Projet modèle linéaire : Étude du prix de vente des maisons dans la ville de Bothell, Washington, États-Unis

Oumnia Tazi, Mossad Abdelsalam & Ethan Miran

2020-12-20

1- Introduction et description des données

Plus que jamais, le marché de l'immobilier est en plein essor, ce qui pousse les sociétés immobilières à faire appel à des techniques de machine learning pour estimer et prévoir le prix de vente des biens immobiliers. L'objectif de notre projet consiste à identifier le meilleur modèle pour expliquer et prédire le prix de vente des maisons individuelles dans la ville de Bothell, dans l'état américain de Washington. Pour ce faire nous allons :

- Expliquer, affiner et décrire nos données grâce à des statistiques descriptives.
- Réaliser un modèle de régression linéaire simple et multiple.
- Réaliser une prédiction à partir du modèle final obtenu.

Dans ce cadre le jeu de données initial considéré est composé de 21 variables qui sont les suivantes :

- `id` : le numéro d'identification de la maison
- `date` : la date à laquelle la maison a été vendue
- `price` : le prix de vente de la maison (en dollars)
- `bedrooms` : le nombre de chambres dans la maison
- `bathrooms` : le nombre de salles de bains dans la maison
- `sqft_living` : la superficie habitable de la maison (en pieds carrés)
- `sqft_lot` : la superficie du terrain (en pieds carrés)
- `floors` : le nombre d'étages dans la maison
- `waterfront` : variable binaire indiquant si la maison est au bord de l'eau (1) si oui, (0) si non
- `view` : le nombre de fois que la maison a été visitée
- `condition` : l'état général de la maison noté de 1 à 5. (1) si mauvais , (5) si excellent
- `grade` : une note donnée à la maison selon une notation du comté de King, entre 1 et 13. (1) si très mauvaise , (5) si excellente
- `sqft_above` : la superficie habitable de la maison sans compter le sous-sol (en pieds carrés)
- `sqft_basement` : la superficie du sous-sol (en pieds carrés)
- `yr_built` : l'année de construction de la maison
- `yr_renovated` : l'année de rénovation de la structure de la maison
- `zipcode` : le code postal afférant à la maison
- `lat` : la latitude de la maison
- `long` : la longitude de la maison
- `sqft_living15`: la superficie habitable moyenne des 15 maisons les plus proches (en pieds carrés)
- `sqft_lot15` : la superficie moyenne des terrains des 15 maisons les plus proches (en pieds carrés)

La complexité du jeu de données initial composé de 21597 occurrences nous a conduit à mener des actions de nettoyage :

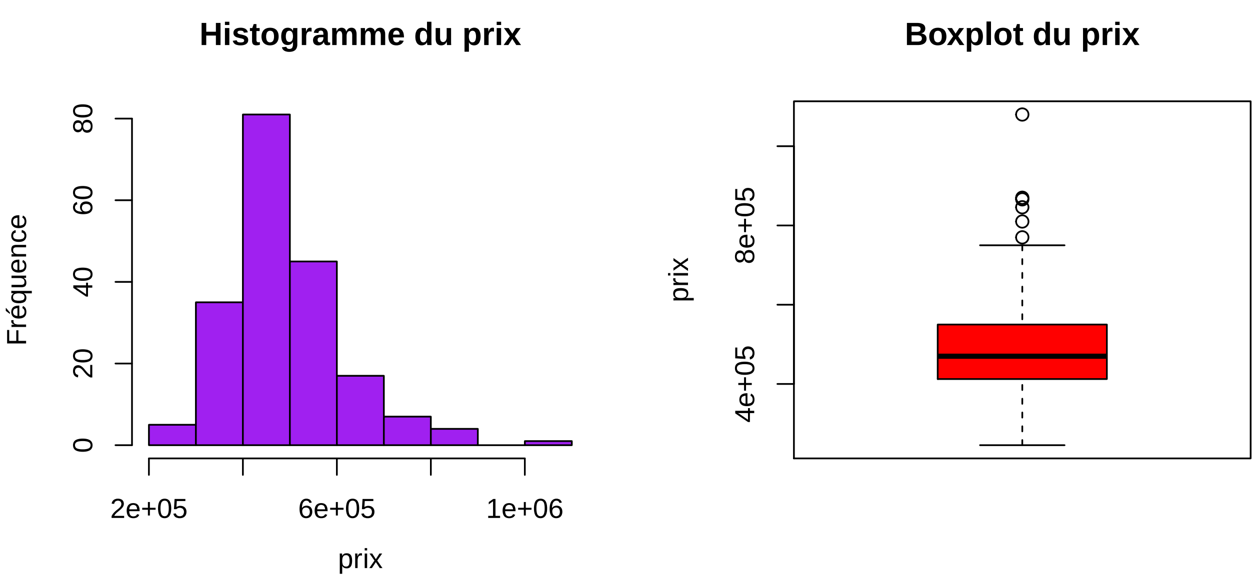
- Restreindre l'application des modèles à un périmètre constitué des maisons dont le code postal est le 98011. En effet, ce choix s'explique par la disparité des valeurs initiales en fonction de la zone géographique (ex : le prix dans Seattle sera plus élevée que dans une région rurale du comté). Cela nous a amené à un jeu comprenant 195 observations.

- Substituer la variable `yr_built` par une nouvelle variable nommée `age` qui correspond à l'âge de la maison. Cette dernière est plus facilement interprétable pour R.

- Retirer les variables, d'une part les variables qualitatives ou catégorielles : `id`, `date`, `waterfront`, `condition`, `grade` ; et d'autre part celles qui peuvent entrainer une incompréhension par le langage R : `lat`, `long`, `yr_renovated`. Pour illustrer, la variable `yr_renovated` indique l'année de rénovation lorsqu'elle existe et 0 sinon ce qui implique une fausse interprétation par R (0 est considéré comme l'an 0).

