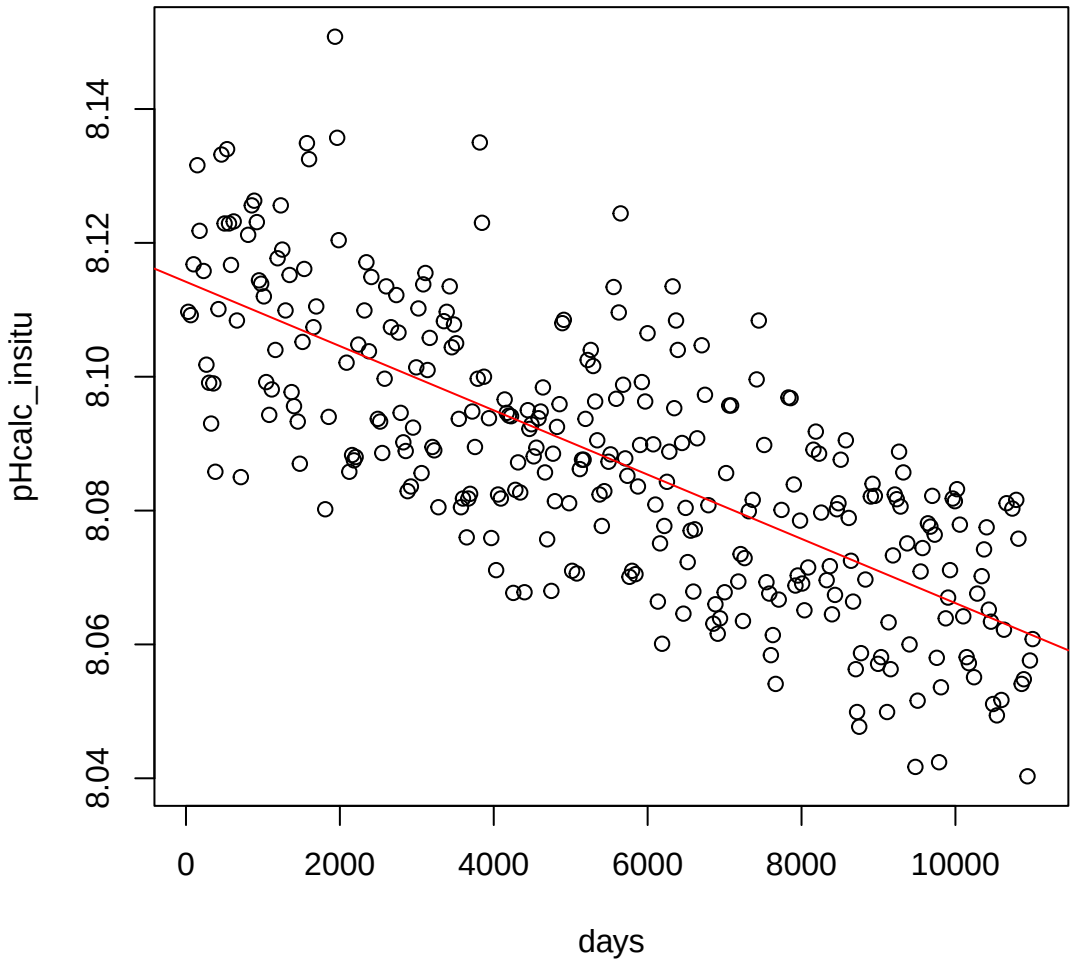
```
mod1 = lm(pHcalc_insitu~days, data = Acid)
summary(mod1)
```

```
##
## Call:
## lm(formula = pHcalc_insitu ~ days, data = Acid)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.026969 -0.010493 -0.000187  0.010587  0.045919
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  8.114e+00  1.597e-03 5079.97   <2e-16 ***
## days        -4.801e-06  2.520e-07  -19.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01359 on 297 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.5499, Adjusted R-squared:  0.5484
## F-statistic: 362.9 on 1 and 297 DF,  p-value: < 2.2e-16
```

1-a) Droite de régression et données sur le même graphique

```
plot(pHcalc_insitu~days, data = Acid)
abline(mod1$coefficients, col = 'red')
```



1-b) Pourcentage de variance expliqué par la régression

Le coefficient de détermination R2 mesure la part de la variance du pH expliquée par la variable days. Le modèle explique donc environ 54.99 % de la variabilité observée dans le pH de l'océan.

1-c) Analyse du test de student

On pose les hypothèses suivantes :

H0 : $\beta_{days} = 0$, c'est-à-dire que la variable days n'a pas un impact significatif sur la variable pH.

H1 : $\beta_{days} \neq 0$, la variable days a un impact significatif sur la variable pH.

Sous l'hypothèse H0, la statistique du test est donnée par :

$$t_{days} = \frac{\hat{\beta}_{days}}{\sqrt{\hat{\sigma}^2 \cdot [(X^T X)^{-1}]_{days,days}}}$$

Où :

- $\hat{\beta}_{days}$ est l'estimation du coefficient β_{days} obtenue par les moindres carrés.
- $\hat{\sigma}^2$ est l'estimation de la variance des erreurs.
- $[(X^T X)^{-1}]_{days,days}$ est l'élément de la diagonale de la matrice $(X^T X)^{-1}$ correspondant à la variable days.

Et ce car sous l'hypothèse H0, T suit une loi de Student à (n-p-1) degré de liberté.

On rejette H0 si la P-Value associée au coefficient days est inférieure à 0.05

Ici la P-Value $<$ 2e-16 est très faible devant le seuil 0.05 donc on rejette H0.

Cela signifie que la variable days est bien significative vis-à-vis des variations de pH.

2) Prédiction du pH moyen en 2050 et intervalle de confiance à 95%

```
predict(mod1, newdata = data.frame(days = 19000), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 8.022966 8.016103 8.029829
```

D'après la prédiction effectuée, on peut prévoir un pH moyen en 2050 de 8.022966.

L'intervalle de confiance à 95% pour le pH en 2050 est [8.016103 ; 8.029829].

