Modelisation Statistique.

BOUCHERRAB meziane ,DIALLO Abdoul

08/05/2022

Table Des Matières

1 Problématique.

une agence de revente de voitures veut mettre en place une stratégie qui lui permettra d'améliorer sa compétitivité et ce en ayant le meilleur rapport qualité-prix, pour cela le directeur de l'agence veut comprendre la variabilité du prix en fonction des caractéristiques des véhicules pour maximiser la rentabilité de son activité, en se basant sur les données des véhicules vendus lors des 3 dernières années, le jeu de données a été donné par l'agance .

2 Description du jeu de donné.

Le jeu de données comporte au total, 494 observations et 10 caractéristiques (variables); la colonne qui est la variable d'intérêt et **Prince Euro** c'est une variable quantitative continue c'est la variable à expliquer en fonction des autres variables qui représentent les caractéristiques des différents véhicules . L'ensemble de données n'inclut aucune valeur Manquante le jeu donné se compose principalement de données numériques.

2.1 Variables explicatives.

Les variables suivantes sont les variables qu'on utilisera pour expliquer notre variable d'intérêt **Price_euro** Qui est caractérisée de la sorte:

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    366.4   2958.0   5697.6   7904.3   9983.4 106957.8
```

variance:

On remarque que la variance **Var(price_euro)=83191802** est élevée donc il est suspectible que cette variable ait des valeurs aberrantes ou isolées.

- Variables quantitatives
continues

- Mileage : Nombre de kilomètres parcourue par le véhicule.
- Manufacture_year : Année de première mise en circulation du véhicule.
- Engine_displacement : Cylindré du véhicule en CM^3.
- Engin_Power : Puissance du véhicule en chevaux

