

projet stat MOALI Sarah & LAFDHAL Ahmed

Contents

Anayse des données	2
Résumée des variables	2
Densité des variable	3
Scatter plots	4
Modélisation linéaire simple	5
premiere modèle simple (juste pour verifier les observations)	5
Modélisation linéaire multiple	7
premier essaie	7
Amélioration du modèle 1	9
Amélioration du modèle 2	10

Anayse des données

```
library(readr)
abalone <- read_csv("/Users/ahmedlafdhala/Desktop/projet/abalone.csv", show_col_types = FALSE)
kable(abalone[1:10,], digits = 4,format = 'markdown')
```

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150	15
M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070	7
F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210	9
M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155	10
I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055	7
I	0.425	0.300	0.095	0.3515	0.1410	0.0775	0.120	8
F	0.530	0.415	0.150	0.7775	0.2370	0.1415	0.330	20
F	0.545	0.425	0.125	0.7680	0.2940	0.1495	0.260	16
M	0.475	0.370	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.550	0.440	0.150	0.8945	0.3145	0.1510	0.320	19

Résumée des variables

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
F:1307	Min.	Min.	Min.	Min.	Min.	Min.	Min.	Min. :
:0.075	:0.0550	:0.0000	:0.0020	:0.0010	:0.0005	:0.0015	:0.0015	1.000
I:1342	1st	1st	1st	1st	1st	1st	1st	1st Qu.:
Qu.:0.450	Qu.:0.3500	Qu.:0.1150	Qu.:0.4415	Qu.:0.1860	Qu.:0.0935	Qu.:0.1300	Qu.:0.1300	8.000
M:1528	Median	Median	Median	Median	Median	Median	Median	Median :
:0.545	:0.4250	:0.1400	:0.7995	:0.3360	:0.1710	:0.2340	:0.2340	9.000
NA	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean :
:0.524	:0.4079	:0.1395	:0.8287	:0.3594	:0.1806	:0.2388	:0.2388	9.934
NA	3rd	3rd	3rd	3rd	3rd	3rd	3rd	3rd
Qu.:0.615	Qu.:0.4800	Qu.:0.1650	Qu.:1.1530	Qu.:0.5020	Qu.:0.2530	Qu.:0.3290	Qu.:0.3290	Qu.:11.000
NA	Max.	Max.	Max.	Max.	Max.	Max.	Max.	Max.
:0.815	:0.6500	:1.1300	:2.8255	:1.4880	:0.7600	:1.0050	:1.0050	:29.000

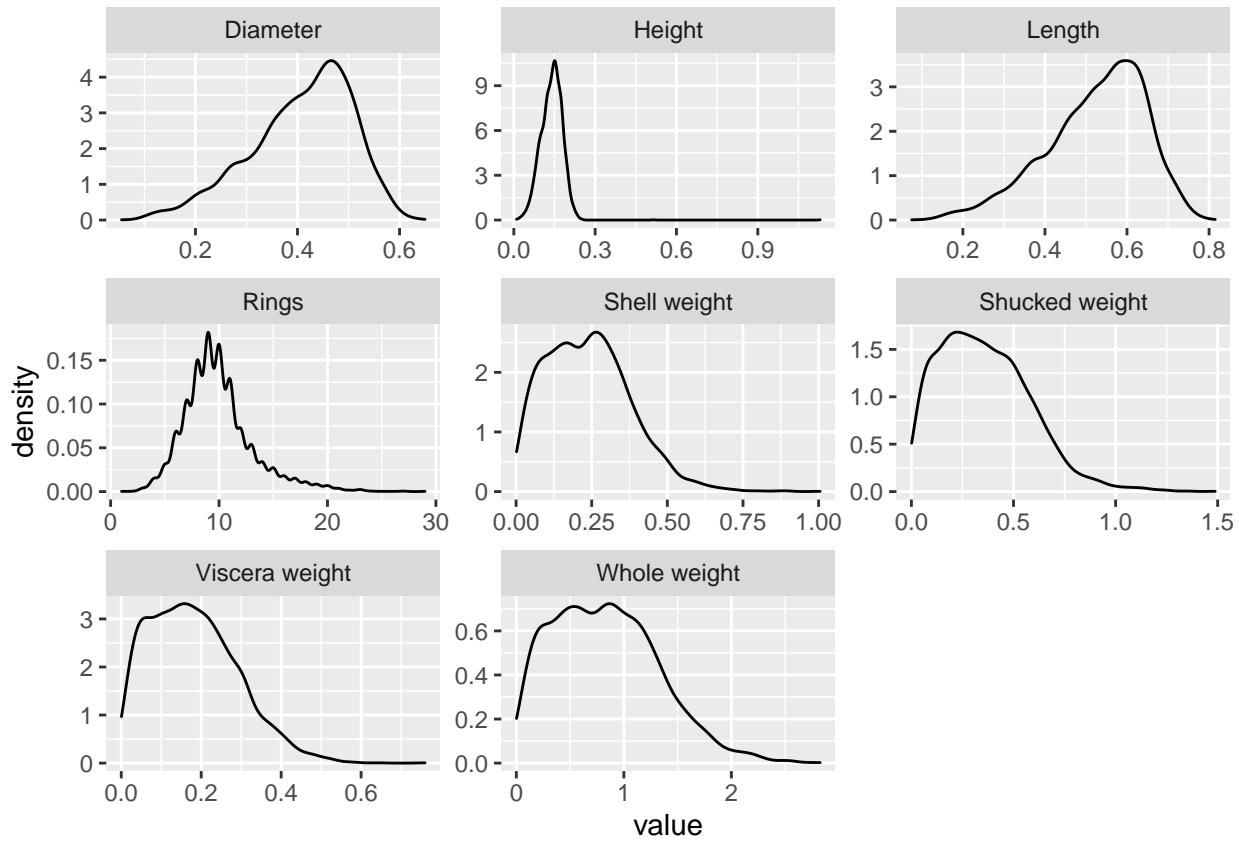
-Nous pouvons voir d'après le summary de l'ensemble de données, que les données sont presque également distribuées entre les trois sexes : male, female et infant.

