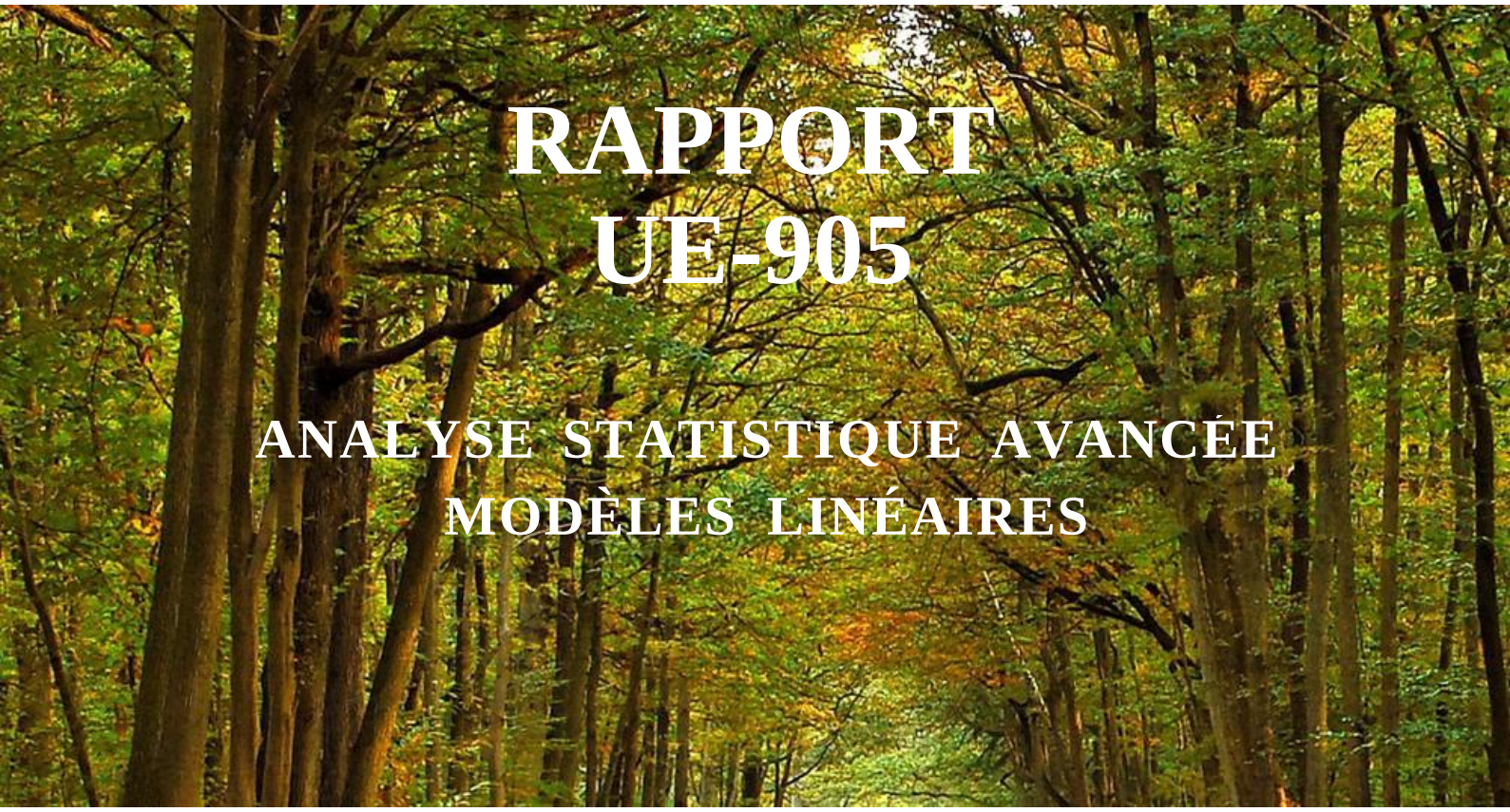


# RAPPORT
# UE-905

## ANALYSE STATISTIQUE AVANCÉE
## MODÈLES LINÉAIRES


**Réalisé par :**

**Abir BEN ABDELGHAFFAR**

**Alban DUMONT**

**Mohamed LAMGHARI**

# Table des Matières

# Liste de figures

# 1. Introduction

Ce projet s'inscrit dans le cadre d'une analyse statistique approfondie d'un jeu de données forestières, en utilisant des approches avancées telles que les modèles linéaires (modèles de régression, modèles mixtes et généralisés), l'ANOVA et les modèles emboités.
L'objectif principal est de comprendre les variations du diamètre des arbres, qui constituent la variable dépendante, en fonction de divers facteurs explicatifs, notamment le relevé, la dernière année de coupe, l'espèce d'arbre et des variables topographiques telles que l'altitude.

En analysant l'effet de chacun de facteurs explicatifs étudiés sur la variable réponse, il devient possible d'estimer et de prédire le diamètre des arbres à l'échelle de la population.

# 2. Méthodologie

## 2.1 Installation de packages nécessaires

Pour effectuer notre analyse ANOVA, nous avons installé les packages suivants pour accéder aux fonctions nécessaires :

- **lme4** : Permet d'ajuster des modèles linéaires mixtes via la fonction « lmer», adaptée à l'analyse de données avec effets fixes et aléatoires.
- **MuMIn** : Utilisé pour calculer les coefficients de détermination ($R^2$ marginal et conditionnel) des modèles mixtes avec la fonction « r.squaredGLMM »
- **car** : Fournit des outils avancés pour l'analyse de régression, notamment les tests ANOVA.
- **ggplot2** : Facilite la création de graphiques clairs et personnalisables pour visualiser et interpréter les résultats.
- **DHARMa** : Sert à évaluer la qualité d'ajustement des modèles linéaires généralisés binomiaux, notamment via des diagnostics de résidus.

## 2.2 Importation des Données sur RStudio

Les données ont été importées dans RStudio en utilisant la fonction **read.csv (),** permettant de travailler avec le fichier "dataProjet_2025.csv".

### 2.3 Filtrage des Données

Pour chacune de ces questions, nous avons travaillé sur un jeu de données filtré à partir de base de données initiale « **dataProjet_2025.csv** ». En effet, pour la première question, le jeu de données a été filtré selon la valeur de champ « recherche_esp_lb_nom_plantae » pour ne conserver que les observations correspondant à l'espèce « **Carpinus betulus L., 1753** ».

## 3. Analyse des Résultats

### 3.1 Question 1 : Modèle ANOVA pour les Charmes

**Objectif :**

L'objectif de cette analyse était d'examiner l'effet des relevés sur le diamètre des arbres de l'espèce ***Carpinus betulus L., 1753*** (charme) en utilisant un simple modèle ANOVA.

**Exploration de jeu de données :**

Avant d'appliquer le modèle ANOVA pour analyser la variation du diamètre (DBH) en fonction de la variable explicative "releve" une exploration préliminaire des données a été effectuée. Cette étape visait à vérifier la qualité et la structure des données, ainsi qu'à identifier les éventuelles tendances ou anomalies. En filtrant les données pour l'espèce **Carpinus betulus L., 1753**, nous avons examiné les distributions des variables et généré une fonction de répartition empirique pour le DBH (Figure 1).

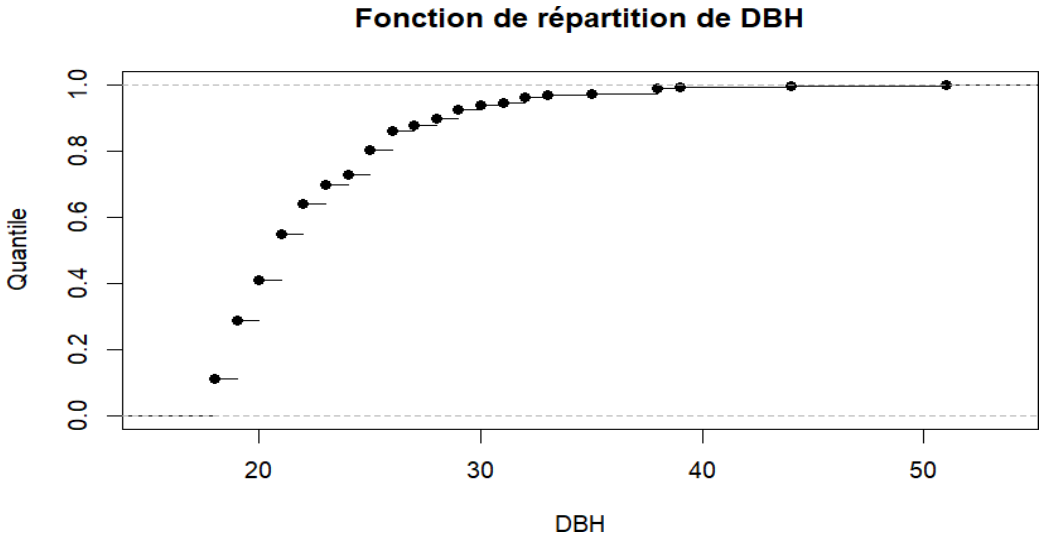



*Figure 1 : Graphique de la fonction de répartition de DBH*

Les résultats montrent que la majorité des arbres de l'échantillon possède un diamètre compris entre **18 cm et 30 cm**, avec une concentration notable autour de 21 cm (médiane). Très peu d'arbres dépassent un diamètre de **50 cm**, ce qui est confirmé par la courbe cumulative atteignant un plateau à ces valeurs. La visualisation met également en évidence la régularité des données, sans anomalies significatives. Enfin, le résumé statistique obtenu à l'aide de la fonction "summary" indique une moyenne de **22,6 cm** pour le DBH, avec des valeurs minimales de **18 cm** et maximales de **51 cm**.

**Application du modèle ANOVA :**

Un modèle ANOVA initial a été ajusté avec la variable « **releve** » comme facteur explicatif et le diamètre DBH comme variable dépendante.

```
modAnova <- lm (DBH~0+releve, data=data_filtred)
```

Une fois que le modèle décrit précédemment est appliqué, une vérification de l'adéquation des erreurs (ou des résidus) aux différentes hypothèses de validité d'un modèle ANOVA a été réalisée.

****Normalité des Résidus**

Les tests graphiques, comprenant un Qqplot et un histogramme (Figures 2 et 3) ont révélé une asymétrie notable des résidus. Les points du Qqplot (Figure 2) s'écartaient de la droite de référence, indiquant une déviation de la normalité. L'histogramme (Figure 3) des résidus a confirmé une asymétrie positive. Les résidus n'étaient donc pas compatibles avec une distribution normale de la loi Gaussienne.