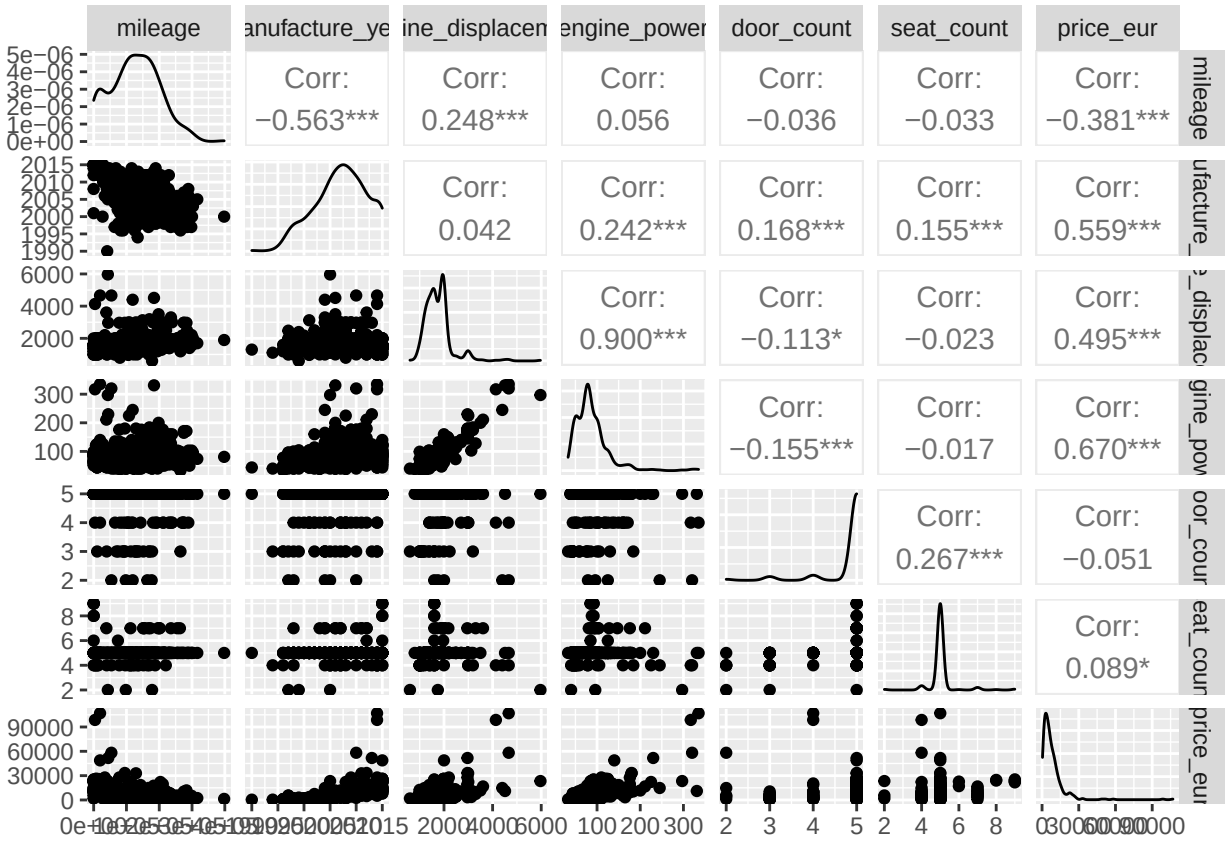
discrétes

- door_count : Nombre de porte du véhicule.
- seat_count : Nombre de siège du véhicule.

- Variables qualitatives

- Maker : la marque du véhicule
- Transmission : le type de transmission du véhicule "Manuelle-Automatique"
- fuel_type : Type du combustible compatible avec le moteur "Diesel-essence".

2 .2 Analyse descriptive des variables -Les données transmission et fuel_type sont des variables qualitatives on peut donc les enlever lors de notre première étude.



Analyse des données avec ggpairs.

-On remarque d'une part en regardant les nuages de points que les variables door_count et seat_count sont des variables quantitatives discrètes et d'autre part que leurs corelations avec la variable d'intérêt sont inférieurs à 0.1 donc elles sont moins corrélées avec la variable à expliquer.

-En regardant le nuage de points entre la variable **engine_power et engine_displacement** qui est presque aligné sur la première bissectrice donc elles paraient linéairement dependant et de même leur coefficient de corrélation qui est **0.9** nous le confirme.

-En regardant les nuages de points se **mileage, engin_power, manifacture_year et engine_dispolacement** présente une forme allongée donc elles sont corrélées avec la variable **price_euro** , leurs coeficients de corrélations respectives **0.38,0.67,0.559,0.495** avec **price_euro** sont importantes donc cela confirme qu'elles sont bien corrélées avec notre variable d'intérêt.

3. Régression linéaire multiple et son étape de validation
3.1 Régression linéaire multiple

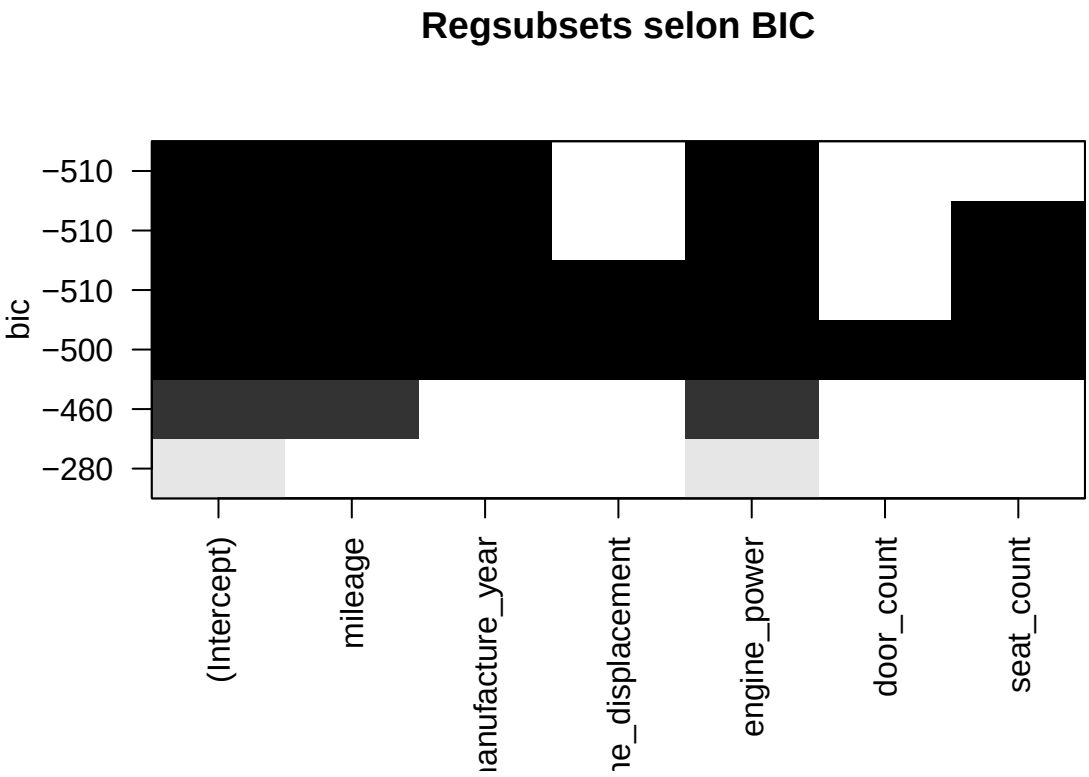
On va effectuer une régression linéaire multiple avec toutes nos variables explicatives. On rappelle qu'on cherche à expliquer le **price_eur** .