-Nous avons quatre types de poids différents : Whole_weight, Shucked_weight, Viscera_weight et Shell.weight. Si nous regardons la moyenne de ces poids, nous pouvons supposer que Whole_weight est peut-être la somme d'autres prédicteurs de poids avec une erreur de masse inconnue.

-Nous constatons également que la colonne Height a un minimum de 0 ce qui est impossible. Il s'agit probablement d'une erreur de données. Afin d'éviter les valeurs aberrantes, nous devons supprimer ces observations.

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
I	0.430	0.34	0	0.428	0.2065	0.0860	0.1150	8
I	0.315	0.23	0	0.134	0.0575	0.0285	0.3505	6

Densité des variables

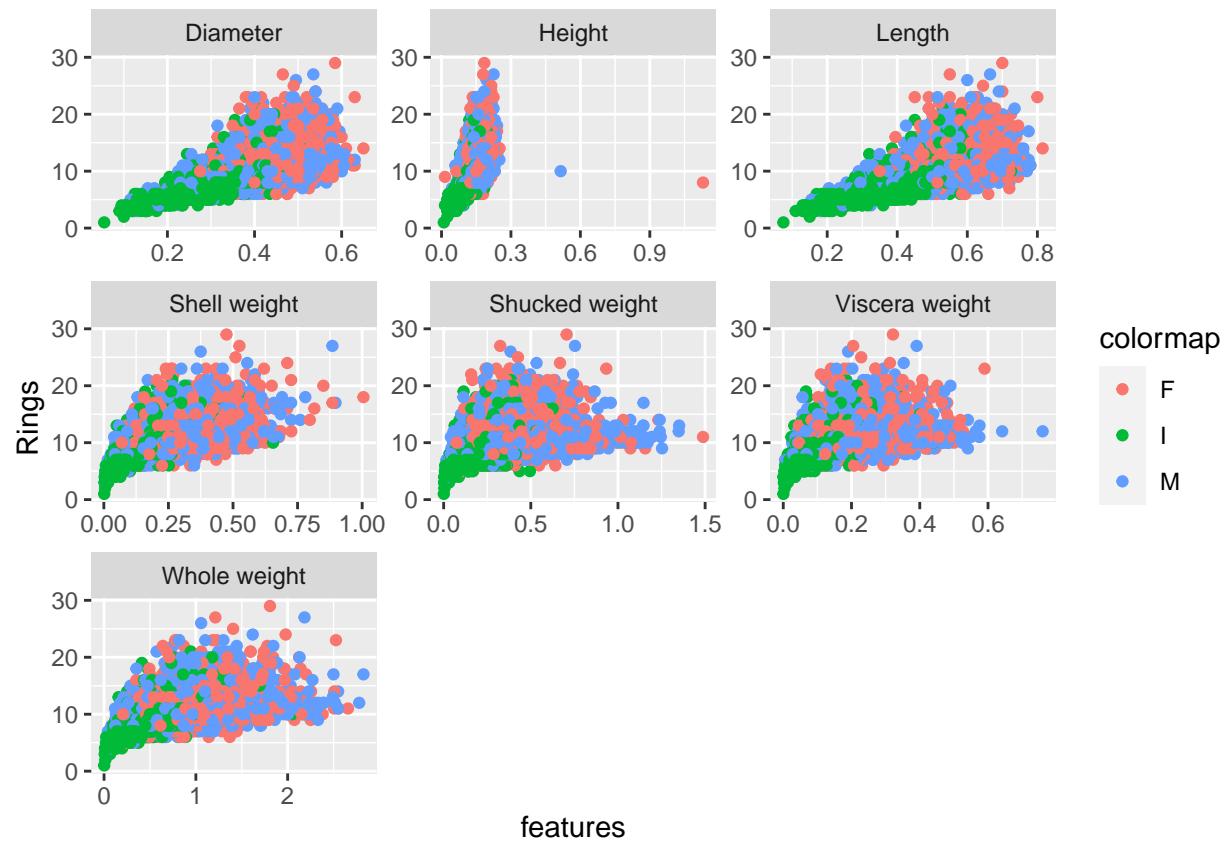


Le nombre de rings varie de 1 à 29, la plupart des observations se situant entre 5 et 15. La distribution est légèrement biaisée positivement.

