BE1- Régression linéaire

Tulio NAVARRO TUTUI, Filipe PENNA CERAVOLO SOARES

12 October, 2022

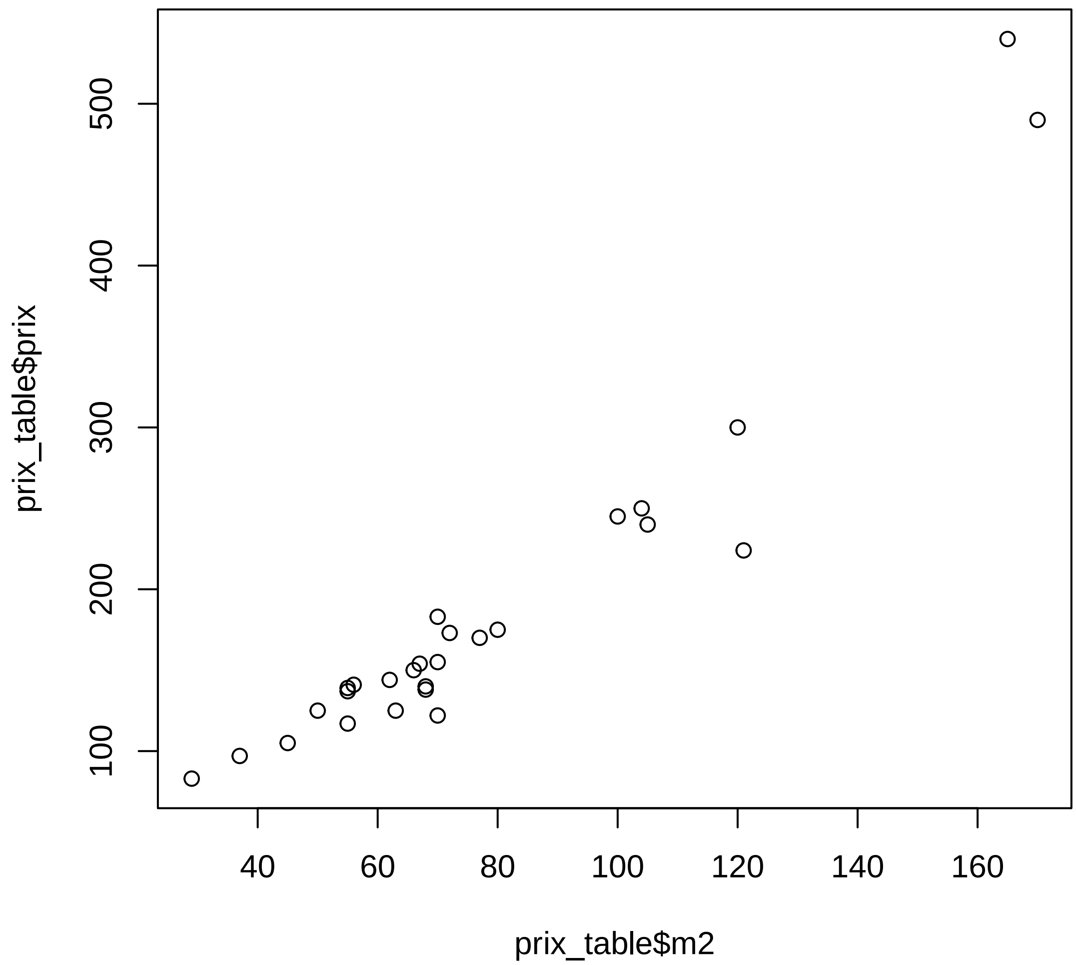
Exercice 01 - Prix de mise en vente des appartements à Grenoble

Dans ce éxercice, on s'interesse à étudier la rélation entre prix et surface des immeubles à Grenoble.

```
prix_table = read.table(file = "immo.txt", header = TRUE)
head(prix_table)
```

```
##   m2 prix
## 1 29   83
## 2 37   97
## 3 45  105
## 4 70  122
## 5 50  125
## 6 55  117
```

```
plot(prix_table$m2, prix_table$prix)
```



1. Proposer un premier modèle de régression

On commence par essayer de modéliser le problème avec une régression lineaire, avec l'estimateur de moindres carrés.

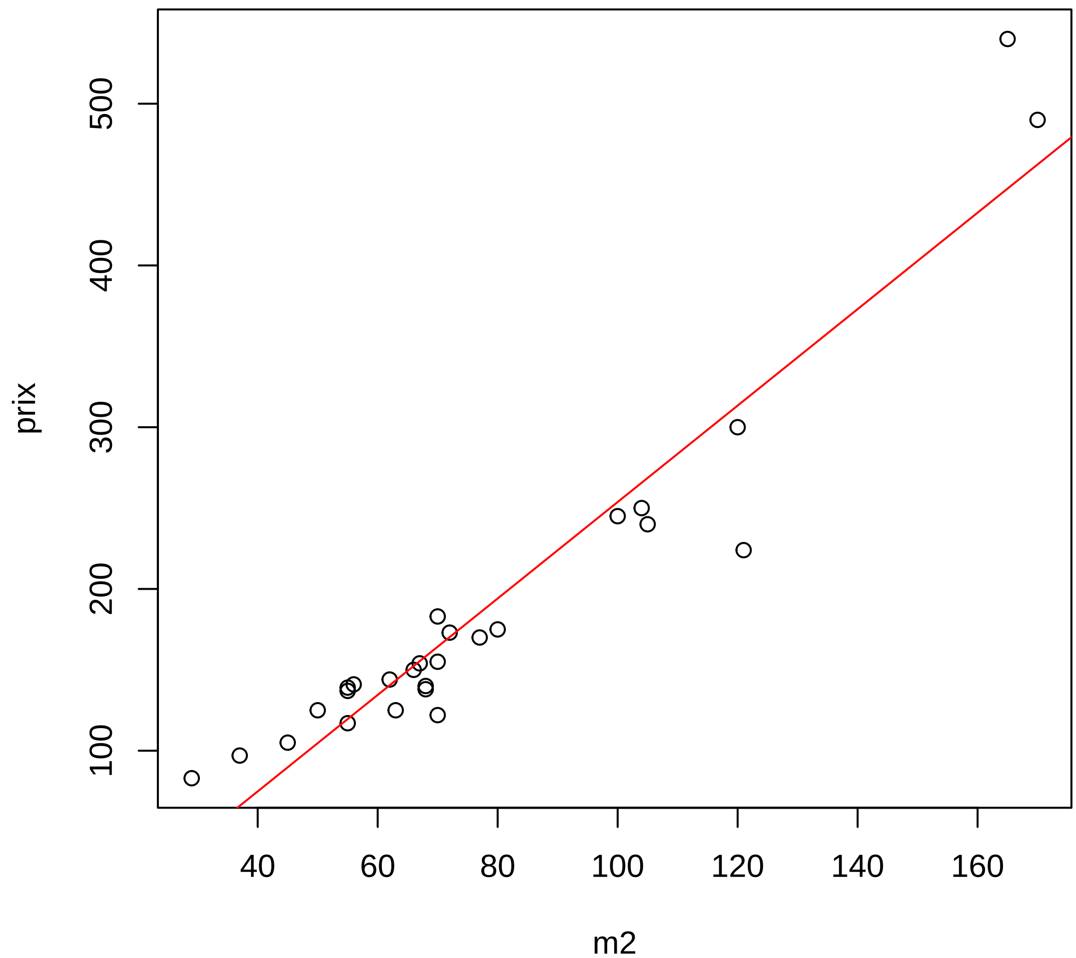
```
model1 = lm(prix~., data = prix_table)
summary(model1)
```

```
##
## Call:
## lm(formula = prix ~ ., data = prix_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.347 -16.996  -2.367  18.578  92.470
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.4093    16.0127  -2.773   0.0103 *
## m2            2.9815     0.1889  15.786 1.65e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.11 on 25 degrees of freedom
## Multiple R-squared:  0.9088, Adjusted R-squared:  0.9052
## F-statistic: 249.2 on 1 and 25 DF,  p-value: 1.646e-14
```