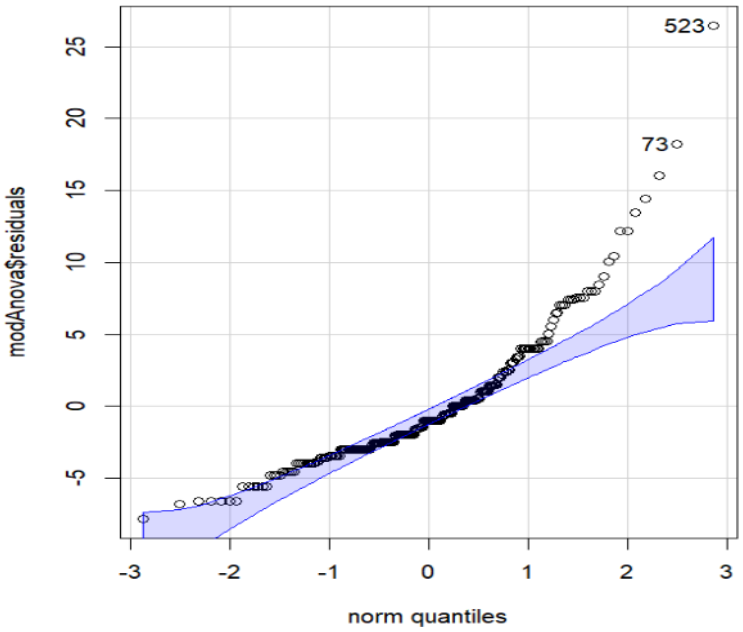



*Figure 2 : Graphique Qqplot témoignant une distribution asymétrique des résidus*

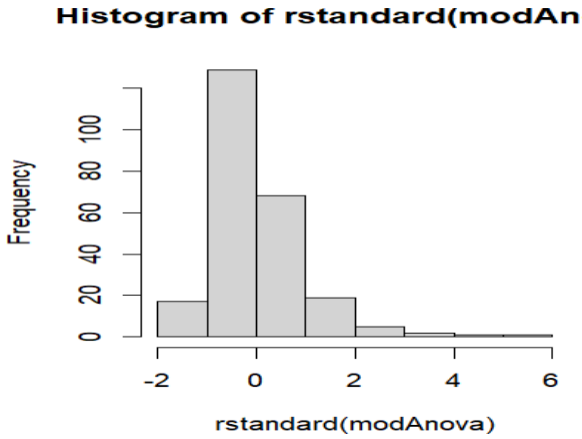



*Figure 3 : Histogramme des résidus standardisés*

****Homoscedasticité des Résidus**

De même, pour l'hypothèse d'homogénéité de variance, un scale-location plot (Figure 04) a révélé une tendance à une dispersion accrue des résidus pour des valeurs ajustées élevées, indiquant une violation de l'hypothèse d'homoscedasticité (hétérogénéité des variances).

Par conséquent, nous avons décidé d'appliquer une transformation Box-Cox, qui s'est révélée efficace pour stabiliser les variances et optimiser l'ajustement du modèle.

****Transformation Box-Cox**

Après l'application d'une transformation Box-Cox dont la valeur de la meilleure puissance lambda égale à -2.764963, nous avons réinitialisé le modèle avec la variable transformé comme l'indique les lignes de code suivantes :

```
modAnova_pT <- powerTransform (modAnova)

modAnova_pT$lambda

modAnova <- lm (DBH**modAnova_pT$lambda~0+releve, data=data_filtred)
```

****Test de normalité des résidus après transformation**

Le nouveau Q-Q plot des résidus du modèle transformé, illustré dans la figure 04, montre une amélioration de la normalité suite à la transformation appliquée.

Les résidus suivent globalement une distribution normale, malgré de légères déviations aux extrémités. L'analyse visuelle des Q-Q plots et de l'histogramme des résidus standardisés valide l'adéquation du modèle ANOVA, bien que certaines valeurs extrêmes méritent une attention particulière.

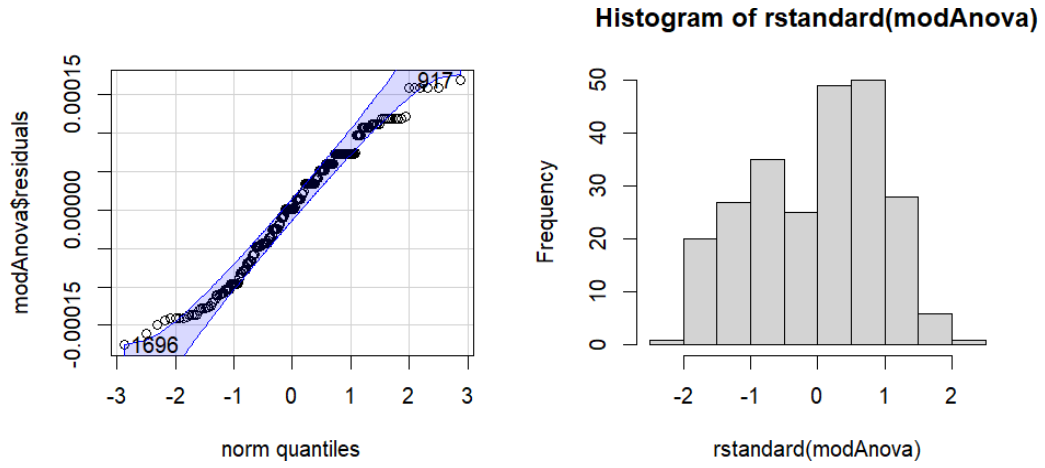



*Figure 4 : Résultats d'un test de normalité des résidus après l'application d'une transformation Box-Cox*

**** Test d'homoscedasticité des résidus après transformation**

Afin d'évaluer l'homoscedasticité des erreurs, on peut se baser sur une visualisation de diagramme de localisation d'échelle (**scale-location plot** ou **SL-plot**).

En effet, les résultats obtenus dans la figure 5, montrent que les résidus semblent être répartis de manière relativement homogène autour de la ligne bleue, bien que quelques points extrêmes soient présents. Les lignes de tendance rouge et bleue ne montrent pas de tendance marquée à augmenter ou à diminuer systématiquement avec les valeurs ajustées, suggérant une variance constante des résidus.

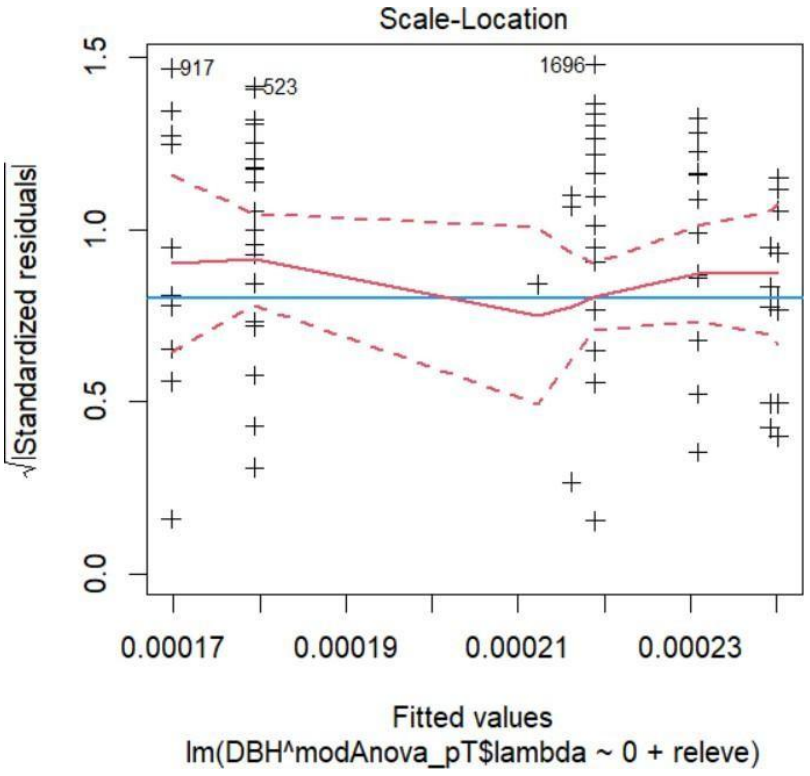



*Figure 5 : Scale location-plot illustrant la répartition des résidus standardisés après une transformation Box- Cox de variable réponse*

Ainsi, l'hypothèse d'homoscedasticité est raisonnablement valide pour ce modèle, bien qu'on observe quelques points extrêmes.

On suppose que notre modèle est bien validé, autrement dit le relevé a un effet significatif sur le diamètre des arbres (DBH) et les résidus sont compatibles avec les hypothèses du modèle linéaire (homoscedasticité et normalité). On passe maintenant à l'analyse des sorties du modèle.

****Analyse de sorties du modèle :**

* Résumé du modèle et récupération des coefficients estimés

```
Call:
lm(formula = DBH ~ 0 + releve, data = data_filtred)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8000 -2.9684 -0.9684  1.5250 26.4386

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
releveBLO_1   24.5614     0.6163  39.856  < 2e-16 ***
releveBLO_12  21.0000     1.0966  19.149  < 2e-16 ***
releveBLO_13  21.6667     2.6862   8.066 3.77e-14 ***
releveBLO_21  21.9684     0.4774  46.022  < 2e-16 ***
releveBLO_24  21.5000     3.2899   6.535 3.93e-10 ***
releveBLO_27  21.4667     0.6936  30.951  < 2e-16 ***
releveBLO_4   25.8000     1.2013  21.477  < 2e-16 ***
releveBLO_9   20.5714     1.7585  11.698  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.653 on 234 degrees of freedom
Multiple R-squared:  0.9608,    Adjusted R-squared:  0.9595
F-statistic: 717.1 on 8 and 234 DF,  p-value: < 2.2e-16

 releveBLO_1 releveBLO_12 releveBLO_13 releveBLO_21 releveBLO_24 releveBLO_27
    24.56140     21.00000     21.66667     21.96842     21.50000     21.46667
releveBLO_4  releveBLO_9
    25.80000     20.57143
```

*Figure 6 : Résumé du modèle et résultats des coefficients estimés pour chaque relevé*

Ces résultats montrent que tous les coefficients des relevés sont bien significatifs ($p < 2e\text{-}16$), indiquant un effet notable de chaque relevé sur le diamètre des arbres (DBH). Cependant, on note une légère différence dans la contribution moyenne de chacun de relevé reflétant des variations possiblement liées à des caractéristiques locales spécifiques à chaque site.

Le modèle présente une bonne qualité d'ajustement, avec un **$R^2$ multiple de 0.9608** et un $R^2$ **ajusté de 0.9595**, expliquant 96 % de la variance totale. De plus, les résidus, sont bien centrés autour de zéro (-0.9684) et la statistique F de Test de Fisher (717.1 ; $p < 2.2e\text{-}16$) confirme bien la significativité globale du modèle.

* Intervalle de confiance :

```
               2.5 %    97.5 %
releveBLO_1  23.34728 25.77553
releveBLO_12 18.83945 23.16055
releveBLO_13 16.37443 26.95890
releveBLO_21 21.02797 22.90888
releveBLO_24 15.01836 27.98164
releveBLO_27 20.10022 22.83312
releveBLO_4  23.43324 28.16676
releveBLO_9  17.10685 24.03601
```

En regardant ce résultat, les intervalles de confiance confirment la significativité des coefficients et mettent en évidence des différences dans les contributions moyennes au DBH.

En effet, Les relevés avec des valeurs plus au moins faibles dans leur intervalle (exemple : **releveBLO_24**) ont un effet moyen plus faible sur la variable dépendante (DBH). À l'inverse, **releveBLO_4** et **releveBLO_1** semblent avoir des contributions plus élevées à **DBH**.

## 3.2 Question 2 : Régression Linéaire pour les Chênes

**Objectif :**

L'objectif de cette analyse était d'examiner l'effet des relevés sur le diamètre des arbres de l'espèce ***Quercus L., 1753*** (chêne) en fonction de cette variable 'lastLog' en utilisant un modèle de régression linéaire.