Il semble que les hauteurs soient distribuées normalement avec seulement deux observations supérieures à 0.5 (voir le tableau ci-dessous). Ces observations sont probablement des valeurs aberrantes.

Sex	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
M	0.705	0.565	0.515	2.210	1.1075	0.4865	0.5120	10
F	0.455	0.355	1.130	0.594	0.3320	0.1160	0.1335	8

Scatter plots



La première chose à constater est que le nombre d'abalone semble être fortement corrélé et augmenter avec chacune des variables. Comme prévu, nous constatons logiquement que les enfants ont un petit âge . Cependant, les attributs masculins et féminins ne semblent pas avoir un effet clair sur l'âge des abalone. L'hypothèse des biologistes n'est pas totalement fausse. Mais d'après le nuage de points, nous pouvons nous attendre à des résidus importants d'une régression linéaire entre la hauteur des abalone et leur âge. À moins que toutes les caractéristiques ne soient parfaitement corrélées.

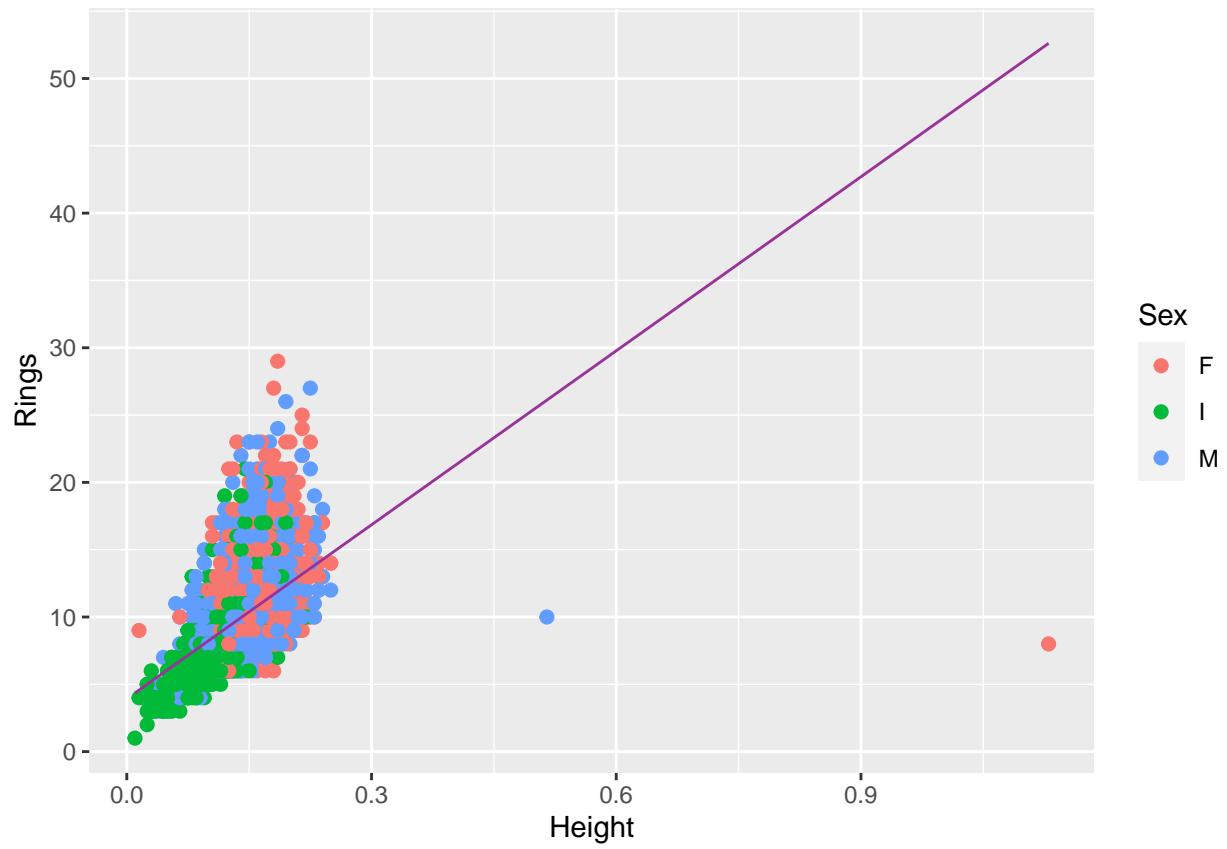
Modélisation linéaire simple

premiere modèle simple (juste pour vérifier les observations)