1.a La droite de régression

```r
plot(prix~m2, data = prix_table)
abline(model1$coefficients, col = 'red')
```



1.b Quel est le pourcentage de variance expliquée par cette régression ?

```
sse_1 <- sum((fitted(model1) - prix_table$prix)^2)

ssr_1 <- sum((fitted(model1) - mean(prix_table$prix))^2)

sst_1 <- ssr_1 + sse_1
ssr_1/sst_1
```

```
## [1] 0.9088217
```

Pour ce calcul, on peut utiliser directment le R^2, ou calculer les SSR et SST pour trouver SSR/SST. Ainsi, on a que le pourcentage de variance expliquée par cette régression est 90,88%.

1.c Analyse du test de student

H_0 : Le variable (m^2) n'a pas un relation lineére avec le variable réponse(prix) Pour cette hypothèse, on utilise la valeur $\alpha = 5$, et si la p_{value} est inferieur a α, on rejjete H_0.

```
summary(model1)
```

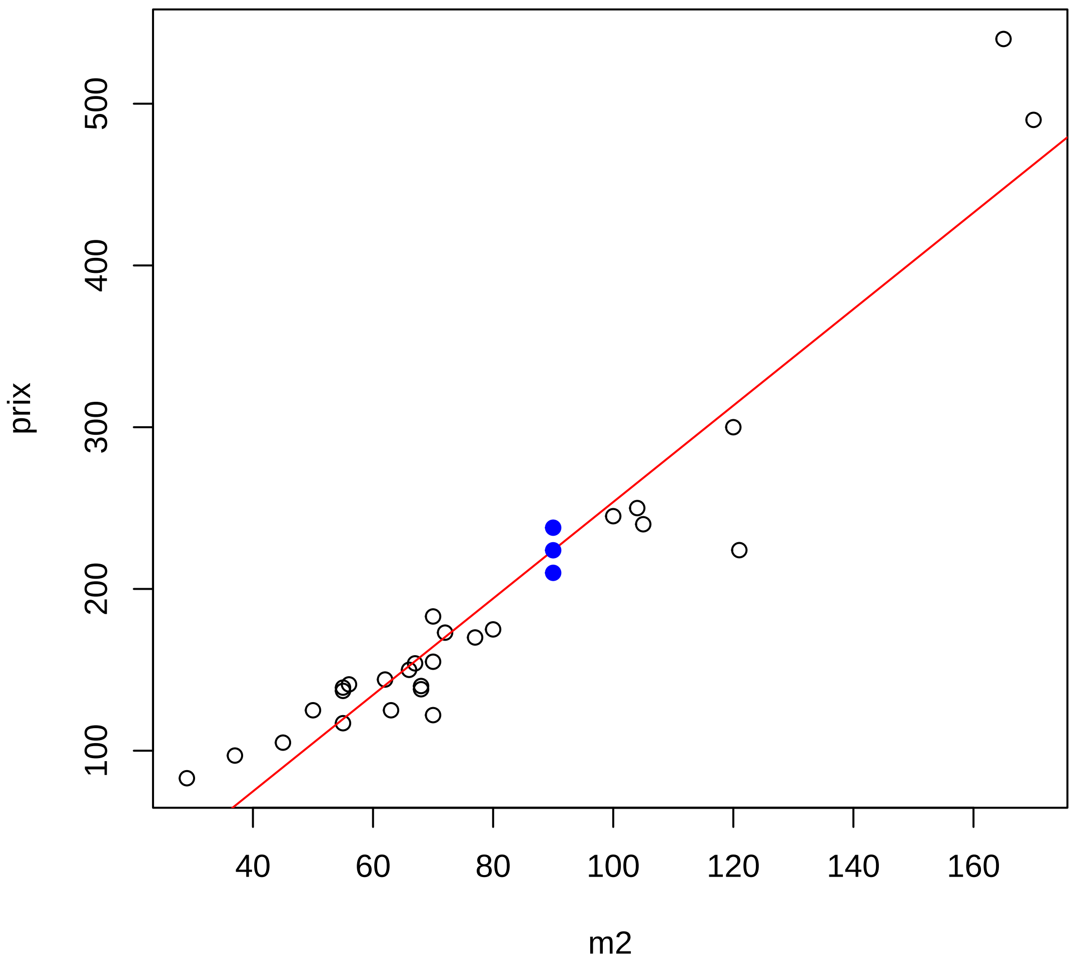
```
##
## Call:
## lm(formula = prix ~ ., data = prix_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.347 -16.996  -2.367  18.578  92.470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -44.4093    16.0127  -2.773   0.0103 *
## m2            2.9815     0.1889  15.786 1.65e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.11 on 25 degrees of freedom
## Multiple R-squared:  0.9088, Adjusted R-squared:  0.9052
## F-statistic: 249.2 on 1 and 25 DF,  p-value: 1.646e-14
```

Dans ce cas, la p_{value} de la variable m2 est de l'ordre de 10^{-14}, donc ce variable(m2) a une relation lineére avec prix.

2. Intervalle de confiance

Pour un niveau de confiance de 95% et un appartement de 90m², on prevoit la valeur $223,915 \pm [209,962; 237.8809]$