```
##
## Call:
## lm(formula = price_eur ~ ., data = Donne)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2063 -0.2279 -0.0329  0.1928  6.2363
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.090e-15  2.612e-02   0.000   1.0000
## mileage             -2.673e-01  3.413e-02  -7.830 3.07e-14 ***
## manufacture_year     2.415e-01  3.583e-02   6.742 4.45e-11 ***
## engine_displacement -5.672e-02  6.953e-02  -0.816   0.4150
## engine_power         6.753e-01  7.070e-02   9.552  < 2e-16 ***
## door_count          -1.858e-02  2.837e-02  -0.655   0.5128
## seat_count           5.790e-02  2.735e-02   2.117   0.0348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5805 on 487 degrees of freedom
## Multiple R-squared:  0.6671, Adjusted R-squared:  0.663
## F-statistic: 162.7 on 6 and 487 DF,  p-value: < 2.2e-16
```

Le test de Fisher global **F-statistique=162.7** est significatif (la p-value est quasi nulle), donc on en conclut qu'il y a au moins une variable qui a un coefficient non nul. On constate que le test de nullité des coefficients est significatif pour les variables:**Engine_power, manufacturing_year, mileage**, donc pour ces variables l'hypothèse nulle est rejetée et pour les variables **door_count, engine_displacement, seat_count** et la constante leurs **p_value > 0.01** donc on ne peut pas rejeter l'hypothèse nulle. **Le coefficient de détermination:** R^2=0.6671 donc ces variables semblent expliquer bien le modèle. poussons un peut notre études pour savoir le modéle le plus significatif à l'aide du critère BIC:

3.2 La sélection des variables

Le jeu de données contient de nombreuses variables explicatives et certaines sont fortement corrélées. Il est possible de faire un choix de modèle qui expliquerait le mieux notre variable. Pour se faire nous allons appliquer la fonction **regsubsets** selon le coefficient **BIC** Nous allons essayer de faire le **BIC** pour voir quel modèle présente un **BIC** minimale.



On retient juste le modèle qui minimise le **BIC** donc on peut considérer le modèle avec les variables explicatives à savoir:
engine_power, **manifacturing_year**, **mileage** et bien sûr aussi avec la constante **intercep**pour notre modèle.
Donc en ne conservant que ces variables, on obtient la régression suivante.

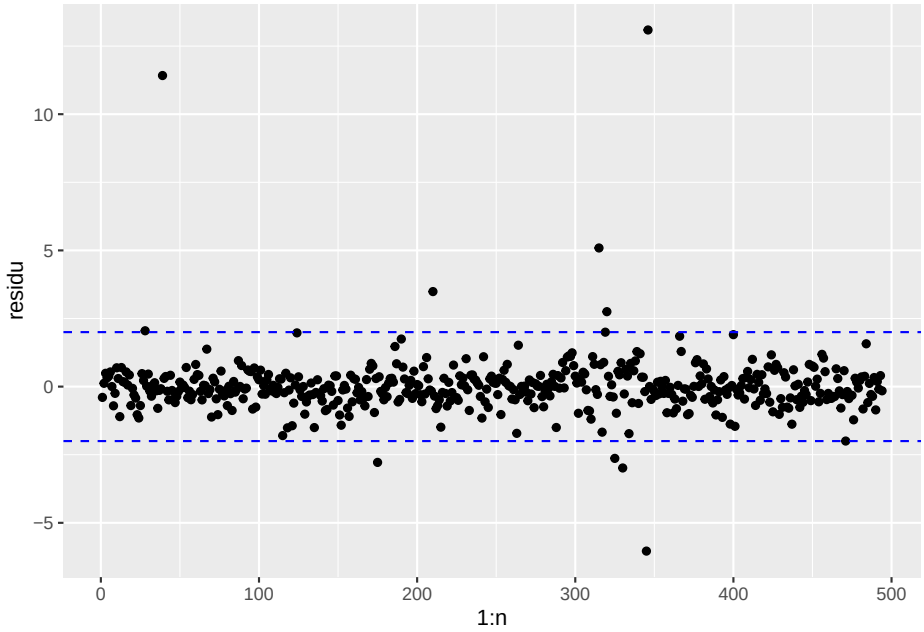
```
##
## Call:
## lm(formula = Donne$price_eur ~ Donne$mileage + Donne$manufacture_year +
##     Donne$engine_power)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1701 -0.2354 -0.0379  0.1890  6.2655
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.092e-15  2.618e-02   0.000        1
## Donne$mileage           -2.722e-01  3.265e-02  -8.337 7.77e-16 ***
## Donne$manufacture_year   2.548e-01  3.360e-02   7.584 1.70e-13 ***
## Donne$engine_power       6.232e-01  2.782e-02  22.404  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5818 on 490 degrees of freedom
## Multiple R-squared:  0.6635, Adjusted R-squared:  0.6615
## F-statistic: 322.1 on 3 and 490 DF,  p-value: < 2.2e-16
```

Le test de Fisher global **F-statistique=322.1** est significatif pour le modéle choisi (la p-value est quasi nulle), donc on en conclut qu'il y a au moins une variable qui a un coefficient non nul. Ici on remarque que le test de nullité est significatif pour toutes les variables.
Le coefficient de détermination **Multiple R-squared: 0.6635** est élevé et presque égale au **Multiple R-squared: 0.6671** du modèle avec toutes les variables.

3.3 Analyse des résidus

Cette étape concerne la vérification des hypothèses énoncées sur l'erreur estimée. Il s'agit du test de normalité des résidus. La validation du modèle est attachée à l'idée selon laquelle les résidus sont indépendants et identiquement distribués suivant la loi normale centrée avec une variance constante. Cette hypothèse sera vérifiée à travers une représentation graphique des erreurs et de la densité de la loi normale.

Résidus studentisées



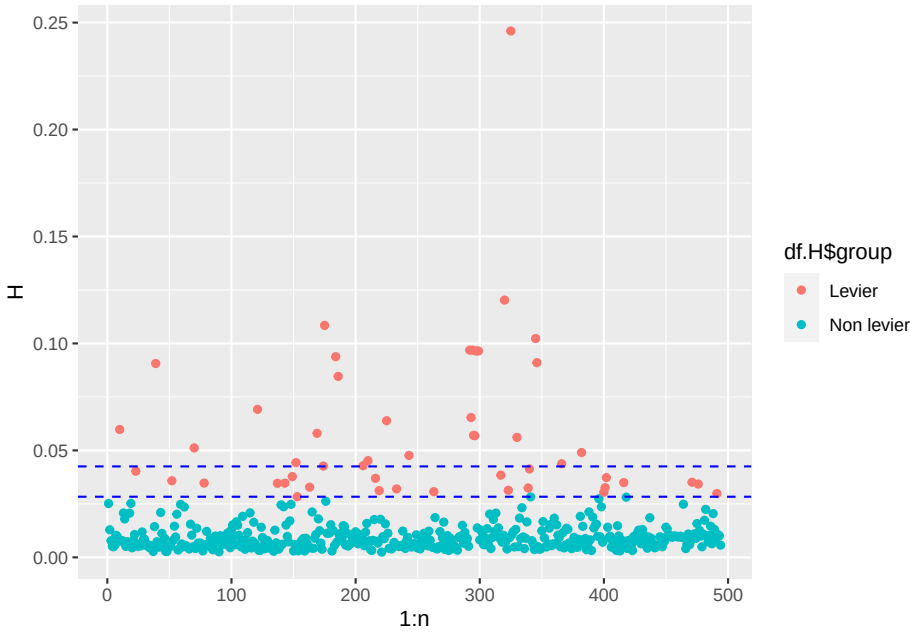
le pourcentage de points en dehors de l'intervalle est :