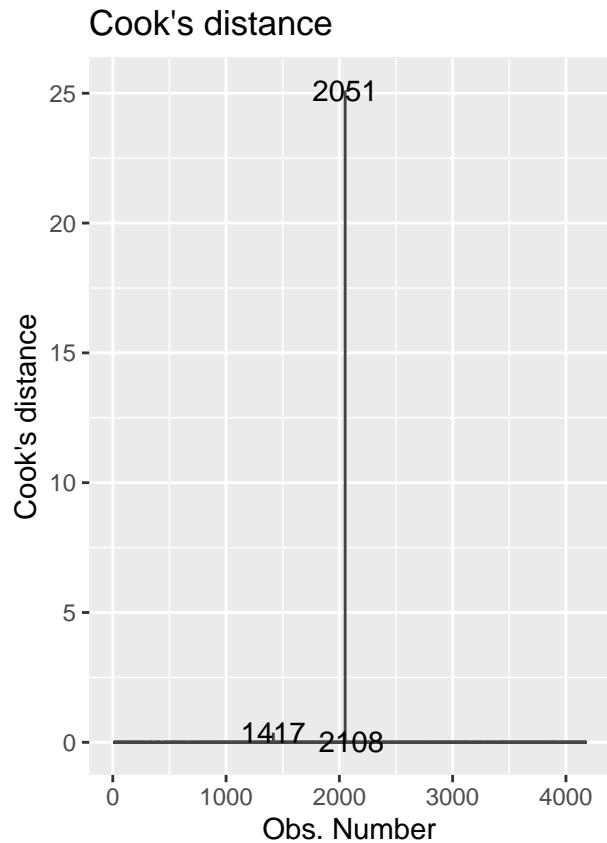
```
##  
## Call:  
## lm(formula = Rings ~ Height, data = abalone)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -44.611  -1.644  -0.644   0.832  17.108  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  3.9206    0.1447   27.1   <2e-16 ***  
## Height       43.0891   0.9930   43.4   <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 2.677 on 4173 degrees of freedom  
## Multiple R-squared:  0.3109, Adjusted R-squared:  0.3108  
## F-statistic: 1883 on 1 and 4173 DF, p-value: < 2.2e-16
```

les deux sont statistiquement très significatif mais le model explique que 31% de la variabilité des données.

Sur le graphique ci-dessous, nous voyons que la ligne de régression ne passe pas exactement là où nous le souhaiterions. Sa pente est faible. Ceci est probablement dû aux deux points extrêmes de très grande hauteur. Pour vérifier cette hypothèse, nous allons utiliser la distance de Cook pour détecter les points aberrants.



Comme nous l'avions imaginé, le graphique de la distance COOK ci-dessous montre que les points 1417 et 2051 ont une distance relativement très grande (supérieure à 1 pour 2051 et proche de 1 pour 1417). Nous allons supprimer ces points aberrants Pour la suite



