Hospital Data Analysis - Regression

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 of the outliers, precisely :
  - Those that we found in function of the admissions' TOTAL_COST;
  - Those that we found in function of the AGE of the patient concerned by the admissions.

II- Regression :

II.1 – 1st Iteration :

```
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
```

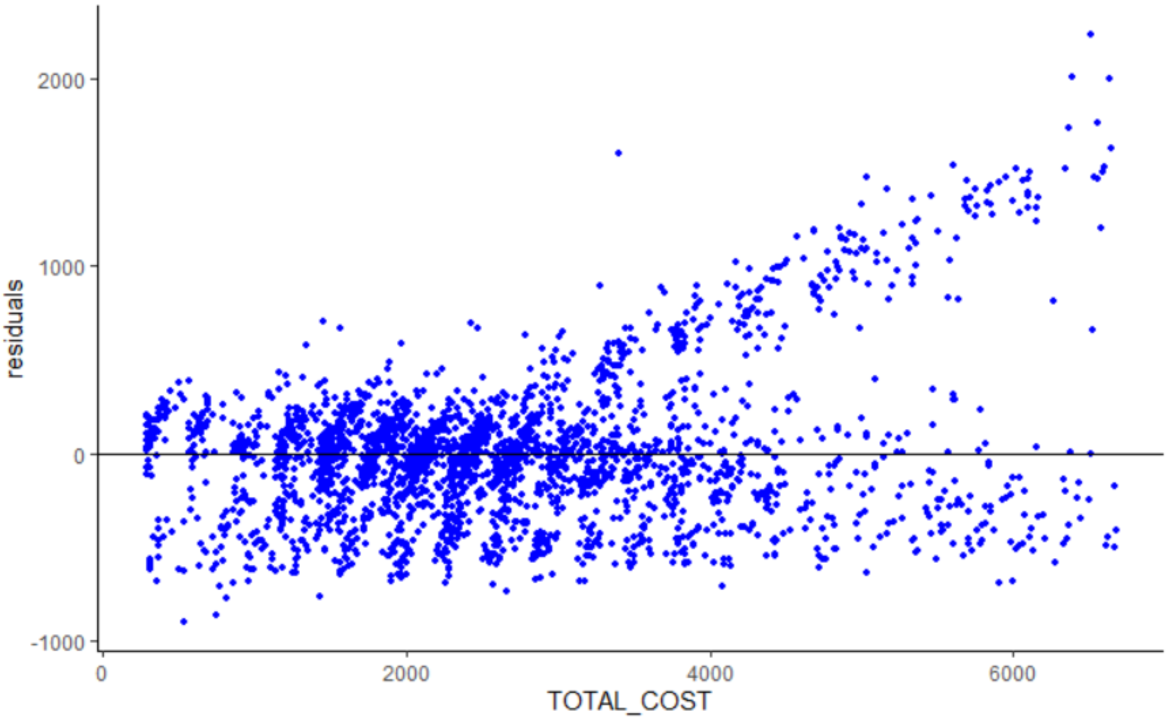
- The variables LOS, AGE, WARD_ADMISSION and CCI are used as *Predictors* (independent variables) for the *prediction* of the TOTAL_COST's values;
- The minimum value of residual is -900.45 € whereas the maximum one is 2 239.12 €, there is a clear difference between the two values (therefore, a high probability of having outliers);
- The Median's value (-20.21 €) is relatively far from 0, we therefore have an Asymmetry to the right when it comes to the distribution of residuals;
- The model's coefficients are all obviously significant, and also, given than the p-value corresponding to the F-Test is well below 1%, the model itself is globally significant.

Residual analysis:

