Analyse de données hospitalières - Régression

I- Setup :

```
df <- df_origine %>%
  filter(TOTAL_COST <= TTC_summary['3rd Qu.'] + 1.5*(TTC_summary['3rd Qu.']
-TTC_summary['1st Qu.'])) %>%
  filter(AGE <= AGE_summary['3rd Qu.'] + 1.5*(AGE_summary['3rd Qu.']-AGE_su
mmary['1st Qu.']))
```

- Exclusions de données aberrantes, précisément :
  - Celles ayant été constatées par rapport au TOTAL_COST des admissions ;
  - Celles ayant été constatées par rapport à l'AGE du patient concerné par les admissions.

II- Regression :

II.1 – Itération 1 :

```
df %>%
  mutate(RISKDEATH = as.character(RISKDEATH),
         WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'G
eneralist', 'Specialist')
         ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION + CCI,
      data = .
  ) -> reg

summary(reg)
```

```
Call:
lm(formula = TOTAL_COST ~ LOS + AGE + WARD_ADMISSION + CCI, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-900.45 -225.41  -20.21  133.15 2239.12

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               758.3634    53.9657  14.053  < 2e-16 ***
LOS                       311.6288     1.9069 163.421  < 2e-16 ***
AGE                       -10.6454     0.6393 -16.652  < 2e-16 ***
WARD_ADMISSIONSpecialist 704.4096    16.4291  42.876  < 2e-16 ***
CCI                        22.2065     7.2001   3.084  0.00206 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 390.4 on 2593 degrees of freedom
```

```
Multiple R-squared:  0.9166,    Adjusted R-squared:  0.9164
F-statistic:  7121 on 4 and 2593 DF,  p-value: < 2.2e-16
```

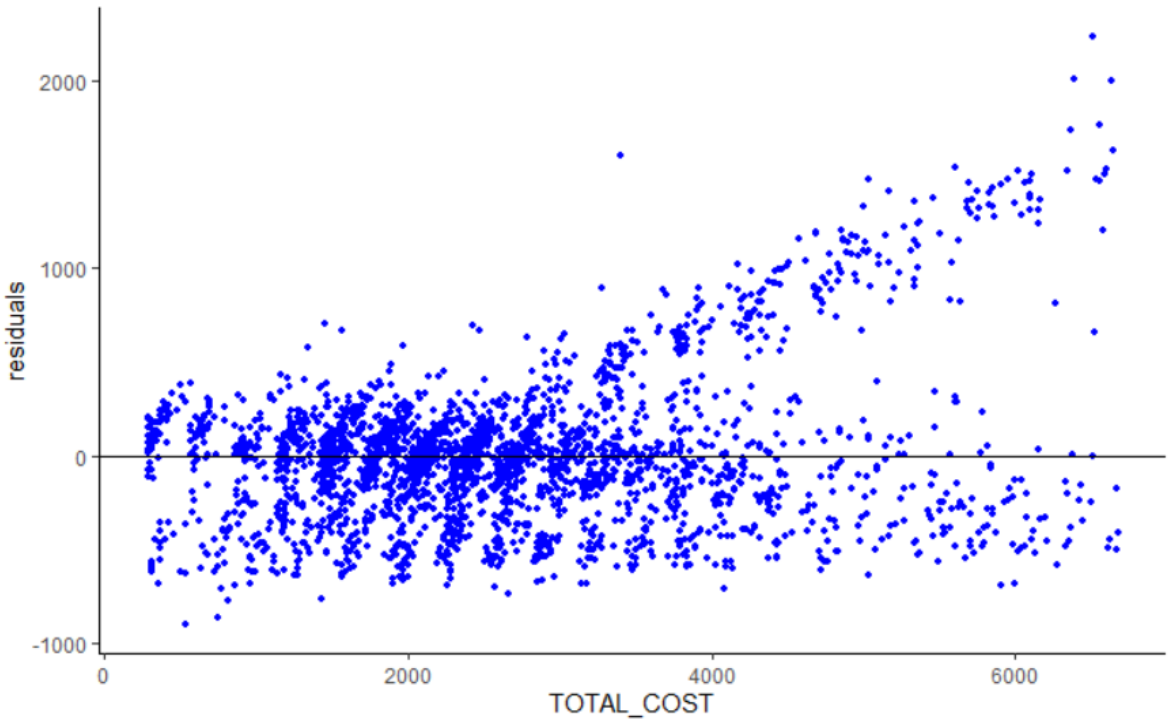
- Les variables LOS, AGE, WARD_ADMISSIONSpecialist et CCI sont utilisée comme *Predictors* (variables indépendantes) pour *prédire* les valeurs de TOTAL_COST.
- La valeur minimale de résiduels est de -900.45£ tandis que la valeur maximale de residuals est de 2 239.12£, il existe une nette différence entre les deux valeurs (par conséquent, grande possibilité d'avoir des données aberrantes).
- La valeur de la médiane (-20.21 £) est quant à elle relativement éloignée de 0, nous avons alors une Asymétrie vers la droite en ce qui concerne la distribution des résiduels.
- Les coefficients du modèle sont clairement tous significatifs, et aussi, vu que la p-value associée au F-Test est bien inférieure à 1%, le modèle lui-même est globalement significatif.

Analyse résiduelle :

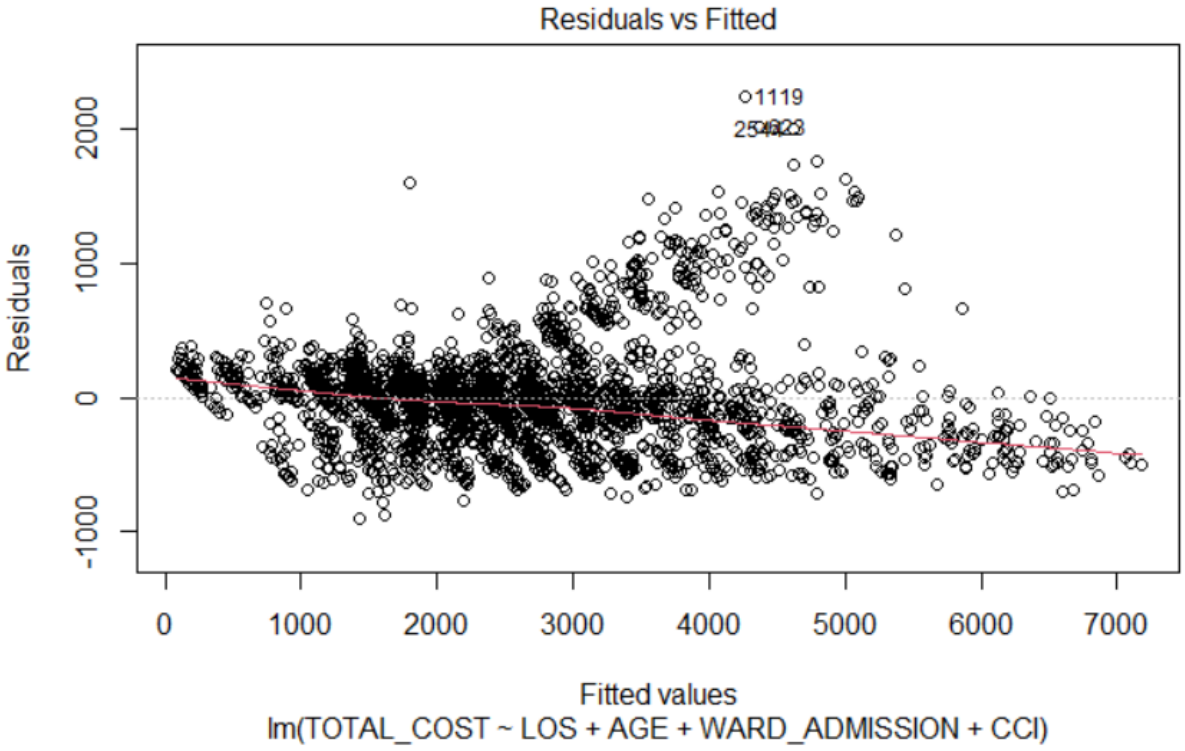
```
df$residuals <- residuals(reg)
```

```
shapiro.test(df$residuals)

Shapiro-Wilk normality test

data:  df$residuals
W = 0.90298, p-value < 2.2e-16
```