```
## [1] 2.024291
```

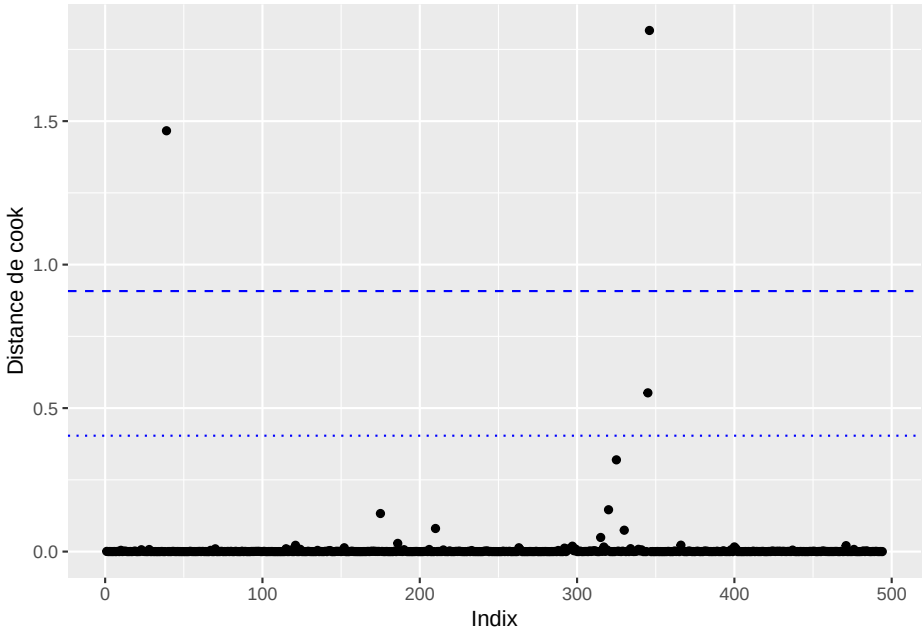
On constate qu'il y a peu de points qui sont en dehors de l'intervalle [-2,2], mais deux entre eux sont quand même loin des bornes. Mais quand même vu la taille de notre échantillon **n=494** donc on pourrait dire que c'est acceptable.

Poussons plus loin notre analyse en regardant les points leviers.

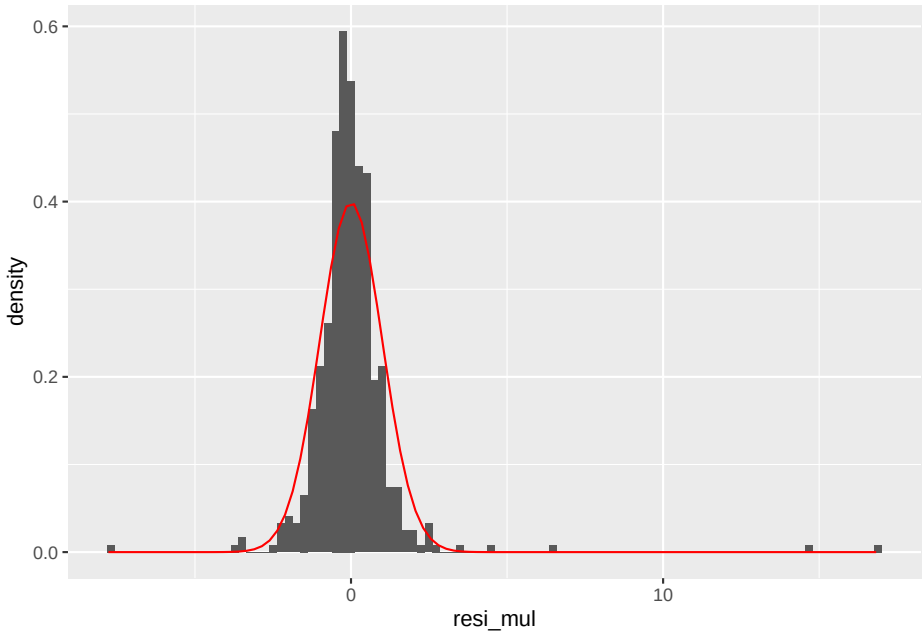
seuil 1=0.01619433
seuil 2=0.0242915



On observe qu'il y'a 22 valeurs aberrantes dont 11 points levier (leurs poids dépassent le deuxième seuil) pour nous assurer de l'influence des valeurs aberrante on regarde la distance de cook des résidus.



Nous constatons qu'il y a deux points qui ont une distance de cook loin du deuxième seuil nous pouvons donc conclure que ces deux points influent sur la justesse du modéle on peut donc enlever les point 39 et 346 qui présentent un poids élevé. vérifions à présent que les résidus suivent une loi normale centrée. Pour se faire nous allons représenter l'histogramme des résidus que nous superposerons avec une densité d'une loi normale centrée réduite.



En regardant l'histogramme on remarque une forte ressemblance avec un histogramme d'une loi symétrique. Pour avoir plus de clarté nous allons y superposer la densité de la loi normale. On remarque l'allure d'une courbe de gaussienne centrée réduite de l'histogramme, donc on peut se dire qu'il y a adéquation du fait qu'on a supposé que les erreurs pourraient suivre une loi normale centrée. Pour pousser plus le test de normalité faisons un QQ_plot pour voir plus clair.

Nous observons que l'alignement des points sur la première bissectrice est acceptable sauf au niveau des queues où on voit une inclinaison dont **05 points de la droite** et **04 points de la gauche**, ce qui confirme l'hypothèse selon laquelle les résidus théoriques ε_i suivent la loi normale $\mathcal{N}(0, \sigma^2)$

4. Analyse de la variance

Comme notre jeu de données ne contient pas que de variables quantitatives, il contient aussi des variables qualitatives qui sont maker, transmission et fuel_type. Nous allons nous intérésser à leurs effets sur notre variable d'intérêt.
Donc nous allons faire une **ANOVA** avec un modèle à trois facteurs .