Nous allons jeter un coup d'oeil aux caractéristiques des 12 variables que nous allons utiliser au cours de notre étude et tracer un histogramme et un boxplot de la variable `price` :

```
##      price            bedrooms       bathrooms      sqft_living
##  Min.   : 245500   Min.   :1.000   Min.   :1.000   Min.   : 790
##  1st Qu.: 412400   1st Qu.:3.000   1st Qu.:1.875   1st Qu.:1700
##  Median : 470000   Median :3.000   Median :2.500   Median :2200
##  Mean   : 490377   Mean   :3.549   Mean   :2.278   Mean   :2253
##  3rd Qu.: 550000   3rd Qu.:4.000   3rd Qu.:2.500   3rd Qu.:2660
##  Max.   :1080000   Max.   :6.000   Max.   :3.500   Max.   :4890
##     sqft_lot          floors           view          sqft_above
##  Min.   :  2801   Min.   :1.000   Min.   :0.00000   Min.   : 790
##  1st Qu.:  7282   1st Qu.:1.000   1st Qu.:0.00000   1st Qu.:1380
##  Median :  8947   Median :1.500   Median :0.00000   Median :1850
##  Mean   : 11314   Mean   :1.503   Mean   :0.06154   Mean   :1955
##  3rd Qu.: 10658   3rd Qu.:2.000   3rd Qu.:0.00000   3rd Qu.:2420
##  Max.   :209959   Max.   :2.000   Max.   :4.00000   Max.   :4140
##  sqft_basement    sqft_living15   sqft_lot15         age
##  Min.   :   0.0   Min.   :1080   Min.   : 2723   Min.   :  1.00
##  1st Qu.:   0.0   1st Qu.:1890   1st Qu.: 7582   1st Qu.: 21.50
##  Median :   0.0   Median :2160   Median : 8970   Median : 32.00
##  Mean   : 298.5   Mean   :2248   Mean   : 9512   Mean   : 32.85
##  3rd Qu.: 590.0   3rd Qu.:2570   3rd Qu.:10188   3rd Qu.: 43.00
##  Max.   :1810.0   Max.   :4590   Max.   :56628   Max.   :102.00
```



Les graphiques nous indiquent que la majorité des prix des maisons se situe entre 250K dollars et 750K dollars. Néanmoins, une valeur beaucoup plus élevée que les autres (à plus d'1M de dollars) ressort nettement via le boxplot.

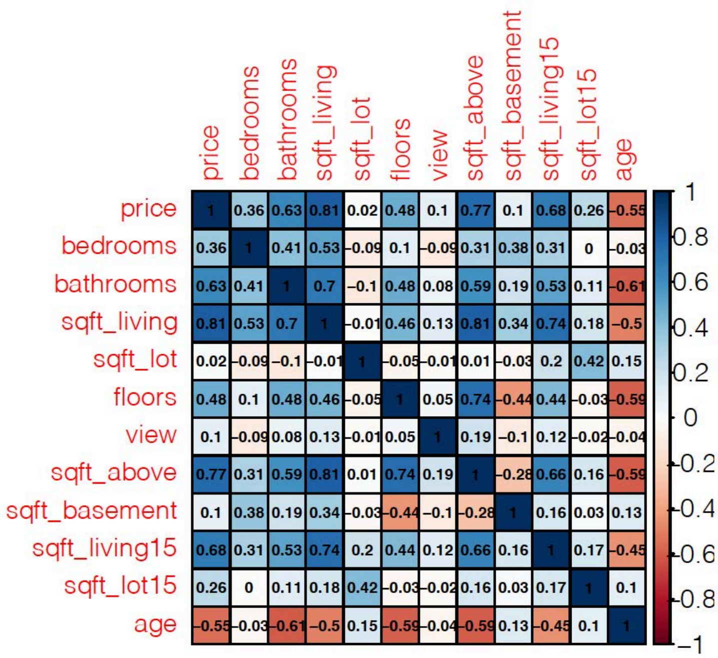
2- Régression linéaire simple

Nous allons tout d'abord réaliser une régression linéaire simple.

On rappelle la forme du modèle linéaire : $y = \beta_1 + \beta_2 x + \varepsilon_i$
y : la variable à expliquer
x : la variable explicative
β_1 et β_2 : des paramètres inconnus
ε_i : erreur que l'on suppose suivre une loi normale centrée de variance σ^2

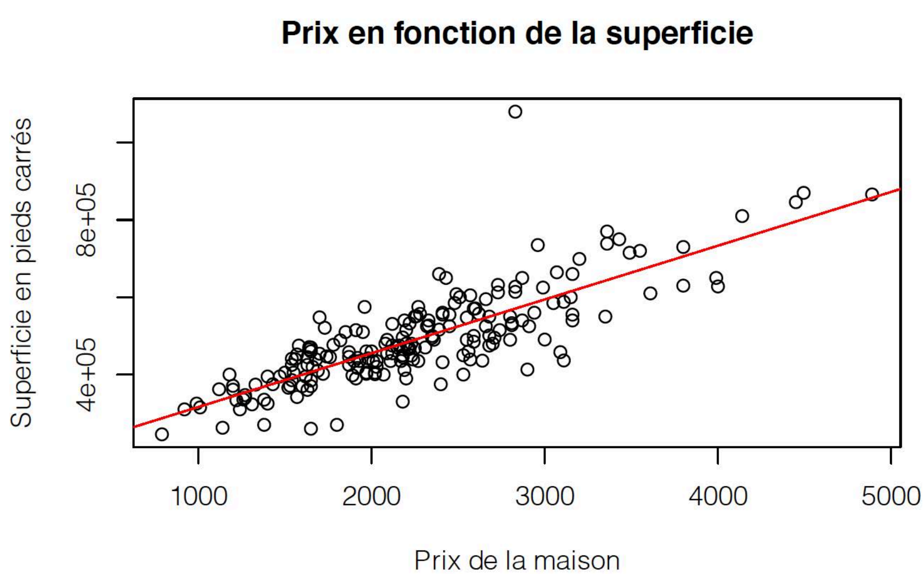
Pour commencer, nous avons constitué la matrice de corrélation entre le prix de vente de la maison et les autres variables quantitatives afin d'identifier les valeurs les plus corrélées pour appliquer notre modèle de régression linéaire simple.