Modélisation linéaire multiple

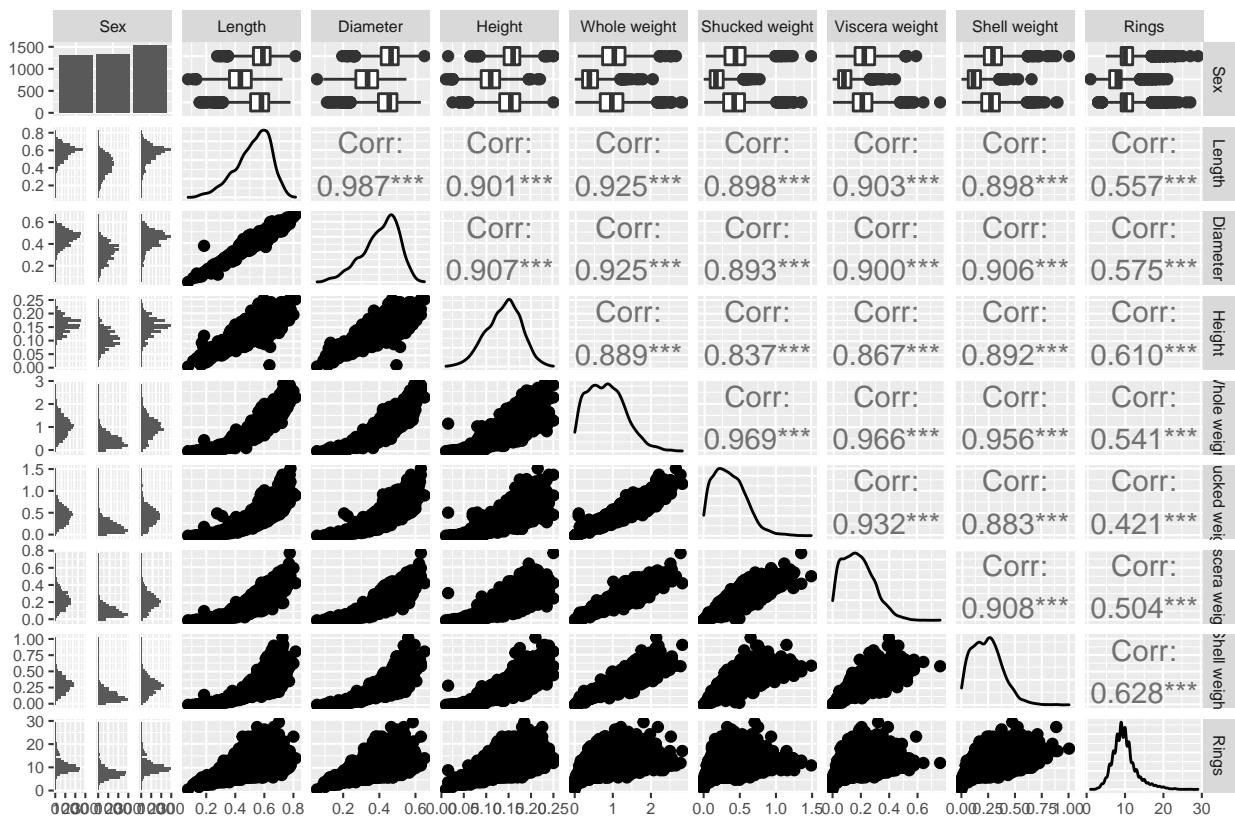
premier essaie

Nous allons d'abord supprimés les valeurs aberrants

```
new_abalone=abalone[-c(which(abalone$Height>=0.5)),]
```

Afin de mieux expliquer la variabilité des âges des abalones, nous allons considérer toutes les variables.

Jeu de donnée abalone



Nous observons dans ce graphe qu'il existe une forte corrélation entre les données. Par exemple, la corrélation entre le length et le height est extrêmement élevée (environ 98,7). Les corrélations entre le whole weight et les trois autres weight sont toutes supérieures à 95,6.

Analysons d'abord une régression multiples en utilisant toutes les variables continues des données.

```
##
## Call:
## lm(formula = Rings ~ Length + Diameter + Height + 'Whole weight' +
##     'Shucked weight' + 'Viscera weight' + 'Shell weight', data = new_abalone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -8.1048 -1.3671 -0.3697  0.9058 13.9659 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 2.6761    0.2710  9.876 < 2e-16 ***
## Length      -2.2594   1.8146 -1.245   0.213    
## Diameter     11.2109   2.2378  5.010 5.68e-07 ***
## Height      25.3618   2.3188 10.938 < 2e-16 ***
## 'Whole weight' 9.0624   0.7332 12.360 < 2e-16 ***
## 'Shucked weight' -19.7265  0.8233 -23.961 < 2e-16 ***
## 'Viscera weight' -10.4129  1.2997 -8.012 1.45e-15 ***
## 'Shell weight' 7.5238   1.1556  6.511 8.35e-11 ***
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.202 on 4165 degrees of freedom
## Multiple R-squared:  0.5345, Adjusted R-squared:  0.5337
## F-statistic: 683.1 on 7 and 4165 DF,  p-value: < 2.2e-16

## [1] 8.000 6596.708

```

La première chose à remarquer dans les résultats ci-dessous est que le length n'est pas significative , la p-value de son test t est de 0,54. Ceci est principalement dû à la forte corrélation entre length et Diameter , Le length n'est ni biologiquement ni statistiquement significative, nous ne l'utiliserons donc pas pour le reste de nos modèles. Cependant, ce n'est pas le cas pour la variable Whole weight. Bien qu'elle soit corrélée avec d'autres poids, elle reste très significative sur le plan statistique. Cela pourrait signifier que le poids total d'un abalone porte une masse inconnue (eau ou sang par exemple) qui a un effet sur l'âge.