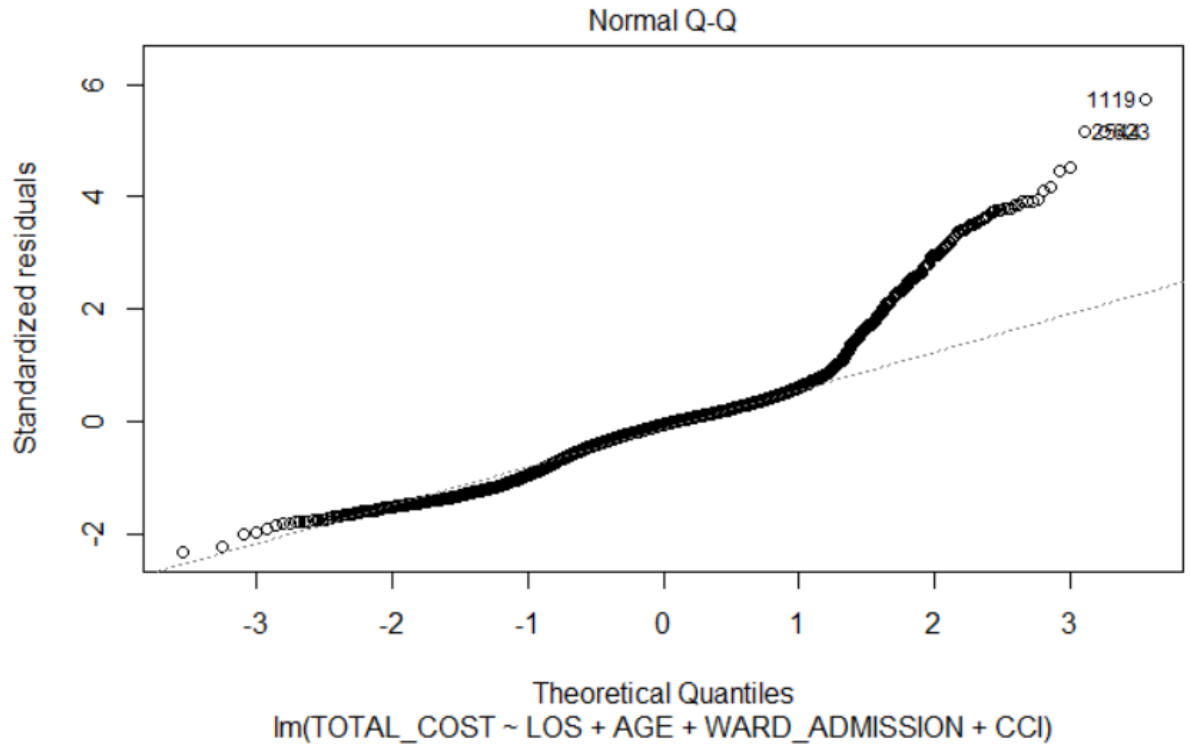
- La distribution des résidus ne suit pas une loi normale (donc, Asymétrique).



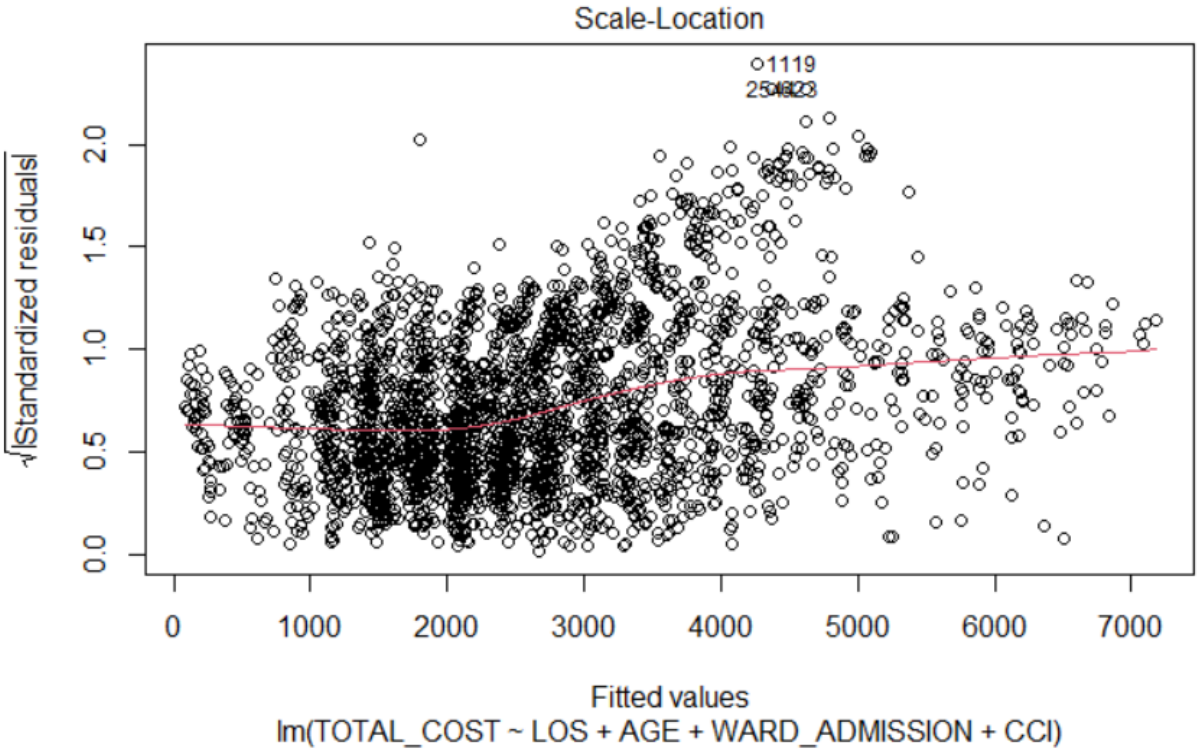
- Ici, nous pouvons remarquer que la relation est relativement Positive.



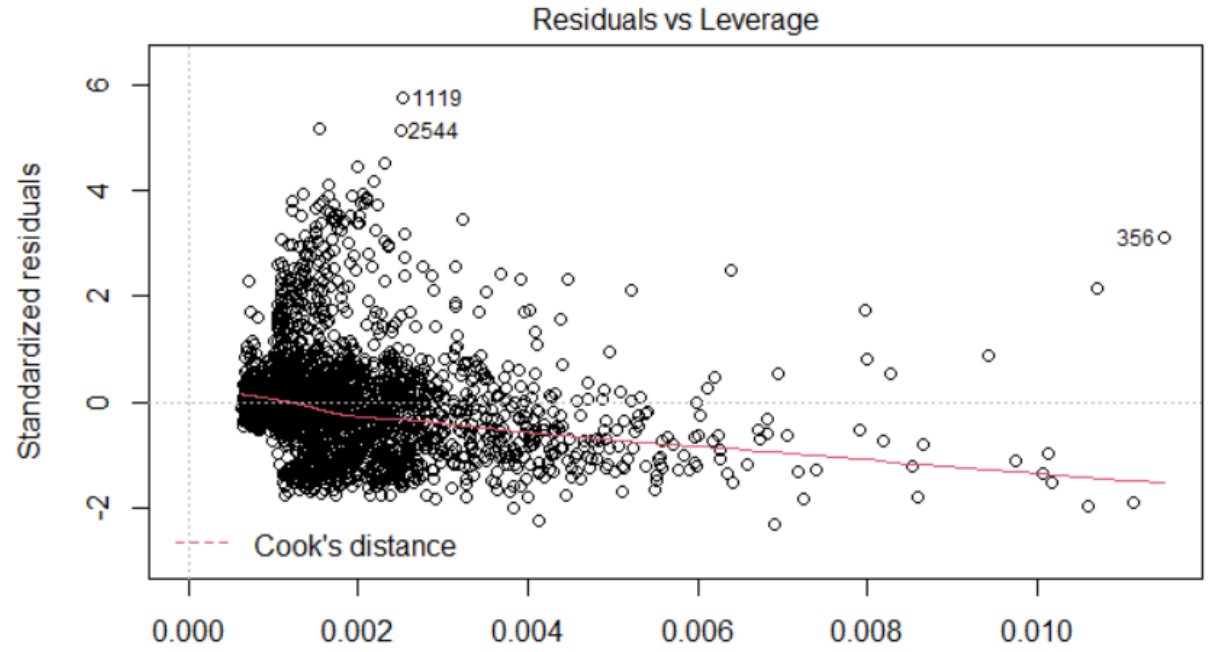
- La linéarité est relativement violée et nous pouvons constater l'existence de données aberrantes par rapport aux résidus, plus précisément aux alentours de la valeur 2 000 £.