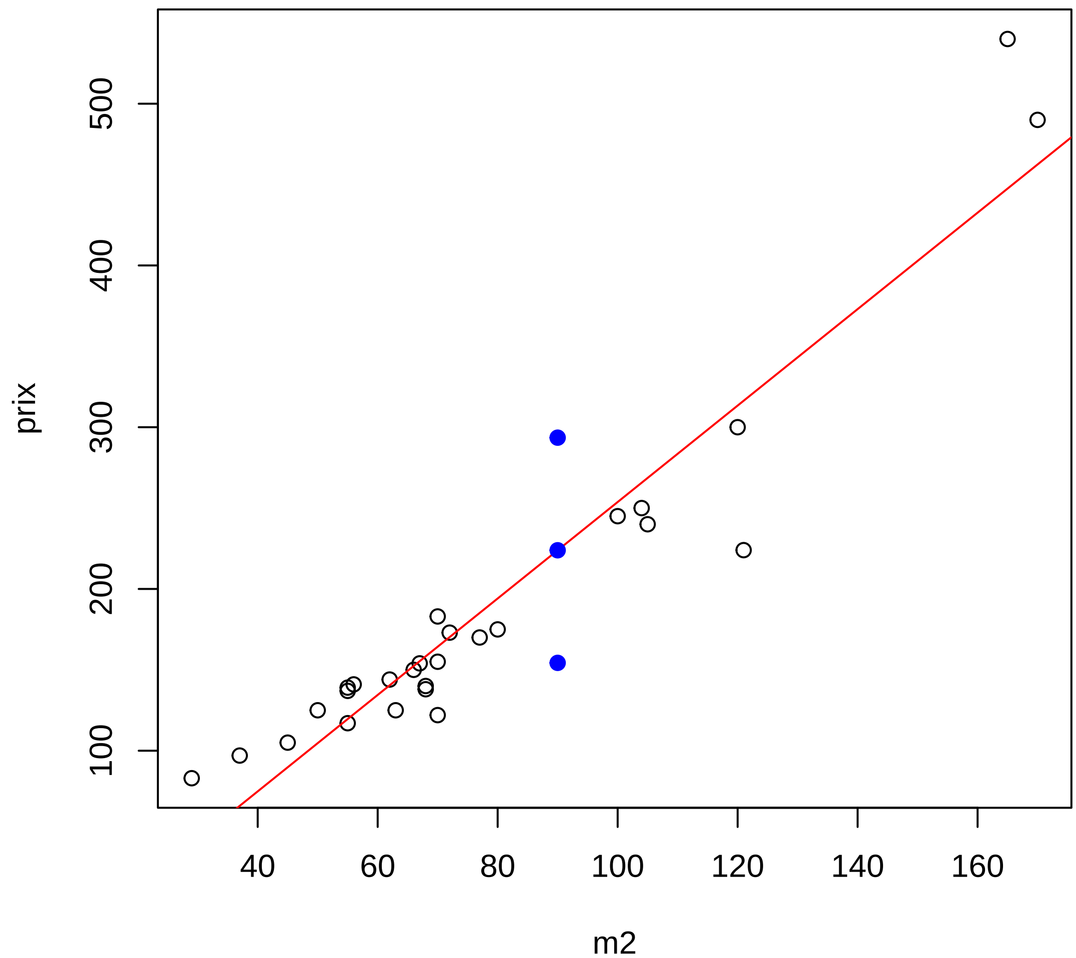
```
nouvelle_apart = data.frame(m2 = 90)
prediction_IC = data.frame(predict(model1, newdata = nouvelle_apart,
                                   interval = 'confidence', level = 0.95))

plot(prix~m2, data = prix_table)
abline(model1$coefficients, col = 'red')
points(nouvelle_apart$m2, prediction_IC$fit,col = 'blue', pch=19)
points(nouvelle_apart$m2, prediction_IC$lwr,col = "blue", pch=19)
points(nouvelle_apart$m2, prediction_IC$upr,col = "blue", pch=19)
```



3. Intervalle de prédiction

Cela donne également la suivante intervalle de prédiction: $223, 915 \pm [154.3086; 293.534]$. En observant cette valeur, on arrive à la conclusion qui *est acceptable de mettre en vente un appartement de* $90m^2$ *à 280 Keuros.*

```
prediction_IP = data.frame(predict(model1, newdata = nouvelle_apart,
                                   interval = 'prediction', level = 0.95))

plot(prix~m2, data = prix_table)
abline(model1$coefficients, col = 'red')

points(nouvelle_apart$m2, prediction_IP$fit,col = 'blue', pch=19)
points(nouvelle_apart$m2, prediction_IP$lwr, col = "blue", pch=19)
points(nouvelle_apart$m2, prediction_IP$upr, col = "blue", pch=19)
```

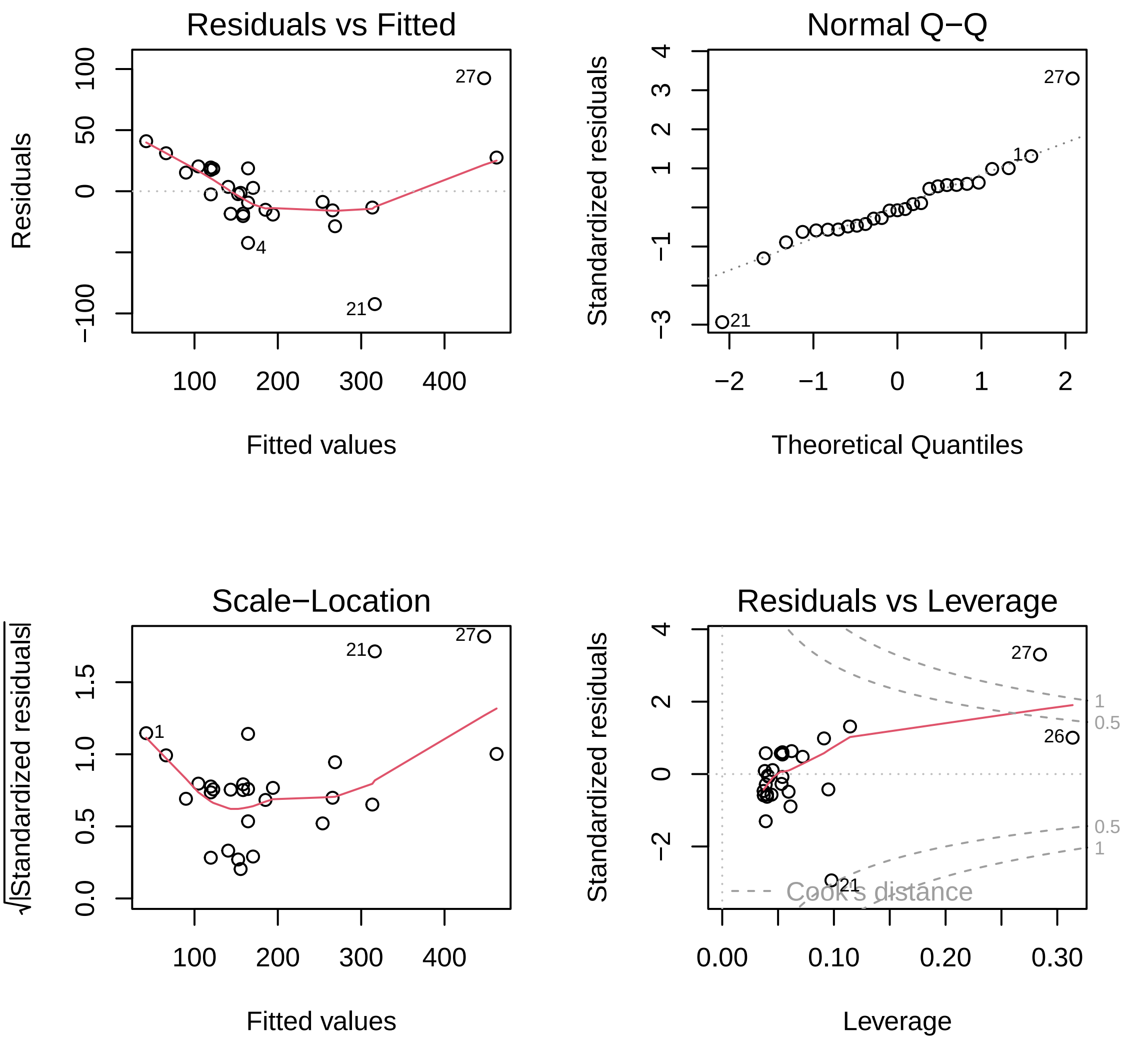


4. Étude des résidus

Les propriétés des estimateurs de la régression linéaire viennent de : Pour tous les i dans {1,. . . ,n}, $E(\epsilon_i) = 0$ $V(\epsilon_i) = \sigma^2 = cte$ Et pour tous les i different de j, $cov(\epsilon_i, \epsilon_j) = 0$