3) Intervalle de prédiction à 95%

```
predict(mod1, newdata = data.frame(days=19000), interval = "prediction", level = 0.95)
```

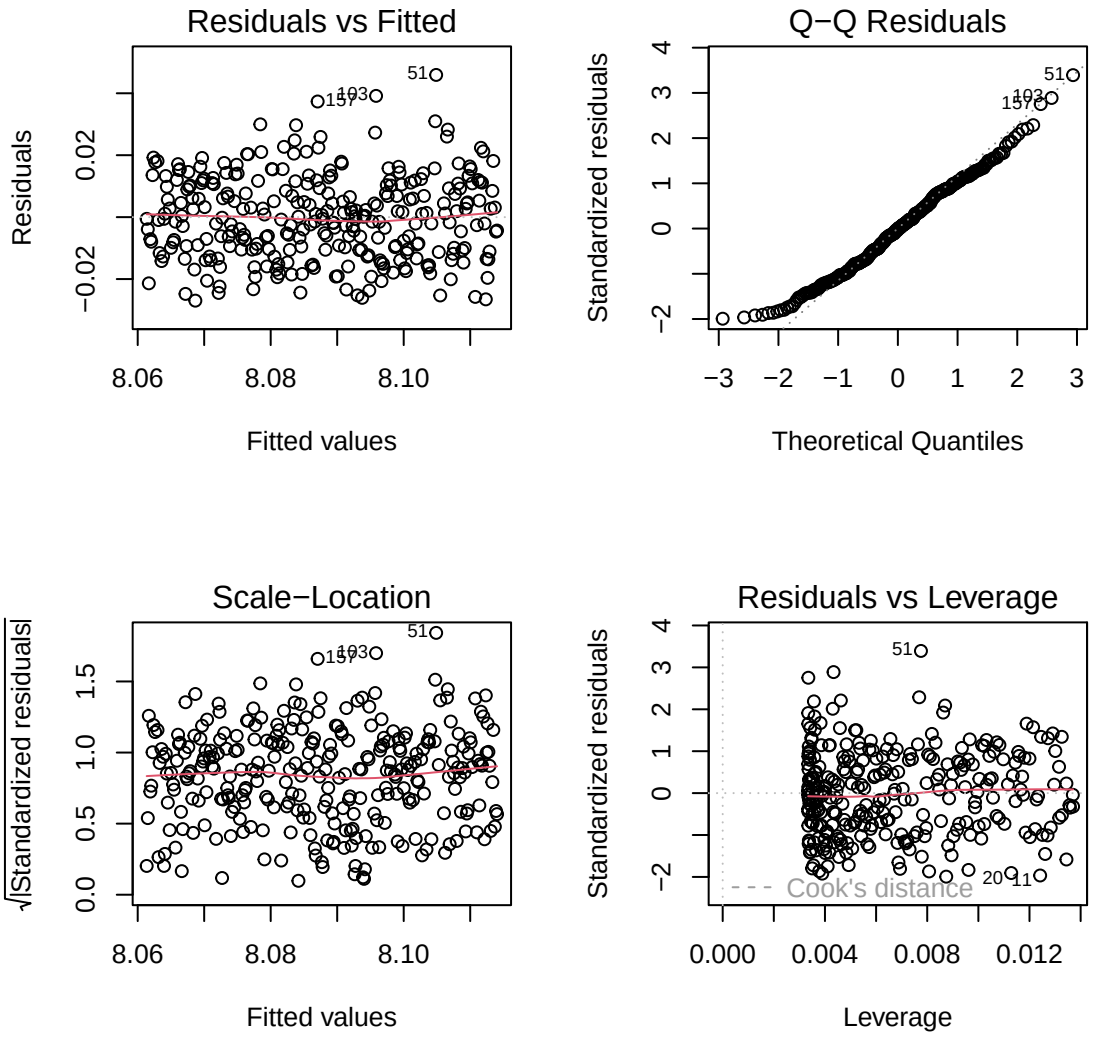
```
##        fit      lwr      upr
## 1 8.022966 7.995358 8.050575
```

L'intervalle de prédiction à 95 % pour le pH en 2050 est [7.995358 ; 8.050575].

Il est donc possible au seuil de 95% d'observer un pH inférieur à 8 (de justesse).

4) Etude des résidus

```
par(mfrow=c(2,2))
plot(mod1)
```



Sur le graphe Residuals vs Fitted, on observe que les résidus sont centrés sur 0, leur esperence semble bien nulle, et les points sont uniformément répartis autour de 0.

Sur le Graphe Q-Q residuals, on remarque que les points suivent dans l'ensemble relativement bien la diagonale et sont allignés, à l'exception des extrémités.

Sur le Graphe Residuals vs Leverage, on observe que l'ensemble des points est regroupé sous une distance de cook inférieure à 0.5 donc pas de valeurs à fort levier.

Le modèle semble donc valide.

5) L'océan se réchauffe t-il ?

