Application de la régression linéaire multiple à des données financières macro-industrielles chinoises

Sommaire

L'objectif de ce document est de présenter la mise en œuvre, sous R, d'une regression linéaire multiple appliquée à des données financières macro-industrielles chinoises. L'étude présentée ici n'est toutefois pas complète, dans la mesure ou l'objectif qui est celui d'aboutir à un modèle LINEAIRE exploitable pour l'explication et la prédiction de la variable dépendente, n'a pas atteint. Ce document se veut donc etre une description des opérations entreprises pour tenter d'aboutir à un tel modèle.

---

1 Description des données traitées

Les données sont issues des rapports du Bureau national des statistiques de Chine (NBSC) contenant des mesures de huit indicateurs macro-industriels sur la période 1998-2017, utiles pour dresser un aperçu de la croissance économique du pays. Elles ont été récupérées sur la plateforme Kaggle (lien vers le jeu de données sur Kaggle).

Le jeu de données originel croise en colonnes 9 variables et en lignes 218 observations, toutefois celui-ci contient des données manquantes (NA), particulièrement pour les mesures entre 1998 et 2000, les observations correspondantes ont donc été omises. Dans ce qui suit, le jeu de données traité est donc de taille de taille 157 x 9.

Table 1: Jeu de données Chinese Macro-industrial finace (les 8 premières observations)

Date	Financial.Costs..Accumulated.Value	Total.Profits..Accumulated.Value	Power.Generation..Accumulated.Value	Inventories..Accumulated.Value	Total.Assets..Accumulated.Value	Interest.Expenses..Accumulated.Value	Administrative.Expenses..Accumulated.Value	Selling.Expenses..Accumulated.Value
2017-10-30	10848.4	62450.8	0	113787.4	1114876	9545.6	36705.9	26319.1
2017-09-29	9787.9	55846.0	0	111925.5	1104710	8588.7	32946.7	23680.9
2017-08-30	8641.4	49213.5	0	111312.0	1092301	7525.9	28947.1	20905.5
2017-07-30	7492.5	42481.2	0	110233.4	1081202	6581.6	25161.6	18179.8
2017-06-29	6409.2	36337.5	0	108929.2	1072033	5653.3	21454.6	15535.7
2017-05-30	5192.1	29047.6	0	109141.2	1062934	4575.4	17396.3	12558.9
2017-04-29	4121.4	22780.3	0	108113.5	1054212	3649.6	13786.5	9904.7
2017-03-30	3080.7	17043.0	0	106122.9	1041788	2720.1	10309.1	7360.3

1.1 Description des variables

Dans ce travail de modélisation, la variable "Date" n'est pas prise en compte, la variable expliquée est "Financial.Costs..Accumulated.Value", et ce, dans la perspective d'une part d'expliquer les facteurs qui participent aux coûts financiers afin les controler de sorte à minimser ces couts et d'autre part de disposer d'un modèle qui permette de les prédire (les couts financiers) pour anticiper les dépenses. Toutes les autres variables sont initialement prises comme variables explicatives.

Table 2: Description des variables du jeu de données

Variable	Description
Date	Date de l'observation
Financial.Costs..Accumulated.Value	Les coûts financiers
Total.Profits..Accumulated.Value	Bénéfices totaux
Power.Generation..Accumulated.Value	La production d'énergie
Inventories..Accumulated.Value	Stocks
Total.Assets..Accumulated.Value	Actif total
Interest.Expenses..Accumulated.Value	Frais d'intérêt
Administrative.Expenses..Accumulated.Value	Dépenses administratives
Selling.Expenses..Accumulated.Value	Frais de vente

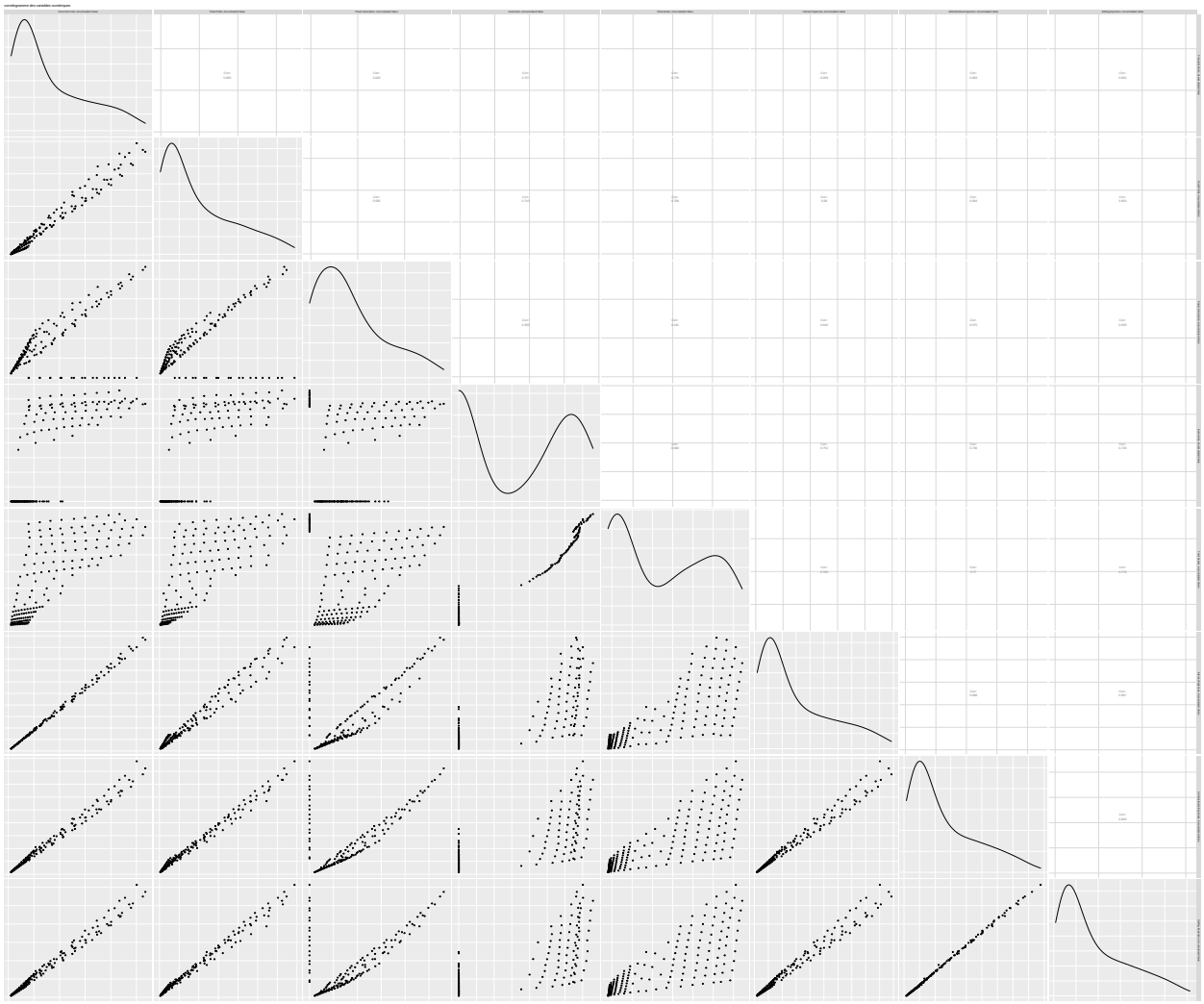
1.2 Analyses préliminaires (Corrélations et distributions des variables)

Coefficients de corrélation entre le variables

```
##                                          Financial.Costs..Accumulated.Value
## Financial.Costs..Accumulated.Value                                 1.0000000
## Total.Profits..Accumulated.Value                                   0.9847974
## Power.Generation..Accumulated.Value                                0.6259008
## Inventories..Accumulated.Value                                     0.7567645
## Total.Assets..Accumulated.Value                                    0.7780268
## Interest.Expenses..Accumulated.Value                               0.9988070
## Administrative.Expenses..Accumulated.Value                         0.9926011
## Selling.Expenses..Accumulated.Value                                0.9920928
##                                          Total.Profits..Accumulated.Value
## Financial.Costs..Accumulated.Value                              0.9847974
## Total.Profits..Accumulated.Value                                1.0000000
## Power.Generation..Accumulated.Value                             0.5848858
## Inventories..Accumulated.Value                                  0.7427434
## Total.Assets..Accumulated.Value                                 0.7681612
## Interest.Expenses..Accumulated.Value                            0.9804094
## Administrative.Expenses..Accumulated.Value                      0.9936430
## Selling.Expenses..Accumulated.Value                             0.9933104
##                                          Power.Generation..Accumulated.Value
## Financial.Costs..Accumulated.Value                                 0.6259008
## Total.Profits..Accumulated.Value                                   0.5848858
## Power.Generation..Accumulated.Value                                1.0000000
## Inventories..Accumulated.Value                                     0.2849396
## Total.Assets..Accumulated.Value                                    0.2417790
## Interest.Expenses..Accumulated.Value                               0.6447162
## Administrative.Expenses..Accumulated.Value                         0.5749610
## Selling.Expenses..Accumulated.Value                                0.5683780
##                                          Inventories..Accumulated.Value
## Financial.Costs..Accumulated.Value                            0.7567645
## Total.Profits..Accumulated.Value                              0.7427434
## Power.Generation..Accumulated.Value                           0.2849396
## Inventories..Accumulated.Value                                1.0000000
## Total.Assets..Accumulated.Value                               0.9679577
## Interest.Expenses..Accumulated.Value                          0.7518333
## Administrative.Expenses..Accumulated.Value                    0.7356854
## Selling.Expenses..Accumulated.Value                           0.7386920
```

```
##                                                 Total.Assets..Accumulated.Value
## Financial.Costs..Accumulated.Value                                    0.7780268
## Total.Profits..Accumulated.Value                                      0.7681612
## Power.Generation..Accumulated.Value                                   0.2417790
## Inventories..Accumulated.Value                                        0.9679577
## Total.Assets..Accumulated.Value                                       1.0000000
## Interest.Expenses..Accumulated.Value                                  0.7690917
## Administrative.Expenses..Accumulated.Value                            0.7696686
## Selling.Expenses..Accumulated.Value                                   0.7759753
##                                                 Interest.Expenses..Accumulated.Value
## Financial.Costs..Accumulated.Value                                         0.9988070
## Total.Profits..Accumulated.Value                                           0.9804094
## Power.Generation..Accumulated.Value                                        0.6447162
## Inventories..Accumulated.Value                                             0.7518333
## Total.Assets..Accumulated.Value                                            0.7690917
## Interest.Expenses..Accumulated.Value                                       1.0000000
## Administrative.Expenses..Accumulated.Value                                 0.9880290
## Selling.Expenses..Accumulated.Value                                        0.9871854
##                                                 Administrative.Expenses..Accumulated.Value
## Financial.Costs..Accumulated.Value                                               0.9926011
## Total.Profits..Accumulated.Value                                                 0.9936430
## Power.Generation..Accumulated.Value                                              0.5749610
## Inventories..Accumulated.Value                                                   0.7356854
## Total.Assets..Accumulated.Value                                                  0.7696686
## Interest.Expenses..Accumulated.Value                                             0.9880290
## Administrative.Expenses..Accumulated.Value                                       1.0000000
## Selling.Expenses..Accumulated.Value                                              0.9994890
##                                                 Selling.Expenses..Accumulated.Value
## Financial.Costs..Accumulated.Value                                        0.9920928
## Total.Profits..Accumulated.Value                                          0.9933104
## Power.Generation..Accumulated.Value                                       0.5683780
## Inventories..Accumulated.Value                                            0.7386920
## Total.Assets..Accumulated.Value                                           0.7759753
## Interest.Expenses..Accumulated.Value                                      0.9871854
## Administrative.Expenses..Accumulated.Value                                0.9994890
## Selling.Expenses..Accumulated.Value                                       1.0000000
```

Corrélogramme des variables



2 Mise en œuvre de la régression linéaire multiple

L'éstimation d'un modèle de régression linéaire multiple dans ce qui suit, a été faite selon le processus suivant: L'éstimation d'un premier modèle prenant en compte toutes les sept variables explicatives, l'élimination des variables non significatives, la vérification du fait que le modèle, ainsi obtenu, est bien le meme que celui retourné par la fonction step (qui automatise l'approche descendente d'élimination de variables selon leur significativité et selon l'AIC).Nous considérons celui-ci comme un modèle de base que nous cherchons à améliorer. Dans cette perspective, d'autres modèles sont estimés, cette fois en incluant ou non des variables sur base de leurs corrélations avec la variable expliquée et sur base du VIF. Nous confirmons finalement que le modèle de base est relativement le meilleur.

Sous R, les coefficients sont par défaut estimés par la méthode des moindres carrés.

2.1 Premier modèle

```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ ., data = donnees)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.08  -49.50   16.13   44.27  204.39
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                              2.1400888 30.4122514   0.070
## Total.Profits..Accumulated.Value        -0.0178714  0.0046032  -3.882
## Power.Generation..Accumulated.Value      0.0027383  0.0010918   2.508
## Inventories..Accumulated.Value           0.0028031  0.0008995   3.116
## Total.Assets..Accumulated.Value         -0.0001639  0.0001305  -1.256
## Interest.Expenses..Accumulated.Value     0.7513404  0.0260077  28.889
## Administrative.Expenses..Accumulated.Value  0.0295174  0.0269162   1.097
## Selling.Expenses..Accumulated.Value      0.1346309  0.0394199   3.415
##                                          Pr(>|t|)
## (Intercept)                              0.943994
## Total.Profits..Accumulated.Value         0.000155 ***
## Power.Generation..Accumulated.Value      0.013212 *
## Inventories..Accumulated.Value           0.002197 **
## Total.Assets..Accumulated.Value          0.211157
## Interest.Expenses..Accumulated.Value      < 2e-16 ***
## Administrative.Expenses..Accumulated.Value 0.274568
## Selling.Expenses..Accumulated.Value      0.000821 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.31 on 149 degrees of freedom
## Multiple R-squared:  0.9993, Adjusted R-squared:  0.9992
## F-statistic: 2.859e+04 on 7 and 149 DF,  p-value: < 2.2e-16

## [1]    8.000 1445.255
```

2.1.1 Lecture des résultats

La première ligne rappelle le modèle estimé.
La section "residuals" donne les quartiles de la variable des résidus. La section "Coefficients" retourne :

- les coefficients estimés (Estimate)
- l'écart type de l'erreur d'estimation du coefficient (Std.Error)
- la valeur de la statistique du test de significativité du coefficient.
- la p-valeur de ce test notée Pr(>|t|) qui est la probabilité d'obtenir une t-valeur aussi élevée ou supérieure à celle observée (t-value) sous H0 par le seul effet du hasard.
- des étoiles de significativité dont l'interprétation est donnée dans la ligne d'après.

2.1.1.1 Test de significativité des coefficients

Le test de significativité des coefficients teste l'hypothèse H0: "Le coefficient est nul" contre H1:" Le coefficient est non nul".

Les résultats de ce test peuvent être lus de trois manières:

Avec la statistique du test :

La valeur de la statistique du test de significativité des coefficients peut être lue de la colonne t value. Il s'agit, pour interpréter le résultat, de comparer cette valeur à la valeur critique de la loi de Student (tabulée) à (nb.obsv - nb.variables - 1) degrés de liberté et au seuil choisi.

Avec la p-valeur du test :

Nous usons ici des p-valeur du test pour en interpréter les résultats. Mis à part ceux des variables "Total.Assets..Accumulated.Value" et "Administrative.Expenses..Accumulated.Value" , tous les coefficients ont des p-valeur du test inférieures à 0.05, l'hypothèse Nulle dans ce cas est rejetée et les coefficients sont considérés comme non nuls et donc les variables associées significatives. Les p-valeur du test de significativité des coefficients des variables "Total.Assets..Accumulated.Value" et "Administrative.Expenses..Accumulated.Value" sont supérieures à 0.1, le résultat du test n'est donc pas statistiquement significatif, l'on ne peut dans ce cas rejeter H0, et concluons donc que ces variables ne sont pas significatives, i.e, ne contribuent probablement pas à l'explication des coûts financiers.

Avec les étoiles de significativité

L'on peut arriver aux mêmes conclusions plus simplement en se fiant aux étoiles de significativité : les variables associées aux coefficients marqués par "***" sont significatives à un seuil de 0.1% ,celles associées à ceux marqués par "**"le sont à un seuil de 1%,celles associées à ceux marqués par "*" le sont à un seuil de 5%, celles associées à ceux marqués par "." le sont à un seuil de 10% et enfin celles associées à ceux marqués par un " " (vide) ne sont pas significatives.

2.1.1.2 Test de significativité globale du modèle

Les trois dernières lignes indiquent dans l'ordre :

- la racine de la moyenne des erreurs au carré.(l'écart-type des erreurs, ayant divisé sur nb.obsv-nb.variables -1)
- le nombre de degrés de libértés, tel que Nb DDL = Nb Observations - Nb variables explicatives - 1 = 157 - 7 -1 = 149.
- la valeur du coefficient de détérmination R2 qui est une mesure du pouvoir prédictif du modèle.
- la valeur du coefficient de détérmination ajusté (qui prend en considération le nombre de degrés de liberté du modèle et qui est plus fiable que R2, ce dernier augmentant dans tout les cas (meme en ajoutant au modèle des variables non significaives))
- la valeur de la statistique de Fisher du test de la significativité globale du modèle.
- le nombre de variables explicatives et, une fois de plus, de degrès de libérté.
- la p-valeur du test de Fisher de sgnificativité globale du modèle.

La valeur du critère d'information d'Akaiké (AIC) est aussi affichée. Celle-ci permet de mesurer la qualité d'un modèle (par rapport à un autre).

Nous lisons pour ce modèle un R2 ajusté de 99.92 %, ce qui signifierait que le modèle explique 99% de la variabilité du coût financier; ce qui est une valeur très élevée, toutefois, une valeur élevée du coefficient de détermination ne signifie pas toujours une bonne qualité du modèle mais peut signifier un modèle bruité(qui inclurait des variables qui n'expliquent pas forcément la variable dépendante). Aussi, la p-valeur du modèle est inférieure à 2.2e-16 ce qui indique qu'on ne peut accepter l'hypothèse H0 du test de Fisher selon laquelle "tous les coefficients du modèle seraient nuls", le modèle est donc globalement significatif (contient au moins un coefficient significativement différent de zéro).

2.2 Elimination des variables non signficatives

Dans ce qui suit nous procédons à l'élimination des variables non significatives, une à une, et réestimons le modèle après chaque élimination.

```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value +
##     Power.Generation..Accumulated.Value + Inventories..Accumulated.Value +
##     Total.Assets..Accumulated.Value + Interest.Expenses..Accumulated.Value +
##     Selling.Expenses..Accumulated.Value, data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -358.83  -49.98   13.86   45.33  205.36
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       19.3268158 26.0807571   0.741 0.459829
## Total.Profits..Accumulated.Value  -0.0172525  0.0045716  -3.774 0.000231 ***
## Power.Generation..Accumulated.Value  0.0026372  0.0010886   2.422 0.016608 *
## Inventories..Accumulated.Value     0.0030444  0.0008727   3.488 0.000638 ***
## Total.Assets..Accumulated.Value   -0.0002128  0.0001227  -1.734 0.085025 .
## Interest.Expenses..Accumulated.Value  0.7554840  0.0257491  29.340  < 2e-16 ***
## Selling.Expenses..Accumulated.Value   0.1735335  0.0172018  10.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.38 on 150 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 3.33e+04 on 6 and 150 DF,  p-value: < 2.2e-16

## [1]    7.000 1444.517
```

On constate que l'AIC a diminué. L'erreur moyenne du modèle a légèrement augmenté sans que ceci soit significatif. On constate par ailleurs que la p-valeur de la variable "Total.Assets..Accumulated.Value" a diminué, la rendant désormais significative si l'on se fixe un seuil de 10 %, nous ne la supprimons donc pas du modèle. Pour confirmer la décision d'arrêter le processus descendant d'élimination des variables nous estimons le modèle avec la fonction "step", qui automatise ce processus [1].