- **Matrice de corrélation et matrice de nuages de points:**



Nous remarquons que la corrélation la plus élevée est celle entre les variables `price` (prix de la maison) et `sqft_living` (superficie habitable). Ceci nous amène à considérer l'application de la régression linéaire simple du prix de vente de la maison en fonction de la superficie habitable (en pieds carrés).

- **Estimation des coefficients :**

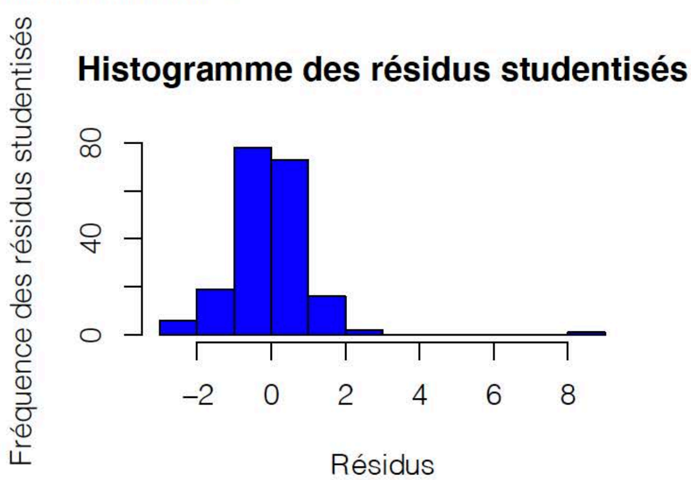


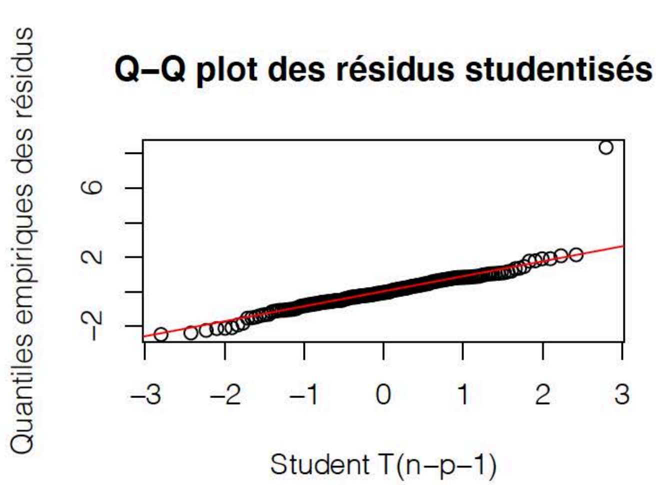
```
##
## Call:
## lm(formula = price ~ sqft_living, data = house_data_kcBT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -173036  -39658   -2993   43455  509400
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.771e+05  1.697e+04   10.43   <2e-16 ***
## sqft_living 1.391e+02  7.183e+00   19.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 71300 on 193 degrees of freedom
## Multiple R-squared:  0.6601, Adjusted R-squared:  0.6583
## F-statistic: 374.8 on 1 and 193 DF,  p-value: < 2.2e-16
```

Cette analyse nous permet de déduire l'estimation des coefficients :
- $\beta_1 = 177100$ ce qui indique que le prix d'une maison avec une superficie habitable de 0 pieds carrés (terrain sans maison) est de 177100 dollars.
- $\beta_2 = 139.1$ ce qui montre que le prix d'une maison augmente de 139.1 dollars lors de l'ajout d'un pied carré à la superficie de la maison.
Les tests nous ont permis également de rejeter H_0 au profit de H_1 D'une part, Le test de significativité montre que la p-valeur de `sqft_living` est très petite ($<2.2.10^{-16}$). Aux différents niveaux (5%, 1%) on rejette $H_0 : \beta_2 = 0$ au profit de $H_1 : \beta_2 \neq 0$. Ceci signifie que la valeur β_2 est significative ce qui nous permet de déduire une relation entre le prix de la maison et la superficie.
D'autre part, La p-valeur de l'intercept est également très petite donc nous rejetons $H_0 : \beta_2 = 0$ au profit de $H_1 : \beta_2 \neq 0$. De plus, la valeur du R2 est correcte : 0.6601, la définition d'un meilleur modèle est donc possible.
Enfin, les graphiques réalisés nous permettent de confirmer la présence d'une valeur éloignée de la droite de régression linéaire.