Amélioration du modèle 1

Nous allons supprimer le length pour voir

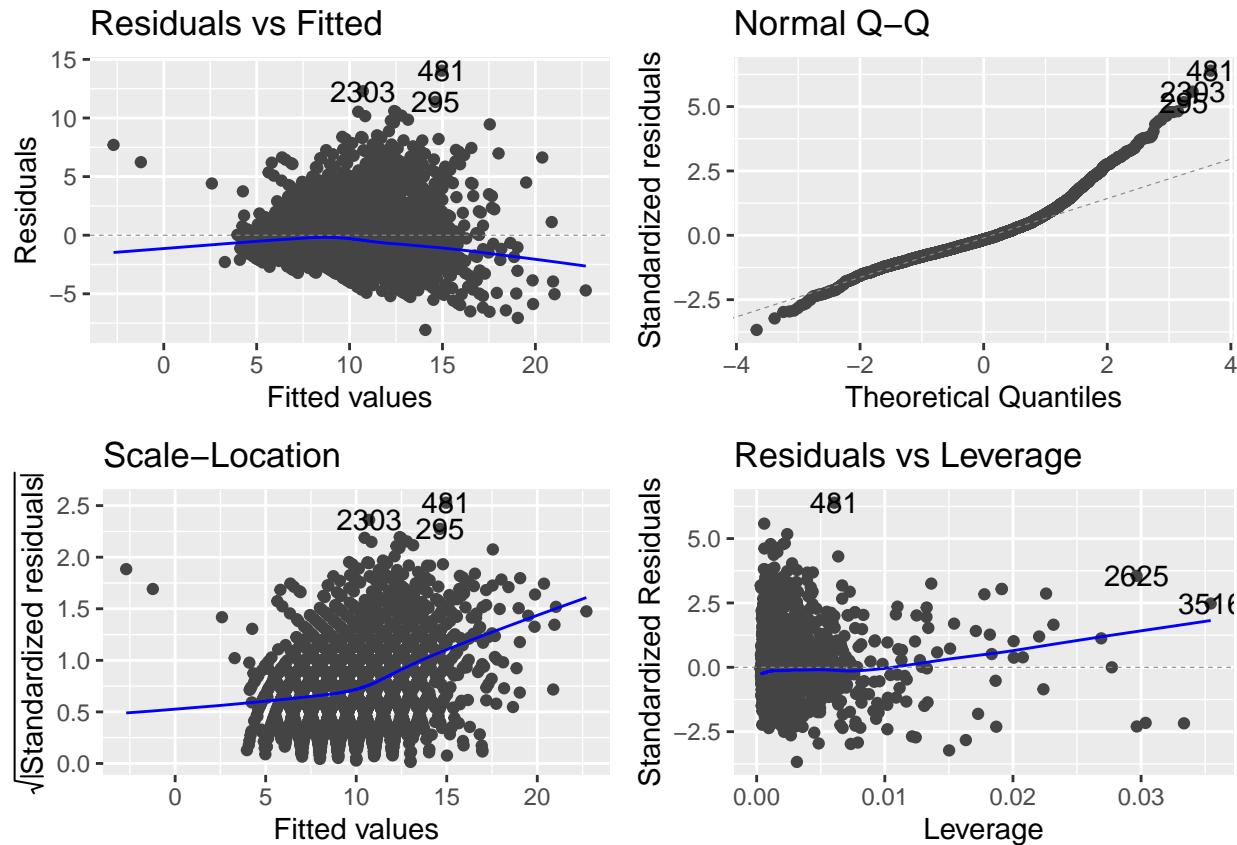
Voici le summary du modèle :

```

##
## Call:
## lm(formula = Rings ~ Diameter + Height + 'Whole weight' + 'Shucked weight' +
##     'Viscera weight' + 'Shell weight', data = new_abalone)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.0817 -1.3645 -0.3736  0.9096 14.0353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.5501    0.2514 10.144 < 2e-16 ***
## Diameter    8.7510    1.0512  8.325 < 2e-16 ***
## Height      25.2026   2.3154 10.885 < 2e-16 ***
## 'Whole weight' 9.0748   0.7332 12.377 < 2e-16 ***
## 'Shucked weight' -19.8113  0.8205 -24.145 < 2e-16 ***
## 'Viscera weight' -10.5530  1.2949  -8.150 4.77e-16 ***
## 'Shell weight' 7.5785    1.1548   6.563 5.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.202 on 4166 degrees of freedom
## Multiple R-squared:  0.5343, Adjusted R-squared:  0.5336
## F-statistic: 796.6 on 6 and 4166 DF,  p-value: < 2.2e-16