2.3 Modèle éstimé par la fonction "step"

Le modèle, ci-dessous, retourné par la fonction "step" est bien le même que celui obtenu précédemment (modèle sans la variable " Administrative.Expenses..Accumulated.Value"). Nous considérons ce modèle comme modèle de base et essaierons dans ce qui suit d'en trouver d'autres meilleurs.

```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value +
##     Power.Generation..Accumulated.Value + Inventories..Accumulated.Value +
##     Total.Assets..Accumulated.Value + Interest.Expenses..Accumulated.Value +
##     Selling.Expenses..Accumulated.Value, data = donnees)
##
## Residuals:
```

[1] Ceci n'est pas prétendu être une étape de la pratique de la régression linéaire et a été fait uniquement par manque de certitude quant à la décision d'arrêt du processus descendent (le choix du seuil de significativité étant arbitraire). Et toujours dans cette optique de vérification, un modèle a été estimé en poursuivant le processus descendent et en supprimant la variable "Total.Assets..Accumulated.Value" et a été, en effet, trouvé d'un AIC supérieur au modèle de base.

```
##     Min      1Q  Median      3Q     Max
## -358.83  -49.98   13.86   45.33  205.36
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        19.3268158 26.0807571   0.741 0.459829
## Total.Profits..Accumulated.Value   -0.0172525  0.0045716  -3.774 0.000231 ***
## Power.Generation..Accumulated.Value  0.0026372  0.0010886   2.422 0.016608 *
## Inventories..Accumulated.Value       0.0030444  0.0008727   3.488 0.000638 ***
## Total.Assets..Accumulated.Value    -0.0002128  0.0001227  -1.734 0.085025 .
## Interest.Expenses..Accumulated.Value  0.7554840  0.0257491  29.340  < 2e-16 ***
## Selling.Expenses..Accumulated.Value   0.1735335  0.0172018  10.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 97.38 on 150 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 3.33e+04 on 6 and 150 DF,  p-value: < 2.2e-16

## [1]    7.000 1444.517

##     Total.Profits..Accumulated.Value  Power.Generation..Accumulated.Value
##                             110.60007                              3.95291
##       Inventories..Accumulated.Value      Total.Assets..Accumulated.Value
##                              27.92074                             28.89358
## Interest.Expenses..Accumulated.Value  Selling.Expenses..Accumulated.Value
##                             111.51363                            277.62086
```

2.4 Modèle sans les variables les moins corrélées avec la variable expliquée

Nous avons, au début de cette étude, visualisé les corrélations entre les variables mais ne les avons jusque-là pas utilisées. Les variables non corrélées avec la variable expliquée peuvent être éliminées dès le début. Dans notre cas, en se référant au coefficient de corrélation de Pearson, toutes les variables sont linéairement fortement corrélées avec la variable expliquée. Nous nous proposons de visualiser les coefficients de corrélations partielles pour évaluer la corrélation entre une variable explicative et la variable expliquée en faisant abstraction de l'influence des autres variables explicatives sur cette corrélation, ceci est motivé par le fait que, dans le contexte des données traitées, les variables explicatives sont reliées entre elles. [2]
Ces coefficients révèlent finalement de faibles corrélations entre "Financial.Costs..Accumulated.Value" et les autres variables, sauf avec "Interest.Expenses..Accumulated.Value" avec laquelle elle a une corrélation de 0.92. Nous réestimons le modèle avec cette seule variable. Le modèle ainsi estimé a un R2 ajusté de 0.9976, ce qui signifie que la variable "Interest.Expenses..Accumulated.Value" a elle seule explique 99% de la variabilité de la variable que l'on cherche à expliquer. De ce fait, nous gardons aussi ce modèle pour les traitements à venir, bien qu'il ait des valeurs d'AIC et d'erreur moyenne supérieures à celle du modèle de base.

```
##             [,1]       [,2]         [,3]        [,4]         [,5]        [,6]
## [1,]  1.00000000 -0.3030947  0.201261510  0.24737471 -0.102338399  0.92114777
## [2,] -0.30309469  1.0000000  0.408011468  0.48924626 -0.366096625  0.11681926
## [3,]  0.20126151  0.4080115  1.000000000 -0.21389502 -0.003727562  0.08594077
## [4,]  0.24737471  0.4892463 -0.213895016  1.00000000  0.928733657 -0.05527167
## [5,] -0.10233840 -0.3660966 -0.003727562  0.92873366  1.000000000 -0.01626926
```

[2] La relation entendue ici n'est pas nécessairement la corrélation linéaire, mais la relation réelle existant entre ces variables comme étant des indicateurs économiques décrivant la même entité.

```
## [6,]  0.92114777  0.1168193  0.085940768 -0.05527167 -0.016269257  1.00000000
## [7,]  0.08947989  0.1434747 -0.100374123  0.21395120 -0.329350061 -0.02610491
## [8,]  0.26944441  0.3217416 -0.211804759 -0.51676684  0.526990556 -0.16532670
##             [,7]       [,8]
## [1,]  0.08947989  0.2694444
## [2,]  0.14347472  0.3217416
## [3,] -0.10037412 -0.2118048
## [4,]  0.21395120 -0.5167668
## [5,] -0.32935006  0.5269906
## [6,] -0.02610491 -0.1653267
## [7,]  1.00000000  0.8390401
## [8,]  0.83904011  1.0000000
```

```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ Interest.Expenses..Accumulated.Value,
##     data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -621.23  -53.22   -9.66   76.59  398.94
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       53.359568  20.807997   2.564   0.0113 *
## Interest.Expenses..Accumulated.Value  1.089099   0.004277 254.644   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 170.8 on 155 degrees of freedom
## Multiple R-squared:  0.9976, Adjusted R-squared:  0.9976
## F-statistic: 6.484e+04 on 1 and 155 DF,  p-value: < 2.2e-16
```

```
## [1]    2.000 1616.106
```

Enfin, nous intronduisons dans le modèle la variable “Total.Profits..Accumulated.Value” dont le coefficient de corrélation partielle est directement inférieur à celui de la variable déja présente dans le modèle, pour voir comment le modèle va évoluer.

L’AIC et l’erreur moyenne diminuent et le R2 ajusté aussi. [3]

```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ Interest.Expenses..Accumulated.Value +
##     Total.Profits..Accumulated.Value, data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -652.70  -51.39  -16.07   57.16  342.14
##
```

[3] N’ayant su interpréter ce résultat(l’AIC suggérant que la variable devrait être incluse et le R2 ajusté suggérant le contraire) et celui-ci étant le même pour toute autre variable (exceptée “Administrative.Expenses..Accumulated.Value”) que l’on introduirait dans le modèle, l’on se propose donc d’arrêter cette démarche, allant ultimement aboutir au modèle de base (si le critère optimisé est l’AIC).

```
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         74.67117   17.20979   4.339 2.58e-05 ***
## Interest.Expenses..Accumulated.Value  0.93596    0.01778  52.642  < 2e-16 ***
## Total.Profits..Accumulated.Value      0.02785    0.00317   8.785 2.85e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 139.9 on 154 degrees of freedom
## Multiple R-squared:  0.9984, Adjusted R-squared:  0.9984
## F-statistic: 4.84e+04 on 2 and 154 DF,  p-value: < 2.2e-16

## [1]    3.000 1554.323
```

2.5 Elimination des variables redondantes (sur base du VIF)

Nous essayons désormais d'estimer un meilleur modèle en usant de la mesure VIF des variables. Le Variation Inflation Factor (VIF) qui évalue si les facteurs sont corrélés les uns aux autres (présence de multi-colinéarité).

Une directive générale est qu'un VIF supérieur à 5 ou 10 est indicateur d'une multi-colinéarité élevée entre les variables explicatives.

```
##           Total.Profits..Accumulated.Value
##                                  112.287529
##        Power.Generation..Accumulated.Value
##                                    3.981287
##             Inventories..Accumulated.Value
##                                   29.698145
##            Total.Assets..Accumulated.Value
##                                   32.712229
##       Interest.Expenses..Accumulated.Value
##                                  113.918126
## Administrative.Expenses..Accumulated.Value
##                                 1339.324345
##        Selling.Expenses..Accumulated.Value
##                                 1459.898260
```

La variable présentant la valeur du VIF la plus elevée est "Selling.Expenses..Accumulated.Value" , toutefois, de part sa p-valeur, celle-ci est significative. L'on constate toutefois que la variable "Administrative.Expenses..Accumulated.Value" présente une valeur VIF presque tout aussi elevée et celle-ci est statistiquement non significative, l'on peut donc supposer que ces deux variables sont collinéaire et que toute la variabilité devant etre expliquée par "Administrative.Expenses..Accumulated.Value" est expliquée par "Selling.Expenses..Accumulated.Value". La mesure de la corrélation (dont les résultats sont donnés plus haut) revèle en effet un taux de corrélation de 0.9994890 entre ces deux variables. De ce fait, nous éliminons "Administrative.Expenses..Accumulated.Value" -décision qui concorde avec celle prise au premier modèle- nous retrouvons donc le modèle de base.

Itérons le processus et voyons maintenant les VIF de celui-ci.

```
##     Total.Profits..Accumulated.Value  Power.Generation..Accumulated.Value
##                               110.60007                               3.95291
##       Inventories..Accumulated.Value     Total.Assets..Accumulated.Value
##                                27.92074                              28.89358
## Interest.Expenses..Accumulated.Value  Selling.Expenses..Accumulated.Value
##                               111.51363                             277.62086
```

Les trois variables “Interest.Expenses..Accumulated.Value”, “Total.Profits..Accumulated.Value” et “Selling.Expenses..Accumulated.Value” présentent des valeurs VIF élevées et de même ordre de grandeur. L’on peut constater en effet d’après la mesure de corrélation qu’elles sont fortement corrélées. Nous nous proposons donc de voir les trois modèles ne contenant qu’une variable parmi celles-ci, à la fois et d’en retenir le meilleur.

Nous constatons toutefois que l’AIC augmente pour les trois modèles.

(Dans ce qui suit, seuls les statistiques globales du modèle et son AIC sont affichés.)

Modèle en ne gardant que “Selling.Expenses..Accumulated.Value”

Table 3: statistiques globales du modèle réestimé en ne gardant que ”Selling.Expenses..Accumulated.Value”

	standard.error	s31.df	R.squared	R.squared.adjusted	F.statistic
value	288.7948	5	0.9933145	0.9931386	5645.946
numdf	288.7948	152	0.9933145	0.9931386	4.000
dendf	288.7948	5	0.9933145	0.9931386	152.000

```
## [1]    5.000 1783.954
```

Modèle en ne gardant que “Interest.Expenses..Accumulated.Value”

Table 4: statistiques globales du modèle réestimé en ne gardant que ”Interest.Expenses..Accumulated.Value”

	standard.error	s32.df	R.squared	R.squared.adjusted	F.statistic
value	146.3662	5	0.9982827	0.9982375	22090.26
numdf	146.3662	152	0.9982827	0.9982375	4.00
dendf	146.3662	5	0.9982827	0.9982375	152.00

```
## [1]    5.000 1570.558
```

Modèle en ne gardant que “Total.Profits..Accumulated.Value”

Table 5: statistiques globales du modèle réestimé en ne gardant que ”Total.Profits..Accumulated.Value”

	standard.error	s33.df	R.squared	R.squared.adjusted	F.statistic
value	527.7444	5	0.9776745	0.9770869	1664.086
numdf	527.7444	152	0.9776745	0.9770869	4.000
dendf	527.7444	5	0.9776745	0.9770869	152.000

```
## [1]    5.000 1973.263
```

L’on procède dans ce qui suit à la suppression d’une seule variable à la fois parmi ces trois et l’on voit si les modèles obtenus sont meilleurs:

Modèle en éliminant la variable “Selling.Expenses..Accumulated.Value”

```
## [1]    6.000 1523.824
```

Le modèle est moins bon que le modèle de base (sur base de l’AIC).

Modèle en éliminant la variable “Interest.Expenses..Accumulated.Value”

Table 6: statistiques globales du modèle réestimé en éliminant "Selling.Expenses..Accumulated.Value"

	standard.error	s3.df	R.squared	R.squared.adjusted	F.statistic
value	125.7391	6	0.998741	0.9986993	23956.9
numdf	125.7391	151	0.998741	0.9986993	5.0
dendf	125.7391	6	0.998741	0.9986993	151.0

Table 7: statistiques globales du modèle réestimé en éliminant "Interest.Expenses..Accumulated.Value"

	standard.error	s4.df	R.squared	R.squared.adjusted	F.statistic
value	251.9482	6	0.9949451	0.9947777	5944.226
numdf	251.9482	151	0.9949451	0.9947777	5.000
dendf	251.9482	6	0.9949451	0.9947777	151.000

```
## [1]    6.000 1742.059
```

Le modèle est moins bon que le modèle de base (sur base de l'AIC).

Modèle en éliminant la variable "Total.Profits..Accumulated.Value"

Table 8: statistiques globales du modèle réestimé en éliminant "Total.Profits..Accumulated.Value"

	standard.error	s5.df	R.squared	R.squared.adjusted	F.statistic
value	101.5572	6	0.9991787	0.9991515	36740.09
numdf	101.5572	151	0.9991787	0.9991515	5.00
dendf	101.5572	6	0.9991787	0.9991515	151.00

```
## [1]    6.000 1456.758
```

Idem, Le modèle est moins bon que le modèle de base (sur base de l'AIC).

Les trois modèles ci-dessus desquels les variables de VIF important ont été supprimées une à la fois, ont tous un AIC supérieur à celui du modèle de base. Nous continuons dans ce qui suit l'élimination des variables suivantes, toujours dont le VIF 10, à savoir,"Inventories..Accumulated.Value" et "Total.Assets..Accumulated.Value".

Modèle en éliminant la variable "Total.Assets..Accumulated.Value"

Table 9: statistiques globales du modèle réestimé en éliminant "Total.Assets..Accumulated.Value"

	standard.error	s6.df	R.squared	R.squared.adjusted	F.statistic
value	98.02176	6	0.9992349	0.9992095	39440.4
numdf	98.02176	151	0.9992349	0.9992095	5.0
dendf	98.02176	6	0.9992349	0.9992095	151.0

```
## [1]    6.000 1445.632
```

Modèle en éliminant la variable "Inventories..Accumulated.Value"

```
## [1]    6.000 1454.764
```

Table 10: statistiques globales du modèle réestimé en éliminant "Inventories..Accumulated.Value"

	standard.error	s7.df	R.squared	R.squared.adjusted	F.statistic
value	100.9144	6	0.999189	0.9991622	37210.07
numdf	100.9144	151	0.999189	0.9991622	5.00
dendf	100.9144	6	0.999189	0.9991622	151.00

On constate que ces deux modèles ont également un AIC supérieur à celui du modèle de base.

Ceci laisse penser que le meilleur modèle est en fait celui obtenu en supprimant uniquement "Administrative.Expenses..Accumulated.Value" qui est le même modèle obtenu avec la fonction "step".

Ci-dessous sont résumées les tentatives d'estimation d'un meilleur modèle que le modèle de base (obtenu en éliminant la seule variable non significative trouvée):

- Elimination des variables sur base des coefficients de corrélations :
  - En éliminant les variables les moins corrélées avec la variable expliquée.
  - En ne gardant que la variable linéairement fortement corrélée avec la variable expliquée (sur base des coefficients de corrélations partielles) (1)
  - En introduisant dans le modèle (1) obtenu à l'étape ci-dessus, la variable suivante dont le coefficient de corrélation partielle est directement inférieur à celui de la variable déjà présente dans le modèle.
- Elimination des variables redondantes (selon le VIF):
  - En supprimant la variable non significative parmi les deux dont le VIF était très élevé (et du même ordre de grandeur).
  - En ne gardant qu'une des trois variables suivantes de VIF élevés et de même ordre de grandeur.
  - En supprimant une variable à la foi parmi ces trois et parmi les variables suivantes de VIF >10.

Dans tous les cas les modèles obtenus étaient moins bons (en terme d'AIC (mais également en terme d'erreur moyenne)) que le modèle de base. Dans les traitements qui suivent nous conservons donc le modèle de base et, comme intentionné plus haut, le modèle (1). Nous rappelons que le modèle de base est donné par :

" Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + Power.Generation..Accumulated.Value + Inventories..Accumulated.Value + Total.Assets..Accumulated.Value + Interest.Expenses..Accumulated.Value + Selling.Expenses..Accumulated.Value "

Tel que tous les coefficients sont significatifs et le modèle explique 98% de la variation des couts finaciers.

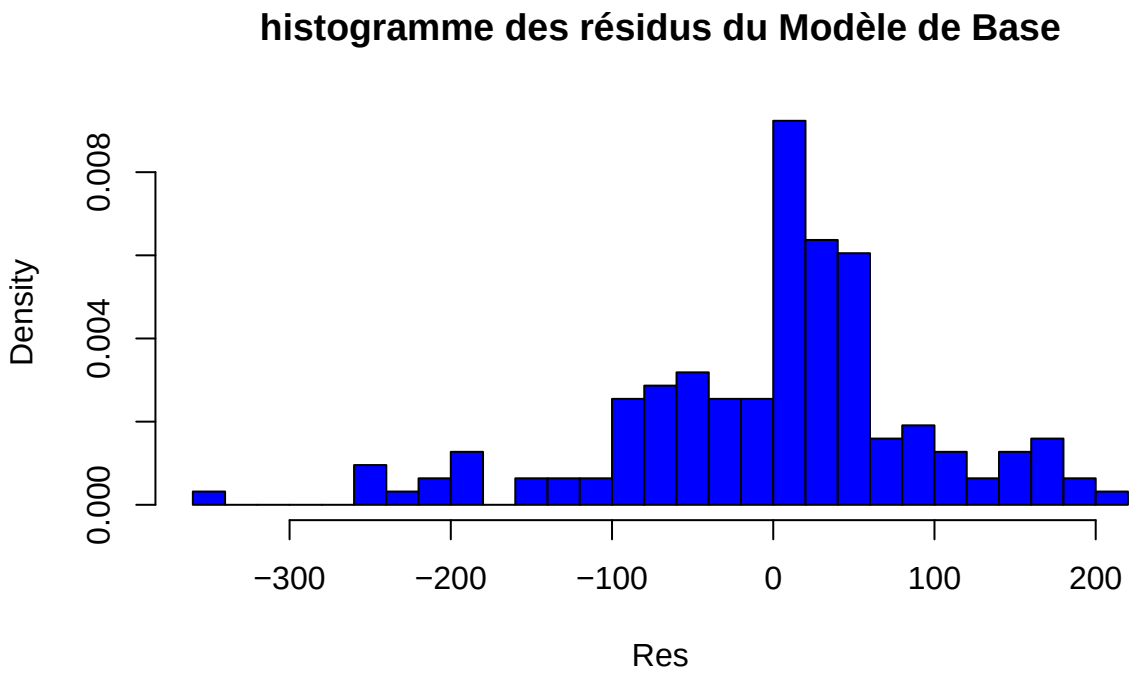
Et le modèle (1) est celui expliquant la variable dépendente par la seule variable lui étant linéairement fortement corréllée ("Interest.Expenses..Accumulated.Value") expliquant 99% de la variabilité des couts finaciers. Il est donné par:

"Financial.Costs..Accumulated.Value ~ Interest.Expenses..Accumulated.Value".