**Exploration de jeu de données :**

Le diamètre des chênes (DBH) présente une médiane de 35 cm, avec des valeurs allant de 18 cm (minimum) à 107 cm (maximum).

Quant à la variable *lastLog*, elle prend les valeurs suivantes : 1846, 1866, 1876, 1871, 1911, 1976, et 1986.

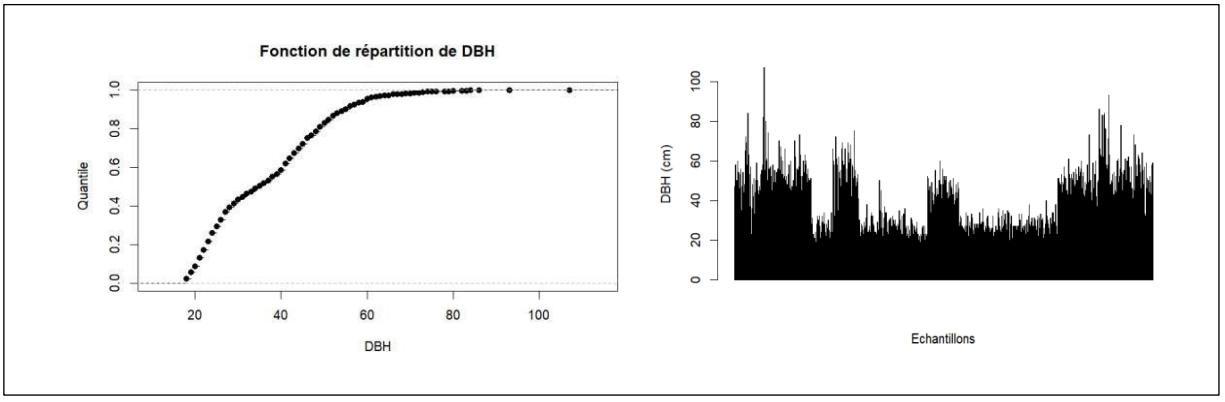



*Figure 7 : Visualisation de la distribution de données à travers la fonction de répartition de DBH et les histogrammes des échantillons*

**Application du modèle régression linéaire :**

En suivant la même démarche que la question précédente, on applique pour cette question un modèle de régression linéaire sur le diamètre en fonction de la variable « lastLog »:

**modReg_Querc <- lm (DBH~lastLog, data=dataIni_Querc)**

***Résultats**

Pour s'assurer de la fiabilité et de l'adéquation du modèle choisi avec notre objectif, nous avons testé la validité des hypothèses de la loi Gaussienne sur les résidus.

***Normalité des Résidus**

Comme avant, on test la normalité de résidus à travers une visualisation de qqplot et des histogrammes des résidus standardisés.

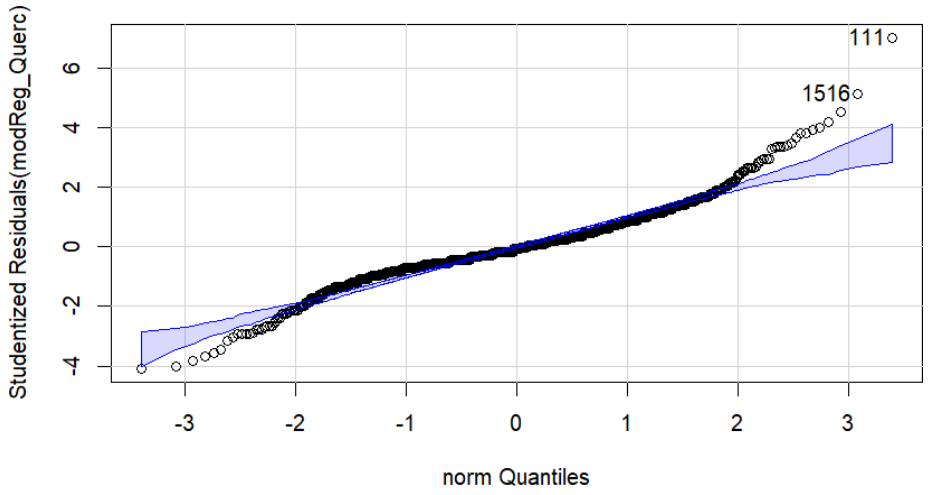



*Figure 8: Qqplot de distribution des résidus en fonction de quantiles normalisés*

Le graphique ci-dessus nous montre un défaut de Kurtosis. La courbe est trop faible au début et trop forte à la fin, cela indique un excès de Kurtosis et donc une distribution forme pointue.

Mais à première vue, le fait de faire la régression linéaire en fonction de lastLog semble pertinent.

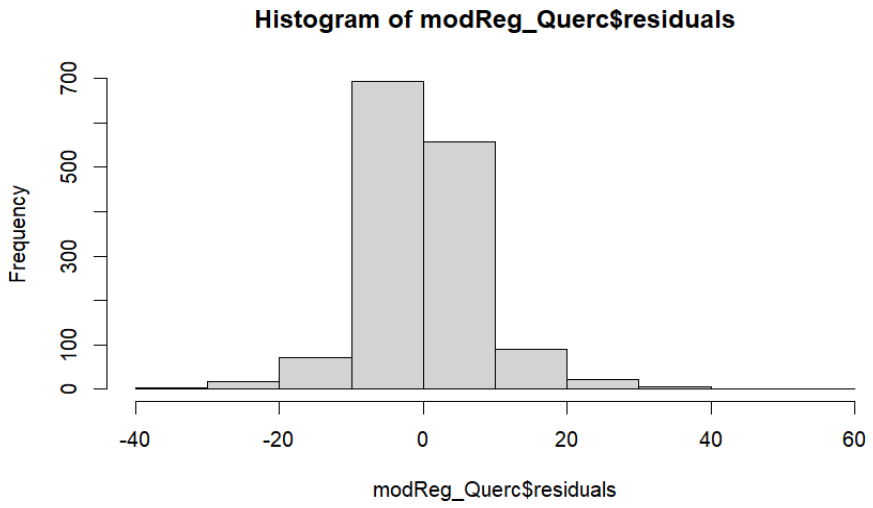



*Figure 9 : Histogrammes de fréquence des résidus standards*

Notre première analyse se confirme avec l'analyse des résidus. On a bien un effet de Kurtosis en forme de pointe. Cependant, ce constat nous permet de continuer l'analyse avec le modèle choisi tant qu'il n'y a pas un problème d'asymétrie et donc la normalité des résidus est supposé valide.

**** Homoscedasticité des Résidus**

En regardant le résultat obtenu dans le scale-location ci-dessous, les points ne sont pas répartis de manière totalement uniforme autour de la ligne rouge. On observe une variation plus faible pour les petites valeurs ajustées et une plus grande dispersion pour les valeurs ajustées élevées.

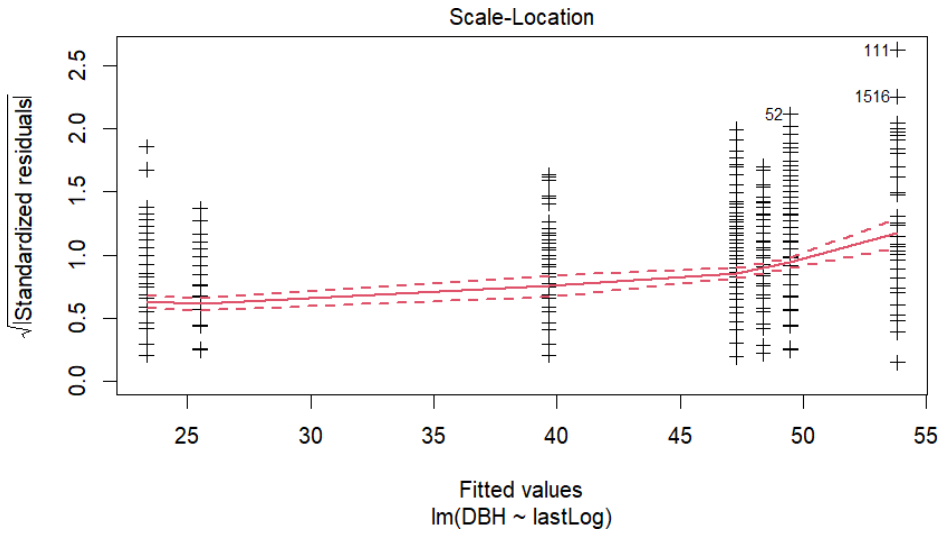



*Figure 10 : Scale-location des résidus standards en fonction des valeurs ajustées*

Cela suggère une possible violation de l'hypothèse d'homogénéité des variances. Donc, une transformation de données est nécessaire pour ajuster le modèle.

Nous avons choisi d'appliquer une transformation logarithmique sur la variable réponse puis nous avons vérifié de nouveau la validité de l'hypothèse d'homogénéité des variances.

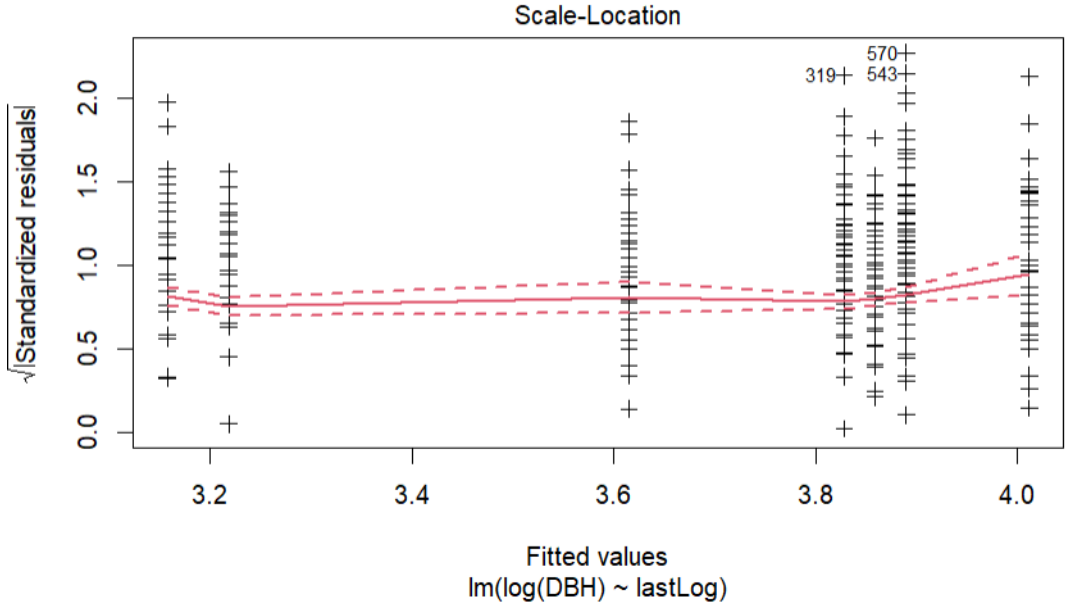



*Figure 11 : Scale-location plot après transformation logarithmique de variable réponse DBH*

Suite à la transformation appliquée, on note une amélioration de la distribution des résidus autour de la ligne rouge. En effet, la dispersion pour les petites valeurs ajustées devient beaucoup plus grande et donc une distribution plus homogène des variances pour les différentes valeurs ajustées.

Et par conséquent, suite à cet ajustement, on peut accepter l'hypothèse d'homoscédasité des variances.