```

```
autoplot(reg)
```



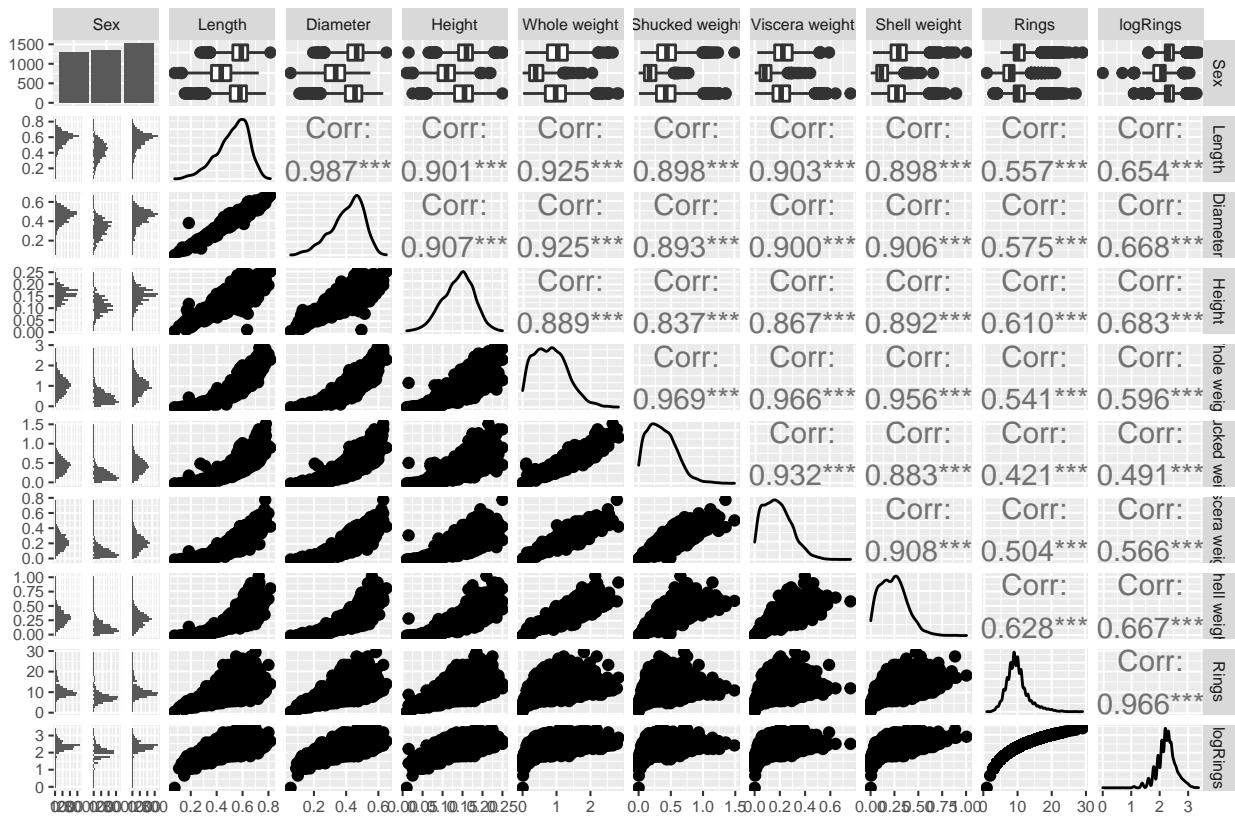
Amélioration du modèle 2

Comme toutes nos hypothèses ne sont pas satisfaites, nous allons essayer de modifier les données pour y porter solution. Si on regarde le graphe Q-Q précédent on constate que si on applique la fonction logarithmique sur les valeurs $(y_i)_{i=1,2,\dots,n}$. Cette transformation va en quelque sorte diminuer les grandes valeurs de $(\hat{\epsilon}_i)_{i=1,2,\dots,n}$ plus que les autres (effet logarithmique), ce qui aura pour effet de réduire la partie finale de la distribution de droite.

```
new_abalone$logRings <- log(new_abalone$Rings)
```

```
ggpairs(new_abalone, title="Jeu de donnée abalone") +
  theme_grey(base_size = 8)
```

Jeu de donnée abalone



En regardant la dernière ligne du graphique ggpairs, nous pouvons voir que les variables Rings et logRings sont proches d'une fonction concave des poids. Ce qui indique que l'application d'une transformation concave (sqrt, log, ...) sur les poids donnerait probablement de meilleurs résultats.

Le meilleure modèle finale est donc:

$$\log(R) = \beta_1 D + \beta_2 D^2 + \beta_3 \sqrt{H} + \beta_4 \sqrt{\text{Whole_weight}} + \beta_5 \sqrt{\text{Shucked_weight}} + \beta_6 \sqrt{\text{Shell_weight}} + \beta_7 \sqrt{\text{Viscera_weight}}$$

d'où R : Rings D : Diameter and H : Height.

Voici le summary du modèle:

```
reg_tot = lm(logRings ~ poly(Diameter, 2) + sqrt(Height) + sqrt(`Whole weight`) + sqrt(`Shucked weight`))
summary(reg_tot)
```

