UE 905 - Analyses statistiques

Master 2 SIGMA - Modèles linéaires

BALLOT Doris, BARBIERO Audrey, HECKENDORN Robin, LIMA Lucas

2025-06-02

1 Question 1: Analyse de l'effet des relevés sur le diamètre des charmes (ANOVA)

1.1 Chargement des données et filtrage

Nous allons commencer par charger le fichier `dataProjet_2025.csv` et filtrer les charmes (`Carpinus betulus L., 1753`).

```
# Charger les données
dataIni <- read.csv("dataProjet_2025.csv")

# Filtrer pour ne garder que les charmes
data_charmes <- dataIni %>%
  filter(recherche_esp_lb_nom_plantae == "Carpinus betulus L., 1753")

# Afficher un aperçu des données filtrées
head(data_charmes)
```

```
    X releve recherche_esp_lb_nom_plantae cav_basses_presence_cavites
1  74 BLO_27    Carpinus betulus L., 1753                         oui
2  81 BLO_27    Carpinus betulus L., 1753                         oui
3 131  BLO_1    Carpinus betulus L., 1753                         oui
4 170 BLO_12    Carpinus betulus L., 1753                         oui
5 173 BLO_12    Carpinus betulus L., 1753                         oui
6 182  BLO_4    Carpinus betulus L., 1753                         oui
  cara_cavites_circonference cara_cavites_avancement cara_cavites_carie TreeID
1                         85                       5            blanche  27-79
2                        110                       5          bois dur 27-126
```

```
3                          70                        5            blanche   1-15
4                         115                        4            blanche   12-4
5                          95                        4            blanche 12-134
6                         170                        5            blanche   4-78
  DBH Status     long     lati    alti lastLog
1  21 vivant 1.779229 44.00650 257.911    1876
2  23 vivant 1.779443 44.00683 246.227    1876
3  19 vivant 1.760582 44.05434 473.737    1866
4  21 vivant 1.725196 44.02283 338.970    1986
5  18 vivant 1.724767 44.02308 347.469    1986
6  44 vivant 1.751003 44.04876 327.044    1911
```

1.2 Ajustement et interprétation du modèle ANOVA

Nous ajustons un modèle ANOVA pour expliquer le diamètre (DBH) en fonction de l'effet du relevé

```r
# Ajuster le modèle ANOVA
mod_anova <- lm(DBH ~ releve, data = data_charmes)

# Afficher le résumé du modèle
summary(mod_anova)
```

```
Call:
lm(formula = DBH ~ releve, data = data_charmes)

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

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.5614     0.6163  39.856  < 2e-16 ***
releveBLO_12  -3.5614     1.2579  -2.831  0.00504 **
releveBLO_13  -2.8947     2.7560  -1.050  0.29464
releveBLO_21  -2.5930     0.7795  -3.326  0.00102 **
releveBLO_24  -3.0614     3.3471  -0.915  0.36133
releveBLO_27  -3.0947     0.9278  -3.336  0.00099 ***
```

```
releveBLO_4    1.2386     1.3502   0.917  0.35989
releveBLO_9   -3.9900     1.8634  -2.141  0.03329 *
---
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:  0.0977,    Adjusted R-squared:  0.07071
F-statistic:  3.62 on 7 and 234 DF,  p-value: 0.001004
```

1.2.1 Interprétation des résultats du modèle ANOVA

Le modèle ANOVA ajusté examine l'effet des différents relevés sur le diamètre des charmes (**DBH**). Voici les conclusions principales :

1. **Effet significatif de certains relevés** :
   - Les relevés **BLO_12, BLO_21, BLO_27 et BLO_9** ont un effet significatif sur le **DBH**, avec des p-valeurs inférieures à **0.05**.
   - Ces relevés entraînent une **réduction significative** du **DBH** par rapport au relevé de référence (**Intercept**, supposé être le relevé de base).
   - Par exemple :
     - **BLO_12** : réduction de **-3.56** unités ($\mathbf{p = 0.005}$).
     - **BLO_27** : réduction de **-3.09** unités ($\mathbf{p < 0.001}$).
     - **BLO_9** : réduction de **-3.99** unités ($\mathbf{p = 0.033}$).
   - En revanche, d'autres relevés (**BLO_13, BLO_24, BLO_4**) **n'ont pas d'effet significatif** ($\mathbf{p > 0.05}$), suggérant que leurs différences avec le groupe de référence pourraient être dues à la variabilité naturelle.
2. **Effet global du modèle** :
   - Le modèle est **globalement significatif** ($\mathbf{F = 3.62, p = 0.001}$), indiquant que les relevés expliquent **une partie** de la variabilité observée dans le **DBH**.
   - Toutefois, la valeur de $\mathbf{R^2}$ **ajusté** $\mathbf{= 7.07\%}$ montre que le modèle ne capture qu'une **faible proportion** de la variabilité totale du **DBH**, ce qui suggère la présence d'autres facteurs explicatifs non inclus dans ce modèle.
3. **Variabilité résiduelle** :

- L'**erreur standard résiduelle** est de **4.653**, indiquant une dispersion relativement importante des observations autour des valeurs prédites.

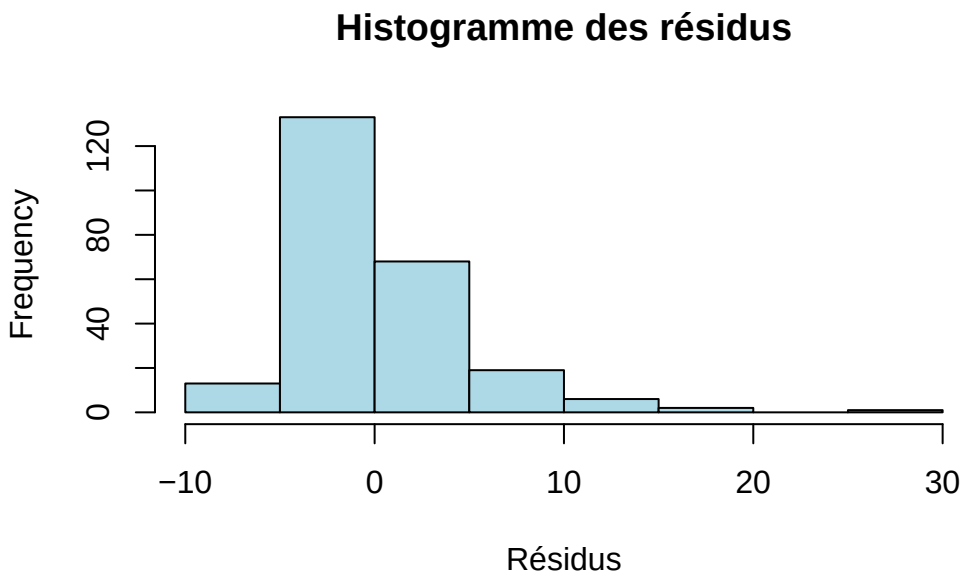
- Cela signifie que, bien que certains effets soient significatifs, **le modèle reste limité en capacité prédictive**.

Conclusion :
- Certains relevés ont un effet significatif sur le diamètre des charmes (**DBH**), mais la faible valeur du **R^2 ajusté** indique que d'autres facteurs non inclus dans le modèle pourraient mieux expliquer la variabilité observée.
- Avant de tirer des conclusions définitives, il est **nécessaire de vérifier les hypothèses du modèle linéaire** (normalité et homoscédasticité des résidus) pour s'assurer de la validité des résultats.

1.3 Diagnostic des hypothèses et normalisation des résidus

D'abord, nous allons analyser l'histogramme des résidus.

```
# Histogramme des résidus
hist(residuals(mod_anova),
     main = "Histogramme des résidus",
     xlab = "Résidus",
     col = "lightblue")
```

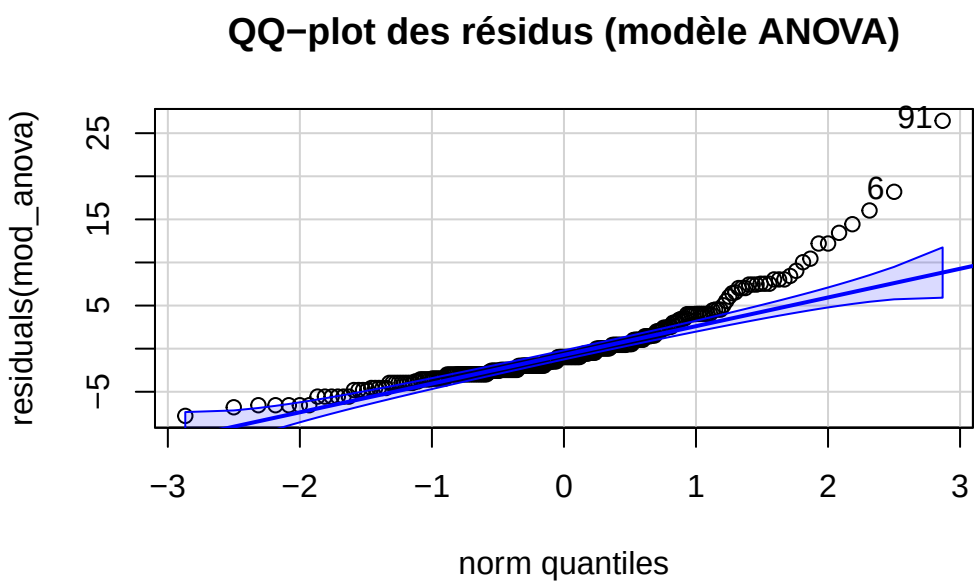


1.3.1 Interprétation de l'histogramme des résidus

1. **Asymétrie à droite** :
   - L'histogramme montre une **queue étendue vers la droite**, suggérant une légère **asymétrie positive** dans la distribution des résidus.
   - Cela pourrait indiquer un écart par rapport à la normalité.
2. **Concentration centrale** :
   - La majorité des résidus sont proches de **zéro**, ce qui est attendu dans un modèle linéaire bien ajusté.
   - Cela suggère que le modèle capture correctement une partie de la tendance globale.
3. **Résidus extrêmes** :
   - Des résidus supérieurs à **10** (jusqu'à 30) sont présents, ce qui pourrait signaler des **valeurs aberrantes** ou des observations influentes dans les données.

Il est important de produire un QQ-plot pour vérifier la normalité des résidus, ce que nous allons faire ci-dessous.

```r
# QQ-plot pour vérifier la normalité des résidus
qqPlot(residuals(mod_anova), main = "QQ-plot des résidus (modèle ANOVA)")
```




```
[1] 91  6
```

1.3.2 Interprétation du QQ-plot des résidus

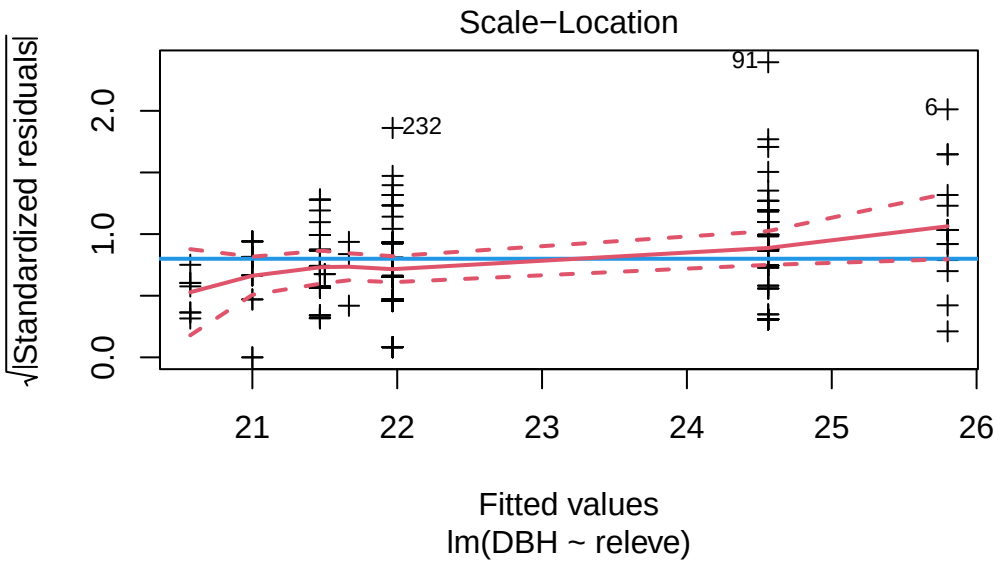
1. **Écarts dans les queues** :
   - Les points s'éloignent nettement de la ligne de référence dans la **queue droite**, confirmant la présence de **valeurs extrêmes** et une **déviation de la normalité**.
2. **Valeurs aberrantes** :
   - Les points **91** et **6** sont identifiés comme des **outliers** influençant la distribution des résidus.
3. **Alignement général** :
   - Les résidus sont globalement alignés sur la ligne centrale pour les quantiles intermédiaires, mais les écarts dans les extrémités indiquent que l'hypothèse de normalité est **partiellement violée**.

Conclusion : La **normalité des résidus** est compromise par des valeurs aberrantes, ce qui pourrait affecter la robustesse des conclusions du modèle ANOVA.