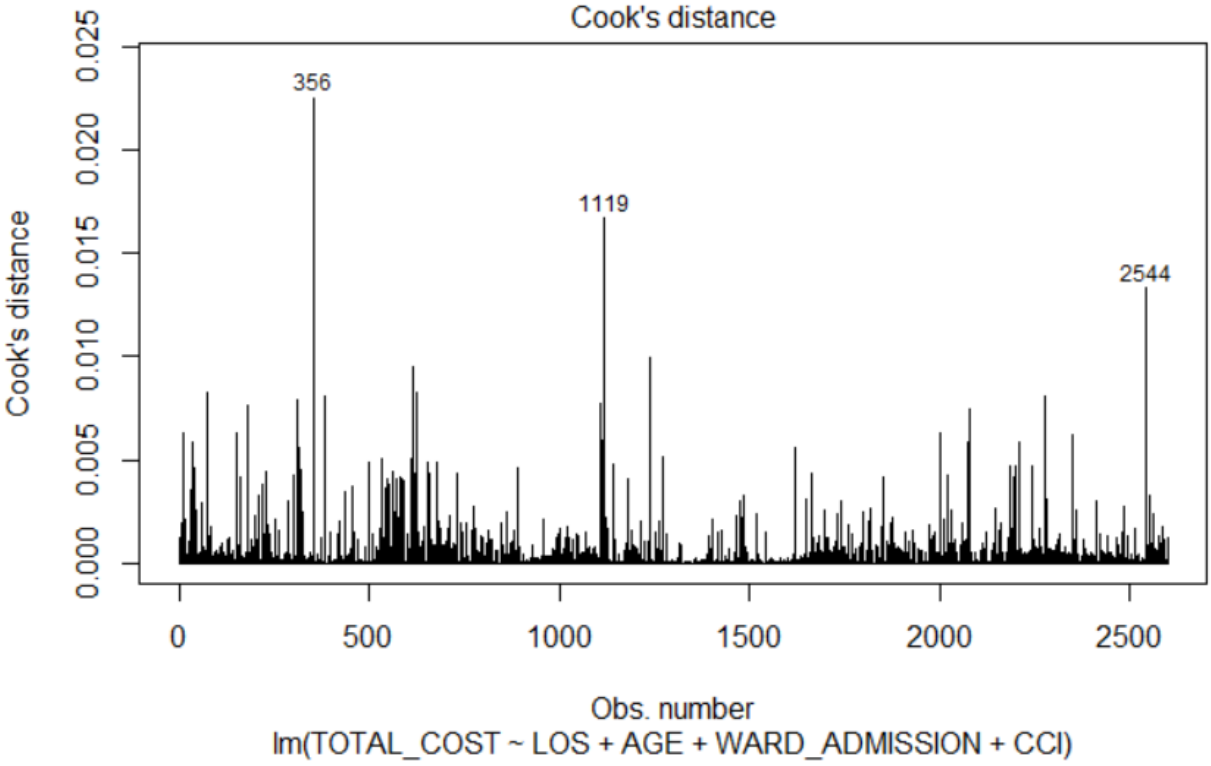


- Le Quantile-Quantile plot ci-dessus nous informe d'avantage sur l'existence de données aberrantes aux extrémités. Des données aberrantes qui s'avèrent être plus importantes au niveau de l'extrémité supérieure.



- A partir du Scale-Location plot ci-dessus, la *ligne rouge* n'est pas encore clairement suffisamment horizontale pour satisfaire *l'hypothèse d'Homoscédasticité* pour notre modèle.
- Les outliers peuvent être observées aux niveaux supérieurs de racines carrées des résidus standardisés.





IDADMISSION <dbl>	TOTAL_COST <dbl>	WARD_ADMISSION <chr>	AGE <dbl>	CCI <dbl>
16005375	6574.57	24	18	2
16000036	6505.56	24	90	1
16018298	6628.54	24	86	1

3 rows

```
summary(df$TOTAL_COST)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    285    1696    2423    2676    3442    6673
```

- En se basant sur les deux précédents Plots de « Residuals vs. Leverage » et de « Cook's distance », puis vérifiés et surtout identifiés à travers le tableau de résumé correspondant, nous pouvons bien confirmer que les résidus #356, #1119 et #2544 (données aberrantes) sont des points d'influences.
- Il serait alors préférable de passer par la suppression de ces données aberrantes avant d'entamer une nouvelle itération (Itération 2).

II.2 – Itération 2 :

```
dfn <- df %>%
  filter(! IDADMISSION %in% c(16005375,16000036, 16018298 ))
```

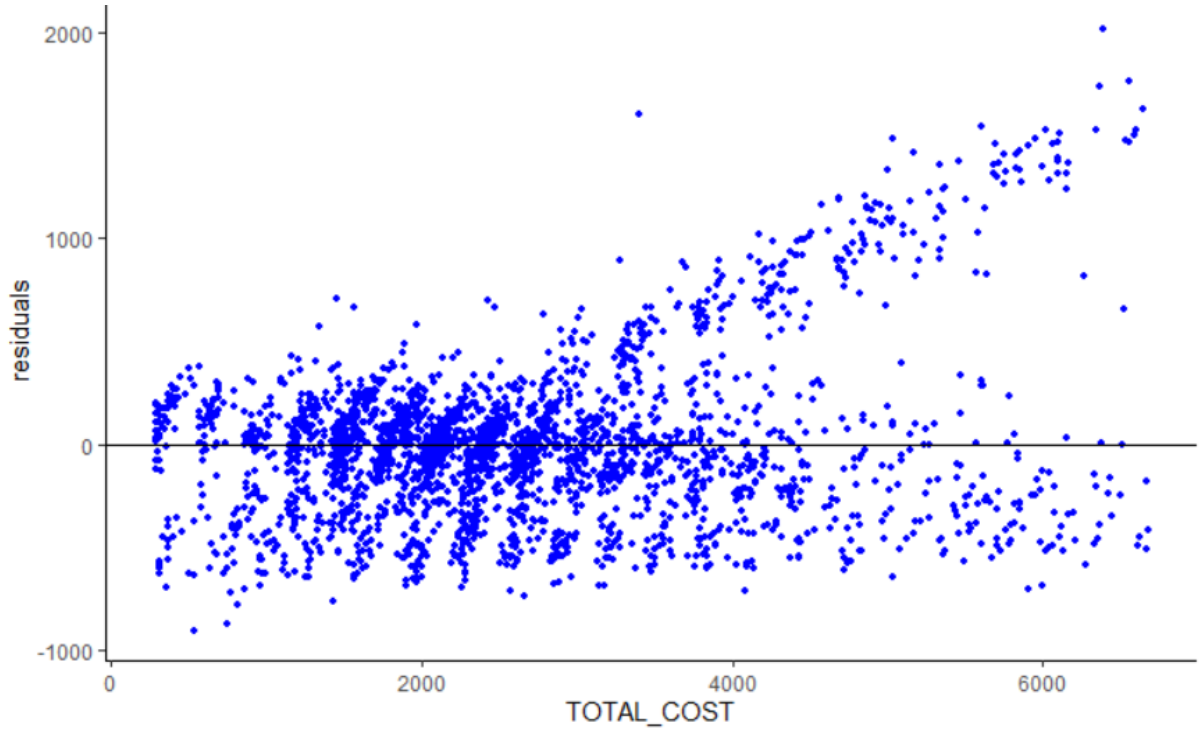
- En suivant la directive déduite en fin de l'Itération 1, les données aberrantes détectées ont été supprimées.

```
dfn %>%
  mutate(RISKDEATH = as.character(RISKDEATH),
         WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'G
eneralist', 'Specialist')
  ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION + CCI,
      data = .
  ) -> regn
summary(regn)

Call:
lm(formula = TOTAL_COST ~ LOS + AGE + WARD_ADMISSION + CCI, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-894.87 -222.12  -19.35  131.28 2024.10

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           754.6761    53.4631  14.116  < 2e-16 ***
```