****Indépendance des erreurs**

Le test d'indépendance des erreurs a montré, d'après la figure 12, une distribution aléatoire des erreurs sans avoir des tendances systématiques. Donc, sur la base de cette observation, on peut accepter l'hypothèse d'indépendance des erreurs.

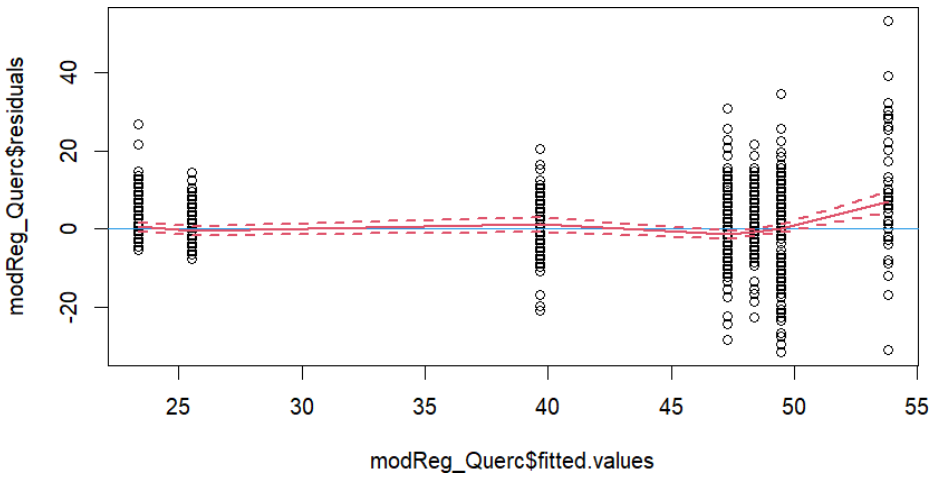



*Figure 12 : Résultats de test d'indépendance des résidus : Scale-location plot des résidus en fonction des valeurs ajustées*

Enfin, après avoir vérifié l'ensemble des hypothèses de validité du modèle linéaire, nous avons visualisé, à l'aide du graphique ci-dessous, l'effet de la variable « lastlog » sur la variable

réponse « DBH ». Cette étape vise à s'assurer de la pertinence de cette variable dans l'explication de la variabilité du diamètre (DBH).

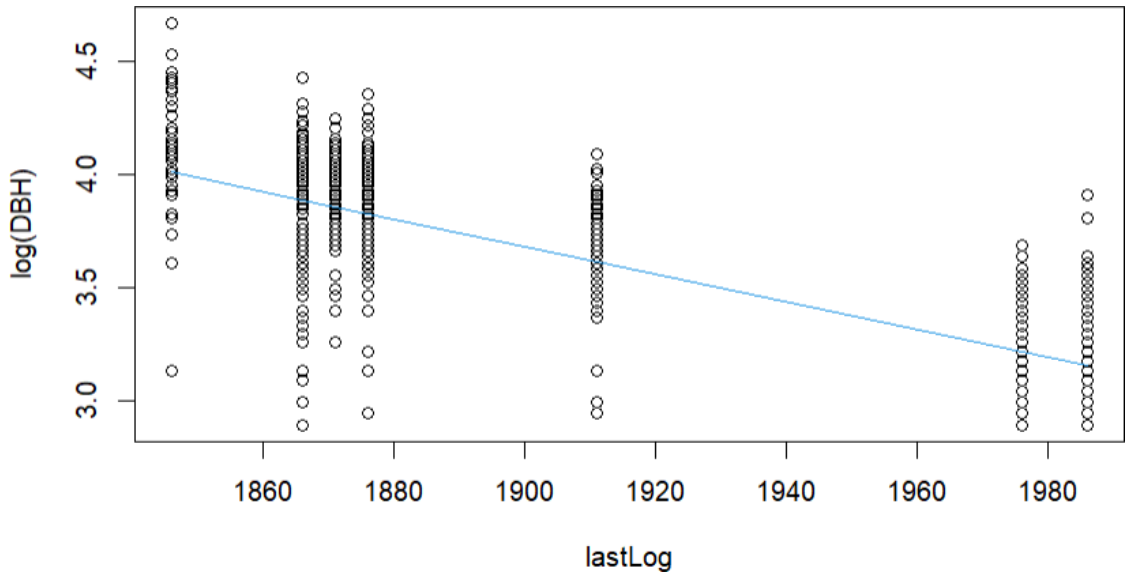



*Figure 13 : Scatter plot de l'évolution de log(DBH) en fonction de la variable explicative lastLog*

Le graphique obtenu montre une tendance à la diminution de log(DBH) au fil des années, avec une variabilité plus élevée dans les années antérieures et une concentration plus forte des données dans les années récentes. Les points sont globalement bien dispersés autour de la ligne bleue, ce qui suggère que la relation entre la dernière année de coupe et le diamètre des arbres est cohérente.

****Analyse des sorties**

```
Call:
lm(formula = log(DBH) ~ lastLog, data = dataIni_Querc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99916 -0.11494  0.00013  0.12272  0.75473

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.528e+01  1.824e-01   83.77   <2e-16 ***
lastLog     -6.102e-03  9.471e-05  -64.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1932 on 1460 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.7398,    Adjusted R-squared:  0.7396
F-statistic:  4151 on 1 and 1460 DF,  p-value: < 2.2e-16
```

*Figure 14 : Résumé du modèle et résultats des coefficients estimés*

A partir de résumé du modèle obtenu (Figure 14), on note que :

- Les résidus sont relativement symétriques autour de la médiane proche de zéro, ce qui indique un bon ajustement du modèle.
- Le coefficient constant (intercept) est estimé à 15.28, indiquant que, théoriquement, en l'absence de dernière année de coupe, le logarithme du diamètre des arbres (DBH) serait de 15.28. Le coefficient pour la variable "lastLog" est négatif (-0.0061), ce qui révèle une relation inverse entre la dernière année de coupe et le diamètre des arbres : les arbres coupés plus récemment ont un diamètre plus petit. La très haute significativité des coefficients (p-value < 2e-16) indique une association significative entre les variables et le DBH.
- Les valeurs de 0.7398 et 0.7396 pour les coefficients de détermination multiple et ajusté montrent que le modèle explique environ 74 % de la variance totale de DBH, ce qui est un bon ajustement.

Et finalement, le test de Fisher (F-Statistic = 4151, avec une p-value très faible) indique que le modèle ajusté est globalement significatif après l'application d'une transformation logarithmique à la variable « DBH ». Cependant, même après cette transformation de la variable réponse, nous observons un déséquilibre dans la distribution autour de la droite de régression aux extrémités de l'axe des ordonnées, notamment pour les dates antérieures à 1860 et postérieures à 1980. Ce phénomène pourrait refléter des évolutions dans les pratiques de coupe ou des variations des conditions de croissance des arbres au cours de ces périodes.

## 3.3 Question 3 : Modèle ANOVA avec Agrégation par Triplets

**** Objectif :**

L'objectif de cette question est de définir des sous-populations basées sur des triplets de relevés, plutôt que d'analyser chaque relevé individuellement, et d'appliquer un modèle ANOVA correspondant à ce niveau d'agrégation. L'objectif final est de déterminer si les relevés peuvent être regroupés en triplets en fonction de leur proximité géographique, afin d'étudier leur effet sur la variabilité du diamètre, plutôt que de considérer l'impact de chaque relevé séparément.

**** Visualisation graphique de triangulation des relèves :**

Après avoir chargé et filtré le jeu de données pour ne garder que les données relatives aux arbres de l'espèce Chêne. Nous avons générée une première visualisation de l'agrégation Triangulaire des relevés (Figure 15).

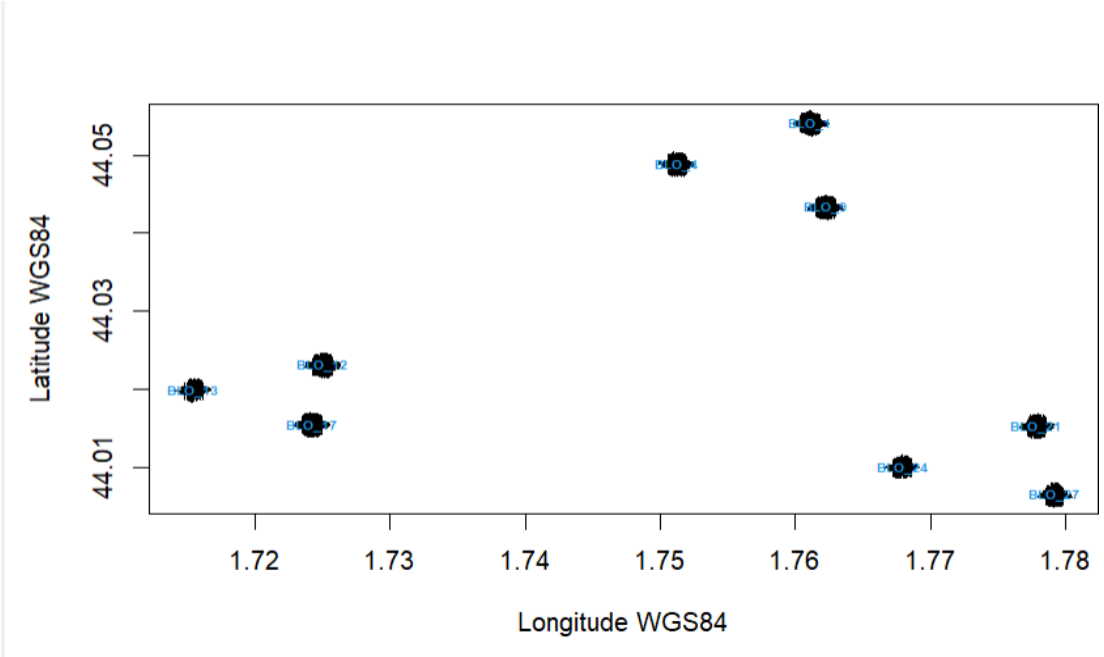



*Figure 15 : Visualisation graphique de la répartition triangulaire des relevés*

Suite à cette première visualisation, nous avons proposé d'ajouter dans notre jeu de donnée filtré un attribut qui indique l'appartenance de chaque relevé à une sous-population (ou un cluster). Pour le faire, nous avons calculé une matrice de distance entre les différents relevés puis effectué un regroupement hiérarchique en se basant sur les distances calculées.

Le graphique présenté ci-dessous montre le résultat de clustering en associant à chaque triplet de relevés un code couleur qui lui spécifié.