1.4 Vérification des hypothèses de l'homoscédasticité

Pour compléter le diagnostic, vérifions si les résidus présentent une **variance constante** (homoscédasticité) en utilisant le **Scale-Location Plot**.

```
# Scale-Location plot pour vérifier l'homoscédasticité
plot(mod_anova, which = 3, pch = 3, add.smooth = FALSE)
abline(h = 0.8, col = 4, lwd = 2)

# Moyenne glissante des points
lo <- loess(sqrt(abs(rstandard(mod_anova))) ~ mod_anova$fitted.values)
vFit <- sort(unique(mod_anova$fitted.values))
predLo <- predict(lo, vFit, se = TRUE)
lines(predLo$fit ~ vFit, col = 2, lwd = 2)

# Enveloppe de confiance
nFit <- length(vFit)
ICBonf <- qnorm(1 - 0.05 / 2 / nFit)
lines(predLo$fit + ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
lines(predLo$fit - ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
```



1.4.1 Interprétation du Scale-Location Plot

1. **Augmentation légère de la variance** :
   - La courbe rouge montre une **tendance croissante modérée** pour les valeurs ajustées les plus élevées.
   - Cela indique une **légère hétéroscédasticité**, avec une variance des résidus qui n'est pas parfaitement constante.
2. **Présence de valeurs aberrantes** :
   - Les observations **91, 6 et 232** sont identifiées comme des **outliers** avec des résidus standardisés particulièrement élevés.
3. **Acceptabilité globale** :
   - Malgré une augmentation modérée de la variance, le modèle reste **globalement interprétable**, mais avec prudence en ce qui concerne les effets des valeurs extrêmes.

Conclusion des diagnostics

1. **Normalité des résidus** : elle est **partiellement compromise** en raison de **valeurs aberrantes** identifiées dans le QQ-plot.

2. **Homoscédasticité** : une **légère hétéroscédasticité** est observée, mais elle reste acceptable pour interpréter le modèle actuel.

1.5 Ajustement du modèle avec une transformation de Box-Cox

En raison d'une légère déviation des hypothèses de normalité et d'homoscédasticité, nous allons appliquer une transformation de Box-Cox. Cette méthode ajuste de manière optimale l'échelle des données afin de stabiliser la variance et d'améliorer l'adéquation des résidus à une distribution normale.

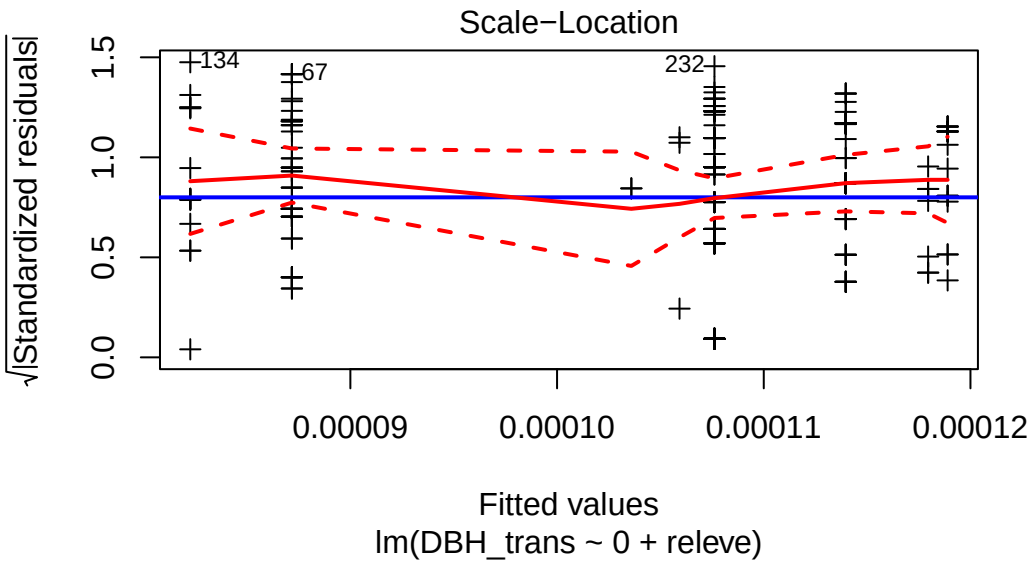
1.5.1 Application de la transformation de Box-Cox

```
# Calculer le lambda optimal
mod_anova_pT <- powerTransform(mod_anova)
lambda <- mod_anova_pT$lambda
lambda
```

```
       Y1
-2.764963
```

```
# Transformation inverse cubique des données
data_charmes$DBH_trans <- 1 / (data_charmes$DBH^3)
mod_anova_BC <- lm(DBH_trans ~ 0 + releve, data = data_charmes)
```

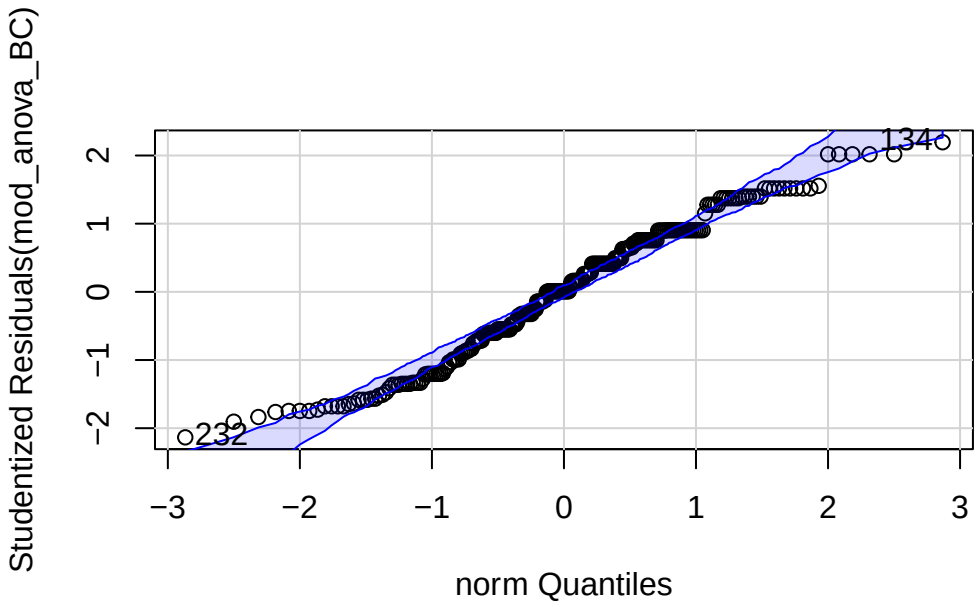
```
# Vérification après transformation (Scale-Location plot)
plot(mod_anova_BC, which = 3, pch = 3, add.smooth = FALSE)
abline(h = 0.8, col = "blue", lwd = 2)
lo <- loess(sqrt(abs(rstandard(mod_anova_BC))) ~ mod_anova_BC$fitted.values)
vFit <- sort(unique(mod_anova_BC$fitted.values))
predLo <- predict(lo, vFit, se = TRUE)
lines(predLo$fit ~ vFit, col = "red", lwd = 2)
lines(predLo$fit + ICBonf * predLo$se.fit ~ vFit, col = "red", lwd = 2, lty = "dashed")
lines(predLo$fit - ICBonf * predLo$se.fit ~ vFit, col = "red", lwd = 2, lty = "dashed")
```



1.5.2 Interprétation du Scale-Location Plot après transformation Box-Cox

1. **Stabilisation de la variance** :
   - La courbe rouge (moyenne glissante) est globalement plate, indiquant une variance relativement constante des résidus.
   - Cela montre que la transformation a corrigé en grande partie le problème d'hétéroscédasticité observé dans le modèle initial.
2. **Valeurs aberrantes** :
   - Quelques observations (ex. **134, 67, 232**) restent influentes, mais elles ne perturbent pas la tendance globale.
3. **Acceptabilité globale** :
   - La variance des résidus est désormais stable, rendant le modèle transformé **valide au regard de l'homoscédasticité**.

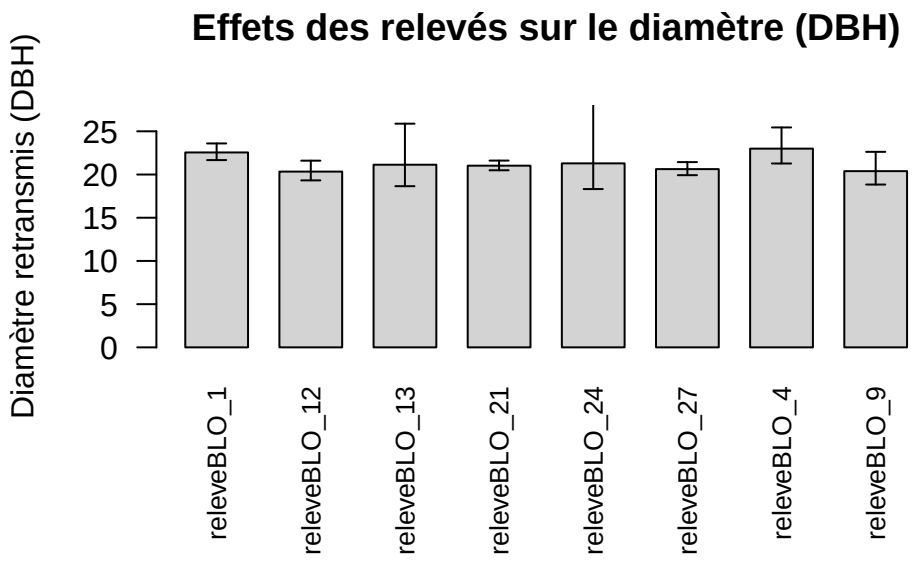
```
# Vérification après transformation (QQ plot)
qqPlot(mod_anova_BC, distribution = "norm", line = "none")
```




```
[1] 134 232
```

1.5.3 Interprétation du QQ-Plot après transformation Box-Cox

1. **Alignement général** :
   - Les points suivent bien la ligne diagonale, ce qui indique que les résidus sont **proches d'une distribution normale**.
2. **Déviation dans les extrémités** :
   - Quelques observations aux extrémités (**134 et 232**) montrent de légers écarts, mais ils restent modérés et n'affectent pas significativement la normalité globale.
3. **Amélioration globale** :
   - Par rapport au modèle sans transformation, la symétrie des résidus est nettement améliorée. La transformation de Box-Cox a réussi à rendre les résidus **plus conformes à une distribution normale**.



`Coefficient de détermination (R²) : 0.8618119`

1.6 Conclusion pour la question 1

La transformation de **Box-Cox** a permis d'améliorer la **normalité des résidus** et, combinée au Scale-Location Plot, valide que le modèle transformé est désormais plus robuste.

Le modèle final explique 86,18 % de la variance (R2 = 0.86), indiquant une bonne qualité d'ajustement.

---

2 Question 2 : Régression linéaire avec 'lastLog'

2.1 Chargement des données et filtrage

Nous allons commencer par filtrer les **chênes** (**Quercus L., 1753**) dans le champ `recherche_esp_lb_nom_plantae`.

```r
# Filtrer pour ne garder que les chênes
data_chenes <- dataIni %>%
  filter(recherche_esp_lb_nom_plantae == "Quercus L., 1753")

# Afficher un aperçu des données filtrées
head(data_chenes)
```

```
   X releve recherche_esp_lb_nom_plantae cav_basses_presence_cavites
1 63 BLO_27             Quercus L., 1753                         oui
2 64 BLO_27             Quercus L., 1753                         oui
3 65 BLO_27             Quercus L., 1753                         oui
4 66 BLO_27             Quercus L., 1753                         oui
5 67 BLO_27             Quercus L., 1753                         oui
6 68 BLO_27             Quercus L., 1753                         oui
  cara_cavites_circonference cara_cavites_avancement cara_cavites_carie TreeID
1                        195                       5              rouge   27-6
2                        190                       5              rouge  27-10
3                        146                       6            blanche  27-15
4                        310                       6            blanche  27-22
5                        204                       5          bois dur  27-23
6                        190                       5            blanche  27-34
  DBH Status     long     lati    alti lastLog
1  47 vivant 1.779575 44.00633 256.980    1876
2  40 vivant 1.779790 44.00623 255.576    1876
3  36 vivant 1.779495 44.00628 259.620    1876
4  58 vivant 1.779172 44.00633 257.551    1876
5  50 vivant 1.779261 44.00631 257.252    1876
6  41 vivant 1.779335 44.00599 259.665    1876
```

2.2 Vérification et préparation des données

Nous devons nous assurer que la variable **lastLog** existe dans le jeu de données filtré.

```r
# Vérifier les noms des colonnes dans le jeu de données filtrées
colnames(data_chenes)
```

```
 [1] "X"                           "releve"
 [3] "recherche_esp_lb_nom_plantae" "cav_basses_presence_cavites"
```

```
 [5] "cara_cavites_circonference"   "cara_cavites_avancement"
 [7] "cara_cavites_carie"           "TreeID"
 [9] "DBH"                          "Status"
[11] "long"                         "lati"
[13] "alti"                         "lastLog"
```

2.3 Ajustement du modèle de régression linéaire

Nous allons maintenant ajuster un **modèle de régression linéaire** pour expliquer le **diamètre** (`DBH`) en fonction de la variable `lastLog` (date de la dernière coupe massive).

```r
# Ajuster le modèle de régression linéaire
mod_lm <- lm(DBH ~ lastLog, data = data_chenes)

# Afficher le résumé du modèle
summary(mod_lm)
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-31.478  -4.342  -0.520   3.658  53.166

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 455.884554   7.286707   62.56   <2e-16 ***
lastLog      -0.217796   0.003785  -57.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.72 on 1460 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.694, Adjusted R-squared:  0.6938
F-statistic:  3312 on 1 and 1460 DF,  p-value: < 2.2e-16
```

2.3.1 Interprétation du modèle de régression linéaire

Nous avons ajusté un modèle de régression linéaire pour expliquer le diamètre (**DBH**) en fonction de la variable `lastLog` (date de dernière coupe massive).