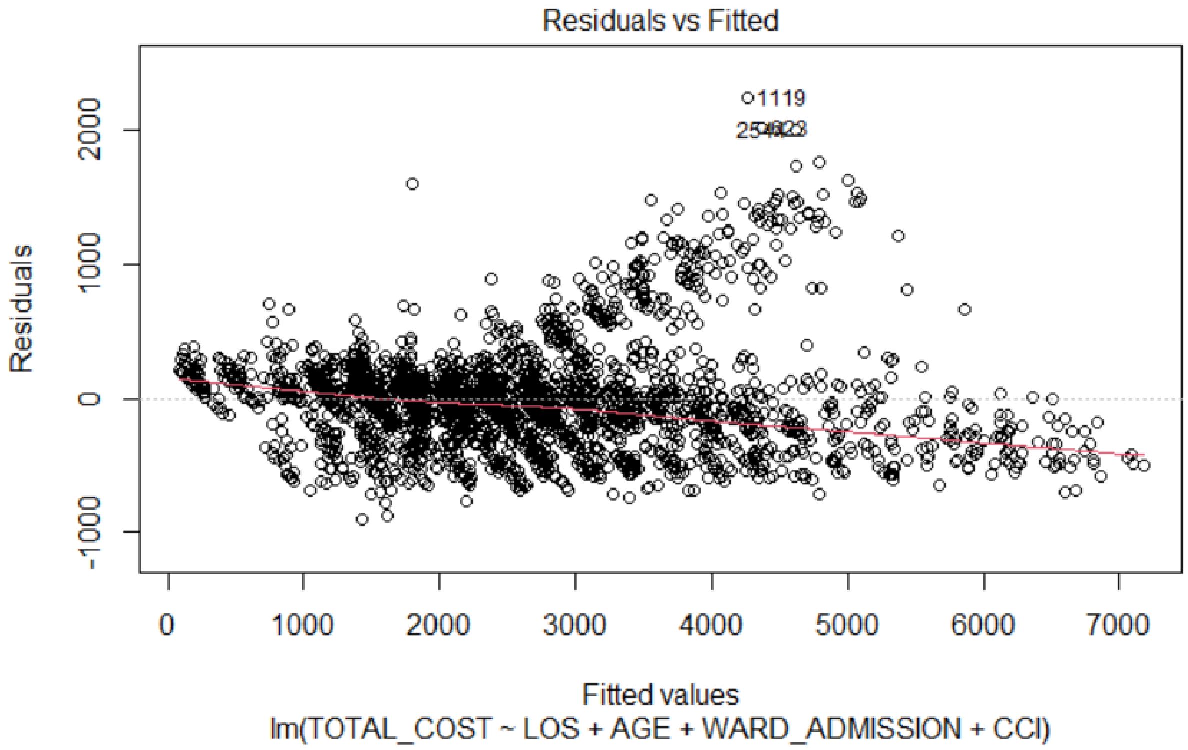
```
df$residuals <- residuals(reg)
```

```
shapiro.test(df$residuals)

Shapiro-Wilk normality test

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

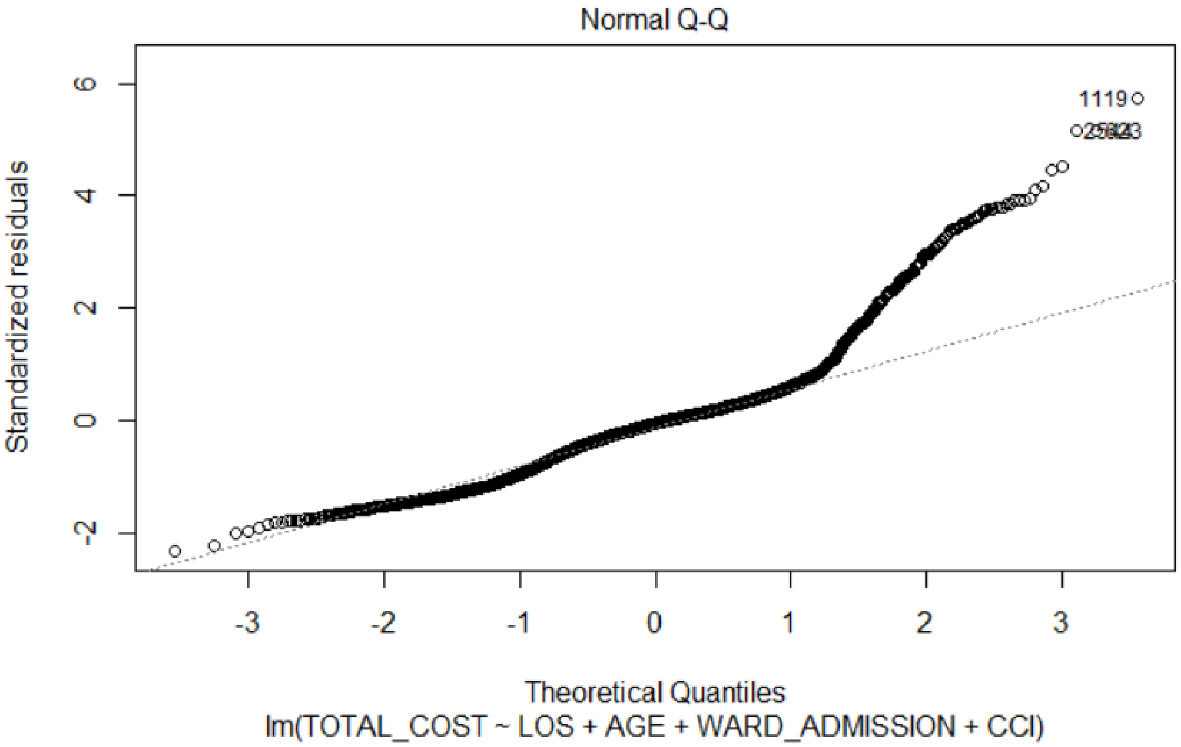
- The residuals, as a variable, doesn't follow a normal distribution;



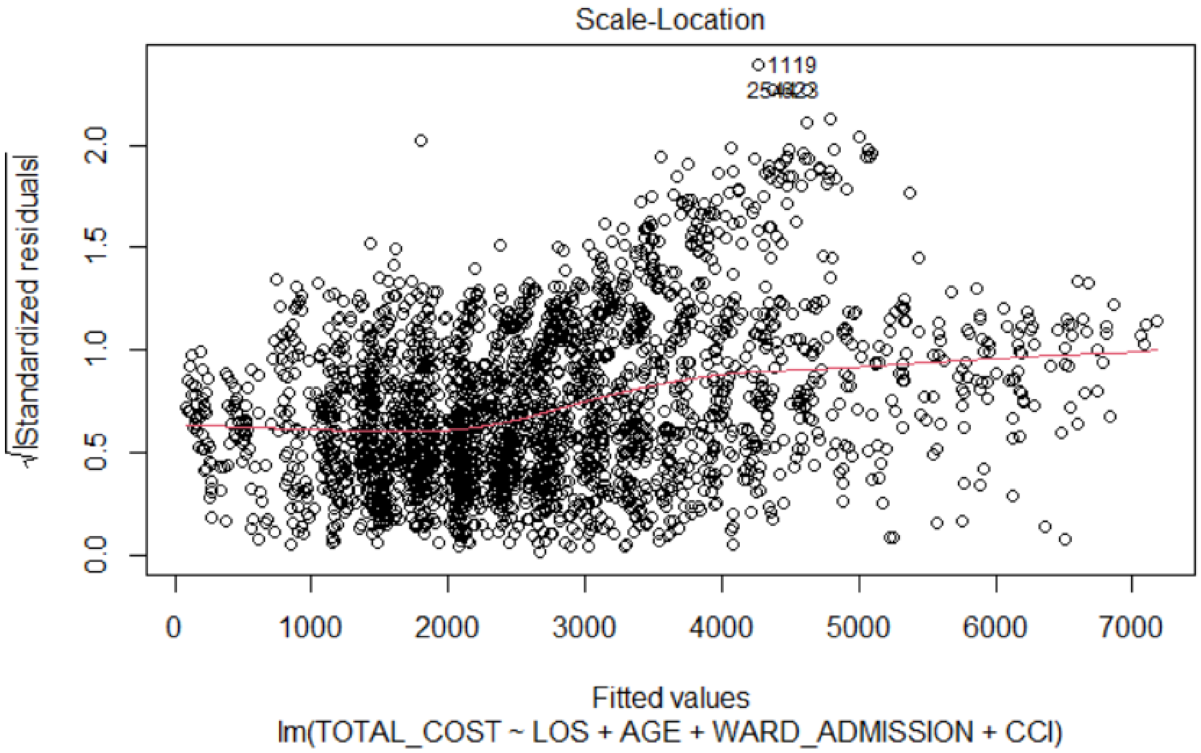
- Here, we can observe that the relationship is relatively Positive;



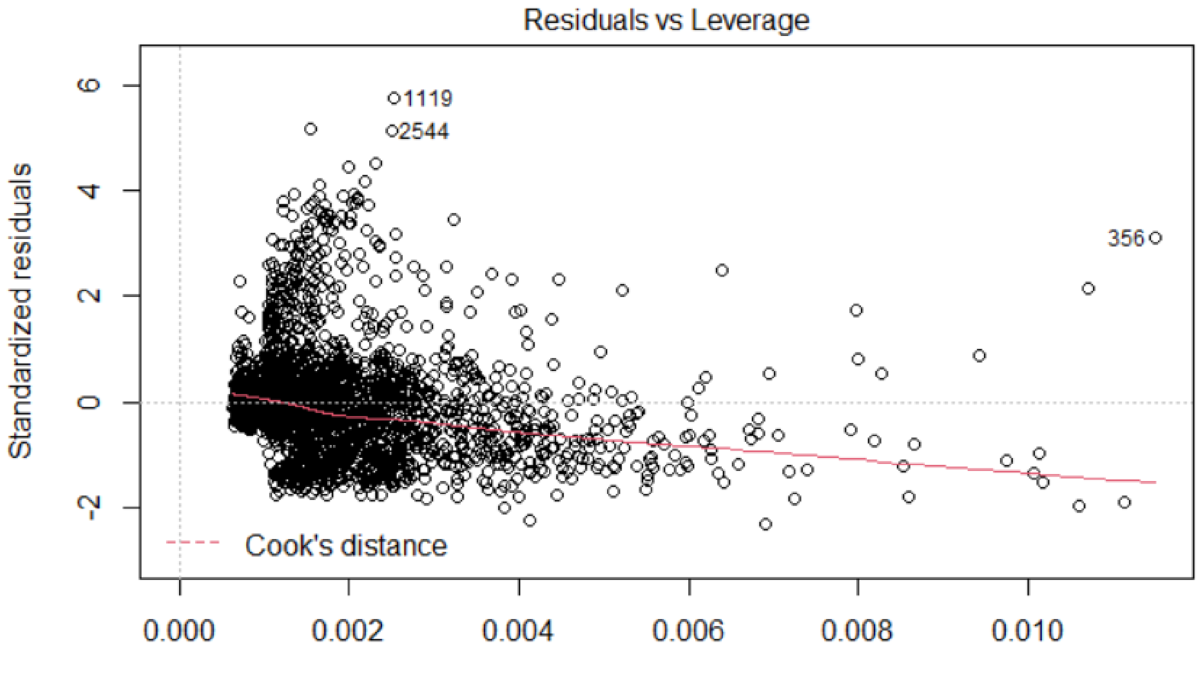
- The linearity is relatively not respected and we can observe the existence of outliers relating to residuals, more precisely around the value 2 000 €;

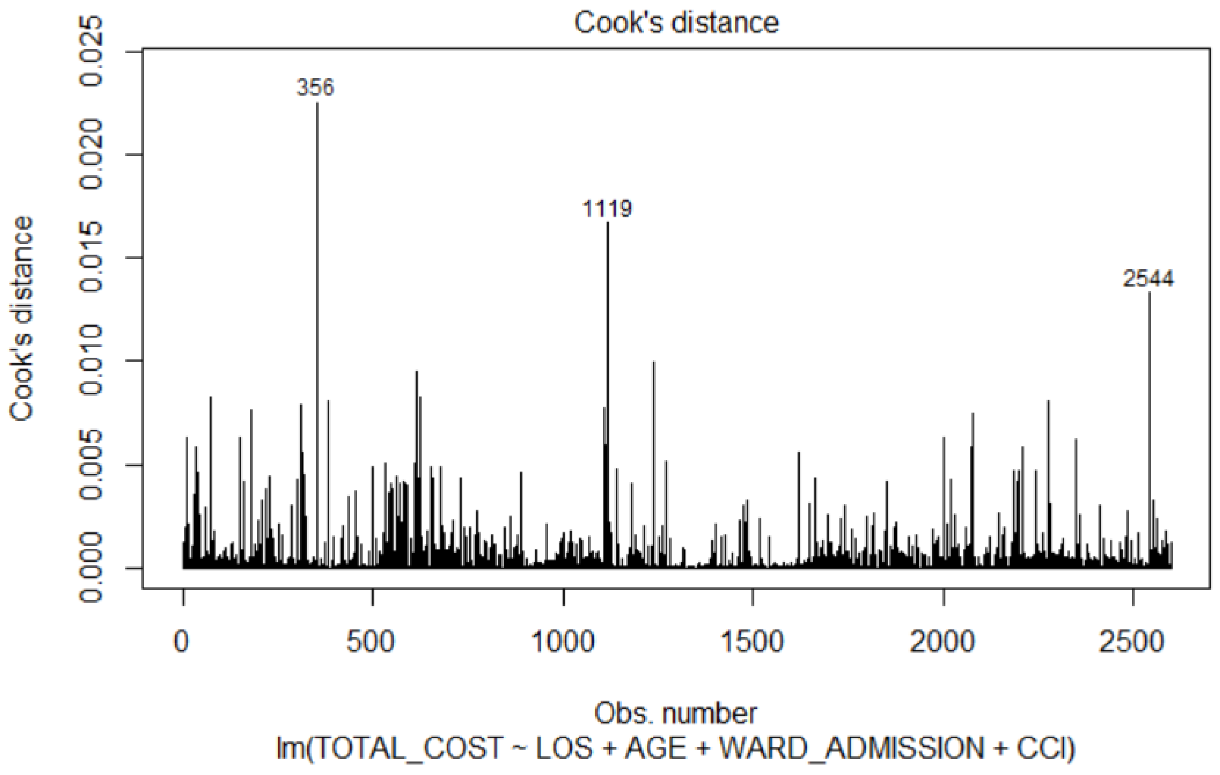


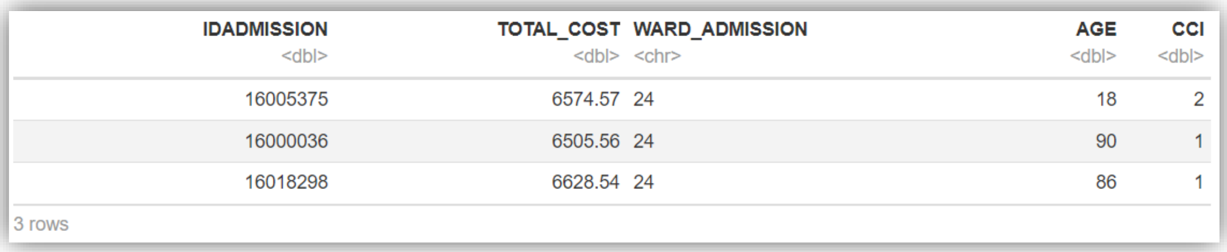
- The Quantile-Quantile plot above tells us more about this existence of outliers at the both ends. Outliers that seem to be more significant at the level of the upper end.



- From the Scale-Location plot above, the *red line* is clearly not enough horizontally straight yet for our model to satisfy the *hypothesis for Homoscedasticity* ;
- Outliers can be observed at the upper levels of the standardized residuals' square roots ;

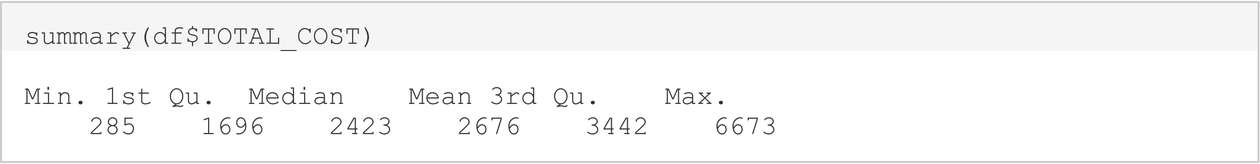






IDADMISSION <dbl>	TOTAL_COST <dbl>	WARD_ADMISSION <chr>	AGE <dbl>	CCI <dbl>
16005375	6574.57	24	18	2
16000036	6505.56	24	90	1
16018298	6628.54	24	86	1

3 rows