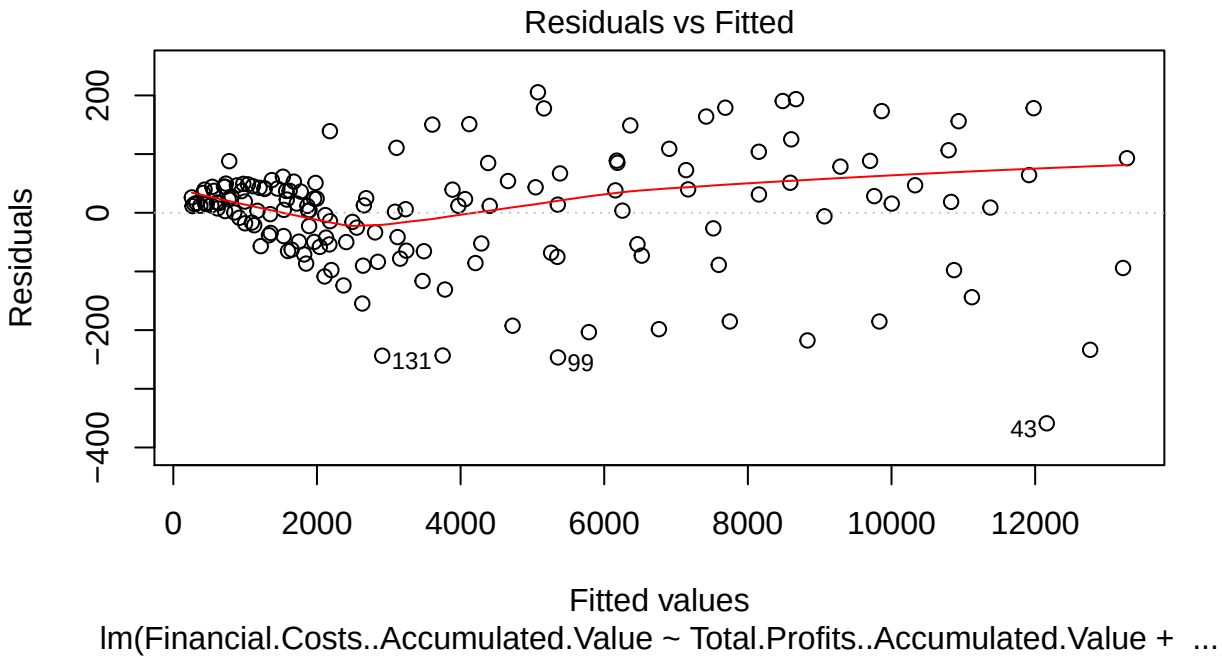
2.6 Vérification des hypothèses sur les résidus

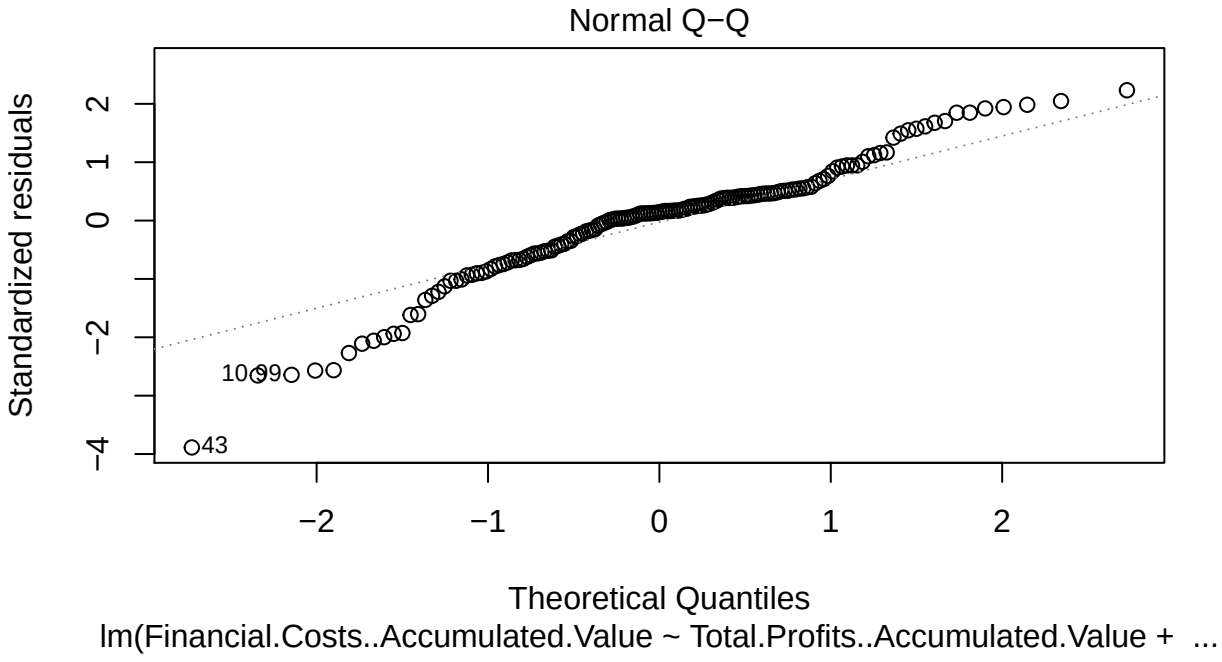
2.6.1 Graphiques et tests des résidus

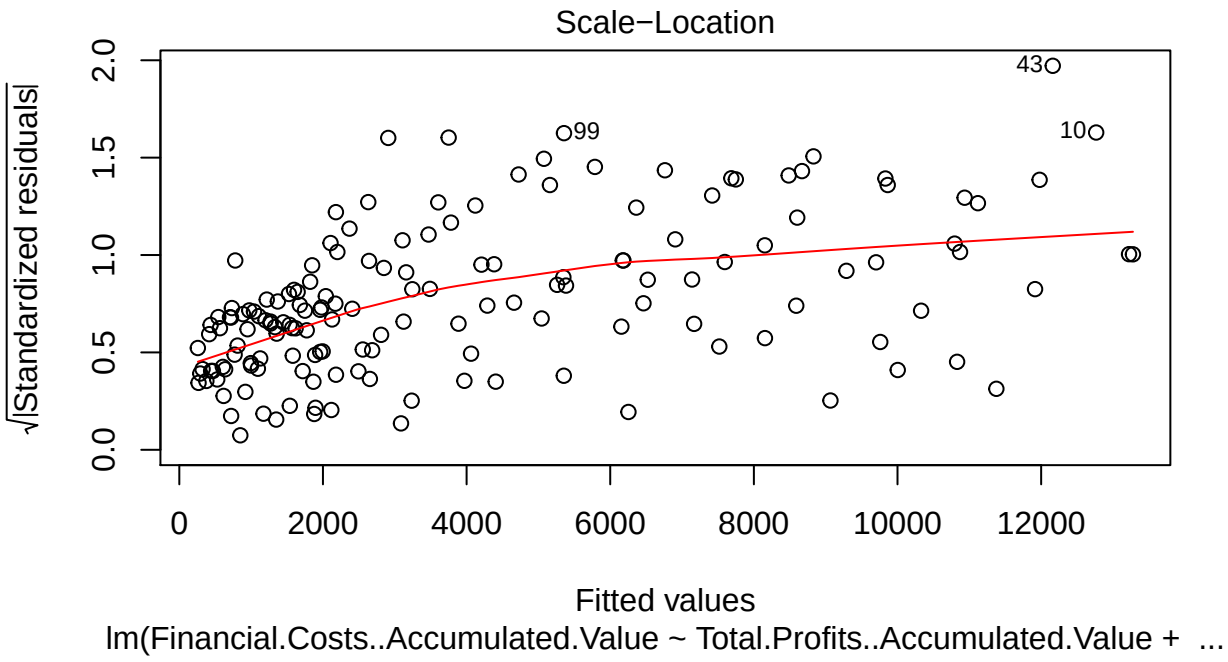
2.6.1.1 Modèle de base

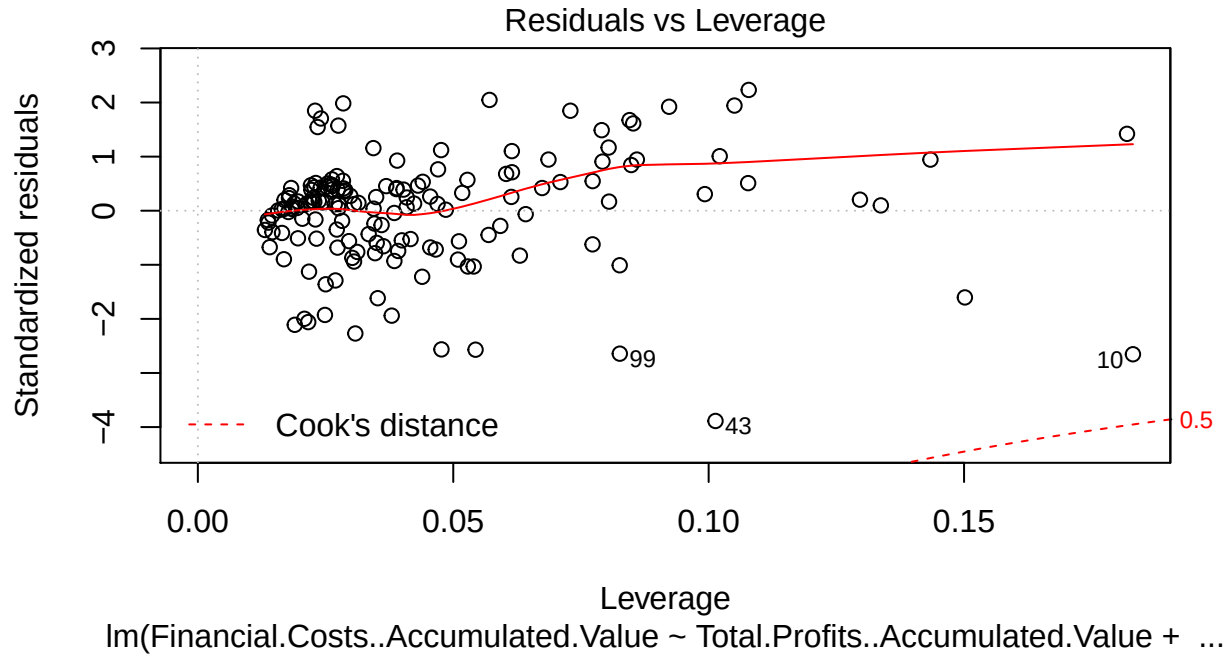



```
##
##  Shapiro-Wilk normality test
##
## data:  Res
## W = 0.95261, p-value = 3.61e-05
```

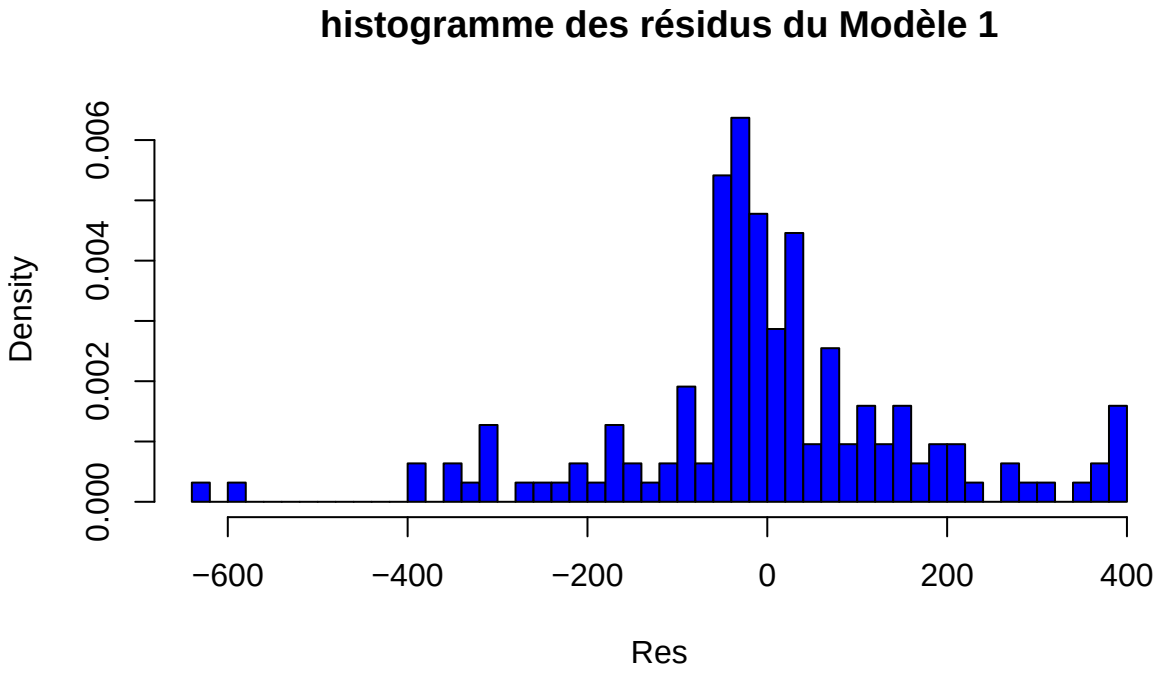

Residuals vs Fitted
Residuals
200
0
−200
−400
131
99
43
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Normal Q−Q
Standardized residuals
2
1
0
−2
−4
99
43
−2
−1
0
1
2
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...