- **Analyse des résidus studentisés et recherche de valeurs abérrantes :**

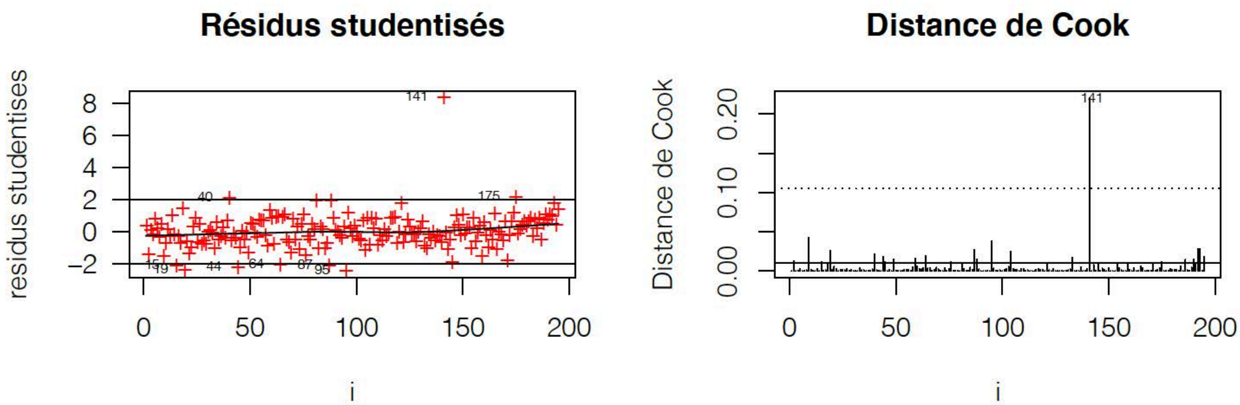
Pour compléter le modèle linéaire, nous procédons à une analyse des résidus studentisés et à la recherche des valeurs aberrantes.





En observant l'histogramme et le Q-Q plot, nous remarquons que la quasi-entièreté des résidus semble suivre une loi normale centrée avec toujours la présence d'une valeur à l'écart.

La réalisation du test de Shapiro-Wilk nous permet de calculer La p-valeur qui s'avère être très petite : 3.566Ö10^{-13}. Ceci réfute l'hypothèse H_0 qui stipule que les résidus suivent une loi normale. Cette conclusion peut s'expliquer par la sensibilité du test de Shapiro-Wilk aux valeurs extrêmes encore confirmées par le Q-Q plot (une valeur est très éloignée de la droite gaussienne).

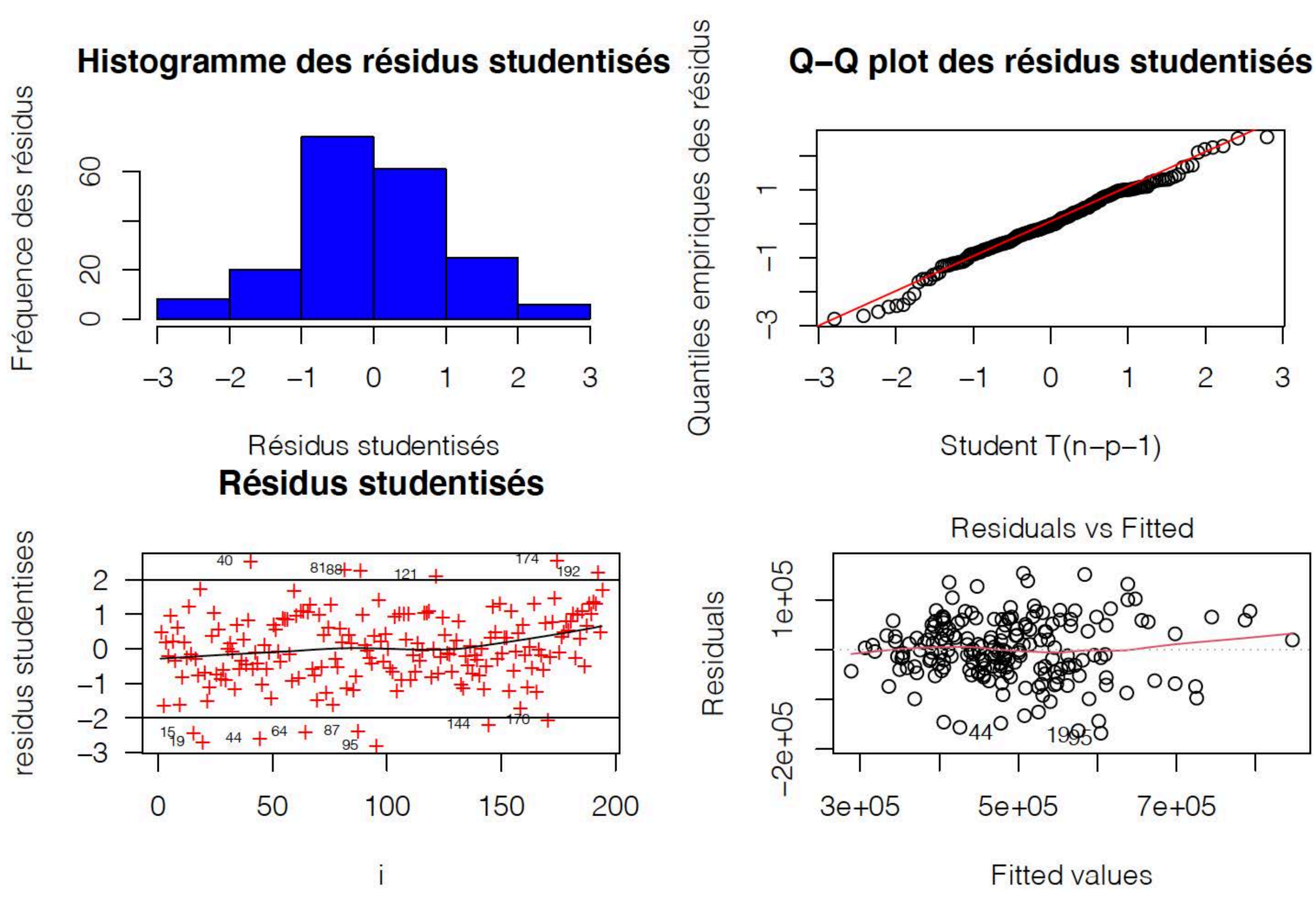



```
## [1] 0.1054181
```

Le graphique des résidus studentisés montre que les points sont inclus dans la bande]-2,2[, hormis quelques points qui restent néanmoins proches de cet intervalle et montre l'éloignement de l'observation 141. Le graphique de la distance de Cook, qui est une mesure qui combine le fait d'être une valeur aberrante et un point levier, confirme ce constat : l'observation 141 est largement supérieure au seuil fixé $f_p^{n-p}(0.1) = 0.106$. Cette observation a une trop grande influence sur les résultats d'estimation de notre régression.
Nous allons donc réaliser un nouveau modèle sans la valeur 141 qui est une valeur aberrante.