1. Résumé des résultats principaux

- **Équation du modèle** :
  [DBH = 455.88 - 0.2178 lastLog]
  - **Intercept** : **455.88** représente le diamètre estimé lorsque `lastLog = 0` (hypothétique).
  - **lastLog** : Le coefficient **-0.2178** indique que le **DBH diminue** de **0.2178 unités par an** après la dernière coupe. Cet effet est **très significatif** ($p < 2e\text{-}16$), suggérant une relation forte entre `lastLog` et `DBH`.

2. Qualité globale du modèle

- **R^2 ajusté = 69.38%** → `lastLog` explique **69.38% de la variabilité du diamètre des chênes**.
- **F-statistic = 3312, p < 2.2e-16** → Le modèle est **globalement significatif**, indiquant que la variable `lastLog` est un bon prédicteur du `DBH`.

3. Variabilité résiduelle

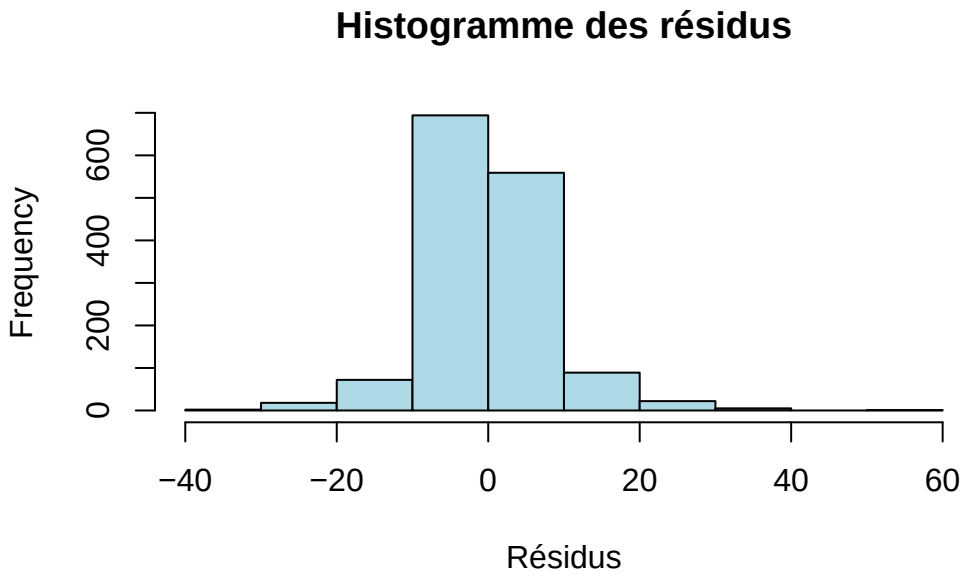
- **Erreur standard résiduelle = 7.72** → Les valeurs de `DBH` sont en moyenne à **±7.72 unités** des prédictions.
- **Étendue des résidus** : **de -31.48 à 53.16**, suggérant des écarts importants qui nécessitent une vérification des hypothèses du modèle.

Prochaine étape : Vérification des hypothèses du modèle via l'analyse des résidus.

2.4 Diagnostic des résidus

Nous allons maintenant **vérifier les hypothèses** du modèle linéaire :
1. **Normalité des résidus**.
2. **Homoscédasticité** (variance constante des résidus).

```r
# Histogramme des résidus
hist(residuals(mod_lm),
     main = "Histogramme des résidus",
     xlab = "Résidus",
     col = "lightblue")
```



2.4.1 Interprétation de l'histogramme des résidus

L'histogramme des résidus permet d'évaluer visuellement la distribution des erreurs du modèle. Voici les principales observations :

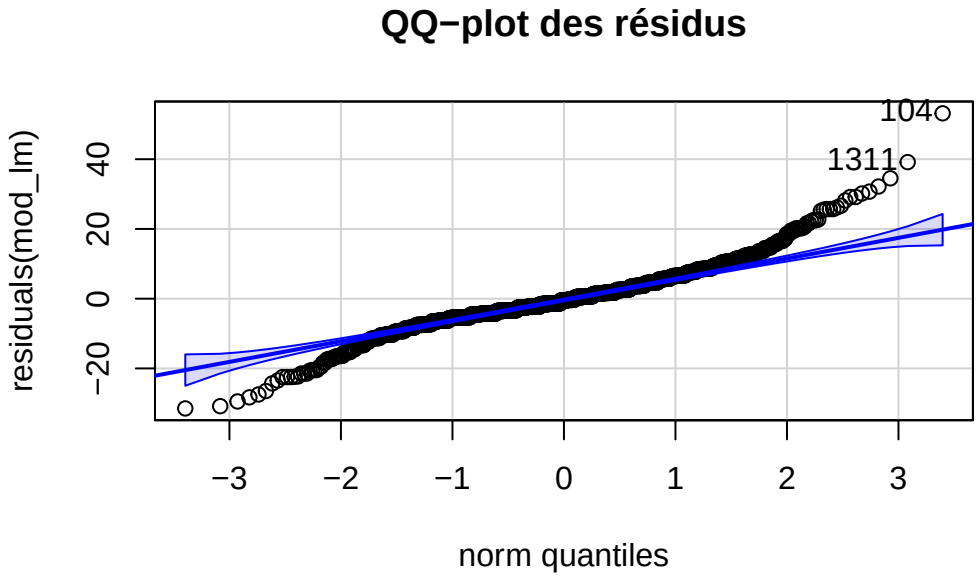
1. **Symétrie imparfaite**
   - Les résidus sont **concentrés autour de zéro**, ce qui correspond à un bon ajustement général du modèle.
   - Une légère **asymétrie vers la droite** est observable, avec quelques valeurs positives élevées.
2. **Présence de valeurs extrêmes**
   - Les résidus couvrent une plage allant de **-40 à 60**, indiquant la possible existence de **valeurs aberrantes** ou influentes qui méritent une analyse plus approfondie.

3. **Distribution en cloche approximative**

    - La forme générale est **proche d'une distribution normale**, mais les **queues étendues** reflètent une déviation potentielle, notamment dans les valeurs extrêmes.

Le **QQ-plot** nous permettra de confirmer visuellement si les résidus suivent une distribution normale.

```
# QQ-plot pour vérifier la normalité des résidus
qqPlot(residuals(mod_lm), main = "QQ-plot des résidus")
```




```
104 1311
104 1307
```

2.4.2 Interprétation du QQ-plot

1. **Déviation aux extrémités** :

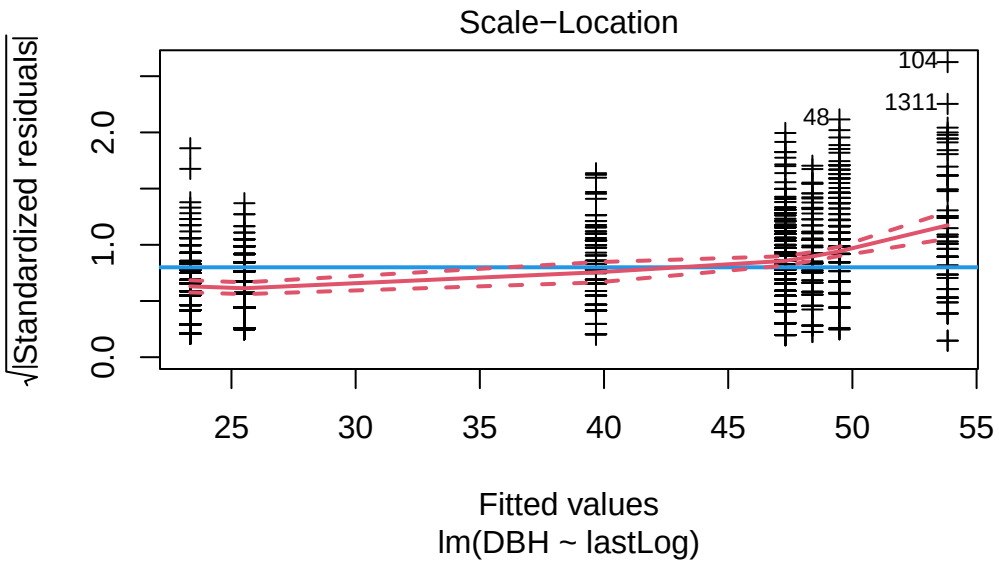
    - Les **queues de la distribution** montrent des écarts significatifs par rapport à la ligne théorique, indiquant la présence de **valeurs aberrantes** ou de **résidus non normaux**.

2. **Alignement central** :

- Les points situés au centre suivent bien la ligne théorique, suggérant que les **résidus centraux sont proches d'une distribution normale**.

3. **Présence d'outliers** :
   - Les observations **104** et **1311** sont identifiées comme des **valeurs extrêmes**, susceptibles d'influencer le modèle.

```
# Scale-Location plot pour vérifier l'homoscédasticité
plot(mod_lm, which = 3, pch = 3, add.smooth = FALSE)
abline(h = 0.8, col = 4, lwd = 2)

# Moyenne glissante des points
lo <- loess(sqrt(abs(rstandard(mod_lm))) ~ mod_lm$fitted.values)
vFit <- sort(unique(mod_lm$fitted.values))
predLo <- predict(lo, vFit, se = TRUE)
lines(predLo$fit ~ vFit, col = 2, lwd = 2)

# Enveloppe de confiance
nFit <- length(vFit)
ICBonf <- qnorm(1 - 0.05 / 2 / nFit)
lines(predLo$fit + ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
lines(predLo$fit - ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
```



2.4.3 Interprétation du Scale-Location Plot

1. **Variance légèrement croissante** :
   - La courbe rouge montre une **augmentation modérée** de la variance des résidus pour les valeurs ajustées les plus élevées.
   - Cela suggère une **légère hétéroscédasticité**, mais elle ne semble pas critique.
2. **Présence de valeurs aberrantes** :
   - Les observations **104** et **1311** sont clairement identifiées comme **outliers**, ce qui pourrait influencer le modèle.
3. **Homoscédasticité** :
   - Le Scale-Location Plot révèle une légère tendance à l'augmentation de la variance, ce qui est une **déviation modérée** de l'hypothèse d'homoscédasticité.

Passons maintenant à l'introduction d'un **terme quadratique** dans le modèle pour vérifier si une **dépendance non-linéaire** améliore l'ajustement.

2.5 Ajustement du modèle quadratique

```r
# Ajuster un modèle quadratique
mod_quad <- lm(DBH ~ lastLog + I(lastLog^2), data = data_chenes)

# Afficher le résumé du modèle quadratique
summary(mod_quad)
```

```
Call:
lm(formula = DBH ~ lastLog + I(lastLog^2), data = data_chenes)

Residuals:
    Min      1Q  Median      3Q     Max
-33.585  -4.033  -0.599   3.721  50.415

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.378e+03  6.495e+02   5.201 2.26e-07 ***
lastLog      -3.256e+00  6.751e-01  -4.822 1.57e-06 ***
```

```
I(lastLog^2)  7.888e-04  1.753e-04   4.500 7.35e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.67 on 1459 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6982,    Adjusted R-squared:  0.6978
F-statistic:  1688 on 2 and 1459 DF,  p-value: < 2.2e-16
```

2.5.1 Interprétation du modèle quadratique

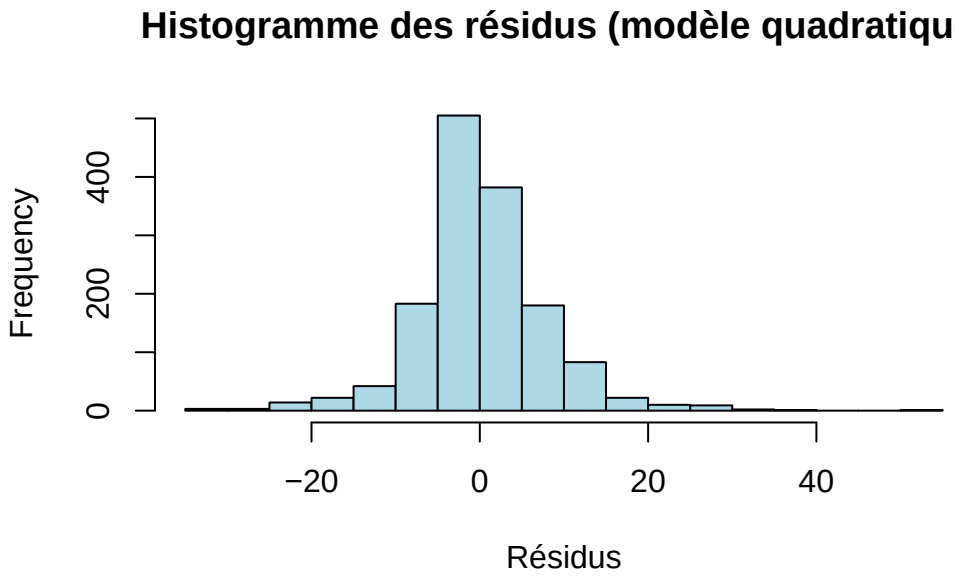
1. **Résumé du modèle ajusté :**
   - **Équation** :
     [DBH = 3378 - 3.256 lastLog + 0.0007888 lastLog^2]
   - Les deux termes, **linéaire** et **quadratique**, sont **significatifs** :
     - **lastLog** : (= -3.256), (p < 0.001)
     - **lastLog²** : (= 0.0007888), (p < 0.001)
2. **Qualité du modèle :**
   - **R² ajusté** : **69.78%**
     L'ajout du terme quadratique améliore légèrement la qualité d'ajustement par rapport au modèle linéaire simple (**R² ajusté = 69.38%**).
   - **Erreur résiduelle standard** : **7.67**, légèrement inférieure à celle du modèle simple (**7.72**).
3. **Signification des coefficients :**
   - Le coefficient **positif** de (lastLog^2) indique une légère **courbure vers le haut** dans la relation entre **DBH** et **lastLog**. Cela reflète une dépendance **non-linéaire**, où l'effet de **lastLog** diminue à mesure que sa valeur augmente.
4. **Significativité globale :**
   - Le modèle est globalement significatif (**F = 1688, p < 2.2e-16**), confirmant que les deux termes expliquent une part significative de la variabilité du **DBH**.