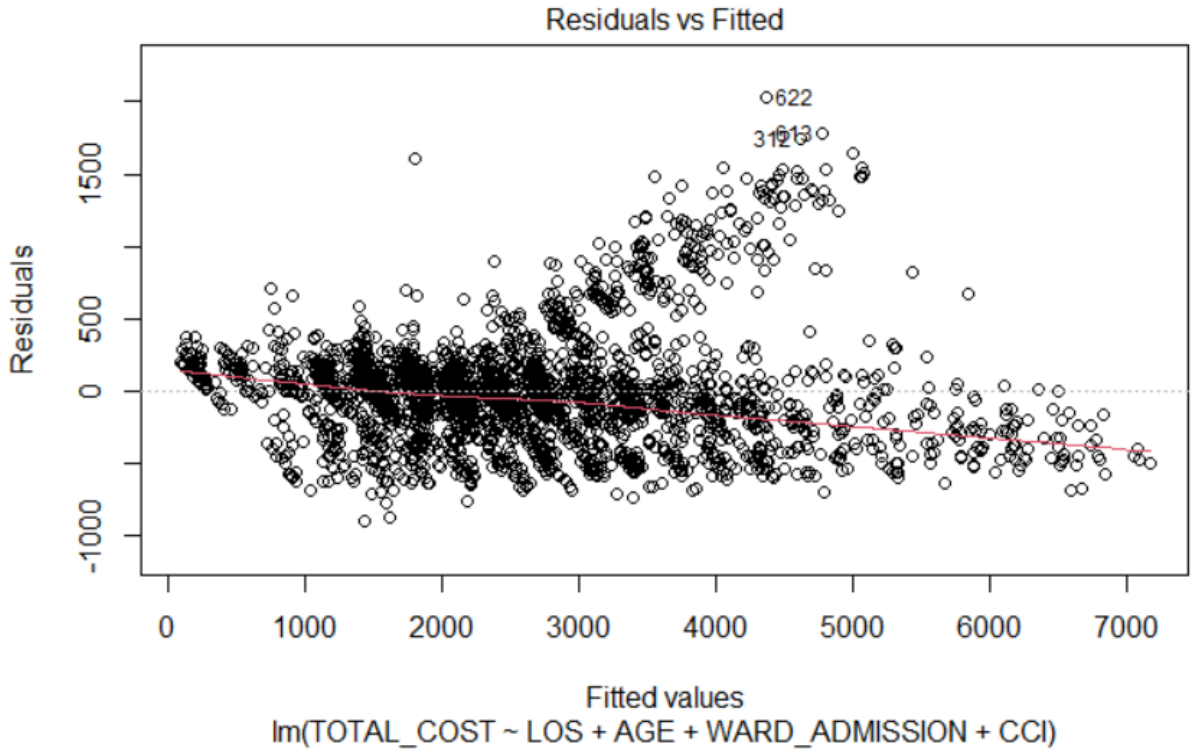
```
LOS                         310.9496     1.8843 165.023  < 2e-16 ***
AGE                         -10.5893     0.6344 -16.692  < 2e-16 ***
WARD_ADMISSIONSpecialist    699.6224    16.2275  43.113  < 2e-16 ***
CCI                          23.8703     7.1134   3.356 0.000803 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 385.4 on 2590 degrees of freedom
Multiple R-squared:  0.918, Adjusted R-squared:  0.9179
F-statistic:  7249 on 4 and 2590 DF,  p-value: < 2.2e-16
```

- La valeur minimale de résiduels a été mise à jour à -894.87 £ (une légère augmentation) tandis que la nouvelle valeur maximale de résiduels est 2 024.10£ (une diminution a été constatée). Néanmoins, il existe toujours une différence assez conséquentes entre les deux valeurs aux extrémités (des données aberrantes sont alors encore présentes).
- La nouvelle valeur de la médiane (-19.35 £) a certes été revue à la hausse, mais demeure encore relativement éloignée de 0, nous avons toujours une Asymétrie vers la droite en ce qui concerne la distribution des résiduels.
- Les coefficients du modèle ont été aussi mis à jour, avec notamment une nette amélioration pour le cas du coefficient lié au CCI.
- La p-value associée au F-Test est plus ou moins restée la même qu'en Itération 1, donc toujours inférieure à 1% : Notre modèle est toujours significatif.

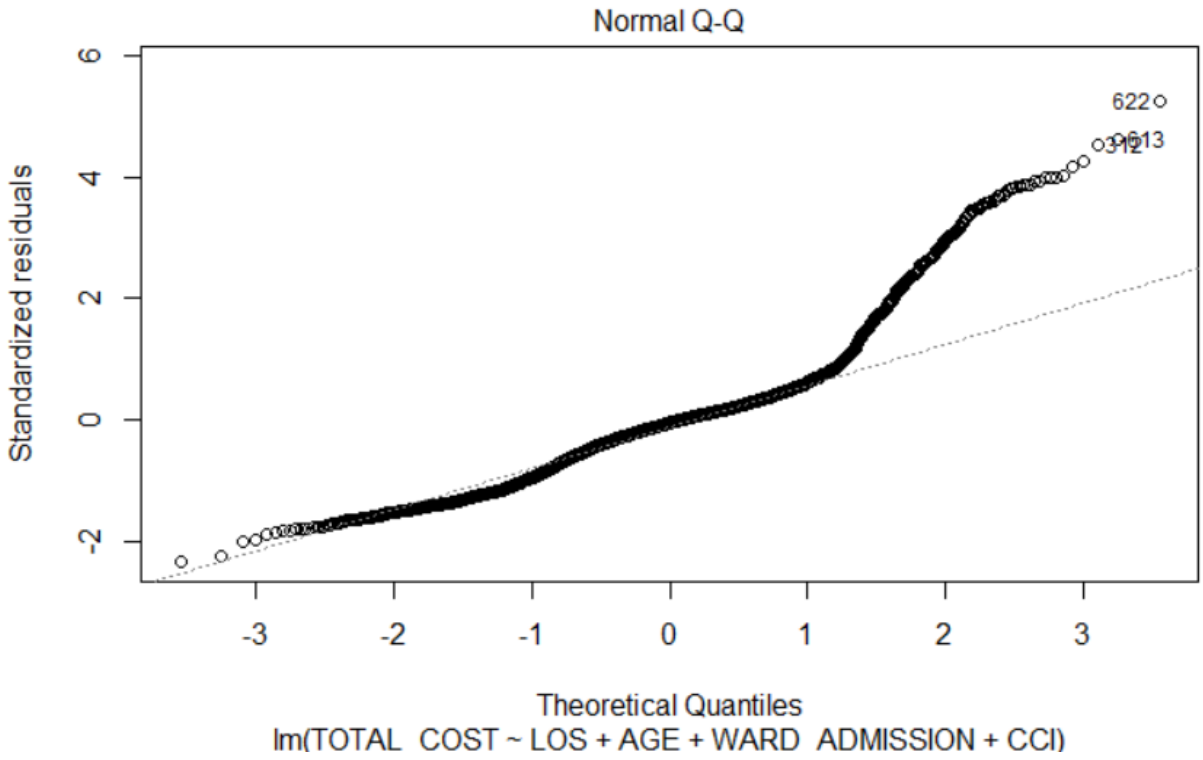
Analyse résiduelle :



- Nous pouvons remarquer que la relation demeure relativement Positive.

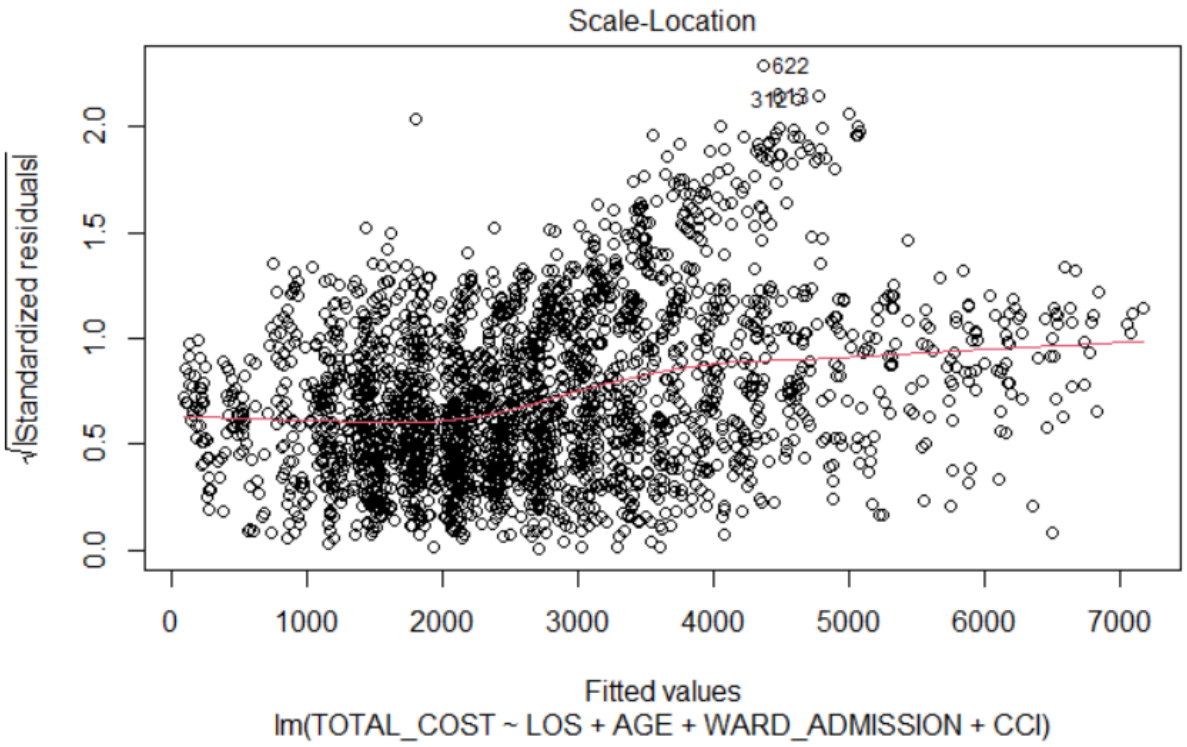


- La linéarité est encore relativement violée et nous pouvons constater l'existence de nouvelles données aberrantes par rapport aux résidus, plus précisément entre les valeurs de 1 500£ et de 1 750£.