- **Application du modèle avec suppression de la valeur aberrante :**

Après suppression de l'observation 141 et en comparaison à notre premier modèle, l'application de la régression linéaire simple fournit des coefficients β_1, β_2 quasiment identiques, les tests de significativité portent les mêmes conclusions mais nous avons une valeur de R^2 considérablement plus élevée (0.7164).



En analysant l'histogramme et le QQ-Plot des résidus, les résidus studentisés semblent maintenant bien suivre une loi normale centrée en 0. De plus, les résidus studentisés sont centrés en 0 et le lisseur n'a pas de tendance, on en déduit donc l'indépendance des résidus. Nous observons également que nos résidus studentisés en fonction des valeurs prédites forment bien un nuage de points quelconques centrés en 0 et le lisseur ne forme pas de tendance non plus, ce qui en atteste de leur homoscédasticité.

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(rls2)
## W = 0.98916, p-value = 0.1482
```

Ces conclusions sont également confirmées par le test de Shapiro-Wilk avec une p-valeur de 0.148 qui valide l'hypothèse H_0 qui stipule que les résidus suivent une loi normale (au risque de 5% et même au risque de 14%). Ceci permet de conclure que les résidus théoriques ε_i suivent une loi normale $\mathcal{N}(0, \sigma^2)$.
Notre modèle de regréssion linéaire simple : `price`$= 181200 + 136.1\times$ `sqft_living` $+ \varepsilon$ est donc validé.

3- Régression linéaire multiple

Pour réaliser une régression linéaire multiple, nous allons effectuer la régression avec toutes nos valeurs quantitatives et toutes nos observations et analyser la corrélation entre les variables en calculant le VIF.

Par ailleurs, les variables `sqft_living`, `sqft_above` et `sqft_basement` ne peuvent pas être considérées simultanément dans le modèle car elles sont naturellement corrélées (`sqft_living` = `sqft_above` + `sqft_basement`).