2.6 Diagnostic des résidus pour le modèle quadratique

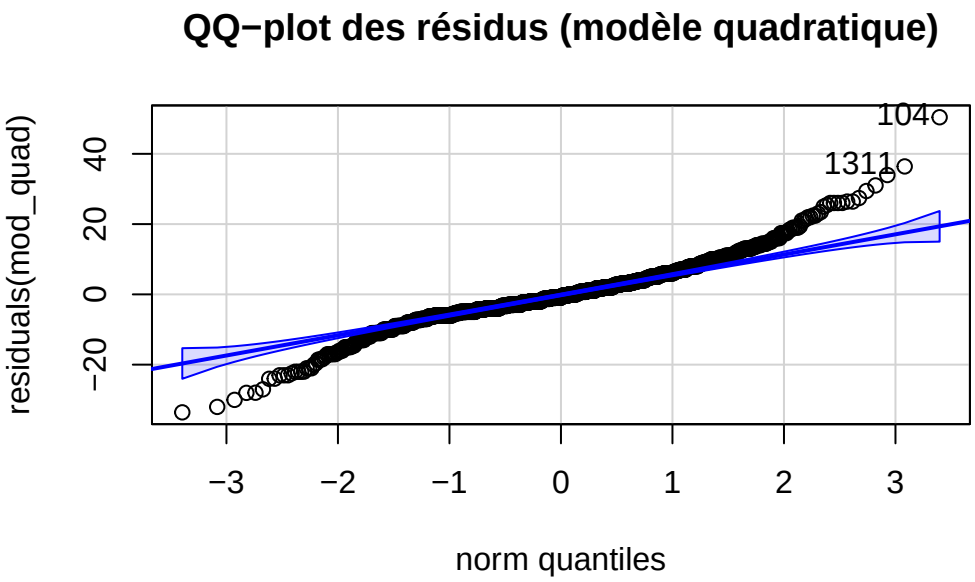
Nous allons maintenant vérifier si l'ajout du terme quadratique a amélioré les hypothèses du modèle, notamment :

1. La normalité des résidus.
2. L'homoscédasticité.

```
# Histogramme des résidus du modèle quadratique
hist(residuals(mod_quad),
     main = "Histogramme des résidus (modèle quadratique)",
     xlab = "Résidus",
     col = "lightblue")
```

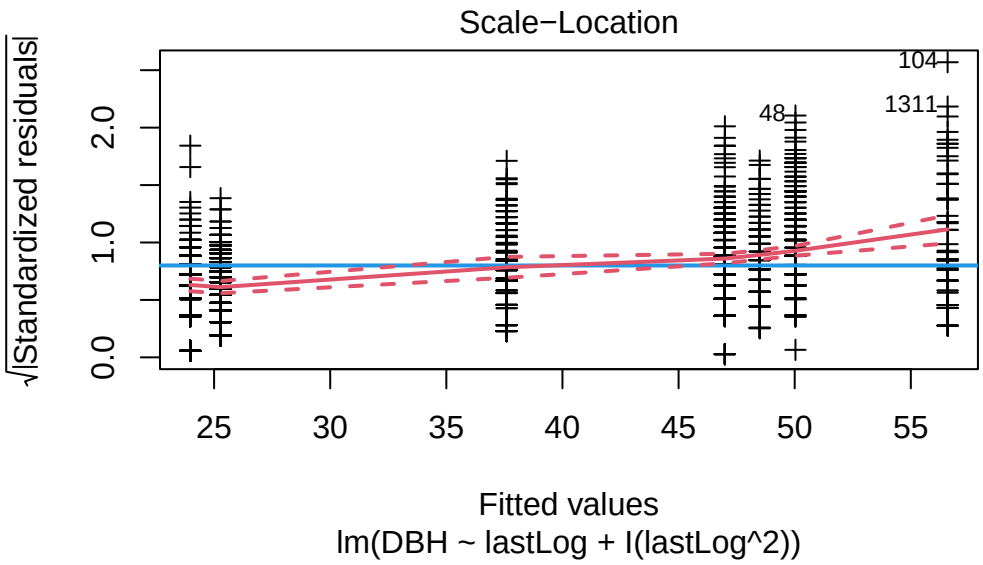


```
# QQ-plot des résidus
qqPlot(residuals(mod_quad), main = "QQ-plot des résidus (modèle quadratique)")
```




```
104 1311
104 1307
```

```r
# Scale-Location plot pour vérifier l'homoscédasticité
plot(mod_quad, which = 3, pch = 3, add.smooth = FALSE)
abline(h = 0.8, col = 4, lwd = 2)

# Moyenne glissante des points
lo <- loess(sqrt(abs(rstandard(mod_quad))) ~ mod_quad$fitted.values)
vFit <- sort(unique(mod_quad$fitted.values))
predLo <- predict(lo, vFit, se = TRUE)
lines(predLo$fit ~ vFit, col = 2, lwd = 2)

# Enveloppe de confiance
nFit <- length(vFit)
ICBonf <- qnorm(1 - 0.05 / 2 / nFit)
lines(predLo$fit + ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
lines(predLo$fit - ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
```



2.6.1 Interprétation des diagnostics pour le modèle quadratique

1. Histogramme des résidus - Symétrie améliorée : Les résidus présentent une **meilleure concentration autour de zéro**, ce qui suggère une distribution plus proche de la normalité. - **Queue droite réduite** : Comparé au modèle linéaire, les queues extrêmes semblent **moins marquées**, indiquant une amélioration.

2. QQ-plot des résidus - Amélioration dans les extrémités : Les écarts aux extrémités sont **moins prononcés** qu'avec le modèle linéaire, bien que les valeurs aberrantes (**104 et 1311**) soient toujours présentes. - **Alignement central satisfaisant** : Les résidus suivent bien la ligne diagonale dans la zone centrale.

3. Scale-Location Plot - Homoscédasticité partiellement améliorée : La pente de la courbe rouge est **moins prononcée**, indiquant une réduction de l'hétéroscédasticité. Cependant, une **légère augmentation de la variance** subsiste pour les valeurs ajustées élevées. - **Valeurs aberrantes** : Les observations **104, 1311 et 48** influencent encore légèrement la variance des résidus.

2.7 Conclusion pour la question 2

L'analyse révèle que le modèle quadratique améliore l'ajustement des données par rapport au modèle linéaire, avec un **R^2 ajusté passant de 69.38% à 69.78%**. Le terme quadratique est significatif (**$p < 0.001$**), indiquant une relation non-linéaire entre `lastLog` et `DBH`, où l'effet négatif de `lastLog` diminue à mesure que sa valeur augmente.

Les diagnostics montrent une **meilleure normalité des résidus** et une **réduction partielle de l'hétéroscédasticité**, bien que des valeurs aberrantes (e.g., **104, 1311**) persistent. Ces observations soulignent l'utilité du modèle quadratique tout en mettant en évidence la nécessité d'un traitement des outliers et d'une potentielle inclusion d'autres variables explicatives pour affiner les prédictions.

---

3 Question 3 : Modèle ANOVA hiérarchique pour le diamètre des chênes

Comme pour les questions précédentes, nous allons commencer par filtrer les **chênes** (**Quercus L., 1753**) à partir de la variable `recherche_esp_lb_nom_plantae`.

3.1 Filtrage et exploration des données

```
# Filtrer pour ne garder que les chênes
data_chenes <- dataIni %>%
  filter(recherche_esp_lb_nom_plantae == "Quercus L., 1753")

# Vérifier un aperçu des données filtrées
head(data_chenes)
```

```
   X releve recherche_esp_lb_nom_plantae cav_basses_presence_cavites
1 63 BLO_27             Quercus L., 1753                         oui
2 64 BLO_27             Quercus L., 1753                         oui
3 65 BLO_27             Quercus L., 1753                         oui
4 66 BLO_27             Quercus L., 1753                         oui
5 67 BLO_27             Quercus L., 1753                         oui
6 68 BLO_27             Quercus L., 1753                         oui
  cara_cavites_circonference cara_cavites_avancement cara_cavites_carie TreeID
1                        195                       5              rouge   27-6
```

```
2                       190                       5              rouge  27-10
3                       146                       6            blanche  27-15
4                       310                       6            blanche  27-22
5                       204                       5          bois dur  27-23
6                       190                       5            blanche  27-34
  DBH Status     long     lati    alti lastLog
1  47 vivant 1.779575 44.00633 256.980    1876
2  40 vivant 1.779790 44.00623 255.576    1876
3  36 vivant 1.779495 44.00628 259.620    1876
4  58 vivant 1.779172 44.00633 257.551    1876
5  50 vivant 1.779261 44.00631 257.252    1876
6  41 vivant 1.779335 44.00599 259.665    1876
```

```r
# Vérifier les colonnes disponibles
colnames(data_chenes)
```

```
 [1] "X"                           "releve"
 [3] "recherche_esp_lb_nom_plantae" "cav_basses_presence_cavites"
 [5] "cara_cavites_circonference"   "cara_cavites_avancement"
 [7] "cara_cavites_carie"           "TreeID"
 [9] "DBH"                          "Status"
[11] "long"                         "lati"
[13] "alti"                         "lastLog"
```

3.1.1 Résumé des résultats :

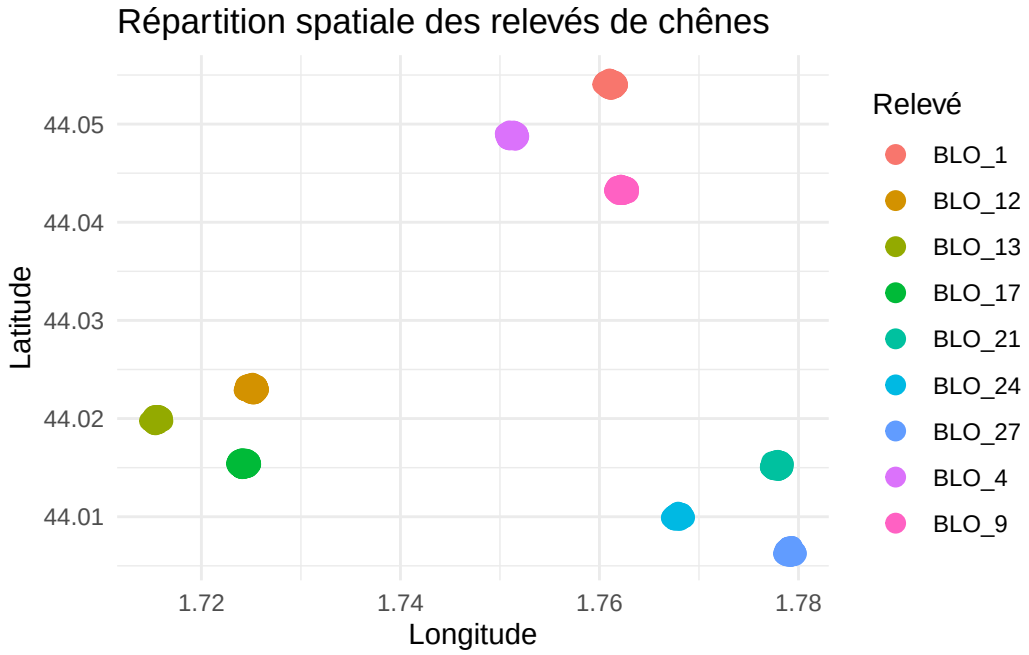
1. **Colonnes pertinentes** :
   - `releve` : Identifie les relevés individuels.
   - `long`, `lati` : Coordonnées géographiques des relevés.
   - `DBH` : Diamètre des arbres, la variable cible pour le modèle ANOVA.
2. **Nombre de relevés** :
   - Les relevés filtrés incluent : **BLO_1, BLO_4, BLO_9, BLO_12, BLO_13, BLO_17, BLO_21, BLO_24, BLO_27**.

L'apperçu des données ne permet pas de visualiser quels groupes forment les triangles. Pour cela, il est nécessaire de les représenter graphiquement afin de comprendre la relation entre les relevés.