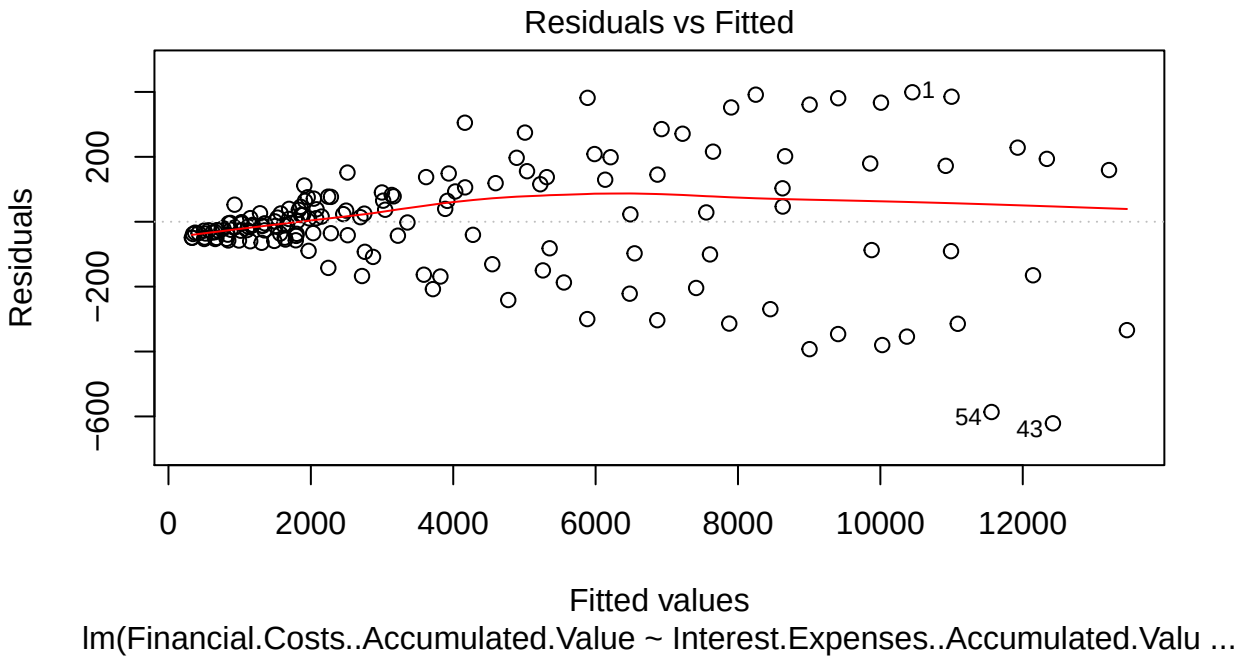


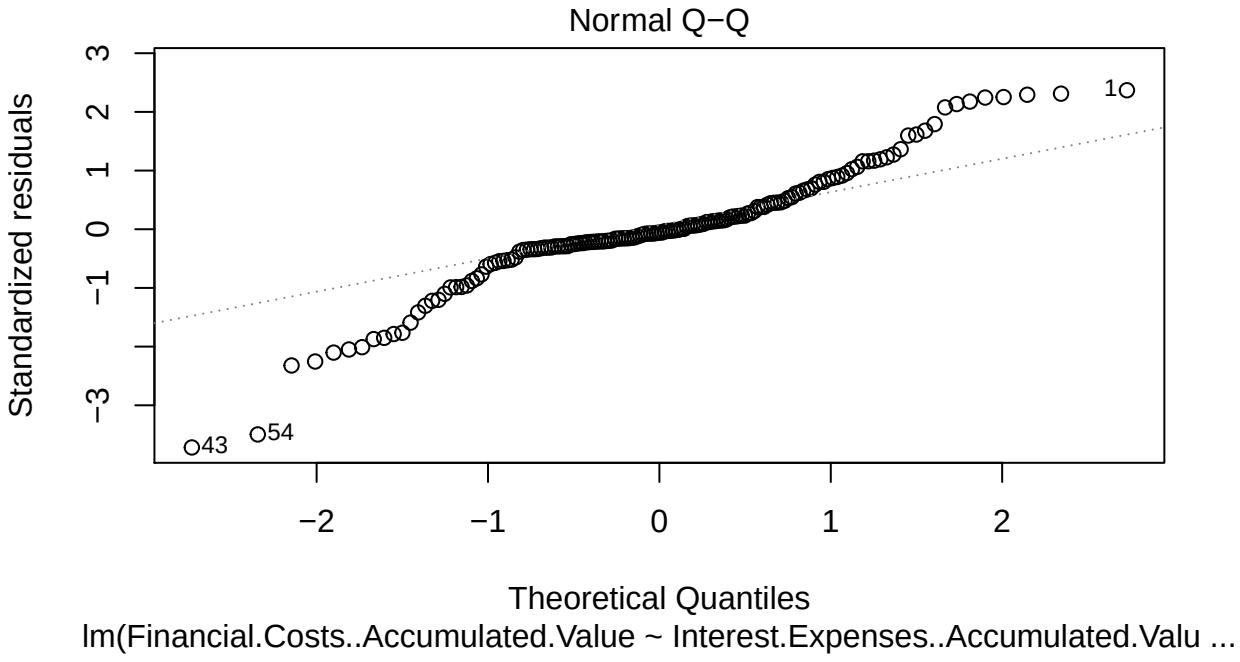


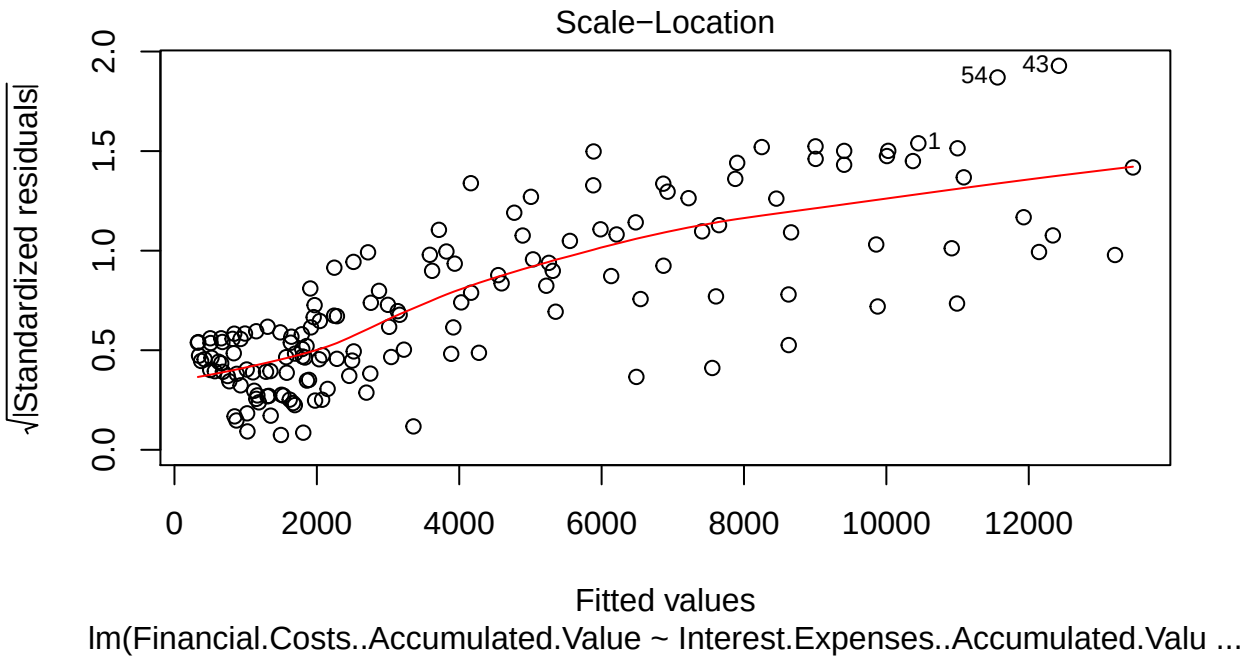
2.6.1.2 Modèle 1

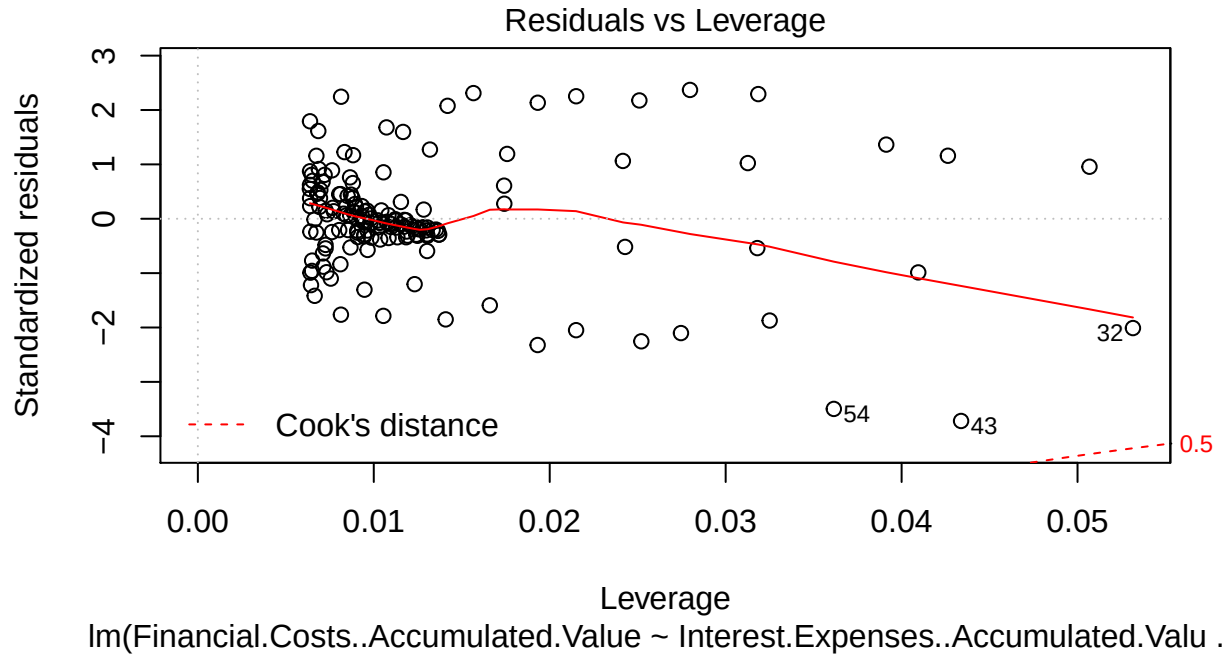



```
##
##  Shapiro-Wilk normality test
##
## data:  Res
## W = 0.94203, p-value = 4.787e-06
```


Residuals vs Fitted
Residuals
200
−200
−600
1
54
43
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Interest.Expenses..Accumulated.Valu ...


Normal Q−Q
Standardized residuals
3
2
1
0
−1
−3
1
43
54
−2
−1
0
1
2
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Interest.Expenses..Accumulated.Valu ...





2.6.2 Normalité

2.6.2.1 Modèle de Base

La p-valeur du test de Shapiro-Wilks sur les résidus observés est de 3.61e-05. L'on peut donc rejeter l'hy-

pothèse nulle selon laquelle les résidus sont normalement distribué. i.e. qu'il est improbable d'obtenir de tels résultats en supposant que les résidus observés soient normalement distribués. Le diagramme Quantile-Quantile montre aussi des résidus non alignés sur la première bissectrice.

2.6.2.2 Modèle 1

Les mêmes résultats et conclusions peuvent être faits sur les résidus observés du modèle 1.

2.6.3 Homoscédasticité

2.6.3.1 Modèle de Base

Le nuage de points des résidus observés en fonction des valeurs prédites présente plus ou moins une distribution conique (des résidus proches de zéro pour des valeurs moindres des prédictions et qui s'éloignent de zéro pour des valeurs supérieures des prédictions) indiquant probablement une hétéroscédasticité des résidus observés (une variance non constante de la variable "résidus"). Ce graphique et les graphiques "Scale Location" et "Residuals vs Leverage" laissent voir une tendance non linéaire des points, ce qui suggère qu'une certaine variabilité de la variable expliquée n'est pas capturé par le modèle (les variables explicatives) ou que le modèle adéquat n'est pas un modèle linéaire. [4]

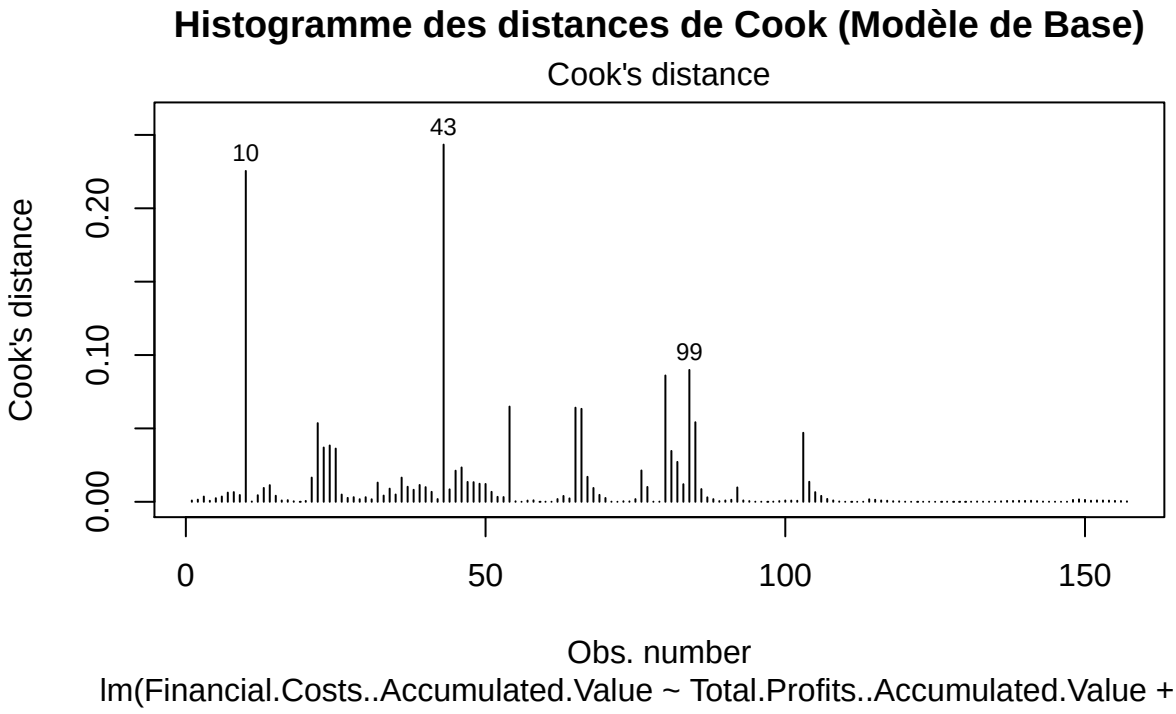
2.6.3.2 Modèle 1

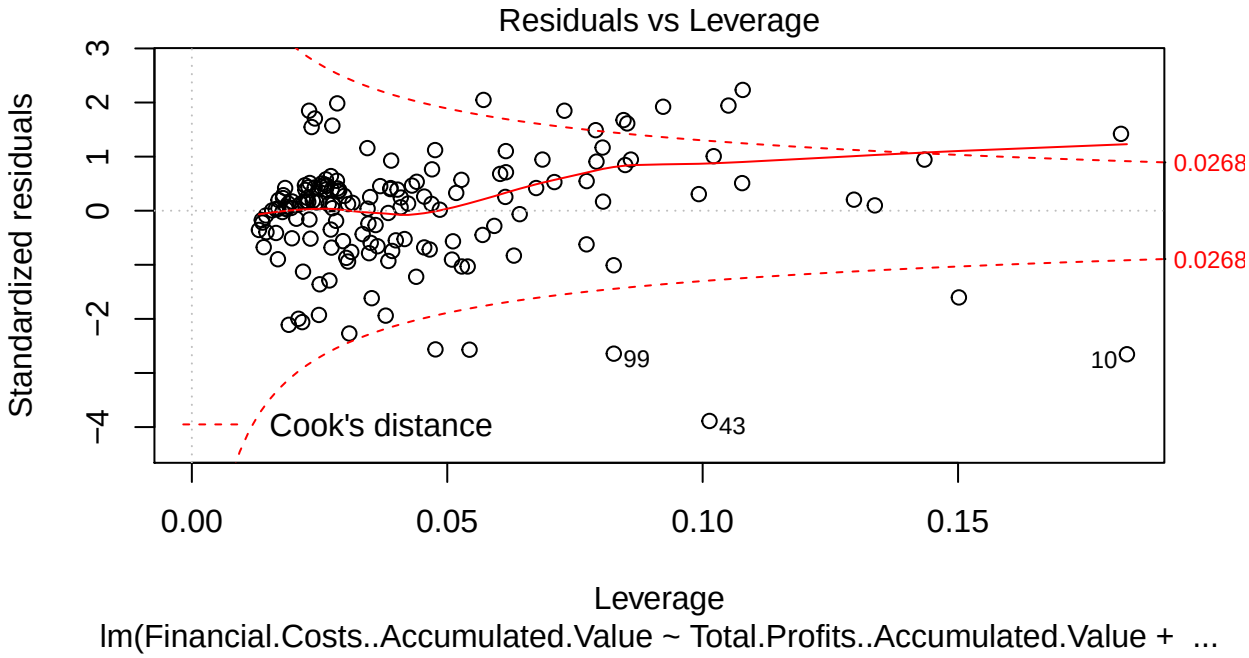
Les graphiques relatifs au modèle 2 laisse arriver aux même conclusions que pour le modèle de base, avec, cette fois une forme conique et une non linéarité plus prononcées.

2.7 Points influents

Les traitements qui suivent ont été entrepris dans l'intention de remédier aux résultats trouvés plus en haut (en terme de distribution des résidus). Ceux-ci consistent en l'élimination des points influents du jeu de données, i.e les points sans lesquels le modèle aurait été considérablement différent. La caractérisation de ces points est faite avec la distance de Cook. La distance de Cook de chaque observation est donnée dans l'histogramme suivant duquel les points qui ont des distances s'éloignant significativement de celles des autres observations, sont identifiés comme points influents et supprimés du jeu de données. Bien que les modèles réestimés après chaque élimination tendent à avoir une distribution normale des résidus observés (selon le diagramme Q-Q) et des valeurs d'AIC et d'erreur moyenne inférieures ainsi qu'un R2 ajusté supérieur à ceux du modèle de base, ce processus a été arrêté, car suggérait l'élimination d'un nombre important d'observations et bien que non continué, allait probablement suggérer l'élimination des points présentant l'amoncellement à gauche du graphique. Or, la totalité de ces points est ce qui définit l'échantillon et ne sont pas qu'une minorité d'observations. Seuls quelques graphiques et modèles réestimés parmi ceux traités sont inclus ci-dessous et le processus n'a été opéré que sur le modèle de base.

[4] Un test d'hétéroscédasticité aurait pu etre opéré.





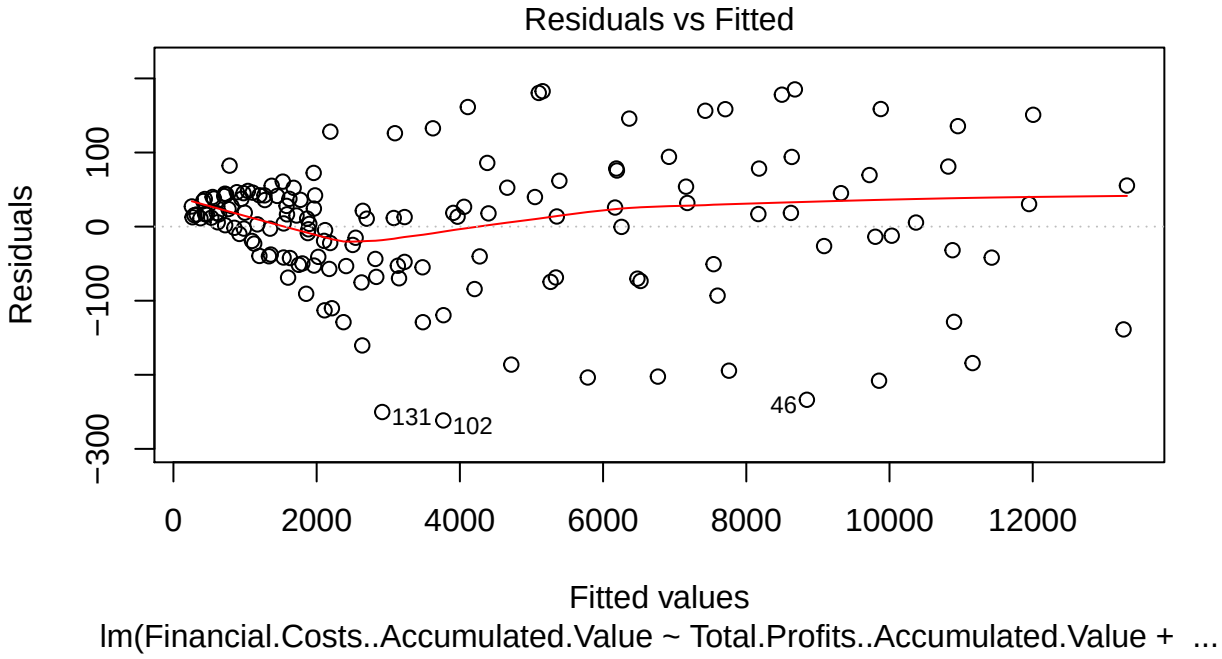
Les observations “43”,“10” et, à moindre degré,“99” sont extremes (en terme de distance de Cook et relativement aux autres observations (puisqu’elles n’ont pas une distance >0.5)) et sont éliminées. Le modèle est réestimé après l’élimination.

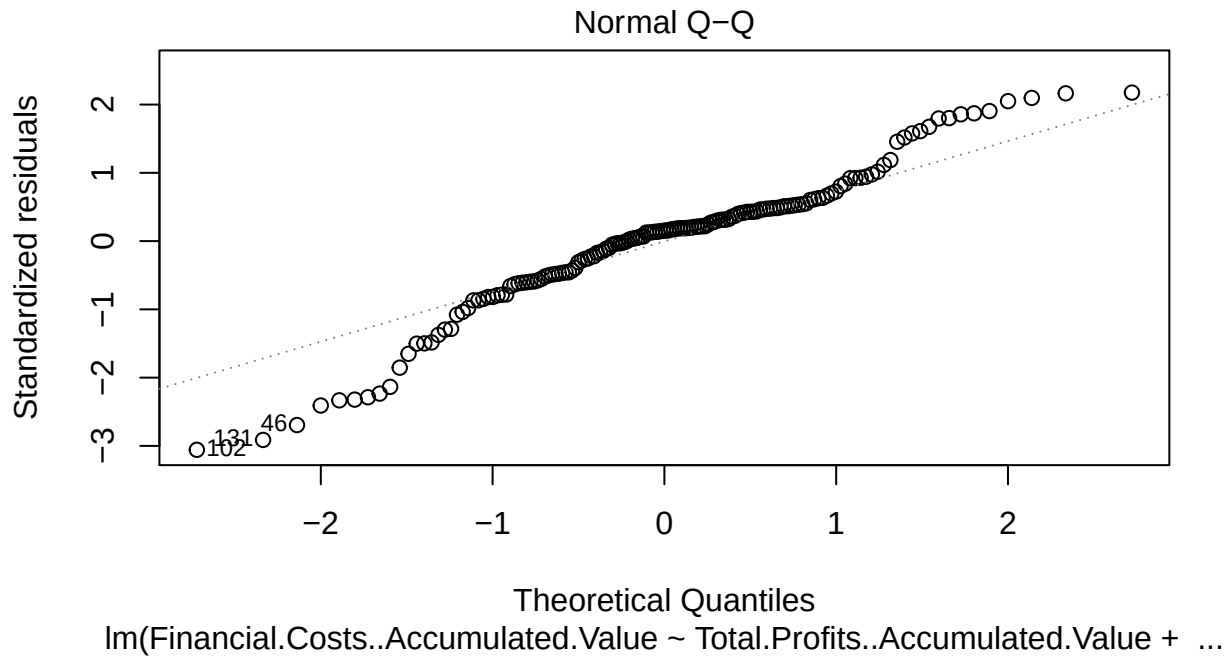
(Dans ce qui suit seul l’histogramme des distances est affiché).

