RAPPORT DE PROJET DE RÉGRESSION ET CLASSIFICATION MAP569

Projet : Étude statistique d'une base de donnée d'ormeaux

20 mars 2022

Paul Landrup, Christian Loroux, Yanis Boudjema

ÉCOLE POLYTECHNIQUE

IP PARIS

1 PARTIE I : RÉGRESSION LINÉAIRE ET VALIDATION DE MODÈLE

Le dataset proposé est un tableau où chaque ligne correspondait à un ormeaux. En colonne, on peut voir plusieurs caractéristiques mesurables sur l'abalone, par exemple :

- le sexe : M, F ou I
- l'age : représenté par le nombre d'anneaux
- la taille

Nous nous posons la question de déterminer un lien entre une ou plusieurs de ces caractéristiques mesurables sur l'ormeau. Plus précisément, puisque les biologistes témoignent de leur intérêt à connaître l'âge des ormeaux (qui plus est, n'est pas la caractéristique la plus aisée à mesurer), nous allons, dans ce projet, nous intéresser à la relation qui peut exister entre l'âge des abalones et leurs autres caractéristiques (taille, sexe, etc.).

1.1 Question 1

Dans cette première question, nous allons nous écrire la relation mathématique entre l'âge des ormeaux et leurs autres caractéristiques, dans le cas où l'on souhaite effectuer une régression linéaire. Introduisons d'abord les notations suivantes :

- y : nombre d'anneaux de l'ormeau (qui est directement proportionnel à son âge)
- x_1 : longueur de l'ormeau
- x_2 : diamètre de l'ormeau
- x_3 : hauteur de l'ormeau
- x_4 : poids (entier) de l'ormeau
- x_5 : poids de l'ormeau sans la coquille
- x_6 : poids des viscères de l'ormeau
- x_7 : poids de la coquille de l'ormeau
- x_8 : sexe de l'ormeau (0 pour une femelle, 1 pour un mâle)

Nous voulons effectuer une régression linéaire de y par rapport aux x_j, $j = 1, ..., 8$. On pose d'abord $n = 4176$ et l'on écrit la relation suivante :

$$y_i = \sum_{j=1}^{8} \beta_j x_{i,j} + \epsilon_i$$

Où :

- y_i, $i = 1, ..., n$: $i^{\text{ème}}$ mesure de la variable y
- $x_{i,j}$, $i = 1, ..., n$: $i^{\text{ème}}$ mesure de la variable x_j

— $\beta = (\beta_1, ..., \beta_8) \in \mathbf{R}^8$: vecteur des coefficients de la combinaison linéaire
— ϵ_i, $i = 1, ..., n$: erreur dans l'approximation de y_i par $\sum_{j=1}^{8} \beta_j x_{i,j}$

Nous faisons également l'hypothèse que les variables aléatoires $\epsilon_1, ...\epsilon_n$ vérifient les hypothèses [P1]-[P4] du cours.

Nous pouvons calculer, grâce à R, un estimateur de β, par la méthode des moindres carrés :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Afin d'étudier la pertinence statistique de nos résultats, nous pouvons effectuer, sur R, un test de Fisher qui permet de tester le fait que tous les coefficients soient non nuls, contre le fait d'avoir un coefficient nul. Nous pouvons ensuite effectuer un test afin de vérifier si le bruit est gaussien. Cette hypothèse de bruit gaussien est très intéressante pour établir des intervalles de confiance. En effet, une fois que l'on disposera des coefficients β_j de la régression linéaire et que l'on s'en servira pour que, connaissant $x_1, ..., x_8$ pour un ormeau, on puisse estimer son âge. Il faudra alors fournir des intervalles de confiance. Puisque l'on connaît bien les intervalles de confiances des variables gaussiennes, nous seront alors capable de calculer un intervalle de confiance, disons à 95%, de l'âge de l'ormeau, à partir des mesures de $x_1, ..., x_8$ qui ont été faites.

1.2 Question 2

Soit $j \in [\![1, 8]\!]$. Alors :
— La moyenne empirique de la variable x_j est $\bar{x_i} = \frac{x_{1,j}+x_{2,j}+...+x_{n,j}}{n}$
— La variance empirique de la variable x_j est $\hat{\sigma_j} = \frac{(x_{1,j}-\bar{x_i})^2+(x_{2,j}-\bar{x_i})^2+...+(x_{n,j}-\bar{x_i})^2}{n}$
— L'intervalle auquel appartient la variable x_j est $[\min_{1\leq i\leq n} x_{i,j},\ \max_{1\leq i\leq n} x_{i,j}]$

Nous obtenons donc les résultats suivants pour chaque variables étudiée, y et les $x_1, ..., x_8$:

1.2.1 • Moyenne des caractéristiques de l'ormeau

Length <dbl>	Diameter <dbl>	Height <dbl>	Whole_weight <dbl>	Shucked_weight <dbl>	Viscera_weight <dbl>	Shell_weight <dbl>	Rings <dbl>
104.8017	81.5783	27.90541	165.7635	71.87996	36.12253	47.77043	9.932471

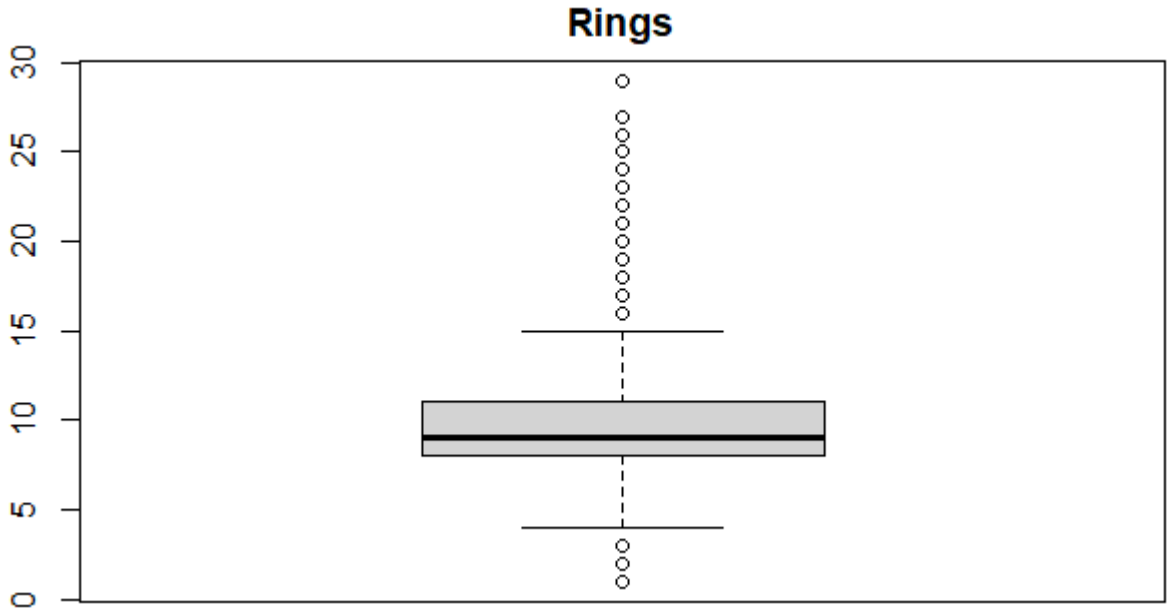
1.2.2 • Écart-type des caractéristiques de l'ormeau

Length <dbl>	Diameter <dbl>	Height <dbl>	Whole_weight <dbl>	Shucked_weight <dbl>	Viscera_weight <dbl>	Shell_weight <dbl>	Rings <dbl>
24.02051	19.84991	8.365278	98.08471	44.39594	21.92409	27.84251	3.223601

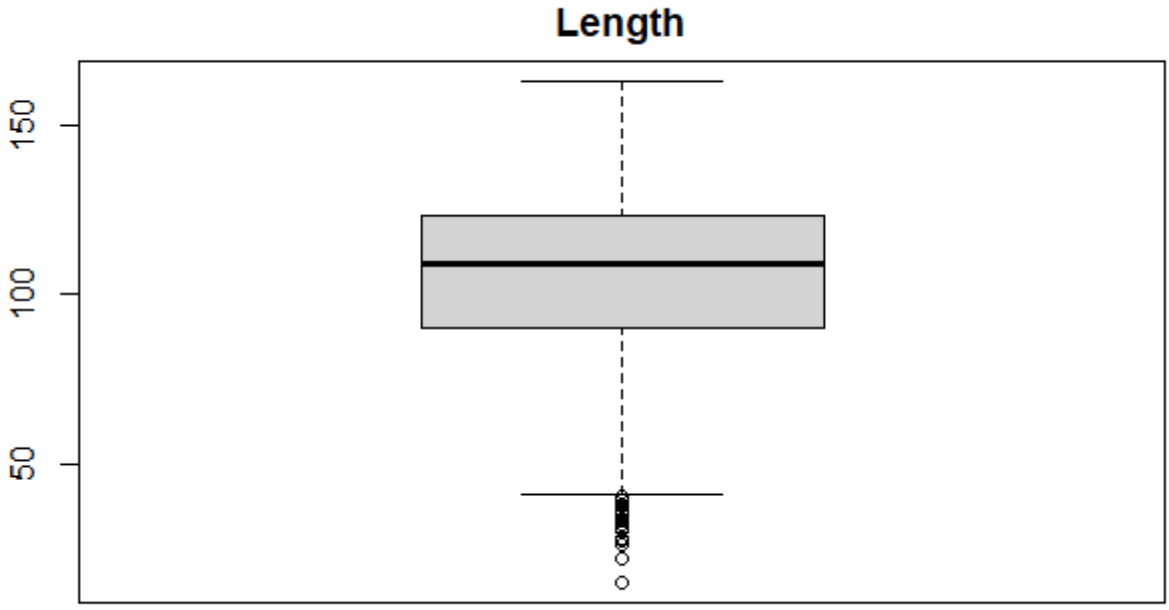
1.2.3 • Intervalle de valeurs pour les caractéristiques de l'ormeau