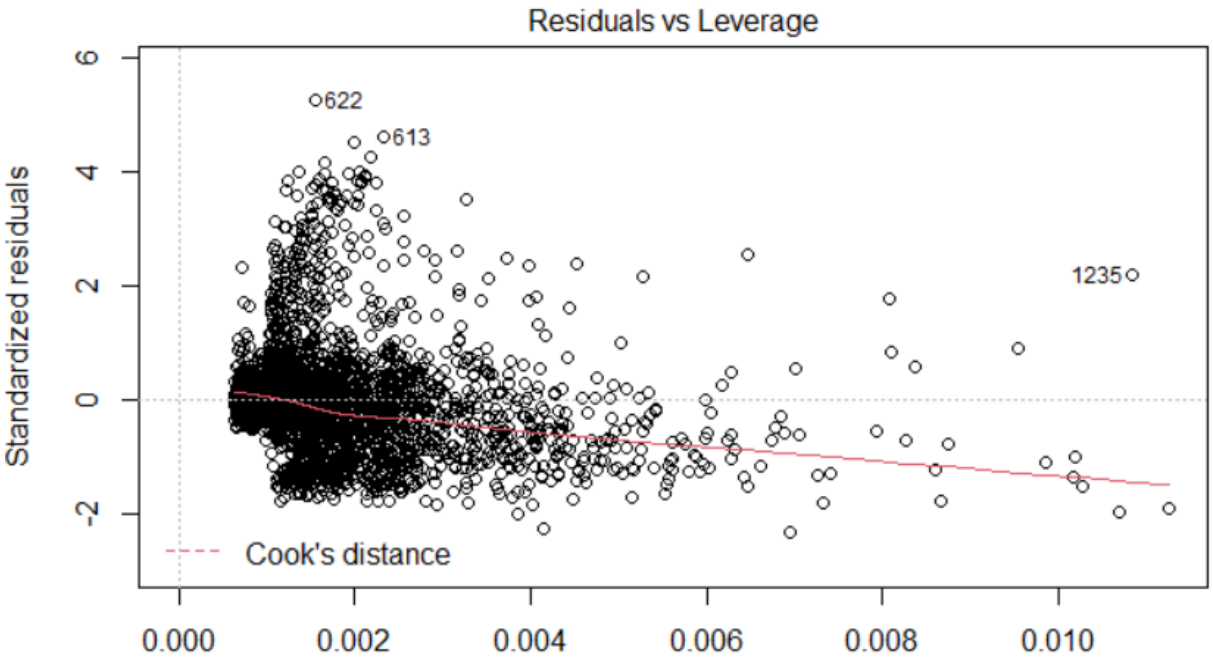


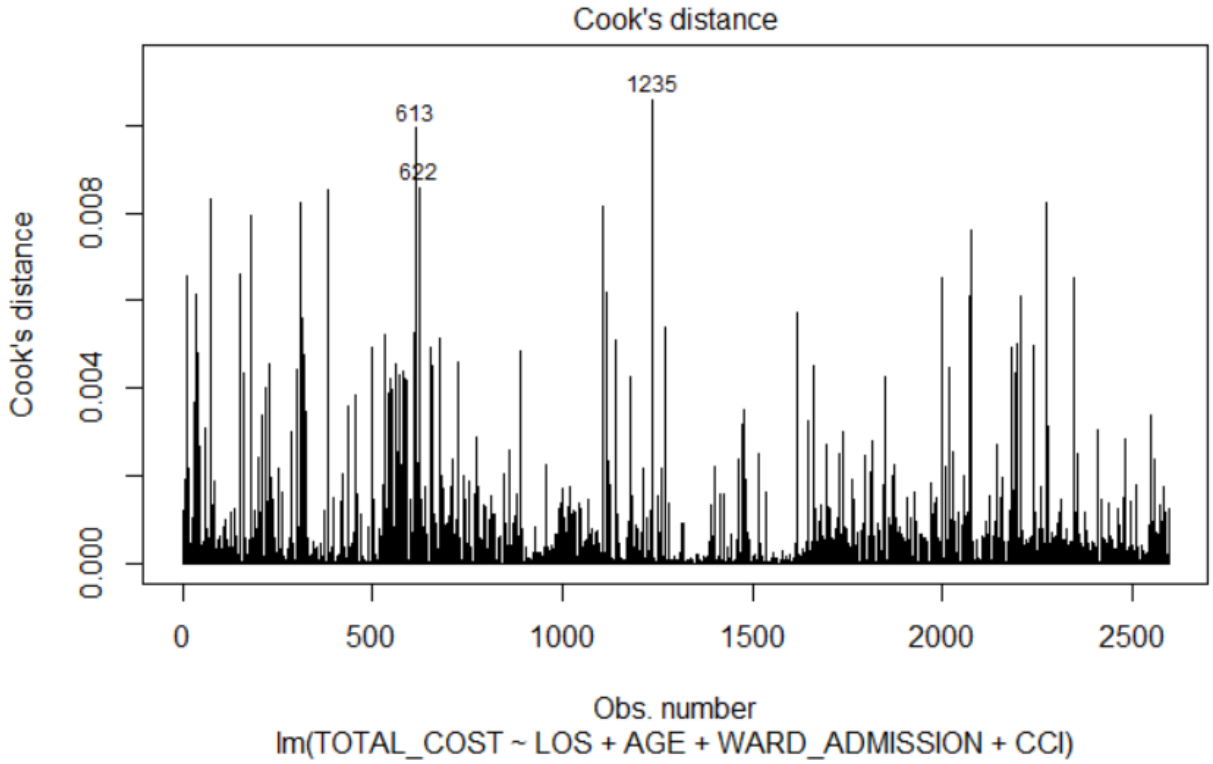
- Le Quantile-Quantile plot ci-dessus nous informe d'avantage sur l'existence de ces nouvelles données aberrantes aux extrémités.

- Toujours le même comportement observé, les données aberrantes s'avèrent être plus importantes au niveau de l'extrémité supérieure.



- A partir du Scale-Location plot ci-dessus, on peut clairement remarquer que la ligne rouge n'est pas encore suffisamment horizontale pour satisfaire *l'hypothèse d'Homoscédasticité* pour notre modèle.
- Les données aberrantes peuvent être observées aux niveaux supérieurs des racines carrées des résidus standardisés.





- En se basant sur les deux précédents Plots de « Residuals vs. Leverage » et de « Cook's distance », nous pouvons bien confirmer que les résidus #1235, #613 et #622 (données aberrantes) peuvent être considérés comme étant des points d'influences.
- De nouveaux travaux de suppressions d'observations sont alors recommandés avant de passer à la prochaine Itération

II.3 – Itération 3 :

```
df2 <- df %>%
  filter(residuals <= 300) %>%
  select(all_of(column_origine))
```

- Ayant été abstraitement annoncée en fin de l'Itération 2, la suppression des observations dont la valeur des résidus est supérieure à 300 a été réalisée.

```
WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'Generalist
', 'Specialist')
  ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION,
      data = .
  ) -> reg2
summary(reg2)
```