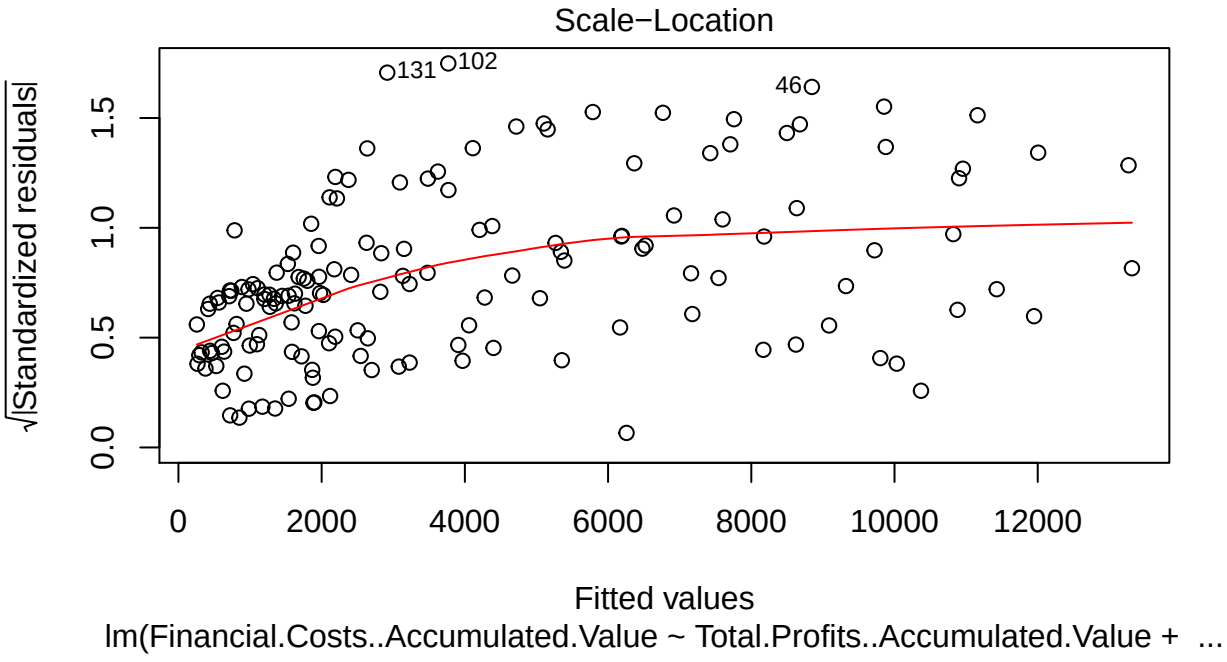
Modèle réestimé

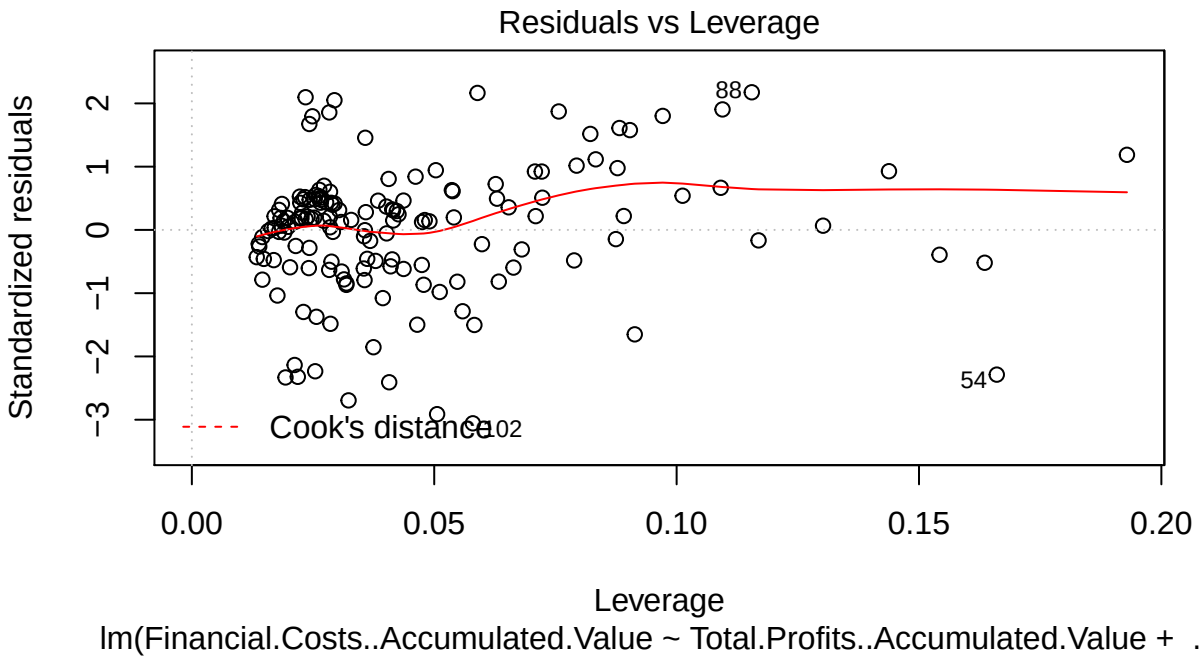
	standard.error	s6.df	R.squared	R.squared.adjusted	F.statistic
value	88.17335	7	0.9993511	0.9993247	37734
numdf	88.17335	147	0.9993511	0.9993247	6
dendf	88.17335	7	0.9993511	0.9993247	147

```
## [1]    7.000 1386.462
```



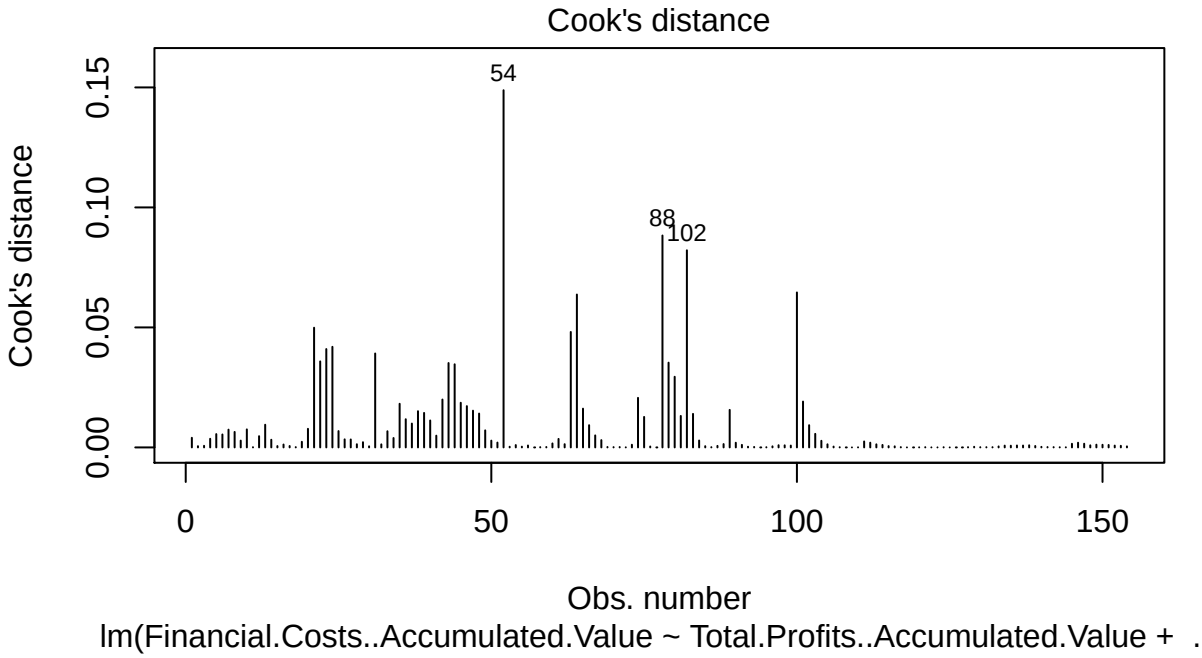

Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
131
46
102
−2
−1
0
1
2
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
131
102
46
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...



On constate que l'AIC et l'erreur moyenne ont diminué. Nous inspectons l'histogramme des distances de Cook du nouveau modèle.

ogramme des distances de Cook (Modèle après élimination de 3 obser

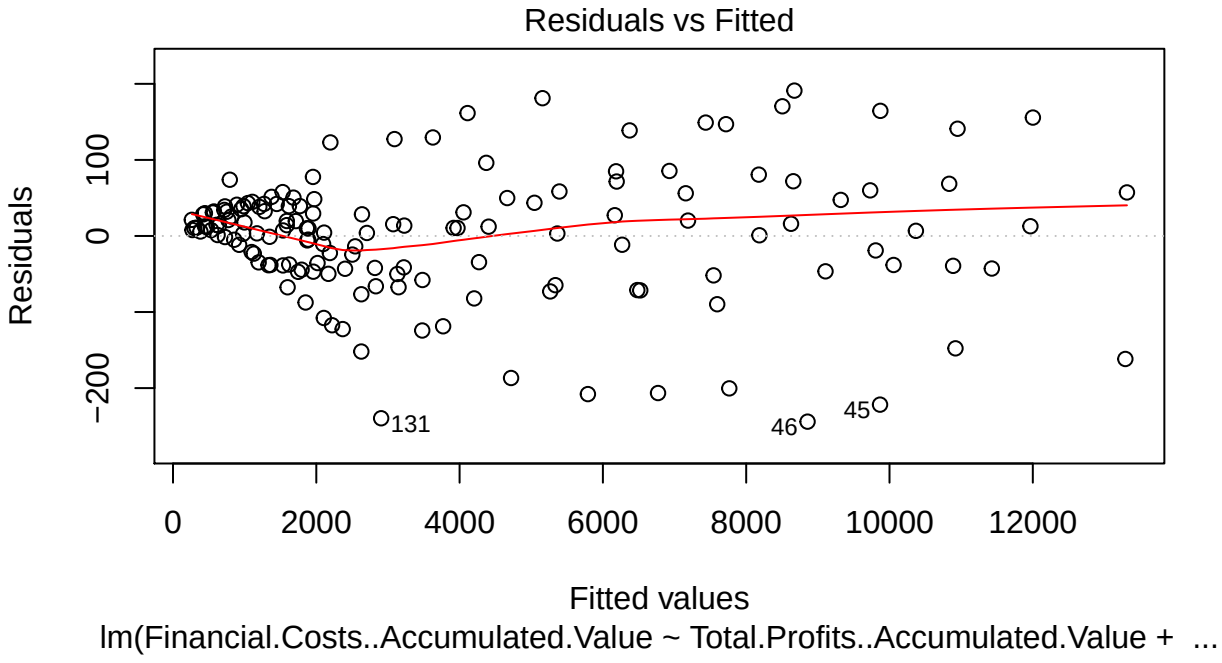


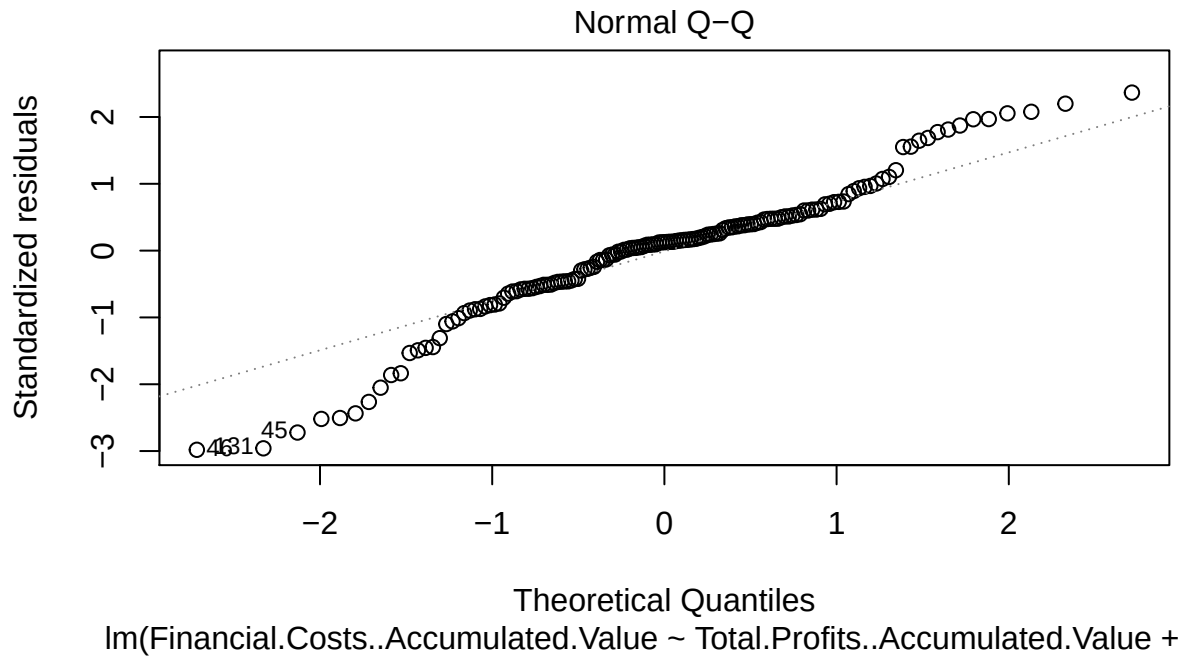
Les observations "54","88","102" sont éliminées et le modèle réestimé.

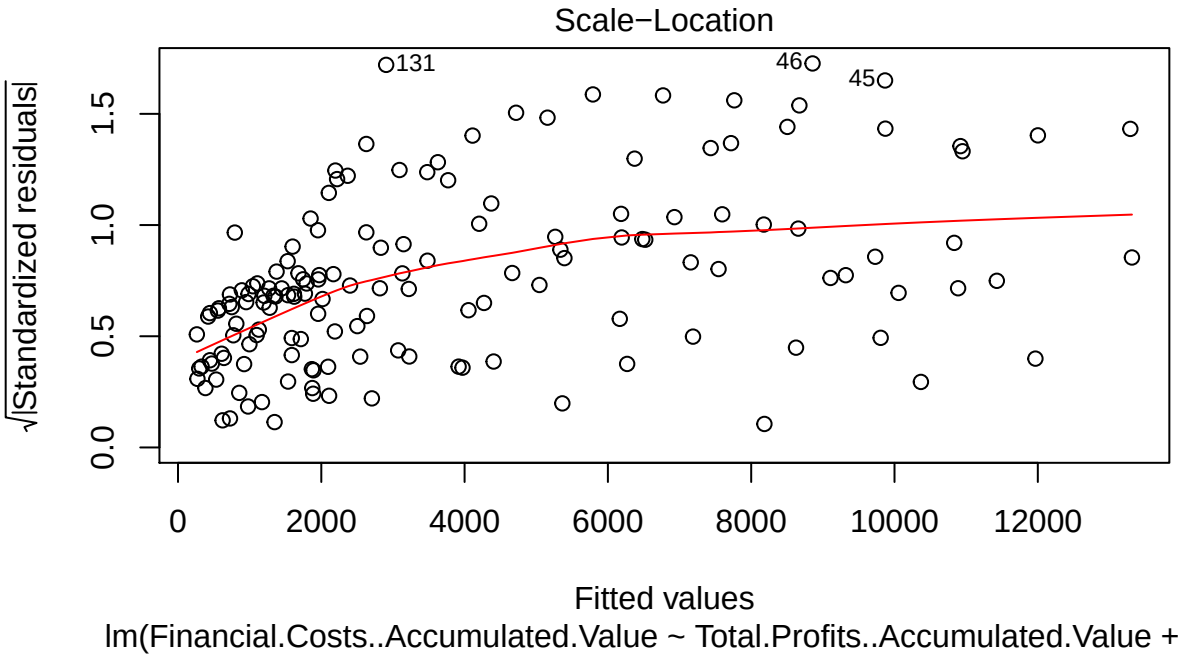
Modèle réestimé

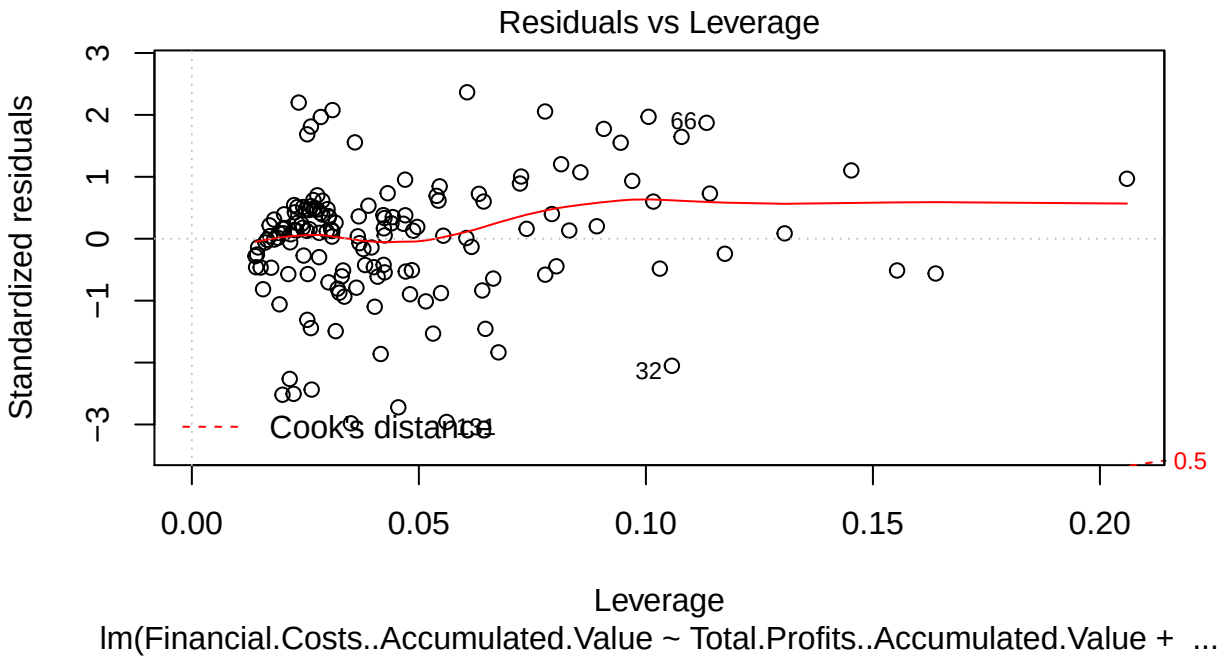
```
##
## Call:
## lm(formula = Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value +
##     Power.Generation..Accumulated.Value + Inventories..Accumulated.Value +
##     Total.Assets..Accumulated.Value + Interest.Expenses..Accumulated.Value +
##     Selling.Expenses..Accumulated.Value, data = donnees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244.12  -40.28   10.38   40.15  191.10
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        18.8864601 23.7162970   0.796  0.42714
## Total.Profits..Accumulated.Value   -0.0125704  0.0042163  -2.981  0.00337 **
## Power.Generation..Accumulated.Value  0.0015714  0.0010006   1.570  0.11851
## Inventories..Accumulated.Value       0.0022269  0.0008073   2.758  0.00656 **
## Total.Assets..Accumulated.Value    -0.0001653  0.0001127  -1.467  0.14454
## Interest.Expenses..Accumulated.Value  0.7824760  0.0236894  33.031  < 2e-16 ***
## Selling.Expenses..Accumulated.Value   0.1563710  0.0161722   9.669  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 83.36 on 144 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 4.098e+04 on 6 and 144 DF,  p-value: < 2.2e-16

## [1]    7.000 1342.623
```