Length <int>	Diameter <int>	Height <int>	Whole_weight <dbl>	Shucked_weight <dbl>	Viscera_weight <dbl>	Shell_weight <dbl>	Rings <int>
15	11	0	0.4	0.2	0.1	0.3	1
163	130	226	565.1	297.6	152.0	201.0	29

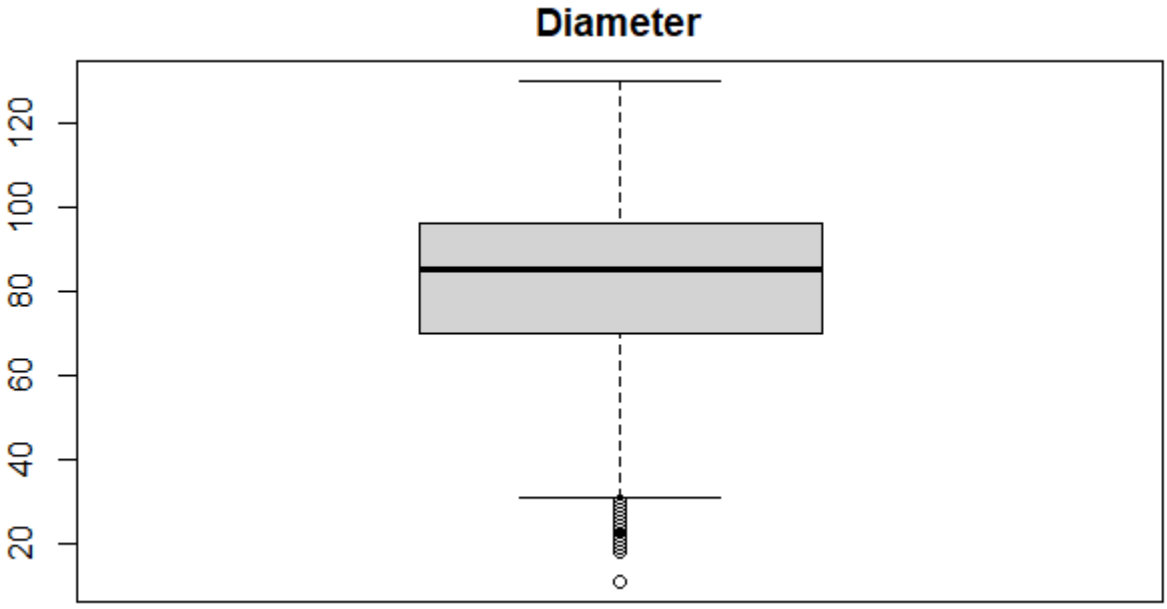
1.2.4 • nombre d'anneaux de l'ormeau



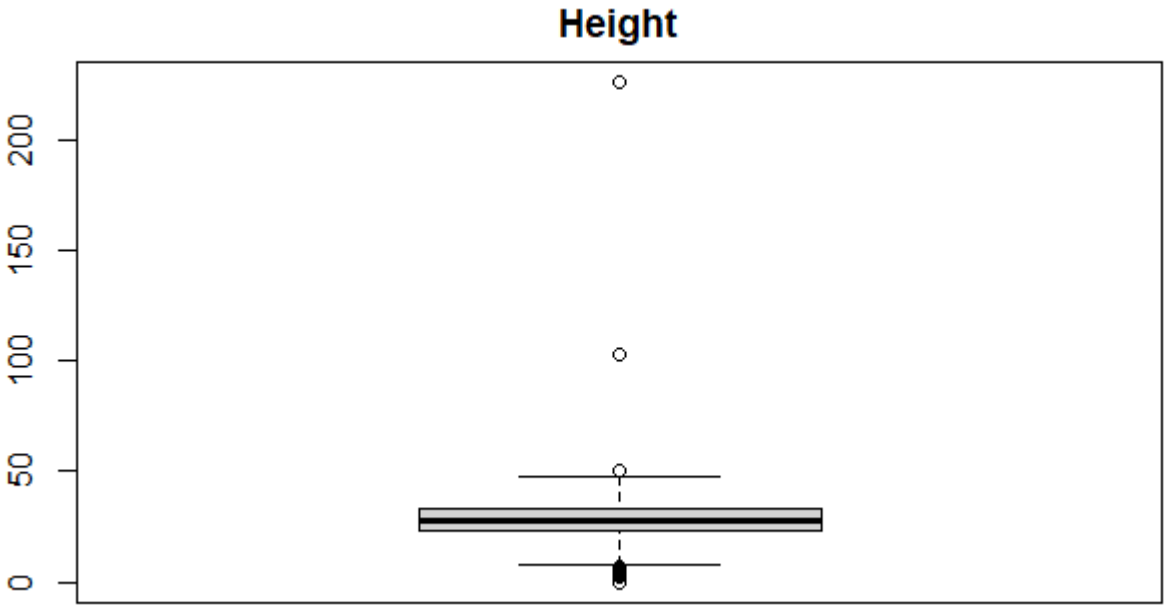
1.2.5 • longueur de l'ormeau



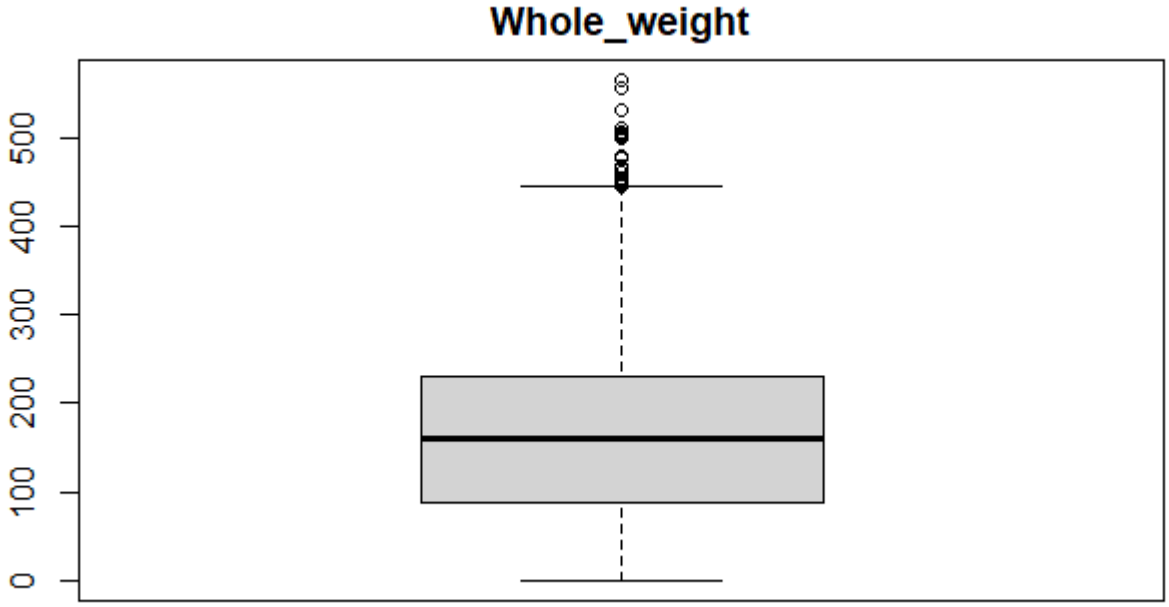
1.2.6 • diamètre de l'ormeau



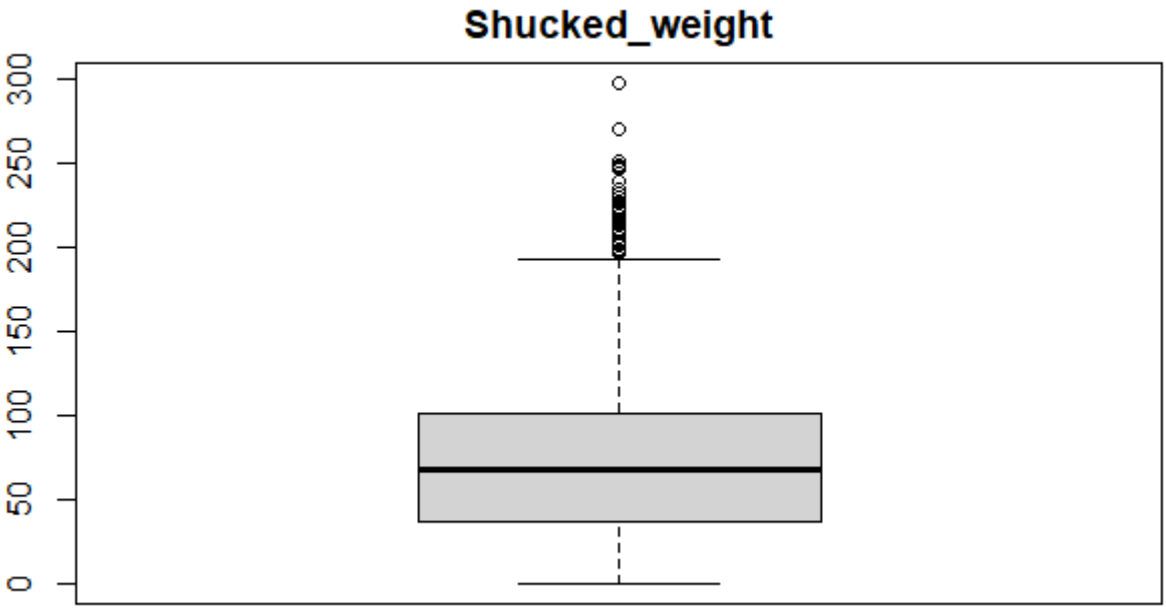
1.2.7 • hauteur de l'ormeau



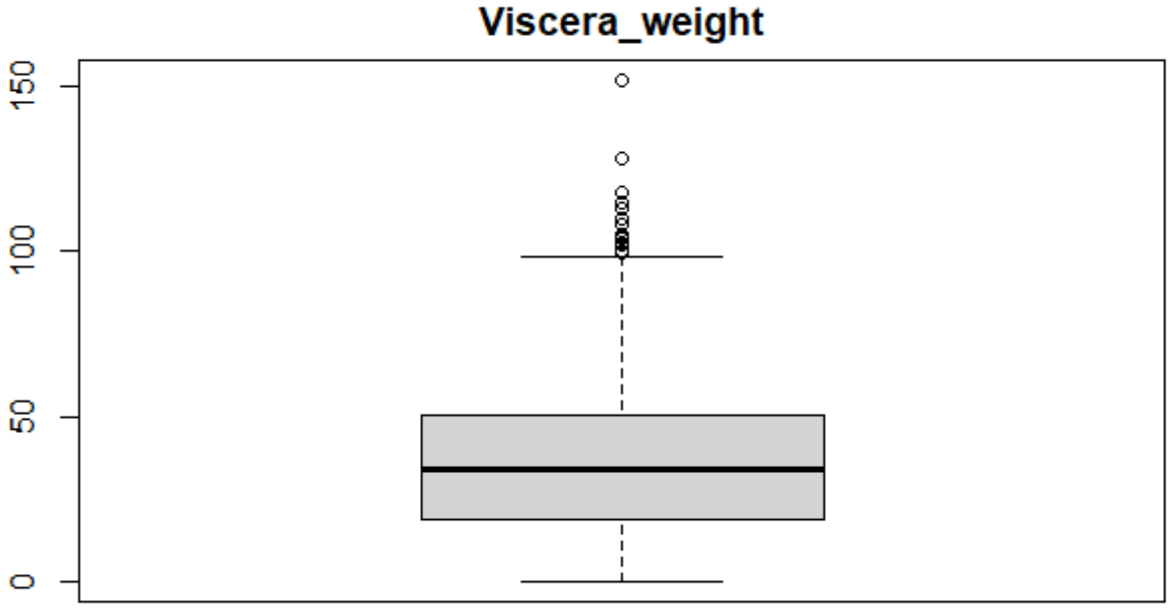
1.2.8 • poids (entier) de l'ormeau



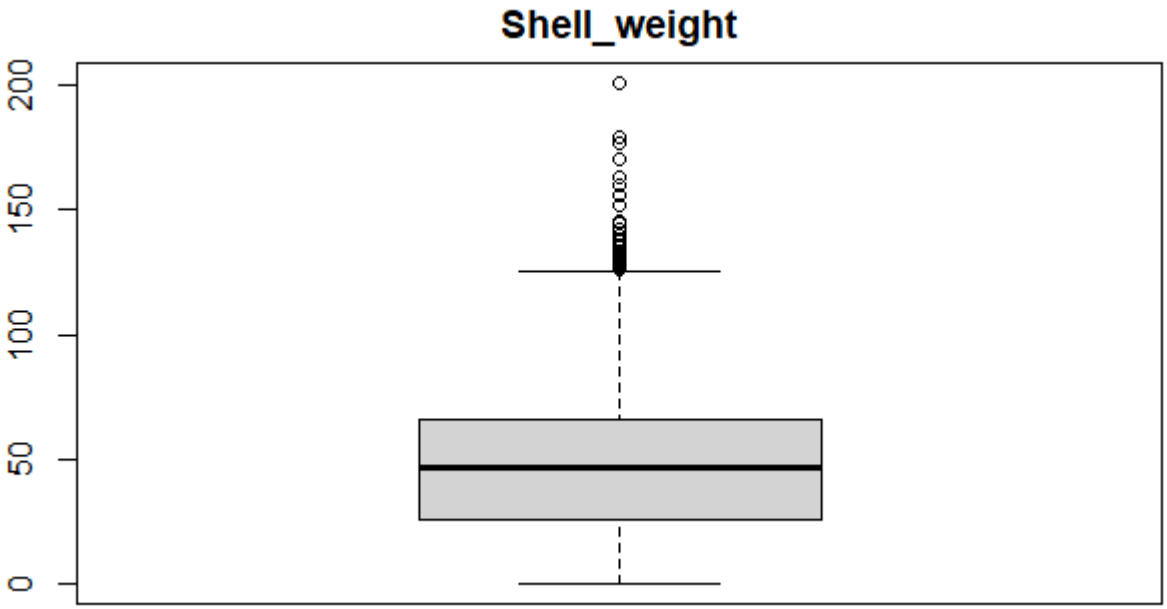
1.2.9 • poids de l'ormeau sans la coquille



1.2.10 • poids des viscères de l'ormeau



1.2.11 • poids de la coquille de l'ormeau



1.3 Question 3

Nous présentons ici les différentes représentations graphiques, représentant une variable caractéristique de l'ormeau, par rapport à une autre.

Tout d'abord, remarquons que la relation linéaire entre le nombre d'anneaux des ormeaux (donc l'âge) et la hauteur de l'ormeau, il n'y a pas vraiment de relation linéaire. Nous constatons même que, pour une même hauteur d'ormeau, plusieurs nombres d'anneaux (donc âges) différents correspondent.

1.4 Question 4

Nous effectuons ici la régression linéaire pour exprimer le nombre d'anneaux d'un ormeau en fonction de sa hauteur. Il s'agit là de la régression entre la variable y et la variable x_3 sous la forme : $y = ax_3 + b + \epsilon$. Avec toutes les données du dataset, nous pouvons calculer des estimateurs de a et b comme suit :