Pour ce faire, on étudie le graphe des résidus.

```
par(mfrow=c(2, 2))
plot(model1)
```



On peut vérifier visuellement que ce modèle ne satisfait pas ces hypothèses. D'abord, on s'attendait une ligne horizontale par le graphe des résidus x la régression (ça veut dire une moyenne de ϵ égale à zéro). En outre, on s'attendait le même par le graphe de la racine des résidus x la régression. Finalement, on observe des points à l'extérieur de la ligne pointillé. Ce que veut dire qu'il a des points à être enlevés (27).

```
ecarts = abs(prix_table$prix - model1$fitted.values)
indice = ecarts == max(ecarts)
prix_table[indice,]
```

```
##     m2 prix
## 27 165  540
```

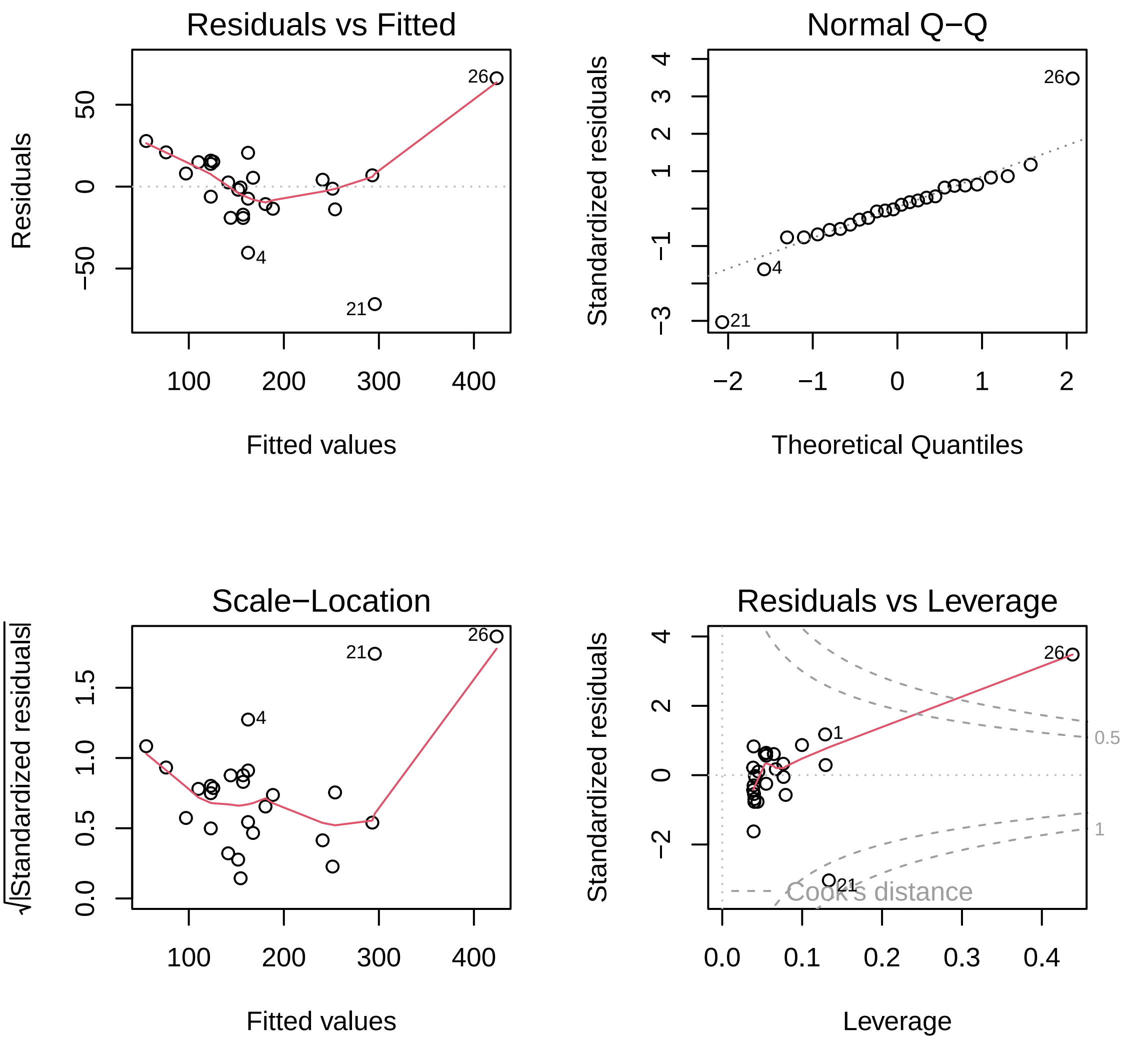
```
c(prix_table$prix[indice], model1$fitted.values[indice])
```

```
##                 27
## 540.0000 447.5304
```

```r
filtered_table = prix_table[-c(27),]
model2 = lm(prix~., data = filtered_table)
summary(model2)
```

```
##
## Call:
## lm(formula = prix ~ ., data = filtered_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.709 -12.794   1.023  14.668  66.170
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -20.672     13.454  -1.536    0.138
## m2             2.615      0.168  15.568 4.82e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.38 on 24 degrees of freedom
## Multiple R-squared:  0.9099, Adjusted R-squared:  0.9061
## F-statistic: 242.4 on 1 and 24 DF,  p-value: 4.82e-14
```

```r
par(mfrow=c(2, 2))
plot(model2)
```