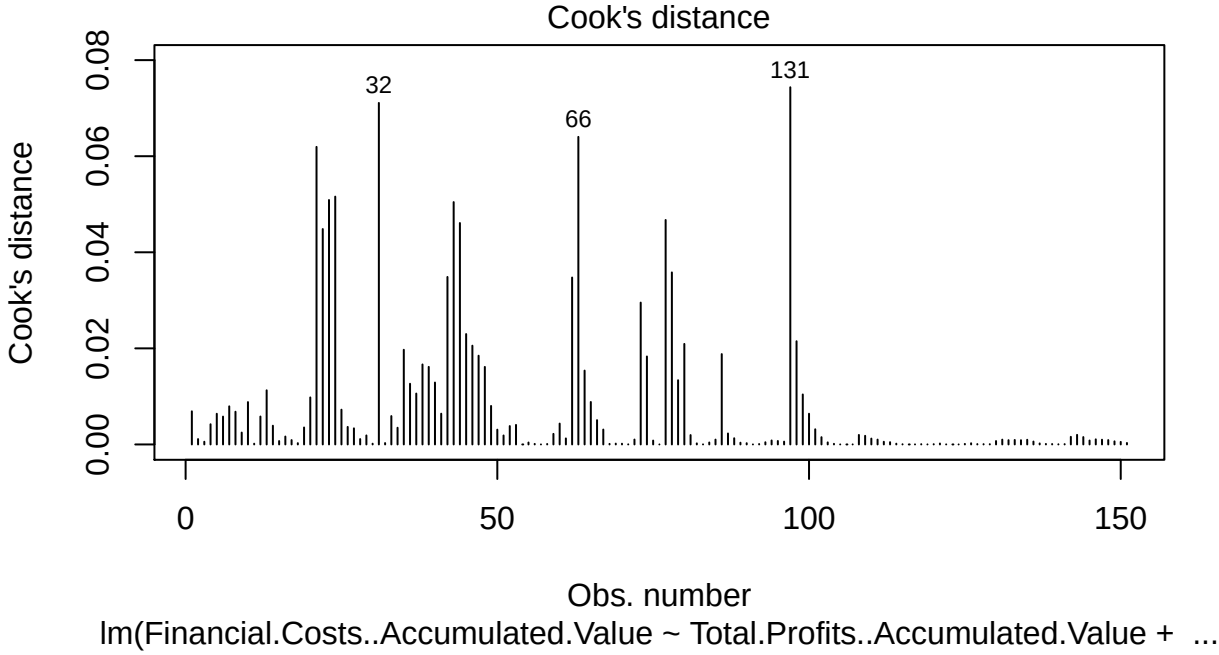

Normal Q−Q
Standardized residuals
−3
−2
−1
0
1
2
46
131
45
−2
−1
0
1
2
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
131
46
45
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...



l'AIC et l'erreur ont baissé et le R2 ajusté a augmenté. Après l'élimination de ces 6 observations les variables "Power.Generation..Accumulated.Value" et "Total.Assets..Accumulated.Value" sont devenues non significatives, nous réestimons le modèle sans celles-ci et inspectons le nouvel histogramme des distances.

ogramme des distances de Cook (Modèle après élimination de 6 obser

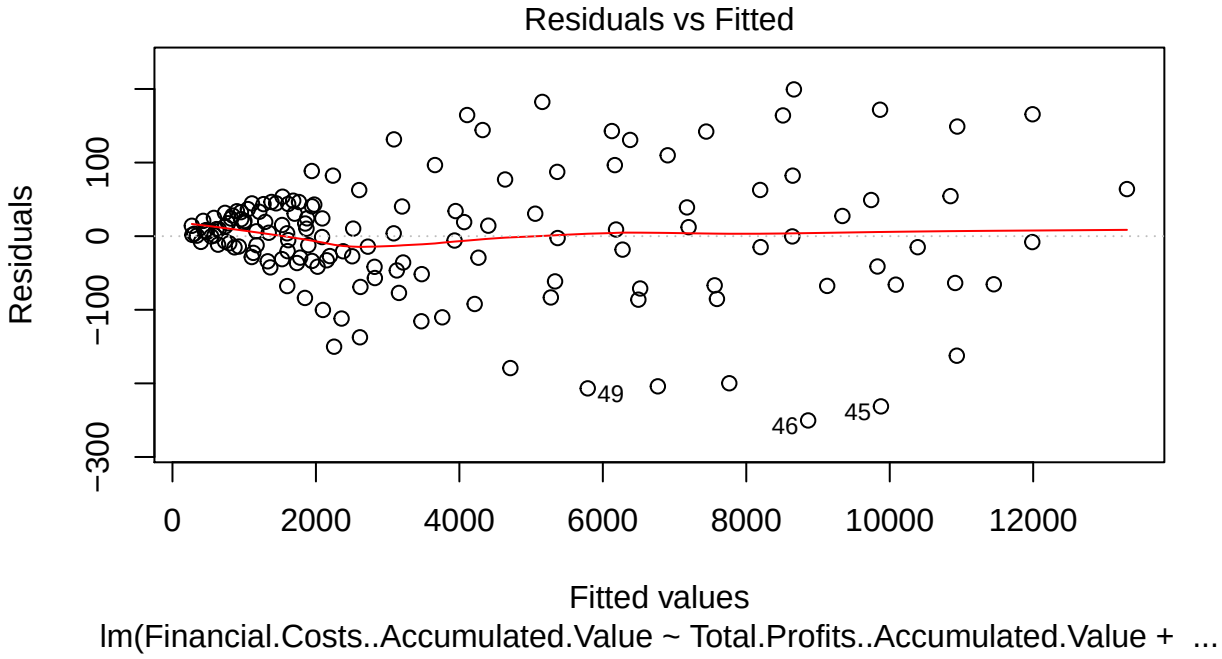


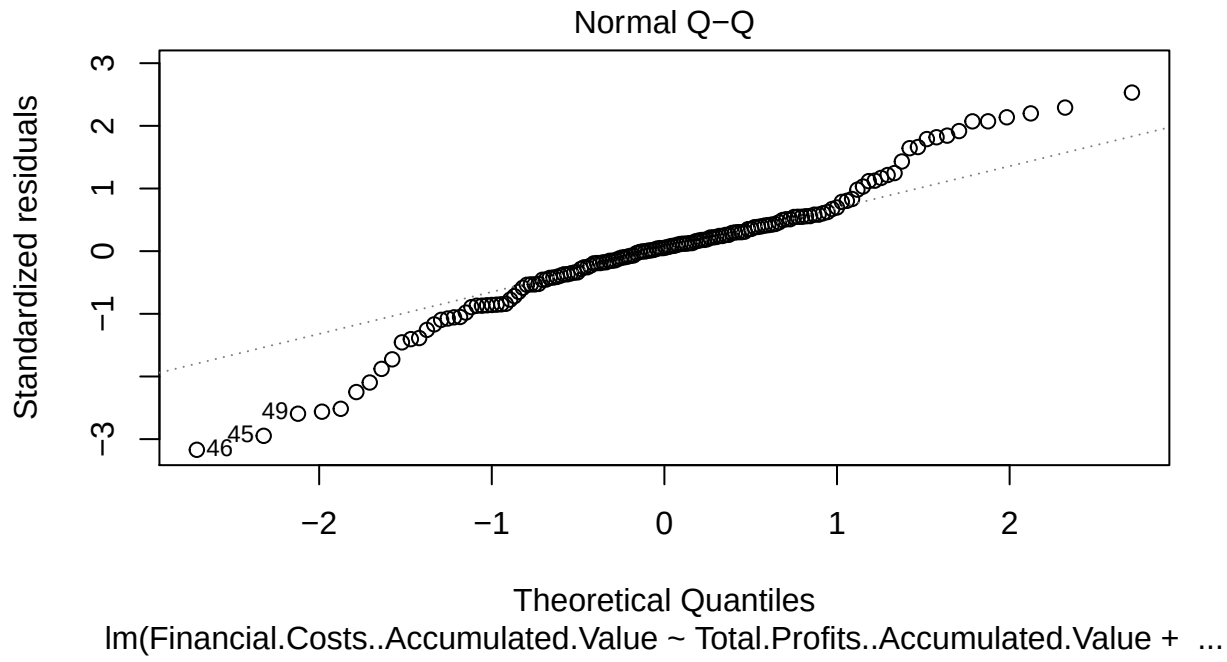
Les observations "32","66","131" sont éliminées et le modèle réestimé.

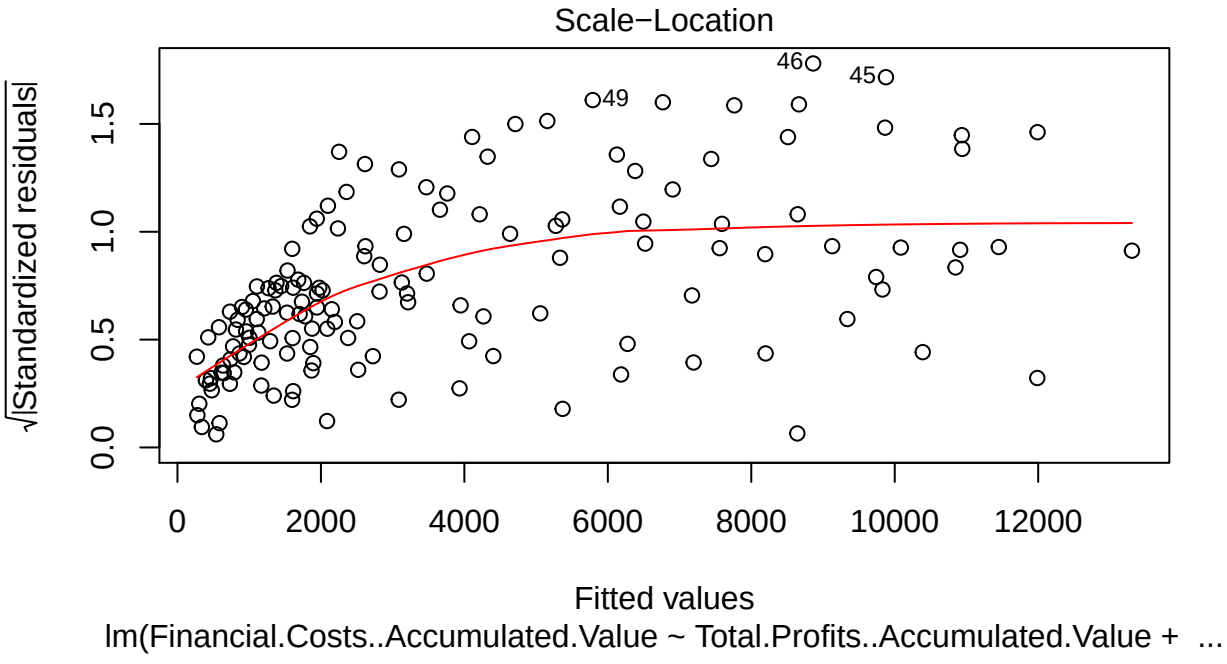
Modèle réestimé

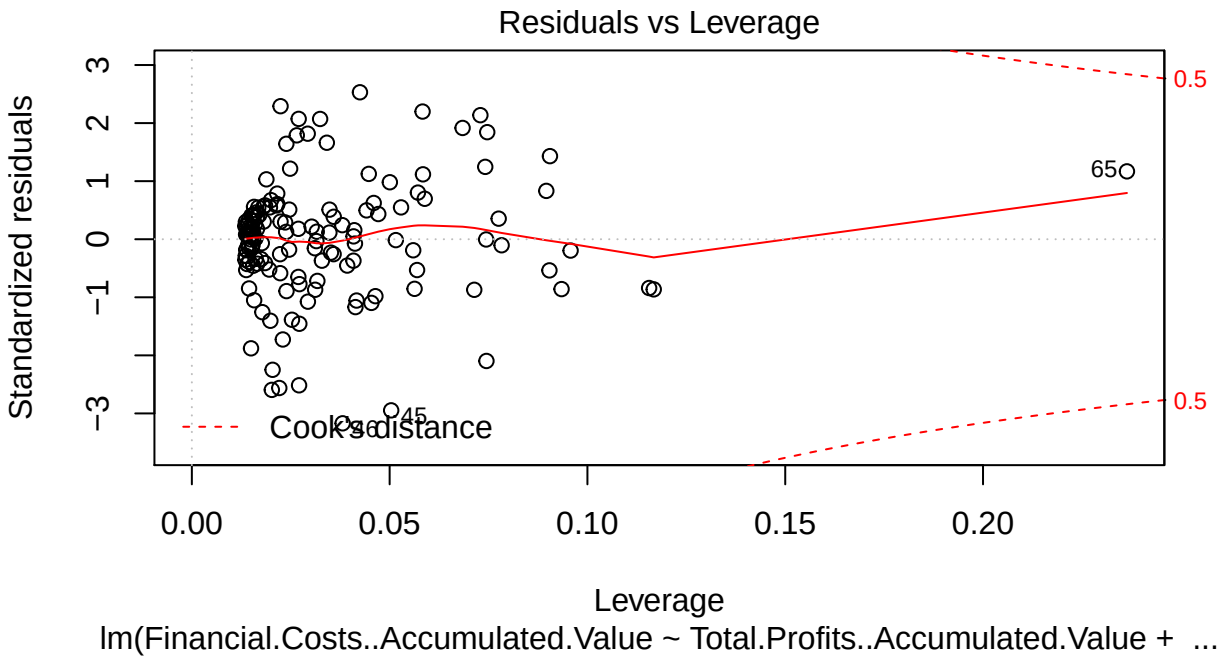
	standard.error	s7.df	R.squared	R.squared.adjusted	F.statistic
value	80.5219	7	0.9994226	0.9994065	61884.17
numdf	80.5219	144	0.9994226	0.9994065	4.00
dendf	80.5219	7	0.9994226	0.9994065	143.00

```
## [1]    5.000 1303.918
```



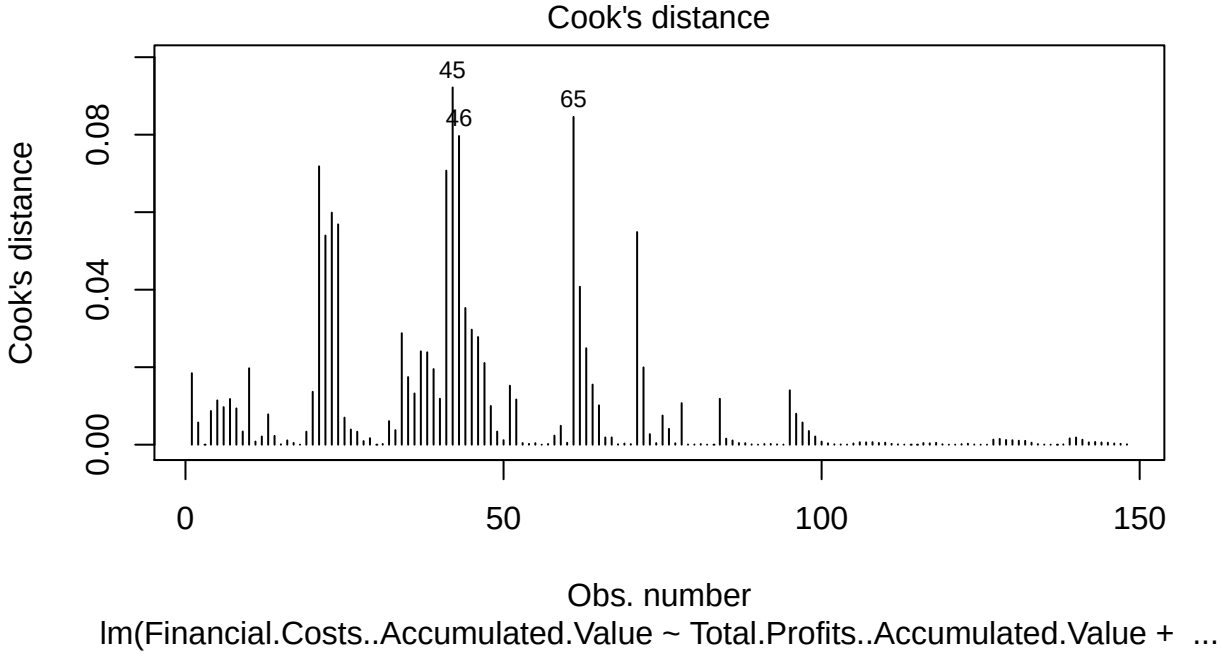

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−3
46
45
49
−2
−1
0
1
2
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...

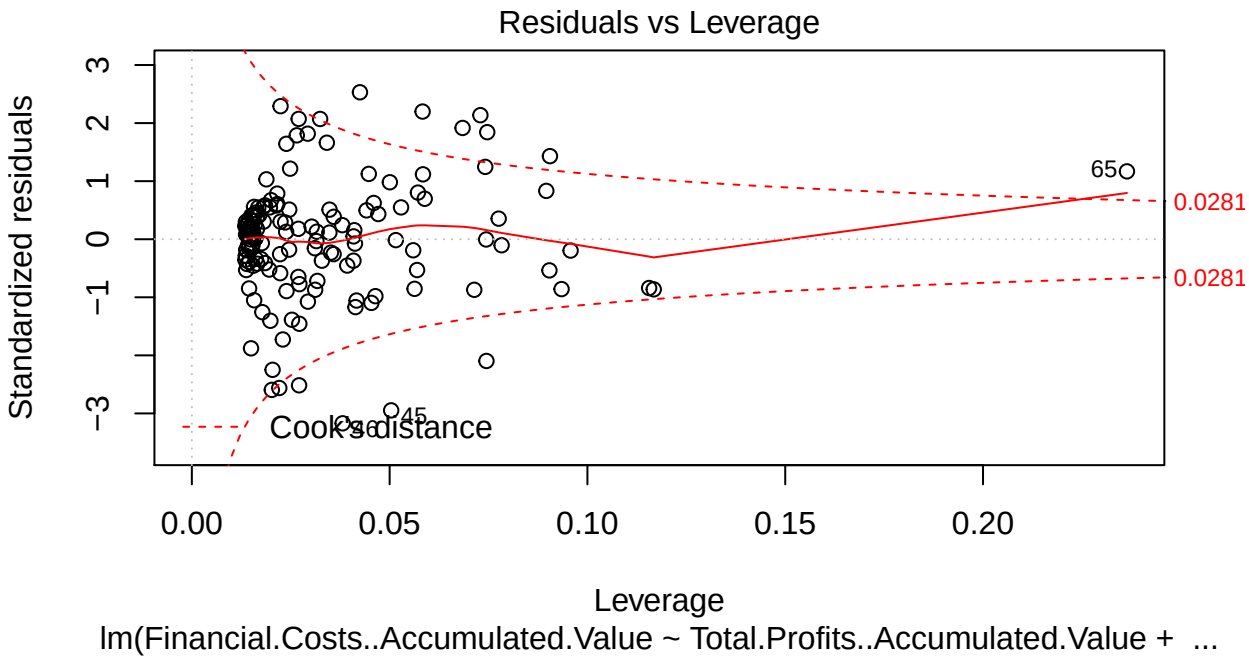

Scale−Location
√|Standardized residuals|
1.5
1.0
0.5
0.0
46
45
49
0
2000
4000
6000
8000
10000
12000
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...