3.2 Visualisation spatiale des relevés

1. **Définir les triangles** :
   - Chaque relevé sera associé à un triangle selon sa position géographique.
2. **Créer une visualisation** :
   - Un graphique spatial sera généré pour représenter les relevés colorés par triangle.

```
ggplot(data_chenes, aes(x = long, y = lati, color = releve)) +
  geom_point(size = 3) +
  theme_minimal() +
  labs(title = "Répartition spatiale des relevés de chênes",
       x = "Longitude", y = "Latitude", color = "Relevé")
```



En analysant le graphique ci-dessus, nous pouvons identifier les trois triangles. Il est maintenant temps de les regrouper.

3.3 Regroupement et visualisation de la structure hiérarchique

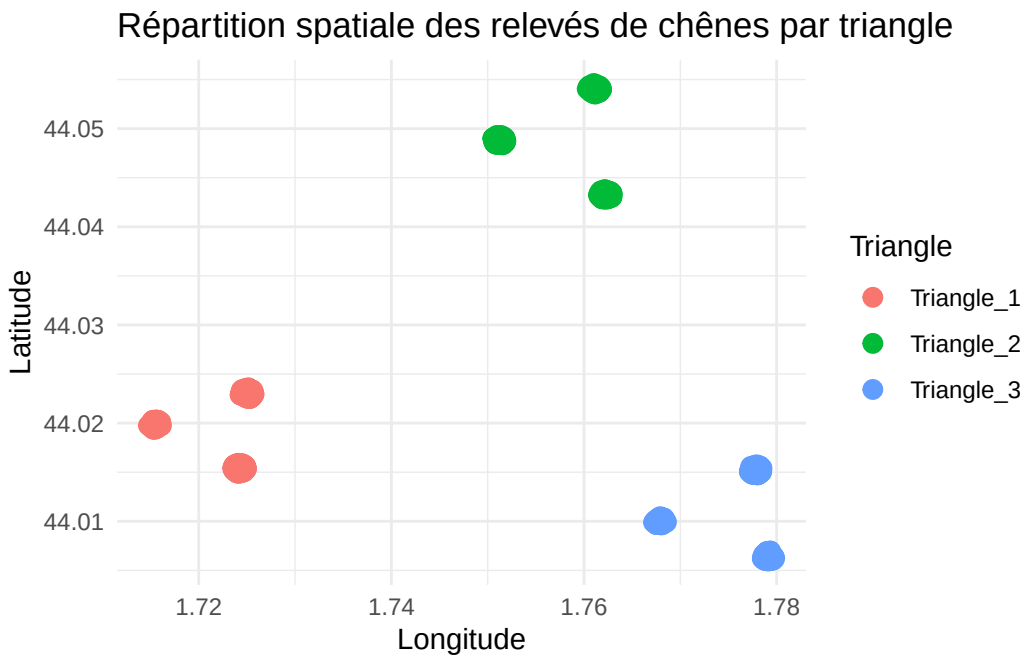
Nous allons regrouper les relevés en trois triangles distincts selon leur position géographique. Voici la correspondance que nous avons établie :

- **Triangle 1** : BLO_12, BLO_13, BLO_17
- **Triangle 2** : BLO_1, BLO_4, BLO_9
- **Triangle 3** : BLO_21, BLO_24, BLO_27

```
# Définir les groupes de triangles
data_chenes <- data_chenes %>%
  mutate(
    Triangle = case_when(
      releve %in% c("BLO_17", "BLO_12", "BLO_13") ~ "Triangle_1",
      releve %in% c("BLO_4", "BLO_9", "BLO_1") ~ "Triangle_2",
      releve %in% c("BLO_21", "BLO_24", "BLO_27") ~ "Triangle_3",
      TRUE ~ NA_character_
    )
  )
```

Maintenant il faut tracer une carte des relevés, colorés par triangle, pour valider cette structure.

```
library(ggplot2)

# Visualisation de la répartition des triangles
ggplot(data_chenes, aes(x = long, y = lati, color = Triangle)) +
  geom_point(size = 3) +
  labs(
    title = "Répartition spatiale des relevés de chênes par triangle",
    x = "Longitude",
    y = "Latitude",
    color = "Triangle"
  ) +
  theme_minimal()
```



3.4 Analyse de la variance (ANOVA) selon les triangles

Nous allons maintenant ajuster un modèle ANOVA pour expliquer le diamètre des chênes (**DBH**) en utilisant les **triangles** comme base de définition des sous-populations. Cela permettra d'évaluer si la variation du diamètre est significativement influencée par les groupes spatiaux définis par les triangles.

```r
# Ajuster le modèle ANOVA avec Triangle comme facteur
mod_anova_triangle <- aov(DBH ~ Triangle, data = data_chenes)

# Résumé du modèle ANOVA
summary(mod_anova_triangle)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Triangle      2 122018   61009   548.1 <2e-16 ***
Residuals  1459 162393     111
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
5 observations deleted due to missingness
```

3.4.1 Interprétation des résultats de l'ANOVA

1. Résultats principaux : - **Effets du facteur Triangle** : - Le facteur **Triangle** est **hautement significatif** avec une p-value $<$ 2e-16. - Cela signifie que le diamètre des chênes (**DBH**) varie significativement entre les trois triangles.

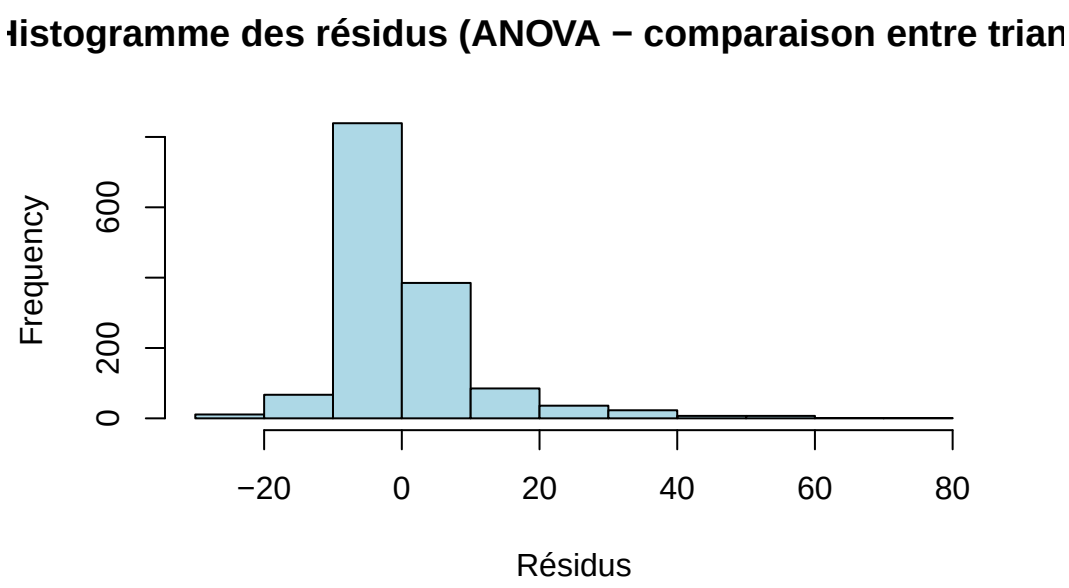
- **F-statistic** :
  - La valeur de $\mathbf{F = 548.1}$ indique un effet très fort des triangles sur le diamètre des chênes.
- **Variance expliquée** :
  - La somme des carrés pour le facteur **Triangle** (122018) représente une proportion importante de la variance totale (122018 + 162393), ce qui montre que les triangles expliquent une grande partie de la variation dans les diamètres.

3.5 Diagnostic des résidus pour le modèle ANOVA

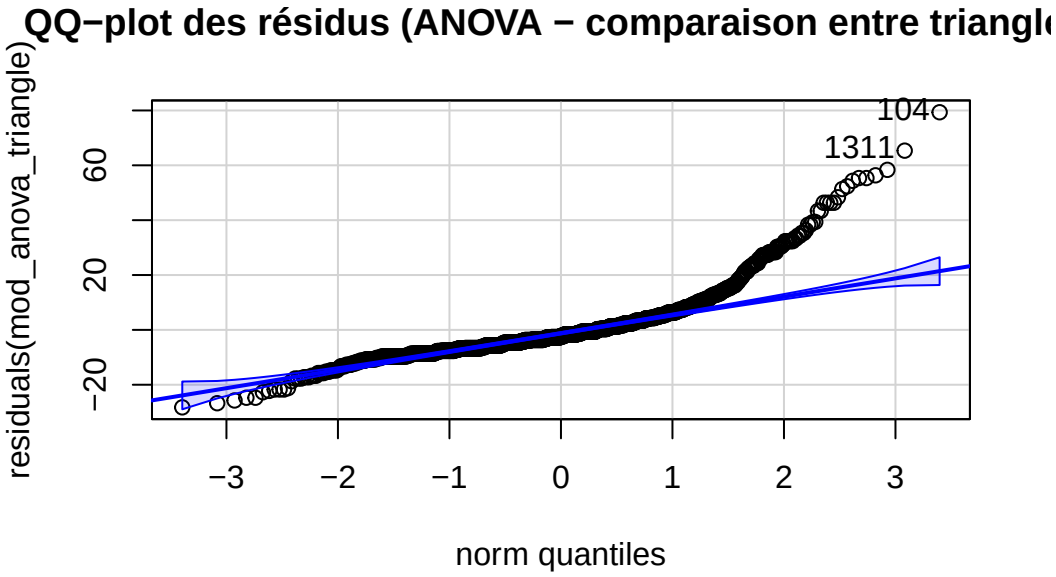
Pour garantir la validité des résultats de l'ANOVA, nous devons vérifier que les hypothèses du modèle sont respectées. Nous allons donc analyser :
1. **La normalité des résidus** à l'aide d'un histogramme et d'un QQ-plot.
2. **L'homoscédasticité** (égalité des variances) à l'aide d'un Scale-Location Plot et du test de Levene.

Ces analyses nous permettront d'évaluer si l'ANOVA est appropriée ou si des ajustements sont nécessaires.

```
# Histogramme des résidus du modèle quadratique
hist(residuals(mod_anova_triangle),
     main = "Histogramme des résidus (ANOVA - comparaison entre triangles)",
     xlab = "Résidus",
     col = "lightblue")
```

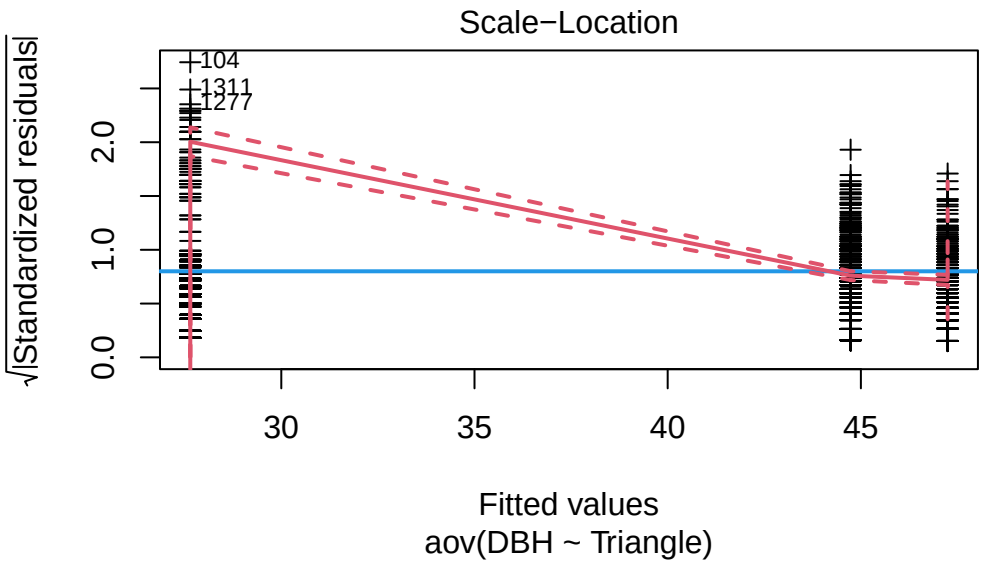


```
# QQ-plot des résidus
qqPlot(residuals(mod_anova_triangle), main = "QQ-plot des résidus (ANOVA - comparaison entre
```




```
104 1311
104 1307
```

```r
# Scale-Location plot pour vérifier l'homoscédasticité
plot(mod_anova_triangle, which = 3, pch = 3, add.smooth = FALSE)
abline(h = 0.8, col = 4, lwd = 2)

# Moyenne glissante des points
lo <- loess(sqrt(abs(rstandard(mod_anova_triangle))) ~ mod_anova_triangle$fitted.values)
vFit <- sort(unique(mod_anova_triangle$fitted.values))
predLo <- predict(lo, vFit, se = TRUE)
lines(predLo$fit ~ vFit, col = 2, lwd = 2)

# Enveloppe de confiance
nFit <- length(vFit)
ICBonf <- qnorm(1 - 0.05 / 2 / nFit)
lines(predLo$fit + ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
lines(predLo$fit - ICBonf * predLo$se.fit ~ vFit, col = 2, lwd = 2, lty = "dashed")
```



Pour garantir la validité des résultats de l'ANOVA, les hypothèses fondamentales du modèle ont été analysées.