```
##
## Call:
## lm(formula = qual$price_eur ~ qual$maker + qual$transmission +
##     qual$fuel_type)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24563  -3096  -1030   2225  82251
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              13999.35    1671.45   8.376 6.54e-16 ***
## qual$makeraudi            8255.47    2411.29   3.424 0.000672 ***
## qual$makerbmw             1366.48    2308.62   0.592 0.554203
## qual$makerchevrolet       2182.71    3732.88   0.585 0.559016
## qual$makerchrysler       -2939.67    5764.55  -0.510 0.610325
## qual$makercitroen        -1559.73    1916.04  -0.814 0.416042
## qual$makerdodge           4879.51    8111.44   0.602 0.547760
## qual$makerfiat           -1803.57    2338.65  -0.771 0.440978
## qual$makerhonda           3898.72    3199.27   1.219 0.223604
## qual$makerhyundai         2201.62    2070.12   1.064 0.288097
## qual$makerjeep           -4409.53    5742.14  -0.768 0.442923
## qual$makerkia             1848.45    2573.12   0.718 0.472890
## qual$makerlancia         -2731.80    8007.24  -0.341 0.733133
## qual$makermazda           -817.09    3191.61  -0.256 0.798056
## qual$makermercedes-benz 12932.73    2781.54   4.649 4.34e-06 ***
## qual$makermitsubishi     -1067.03    7996.13  -0.133 0.893900
## qual$makernissan           459.78    2736.74   0.168 0.866655
## qual$makeropel            6220.23    1669.73   3.725 0.000219 ***
## qual$makerporsche         -708.57    8111.44  -0.087 0.930428
## qual$makerrover          10150.21    4126.12   2.460 0.014257 *
## qual$makerseat             -22.19    2643.05  -0.008 0.993304
## qual$makerskoda            779.00    1238.85   0.629 0.529784
## qual$makersmart          -9927.67    8111.44  -1.224 0.221606
## qual$makersubaru          4660.38    4152.43   1.122 0.262303
## qual$makersuzuki          -212.19    3739.64  -0.057 0.954777
## qual$makertoyota          1552.26    2323.65   0.668 0.504447
## qual$makervolvo           2446.46    4726.99   0.518 0.605017
## qual$transmissionman     -7751.01    1254.93  -6.176 1.43e-09 ***
## qual$fuel_typegasoline   -2224.91     760.16  -2.927 0.003592 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7916 on 465 degrees of freedom
## Multiple R-squared:  0.2895, Adjusted R-squared:  0.2468
## F-statistic: 6.768 on 28 and 465 DF,  p-value: < 2.2e-16
```

On constate ici les effets globaux de chaque facteur. La variables marque **maker** est codé par **R** en 27 variables binaires et nous avons choisi comme modalité de référence **ford** donc on interprète l'effet des autres par rapport à celui de ford .
On voit par exemple que que la marque **makeraudi** en moyenne influence 8255.47 de plus que **ford** sur le prix et la p_value est quasi nulle donc on rejette l'hypothèse selon laquelle elle a même effet que la marque **ford** sur le prix.
La marque **jeep** influence en moyenne **-4409.53** moins que **ford** sur le prix mais on voit que le test de nullité est signicatif ce qui veut dire qu'il a même effet que **ford** sur le prix.
La variable **transmisson man** influence en moyenne **-7751.01** moins que la variable **transmission aut** sur le prix et aussi le test de nullité est quasi nul donc son effet est different de celui de la transmission automatique.