```
summary(df$TOTAL_COST)

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

- Based on the two previous Plots of « Residuals vs. Leverage » and « Cook's distance », then verified and most importantly identified through the corresponding table of summary, we can confirm with confidence that the residuals #356, #1119 and #2544 (outliers) constitute influential points;
- It would be then preferable to remove those outliers before entering a new iteration (2nd Iteration).

II.2 – 2nd Iteration :

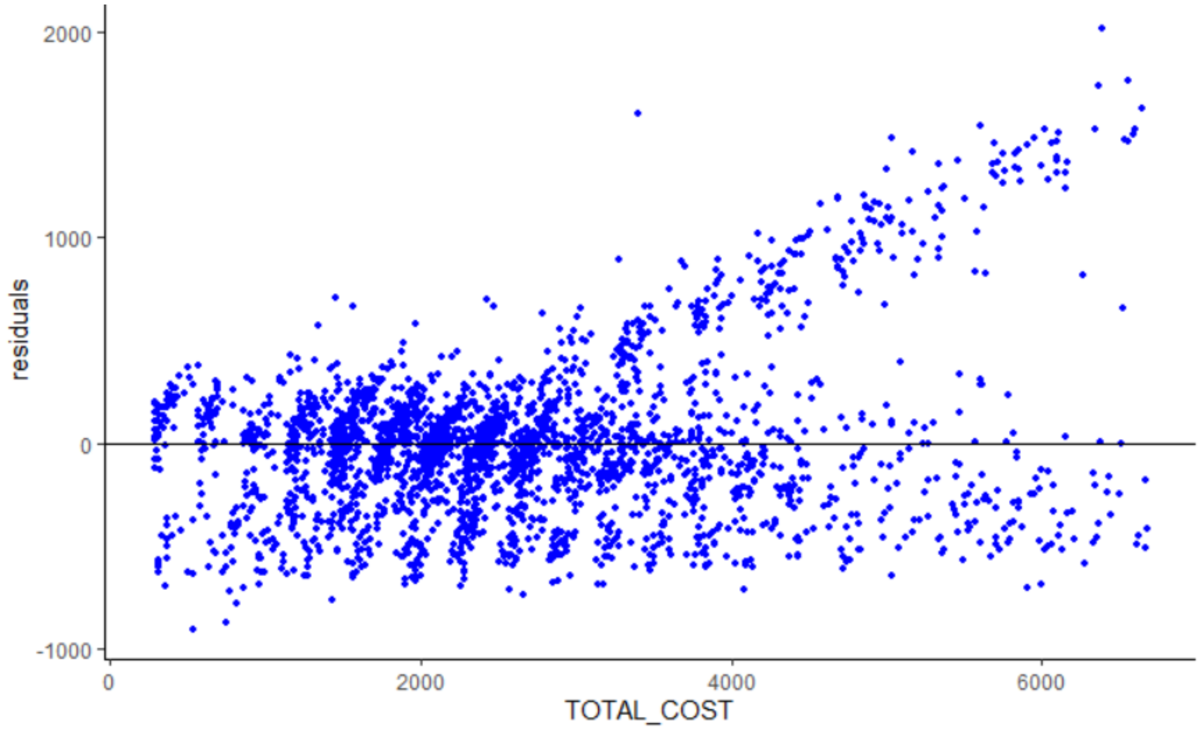
```
dfn <- df %>%
  filter(! IDADMISSION %in% c(16005375,16000036, 16018298 ))
```

- By following the directive deduced at the end of the 1st Iteration's part, the detected outliers have been removed;

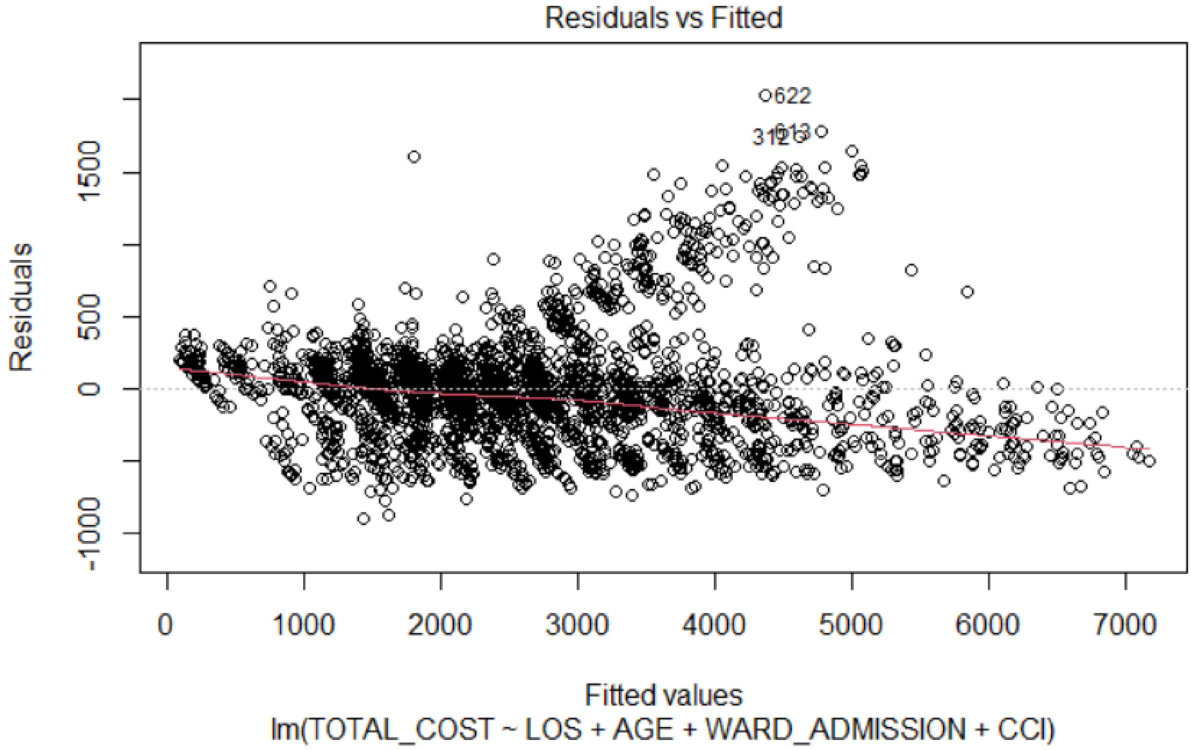
```
dfn %>%
  mutate(RISKDEATH = as.character(RISKDEATH),
         WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'G
eneralist', 'Specialist')
  ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION + CCI,
      data = .
  ) -> regn
summary(regn)

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
```