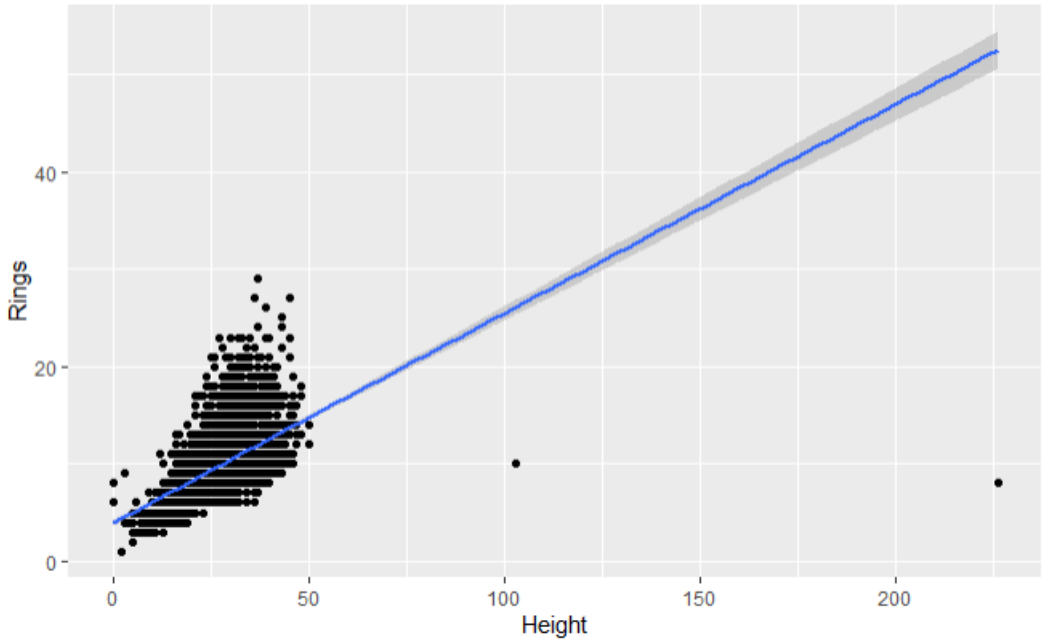
$$\hat{a} = \frac{(\frac{1}{n}\sum_{i=1}^{n} y_i)(\frac{1}{n}\sum_{i=1}^{n} x_{i,3}^2) - (\frac{1}{n}\sum_{i=1}^{n} x_{i,3})(\frac{1}{n}\sum_{i=1}^{n} x_{i,3}y_i)}{\frac{1}{n}\sum_{i=1}^{n} x_{i,3}^2 - (\frac{1}{n}\sum_{i=1}^{n} x_{i,3})^2}$$

$$\hat{b} = \frac{n(\frac{1}{n}\sum_{i=1}^{n} x_{i,3}y_i) - (\frac{1}{n}\sum_{i=1}^{n} x_{i,3})(\frac{1}{n}\sum_{i=1}^{n} y_i)}{\frac{1}{n}\sum_{i=1}^{n} x_{i,3}^2 - (\frac{1}{n}\sum_{i=1}^{n} x_{i,3})^2}$$

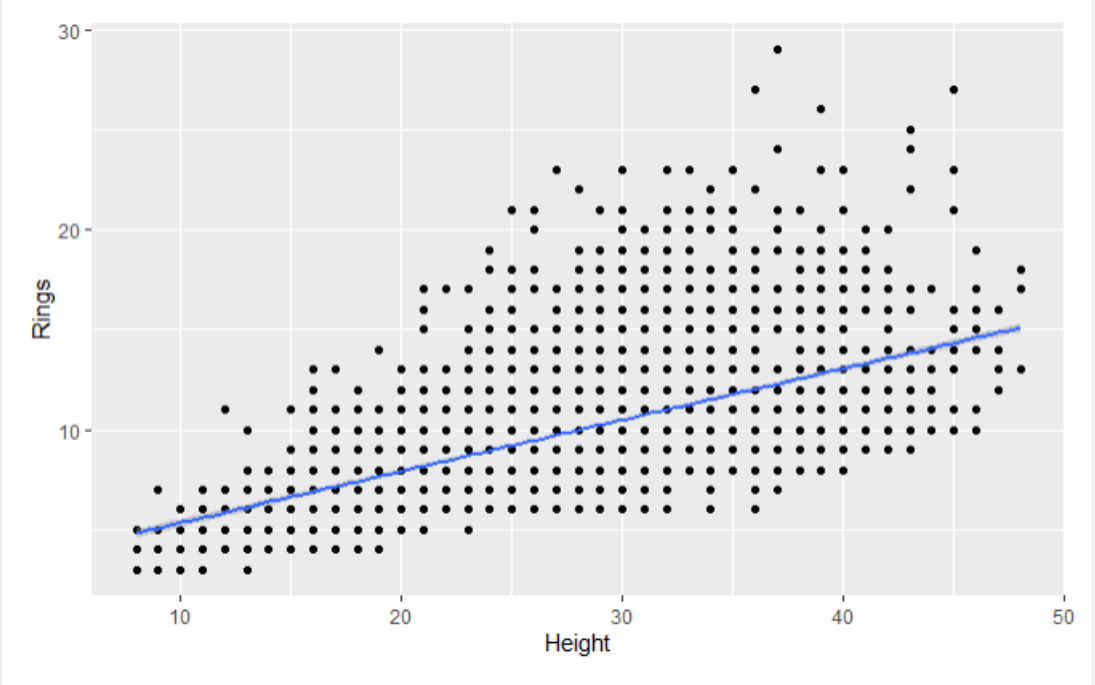
1.5 Question 5

Nous affichons ici, toujours dans le cadre de la régression linéaire du nombre d'anneaux d'un ormeau par sa hauteur, le graphe contenant le nuage de points des différentes mesures présentes dans le dataset, ainsi que la courbe de régression linéaire.

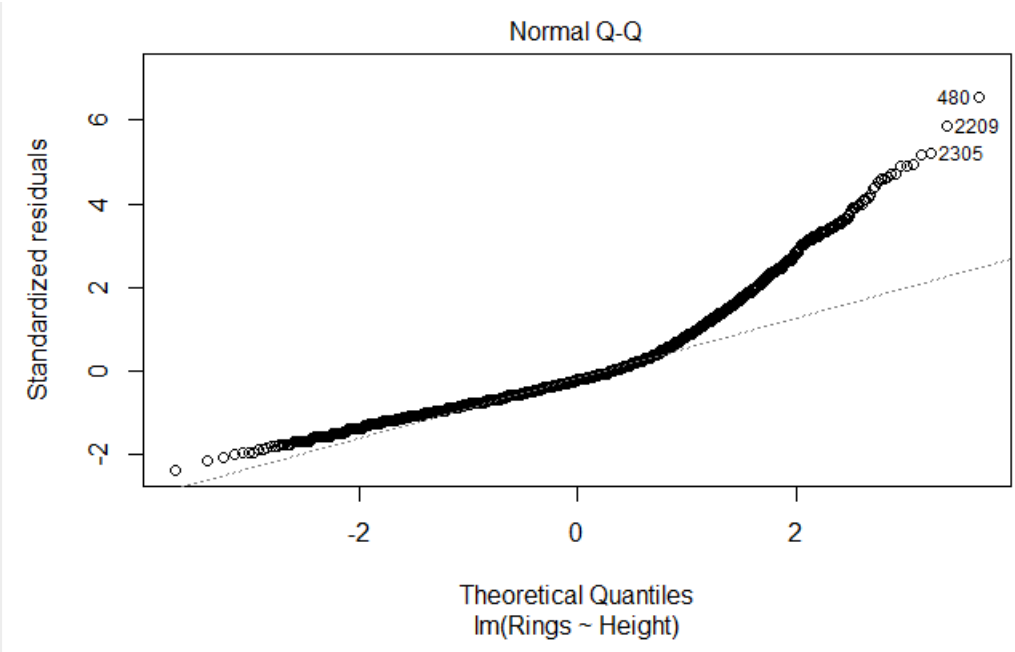
Avec les valeurs aberrantes :



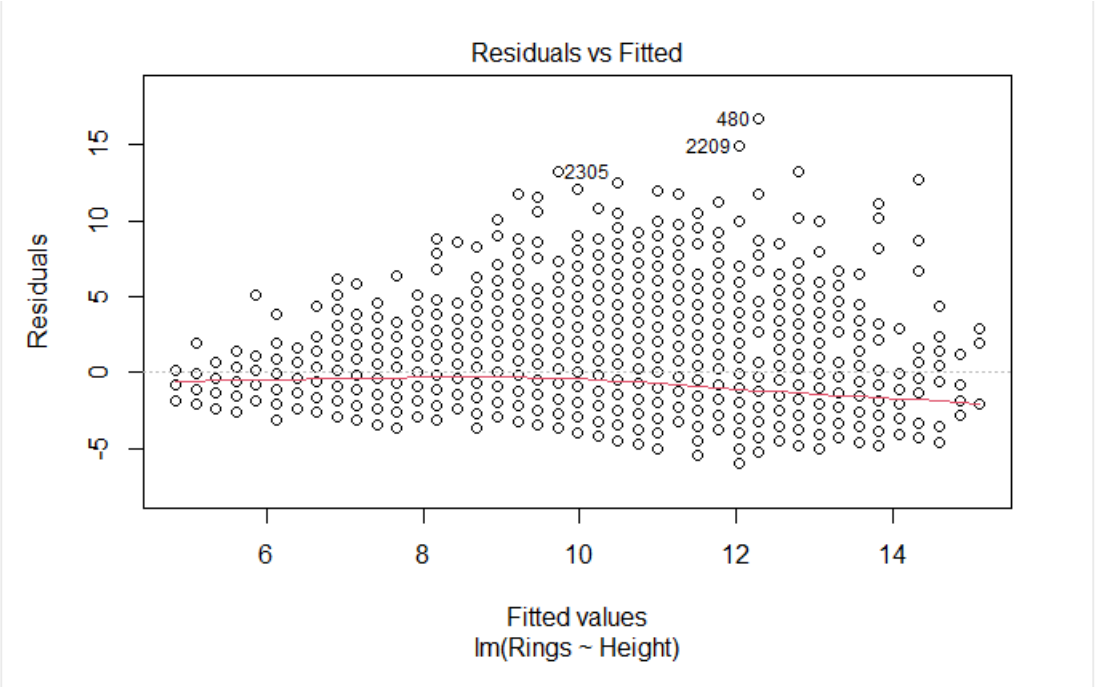
Sans les valeurs aberrantes :