La variable **fuel_type gasoline** influence en moyenne **-2224.91** moins que le **fuel_typediesel** sur le prix et le test de nullité est inférieur à **0.01** donc leurs effefs sont quasiment égaux. Ici l'étude n'est pas trop claire car on aimerait bien voir les effets principaux des variables à savoir **maker, fuel_type et transmission sur le prix en euro**, donc pour se faire nous allons utiliser la fonction **anova** de R.

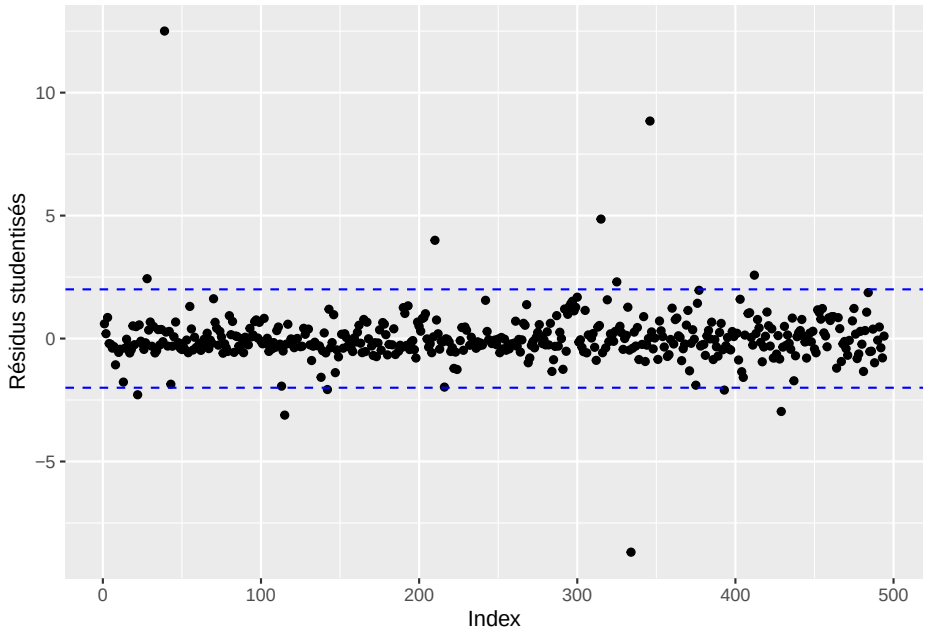
```
## Analysis of Variance Table
##
## Response: qual$price_eur
##                     Df     Sum Sq    Mean Sq F value    Pr(>F)
## qual$maker          26 8.7353e+09  335974297  5.3615 6.609e-15 ***
## qual$transmission    1 2.6027e+09 2602735717 41.5349 2.900e-10 ***
## qual$fuel_type       1 5.3682e+08  536816373  8.5666  0.003592 **
## Residuals          465 2.9139e+10   62663817
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nous constatons que la variable marque **maker** a une **p_valu =6.609e-15 qui est nulle** donc elle a un effet principal sur le prix.
La variable transmission a une **p_value=2.900e-10 qui est nulle** donc elle a un effet principal sur le prix.
La variable fuel_type a une **p_value=0.003592>0.001** donc on peut dire que cette variable n'a pas d'effet principal sur le prix. La marque et le mode de transmission ont des effets principaux sur le prix mais par contre le type de carburant n'a pas un effet principal sur le prix. Mais cela ne veut pas dire qu'il ne peut avoir d'effet d'interaction. Nous allons à présent voir les effets interactions.

```
## Analysis of Variance Table
##
## Response: qual$price_eur
##                                              Df     Sum Sq    Mean Sq F value
## qual$maker                                   26 8.7353e+09  335974297  6.4383
## qual$transmission                             1 2.6027e+09 2602735717 49.8764
## qual$fuel_type                                1 5.3682e+08  536816373 10.2871
## qual$maker:qual$transmission                 15 1.0133e+09   67551206  1.2945
## qual$maker:qual$fuel_type                    16 4.2137e+09  263359131  5.0468
## qual$transmission:qual$fuel_type              1 4.1283e+07   41283209  0.7911
## qual$maker:qual$transmission:qual$fuel_type   5 1.5358e+09  307150548  5.8859
## Residuals                                   428 2.2335e+10   52183702
##                                                Pr(>F)
## qual$maker                                   < 2.2e-16 ***
## qual$transmission                           6.665e-12 ***
## qual$fuel_type                                0.00144 **
## qual$maker:qual$transmission                  0.20149
## qual$maker:qual$fuel_type                   1.225e-09 ***
## qual$transmission:qual$fuel_type              0.37426
## qual$maker:qual$transmission:qual$fuel_type 2.827e-05 ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

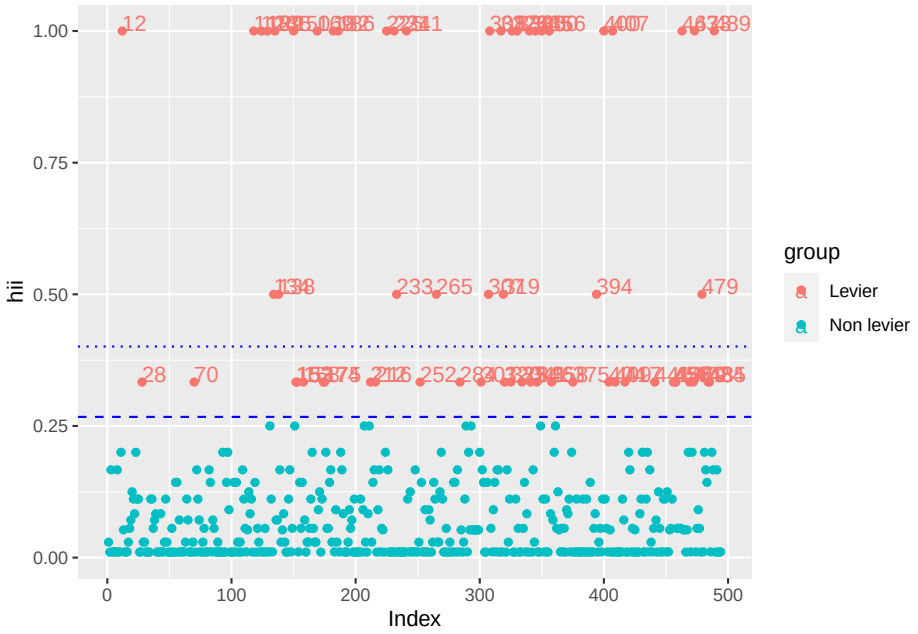
On constate que l'effet d'interaction entre la marque **maker** et le type de carburant **fuel_type** a une *p_value=1.225e-09 « 0.001* donc il y a bel et bien effet d'interaction. L'effet interaction entre **fuel_type, maker et transmission** a une **p_value=2.827e-05 « 0.001** donc il y a bien un effet d'interaction entre les trois variables donc on considère le modèle avec les trois variables.

Analyse des résidus