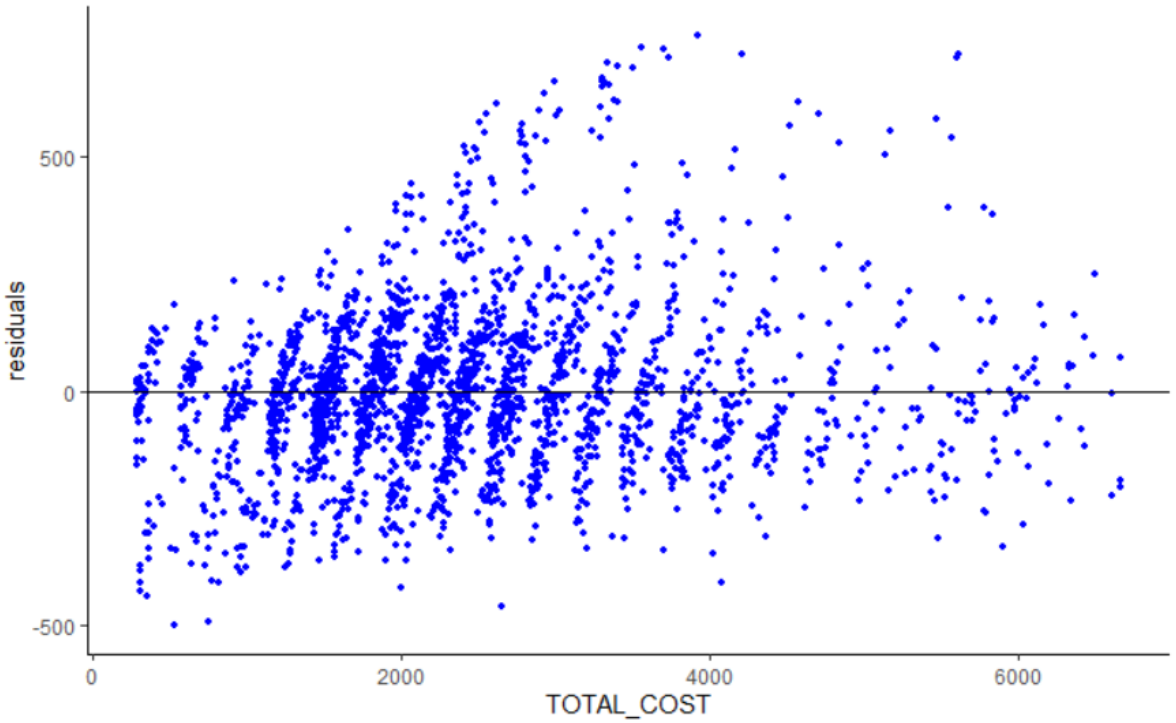
```
Call:
lm(formula = TOTAL_COST ~ LOS + AGE + WARD_ADMISSION, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-498.82 -105.65  -12.86   87.74  759.30

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               562.4525    26.1655   21.50   <2e-16 ***
LOS                       296.3324     0.9051  327.40   <2e-16 ***
AGE                        -6.2157     0.3099  -20.06   <2e-16 ***
WARD_ADMISSIONSpecialist 385.5801     8.1566   47.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 178.2 on 2265 degrees of freedom
Multiple R-squared:  0.9793,    Adjusted R-squared:  0.9793
F-statistic: 3.579e+04 on 3 and 2265 DF,  p-value: < 2.2e-16
```

- La valeur minimale de résidus a été mise à jour à -498.82 £ (une nette augmentation) tandis que la nouvelle valeur maximale de résidus est de 759.30 £ (une nette diminution a été constatée). Cependant, la différence entre les deux valeurs aux extrémités semble encore assez significative (possibilité d'existence de données aberrantes).
- La nouvelle valeur de la médiane (-12.86 £) a certes, encore une fois, été revue à la hausse, mais demeure encore relativement éloignée de 0, nous avons toujours une Asymétrie vers la droite en ce qui concerne la distribution des résidus.
- Les coefficients du modèle sont restés à peu près les mêmes qu'en Itération 2 (toujours significatifs pour leur part dans ce cas), remarquons juste le fait que la CCI n'a plus été prise en compte en tant que Predictor pour la *prédiction* de valeurs de TOTAL_COST dans cette Itération 3.
- La p-value associée au F-Test est plus ou moins restée la même qu'en Itération 2, donc toujours inférieure à 1% : Notre modèle est toujours significatif.

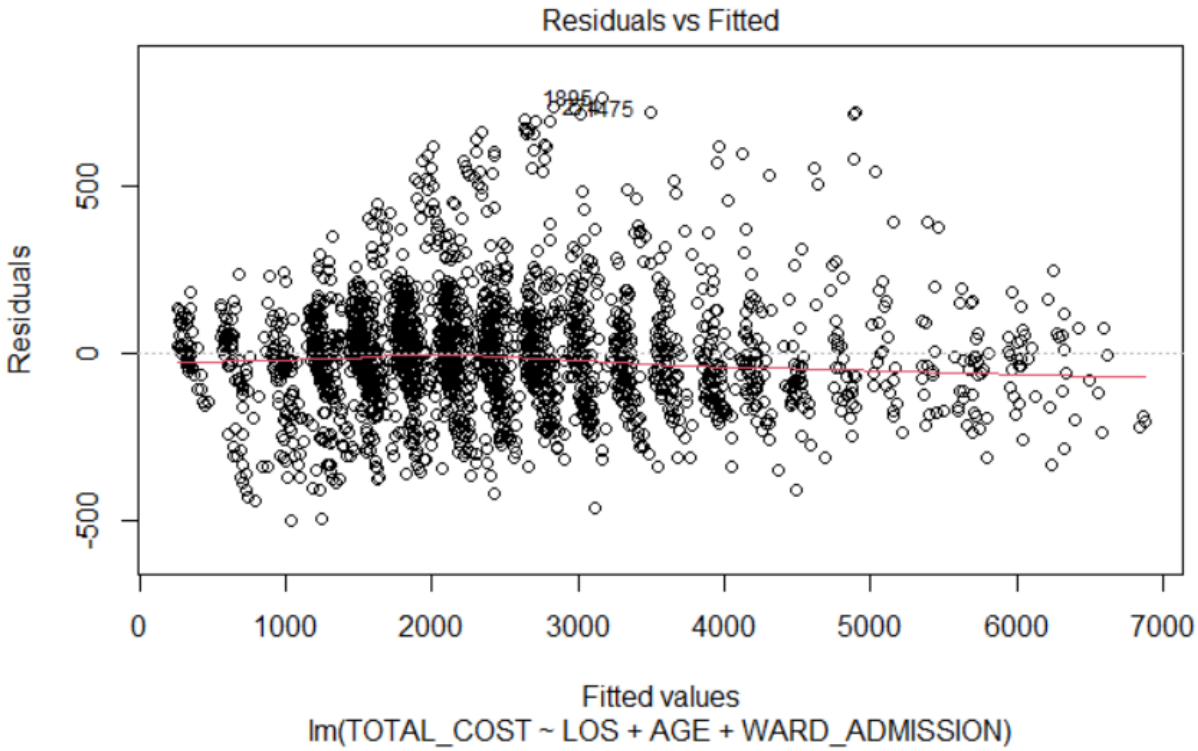
Analyse résiduelle :

```
shapiro.test(df2$residuals)

    Shapiro-Wilk normality test

data:  df2$residuals
W = 0.95338, p-value < 2.2e-16
```

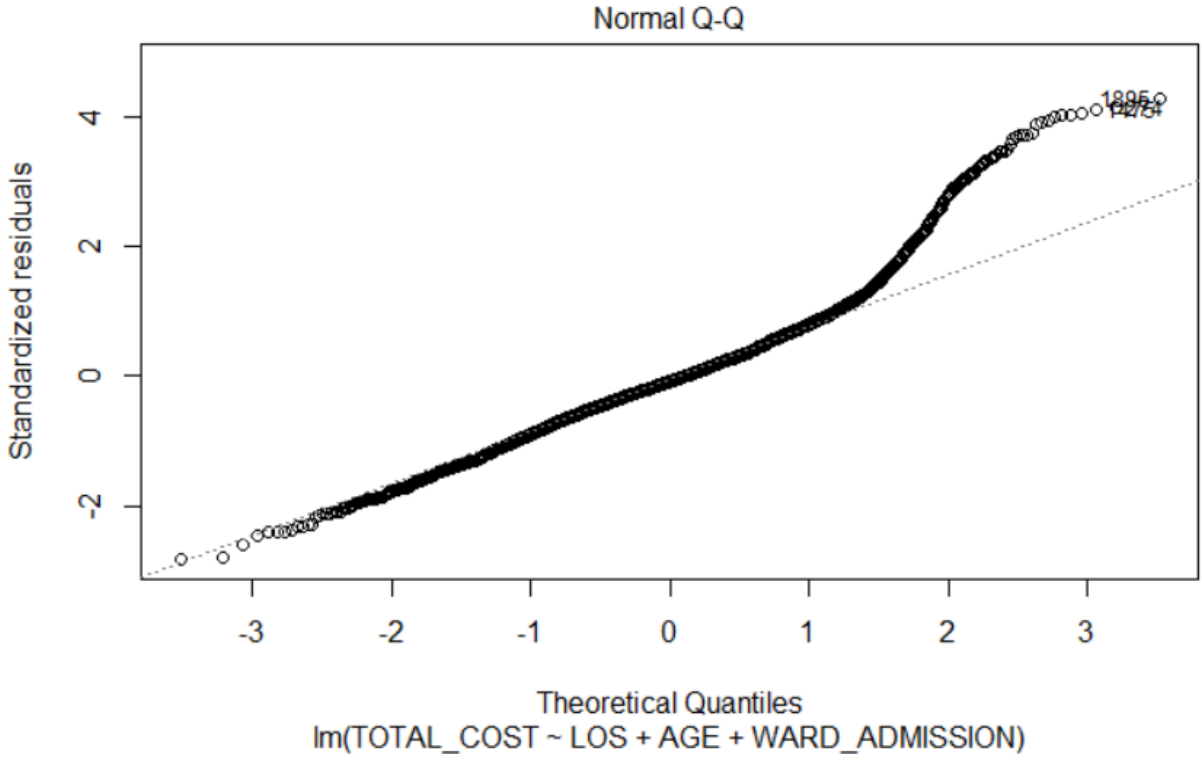
- La distribution des résidus ne suit pas une loi normale (donc, asymétrique).