- The minimum value of residual has been updated to -894.87 € (we have a slight increase) whereas the new maximum value of residual is 2 024.10 € (a decrease has been observed). Nevertheless, there is still an enough significant difference between the two values at the ends (therefore, outliers are still existing);
- Although the new value of the median (-19.35 €) is superior than the old one, it still remains relatively far from 0, we still have an Asymmetry to the right when it comes to the residuals' distribution;
- The model's coefficients have also been updated, with in particular a clear improvement for the particular case of the coefficient associated with the CCI;
- The p-value associated with the F-Test has more or less remained the same as in the 1st Iteration, therefore, still less than 1 %: Our model is still significant.

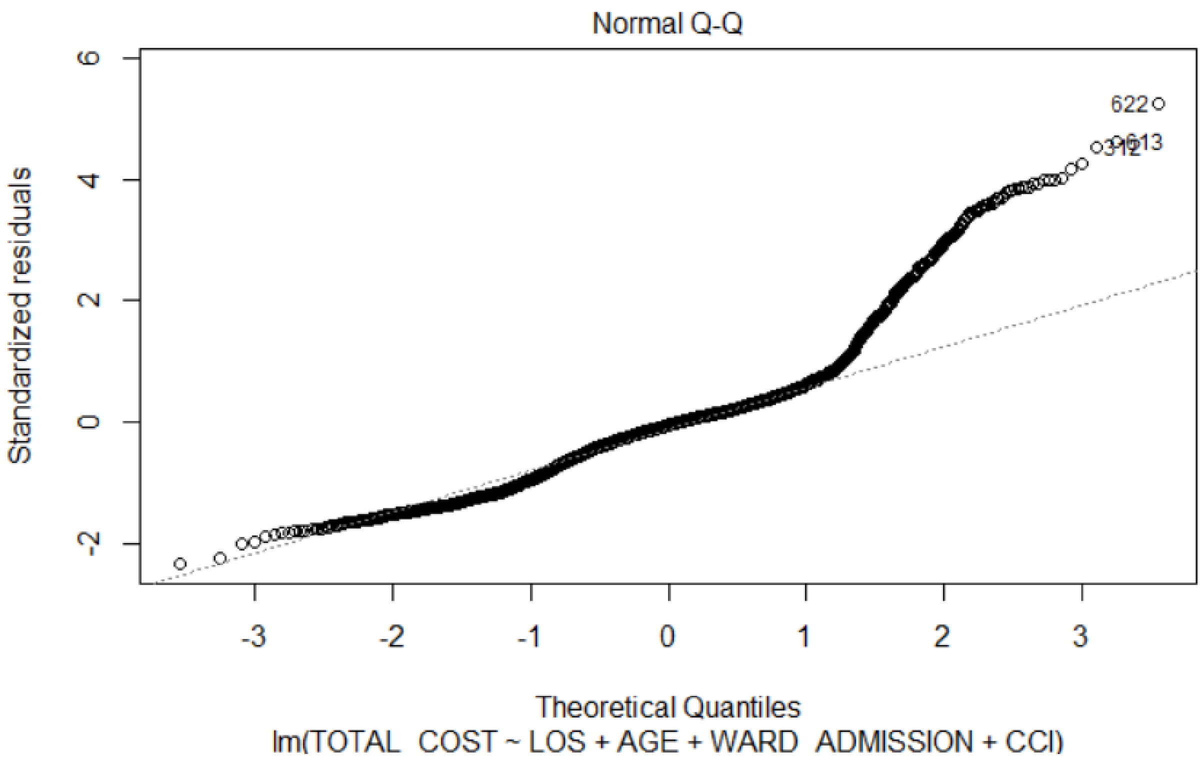
Residual analysis :



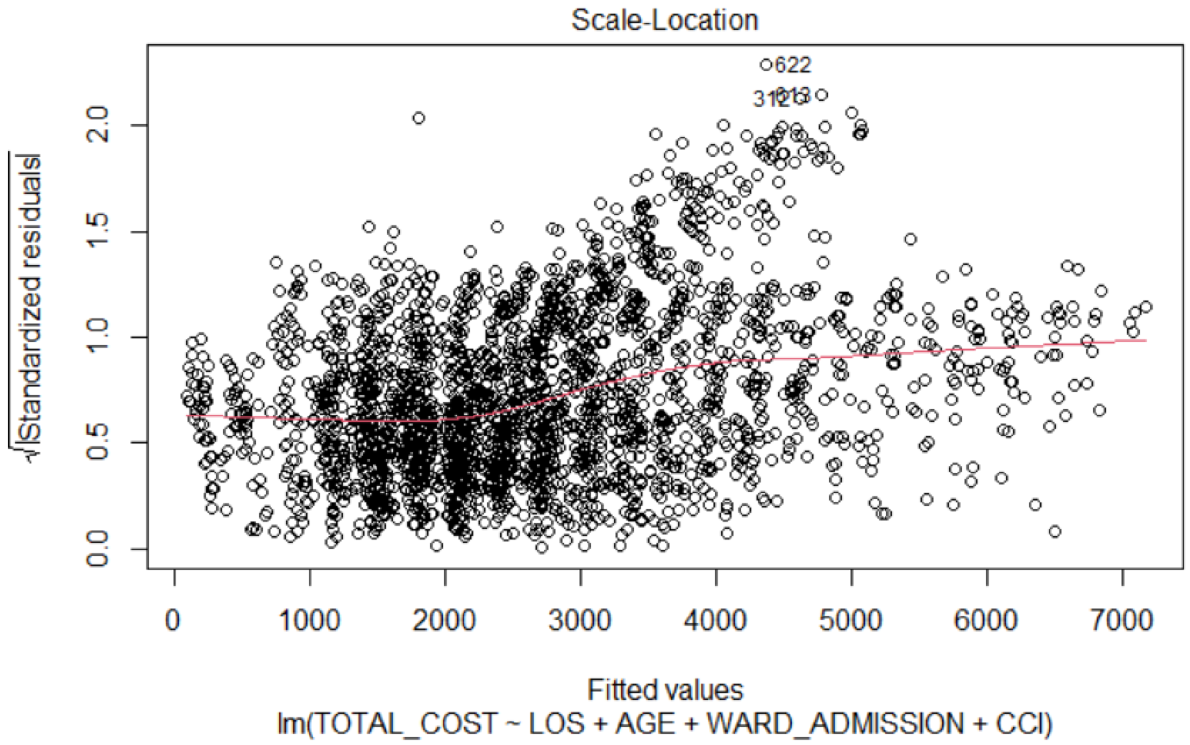
- We can observe that the relationship remains relatively Positive;



- The linearity still remains relatively not respected and we can observe the existence of new outliers associated with the residuals, more precisely between the values 1 500 € and 1 750 €;