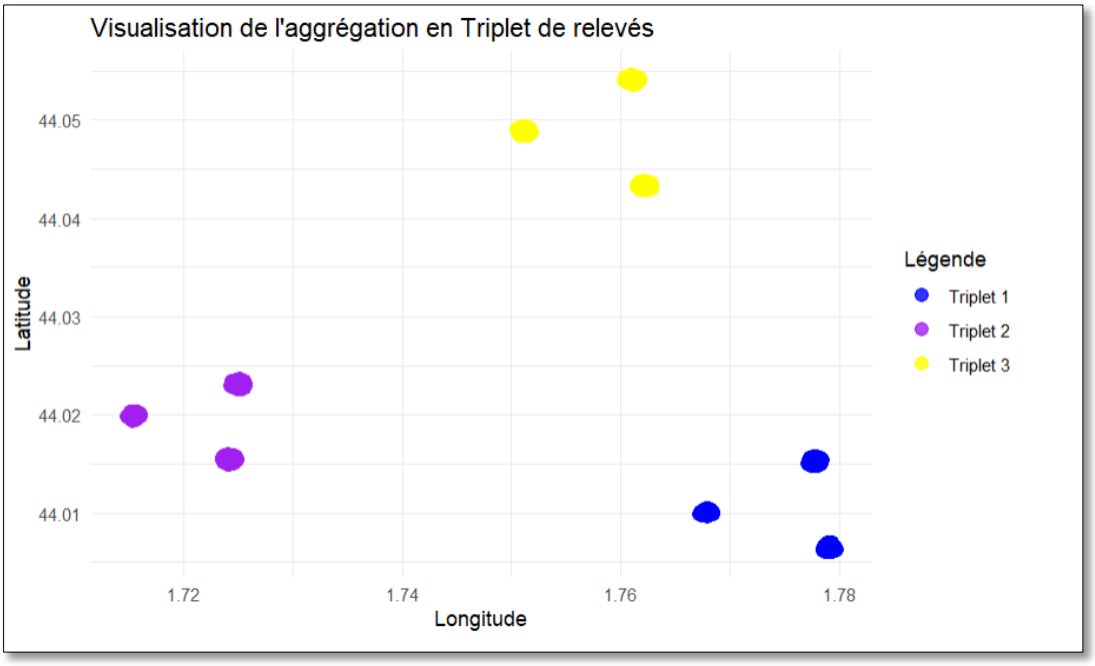



*Figure 16 : Répartition en cluster de trois relevés*

**** Application de L'ANOVA**

Pour répondre à l'objectif précédemment défini, nous avons élaboré un modèle linéaire d'ANOVA dans lequel la variable récemment ajoutée, « cluster », est utilisée comme variable explicative, tandis que le diamètre des arbres « DBH » est pris comme variable dépendante.

**modAnova ← lm (data_Chêne$DBH ~ 0 + cluster, data = data_Chêne)**

**** Diagnostic de validité de modèle**

Comme précédemment, avant d'analyser les résultats du modèle, nous vérifions si les résidus respectent les hypothèses fondamentales d'un modèle linéaire : normalité, homoscedasticité et indépendance des erreurs.

****Test de Normalité :**

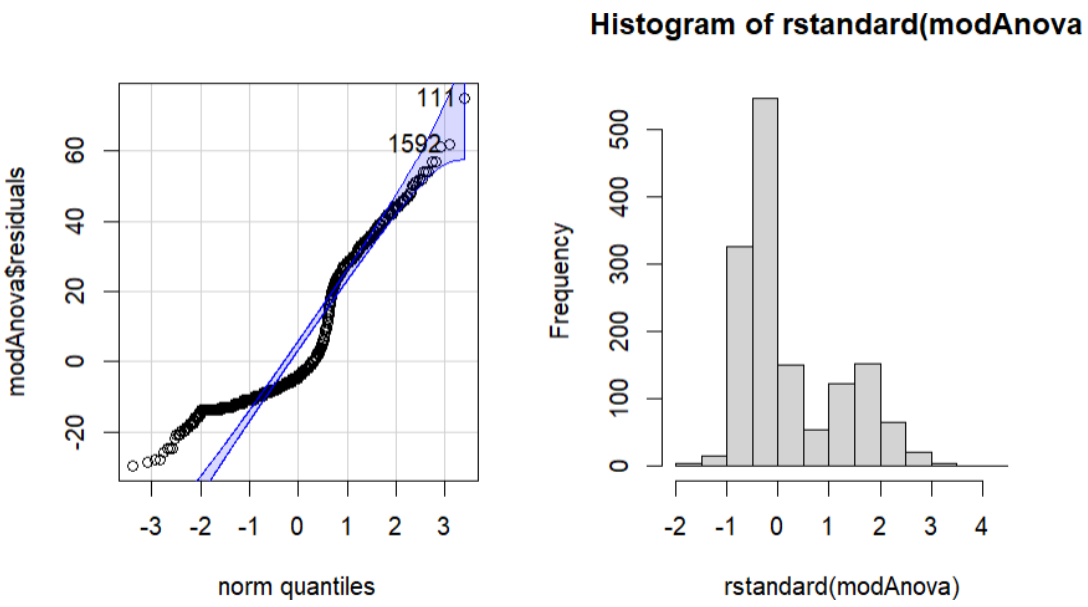



*Figure 17 : Résultats de test de normalité des résidus : qqplot et les histogrammes de résidus standards*

L'analyse des graphiques ci-dessus montre que les points sur le qqplot s'alignent principalement le long de la ligne diagonale, bien qu'ils présentent des écarts notables aux extrémités et au centre. Cela suggère que, dans l'ensemble, les résidus suivent une distribution normale, avec des déviations aux valeurs extrêmes et quelques anomalies au centre. L'histogramme indique également une distribution légèrement asymétrique, présentant une forme qui n'est pas parfaitement en cloche. Ces observations suggèrent une légère déviation de la normalité des résidus.

Afin de corriger cette asymétrie, une transformation de la variable dépendante pourrait être envisagée.

**** Transformation de données**

Avant d'appliquer une transformation, nous avons vérifié la présence de valeurs aberrantes et identifié 29 échantillons sur 1 467. Ces données ont été retirées du jeu de données, puis le modèle ANOVA a été recalculé après application d'une transformation logarithmique à la variable dépendante « DBH ». Toutefois, cette transformation, bien qu'effectuée sur les données filtrées, n'a pas résolu le problème, comme le révèlent les résultats du test de normalité (Figure 18).

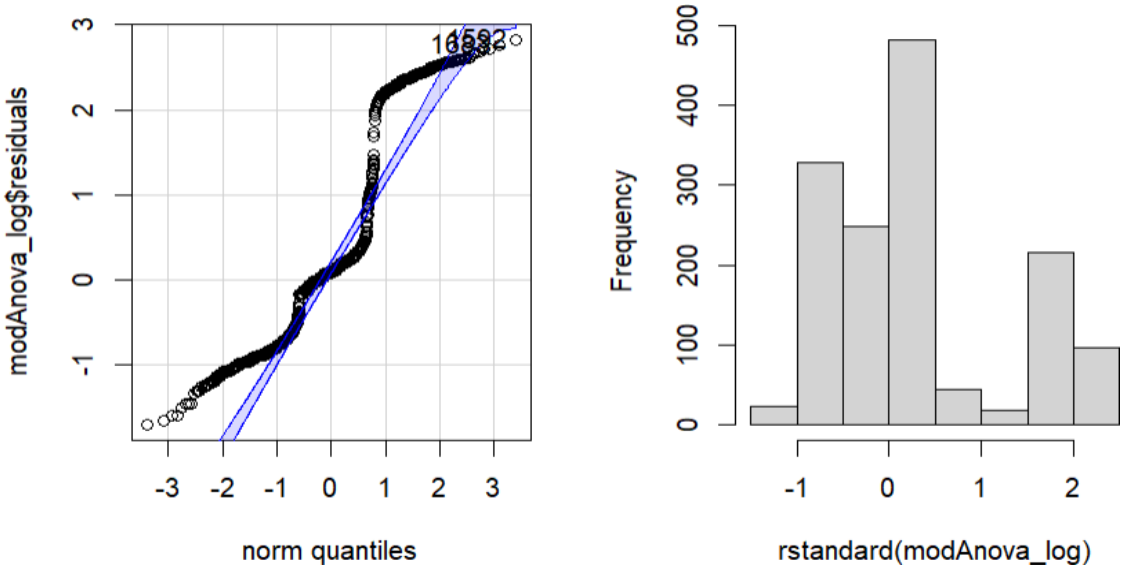



*Figure 18 : Résultats de Test de normalité après l'application d'une transformation logarithmique*

Nous avons testé par la suite une deuxième transformation de type Box-Cox avec une valeur de lambda égale à 1.349468 en espérant qu'elle peut résoudre ce problème d'asymétrie.

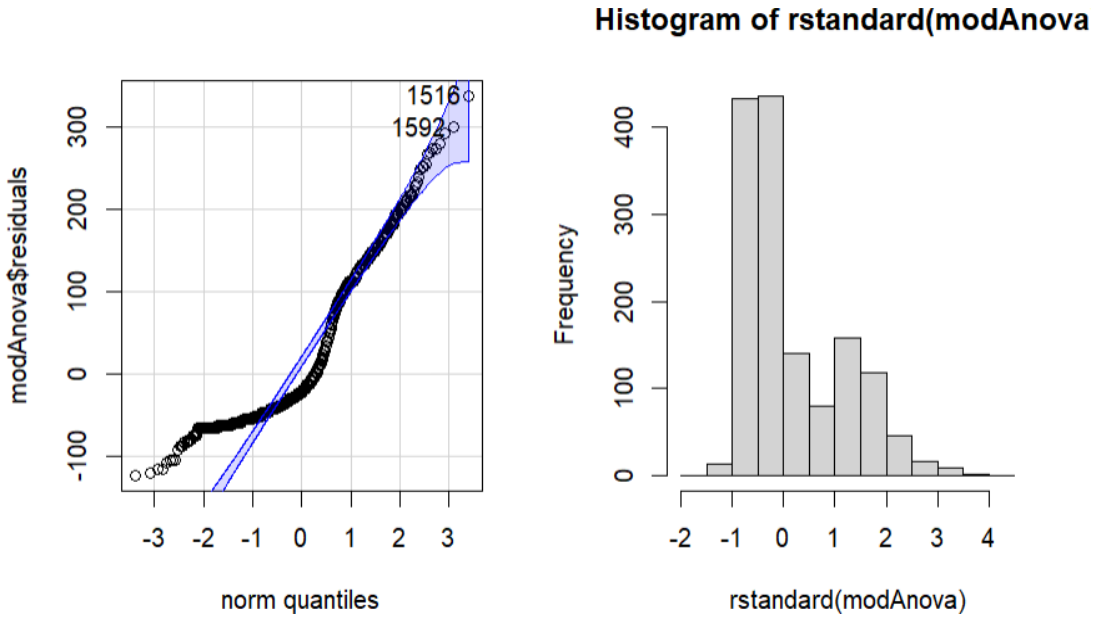


*Figure 19 : Résultats de Test de normalité des résidus après l'application d'une transformation Box-Cox*

Les résultats montrent que, comme la valeur de la puissance lambda est proche de 1, la transformation de Box-Cox appliquée n'apporte aucun avantage et ne parvient pas à ajuster la distribution des résidus à une distribution normale. Par conséquent, la solution proposée consiste à tester un modèle emboîté afin de déterminer si les sous-populations peuvent effectivement être regroupées au sein d'un même cluster.

****Application d'un modèle emboité :**

L'application de ce modèle a pour objectif de comparer le modèle actuel avec un autre modèle dans lequel la variable « releve » est ajoutée à la variable « cluster » en tant que facteur explicatif.

**** Hypothèses testées :**

- **$H_0$ (hypothèse nulle)** : Le facteur "releve" n'apporte pas d'information supplémentaire pour expliquer la variabilité de "DBH".
- **$H_1$ (hypothèse alternative)** : Le facteur "releve" apporte une information supplémentaire significative.