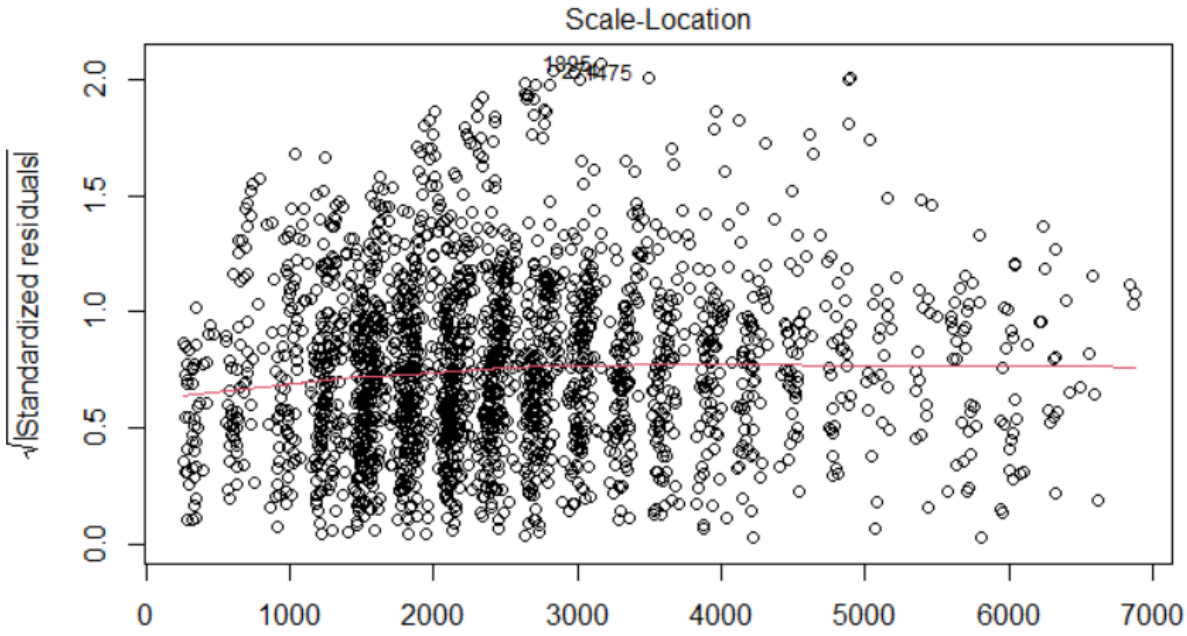
- Nous pouvons remarquer que la relation est devenue moins Positive qu'elle ne l'était auparavant, c'est–à-dire, comparée à ces précédentes versions correspondant respectivement aux précédentes Itérations.



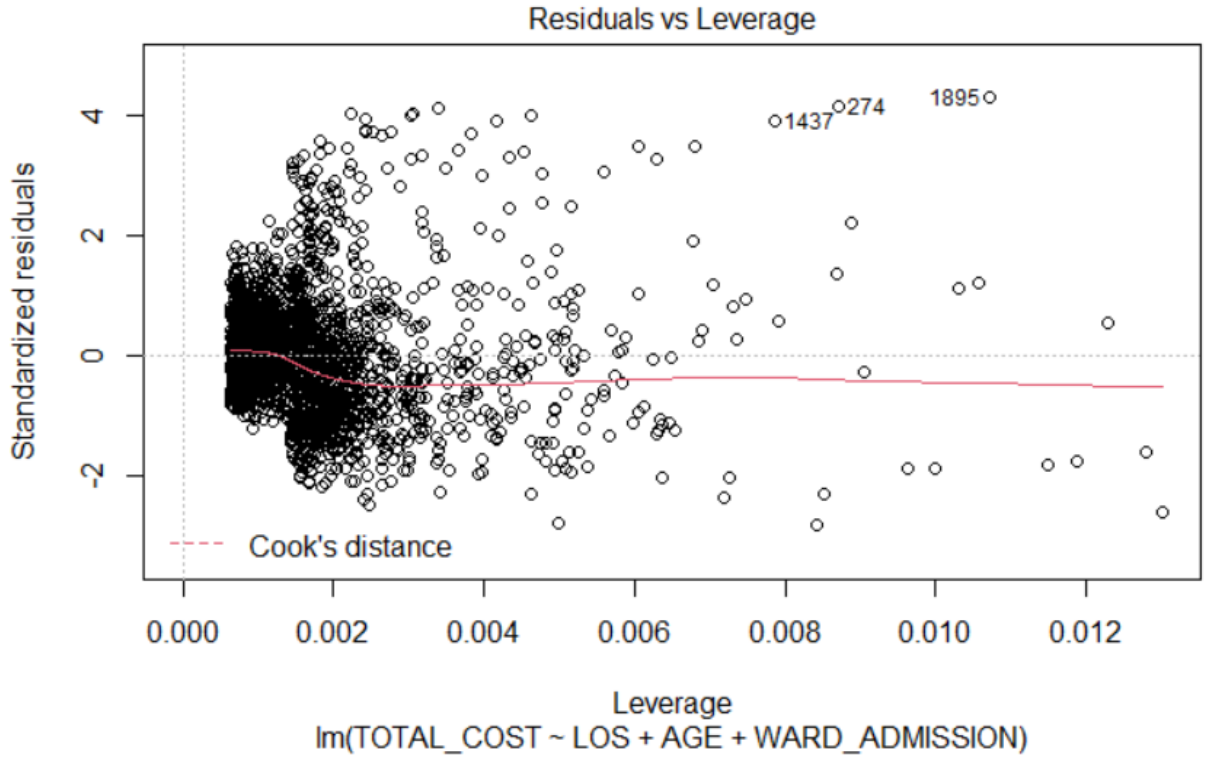
- La linéarité est désormais moins violée, mais nous pouvons cependant encore constater l'existence de nouvelles données aberrantes par rapport aux résidus, plus précisément aux environs de la valeur 750 £.



- Le Quantile-Quantile plot ci-dessus nous informe d'avantage sur l'existence de ces nouvelles données aberrantes aux extrémités, particulièrement au niveau de l'extrémité supérieure.



- A partir du Scale-Location plot ci-dessus, la ligne rouge est désormais toute proche d'être suffisamment horizontale pour satisfaire *l'hypothèse d'Homoscédasticité* pour notre modèle.
- Des données aberrantes, cependant, peuvent être encore observées aux niveaux supérieurs de racines carrées des résidus standardisés.



- En se basant sur les Plot de « Residuals vs. Leverage », nous pouvons toujours, soi-disant, confirmer que les résidus #1427, #274 et #1895 (données aberrantes) peuvent être considérés comme étant des points d'influences.

- A nouveau, des travaux de suppressions d'observations sont alors recommandés avant de passer à la prochaine Itération.