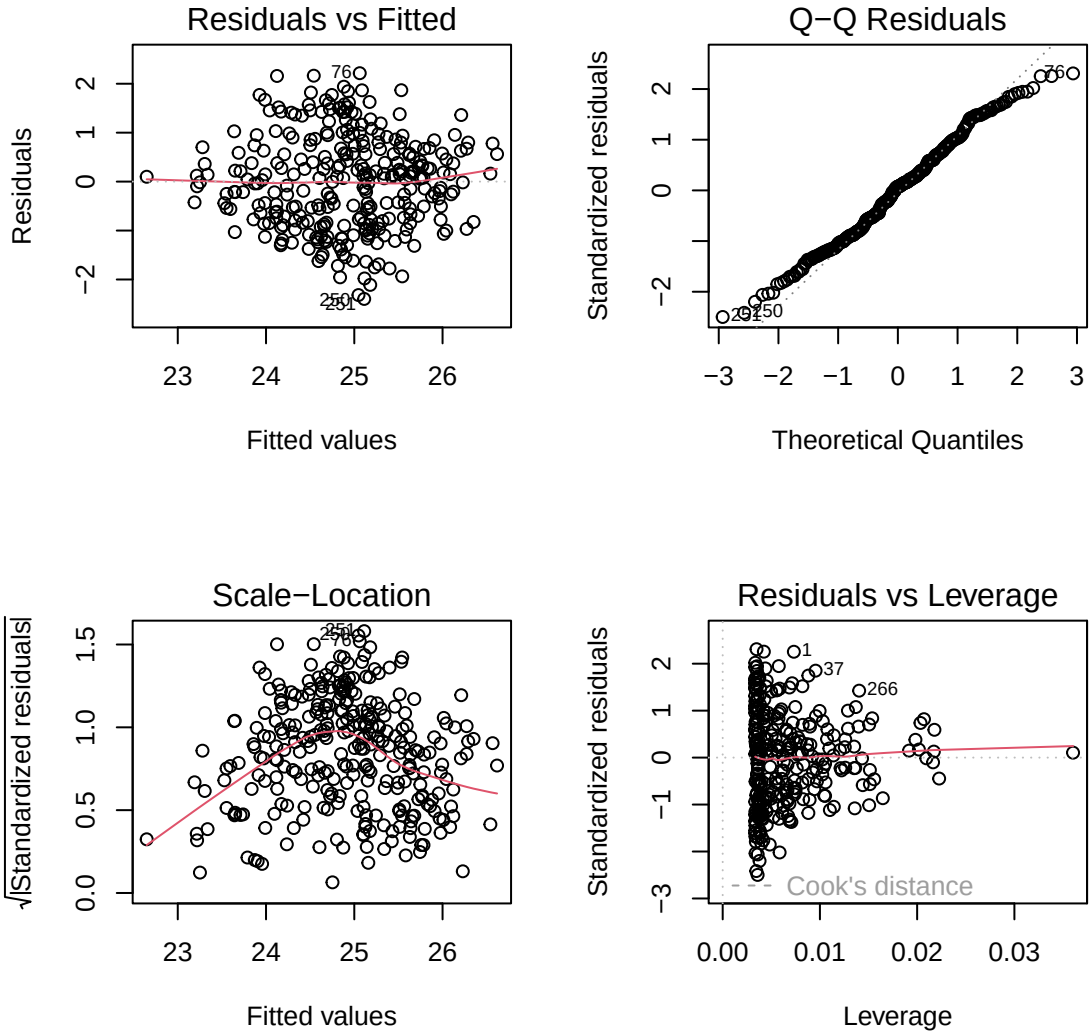
On commence par créer un modèle de régression linéaire pour prédire la température.

Au vu des relations entre les différentes variables, on choisit d'expliquer la température en fonction du pH :

```
mod_temp = lm(temp~pHcalc_insitu, data = Acid)
summary(mod_temp)
```

```
##
## Call:
## lm(formula = temp ~ pHcalc_insitu, data = Acid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.39717 -0.77155  0.07192  0.68099  2.21512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    315.441     22.271   14.16   <2e-16 ***
## pHcalc_insitu  -35.922      2.754  -13.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9612 on 297 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.3643, Adjusted R-squared:  0.3621
## F-statistic: 170.2 on 1 and 297 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod_temp)
```



Le modèle n'explique que 36.43% de la variance, mais cela s'explique par le fait que les données semblent très bruitées.

Le test de student sur le coefficient de la variable pH donne une P-Value très largement inférieure à 0.05, la variable est donc significative.

L'analyse des résidus quand à elle, similaire à la question 4, semble concluante et laisse penser que le modèle est valide. (Suivi d'une loi normale, résidus uniformément répartis autour de 0 et pas de levier)

Ce modèle prédit que la température diminue quand le pH augmente. Or on a vu précédemment que le pH diminue en fonction du temps. On peut donc en conclure que la température des océans augmente.

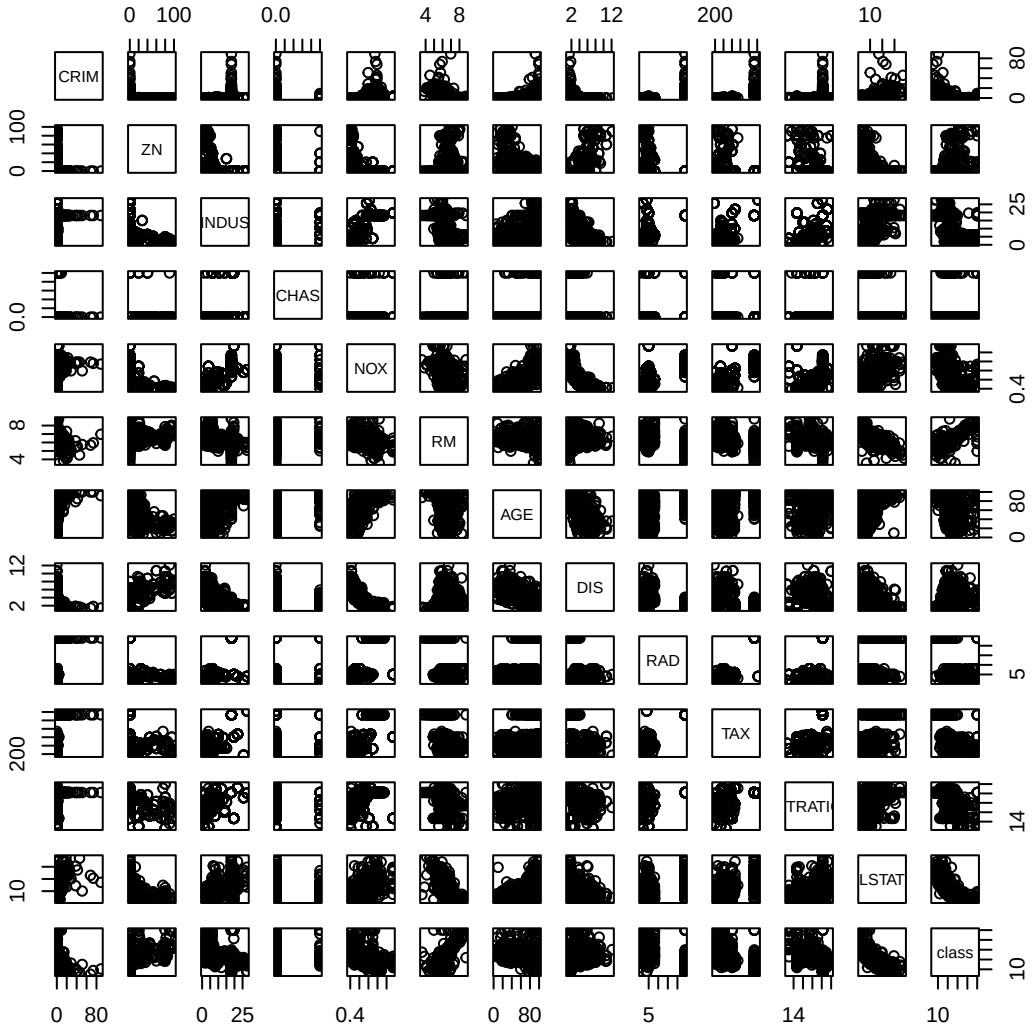
Exercice 2 - Valeur des logements des villes aux alentours de Boston