```
##      bedrooms     bathrooms   sqft_living      sqft_lot        floors
##      1.828770      2.715820      5.174547      1.414452      3.124586
##          view sqft_basement sqft_living15    sqft_lot15           age
##      1.142032      2.640308      2.617221      1.462471      2.458311
```

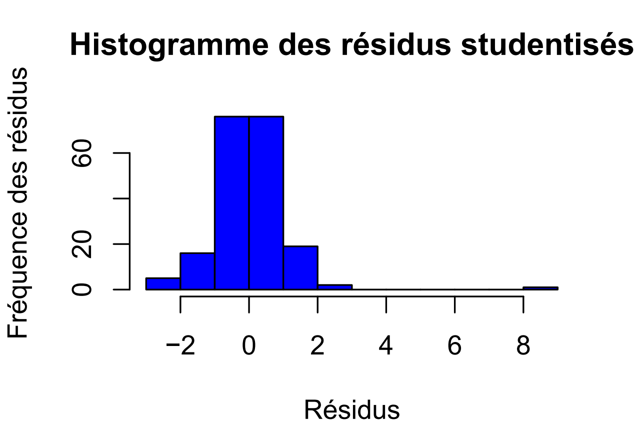
Le VIF de `sqft_living` est de 5.17, valeur supérieure au seuil fixé au préalable (égal à 5), nous sommes contraints de retirer cette variable du modèle de régression linéaire multiple. L'observation des p-valeurs permet également d'identifier les valeurs insignifiantes dans le modèle. Ceci nous conduit à les supprimer selon l'ordre décroissant de façon itérative afin d'avoir un modèle uniquement avec des variables significatives. Le modèle optimal identifié est le suivant :

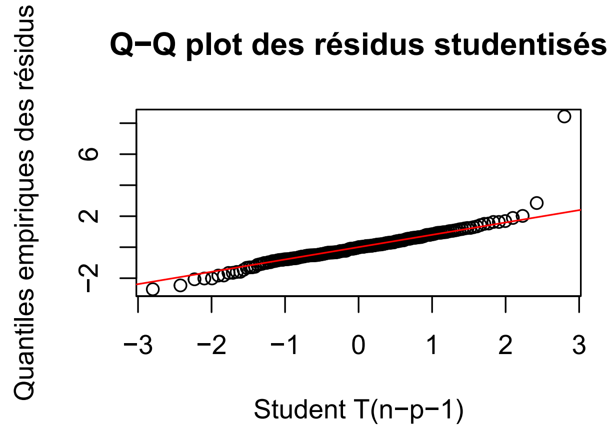
```
##
## Call:
## lm(formula = price ~ sqft_above + sqft_basement + sqft_lot15 +
##     age, data = house_data_kcBT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -171027  -33791      59   34670  458671
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.169e+05  2.507e+04   8.653 2.10e-15 ***
## sqft_above     1.277e+02  8.879e+00  14.383  < 2e-16 ***
## sqft_basement  8.915e+01  1.109e+01   8.041 9.31e-14 ***
## sqft_lot15     3.570e+00  9.616e-01   3.713 0.000269 ***
## age           -1.118e+03  3.117e+02  -3.586 0.000427 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64400 on 190 degrees of freedom
## Multiple R-squared:  0.727,  Adjusted R-squared:  0.7212
## F-statistic: 126.5 on 4 and 190 DF,  p-value: < 2.2e-16
```

Le test de Fisher global permet le rejet de l'hypothèse H_0 qui indique que les variables `sqft_above`, `sqft_basement`, `sqft_lot15`, `age` n'ont pas d'effet sur le prix `price`. Ceci est confirmé par les tests de significativité qui prouvent qu'aucun coefficient n'est significativement égal à 0 au seuil de 5%. La régression multiple nous permet d'améliorer la valeur du R^2 qui vaut 0.7212 dans ce modèle.

- **Analyse des résidus studentisés :**

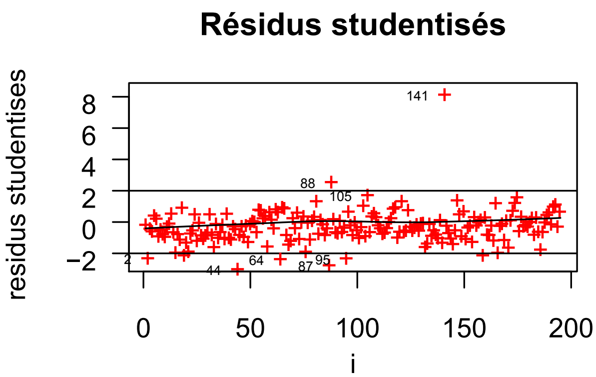
Nous allons désormais procéder à une analyse des résidus en traçant l'histogramme et le Q-Q plot.

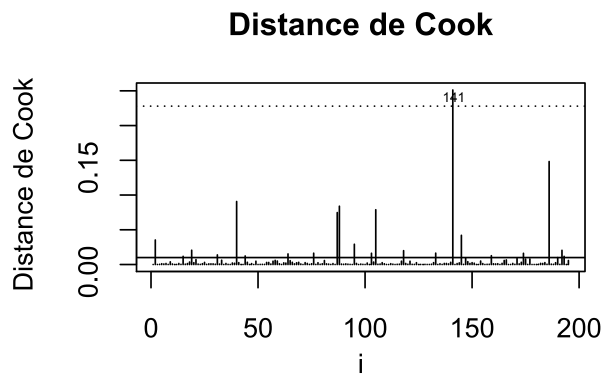




Les deux graphiques montrent la même problématique que le modèle de régression simple : l'observation 141 semble à l'écart des autres valeurs. Le test de Shapiro-Wilk indique une p-valeur de 2.801Ö10^{-13}. Nous rejetons l'hypothèse H_0 qui énonce que les résidus suivent une loi normale.

- **Recherche de valeurs aberrantes et points leviers :**





Les graphiques des résidus studentisés et de la distance de Cook montrent respectivement que la valeur 141 est à l'écart de la bande]-2,2[, et qu'elle est au-dessus du seuil $f_p^{n-p}(0.1) = 0.228$. Pour avancer sur l'analyse, nous allons retirer cette valeur du modèle.

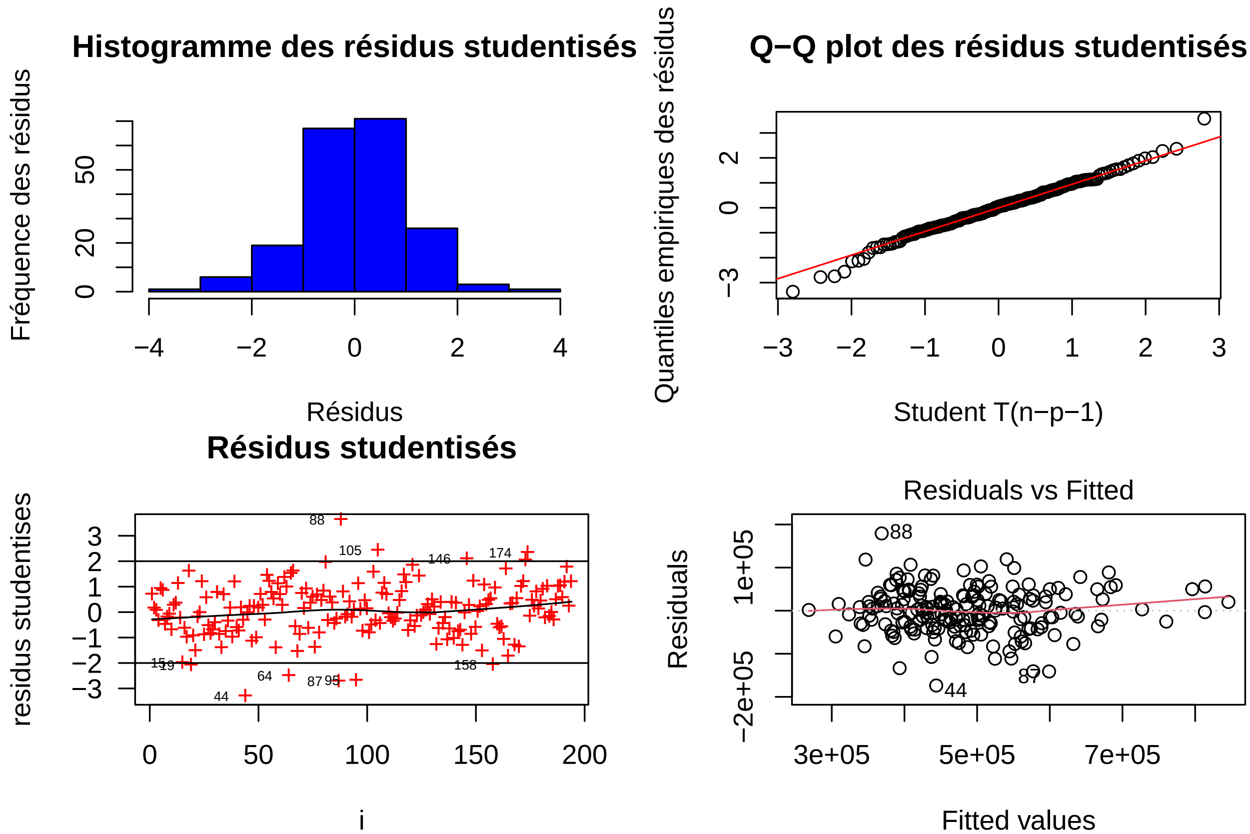
- **Application du modèle avec suppression de la valeur aberrante :**