L'analyse comparative de deux modèles nous a donné les résultats suivants :

```
Analysis of Variance Table

Model 1: data_Chêne$DBH ~ 0 + cluster
Model 2: data_Chêne$DBH ~ cluster + releve
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1460 489810
2   1452  81554  8    408256 908.58 < 2.2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Étant donné que la p-valeur est très inférieure à 0,05 (< 2.2e-16), nous rejetons l'hypothèse nulle ($H_0$). Cela indique que le facteur "releve" contribue de manière significative à expliquer la variabilité de "DBH" au sein des clusters.

En d’autres termes, les relevés individuels à l'intérieur des clusters (triplets) représentent une source importante de variation dans les données. Par conséquent, pour mieux expliquer la variabilité du diamètre des arbres, il est préférable de construire un modèle ANOVA en intégrant la variable "releve" comme variable explicative, considérée individuellement, plutôt que regroupée sous forme de clusters.

## 3.4 Question 4 : Modèle Mixte pour Étudier l'Effet de lastLog

**Objectif :**

Pour cette question, on vise à appliquer un modèle linéaire mixte dans l’objectif d’évaluer à quelles mesure la variable dernière année de coupe « lastlog » influence l’effet de relevé sur la variation de diamètre des arbres.

**Modèle linéaire mixte :**

Pour répondre à cet objectif, nous avons travaillé sur le même jeu de donnée utilisé dans la question précédente en faisant un filtre pour ne garder que les données relatives à l’espèce Chêne.

Le modèle que nous avons appliqué est un modèle linéaire mixte dont la variable réponse est le diamètre des arbres, « lastlog » représente l’effet fixe et la variable « releve » représente l’effet aléatoire.

**mod_mixed <- lmer(DBH ~ lastLog + (1 | releve), data = data_chene_clean)**

**Diagnostic des hypothèses :**

**** Diagnostics sur les erreurs résiduelles**

*****Normalité des résidus**

Les résultats de test de normalité, nous a montré une distribution des résidus semble relativement symétrique et en cloche, ce qui est cohérent avec une distribution normale. Toutefois, il peut y avoir des écarts subtils qu'on les voit surtout sur le graphique Qqplot (Figure 20). Donc, globalement l'hypothèse de normalité on peut le considérer comme valide malgré la présence des quelques valeurs aberrantes sur les extrémités.

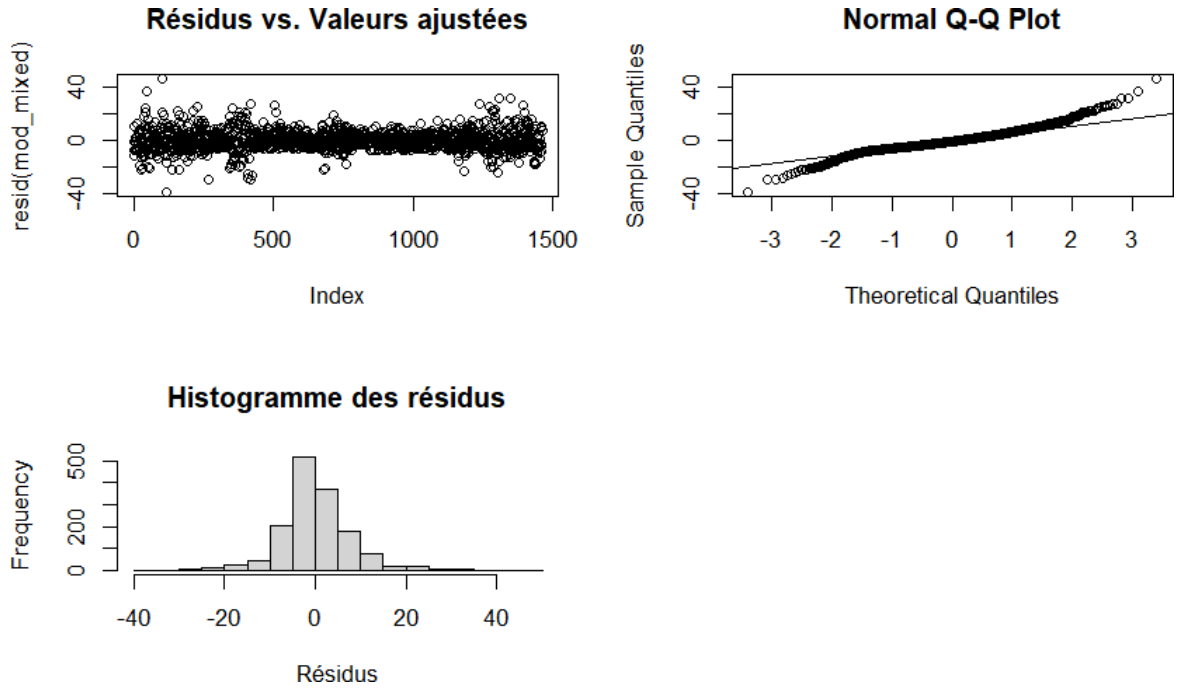



*Figure 20 : Résultats de test de normalité_modèle linéaire mixte*

*****Homogénéité des résidus**

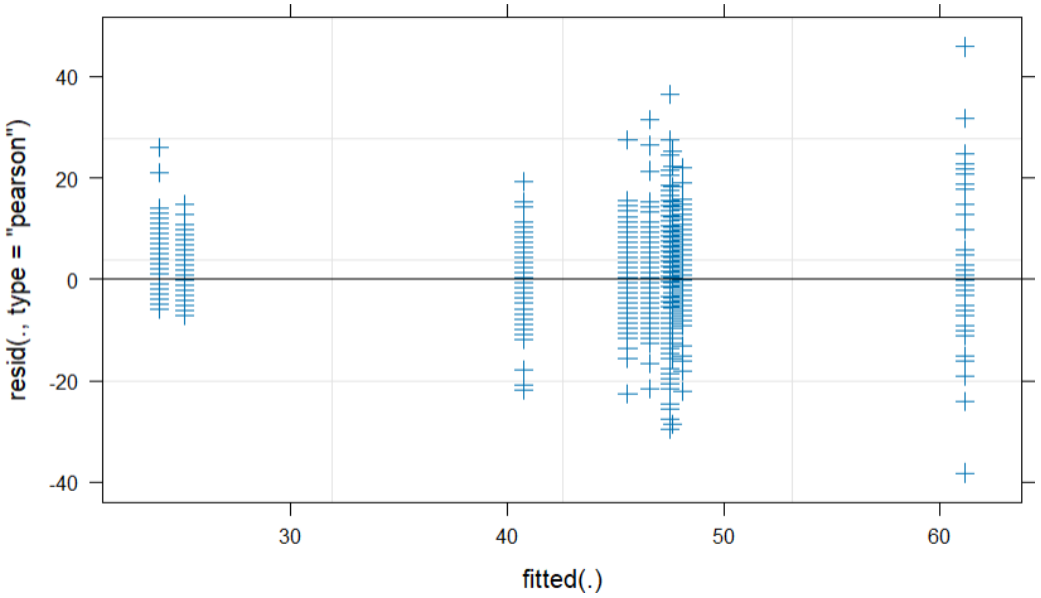



*Figure 21 : Résultats de test d'homogénéité de variances_ modèle linéaire mixte*

En analysant le scale-location plot présenté ci-dessus, on peut noter une distribution symétrique et homogène départ et d'autre de l'axe zéro pour certaines valeurs ajustées (40, entre 40 et 50 et au niveau de la valeur 60). Cependant pour les valeurs ajustées moins de 30, l'hypothèse d'homogénéité de la variance n'est plus valide. Donc, une interprétation visuelle de ce graphique ne nous permet pas de s'assurer de la validité de l'hypothèse. Donc, on propose d'appliquer un test statistique comme le Test de Leuven.

*****Test de Leuven**

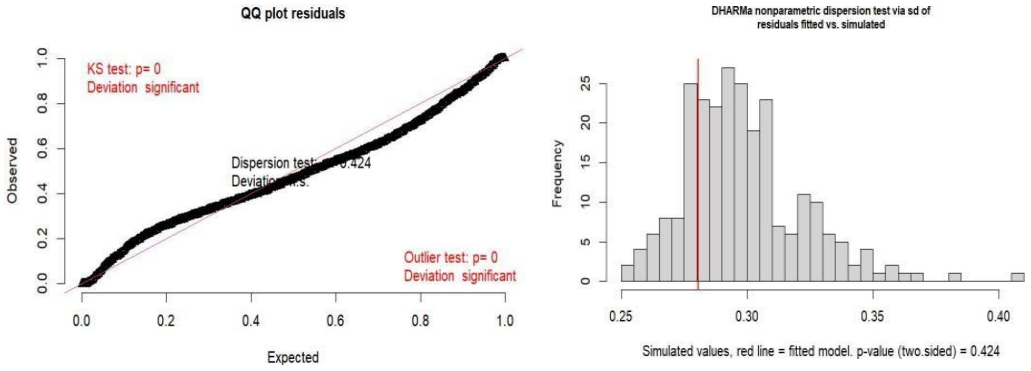



*Figure 22 : Résultats de Test de Leuven*

Les résultats de Test de Leuven suggèrent une distribution homogène des résidus. En effet, La valeur de dispersion est de 0.93938. Cela signifie que les variances des résidus simulés et des résidus ajustés du modèle sont similaires (près de 1, qui indiquerait une correspondance parfaite). De plus, une p-valeur = 0.424 > 0.05 indique que nous ne rejetons pas l'hypothèse nulle d'homogénéité des variances. Cela signifie que les variances des résidus ne diffèrent pas de manière significative.

***** Independence des erreurs**

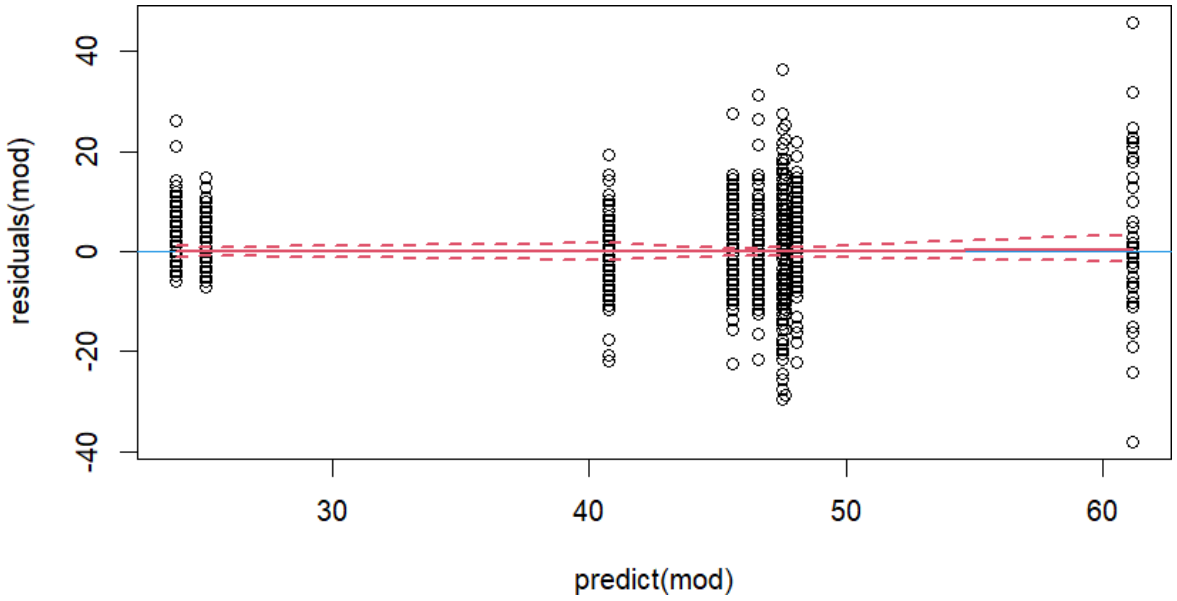



*Figure 23 : Résultats de test d'indépendance des erreurs _ modèle linéaire mixte*

D'après ces résultats, il semble que les résidus sont plutôt bien répartis autour de la ligne Y = 0 sans montrer des tendances particulières, ce qui suggère que l'hypothèse d'indépendance des résidus pourrait être valide.

****Diagnostic sur les effets aléatoires « releve»**

De la même manière, nous testons la validité de l'ensemble des hypothèses (normalité, homoscédasité et indépendance des erreurs) sur les résidus pour les effets aléatoires « releve ».

*****Normalité des résidus**

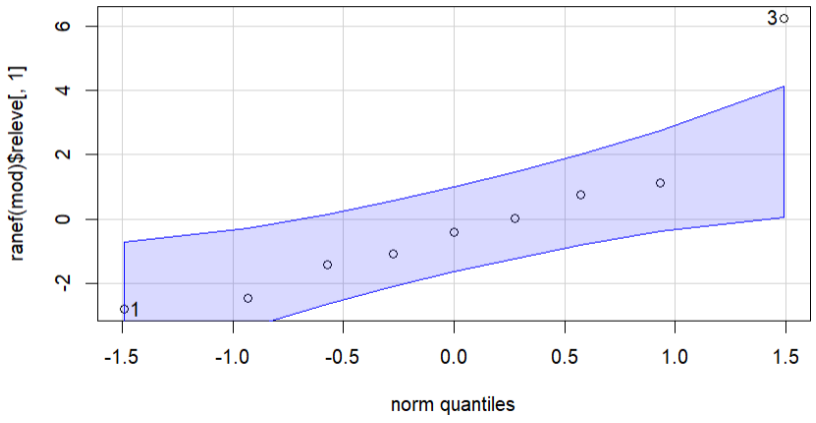
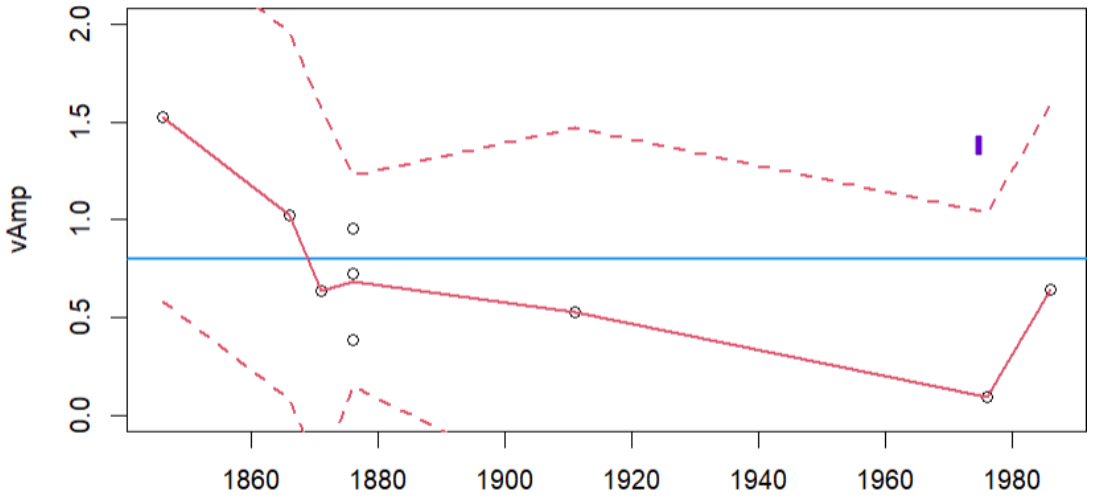



*Figure 24 : Résultats de test de normalité des résidus_ diagnostic sur les effets aléatoires*

Les points dans la figure ci-dessus semblent suivre une tendance linéaire croissante, ce qui suggère que les effets aléatoires sont normalement distribués. De plus, presque la totalité des points se trouve à l'intérieur de zone de confiance (fuseau en bleu). Donc d'après ce constat, on peut déduire que les effets aléatoires « releve » dans le modèle étudié semblent bien suivre une distribution normale malgré le nombre limité de points qui peut biaiser ce résultat.

*****Homogénéité des variances**




Le graphique ci-dessus suggère que les résidus sont assez bien répartis sans motif distinct et ils sont tous inclus dans l'intervalle de confiance (délimité par les deux traits en rouge) soutenant ainsi l'hypothèse d'homogénéité de variance.

*****Independence des erreurs**

On obtient dans le résultat de ce test (Figure 26) un nombre faible de points donc une faible puissance pour détecter d'éventuelles anomalies. Cependant, aucun problème est détecté, les résidus aléatoires sont répartis d'une façon aléatoire et indépendante et ils sont bien inclus dans la zone de confiance donc on peut supposer que l'hypothèse d'indépendance des résidus aléatoires est valide.

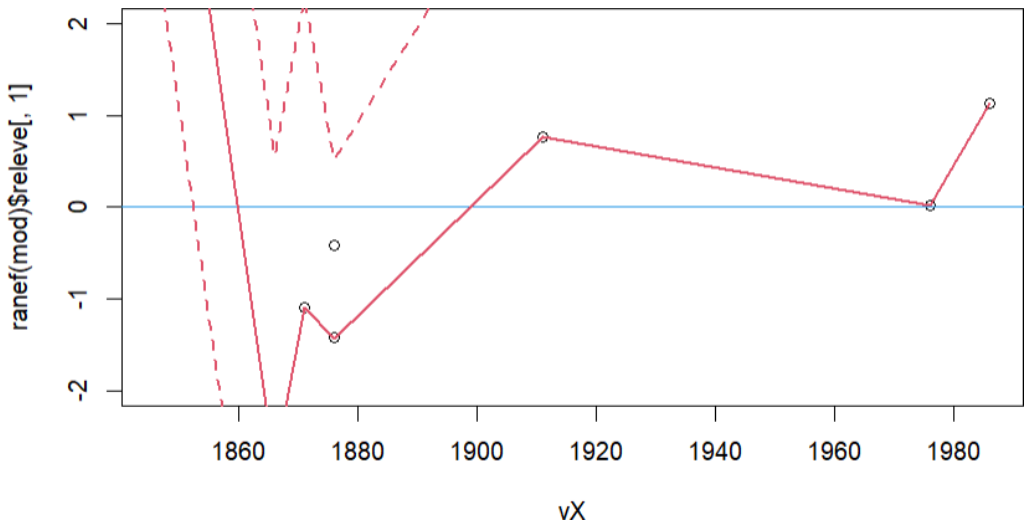



*Figure 25 : Résultat de test d'indépendance des erreurs _diagnostic des effets aléatoires*

****Analyse de sortie**

*<u>Résumé du modèle</u>