Visuellement, nous remarquons que la régression linéaire n'est pas pertinente. Les points ne se répartissent pas selon une droite. Bien plus, nous voyons même que la tendance générale du nuage de point ne suit pas vraiment la droite. D'ailleurs, les courbes suivantes nous convainquent dans l'idée que la régression linéaire n'est pas pertinente :
Courbe Q-Q :

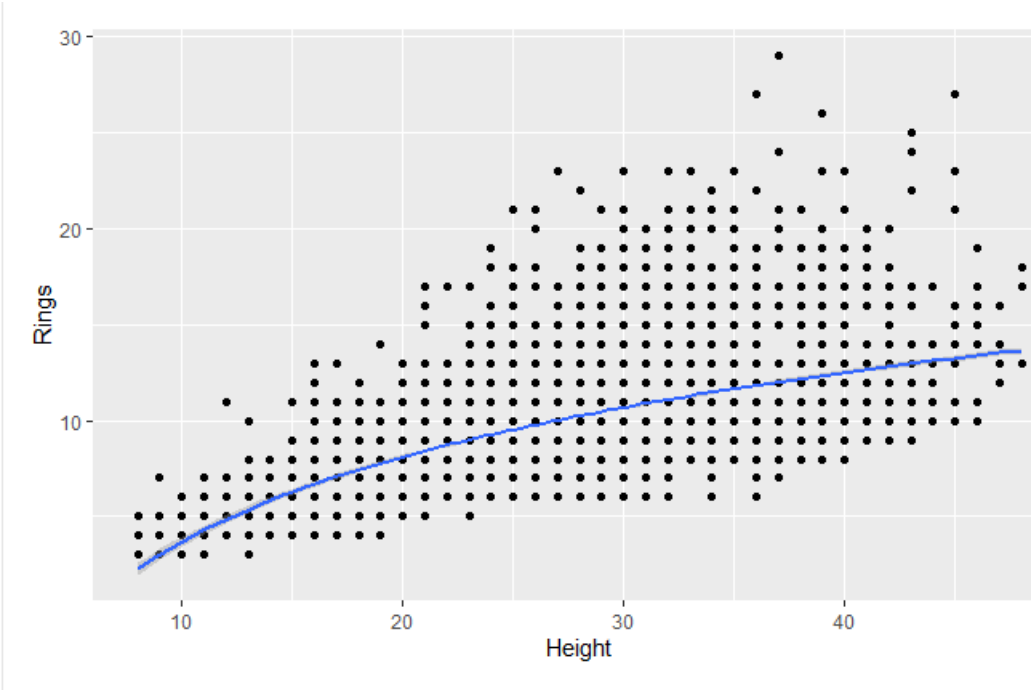


Courbe des résidus :

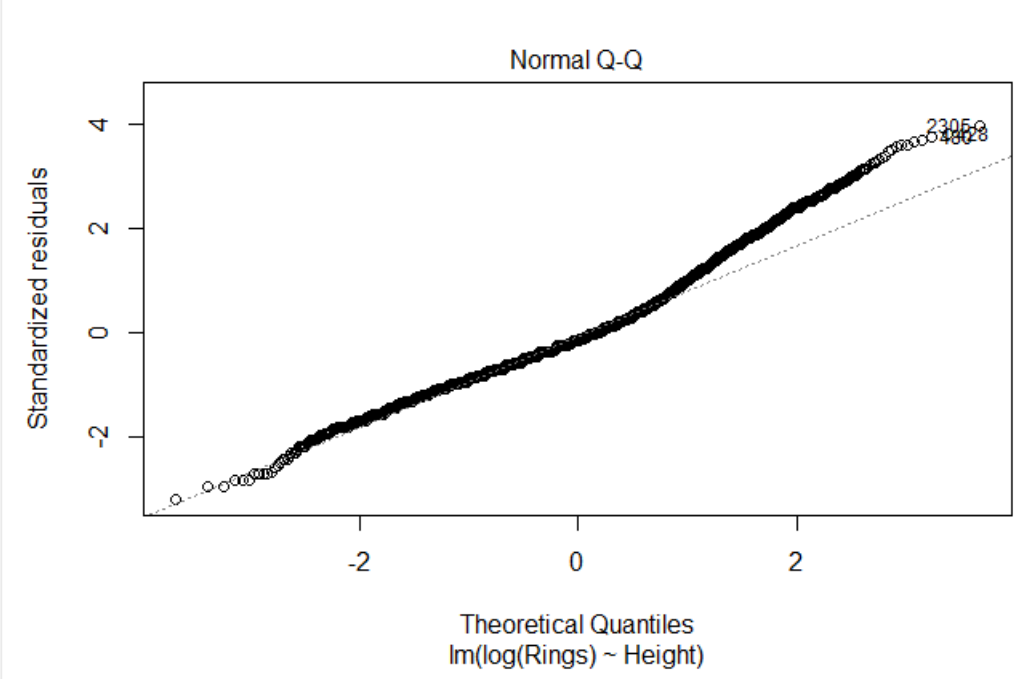


1.6 Question 6 et 7

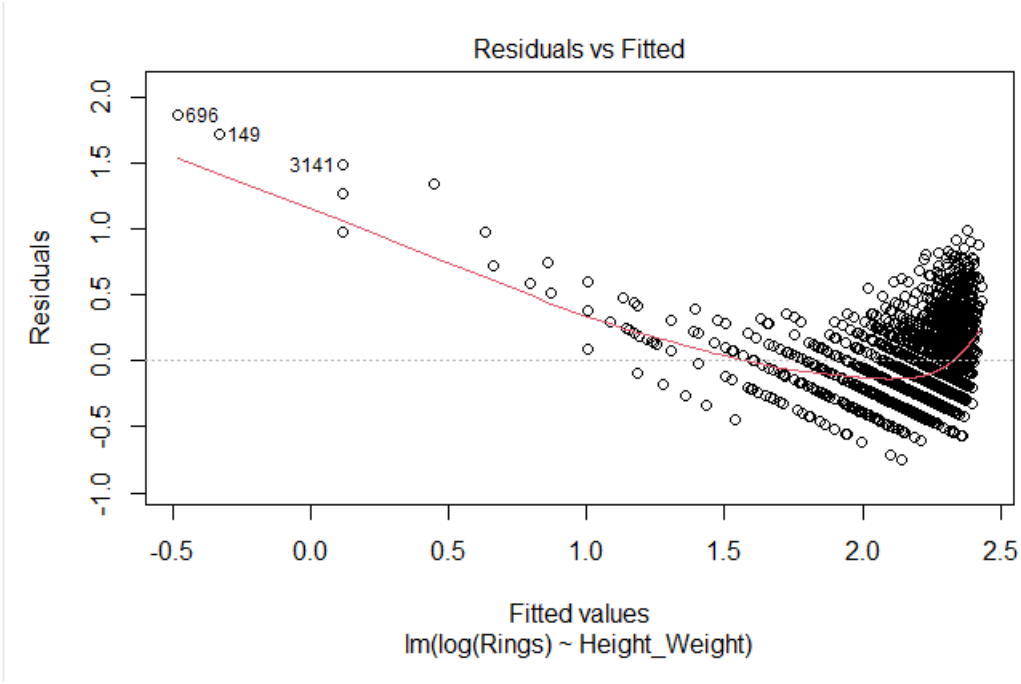
Nous avons remarqué dans les questions précédentes que le modèle linéaire entre l'âge de l'ormeau et sa hauteur n'était pas pertinent. Nous allons donc changer ce modèle. Le nuage de point du nombre d'anneaux en fonction de la hauteur, nous suggère pluôt une régression avec le logarithme du nombre d'anneaux. Voici ce que nous obtenons :



Nous traçons les courbes suivantes également :
Courbe Q-Q :