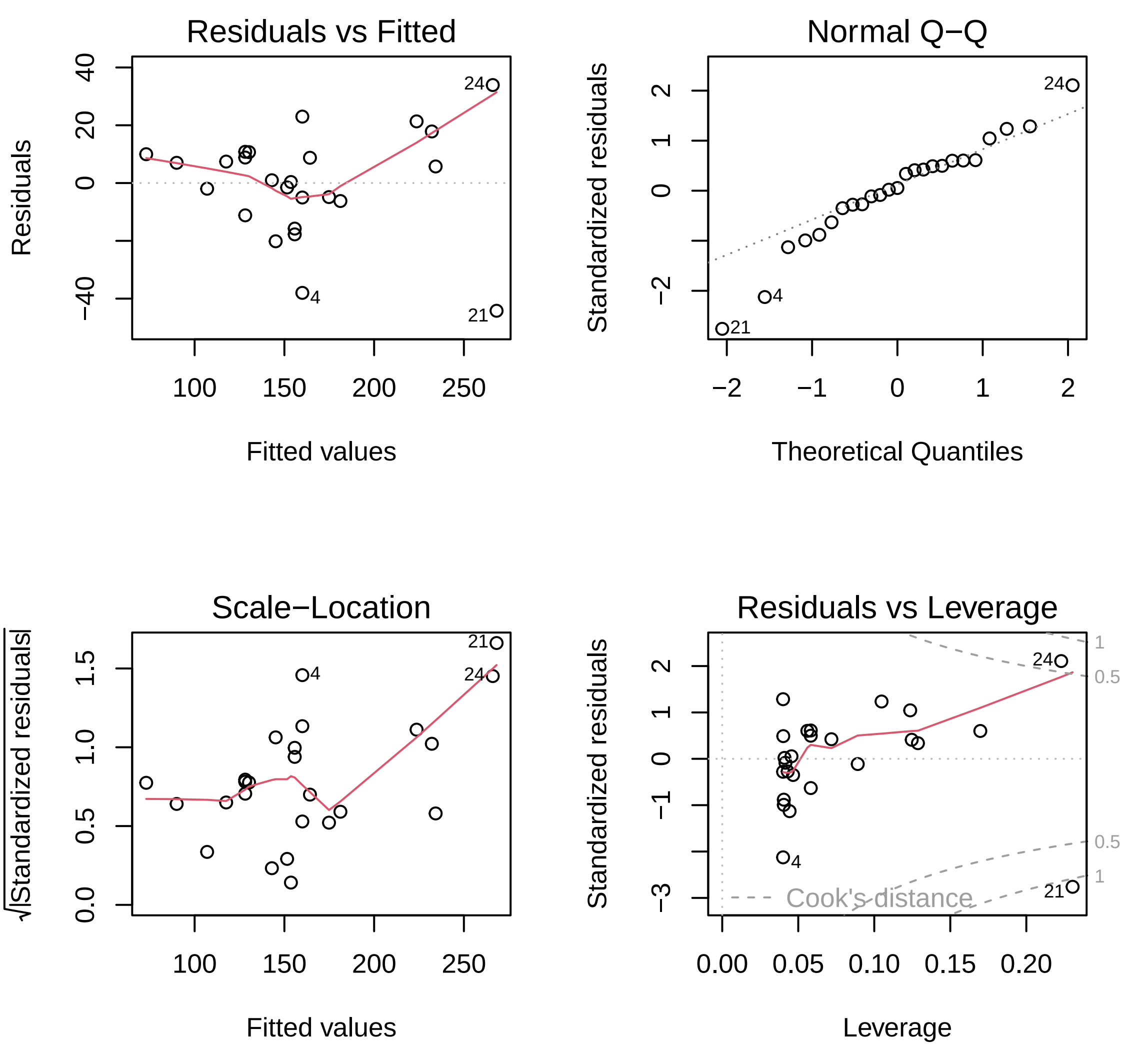
On observe que le modèle continue a avoir une mauvaise résultat. Donc on continue à enlever des points.

```
filtered_table_2 = prix_table[-c(27, 26),]
model3 = lm(prix~., data = filtered_table_2)
summary(model3)
```

```
##
## Call:
## lm(formula = prix ~ ., data = filtered_table_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.208  -6.223   0.966   9.978  33.913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.4956    11.7415   0.979    0.338
```

```
## m2             2.1216     0.1581  13.422 2.29e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.26 on 23 degrees of freedom
## Multiple R-squared:  0.8868, Adjusted R-squared:  0.8819
## F-statistic: 180.2 on 1 and 23 DF,  p-value: 2.291e-12
```

```
par(mfrow=c(2, 2))
plot(model3)
```

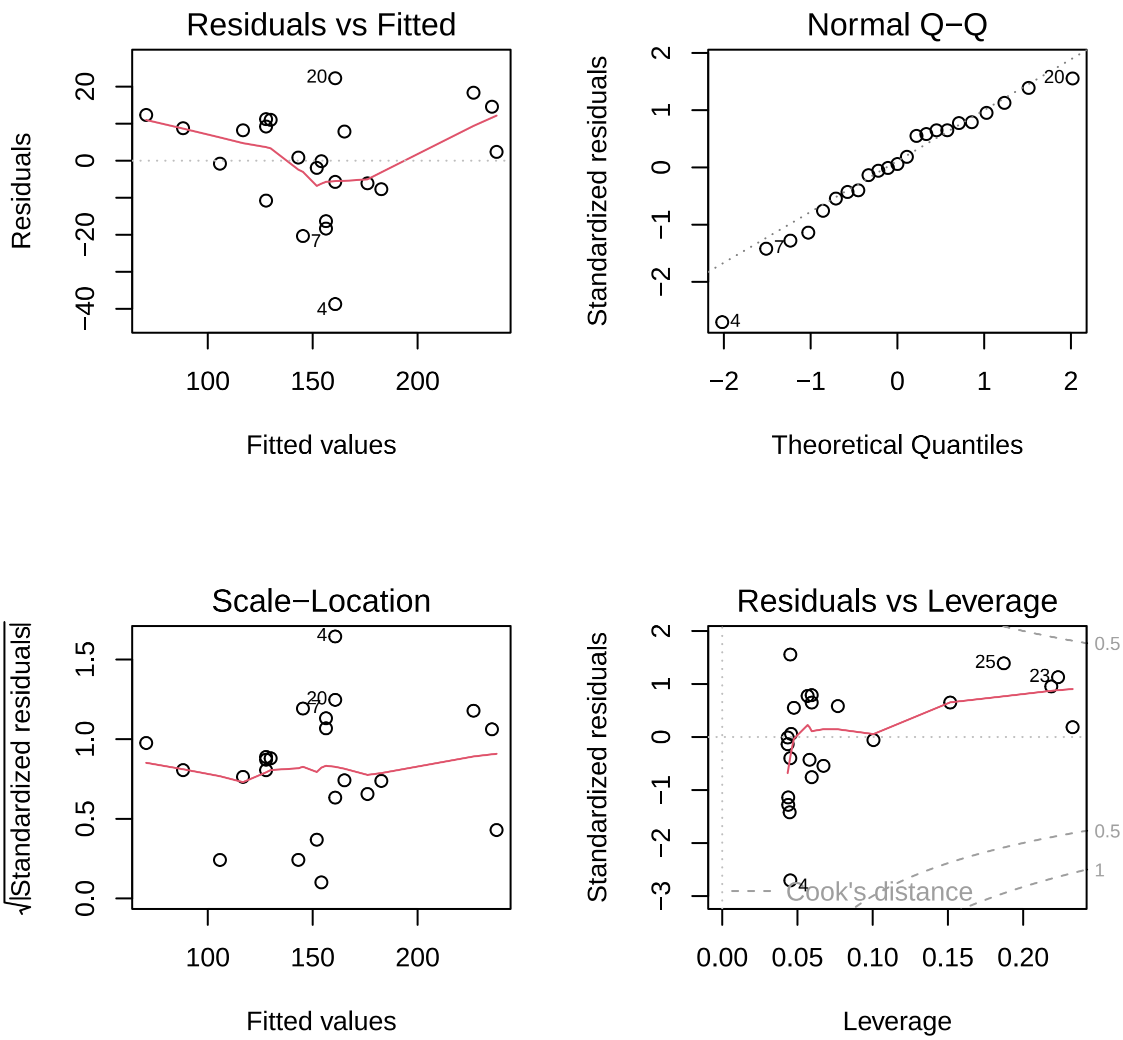



```
filtered_table_3 = prix_table[-c(27, 26, 24, 21),]
model4 = lm(prix~., data = filtered_table_3)
summary(model4)
```

```
##
```

```
## Call:
## lm(formula = prix ~ ., data = filtered_table_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.737  -6.911   0.839  10.120  22.263
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.9476    11.3287   0.613    0.546
## m2            2.1970     0.1646  13.345    1e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.66 on 21 degrees of freedom
## Multiple R-squared:  0.8945, Adjusted R-squared:  0.8895
## F-statistic: 178.1 on 1 and 21 DF,  p-value: 1.003e-11
```

```r
par(mfrow=c(2, 2))
plot(model4)
```