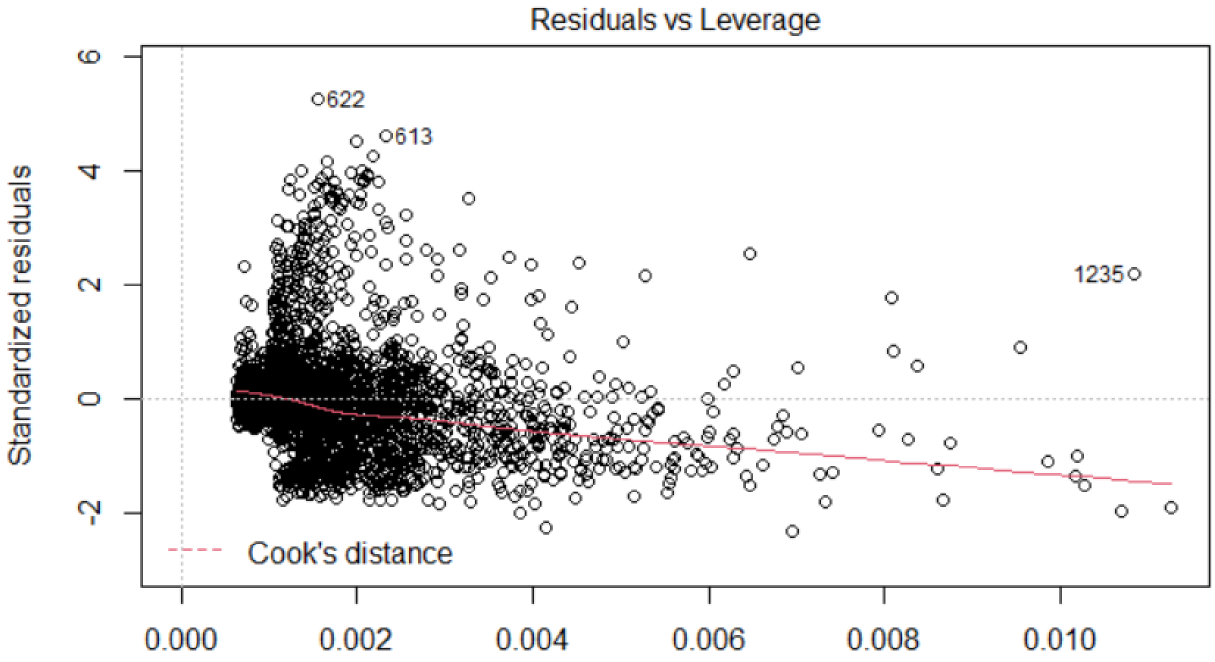


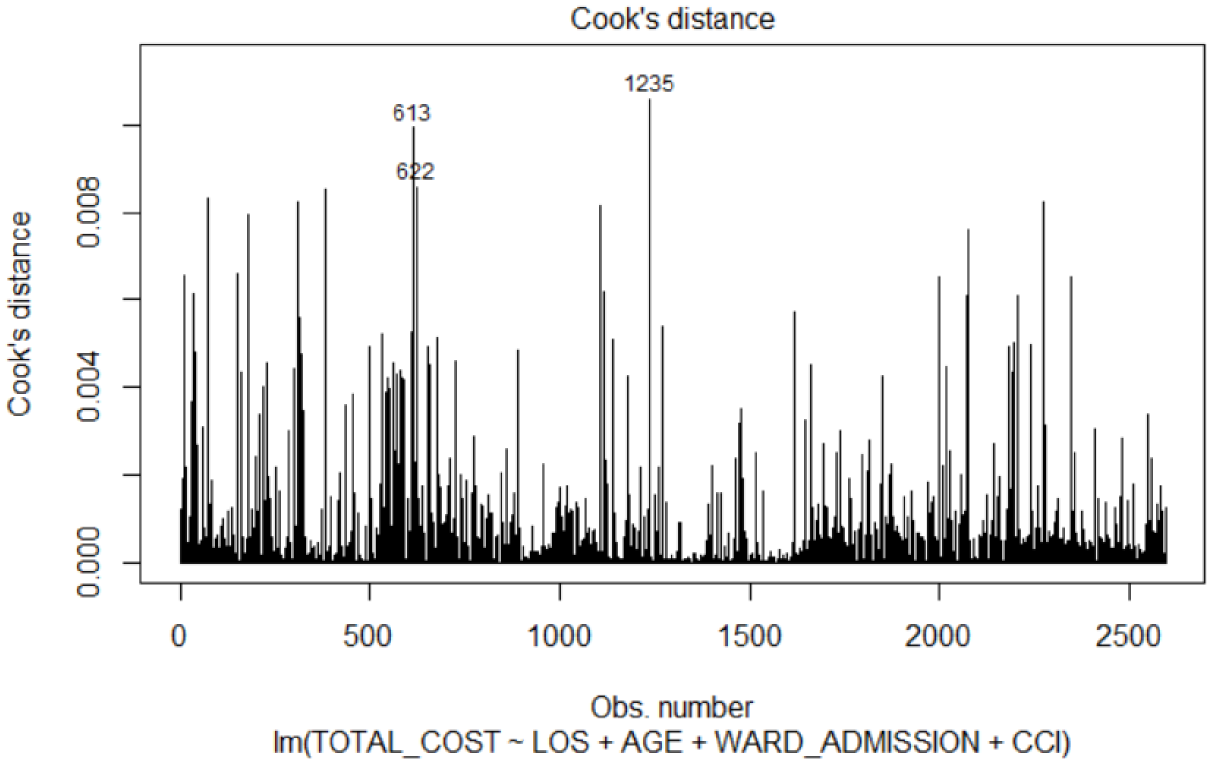
- The Quantile-Quantile plot above tells us more about the existence of new outliers at the ends;
- It’s always the same behavior that is observed: outliers seem to be more significant at the level of the upper end;



- From the Scale-Location plot above, we can clearly notice that the red line is not enough horizontally straight yet to satisfy the hypothesis for Homoscedasticity;

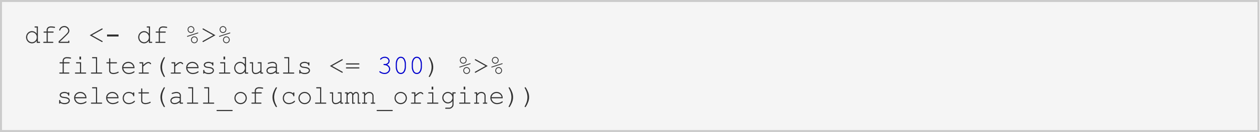
- The outliers can be observed at the upper levels of the standardized residuals’ square roots;





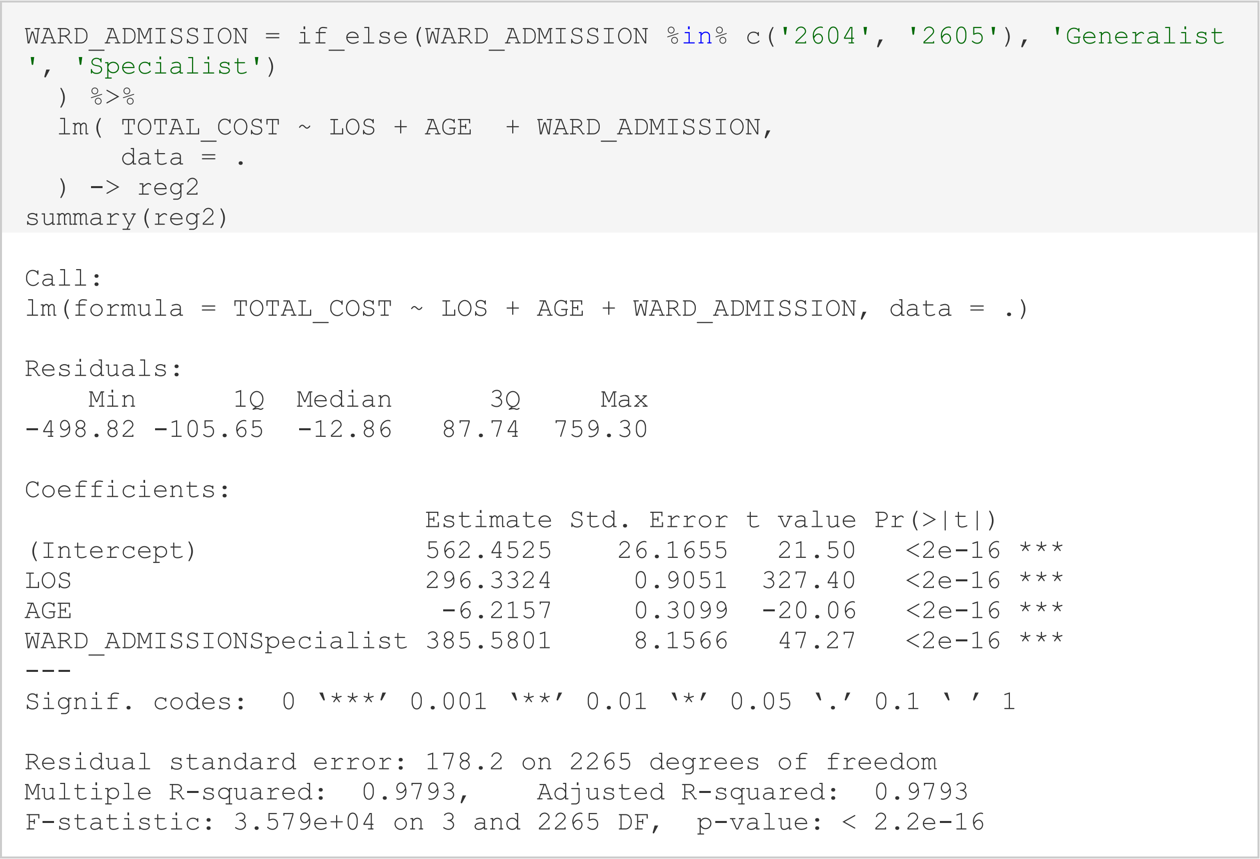
- Based on the two previous Plots of « Residuals vs. Leverage » and « Cook's distance », we can confirm with confidence that the residuals #1235, #613 and #622 (outliers) can be considered as being influential points;
- New actions of observations removals are then recommended before entering the next Iteration.