3.5.1 Interprétation des résidus de l'ANOVA selon les triangles

1. **Histogramme des résidus**
   - Les résidus sont concentrés autour de zéro, ce qui est généralement un bon indicateur.
   - Cependant, une asymétrie légère à droite est observée, indiquant des écarts potentiels dans les valeurs extrêmes.
2. **QQ-plot des résidus**
   - Les points sont alignés sur la ligne diagonale au centre, confirmant que la normalité est respectée pour la majorité des résidus.
   - Des écarts significatifs aux extrémités révèlent la présence de valeurs aberrantes qui pourraient influencer les résultats.
3. **Scale-Location Plot**

- La variance des résidus est globalement homogène, mais le comportement des résidus pour les valeurs ajustées basses (à gauche) est inhabituel.

- La pente descendante indique une possible influence des observations extrêmes (104, 1311 et 1307), nécessitant une vérification approfondie.

- Ce comportement pourrait refléter une hétéroscédasticité localisée ou un effet spécifique de ces points.

3.6 Teste du modèle emboîté : Comparaison entre les modèles

Maintenant que nous avons ajusté un modèle ANOVA le facteur **triangle** et validé les hypothèses du modèle, nous allons comparer ce modèle à un modèle plus simple (sans les groupes définis par les triangles) pour vérifier si l'agrégation des données par triangles améliore significativement l'ajustement. Pour ce faire, nous allons effectuer un test F pour comparer les deux modèles.

Ajustement du modèle simple (sans les triangles) Nous allons d'abord ajuster un modèle plus simple, qui ne prend pas en compte les triangles, mais seulement l'effet global de la moyenne sur le diamètre des chênes (modèle sans facteur).

3.6.1 Ajustement du modèle simple (sans les triangles)

Nous allons d'abord ajuster un modèle plus simple, qui ne prend pas en compte les **triangles**, mais seulement l'effet global de la moyenne sur le diamètre des chênes (modèle sans facteur).

```
# Ajuster un modèle simple sans facteur 'Triangle'
mod_simple <- aov(DBH ~ 1, data = data_chenes)

# Résumé du modèle simple
summary(mod_simple)
```

```
              Df Sum Sq Mean Sq F value Pr(>F)
Residuals   1461 284411   194.7
5 observations deleted due to missingness
```

3.6.2 Comparaison des modèles avec un test emboîté

Ensuite, nous utiliserons la fonction anova() pour effectuer un test F entre le modèle complet (avec triangles) et le modèle simple (sans triangle). Cela nous permettra de déterminer si l'inclusion du facteur **triangle** améliore significativement l'ajustement du modèle.

```
# Comparer les deux modèles avec un test F
anova(mod_simple, mod_anova_triangle)
```

```
Analysis of Variance Table

Model 1: DBH ~ 1
Model 2: DBH ~ Triangle
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1461 284411
2   1459 162393  2    122018 548.13 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.7 Conclusion pour la question 3

L'agrégation des relevés en trois groupes définis par les triangles est non seulement logique mais aussi statistiquement justifiée, car elle améliore de manière significative l'ajustement du modèle.

---

4 Question 4 : Modèle mixte pour analyser l'effet de 'lastLog' sur les relevés

Dans cette section, nous allons examiner si la variable **'lastLog'** (date de dernière coupe massive) explique l'effet **relevé** en utilisant un **modèle mixte linéaire**.

4.1 Ajustement du modèle mixte avec lastLog´

```r
# Ajuster le modèle mixte avec lastLog comme effet fixe et releve comme effet aléatoire
mod_mixed <- lmer(
  I(log10(DBH)) ~ lastLog + alti + I(alti^2) + (1 | releve),
  data = data_chenes,
  subset = NULL,  # Utilise toutes les données (aucun sous-ensemble)
  weights = NULL,  # Pas de pondération
  na.action = na.omit,  # Supprime les valeurs manquantes
  offset = NULL  # Pas de décalage
)

# Résumé du modèle mixte
summary(mod_mixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: I(log10(DBH)) ~ lastLog + alti + I(alti^2) + (1 | releve)
   Data: data_chenes

REML criterion at convergence: -3058.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.9849 -0.6045 -0.0004  0.6431  3.9784

Random effects:
 Groups   Name        Variance  Std.Dev.
 releve   (Intercept) 0.0007626 0.02762
 Residual             0.0068068 0.08250
Number of obs: 1461, groups:  releve, 9

Fixed effects:
             Estimate Std. Error t value
(Intercept)  6.846e+00  4.005e-01  17.096
lastLog     -2.731e-03  2.322e-04 -11.760
alti        -1.732e-04  1.083e-03  -0.160
I(alti^2)    1.819e-08  1.499e-06   0.012

Correlation of Fixed Effects:
          (Intr) lastLg alti
lastLog   -0.904
alti       0.024 -0.446
I(alti^2) -0.058  0.469 -0.994
```

```
fit warnings:
Some predictor variables are on very different scales: consider rescaling
```

4.1.1 Interprétation des résultats du modèle mixte

1. **Effets aléatoires** :
   - La variance associée à l'effet aléatoire (**releve**) est très faible (**0.0007626**), ce qui indique que l'effet de regroupement par relevé a peu d'impact sur la variabilité totale de la réponse.
   - La variance résiduelle (**0.0068068**) est beaucoup plus importante, suggérant que la majeure partie de la variabilité est due à des facteurs non modélisés ou à des erreurs aléatoires.
2. **Effets fixes** :
   - **Intercept (6.846)** : Représente la valeur moyenne du **log10(DBH)** lorsque toutes les variables explicatives sont à

     0.
   - **lastLog (-0.002731, $p < 0.001$)** : Cet effet est significatif. Il indique une légère diminution de **log10(DBH)** lorsque la variable `lastLog` augmente.
   - **alti et I(alti^2)** : Ces deux termes ne sont pas significatifs ($p > 0.05$), suggérant que l'altitude n'a pas d'effet significatif sur la réponse dans ce modèle.
3. **Critère REML** :
   - Le critère REML à convergence est de **-3058.8**. Ce score peut être utilisé pour comparer plusieurs modèles (par exemple, avec ou sans certaines variables).
4. **Avertissements** :
   - L'avertissement mentionne que certaines variables explicatives ont des échelles très différentes. Il est conseillé de les standardiser pour améliorer l'estimation des coefficients et réduire la colinéarité.

4.2 Standardisation des variables explicatives

Avant de poursuivre avec le modèle, effectuons la standardisation de **lastLog** et **alti** afin de corriger les avertissements liés aux échelles différentes.

```r
# Standardiser les variables explicatives
data_chenes <- data_chenes %>%
  mutate(
    lastLog_std = scale(lastLog),
    alti_std = scale(alti)
  )
```

Nous avons de nouveau ajusté le modèle en intégrant les versions standardisées de **lastLog** et **alti**.

```r
# Ajuster le modèle mixte avec variables standardisées
mod_mixed_std <- lmer(
  I(log10(DBH)) ~ lastLog_std + alti_std + I(alti_std^2) + (1 | releve),
  data = data_chenes,
  subset = NULL,  # Utilise toutes les données (aucun sous-ensemble)
  weights = NULL,  # Pas de pondération
  na.action = na.omit,  # Supprime les valeurs manquantes
  offset = NULL  # Pas de décalage
)
# Résumé du modèle
summary(mod_mixed_std)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: I(log10(DBH)) ~ lastLog_std + alti_std + I(alti_std^2) + (1 |
    releve)
   Data: data_chenes

REML criterion at convergence: -3091.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.9849 -0.6045 -0.0004  0.6431  3.9784

Random effects:
 Groups   Name        Variance  Std.Dev.
 releve   (Intercept) 0.0007626 0.02762
 Residual             0.0068068 0.08250
Number of obs: 1461, groups:  releve, 9

Fixed effects:
                Estimate Std. Error t value
```

```
(Intercept)    1.536e+00  1.207e-02 127.189
lastLog_std   -1.458e-01  1.240e-02 -11.760
alti_std      -1.002e-02  9.386e-03  -1.068
I(alti_std^2)  7.026e-05  5.790e-03   0.012

Correlation of Fixed Effects:
            (Intr) lstLg_ alt_st
lastLog_std  0.157
alti_std     0.289 -0.147
I(lt_std^2) -0.447  0.469 -0.619
```

4.2.1 Interprétation des deux modèles

Modèle sans standardisation

1. **Effets aléatoires :** - La variance de l'effet aléatoire pour `releve` est très faible (0.0007626), indiquant que l'effet de groupe dû aux relevés est négligeable. - La variance résiduelle est beaucoup plus importante (0.0068068), ce qui suggère que la majorité de la variabilité n'est pas expliquée par l'effet aléatoire.
2. **Effets fixes :**
   - **lastLog** : Effet significatif ($p < 0.001$) avec une légère influence négative sur **log10(DBH)** ((-0.002731)).
   - **alti** et **I(alti^2)** : Pas significatifs, indiquant que l'altitude n'explique pas de manière importante la variabilité de DBH.
3. **Critère REML :**
   - Le critère REML à convergence est de **-3058.8**.

Modèle avec standardisation

1. **Effets aléatoires :** Résultats identiques au modèle sans standardisation, confirmant que l'effet aléatoire `releve` reste négligeable.
2. **Effets fixes :**
   - **lastLog_std** : Effet toujours significatif ($p < 0.001$) avec une influence négative ((-0.1458)), confirmant l'importance de cette variable.
   - **alti_std** et **I(alti_std^2)** : Toujours non significatifs.
3. **Critère REML :**

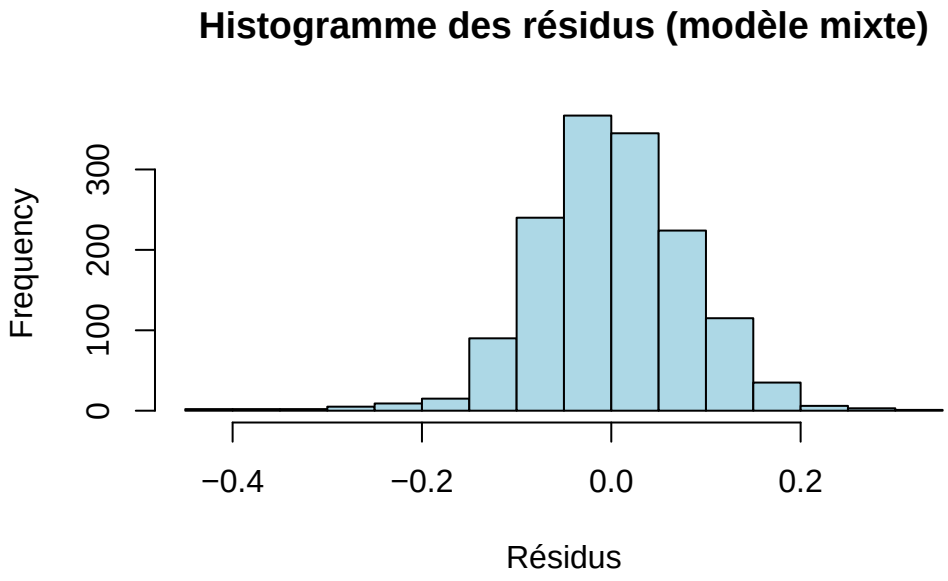
- Le critère REML est amélioré ((-3091.6) contre (-3058.8)), indiquant que la standardisation améliore légèrement l'ajustement global du modèle.

- La variable `lastLog` (même standardisée) a un effet significatif sur **log10(DBH)**, tandis que l'altitude n'en a pas.
- La standardisation a permis d'améliorer l'interprétation et la précision des coefficients, ainsi que l'ajustement global du modèle.

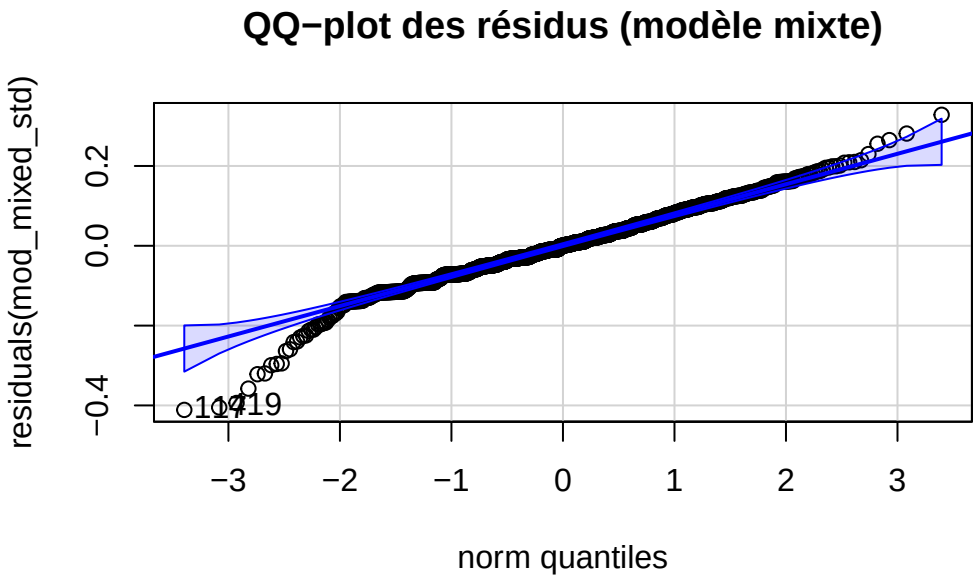
4.3 Vérification des hypothèses du modèle mixte

Avant d'interpréter les résultats du modèle, il est essentiel de vérifier si les hypothèses de **normalité des résidus** et **homoscédasticité** sont respectées. Pour cela, nous analyserons l'histogramme des résidus, le QQ-plot et le Scale-Location plot. Ces diagnostics permettront d'évaluer la validité des conclusions issues du modèle mixte.