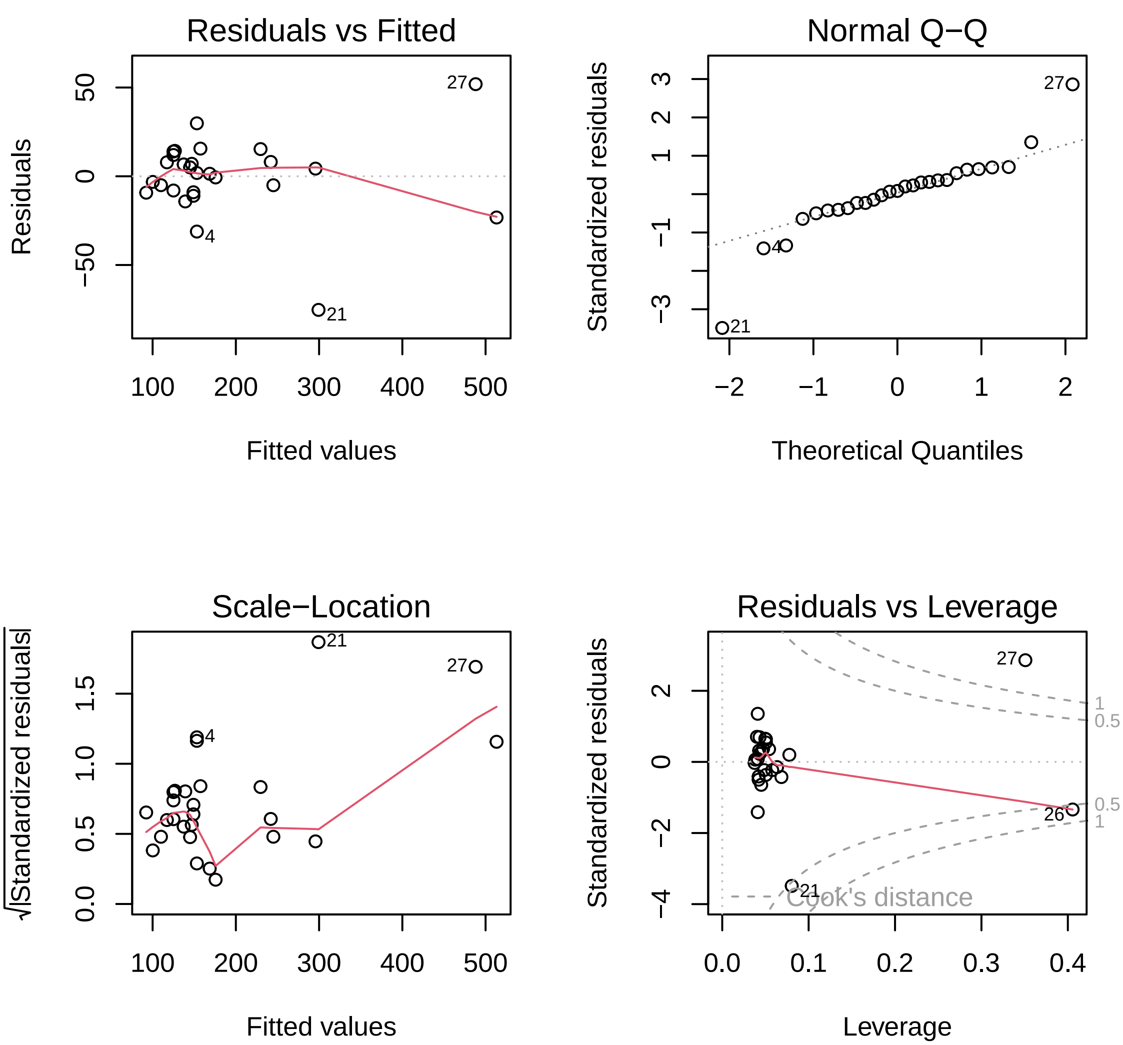
Même avec 4 points, on arrive pas a avoir des bonnes résultats. Donc on assume que cette stratégie ne va pas marcher. On vérifie par le graphe des résidus x régression qu'il est possible qu'il ait une rélation quadatique. On l'étude.

```r
sqr_table = prix_table
sqr_table['m2'] = sqr_table$m2^2

model_sqr = lm(prix~., data = sqr_table)

summary(model_sqr)
```

```
##
## Call:
## lm(formula = prix ~ ., data = sqr_table)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -75.303  -8.520   1.844  10.030  51.897
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.964e+01  6.264e+00   12.71 2.06e-12 ***
## m2          1.500e-02  6.294e-04   23.84  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.51 on 25 degrees of freedom
## Multiple R-squared:  0.9579, Adjusted R-squared:  0.9562
## F-statistic: 568.2 on 1 and 25 DF,  p-value: < 2.2e-16
```