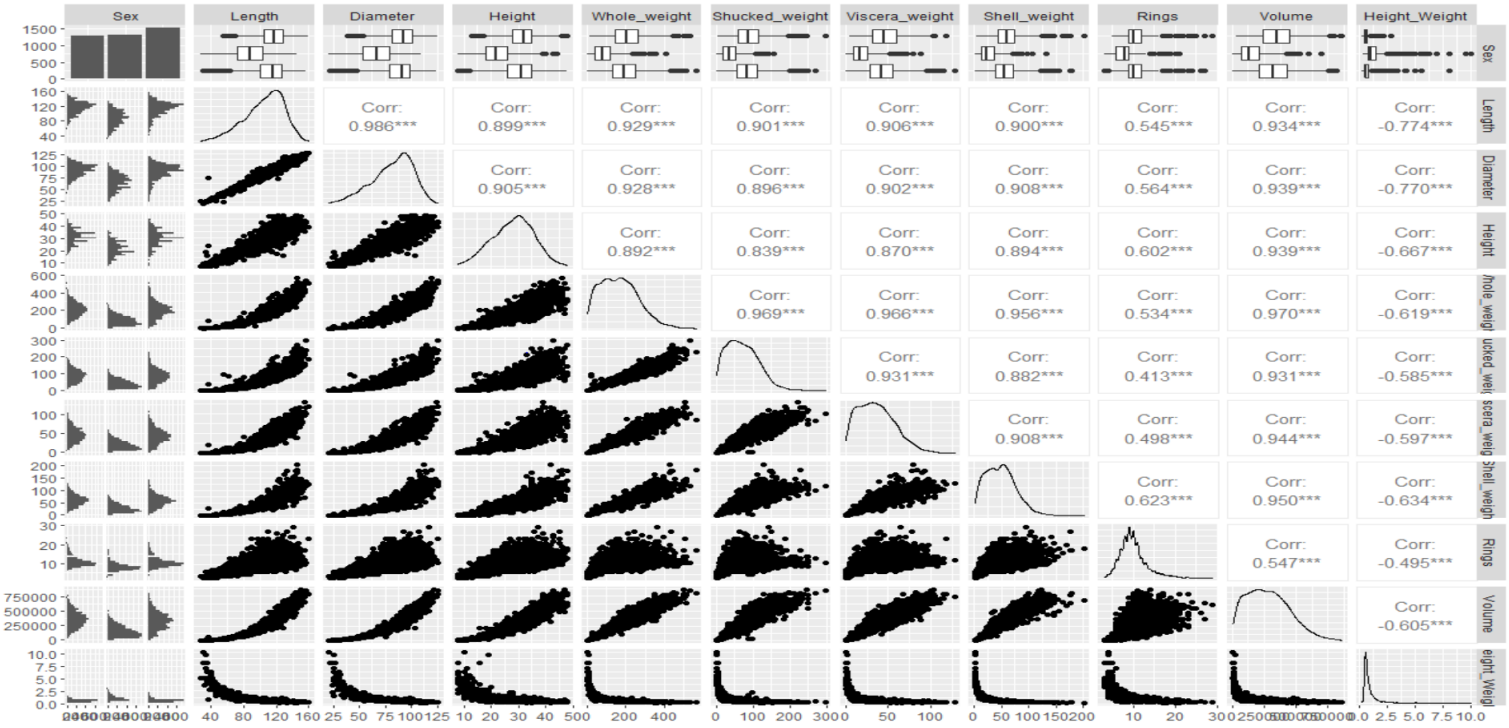
Courbe des résidus :



Une régression par une fonction logarithmique est un peu mieux qu'une régression linéaire. Néanmoins, le résultat ne reste pas vraiment intéressant et il ne nous semble pas pertienent de déterminer une relation (linéaire ou non) entre l'âge des ormeaux et leur hauteur.

1.7 Question 8

Cette fois, nous conservons toutes les variables $x_1, ..., x_8$ et pas seulement x_3. Nous traçons, grâce à GGally :ggpairs, les courbes suivantes :



Par suite, nous cherchons une relation linéaire qui nous permette de prédire, en pratique, l'âge de l'ormeau en fonction des caractéristiques $x_1, ..., x_8$. Nous voulons en fait exprimer l'âge de l'ormeau à partir de caractéristiques que l'on peut mesurer facilement chez l'ormeau (tout en pouvant le garder en vie) : sa taille, son diamètre, sa hauteur et son poids total. Par suite, nous allons en fait effectuer une régression linéaire entre y et x_1, x_2, x_3, x_4. Nous allons donc effectuer la régression suivante :

$$y = \sum_{j=1}^{4} \beta_j x_j + \epsilon$$

Nous allons ensuite séparer le dataset en deux partie :
— 70% pour calculer les coefficients de la combinaison linéaire
— 30% pour tester la prédiction de y par les x_j grâce à la régression linéaire trouvée

2

PARTIE II : ANOVA, ANCOVA

2.1 Question 9

ANOVA - ANCOVA

```
{r}
anova(linear_model, multiple_model)
```

```
Analysis of Variance Table

Model 1: log(Rings) ~ Weight_per_Volume
Model 2: log(Rings) ~ Weight_per_Volume + Viscera_weight + Shell_weight
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3956 357.33
2   3954 202.44  2    154.89 1512.6 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

D’après les résultats de l’analyse Anova le modèle multiple est à préferer au modèle linéaire classique.

2.2 Question 10

Afin de vérifier l’influence de la covariable sexe nous avons effectué une analyse en regardant les différents effets d’un ajout de la considération d’un effet de cette dernière au niveau de l’intercept de chacune des autres variables.

```
Call:
lm(formula = log(Rings) ~ Weight_per_Volume + Weight_per_Volume:Sex +
    Viscera_weight + Viscera_weight:Sex + Shell_weight + Shell_weight:Sex +
    Sex, data = x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.94745 -0.13494 -0.02212  0.10914  0.81288