0) Chargement des données et création du modèle

```
Housing = read.table("Data/housing_new.txt", header = TRUE)
head(Housing)
```

```
##    CRIM ZN INDUS CHAS   NOX    RM  AGE   DIS RAD TAX PTRATIO LSTAT class
## 1 0.006 18  2.31    0 0.538 6.575 65.2 4.090   1 296    15.3  4.98  24.0
## 2 0.027  0  7.07    0 0.469 6.421 78.9 4.967   2 242    17.8  9.14  21.6
## 3 0.027  0  7.07    0 0.469 7.185 61.1 4.967   2 242    17.8  4.03  34.7
## 4 0.032  0  2.18    0 0.458 6.998 45.8 6.062   3 222    18.7  2.94  33.4
## 5 0.069  0  2.18    0 0.458 7.147 54.2 6.062   3 222    18.7  5.33  36.2
## 6 0.030  0  2.18    0 0.458 6.430 58.7 6.062   3 222    18.7  5.21  28.7
```

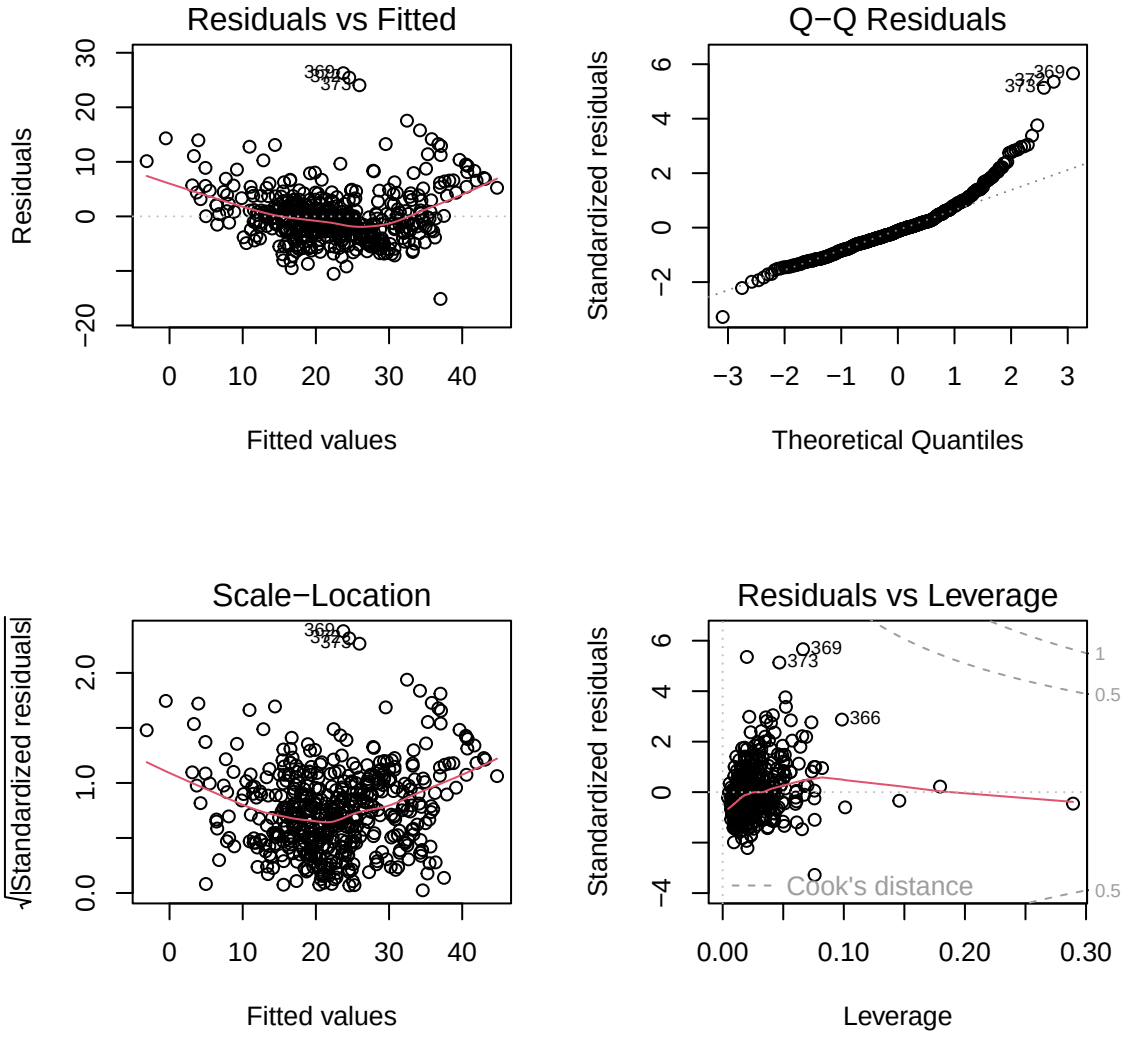
```
pairs(Housing)
```



```r
mod_full <- lm(class ~ ., data = Housing)
summary(mod_full)
```

```
##
## Call:
## lm(formula = class ~ ., data = Housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1299  -2.7673  -0.5816   1.9411  26.2520
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.617992   4.936061   8.431 3.79e-16 ***
## CRIM         -0.121391   0.033000  -3.678 0.000260 ***
## ZN            0.046962   0.013879   3.384 0.000772 ***
## INDUS         0.013462   0.062145   0.217 0.828591
## CHAS          2.840102   0.870005   3.264 0.001173 **
## NOX         -18.758738   3.851359  -4.871 1.50e-06 ***
## RM            3.658127   0.420244   8.705  < 2e-16 ***
## AGE           0.003610   0.013329   0.271 0.786658
## DIS          -1.490767   0.201622  -7.394 6.17e-13 ***
## RAD           0.289397   0.066908   4.325 1.84e-05 ***
## TAX          -0.012681   0.003801  -3.336 0.000913 ***
## PTRATIO      -0.937562   0.132207  -7.092 4.62e-12 ***
## LSTAT        -0.552013   0.050659 -10.897  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.798 on 493 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7278
## F-statistic: 113.5 on 12 and 493 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(mod_full)
```



1) Part de la variance expliquée par le modèle

Le modèle explique environ R2 = 73.43 % de la variabilité observée dans la valeur des logements.