```
# Histogramme des résidus du modèle quadratique
hist(residuals(mod_mixed_std),
     main = "Histogramme des résidus (modèle mixte)",
     xlab = "Résidus",
     col = "lightblue")
```



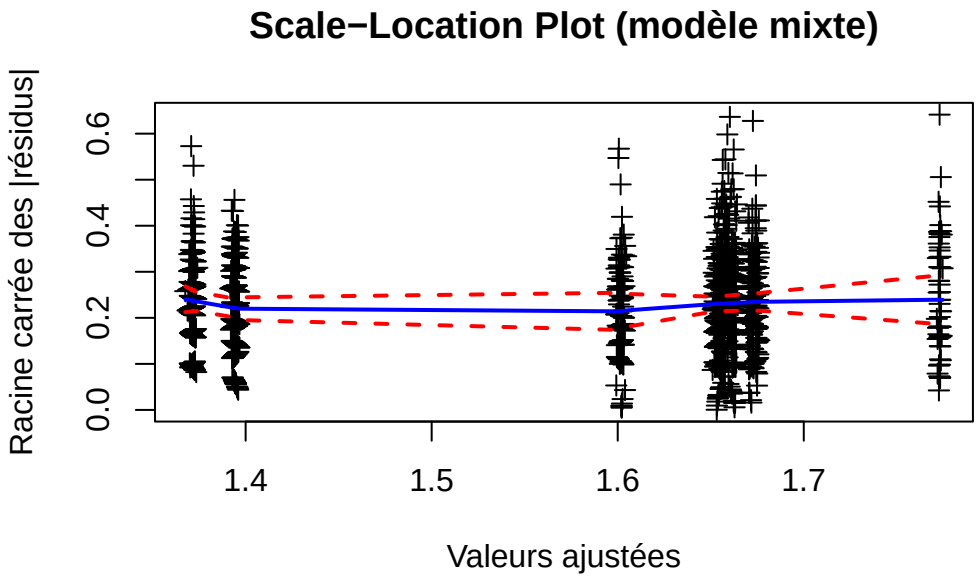
```r
# QQ-plot des résidus
qqPlot(residuals(mod_mixed_std), main = "QQ-plot des résidus (modèle mixte)")
```




```
[1] 117 419
```

```r
# Extraire les résidus et les valeurs ajustées
residus <- residuals(mod_mixed_std)  # Résidus bruts
valeurs_ajustees <- fitted(mod_mixed_std)  # Valeurs ajustées du modèle

# Tracer le Scale-Location Plot
plot(valeurs_ajustees, sqrt(abs(residus)),
     main = "Scale-Location Plot (modèle mixte)",
     xlab = "Valeurs ajustées",
     ylab = "Racine carrée des |résidus|",
     pch = 3, col = "black")

# Ajouter la ligne bleue horizontale attendue
```

```r
abline(h = 0.8, col = "blue", lwd = 2)

# Moyenne glissante des points
lo <- loess(sqrt(abs(residus)) ~ valeurs_ajustees)
vFit <- sort(unique(valeurs_ajustees))
predLo <- predict(lo, vFit, se = TRUE)

# Tracer la moyenne glissante
lines(vFit, predLo$fit, col = "blue", lwd = 2)

# Enveloppe de confiance autour de la moyenne glissante
ICBonf <- qnorm(1 - 0.05 / 2 / length(vFit))
lines(vFit, predLo$fit + ICBonf * predLo$se.fit, col = "red", lwd = 2, lty = "dashed")
lines(vFit, predLo$fit - ICBonf * predLo$se.fit, col = "red", lwd = 2, lty = "dashed")
```



4.3.1 Interprétation de la normalité et de la homoscédasticité pour le modèle mixte

1. Ajustement des modèles :

- **lastLog** a un effet significatif et négatif sur log10(DBH), confirmant que cette variable explique une partie de l'effet entre relevés. Cela indique que plus la date de dernière coupe est ancienne, plus le diamètre (DBH) tend à être important.

- La standardisation améliore l'interprétation et la stabilité des coefficients.

2. Vérification des hypothèses :

- **Histogramme des résidus :** Les résidus suivent une distribution globalement normale.
- **QQ-Plot :** Bonne adéquation avec la normalité, excepté quelques valeurs extrêmes.
- **Scale-Location Plot :** Les variances résiduelles sont homogènes, confirmant l'homoscédasticité.

4.4 Conclusion pour la question 4

- La variable `lastLog` explique significativement l'effet relevé, validant son inclusion dans le modèle. Cependant, l'altitude (`alti` et `alti^2`) ne joue pas un rôle significatif dans cette analyse.
- Le modèle mixte permet de mieux capturer l'effet entre relevés par rapport à un modèle linéaire classique.

---

5 Question 5 : Modèle linéaire généralisé binomial avec le diamètre des arbres

Dans cette dernière analyse, nous allons étendre le modèle linéaire généralisé binomial étudié en cours en ajoutant **le diamètre des arbres (DBH)** comme variable explicative de la **présence de cavités basses**.

5.1 Transformation de la variable cible cavPA

On commence par ajuster le modèle sans le diamètre afin de pouvoir comparer les effets par la suite. Pour cela il faut transformer le `cav_basses_presence_cavites` en binaire (cavPA)

```
# Transformer `cav_basses_presence_cavites en binaire` (cavPA)
data_chenes <- data_chenes %>%
  mutate(cavPA = ifelse(cav_basses_presence_cavites == "oui", 1, 0))

# Vérifier la conversion
cat('Vérifier la conversion : ')
```

```
Vérifier la conversion :
```

```
table(data_chenes$cavPA, useNA = "always")
```

```
   0    1 <NA>
1293  174    0
```

```
cat('Vérifier la répartition de cavPA :')
```

```
Vérifier la répartition de cavPA :
```

```
table(data_chenes$releve, data_chenes$cavPA)
```

```
          0   1
  BLO_1   89  37
  BLO_12 320   7
  BLO_13  38  24
  BLO_17 346   4
  BLO_21  99  34
  BLO_24  54  20
  BLO_27  93  18
  BLO_4  109   7
  BLO_9  145  23
```

5.1.1 Interprétation de la transformation de `cavPA` par `releve`

- Le tableau montre que chaque relevé contient **au moins un 0 et un 1**, ce qui signifie que notre modèle binomial pourra bien converger **sans problème de séparation complète**.
- Certains relevés (ex: **BLO_12 et BLO_17**) ont une très faible quantité de `1`, ce qui peut causer un **déséquilibre** dans l'estimation.

- D'autres relevés (ex: **BLO_1, BLO_21, BLO_13, BLO_9**) ont une meilleure répartition.

5.2 Ajustement du modèle initial (sans DBH)

```r
# Filtrer les données pour exclure les observations avec des valeurs manquantes
data_chenes_complete <- na.omit(data_chenes[, c("cavPA", "alti", "releve", "DBH")])


# Ajuster le modèle binomial AVEC DBH
mod_bin_cav <- glm(
  cavPA ~ alti + releve,
  data = data_chenes_complete,
  family = binomial(link = "logit")
)

# Résumé du modèle
summary(mod_bin_cav)
```

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

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   7.16431    5.12412   1.398 0.162066
alti         -0.01712    0.01092  -1.569 0.116741
releveBLO_12 -5.23442    1.52156  -3.440 0.000581 ***
releveBLO_13  0.34906    0.33231   1.050 0.293541
releveBLO_17 -6.30756    1.81975  -3.466 0.000528 ***
releveBLO_21 -4.17195    2.55434  -1.633 0.102410
releveBLO_24 -3.11507    1.93436  -1.610 0.107314
releveBLO_27 -4.39444    2.33271  -1.884 0.059587 .
releveBLO_4  -4.14150    1.51203  -2.739 0.006162 **
releveBLO_9  -3.56711    1.68480  -2.117 0.034240 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1066.88  on 1460  degrees of freedom
Residual deviance:  865.54  on 1451  degrees of freedom
AIC: 885.54
```

```
Number of Fisher Scoring iterations: 7
```

5.2.1 Interprétation du modèle sans DBH (`mod_bin_cav`)

1. **Effets des variables explicatives** :
   - **Relevé (`releve`)** :
     - Plusieurs relevés ont un effet significatif sur la probabilité de présence de cavités basses (ex. BLO_12, BLO_17, BLO_4 et BLO_9). Ces sites montrent une probabilité significativement plus faible de présence de cavités basses.
     - Quelques relevés (ex. BLO_27) sont marginalement significatifs ($p \quad 0.06$), indiquant des effets potentiels à vérifier.
   - **Altitude (`alti`)** :
     - Son effet n'est pas significatif ($p = 0.11$), ce qui indique qu'elle n'a pas d'influence détectable sur la présence de cavités basses dans ce modèle.
2. **Qualité du modèle** :
   - Une **deviance résiduelle** de `865.54` indique que le modèle explique une partie de la variabilité observée.
   - L'**AIC** de `885.54` sert de référence pour évaluer si l'ajout de nouvelles variables, comme `DBH`, apporte des améliorations.

5.3 Ajustement du modèle avec DBH

```r
# Ajustement du modèle binomial avec DBH
mod_bin_cav_DBH <- glm(
  cavPA ~ alti + releve + DBH,
  data = data_chenes_complete,
  family = binomial(link = "logit")
)

# Résumé du modèle
summary(mod_bin_cav_DBH)
```

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

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
```

5.3.1 Interprétation du modèle avec DBH (`mod_bin_cav_DBH`)

1. **Effets des variables explicatives** :
   - **Relevé (`releve`)** :
     - L'effet des relevés reste inchangé par rapport au modèle sans DBH, confirmant leur rôle clé dans l'hétérogénéité spatiale.
   - **Altitude (`alti`)** :
     - Toujours non significative, son inclusion dans le modèle n'apporte aucune amélioration.
   - **Diamètre (`DBH`)** :

– Non significatif ($p = 0.70$), il n'ajoute aucune information pertinente pour expliquer la présence de cavités basses.

2. **Qualité du modèle** :
   - Une **deviance résiduelle** quasi identique (`865.4`) à celle du modèle sans DBH, avec une légère augmentation de l'**AIC** (`887.4`), indique que l'ajout de `DBH` n'améliore ni l'ajustement ni la prédiction du modèle.

5.4 Comparaison des modèles avec et sans DBH

Pour évaluer si l'ajout du diamètre améliore significativement l'ajustement du modèle, nous comparons les modèles avec et sans `DBH` à l'aide d'un **test du rapport de vraisemblance** :

```
# Comparaison des modèles avec et sans DBH
anova(mod_bin_cav, mod_bin_cav_DBH, test = "Chisq")
```

```
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
```

5.4.1 Interprétation de la comparaison des modèles avec et sans DBH

Le test du rapport de vraisemblance montre que l'ajout de `DBH` **n'améliore pas significativement** l'ajustement du modèle.

1. **Valeur du test**
   - **Deviance résiduelle du modèle sans DBH** : `865.54`
   - **Deviance résiduelle du modèle avec DBH** : `865.40`
   - **Différence de déviance** : `0.14472` avec **1 degré de liberté**
   - **p-value = 0.7036**

2. **Interprétation**

   - La **p-value élevée (> 0.05)** indique que l'ajout de `DBH` **n'apporte pas d'amélioration statistiquement significative** au modèle.

   - La réduction de la déviance est **très faible**, confirmant que `DBH` **n'explique pas la présence de cavités basses**.

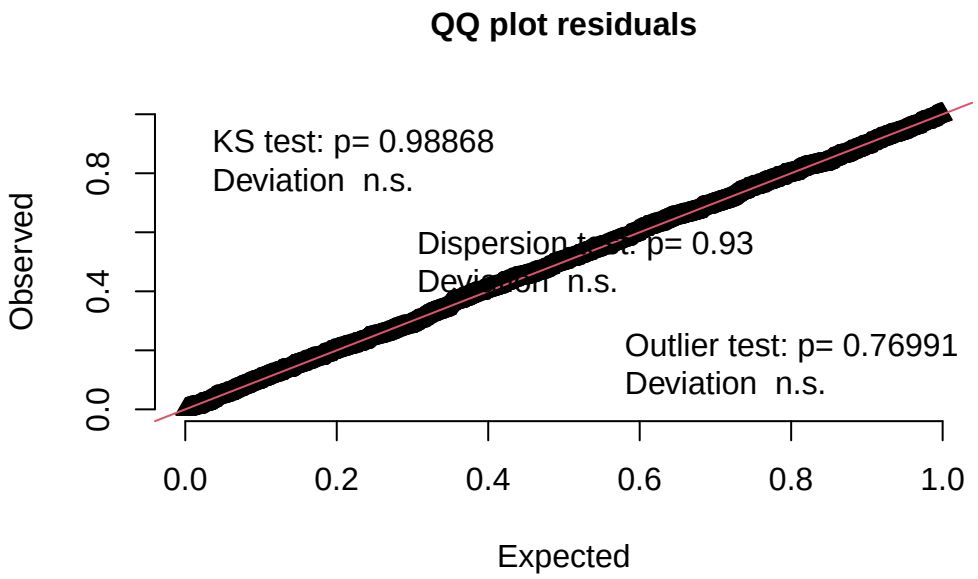
Conclusion

- L'**effet du relevé est confirmé** comme un facteur important, avec des différences significatives entre sites.
- L'**altitude ne joue pas un rôle significatif**.
- **Le diamètre (`DBH`) ne contribue pas significativement à l'explication de la présence de cavités basses**.
- **L'ajout de DBH n'améliore pas le modèle**, ce qui suggère que d'autres variables explicatives pourraient être plus pertinentes pour comprendre la présence de cavités.

5.5 Vérification des hypothèses du modèle avec DHARMa

Afin d'assurer la validité des conclusions du modèle binomial généralisé, il est essentiel de vérifier si ses hypothèses sont respectées. Nous utilisons le package **DHARMa** pour analyser les résidus simulés et détecter d'éventuelles violations des hypothèses, notamment en ce qui concerne **l'uniformité, la dispersion et l'indépendance des résidus**. Ces tests permettront d'évaluer la fiabilité du modèle avant d'interpréter les résultats finaux.