```
## Call:
## lm(formula = price ~ sqft_above + sqft_basement + sqft_lot +
##     sqft_living15 + sqft_lot15 + age, data = house_data_kcBT[-c(141),
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -173701  -33239    1948   33576  179301
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.800e+05  2.421e+04   7.433 3.69e-12 ***
## sqft_above     1.077e+02  9.596e+00  11.224  < 2e-16 ***
## sqft_basement  7.447e+01  1.067e+01   6.981 4.96e-11 ***
## sqft_lot      -9.150e-01  2.854e-01  -3.206  0.00158 **
## sqft_living15  3.381e+01  1.130e+01   2.993  0.00314 **
## sqft_lot15     3.869e+00  8.718e-01   4.438 1.55e-05 ***
## age           -8.332e+02  2.625e+02  -3.174  0.00176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53310 on 187 degrees of freedom
## Multiple R-squared:  0.7905, Adjusted R-squared:  0.7838
## F-statistic: 117.6 on 6 and 187 DF,  p-value: < 2.2e-16
```

```
##    sqft_above sqft_basement      sqft_lot sqft_living15    sqft_lot15
##      3.037434      1.476303      1.386935      2.572653      1.314061
##           age
##      1.699395
```

H_0 ce qui indique que `sqft_above`, `sqft_basement`, `sqft_lot`, `sqft_living15`, `sqft_lot15` et `age` ont vraisemblablement un effet significatif sur price. La valeur du R^2 ajusté de notre nouveau modèle s'est améliorée, elle atteint désormais 0.7838, ce qui est satisfaisant.

- **Analyse des résidus studentisés :**



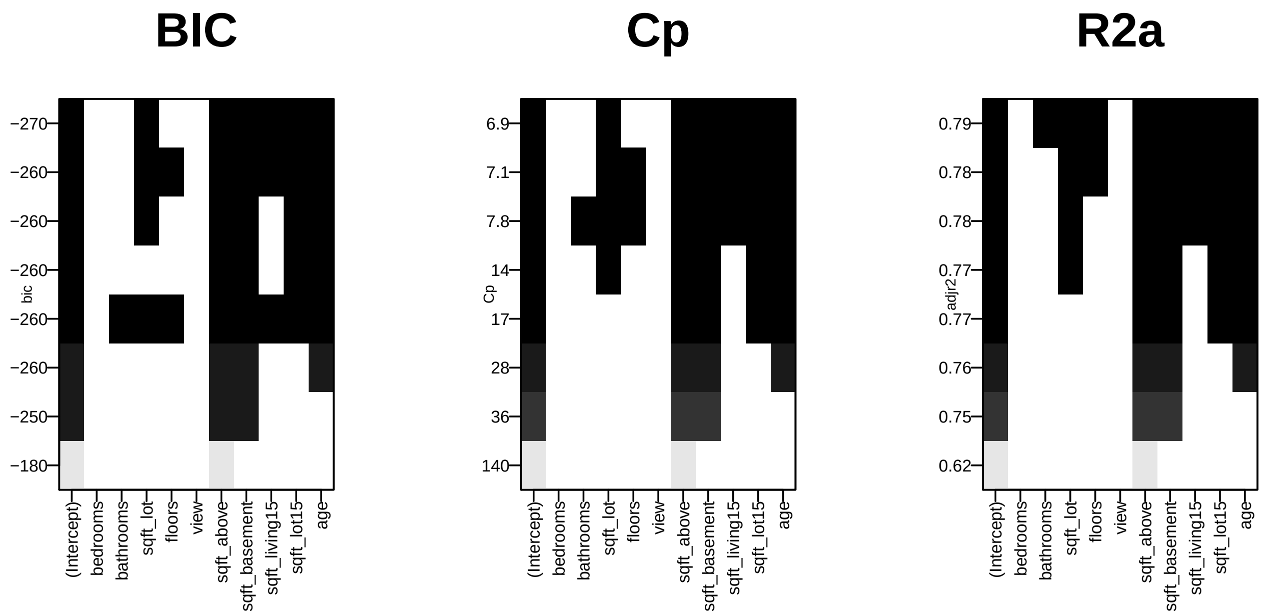