```
Linear mixed model fit by REML ['lmerMod']
Formula: DBH ~ lastLog + (1 | releve)
   Data: data_chene_clean

REML criterion at convergence: 10068.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.0928 -0.5483 -0.0919  0.4578  6.1165

Random effects:
 Groups   Name        Variance  Std.Dev.
 releve   (Intercept)  8.788    2.964
 Residual             56.157    7.494
Number of obs: 1462, groups:  releve, 9

Fixed effects:
            Estimate Std. Error t value
(Intercept) 478.4266    40.6459   11.77
LastLog      -0.2294     0.0214  -10.72

Correlation of Fixed Effects:
        (Intr)
LastLog -1.000
```

* Évaluation de l'effet fixe (LastLog)

La valeur estimée de l'intercept est de 478.4266, avec une erreur standard de 40.6459 et une valeur t de 11.77. Cela indique que lorsque lastLog est égal à zéro, la valeur prédite de DBH est de 478.4266. La valeur t élevée et l'erreur standard relativement faible suggèrent que cet estimateur est statistiquement significatif.

La pente associée à lastLog est de -0.2294, avec une erreur standard de 0.0214 et une valeur t de -10.72. Cela indique une relation négative significative entre lastLog et DBH.

* Évaluation des effets aléatoires

La variance de l'intercept pour releve est de 8.788 avec un écart-type de 2.964, tandis que la variance résiduelle est de 56.157 avec un écart-type de 7.494. Ces résultats montrent que le modèle tient compte de la variabilité entre les groupes releve et capture également la variation résiduelle.

****Intervalles de Confiance**

| | 2.5 % | 97.5 % |
|---|---|---|
| .sig01 | 1.5924566 | 4.5926577 |
| .sigma | 7.2297107 | 7.7752483 |
| (Intercept) | 400.4385056 | 557.0567789 |
| lastLog | -0.2708158 | -0.1883675 |

L'écart-type (.sig01) des variations entre les relevés (effet aléatoire) est compris entre 1.592 et 4.593, indiquant une variabilité significative mais modérée du diamètre des arbres entre les relevés.

L'écart-type (.sigma) reflète la variabilité inexpliquée après prise en compte des effets fixes (lastLog) et aléatoires (relevé). L'intervalle de confiance étroit suggère une estimation précise de cette variabilité.

Pour la variable *lastLog*, l'intervalle de confiance montre qu'une augmentation d'une unité de *lastLog* entraîne une diminution moyenne du diamètre des arbres comprise entre -0.271 mm et -0.188 mm. Cela confirme un effet significatif et négatif de *lastLog* sur le diamètre des arbres.

Ainsi, une partie de la variabilité est attribuée aux différences entre relevés (effet aléatoire), bien qu'une variabilité résiduelle (.sigma) persiste, suggérant l'influence d'autres facteurs non modélisés.

*Diagnostic des coefficients de détermination marginal et conditionnel