2) Significativité du modèle

On teste H0 : tous les coefficients des variables explicatives sont nuls. La statistique F suit une loi de Fisher F12,493 sous H0. Avec une p-value $< 2.2 \times 10^{\wedge}(-16)$, très inférieure au risque de première espèce alpha = 1 %, on rejette H0.

Le modèle global est donc hautement significatif et les variables explicatives ont un effet sur la valeur des logements.

3) Variables significatives dans le modèle

En considérant un seuil de significativité alpha = 1 %, les variables significatives dans le modèle sont CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO et LSTAT (Ce sont toutes les caractéristiques qui ont une p-value < 0.01)

On ne peut pas être sûr que les autres variables n'aient aucun effet, car certains de ces effets peuvent être masqués par la colinéarité dans les variables explicatives.

4) Simplification du modèle

Pour simplifier le modèle, on utilise la méthode backward, qui supprime progressivement les variables les moins significatives afin de minimiser le critère AIC.

```
mod_simpl = step(mod_full, direction = "backward", trace = FALSE)
summary(mod_simpl)
```

```
##
## Call:
## lm(formula = class ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + LSTAT, data = Housing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.1809  -2.7623  -0.6239   1.8454  26.3913
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.452529   4.903305   8.454 3.17e-16 ***
## CRIM         -0.121667   0.032919  -3.696 0.000244 ***
## ZN            0.046190   0.013673   3.378 0.000787 ***
## CHAS          2.871969   0.862589   3.329 0.000935 ***
## NOX         -18.263334   3.565266  -5.123 4.33e-07 ***
## RM            3.672962   0.409125   8.978  < 2e-16 ***
## DIS          -1.515957   0.187672  -8.078 5.07e-15 ***
## RAD           0.283927   0.063945   4.440 1.11e-05 ***
## TAX          -0.012291   0.003407  -3.608 0.000340 ***
## PTRATIO      -0.930993   0.130423  -7.138 3.39e-12 ***
## LSTAT        -0.546504   0.047442 -11.519  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.789 on 495 degrees of freedom
## Multiple R-squared:  0.7342, Adjusted R-squared:  0.7289
## F-statistic: 136.8 on 10 and 495 DF,  p-value: < 2.2e-16
```