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

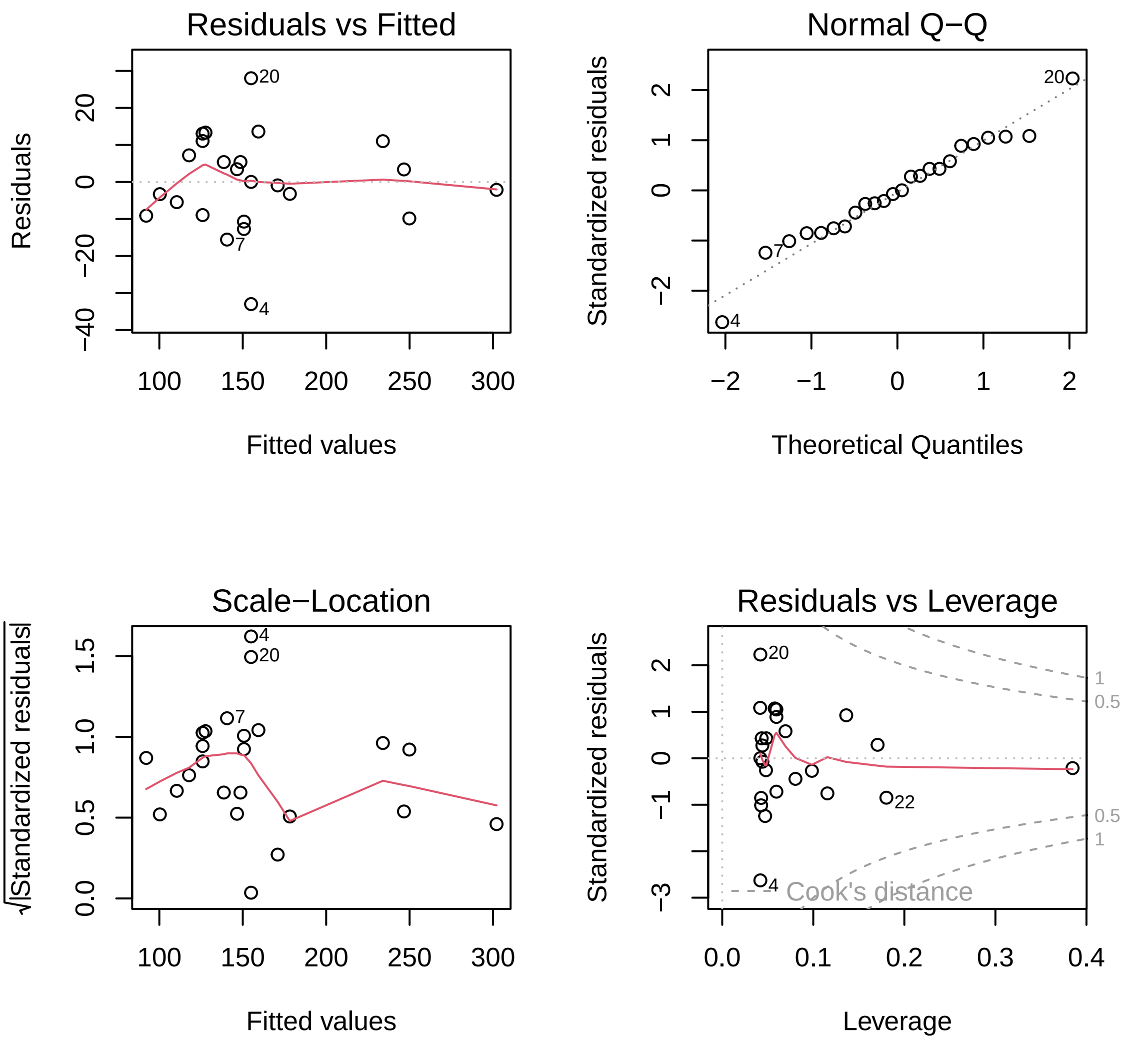


On observe qu'il a des points que a être enlevés (27,16 et 21) avec la rélation quadratique. On les enlève.

```r
filtered_sqr_table = sqr_table[-c(27, 26, 21),]
model_sqr2 = lm(prix~., data = filtered_sqr_table)
summary(model_sqr2)
```

```
##
## Call:
## lm(formula = prix ~ ., data = filtered_sqr_table)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.984  -8.985  -0.453   8.148  28.016
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.088274   4.920870   16.07 1.22e-13 ***
## m2           0.015489   0.000811   19.10 3.49e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.82 on 22 degrees of freedom
## Multiple R-squared:  0.9431, Adjusted R-squared:  0.9405
## F-statistic: 364.7 on 1 and 22 DF,  p-value: 3.487e-15
```

```r
par(mfrow=c(2, 2))
plot(model_sqr2)
```




```
nouvelle_apart_sqr = data.frame(m2 = 90*90)
prediction_IP = data.frame(predict(model_sqr2, newdata = nouvelle_apart_sqr,
                                   interval = 'prediction', level = 0.95))
prediction_IP
```

```
##        fit      lwr      upr
## 1 204.5488 176.9511 232.1466
```

En combinant, on arrive a une meilleur modèle.

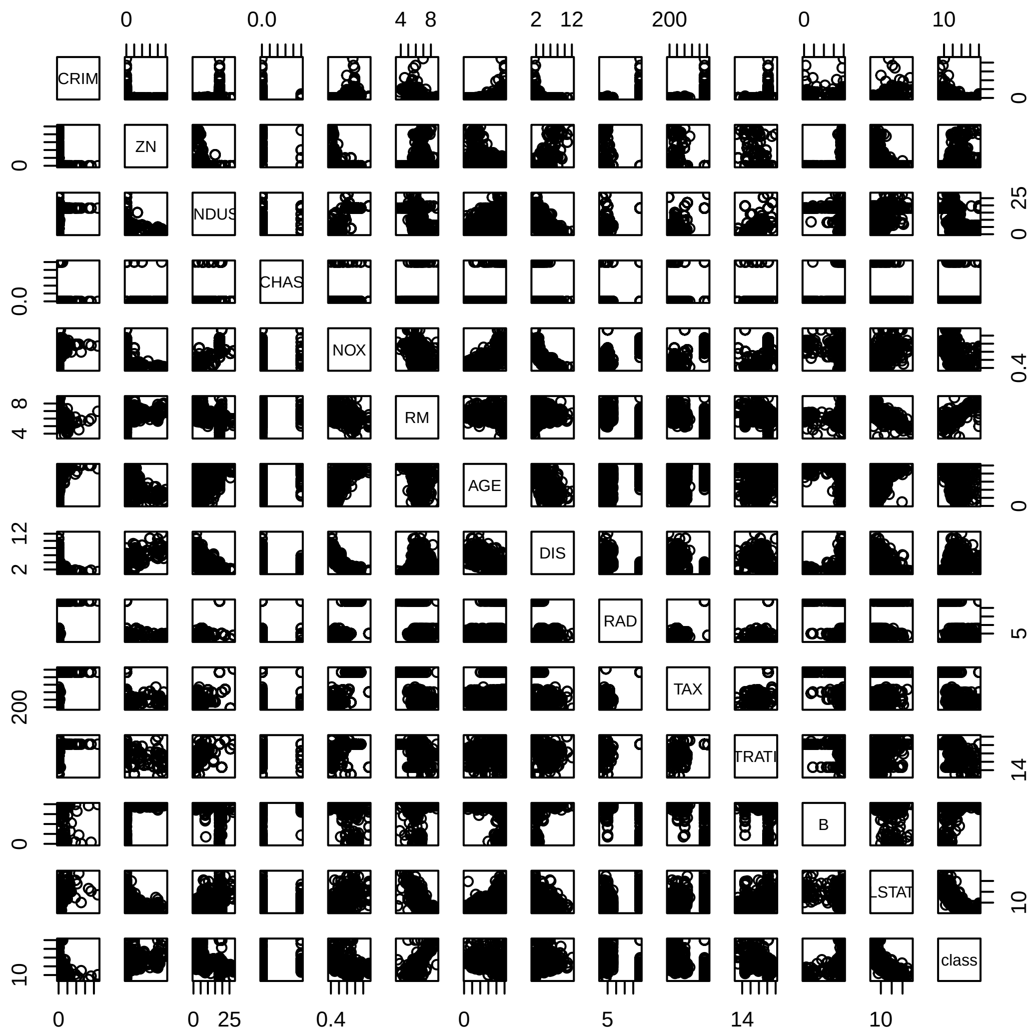
On étude après quel effet ce changement aurait sur la valeur de la prédiction faite. On arrive aux suivantes valeurs : $204.5488 \pm [176.9511; 232.1466]$ (par rapport à $223,915 \pm [154.3086; 293.534]$ avant). On observe que l'incertitude a beaucoup diminué, cela qui indique une précision plus importante du modèle actuelle.