II.4 – Itération 4 :

```
df3 <- df2 %>%
  filter(residuals <= 300) %>%
  select(all_of(column_origine))
```

- Ayant été abstraitement annoncée en fin de l'Itération 3, une nouvelle séance de suppression des observations dont la valeur des résidus est supérieure à 300 a été à nouveau réalisée.

```
df3 %>%
  mutate(RISKDEATH = as.character(RISKDEATH),
         WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'G
eneralist', 'Specialist')
  ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION + CCI,
      data = .
  ) -> reg3
summary(reg3)
```

```
Call:
lm(formula = TOTAL_COST ~ LOS + AGE + WARD_ADMISSION + CCI, data = .)

Residuals:
    Min      1Q  Median      3Q     Max
-331.12  -85.99  -13.86   72.20  476.06

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              326.9121    20.0412  16.312  < 2e-16 ***
LOS                      294.6779     0.6579 447.892  < 2e-16 ***
AGE                       -3.5936     0.2452 -14.654  < 2e-16 ***
WARD_ADMISSIONSpecialist 300.3605     6.3172  47.546  < 2e-16 ***
CCI                       11.7058     2.4647   4.749 2.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127.1 on 2141 degrees of freedom
Multiple R-squared:  0.9895,    Adjusted R-squared:  0.9895
F-statistic: 5.053e+04 on 4 and 2141 DF, p-value: < 2.2e-16
```

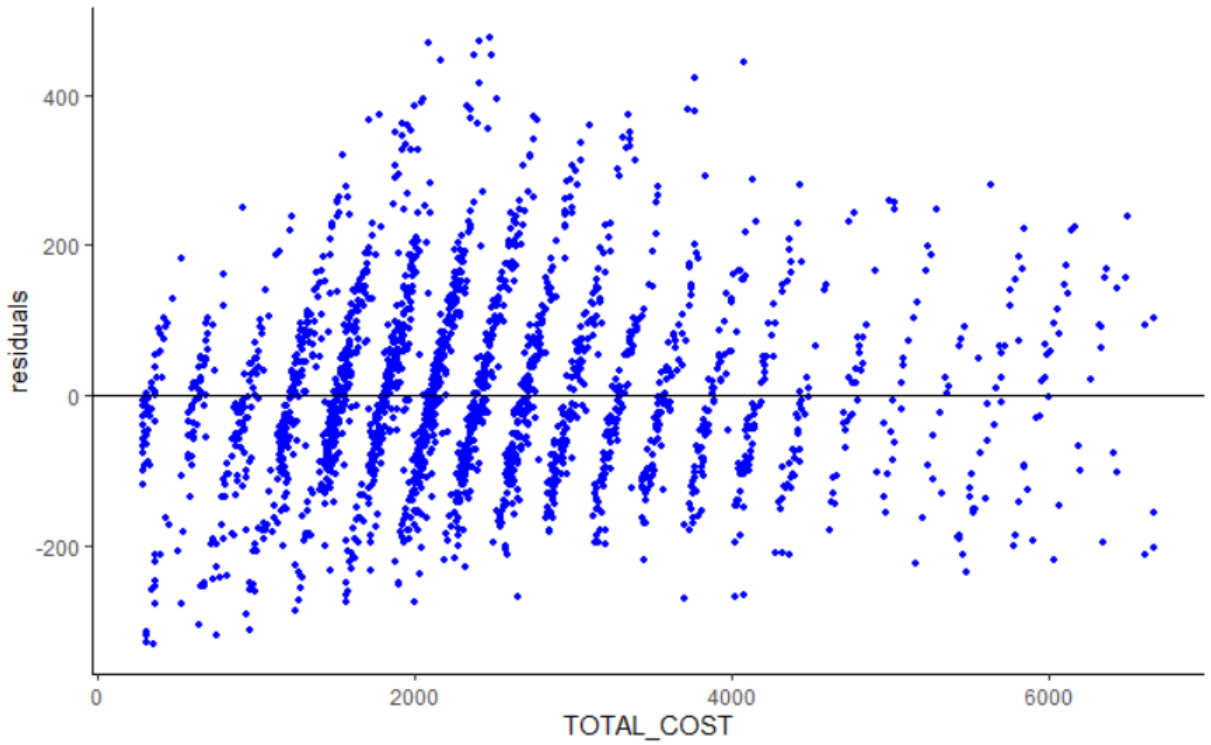
- La valeur minimale de résidus a été mise à jour à -331.12 £ (une augmentation constatée) tandis que la nouvelle valeur maximale de résidus est de 476.06 £ (une diminution a été constatée).

- La nouvelle valeur de la médiane (-13.86 £) a pour sa part légèrement diminuée comparée à celle vue en Itération 3, et donc, logiquement, demeure encore relativement éloignée de 0. Nous avons toujours une Asymétrie vers la droite en ce qui concerne la distribution des résidus.
- Les coefficients du modèle sont restés à peu près les mêmes qu'en Itération 3 (toujours significatifs pour leur part dans ce cas), à noter juste le fait que celui le CCI a de nouveau été prise en compte en tant que Predictor pour la *prédiction* de valeurs de TOTAL_COST dans cette Itération 4, et que le coefficient qui lui correspond s'est nettement amélioré comparé à sa dernière valeur observée en Itération 2.
- La p-value associée au F-Test est plus ou moins restée la même qu'en Itération 3, donc toujours inférieure à 1% : Notre modèle est toujours lui-même significatif.

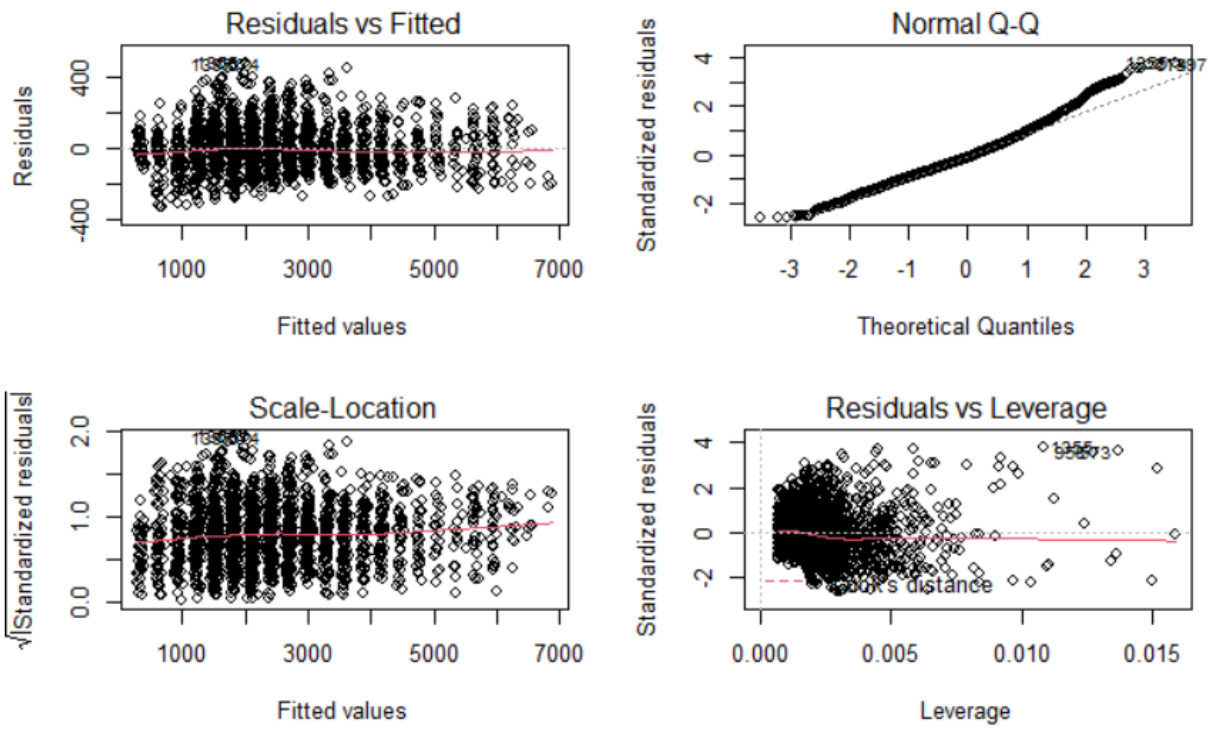
Analyse résiduelle :

```
df3$residuals <- residuals(reg3)
shapiro.test(df3$residuals)

Shapiro-Wilk normality test

data:  df3$residuals
W = 0.98164, p-value = 5.27e-16
```

- La distribution des résidus ne suit pas une loi normale (donc, asymétrique).



- Nous pouvons remarquer que la relation n'est plus que TRES LEGEREMENT positive, non plus comme elle a successivement durant les précédentes Itérations.



- D'après ce que peut nous indiquer le plot sur les « Residuals vs Fitted », la linéarité est désormais *plus respectée* et aussi, les données aberrantes, certes quelques-unes persistent encore, semblent désormais être *plus raisonnables*.
- Le Quantile-Quantile plot aussi confirme cette tendance de linéarité respectée de notre modèle, avec les données aberrantes qui sont désormais, rappelons-le, assez raisonnables.
- A partir du Scale-Location, la ligne rouge est désormais suffisamment horizontale pour satisfaire *l'hypothèse d'Homoscédasticité* pour notre modèle.
- Enfin, en se basant sur le Plots de « Residuals vs. Leverage », nous pouvons bien confirmer que les résidus correspondant aux données aberrantes qui continuent de persister ne constituent plus forcément des points d'influences