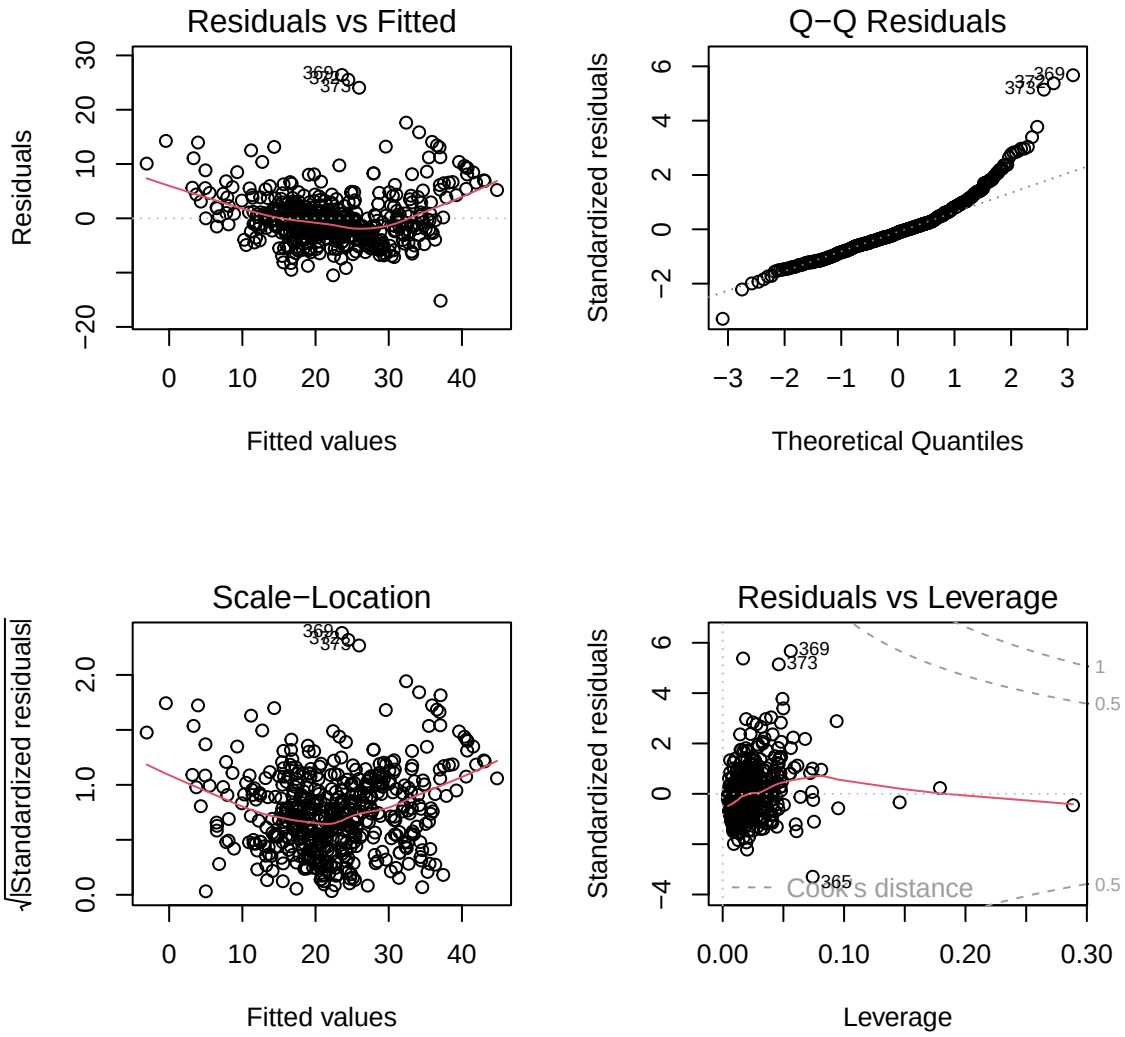
5) Evaluation du modèle obtenu

On compare ici les R2 ajustés pour prendre en compte le nombre de variables. Avec le modèle simplifié : R2 = 72.89%, tandis qu'avec le modèle complet, R2=72.78%. Le modèle est donc tout à fait performant vis-à-vis du modèle initial pour les prédictions.

Ce modèle a moins de variables donc plus de simplicité. Tous les coefficients des variables retenues ont une p-value inférieure à 0.01. Les variables retenues sont donc toutes significatives dans ce nouveau modèle.

Analysons les résidus :

```
par(mfrow=c(2,2))
plot(mod_simpl)
```



Ce modèle semble globalement valide et très similaire au modèle complet au niveau des résidus (malgré une légère tendance en U aux extrémités sur le graphe Residual vs Fitted et une légère déviation aux éxtrémités sur le graphe Q-Q Residuals)$

Ce modèle simple donc satisfaisant.

6) Proposition d'un meilleur modèle

On tente ici d'ajouter des interactions entre les variables explicatives :

```
# modèle avec toutes les interactions
new_full_model <- lm(class ~ (.)^2, data = Housing)
summary(new_full_model)
```

```
##
## Call:
## lm(formula = class ~ (.)^2, data = Housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9484 -1.6550 -0.0683  1.4551 19.3056
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.301e+02  6.051e+01  -2.151 0.032061 *
## CRIM          -1.899e+01  6.695e+00  -2.836 0.004783 **
## ZN            -1.125e-01  3.530e-01  -0.319 0.750107
## INDUS         -2.339e+00  1.550e+00  -1.510 0.131854
## CHAS           5.811e+01  1.918e+01   3.031 0.002590 **
## NOX            3.470e+01  7.593e+01   0.457 0.647895
## RM             2.464e+01  5.753e+00   4.282 2.29e-05 ***
## AGE            9.756e-01  2.577e-01   3.785 0.000175 ***
## DIS           -1.679e+00  3.899e+00  -0.431 0.666847
## RAD            1.241e+00  2.261e+00   0.549 0.583248
## TAX            5.039e-02  1.177e-01   0.428 0.668769
## PTRATIO        3.080e+00  2.607e+00   1.181 0.238132
## LSTAT          1.144e+00  8.232e-01   1.389 0.165489
## CRIM:ZN        3.458e-01  1.834e-01   1.885 0.060103 .
## CRIM:INDUS     6.457e-02  4.581e-01   0.141 0.887982
## CRIM:CHAS      2.126e+00  5.737e-01   3.706 0.000238 ***
## CRIM:NOX      -2.208e+00  8.471e-01  -2.607 0.009456 **
## CRIM:RM        2.476e-01  4.438e-02   5.580 4.29e-08 ***
## CRIM:AGE       6.166e-03  3.377e-03   1.826 0.068545 .
## CRIM:DIS       4.235e-02  8.854e-02   0.478 0.632695
## CRIM:RAD      -4.317e-01  5.955e-01  -0.725 0.468877
## CRIM:TAX       2.120e-02  4.391e-02   0.483 0.629551
## CRIM:PTRATIO   6.211e-01  3.381e-01   1.837 0.066865 .
## CRIM:LSTAT     3.392e-02  5.494e-03   6.174 1.55e-09 ***
## ZN:INDUS      -1.067e-03  4.795e-03  -0.222 0.824037
## ZN:CHAS       -1.948e-02  6.555e-02  -0.297 0.766518
## ZN:NOX         9.731e-02  4.857e-01   0.200 0.841298
## ZN:RM          2.403e-03  2.668e-02   0.090 0.928277
## ZN:AGE         3.906e-05  8.750e-04   0.045 0.964413
## ZN:DIS         8.290e-03  7.670e-03   1.081 0.280383
## ZN:RAD        -3.323e-03  7.155e-03  -0.464 0.642628
## ZN:TAX         4.215e-04  1.810e-04   2.329 0.020336 *
## ZN:PTRATIO    -3.032e-03  7.215e-03  -0.420 0.674540
## ZN:LSTAT      -1.011e-02  4.473e-03  -2.260 0.024306 *
## INDUS:CHAS    -2.902e-01  3.883e-01  -0.747 0.455324
## INDUS:NOX      2.585e+00  1.480e+00   1.747 0.081349 .