II.3 – 3rd Iteration :



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

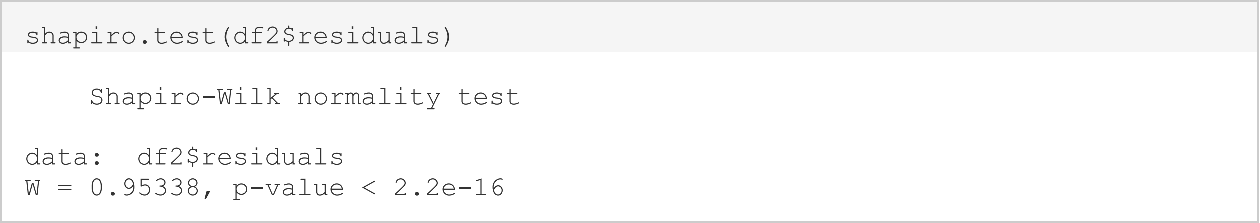
- Abstractly announced at the end of the 2nd Iteration, the removal of the observations which residuals' value is superior to 300 has been realized;



```
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
```

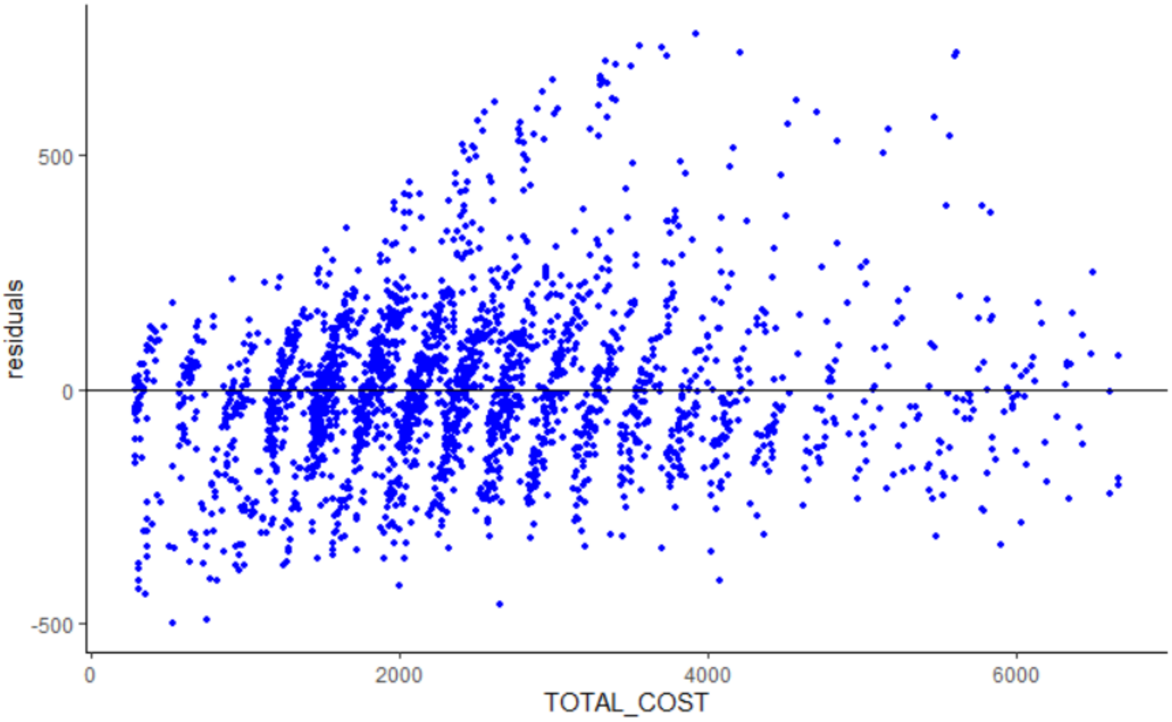
- The minimum value of the residuals has been updated to -498.82 € (a clear increase took place) whereas the new maximum one is 759.30 € (a clear decrease has been noticed). However, the difference between the two values at the ends still seems quite significant (possible existence of outliers);
- Although the new Median's value (-12.86 €) is higher than the previous one, it still remains relatively far from 0, we still have an Asymmetry to the right when it comes to the residuals' distribution;
- The model's coefficients remains about the same as in the 2nd Iteration (so, still significant), let us just notice the fact that the CCI has no longer been considered as a Predictor for the *prediction* of the values of TOTAL_COST within this 3rd Iteration;
- The p-value associated with the F-Test has remained about the same as in the 2nd Iteration, therefore still lower than 1%: Our model is still significant.

Residual analysis :