```
R2m       R2c
[1,] 0.6977458 0.7386439
```

Un R2m de 0.6977 indique que 69,77 % de la variance du diamètre des arbres (DBH) est expliquée uniquement par l'effet fixe *lastLog*. En revanche, un R2c de 0.7386 montre que 73,86 % de la variance totale de DBH est expliquée par l'ensemble des effets du modèle, incluant la variation entre les relevés (effets aléatoires). La légère différence entre R2c et R2m suggère que les effets aléatoires, représentant la variabilité entre les relevés, contribuent légèrement à l'explication supplémentaire de la variance totale de DBH, au-delà de ce qui est capté par l'effet fixe *lastLog*.

## 3.5 Question 5 (Bonus) : Modèle Linéaire Généralisé avec la Présence de Cavité Basse

**Modèle linéaire :**

Pour cette question, nous avons appliqué un modèle linéaire en prenant comme facteurs explicatifs le relevé, l'altitude et le diamètre des arbres. De la même manière que précédemment, on évalue le modèle à travers une vérification de la validité des hypothèses d'un modèle linéaire.

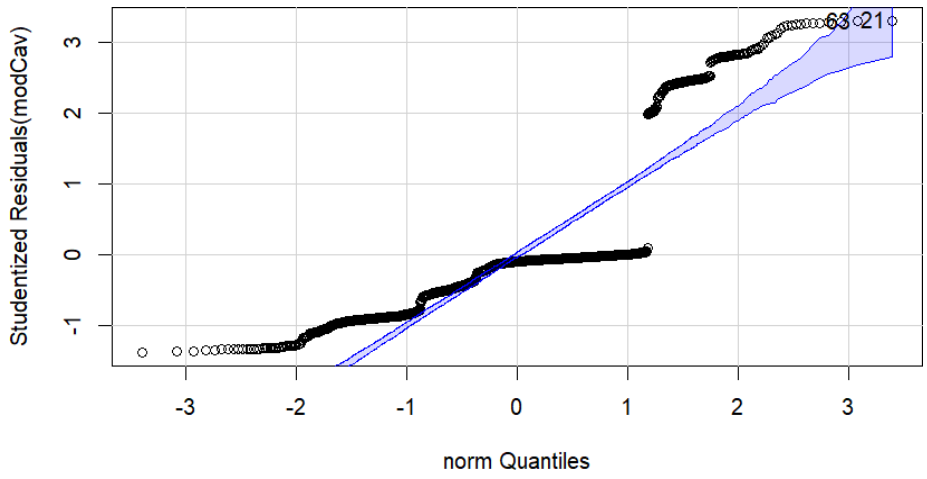



*Figure 26 : Résultat de Test de normalité Qqplot pour un modèle linéaire*

Avec que le Qqplot, on observe bien une discontinuité (comme l'exemple vu dans le cours).

**Modèle linéaire généralisé :**

On applique le modèle linéaire généralisé de type binomial pour étudier l'effet du relevé et de l'altitude et du diamètre de l'arbre sur la présence ou non d'une cavité basse. Grâce à la bibliothèque *DHARMa*, nous pouvons transformer les résidus afin de les rendre « continus, homoscédastiques » et non-biaisés si les hypothèses du modèle sont validés" (cf. cours). Nous appliquons donc cette transformation et testons son uniformité.

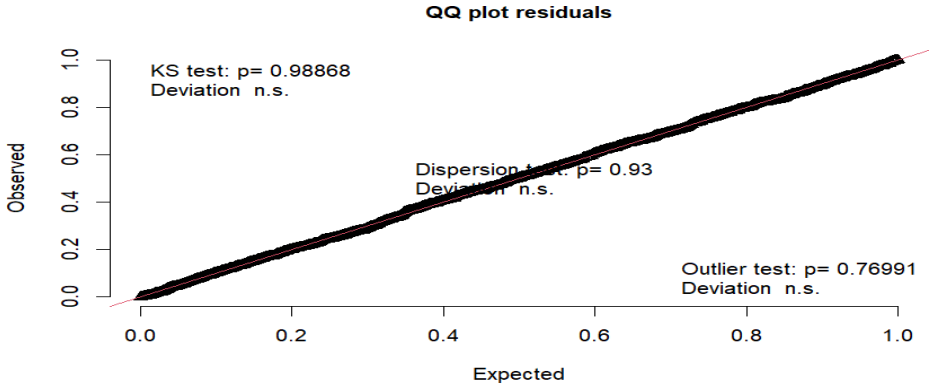



*Figure 27 : Résultat de qqplot résiduels pour un modèle linéaire généralisé de type binomial*

Ce qqplot des résidus confirme la distribution attendue, c'est à dire qu'ils suivent bien une loi uniforme entre 0 et 1.

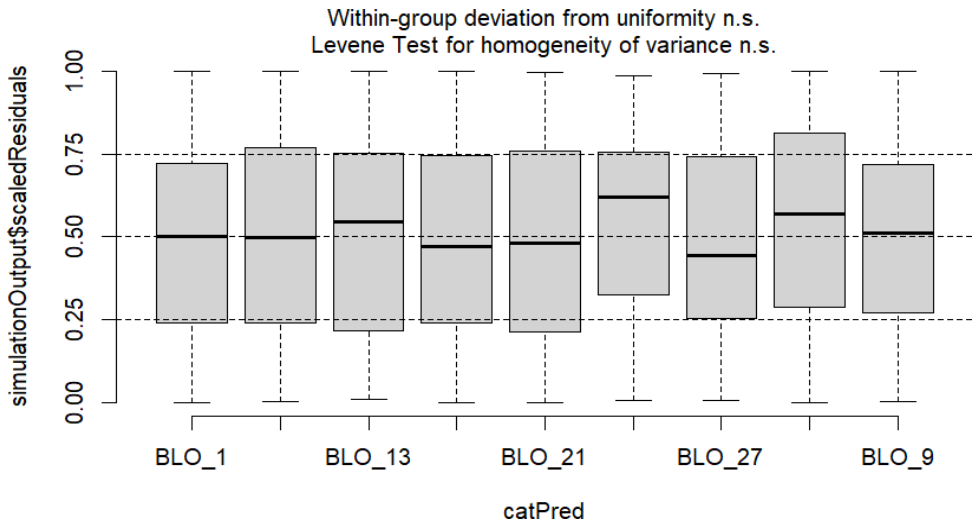



*Figure 28 : Résulats de test d'uniformité des résidus*

Avec la fonction testCategorical, nous observons que nous avons une distribution plus ou moins uniforme.

Vérification de la répartition homogène des résidus selon l'altitude :

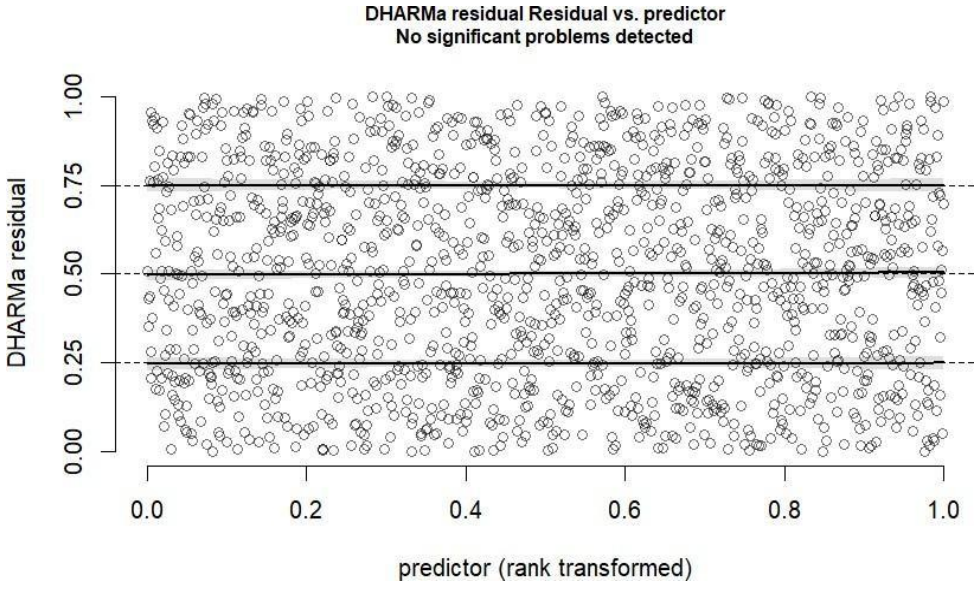



*Figure 29 : Résultat de Test d'homogénéité des résidus*

**Analyse des sorties**

*Résumé du modèle

```
Call:
glm(formula = cavPA ~ alti + releve + DBH, family = binomial(link = "logit"),
    data = dataIni_Querc)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.482972   5.192341   1.441 0.149541
alti         -0.017457   0.010951  -1.594 0.110915
releveBLO_12 -5.358955   1.556375  -3.443 0.000575 ***
releveBLO_13  0.396694   0.355347   1.116 0.264270
releveBLO_17 -6.436675   1.851172  -3.477 0.000507 ***
releveBLO_21 -4.252894   2.563185  -1.659 0.097071 .
releveBLO_24 -3.171173   1.939715  -1.635 0.102077
releveBLO_27 -4.464472   2.339876  -1.908 0.056392 .
releveBLO_4  -4.207911   1.521906  -2.765 0.005694 **
releveBLO_9  -3.624452   1.691620  -2.143 0.032146 *
DBH          -0.003412   0.008976  -0.380 0.703888
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1066.9  on 1460  degrees of freedom
Residual deviance:  865.4  on 1450  degrees of freedom
AIC: 887.4

Number of Fisher Scoring iterations: 7
```

*Figure 29 : Résumé du modèle Linéaire généralisé de type binomial*

*Les intervalles de confiance

```
                  2.5 %       97.5 %
(Intercept)   -2.69746883 17.703977179
alti          -0.03904380  0.003980517
releveBLO_12  -8.42329941 -2.307725582
releveBLO_13  -0.30310878  1.094147372
releveBLO_17 -10.09099925 -2.815772824
releveBLO_21  -9.29381813  0.777605888
releveBLO_24  -6.98212545  0.639520823
releveBLO_27  -9.06859906  0.125390117
releveBLO_4   -7.20914440 -1.229036756
releveBLO_9   -6.95674779 -0.309489338
DBH           -0.02113494  0.014135156
```

*Figure 30 : Résultat des intervalles de confiance*

*Test du modèle emboîté

```
Analysis of Deviance Table

Model 1: cavPA ~ alti + releve
Model 2: cavPA ~ alti + releve + DBH
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1451     865.54
2      1450     865.40  1  0.14472   0.7036
```

On peut en déduire que l'impact du diamètre dans la présence ou non de cavité n'a pas un impact significatif, le P-Value dans le résumé bien supérieur à 0 était déjà un bon indicateur.

*$R^2$ de McFadden :

Le $R^2$ de McFadden "qui quantifie de combien on a progressé vers le meilleur modèle possible en terme de vraisemblance" est similaire à celui du cours (soit sans le diamètre) 0.1889081 alors que celui obtenu avec ce modèle est de 0.1888564.

Comme la mesure de la qualité d'ajustement dans notre cas est inférieure à celle du cours. On pourrait en conclure que le diamètre n'apporte pas de plus-value significative bien au contraire.

# 4. Conclusion

Cette étude a permis d'explorer différentes approches statistiques pour analyser les données forestières, en particulier la variabilité du diamètre des arbres en fonction de divers facteurs explicatifs. En mobilisant des modèles ANOVA, des régressions linéaires, des modèles mixtes et généralisés, nous avons pu identifier les variables influentes et valider les hypothèses sous-jacentes à chaque modèle.

Les résultats ont mis en évidence l'effet significatif du relevé sur le diamètre des arbres, confirmant que l'analyse individuelle de chaque relevé est plus pertinente que l'agrégation par triplets pour expliquer cette variabilité. En effet, l'analyse par clusters n'a pas apporté d'éléments concluants pour expliquer la variation du diamètre, renforçant l'importance d'une approche plus détaillée par relevé.

Par ailleurs, la dernière année de coupe (lastLog) a montré un effet significatif mais négatif sur le diamètre des arbres, suggérant que plus la coupe est récente, plus le diamètre des arbres est réduit. En revanche, l'analyse du diamètre des arbres n'a pas permis de démontrer une relation significative avec la présence ou l'absence de cavités basses, indiquant que ce facteur ne joue pas un rôle déterminant dans cette étude.

Toutefois, certaines limites demeurent, notamment la présence de valeurs extrêmes et la nécessité d'affiner certains modèles pour mieux refléter la complexité des interactions entre facteurs explicatifs. Des perspectives futures pourraient inclure l'intégration d'autres variables environnementales (climat, sol) ou l'utilisation de modèles non linéaires pour capturer des dynamiques plus complexes.