## INDUS:RM       3.548e-01  1.358e-01   2.612 0.009310 **
## INDUS:AGE      8.263e-04  3.744e-03   0.221 0.825412
## INDUS:DIS     -4.160e-02  6.417e-02  -0.648 0.517162
## INDUS:RAD     -1.608e-02  5.118e-02  -0.314 0.753511
## INDUS:TAX      5.210e-04  6.194e-04   0.841 0.400735
## INDUS:PTRATIO -6.148e-02  3.848e-02  -1.598 0.110855
## INDUS:LSTAT   -2.350e-03  1.557e-02  -0.151 0.880060
## CHAS:NOX      -4.362e+01  1.206e+01  -3.616 0.000335 ***
```

```
## CHAS:RM       -5.395e+00  1.181e+00  -4.569 6.43e-06 ***
## CHAS:AGE       3.674e-02  5.985e-02   0.614 0.539613
## CHAS:DIS       4.268e-01  1.359e+00   0.314 0.753623
## CHAS:RAD      -6.095e-01  5.851e-01  -1.042 0.298136
## CHAS:TAX       4.612e-02  3.756e-02   1.228 0.220195
## CHAS:PTRATIO  -5.931e-01  7.157e-01  -0.829 0.407735
## CHAS:LSTAT    -2.284e-01  1.864e-01  -1.225 0.221114
## NOX:RM         3.722e+00  5.408e+00   0.688 0.491714
## NOX:AGE       -7.436e-01  2.324e-01  -3.200 0.001475 **
## NOX:DIS        2.229e+00  3.794e+00   0.588 0.557121
## NOX:RAD        4.322e-01  1.931e+00   0.224 0.822998
## NOX:TAX       -4.754e-02  1.340e-01  -0.355 0.722906
## NOX:PTRATIO   -3.138e+00  3.172e+00  -0.989 0.323152
## NOX:LSTAT      1.723e+00  6.116e-01   2.818 0.005064 **
## RM:AGE        -5.887e-02  2.261e-02  -2.603 0.009553 **
## RM:DIS         2.363e-01  3.352e-01   0.705 0.481144
## RM:RAD        -9.412e-02  1.545e-01  -0.609 0.542839
## RM:TAX        -2.224e-02  1.009e-02  -2.204 0.028025 *
## RM:PTRATIO    -4.893e-01  2.216e-01  -2.208 0.027755 *
## RM:LSTAT      -3.035e-01  4.198e-02  -7.231 2.24e-12 ***
## AGE:DIS       -1.537e-02  9.066e-03  -1.695 0.090728 .
## AGE:RAD        1.359e-02  4.248e-03   3.198 0.001487 **
## AGE:TAX       -2.951e-04  2.220e-04  -1.329 0.184409
## AGE:PTRATIO   -9.323e-03  6.852e-03  -1.361 0.174336
## AGE:LSTAT     -5.298e-03  1.961e-03  -2.702 0.007163 **
## DIS:RAD       -2.802e-02  7.141e-02  -0.392 0.694992
## DIS:TAX       -3.811e-03  2.502e-03  -1.524 0.128348
## DIS:PTRATIO   -5.053e-02  1.016e-01  -0.497 0.619153
## DIS:LSTAT      1.264e-01  4.875e-02   2.594 0.009827 **
## RAD:TAX       -3.544e-04  1.470e-03  -0.241 0.809534
## RAD:PTRATIO   -3.277e-02  8.559e-02  -0.383 0.702012
## RAD:LSTAT     -2.747e-02  1.564e-02  -1.757 0.079657 .
## TAX:PTRATIO    7.359e-03  2.533e-03   2.905 0.003865 **
## TAX:LSTAT     -1.401e-03  9.999e-04  -1.401 0.161937
## PTRATIO:LSTAT  6.569e-03  2.909e-02   0.226 0.821481
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.957 on 427 degrees of freedom
## Multiple R-squared:  0.9126, Adjusted R-squared:  0.8967
## F-statistic: 57.18 on 78 and 427 DF,  p-value: < 2.2e-16
```