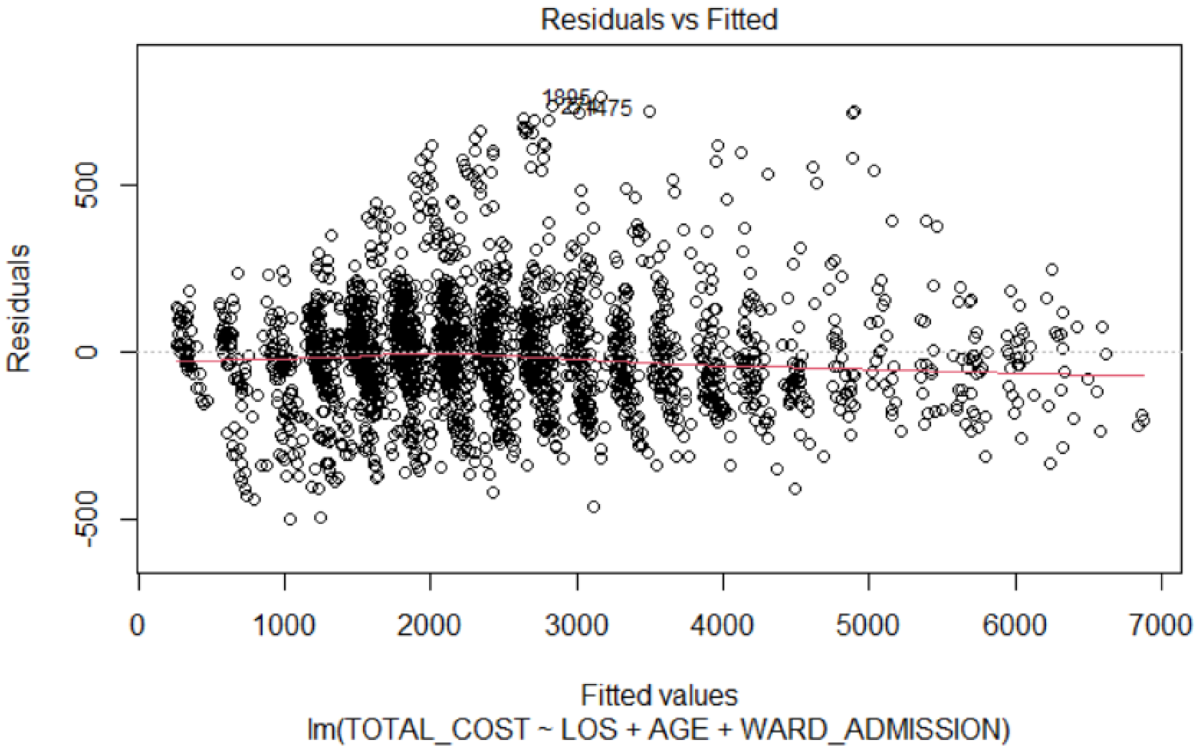


```
shapiro.test(df2$residuals)

    Shapiro-Wilk normality test

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

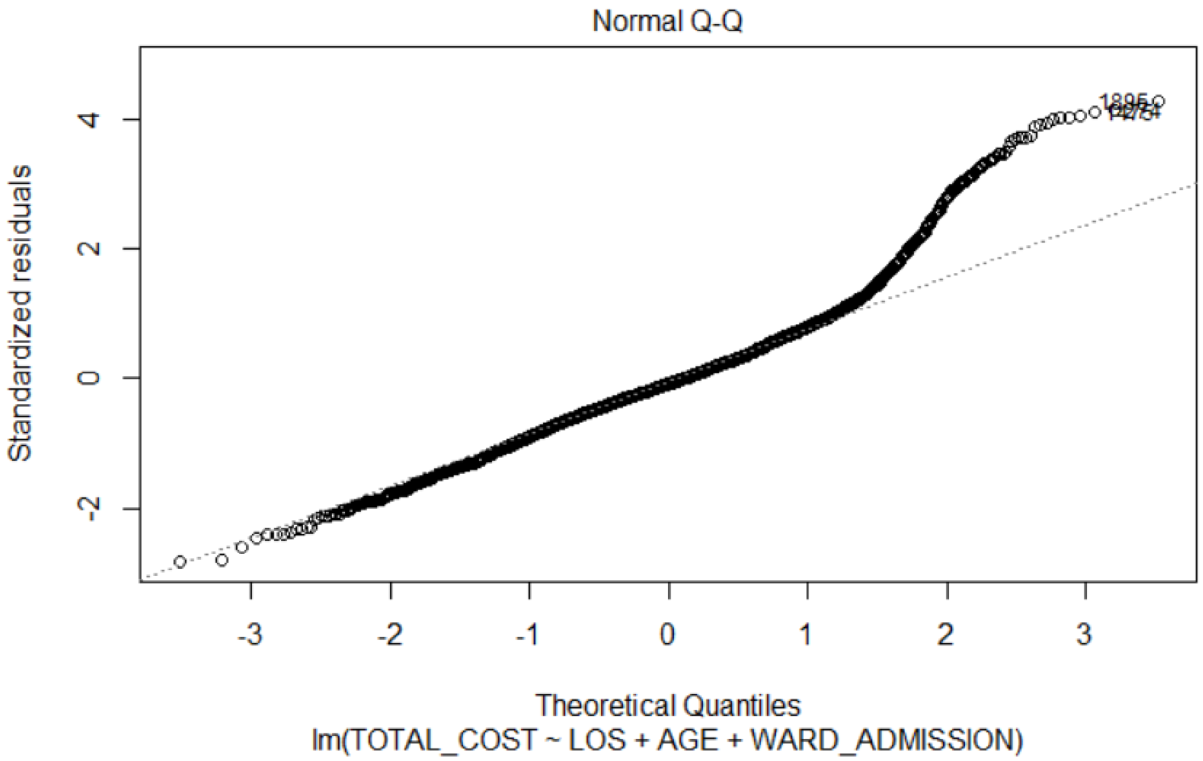
- The residuals (as a variable) doesn’t follow a normal distribution (asymmetrical distribution);



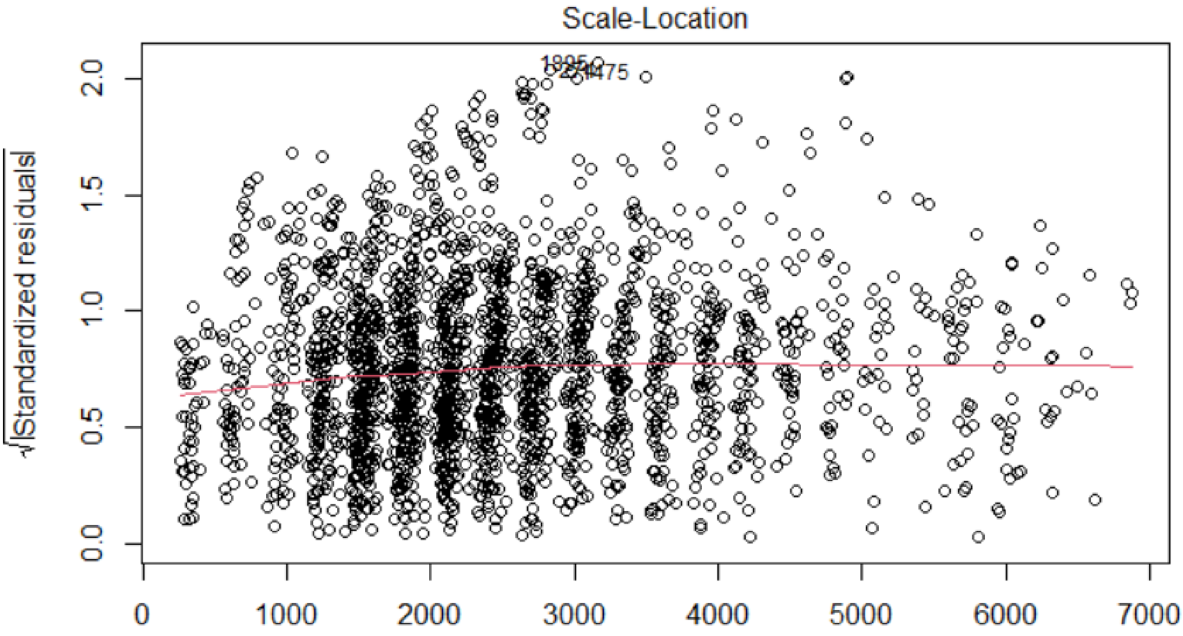
- We can notice that the relationship has become less Positive than it used to be before, compared to the versions which correspond respectively the previous Iterations;



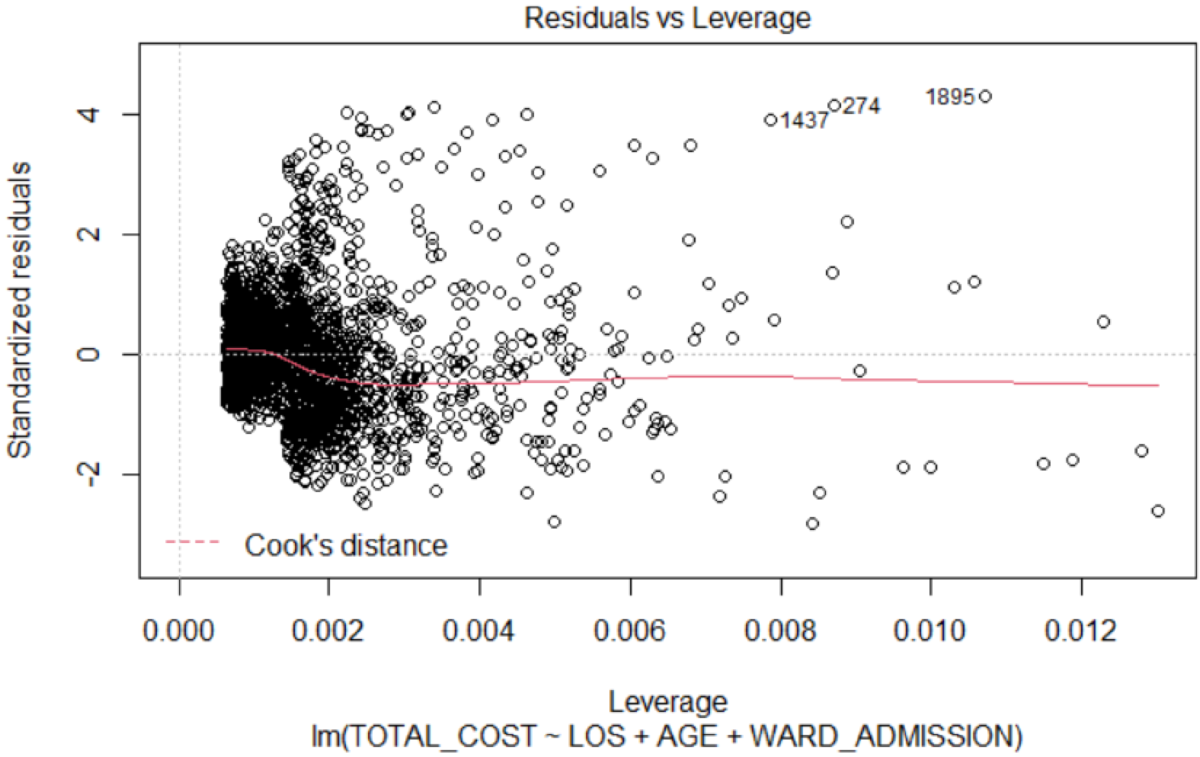
- The linearity is now less violated, but we can still notice the existence of new outliers when it comes to the residuals, more precisely near the value 750 €;



- The Quantile-Quantile plot above tells us more about the existence of these new outliers at the ends, in particular at the level of the upper one;



- From the Scale-Location plot above, the red line is now very near to become enough horizontally straight to satisfy the *hypothesis for Homoscedasticity* of our model;
- However, some outliers can still be observed at the square roots' upper levels of the standardized residuals;



- Based on the « Residuals vs. Leverage » Plot above, we can still confirm that the residuals #1427, #274 and #1895 (outliers) can be considered as influential points;
- Once again, actions of observations removal are then recommended before starting the next Iteration.

II.4 – 4th Iteration :

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

- Abstractly announced at the end of the 3rd Iteration, a new session of removal of observations which residuals value higher than 300 has once again been realized;

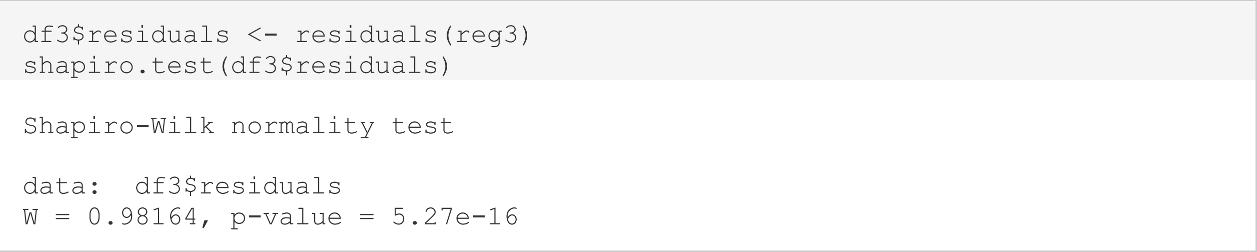
```
df3 %>%
  mutate(RISKDEATH = as.character(RISKDEATH),
         WARD_ADMISSION = if_else(WARD_ADMISSION %in% c('2604', '2605'), 'G
eneralist', 'Specialist')
  ) %>%
  lm( TOTAL_COST ~ LOS + AGE  + WARD_ADMISSION + CCI,
      data = .
  ) -> reg3
summary(reg3)