```
## [1] 2.631579
```

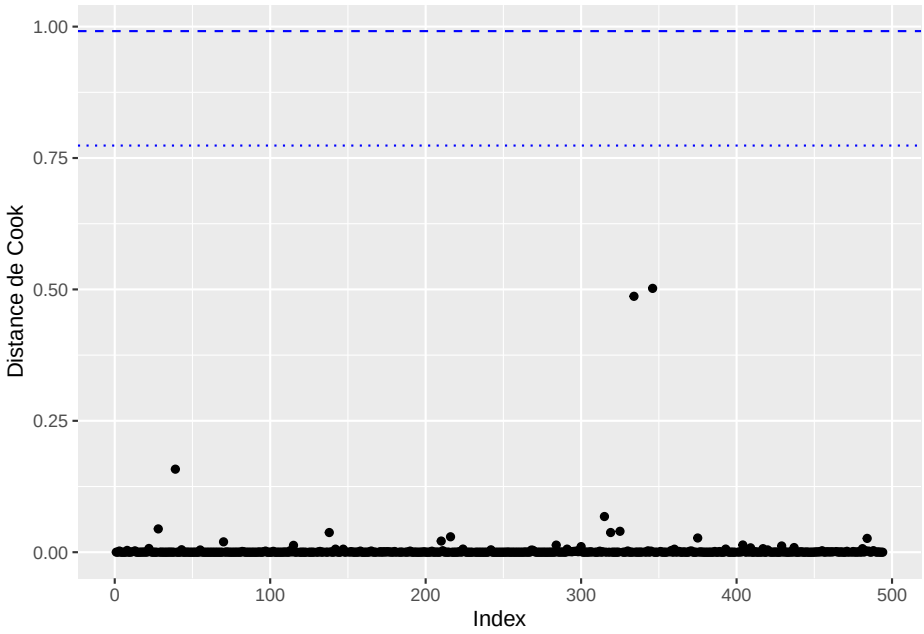
On remarque que **2.631579%** des points qui sont en dehors de l'intervalle $[-2, 2]$ ce qui inférieur à 5% donc acceptables.Nous allons regarder les points leviers.