```
##
##  Shapiro-Wilk normality test
##
## data:  rstudent(rlmfin2)
## W = 0.99033, p-value = 0.2174
```

Sur histogramme les résidus studentisés semblent suivre une distribution normale centrée. De plus, l'alignement du Q-Q plot des résidus studentisés avec la loi gaussienne corrobore également la validité des hypothèses.
Nous observons que les résidus studentisés sont centrés en 0. Le lisseur n'a pas de tendance, on en déduit l'indépendance des résidus. Nous observons également que nos résidus studentisés en fonction des valeurs prédites forment bien un nuage de points quelconques centrés en 0 et le lisseur ne forme pas de tendance non plus, ce qui atteste leur homoscédasticité.
De plus le résultat du test de Shapiro-Wilk est très satisfaisant, nous acceptons l'hypothèse H_0 selon laquelle les residus théoriques ε_i suivent la loi normale $\mathcal{N}(0, \sigma^2)$, au risque 21%. Le test de Kolmogorov des résidus studentisés valide également l'hypothèse H_0 avec une p-valeur de 0.792 nettement supérieure au seuil.

L'ensemble des analyses permet de valider le modèle de régression linéaire multiple : `price` $= 180000 + 107.8\times$ `sqft_above` $+ 74.47\times$ `sqft_basement` - $91.5\times$ `sqft_lot` $+ 33.81\times$ `sqft_living15` $+ 3.869\times$ `sqft_lot15` - $833.3\times$ `age`$+\ \varepsilon$. est validé.

- **Choix des variables :**

Afin de valider le choix des variables pertinentes dans notre modèle linéaire, nous allons identifier les meilleurs modèles selon les critères BIC, $\mathcal{C}_p$ et $\mathcal{R}^2_\alpha$.



Les critères BIC et C_p de Mallows concordent pour le choix du modèle : `price`$= \beta_1 + \beta_2$ `sqft_lot` $+ \beta_3$ `sqft_above` $+ \beta_4$ `sqft_basement` $+ \beta_5$ `sqft_living15` $+ \beta_6$ `sqft_lot15`$+ \beta_7$ `age`$+ \varepsilon$.
Nous remarquons que c'est le même modèle que nous avons réalisé manuellement. L'analyse permet donc de valider notre modèle. Ceci confirme le choix des variables utilisées pour la création du modèle.

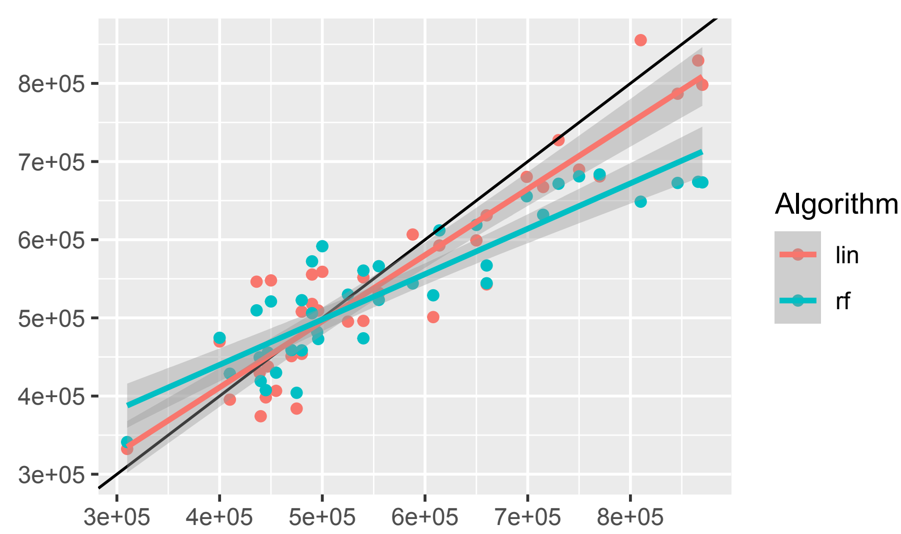
Le critère R_a^2 choisit le modèle : `price` $= \beta_1 + \beta_2$ `bathrooms`$+ \beta_3$ `sqft_lot` $+ \beta_4$ `floors`$+ \beta_5$ `sqft_above` $+ \beta_6$ `sqft_basement` $+ \beta_7$ `sqft_living15` $+ \beta_8$ `sqft_lot15`$+ \beta_9$ `age`$+ \varepsilon$.
En appliquant la régression, la valeur R_a^2 est de 0.7847, légèrement plus élevée (de 0.09) que celle de notre modèle initial. Mais certaines variables de ce modèle ne sont pas significatives (ex : `bedrooms` avec une p-valeur de 0.32).

Cette étude comforte le choix de la régression linéaire multiple pour les données considérées.

4- Prédiction

Pour mener une prédiction avec notre modèle, nous allons diviser notre jeu de données en 2 :
- Les premiers 80% qui servent à l'entrainement du modèle (la partie `train`)
- Les 20% restants qui permettent d'évaluer le modèle (la partie `pred`)

Nous allons comparer l'efficacité de la prédiction de notre modèle avec celle de la méthode du random forest.



Les points rouges sont les points prédits par notre modèle et les points verts ceux du modèle random forest. Nous observons que les deux nuages de points sont similaires. Nous remarquons de plus que la prédiction effectuée avec notre modèle semble meilleure que celle effectuée par le modèle random forest, surtout dans le cas où les prix sont élevés. En effet, la droite des moindres carrés de notre modèle est plus proche de la première bissectrice que celle du random forest.

5- Conclusion

Dans ce projet nous avions pour objectif d'estimer et prédire le prix de vente des maisons dans la ville de Bothell, WA, USA. Pour ce faire nous avons réalisé des modèles de régression linéaire simple et multiple, qui ont tous les deux été validés par les analyses des résidus studentisés. Notre modèle de régression linéaire multiple : `price` = 180000 + 107.8. `sqft_above` + 74.47. `sqft_basement` - 91.5. `sqft_lot` + 33.81. `sqft_living15` + 3.869. `sqft_lot15` - 833.3. `age`+ ε a atteint un R_a^2 de 0.7838, ce qui est décent. La prédiction de notre modèle s'est avérée plutôt bonne, semblant même faire une meilleure prédiction que l'algorithme du random forest.