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
```

- The minimum value of the residuals has been updated to -331.12 € (a decrease has been noticed);
- The new median's value (-13.86 €) has slightly decreased compared to the previous one seen in the 3rd Iteration, therefore, logically, still remains relatively far from 0. We still have an Asymmetry to the right when it comes to the residuals' distribution;
- The model's coefficients remain about the same as in the 3rd Iteration (still significant in that case), let us just notice the fact that the CCI has once again been taken into account as a Predictor for the *prediction* of the TOTAL_COST' s values within this 4th Iteration and that the

coefficient which corresponds to it has improved significantly compared to the previous one observed in the 2nd Iteration;

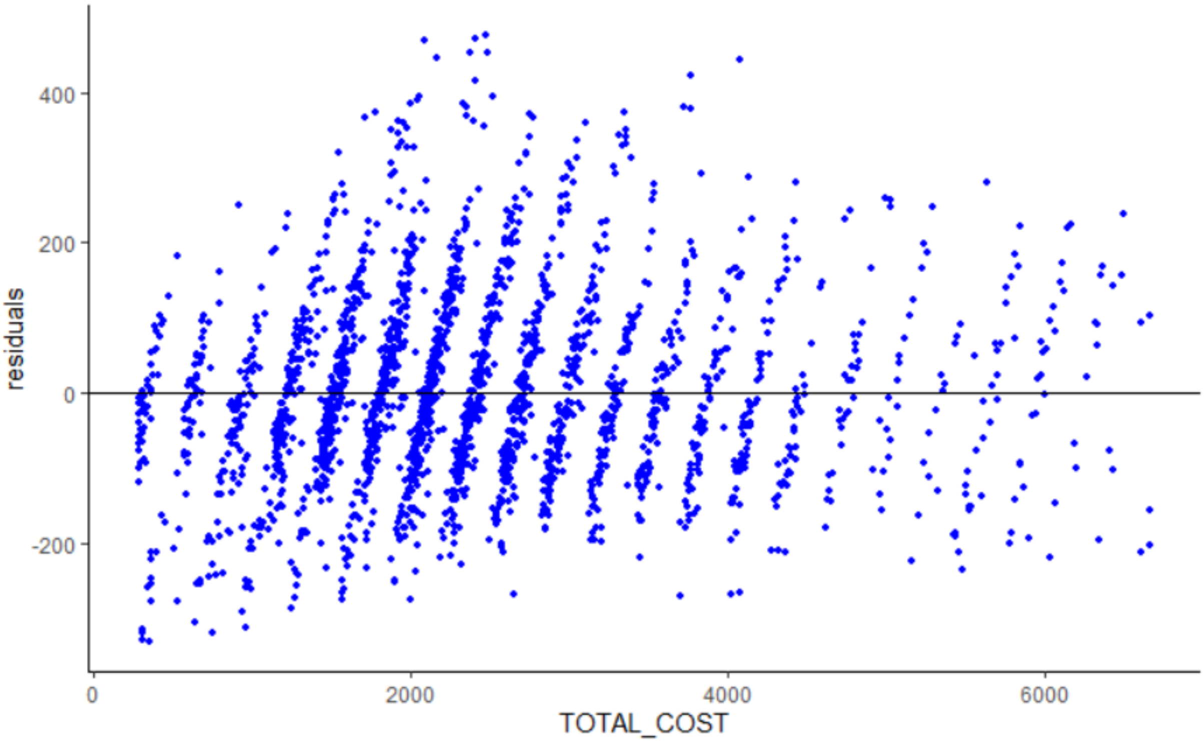
- The p-value associated with the F-Test has more or less remained the same as in the 3rd Iteration, therefore still lower than 1 %: Our Model still remains significant itself.

Residual analysis :

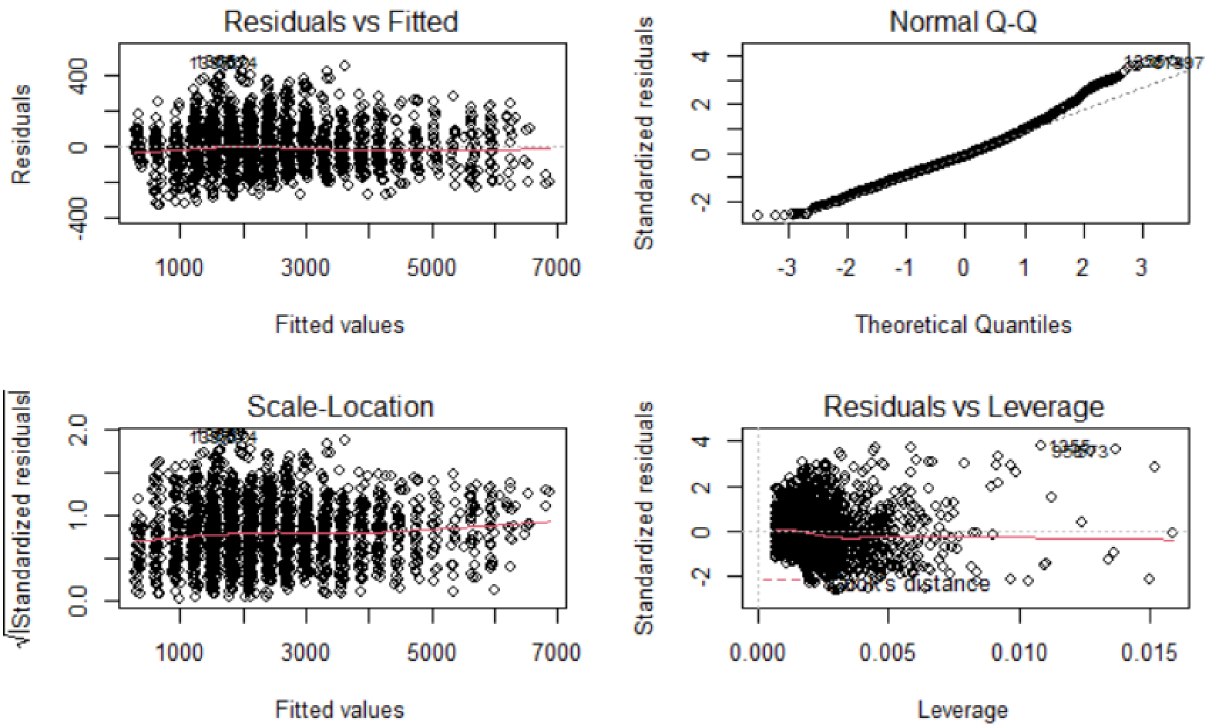


```
df3$residuals <- residuals(reg3)
shapiro.test(df3$residuals)

Shapiro-Wilk normality test

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

- The residuals (as a variable) doesn't follow a normal distribution (asymmetrical distribution);



- We can notice that the relationship is now only very slightly positive, unlike all the previous cases during all the previous iterations;



- Based on what the « Residuals vs Fitted » plot can show us, the linearity is now *more respected* and also, outliers, although some still persist, now seem to be *more acceptable*;
- The Quantile-Quantile plot also confirms this tendency of linearity now respected when it comes to our model, with the outliers, let us remind it, now more acceptable;
- From the Scale-Location plot, the red line is now enough horizontal to satisfy the hypothesis for Homoscedasticity of our model;
- Finally, based on the « Residuals vs. Leverage » plot, we can confirm with confidence that the residuals corresponding to the outliers which still persist no longer necessarily constitute influential points.