Exercice 02 - La valeur des logements des villes aux alentours de Boston

```
logements_table = read.table(file = "housing.txt", header = TRUE)
head(logements_table)
```

```
##    CRIM ZN INDUS CHAS   NOX    RM  AGE   DIS RAD TAX PTRATIO       B LSTAT
## 1 0.006 18  2.31    0 0.538 6.575 65.2 4.090   1 296    15.3 396.898  4.98
## 2 0.027  0  7.07    0 0.469 6.421 78.9 4.967   2 242    17.8 396.898  9.14
## 3 0.027  0  7.07    0 0.469 7.185 61.1 4.967   2 242    17.8 392.828  4.03
## 4 0.032  0  2.18    0 0.458 6.998 45.8 6.062   3 222    18.7 394.629  2.94
## 5 0.069  0  2.18    0 0.458 7.147 54.2 6.062   3 222    18.7 396.898  5.33
## 6 0.030  0  2.18    0 0.458 6.430 58.7 6.062   3 222    18.7 394.119  5.21
##   class
## 1  24.0
## 2  21.6
## 3  34.7
## 4  33.4
## 5  36.2
## 6  28.7
```

```
pairs(logements_table)
```




```r
jpeg("Plot4.jpeg", width = 30, height = 30, units = 'cm', res = 600)
# plot(logements_table$m2, prix_table$prix)
```

1. Quelle est la part de variance expliquée par ce modèle ?

On obtient un modèle avec un R^2 de 74,06%. C'est-à-dire que la part de variance expliquée par ce modèle est de 74,06%.

```r
modelb = lm(class~., data = logements_table)
summary(modelb)
```

```
##
## Call:
## lm(formula = class ~ ., data = logements_table)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5940  -2.7295  -0.5179   1.7767  26.1987
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## CRIM        -1.080e-01  3.287e-02  -3.287 0.001087 **
## ZN           4.642e-02  1.373e-02   3.381 0.000779 ***
## INDUS        2.055e-02  6.150e-02   0.334 0.738352
## CHAS         2.687e+00  8.616e-01   3.119 0.001924 **
## NOX         -1.777e+01  3.820e+00  -4.651 4.24e-06 ***
## RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## AGE          6.915e-04  1.321e-02   0.052 0.958274
## DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
## PTRATIO     -9.528e-01  1.308e-01  -7.283 1.31e-12 ***
## B            9.311e-03  2.686e-03   3.467 0.000573 ***
## LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

2. Le modèle de régression est-il significatif dans son ensemble (prendre un risque de première espèce $\alpha = 1\%$) ? Donner l'hypothèse H0, la statistique du test, sa loi sous H0 et la conclusion.

H_0 : Les variables n'a pas un relation lineére avec le variable réponse(classe). Pour cette hypothèse, on utilise la valeur $\alpha = 1$, et si la p_{value} est inferieur a α, on rejjete H_0. On verifie que $p - value < \alpha = 1$ à partir de la statistique de Fischer, donc on rejjete H0 et on arrive à la conclusion que le modèle est significatif, pour cette risque de première espèce.

3. Quelles sont les variables significatives (prendre un risque de première espèce $\alpha = 1\%$) ? Est-on sûr qu'il n'y en a pas d'autres ?

Pour identifier les variables significatives, on procède par deux stratégies: AIC et BIC.

```
slm_AIC <- step(modelb, direction="backward", k = 2)
```

```
## Start:  AIC=1589.64
## class ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - AGE      1      0.06 11079 1587.6
## - INDUS    1      2.52 11081 1587.8
## <none>                 11079 1589.6
```

```
## - CHAS     1    218.99 11298 1597.5
## - TAX      1    242.23 11321 1598.6
## - CRIM     1    243.23 11322 1598.6
## - ZN       1    257.47 11336 1599.3
## - B        1    270.61 11349 1599.8
## - RAD      1    479.13 11558 1609.1
## - NOX      1    487.19 11566 1609.4
## - PTRATIO  1   1194.28 12273 1639.4
## - DIS      1   1232.42 12311 1641.0
## - RM       1   1871.33 12950 1666.6
## - LSTAT    1   2410.79 13490 1687.3
##
## Step:  AIC=1587.64
## class ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - INDUS    1      2.52 11081 1585.8
## <none>                 11079 1587.6
## - CHAS     1    219.93 11299 1595.6
## - TAX      1    242.21 11321 1596.6
## - CRIM     1    243.21 11322 1596.6
## - ZN       1    260.30 11339 1597.4
## - B        1    272.24 11351 1597.9
## - RAD      1    481.07 11560 1607.2
## - NOX      1    520.90 11600 1608.9
## - PTRATIO  1   1200.28 12279 1637.7
## - DIS      1   1352.27 12431 1643.9
## - RM       1   1959.56 13038 1668.0
## - LSTAT    1   2718.84 13798 1696.7
##
## Step:  AIC=1585.76
## class ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## - CHAS     1    227.23 11309 1594.0
## - CRIM     1    245.38 11327 1594.8
## - ZN       1    257.80 11339 1595.4
## - B        1    270.80 11352 1596.0
## - TAX      1    273.59 11355 1596.1
## - RAD      1    500.90 11582 1606.1
## - NOX      1    541.95 11623 1607.9
## - PTRATIO  1   1206.51 12288 1636.0
## - DIS      1   1448.96 12530 1645.9
## - RM       1   1963.67 13045 1666.3
## - LSTAT    1   2723.45 13805 1695.0
```

```
slm_BIC = step(modelb, direction="backward", k = log(nrow(logements_table)))
```

```
## Start:  AIC=1648.81
## class ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
```

```
##
##           Df Sum of Sq   RSS    AIC
## - AGE      1      0.06 11079 1642.6
## - INDUS    1      2.52 11081 1642.7
## <none>                 11079 1648.8
## - CHAS     1    218.99 11298 1652.5
## - TAX      1    242.23 11321 1653.5
## - CRIM     1    243.23 11322 1653.6
## - ZN       1    257.47 11336 1654.2
## - B        1    270.61 11349 1654.8
## - RAD      1    479.13 11558 1664.0
## - NOX      1    487.19 11566 1664.4
## - PTRATIO  1   1194.28 12273 1694.4
## - DIS      1   1232.42 12311 1696.0
## - RM       1   1871.33 12950 1721.6
## - LSTAT    1   2410.79 13490 1742.2
##
## Step:  AIC=1642.59
## class ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - INDUS    1      2.52 11081 1636.5
## <none>                 11079 1642.6
## - CHAS     1    219.93 11299 1646.3
## - TAX      1    242.21 11321 1647.3
## - CRIM     1    243.21 11322 1647.3
## - ZN       1    260.30 11339 1648.1
## - B        1    272.24 11351 1648.7
## - RAD      1    481.07 11560 1657.9
## - NOX      1    520.90 11600 1659.6
## - PTRATIO  1   1200.28 12279 1688.4
## - DIS      1   1352.27 12431 1694.6
## - RM       1   1959.56 13038 1718.8
## - LSTAT    1   2718.84 13798 1747.4
##
## Step:  AIC=1636.48
## class ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1636.5
## - CHAS     1    227.23 11309 1640.5
## - CRIM     1    245.38 11327 1641.3
## - ZN       1    257.80 11339 1641.9
## - B        1    270.80 11352 1642.5
## - TAX      1    273.59 11355 1642.6
## - RAD      1    500.90 11582 1652.6
## - NOX      1    541.95 11623 1654.4
## - PTRATIO  1   1206.51 12288 1682.5
## - DIS      1   1448.96 12530 1692.4
## - RM       1   1963.67 13045 1712.8
## - LSTAT    1   2723.45 13805 1741.5
```

Par les deux méthodes, on arrive au même emsamble de 11 variables significatives CRIM, ZN, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, B et LSTAT.

C'est-à-dire, que les variables INDUS et AGE ne sont pas significatives et donc ont été enlevés du modèle