```
## [1] 132
```

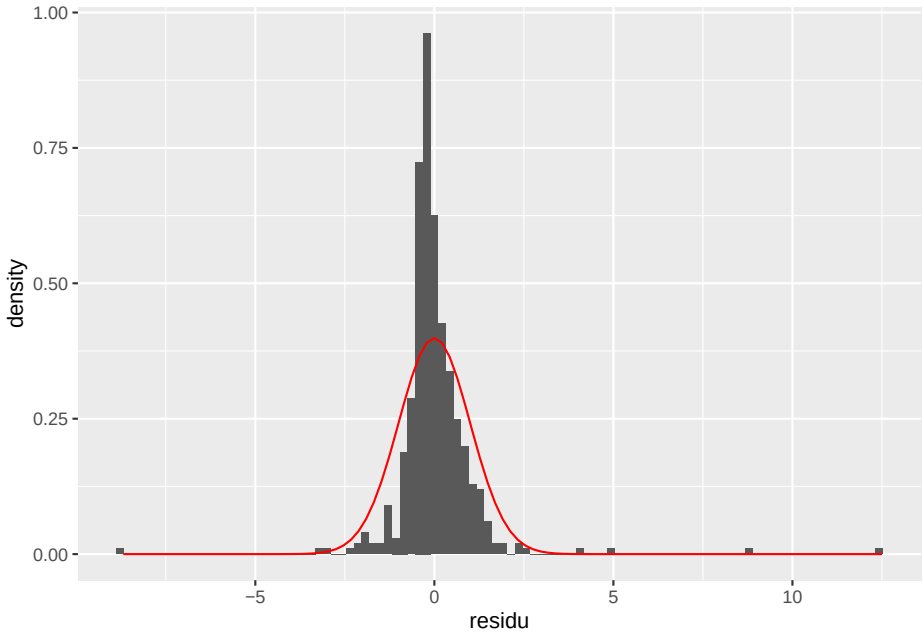
```
## [1] 106
```

On observe qu'il y'a 57 valeurs aberrantes dont 32 points levier (leurs poids dépassent le deuxième seuil) pour nous assurer de l'influence des valeurs aberrante on regarde la distance de cook des résidus.

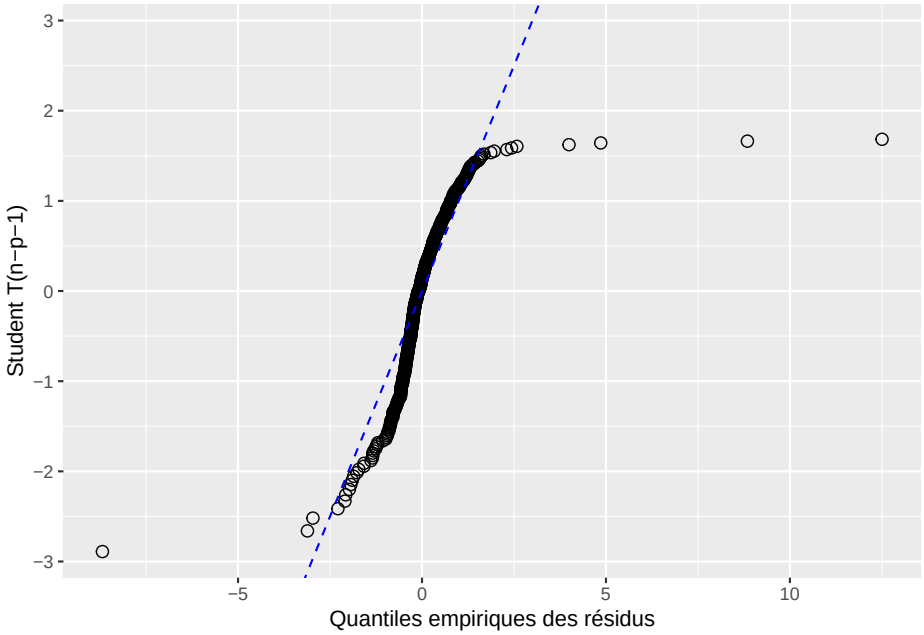


On constate qu'aucun point ne dépasse le premier seuil donc il n'y a pas de valeurs aberrantes, nous pouvons garder toutes les variables à savoir **maker, fuel_type,transmission**.
vérifions à présent que les résidus suivent une loi normale centrée. Pour se faire nous allons représenter l'histogramme des résidus que nous superposerons avec une densité d'une loi normale centrée réduite.

```
## [1] 471
```



On voit que l'histogramme des résidus a la forme d'une dansité gaussienne centrée. Pour voir plus clair faisons un qqplot.



On remarque que l'alignement des points par rapport à la première bissectrice est acceptable sauf au niveaux des queues où l'on remarque une forte inclinaison. Nous pouvons conclure que que notre hypothèse selon laquelle les erreurs théoriques suivent une lois normale centrée avec une variance constante ($\epsilon_i \sim \mathcal{N}(0, \sigma^2)$) Donc on peut conclure qu'on peut garder toutes les variables qualitatives dans notre modèle.