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.297e+00  5.653e-02  40.634  < 2e-16 ***
Weight_per_Volume       -2.536e+02  8.819e+01  -2.875  0.00406 **
Viscera_weight          -5.264e-03  6.126e-04  -8.592  < 2e-16 ***
Shell_weight             7.876e-03  5.003e-04  15.743  < 2e-16 ***
SexI                    -3.864e-01  7.891e-02  -4.897 1.01e-06 ***
SexM                    -2.054e-01  7.742e-02  -2.653  0.00800 **
Weight_per_Volume:SexI -7.588e+01  1.222e+02  -0.621  0.53458
Weight_per_Volume:SexM  1.372e+02  1.205e+02   1.139  0.25476
SexI:Viscera_weight      4.104e-03  1.548e-03   2.650  0.00807 **
SexM:Viscera_weight     -5.876e-04  8.583e-04  -0.685  0.49362
SexI:Shell_weight        5.477e-03  1.184e-03   4.627 3.82e-06 ***
SexM:Shell_weight        2.216e-03  7.004e-04   3.164  0.00157 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2113 on 3946 degrees of freedom
Multiple R-squared:  0.5228,    Adjusted R-squared:  0.5214
F-statistic: 392.9 on 11 and 3946 DF,  p-value: < 2.2e-16
```

Ainsi la variable sexe semble avoir un effet interessant quand on l'ajoute au modèle mais également lorque l'on considère un effet de celle ci au niveau de l'intercept du poids des Viscères et du poids de la coquille ce qui contredit une des hypothèses clé de l'ANCOVA.Ainsi nous utiliserons ces condiérations pour choisir notre nouveaux modèle qui incluera ces considérations.

	df <dbl>	BIC <dbl>
linear_model	3	1738.8453
multiple_model	5	-493.5474
mult_cov	11	-989.4413

3 rows

Comme nous le voyons avec le BIC ce modèle ne semble pas très interessant. Ainsi nous continuerons notre étude avec notre modèle multiple.

3
PARTIE III : SÉLECTION DE MODÈLES ET PRÉDICTION

3.1 Question 11

Plaçons nous d'abord dans le cas d'une régression linéaire entre une variable y et des variables $x_1, ..., x_p$. Nous considérons que le dataset contienne n éléments $\{(x_{i,1}, ..., x_{i,p}), y_i\}_{1 \leq i \leq n}$ et que l'on choisisse les n_{train} premières lignes pour calculer les coefficients de la régression linéaire et les n_{test} lignes suivantes pour tester la prédiction avec les coefficients de régression linéaire calculés. On construit alors les matrices suivantes :

$$y = (y_1, ..., y_{n_{train}}) \in \mathbf{R}^{n_{train}}$$

$$X = \begin{pmatrix} 1 & x_{1,1} & ... & x_{1,p} \\ 1 & x_{2,1} & ... & x_{2,p} \\ ... & .... & ... & ... \\ ... & .... & ... & ... \\ 1 & x_{n_{train},1} & ... & x_{n_{train},p} \end{pmatrix} \in \mathcal{M}_{n_{train},p+1}(\mathbf{R})$$

On calcule ensuite, grâce à ces données :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

On va ensuite calculer un estimateur de l'erreur ϵ, en nous appyant maintenant sur les données gardées pour tester la prédiction de notre modèle. Pour chaque $i \in [\![1, n_{test}]\!]$, on note :

$$\hat{y_i} = \sum_{j=1}^{8} \beta_j x_{i,j}$$

Cette quantité correspond à la valeur prédite de y_i par les x_j, grâce au modèle linéaire. On considère ensuite l'erreur :

$$\hat{\epsilon_i} = y_i - \hat{y_i}$$

Cette quantité correspond à l'erreur dans la prédiction de y_i par les x_j, grâce au modèle linéaire. On construit enfin le vecteur des erreurs des différentes prédictions :

$$\hat{\epsilon} = \begin{pmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_1 \\ ... \\ ... \\ \hat{\epsilon}_{n_{test}} \end{pmatrix} \in \mathbf{R}^{n_{test}}$$

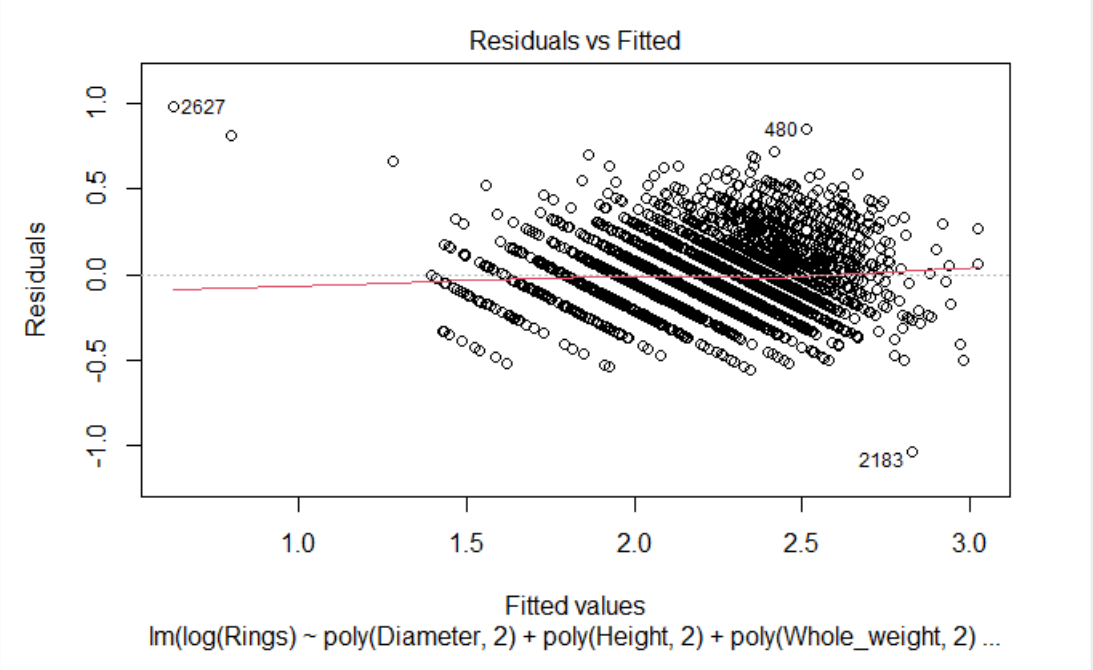
Pour revenir à notre projet, nous effectuons la régression linéaire du nombre d'anneaux de l'ormeau par les variables $x_1, ..., x_8$. Nous avons choisi $n_{train} = 0.7 * n$ et $n_{test} = 0.3 * n$. Après avoir calculé les coefficients de régression linéaire, nous en déduisons le vecteur des erreurs :

On en calcule la norme infinie :

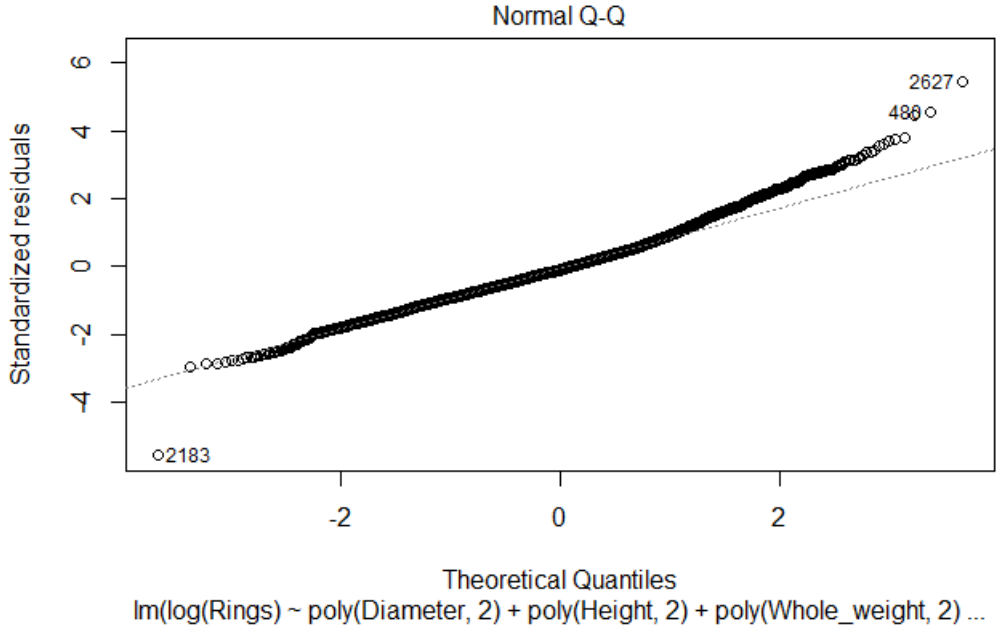
$$\|\hat{\epsilon}\|_\infty = \max_{1 \leq i \leq n_{test}} |\hat{\epsilon}_i|$$

3.2 Question 12

En considérant les différentes variables nous avons conçu principalement 3 modéles polynomiaux. Un premier modèle pour lequel nous avons effetué une regression linéaire du logarithme du nombre d'anneaux en fonction de polynomes de degré de chacune des variables.Ainsi nous avons pu allier la meilleure normalité induite par la focntion logarithme à la plus grande capacité prédictive du modèle induite par le caractère polynomial.



Les résultats sont tout de suite très satisfaisant. La courbe des résidus contre les valeurs ajustées est pratiquement plate, ce qui nous traduit le fait que le modèle est très adapté aux données.



Le QQ-plot nous montre également que la normalité des variables est très satisfaisante.Regardons le résumé du modèle pour voir si l'on peut envisager un modèle plus complexe.

```
    poly(Height, 2) + poly(Whole_weight, 2) + poly(Shucked_weight,
    2) + poly(Viscera_weight, 2) + poly(Shell_weight, 2) + Sex,
    data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0536 -0.1246 -0.0165  0.1029  0.9629