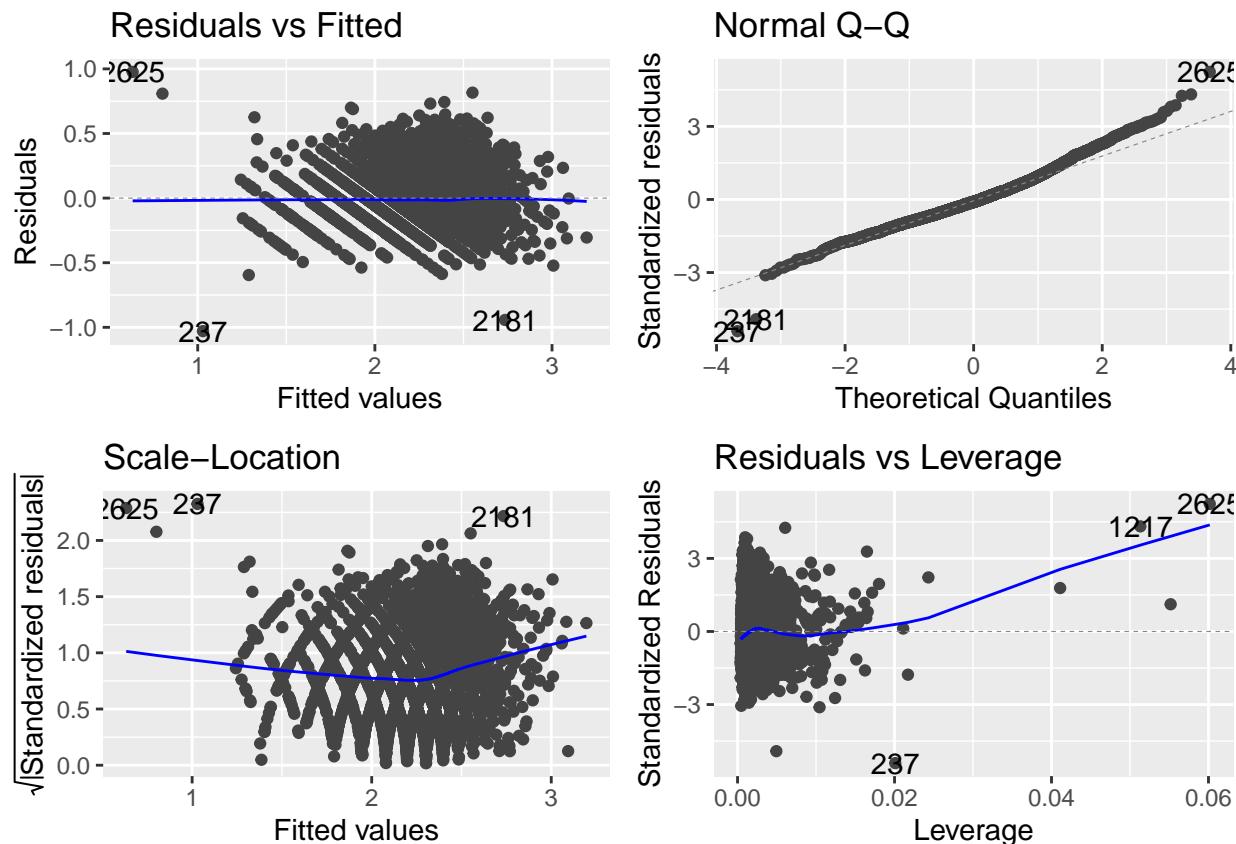
```
##
## Call:
## lm(formula = logRings ~ poly(Diameter, 2) + sqrt(Height) + sqrt('Whole weight') +
##     sqrt('Shucked weight') + sqrt('Viscera weight') + sqrt('Shell weight'),
##     data = new_abalone)
##
## Residuals:
##      Min        1Q    Median        3Q       Max
## -1.02953 -0.12889 -0.01887  0.10879  0.97667
##
## Coefficients:
```

```

##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 1.42560   0.05591 25.499 < 2e-16 ***
## poly(Diameter, 2)1         2.28534   0.89934  2.541  0.0111 *
## poly(Diameter, 2)2        -3.54757   0.21611 -16.416 < 2e-16 ***
## sqrt(Height)                1.12310   0.15870  7.077 1.72e-12 ***
## sqrt('Whole weight')      1.60254   0.13000 12.327 < 2e-16 ***
## sqrt('Shucked weight')   -2.23595   0.09551 -23.410 < 2e-16 ***
## sqrt('Viscera weight')   -0.56819   0.10759 -5.281 1.35e-07 ***
## sqrt('Shell weight')       1.11436   0.11585  9.619 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1924 on 4165 degrees of freedom
## Multiple R-squared:  0.6382, Adjusted R-squared:  0.6376
## F-statistic: 1050 on 7 and 4165 DF, p-value: < 2.2e-16

```

Toutes les pentes sont statistiquement très significatives. Par rapport au modèle précédent le R^2 est passé de 53.36% à 63.74%. Il semble que le modèle soit beaucoup plus performant, mais nous devons encore vérifier sa validité.



En examinant les graphiques ci-dessous, nous pouvons affirmer qu'il s'agit d'un modèle acceptable. Les deux premiers graphiques montrent que les résidus ne sont pas corrélés aux valeurs ajustées et suivent une distribution normale.