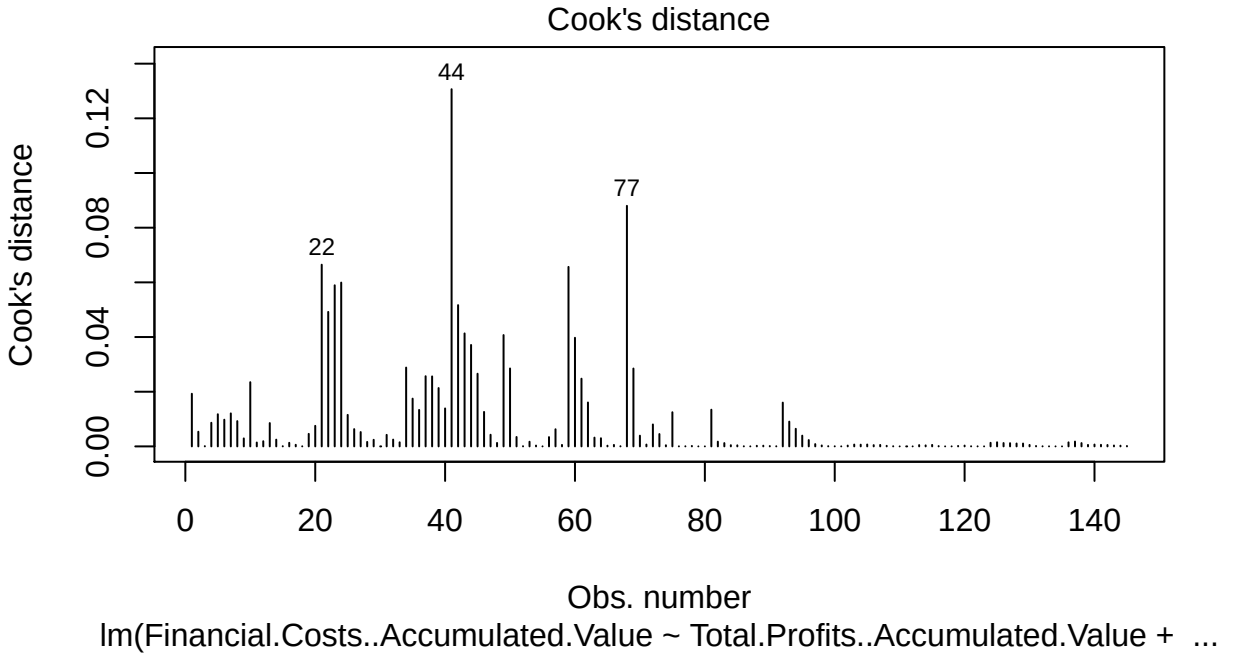


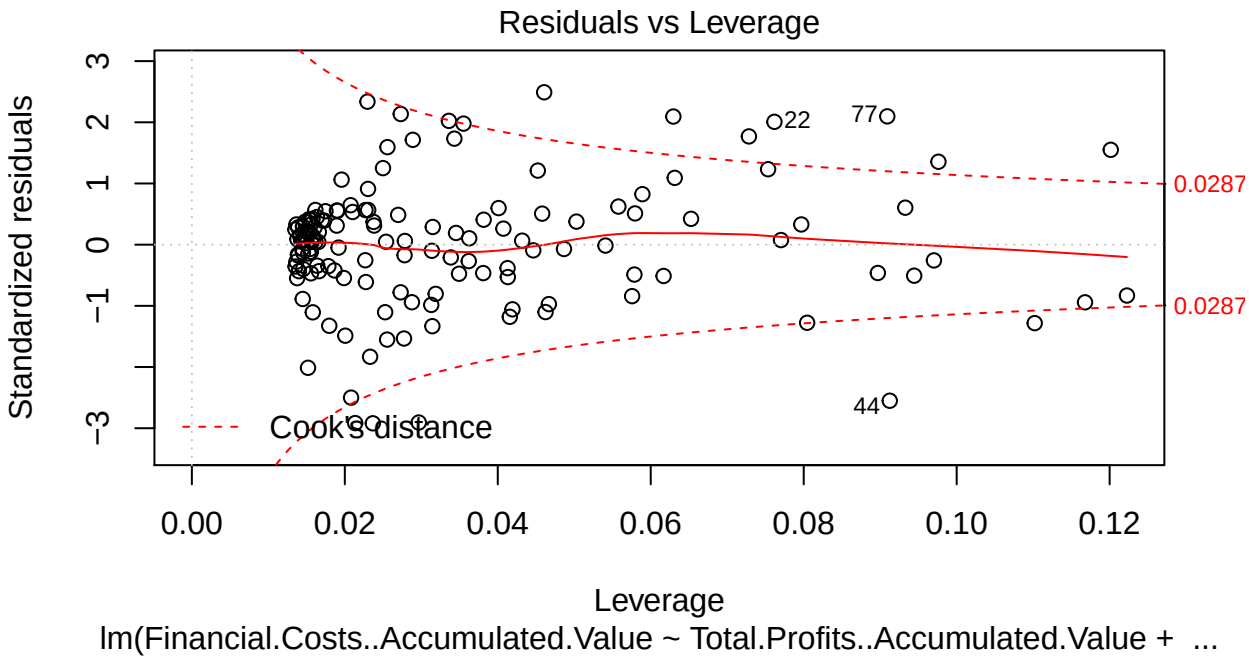
L'AIC et l'erreur moyenne ont diminué et le R2 ajusté augmenté.

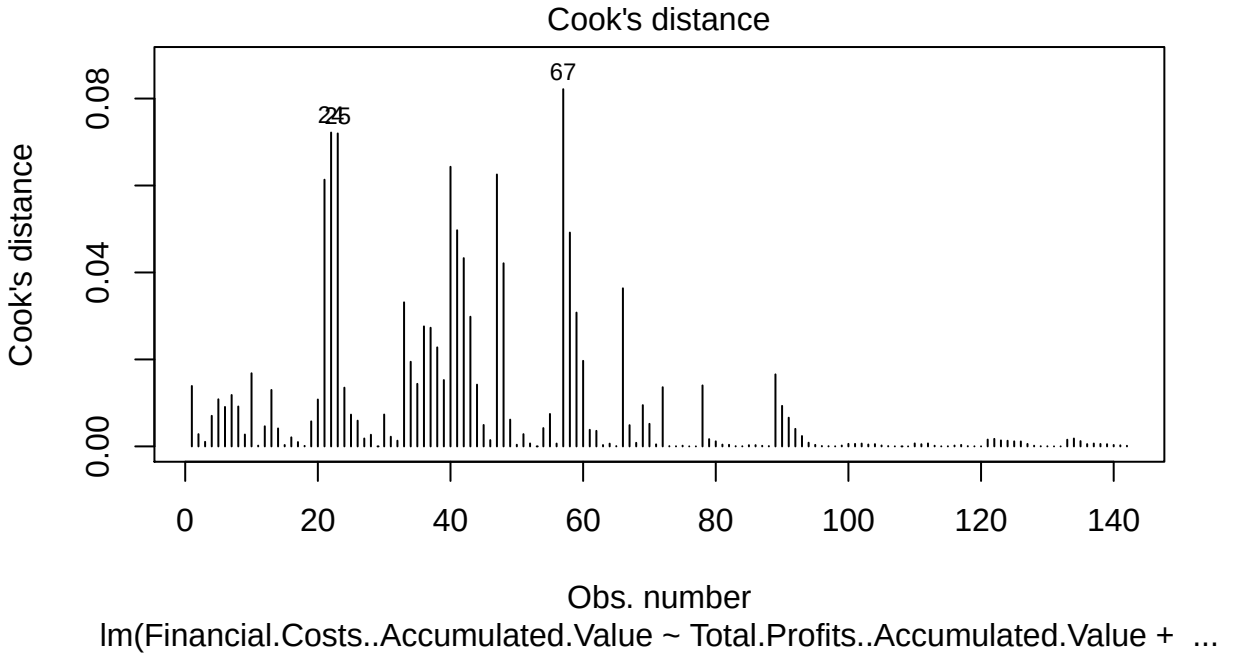
ogramme des distances de Cook (Modèle après élimination de 9 obser

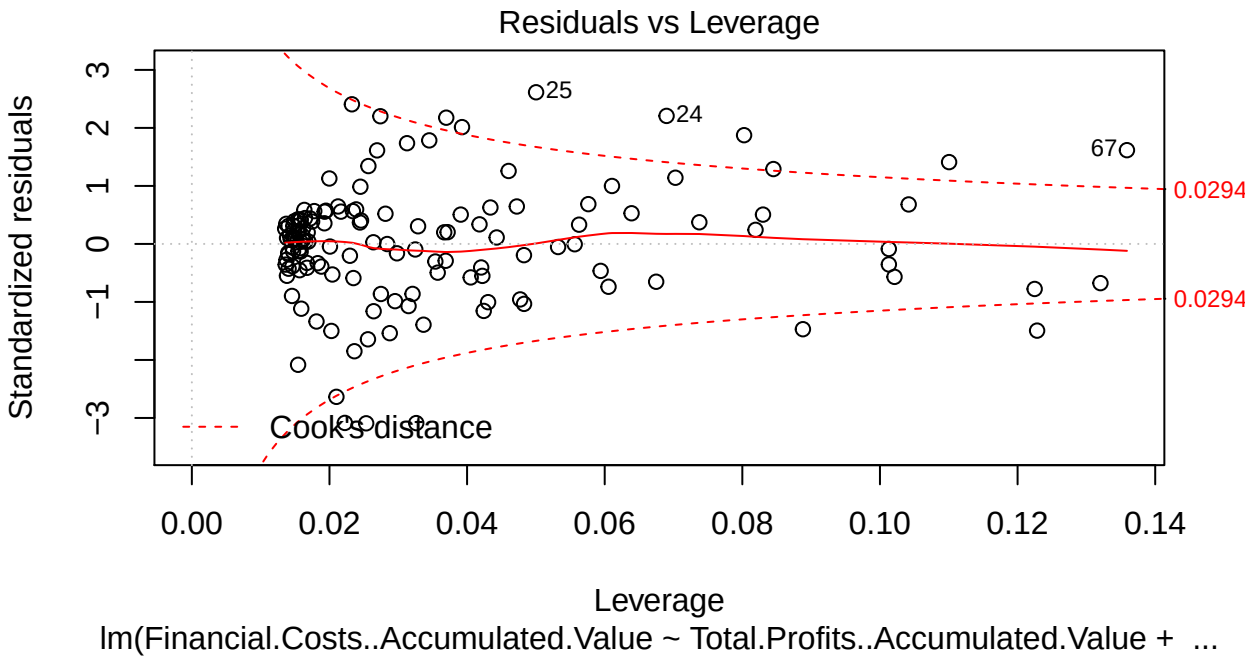



Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−3
65
0.0281
0.0281
Cook's distance
45
0.00
0.05
0.10
0.15
0.20
Leverage
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Cook's distance
Cook's distance
0.12
0.08
0.04
0.00
44
77
22
0
20
40
60
80
100
120
140
Obs. number
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
0.0287
0.0287
22
77
44
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...


Cook's distance
Cook's distance
Obs. number
67
25
24
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...

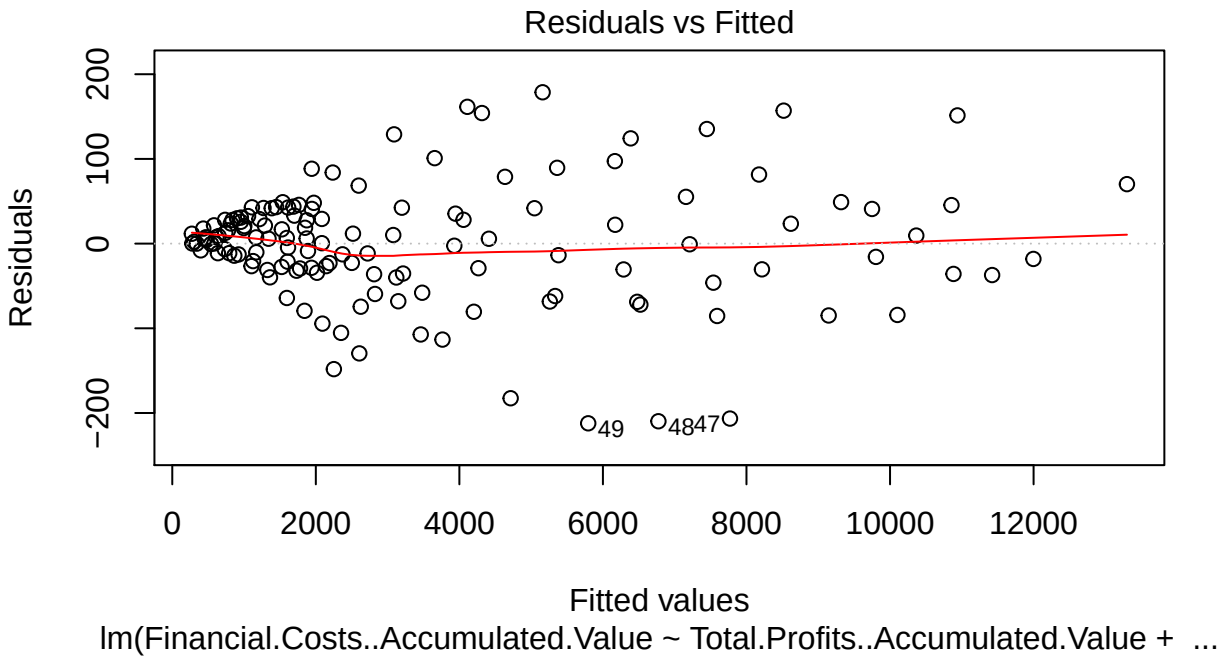


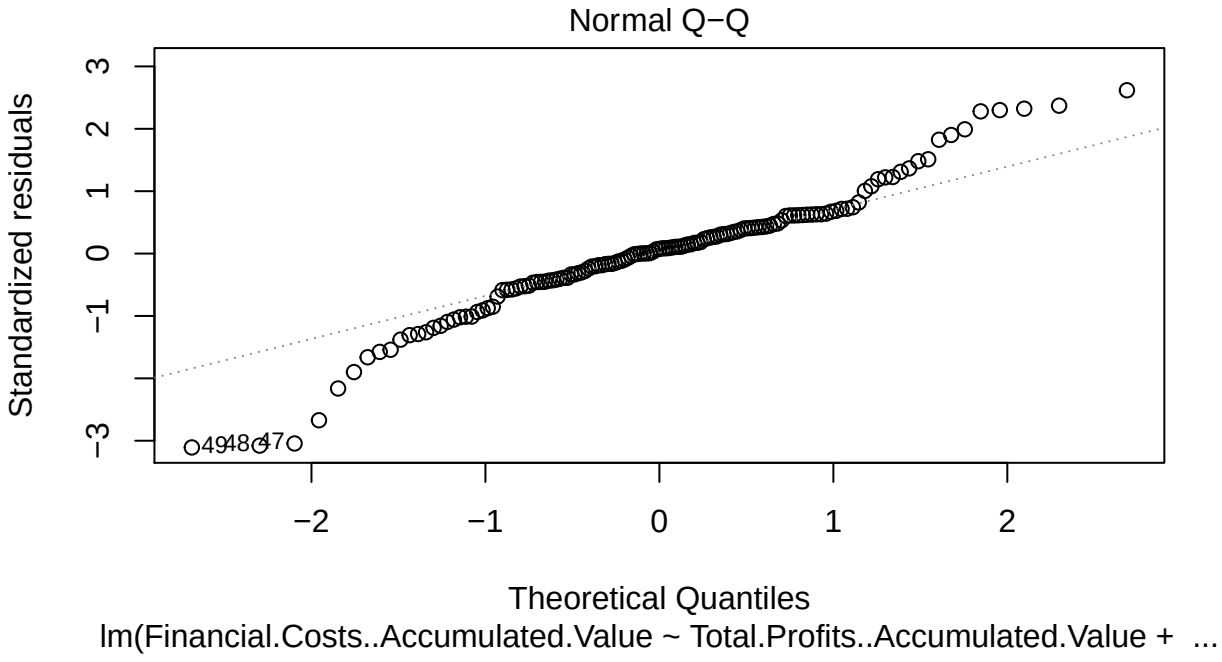
Après l'élimination de 18 observations et enfin de 21 (et de deux variables), l'on constate que la distribution des résidus reste toujours "problématique". Le processus ne fera que suggérer de supprimer les points des extrémités du diagramme Q-Q et peut-être les points de l'amoncellement sur la gauche des différents graphiques, celui-ci est donc arrêté.

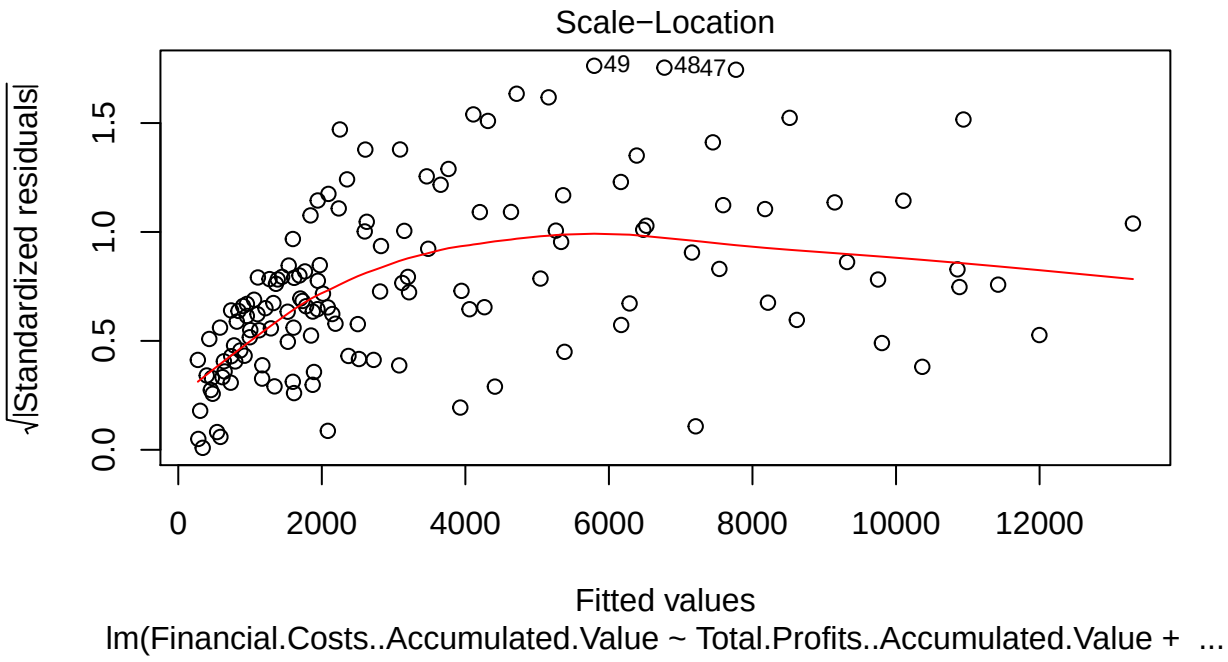
Modèle aprè élimination de 18 observations