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               2.272444   0.005694 399.064  < 2e-16 ***
poly(Length, 2)1         -3.461329   1.543545  -2.242 0.024987 *
poly(Length, 2)2         -2.167350   0.870491  -2.490 0.012822 *
poly(Diameter, 2)1        3.472693   1.586631   2.189 0.028676 *
poly(Diameter, 2)2       -1.478259   0.899024  -1.644 0.100196
poly(Height, 2)1          4.178040   0.606055   6.894 6.30e-12 ***
poly(Height, 2)2         -1.330835   0.371318  -3.584 0.000342 ***
poly(Whole_weight, 2)1   28.946052   2.506390  11.549  < 2e-16 ***
poly(Whole_weight, 2)2   -4.943745   1.085940  -4.553 5.46e-06 ***
poly(Shucked_weight, 2)1 -25.671253   1.194129 -21.498  < 2e-16 ***
poly(Shucked_weight, 2)2   6.040425   0.595853  10.137  < 2e-16 ***
poly(Viscera_weight, 2)1  -6.237877   0.974861  -6.399 1.75e-10 ***
poly(Viscera_weight, 2)2   1.609957   0.511151   3.150 0.001647 **
poly(Shell_weight, 2)1    10.075755   1.170739   8.606  < 2e-16 ***
poly(Shell_weight, 2)2    -1.566679   0.565089  -2.772 0.005590 **
SexI                      -0.068658   0.009333  -7.356 2.29e-13 ***
SexM                       0.002166   0.007342   0.295 0.767945
```

Au regard du résumé du modèle nous voyons que les coefficients de degré 2 ont un impact significatif sur le modèle.Ainsi nous avons successivement comparé des modèles avec des coefficients de degré 3 et 4 en chacune des variables et en regardant les résumés et les différents

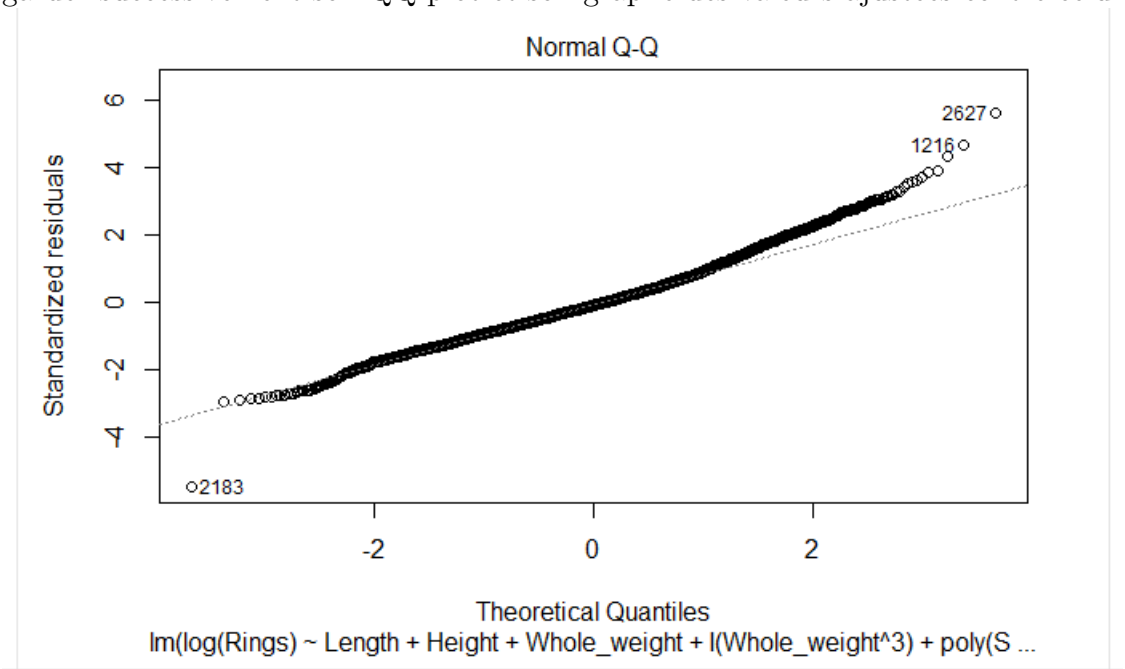
test de Student nous avons finalement opté pour un modèle possédant des coefficients de degré 2 en certaines variables et 3 et 4 pour d'autres.Ainsi nous avons pu obtenir un modèle dont le résumé est le suivant :

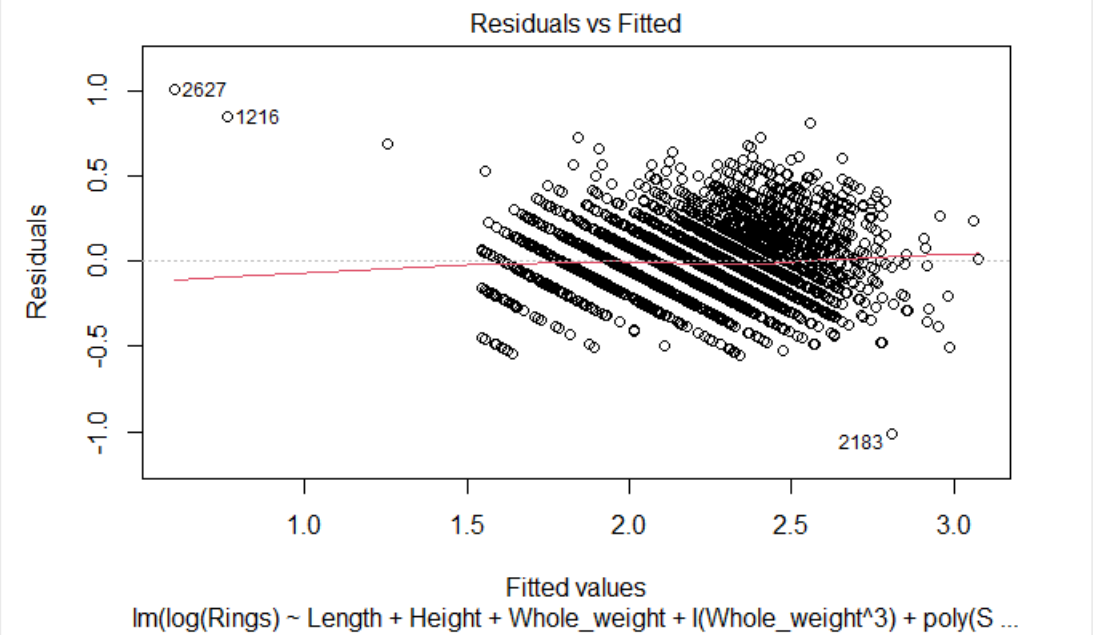
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.01601 -0.12277 -0.01359  0.10386  1.00831

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.571e+00  9.526e-02  16.487  < 2e-16 ***
Length                     -2.897e-03  7.244e-04  -3.999 6.48e-05 ***
Height                      7.192e-03  1.182e-03   6.082 1.30e-09 ***
Whole_weight                6.977e-03  5.773e-04  12.085  < 2e-16 ***
I(Whole_weight^3)          -1.450e-08  2.108e-09  -6.876 7.14e-12 ***
poly(Shucked_weight, 2)1 -2.631e+01  1.177e+00 -22.354  < 2e-16 ***
poly(Shucked_weight, 2)2  6.350e+00  5.770e-01  11.005  < 2e-16 ***
Viscera_weight             -6.670e-03  1.136e-03  -5.871 4.69e-09 ***
I(Viscera_weight^3)         2.906e-07  9.879e-08   2.941  0.00329 **
poly(Shell_weight, 4)1      1.328e+01  1.105e+00  12.021  < 2e-16 ***
poly(Shell_weight, 4)2     -5.828e+00  4.519e-01 -12.897  < 2e-16 ***
poly(Shell_weight, 4)3      3.656e+00  2.530e-01  14.451  < 2e-16 ***
poly(Shell_weight, 4)4     -1.587e+00  2.042e-01  -7.771 9.89e-15 ***
SexI                       -6.119e-02  9.275e-03  -6.597 4.76e-11 ***
SexM                        2.569e-03  7.323e-03   0.351  0.72578
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1879 on 3943 degrees of freedom
Multiple R-squared:  0.6229,    Adjusted R-squared:  0.6216
F-statistic: 465.3 on 14 and 3943 DF,  p-value: < 2.2e-16
```

Comme l'on peut le voir toutes les valeurs utilisées par celui ci sont significative. On peut regarder successivement son QQ-plot et son graphe des valeurs ajustées contre celui des résidus.





Ce modèle possède donc de très interessantes caractéristiques et nous le comparerons à notre modèle précédent que nous avons manuellement conçu.

	df <dbl>	BIC <dbl>
poly_7	16	-1883.8484
multiple_model	5	-493.5474

2 rows

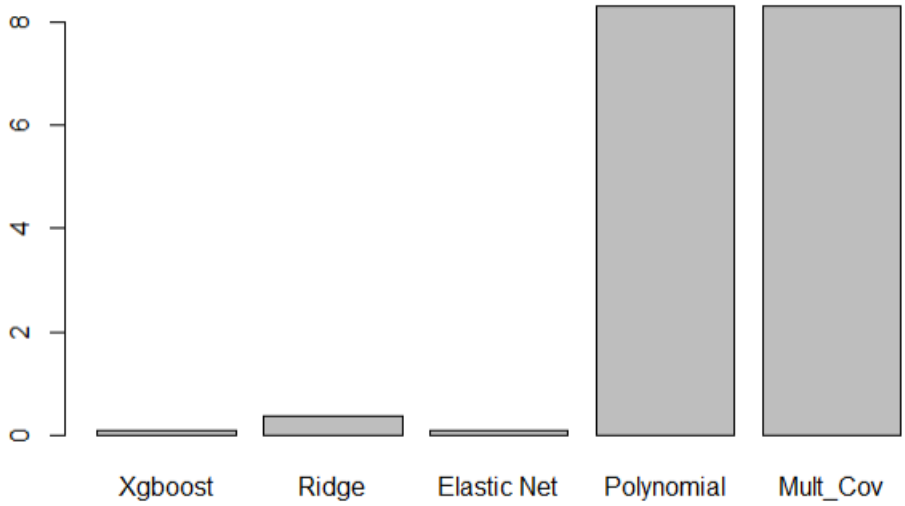
En regardant le critère BIC le modèle polynomial semble moins efficace, nous le garderons néanmoins pour la comparaison finale avec les derniers modèles.

Polynomial	8.3041145
Mult_Cov	8.2994092

En comparant les RMSE on constate également le meme résultat.

3.3 Question 13

Comme modèles nous avons choisi le modèle Ridge , le regresseur XGBOOST et l'Elastic Net. Comme l'on pouvait s'y attendre c'est le modèle XGboost qui possède les meilleures caratéristiques. On obtient en effet en regardant les Root Mean Squared Errord echaque modèle le barplot suivant :



Les RMSE des différents modèles sont :

names <chr>	RMSE <dbl>
Xgboost	0.0802950
Ridge	0.3735094
Elastic Net	0.1005455
Polynomial	8.3041145
Mult_Cov	8.2994092

Ainsi nous voyons que le meilleur choix de modèle est le modèle XGBOOST . Qui l'emporte largement sur les autres.