Puis on simplifie le modèle avec BIC cette fois-ci (car la méthode est plus restrictive que AIC et on a beaucoup de variables)

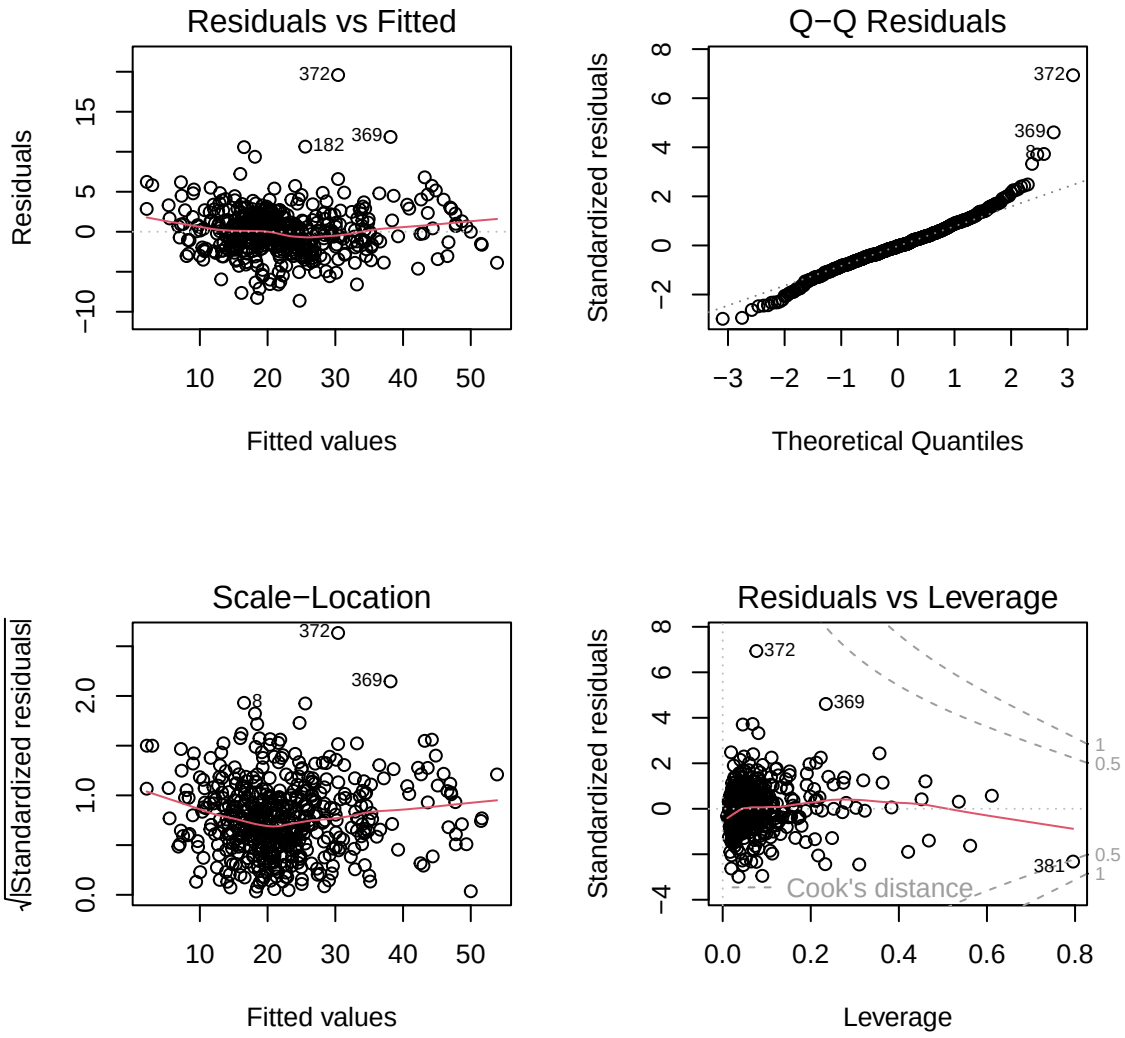
```
best_model <- step(new_full_model, direction = "backward",k = log(nrow(Acid)), trace = FALSE)
summary(best_model)
```

```
##
## Call:
## lm(formula = class ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + TAX + PTRATIO + LSTAT + CRIM:CHAS + CRIM:NOX +
##     CRIM:RM + CRIM:LSTAT + ZN:TAX + ZN:LSTAT + INDUS:NOX + INDUS:RM +
```

```
##     CHAS:NOX + CHAS:RM + CHAS:LSTAT + NOX:AGE + NOX:PTRATIO +
##     NOX:LSTAT + RM:AGE + RM:TAX + RM:PTRATIO + RM:LSTAT + AGE:DIS +
##     AGE:RAD + AGE:LSTAT + DIS:TAX + DIS:LSTAT + RAD:LSTAT + TAX:PTRATIO +
##     TAX:LSTAT, data = Housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6242 -1.6278 -0.0799  1.4598 19.5894
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.893e+02  1.984e+01  -9.544  < 2e-16 ***
## CRIM        -6.945e-01  4.133e-01  -1.681 0.093525 .
## ZN          -7.952e-02  3.667e-02  -2.169 0.030609 *
## INDUS       -3.634e+00  5.635e-01  -6.449 2.81e-10 ***
## CHAS         5.285e+01  8.326e+00   6.348 5.17e-10 ***
## NOX          9.474e+01  1.989e+01   4.763 2.55e-06 ***
## RM           3.153e+01  2.145e+00  14.694  < 2e-16 ***
## AGE          8.672e-01  1.313e-01   6.604 1.09e-10 ***
## DIS          7.115e-01  4.516e-01   1.576 0.115792
## RAD         -1.761e-01  1.539e-01  -1.144 0.253186
## TAX          1.151e-01  3.116e-02   3.694 0.000247 ***
## PTRATIO      4.704e+00  1.068e+00   4.404 1.32e-05 ***
## LSTAT        1.605e+00  3.952e-01   4.062 5.71e-05 ***
## CRIM:CHAS    2.075e+00  2.975e-01   6.975 1.05e-11 ***
## CRIM:NOX    -2.217e+00  6.657e-01  -3.330 0.000938 ***
## CRIM:RM      2.225e-01  3.836e-02   5.801 1.22e-08 ***
## CRIM:LSTAT   3.360e-02  5.093e-03   6.598 1.13e-10 ***
## ZN:TAX       4.792e-04  1.306e-04   3.669 0.000272 ***
## ZN:LSTAT    -9.398e-03  2.662e-03  -3.530 0.000456 ***
## INDUS:NOX    2.170e+00  4.739e-01   4.580 5.98e-06 ***
## INDUS:RM     4.168e-01  7.464e-02   5.585 3.98e-08 ***
## CHAS:NOX    -3.147e+01  5.438e+00  -5.786 1.32e-08 ***
## CHAS:RM     -5.220e+00  1.016e+00  -5.139 4.07e-07 ***
## CHAS:LSTAT  -3.140e-01  1.276e-01  -2.461 0.014197 *
## NOX:AGE     -9.036e-01  1.604e-01  -5.633 3.06e-08 ***
## NOX:PTRATIO -4.416e+00  8.543e-01  -5.170 3.48e-07 ***
## NOX:LSTAT    1.300e+00  4.016e-01   3.238 0.001290 **
## RM:AGE      -5.900e-02  1.379e-02  -4.278 2.29e-05 ***
## RM:TAX      -2.738e-02  3.335e-03  -8.208 2.21e-15 ***
## RM:PTRATIO  -6.787e-01  1.245e-01  -5.451 8.13e-08 ***
## RM:LSTAT    -2.971e-01  3.599e-02  -8.256 1.56e-15 ***
## AGE:DIS     -2.398e-02  5.746e-03  -4.173 3.59e-05 ***
## AGE:RAD      9.982e-03  1.643e-03   6.077 2.54e-09 ***
## AGE:LSTAT   -4.764e-03  1.594e-03  -2.989 0.002946 **
## DIS:TAX     -3.833e-03  1.285e-03  -2.982 0.003009 **
## DIS:LSTAT    9.550e-02  3.055e-02   3.126 0.001885 **
## RAD:LSTAT   -2.361e-02  7.081e-03  -3.334 0.000923 ***
## TAX:PTRATIO  3.817e-03  1.278e-03   2.987 0.002966 **
## TAX:LSTAT   -1.723e-03  4.040e-04  -4.266 2.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.939 on 467 degrees of freedom
```

```
## Multiple R-squared:  0.9056, Adjusted R-squared:  0.8979
## F-statistic: 117.8 on 38 and 467 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(best_model)
```



On obtient finalement un modèle très complexe mais beaucoup plus performant car R2 = 90.56% et R2 ajusté = 89.79%.

L'étude des résidus est également plus convaincante que celle des modèles précédents : en effet La tendance en U dans residuals vs fitted est atténuée et les points sont beaucoup plus alignés sur la diagonale de Q-Q Residuals.

Ce modèle est beaucoup plus complexe que les autres mais s'adapte bien dans un cadre ou la prédiction est favorisée devant l'explicabilité du modèle.

NB : Nous n'avons plus de temps mais il faudrait analyser les variances des variables explicatives : il y a un risque de surapprentissage avec ce modèle.