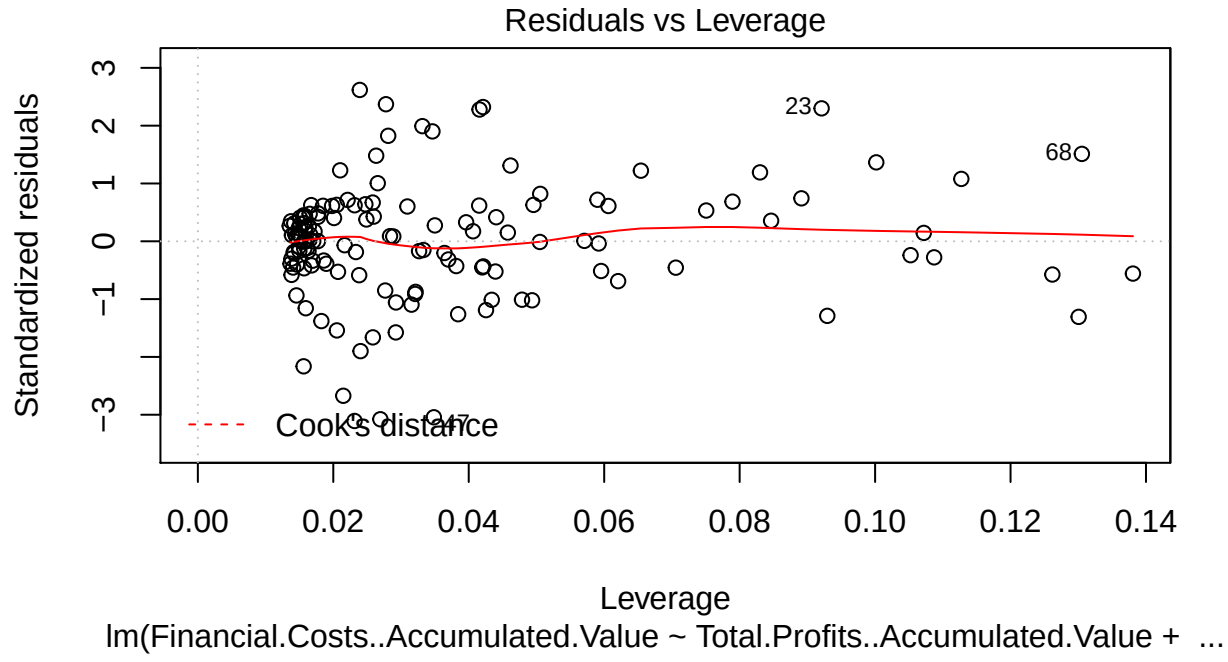
	standard.error	s11.df	R.squared	R.squared.adjusted	F.statistic
value	69.08147	5	0.9995116	0.9994971	68564.21
numdf	69.08147	134	0.9995116	0.9994971	4.00
dendf	69.08147	5	0.9995116	0.9994971	134.00

```
## [1]    5.000 1182.317
```

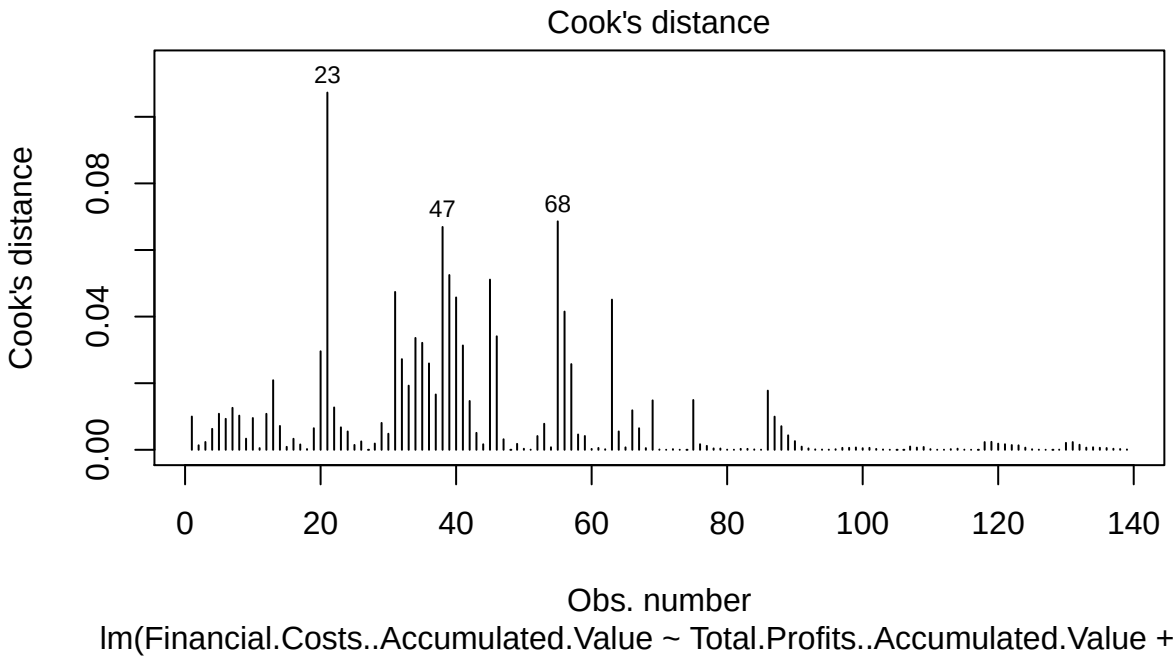

Residuals vs Fitted
Residuals
200
100
0
−200
0
2000
4000
6000
8000
10000
12000
49
48
47
Fitted values
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...

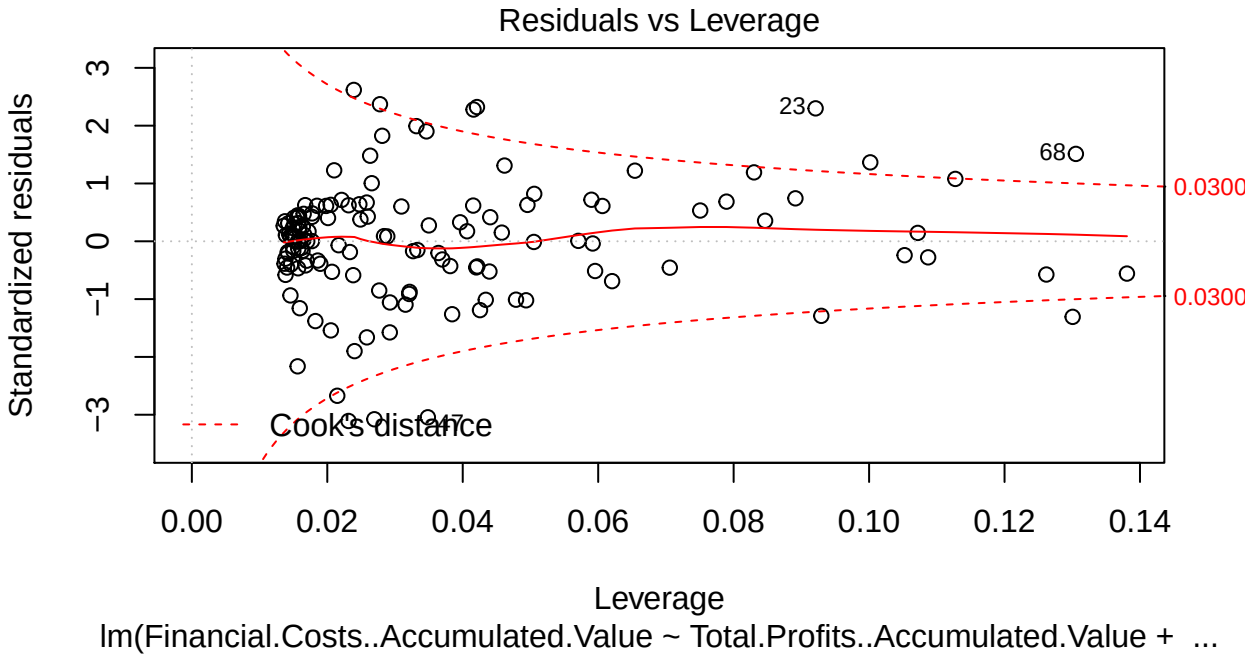

Normal Q−Q
Standardized residuals
3
2
1
0
−1
−3
−2
−1
0
1
2
49
48
47
Theoretical Quantiles
lm(Financial.Costs..Accumulated.Value ~ Total.Profits..Accumulated.Value + ...





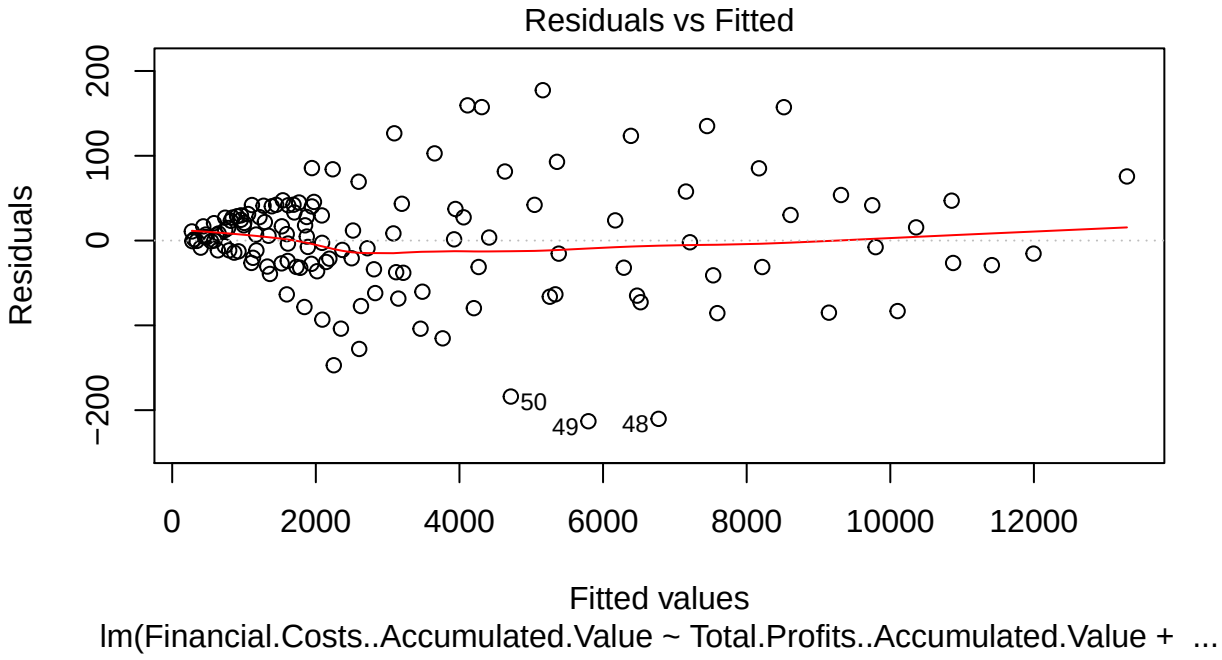
Modèle aprè élimination de 21 observations

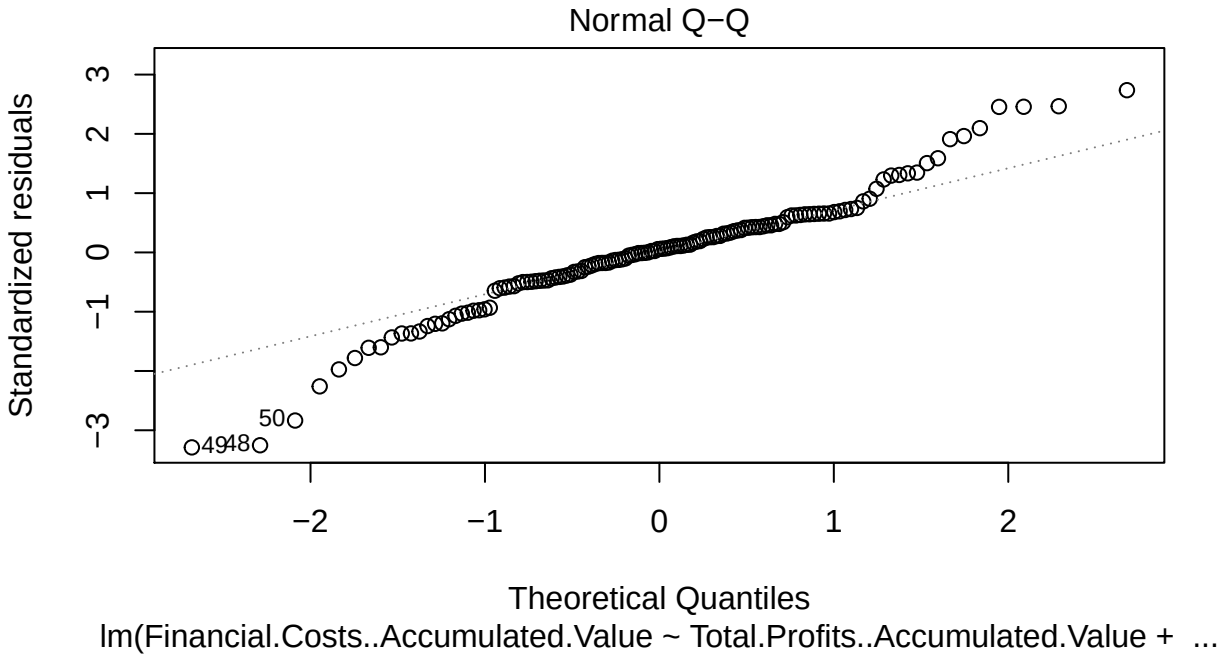


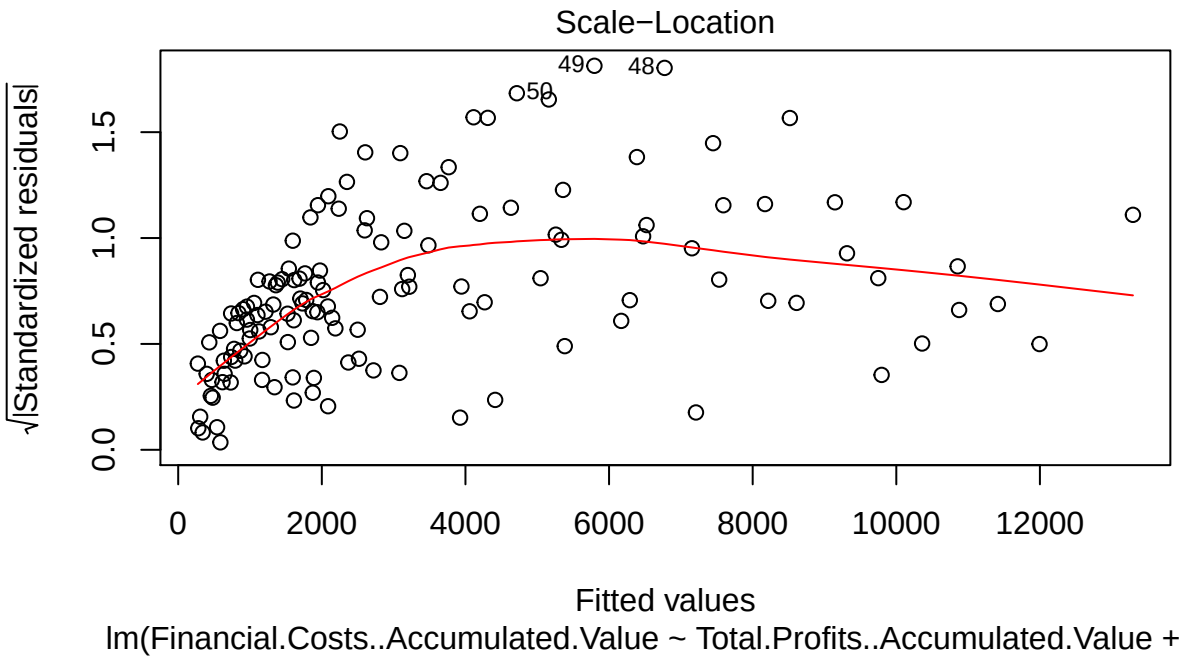


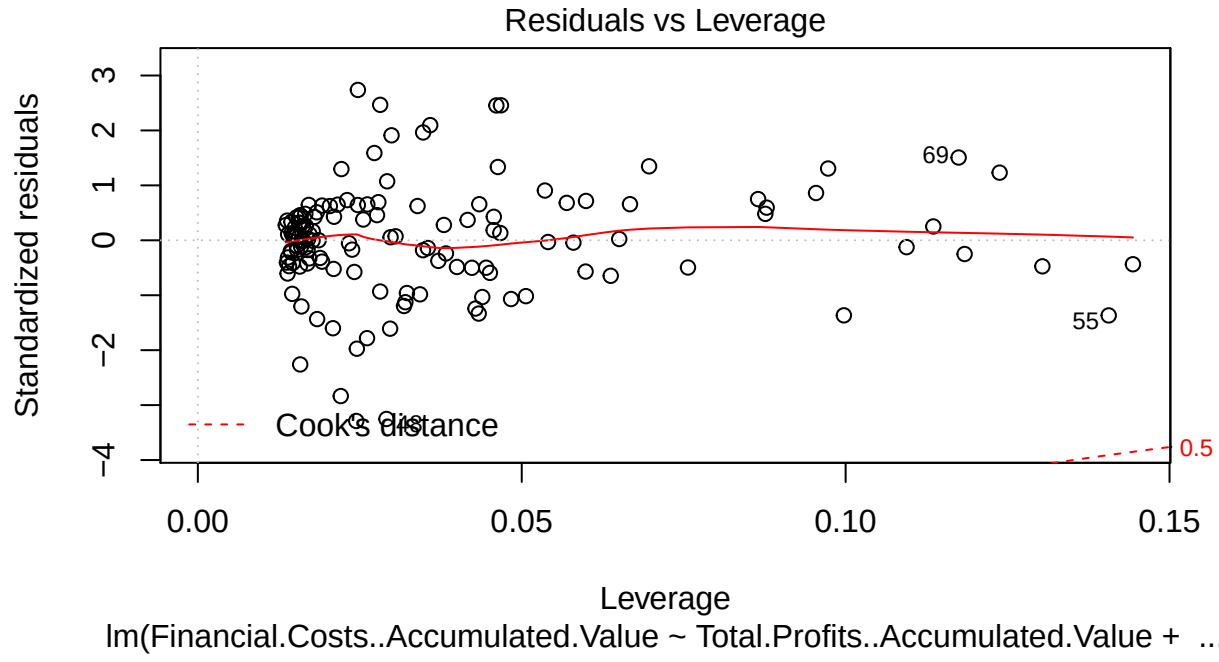
	standard.error	s12.df	R.squared	R.squared.adjusted	F.statistic
value	65.59935	5	0.9995399	0.9995258	71145.11
numdf	65.59935	131	0.9995399	0.9995258	4.00
dendf	65.59935	5	0.9995399	0.9995258	131.00

```
## [1]    5.000 1142.836
```









Il est conclu à cette étape (sur base de ce qui a été opéré dans cette étude) que le modèle obtenu n'est pas un modèle avec lequel l'on s'engagerait à expliquer ou prédir la variable cible, probablement celles-ci n'est pas linéairement modélisable par les variables manipulées ici.

2.8 Interprétation et prédiction

A cette étape le modèle aurait été interprété (interprétation “métier” de comment les variables retenues expliqueraient la variable dépendente et interprétation quantitative de comment la variation d’une variable influe sur la valeur de la variable expliquée), et testé sur de nouvelles données pour prédiction [5].

[5]La pratique courante est d’ “entrainer” le modèle sur une partie des données et de le tester sur une autre partie de celles-ci. Ceci a été essayé, en adoptant une “validation croisée”, toutefois pour ce faire, sous R, il faut utiliser la fonction “train” au lieu de “lm”, la manipulation d’un objet retourné par cette première n’ayant pas été aisé, cette initiative a été abandonnée.