```
# Transformer les résidus avec DHARMa
residusTransfo <- simulateResiduals(mod_bin_cav_DBH, n = 1000)

# Vérification de la distribution uniforme des résidus
testUniformity(residusTransfo)
```




```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  simulationOutput$scaledResiduals
D = 0.011663, p-value = 0.9887
alternative hypothesis: two-sided
```

5.5.1 Interprétation du test d'uniformité des résidus

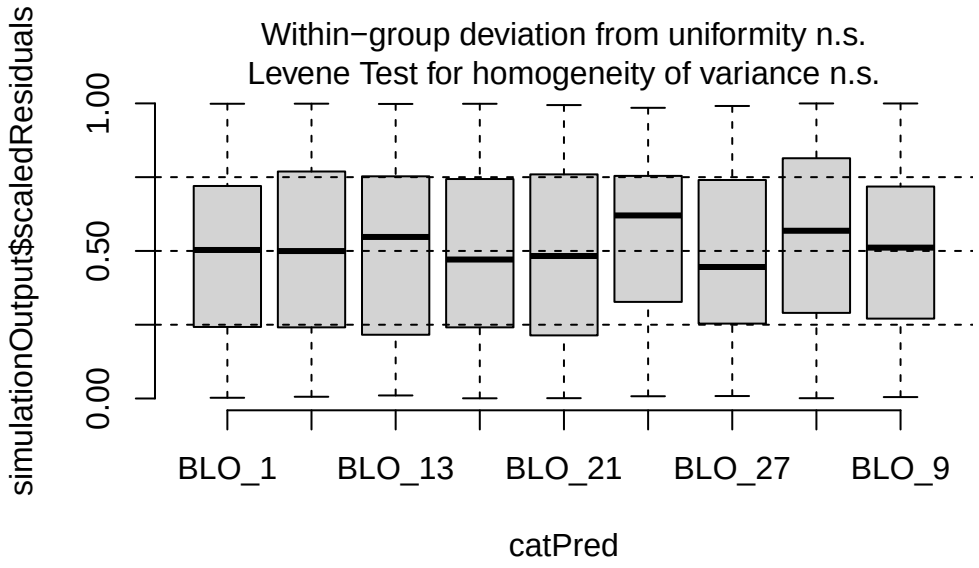
1. **Graphique QQ des résidus** :
   - La ligne des points observés suit de très près la diagonale attendue, ce qui indique que les résidus sont bien distribués de manière uniforme, comme attendu sous les hypothèses du modèle binomial.
2. **Tests statistiques associés** :
   - **Test KS (Kolmogorov-Smirnov)** : *p-value = 0.9887*.
     - Une p-value élevée (> 0.05) signifie que l'on ne rejette pas l'hypothèse nulle selon laquelle les résidus suivent une distribution uniforme. Cela indique qu'aucune déviation significative par rapport à l'uniformité n'a été détectée avec ce test.
   - **Test de dispersion** : $p = 0.93$.

  - Cela montre que les résidus ne sont pas sur- ou sous-dispersés.
- **Test des valeurs aberrantes (outliers)** : $p = 0.76991$.
  - Aucune valeur aberrante significative n'est détectée dans le modèle.

Conclusion : Les résultats montrent que le modèle respecte les hypothèses d'uniformité, de dispersion et d'absence d'outliers. Cela indique que les conclusions tirées du modèle sont fiables et ne souffrent pas de biais liés à des résidus mal ajustés

---

```
# Vérification de la distribution uniforme des résidus par relevé
testCategorical(residusTransfo, mod_bin_cav_DBH$model$releve)
```




```
$uniformity
$uniformity$details
catPred: BLO_1

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.074492, p-value = 0.4866
alternative hypothesis: two-sided
```

```
------------------------------------------------------------
catPred: BLO_12

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.028324, p-value = 0.9567
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_13

    Exact one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.099827, p-value = 0.5441
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_17

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.035155, p-value = 0.783
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_21

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.074815, p-value = 0.4509
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_24

    Exact one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
```

```
D = 0.13318, p-value = 0.1322
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_27

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.058882, p-value = 0.8362
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_4

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.085654, p-value = 0.3624
alternative hypothesis: two-sided

------------------------------------------------------------
catPred: BLO_9

    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dd[x, ]
D = 0.036664, p-value = 0.9776
alternative hypothesis: two-sided


$uniformity$p.value
[1] 0.4865661 0.9567423 0.5441448 0.7829592 0.4509148 0.1321648 0.8362248
[8] 0.3624053 0.9776274

$uniformity$p.value.cor
[1] 1 1 1 1 1 1 1 1 1


$homogeneity
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    8  0.3681 0.9376
```

```
1452
```

5.5.2 Interprétation de la vérification par relevé

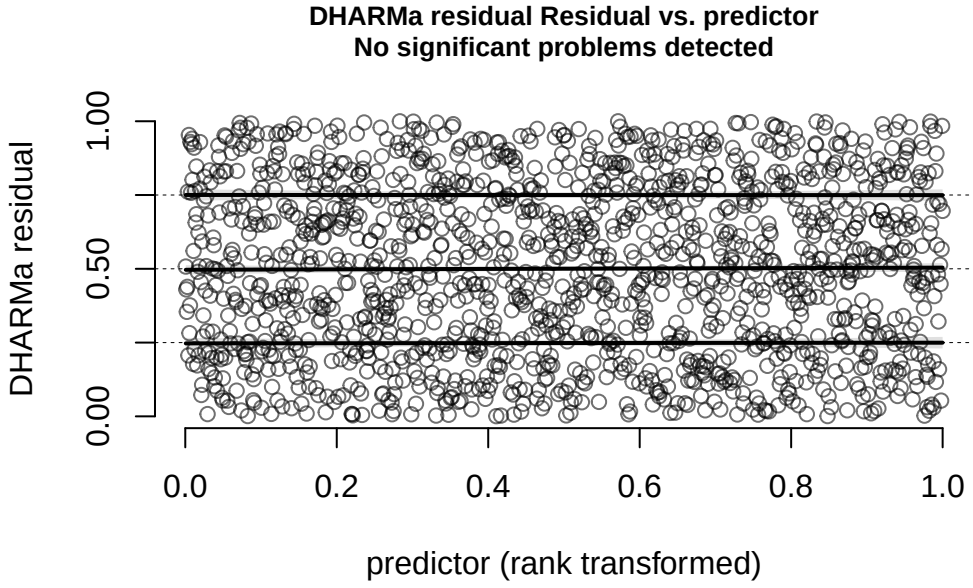
1. **Tests d'uniformité des résidus par relevé** :
   - Les p-values des tests Kolmogorov-Smirnov pour chaque relevé (ex. BLO_1, BLO_13, etc.) sont toutes **supérieures à 0.05**, ce qui indique que la distribution des résidus au sein de chaque relevé ne dévie pas de manière significative d'une distribution uniforme.
   - **p-value corrigée** (multiples tests) : toutes les valeurs corrigées restent à `1`, ce qui confirme que les résidus sont uniformes même après correction pour tests multiples.
2. **Homogénéité de la variance entre les relevés** :
   - Le test de Levene donne une **p-value = 0.9376**, ce qui indique que les variances des résidus sont homogènes entre les relevés. Aucune hétérogénéité marquée des résidus n'est observée.
3. **Graphique des résidus par relevé** :
   - Les boîtes à moustaches montrent une distribution uniforme des résidus dans chaque relevé. Aucune déviation significative ou valeur aberrante n'apparait.

Conclusion :

Les résultats confirment que :

- Les résidus sont uniformément distribués au sein de chaque relevé.
- Les variances des résidus sont homogènes entre les relevés. Ces éléments indiquent que le modèle respecte les hypothèses essentielles de validité concernant la répartition des résidus par catégories.

```r
# Vérification de l'homogénéité des résidus en fonction de l'altitude
testQuantiles(residusTransfo, predictor = mod_bin_cav_DBH$model$alti)
```



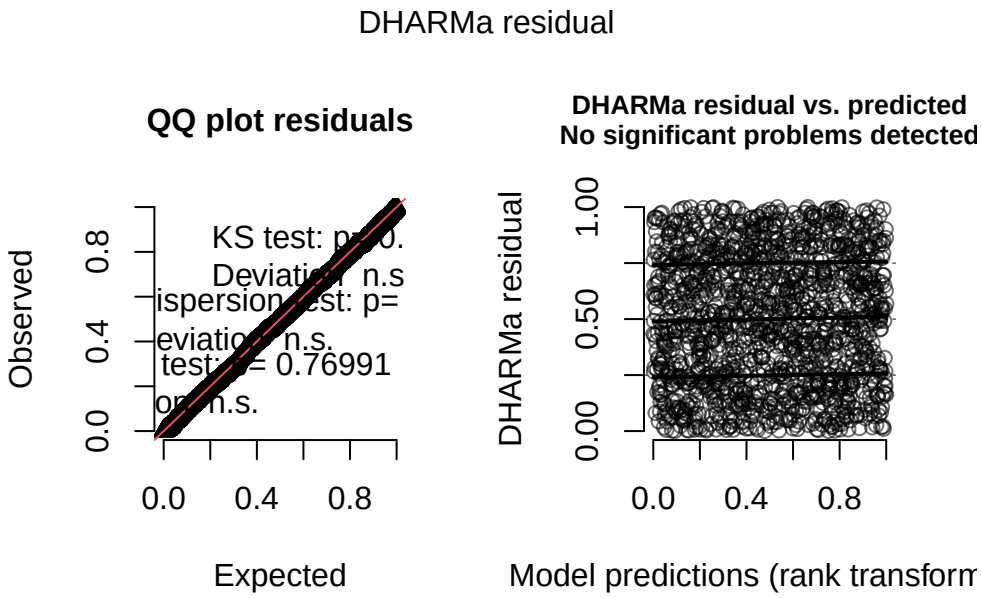

```
    Test for location of quantiles via qgam

data:  res
p-value = 0.9946
alternative hypothesis: both
```

5.5.3 Interprétation de la vérification des résidus par rapport à l'altitude

1. **Test des quantiles** :
   - La **p-value** obtenue est de `0.9946`, ce qui est bien supérieur au seuil de 0.05.
   - Cela indique qu'il n'y a **aucune déviation significative** des résidus en fonction de la variable prédictrice `alti` (altitude). En d'autres termes, l'altitude n'introduit pas de structure systématique dans les résidus, ce qui confirme que cette hypothèse du modèle est respectée.
2. **Graphique des résidus vs altitude** :
   - Les résidus simulés sont uniformément répartis autour des lignes de référence (horizontales), sans tendance apparente ni structure systématique. Cela confirme les résultats du test statistique.

Conclusion : Les résultats montrent que les résidus ne dépendent pas de l'altitude, confirmant que le modèle est robuste par rapport à cette variable. Il n'y a **pas de problème significatif détecté** en termes de relation entre les résidus et l'altitude.

```
# Visualisation des résidus simulés
plot(residusTransfo)
```



5.5.4 Interprétation de la visualisation des résidus simulés

1. **Graphique QQ des résidus simulés** :
   - Le graphique montre que les résidus suivent une ligne diagonale proche de la ligne de référence rouge.
   - Les tests associés (KS test, dispersion test et outlier test) donnent tous des **p-values non significatives** (toutes > 0.05), indiquant que :
     - Les résidus suivent une distribution uniforme.
     - Il n'y a pas de problèmes significatifs de dispersion ou de valeurs aberrantes.
2. **Résidus en fonction des prédictions** :

- Le second graphique affiche les résidus simulés par rapport aux prédictions du modèle.
- Les résidus sont uniformément répartis autour de la ligne médiane (0.5), sans motif systématique visible.
- Le message indique qu'**aucun problème significatif n'a été détecté**.

Conclusion générale : Les deux graphiques confirment que les résidus simulés respectent les hypothèses du modèle. Ces résultats valident la robustesse et la fiabilité de notre modèle binomial généralisé (`mod_bin_cav_DBH`). Il n'y a **aucune violation majeure des hypothèses** qui pourrait remettre en question les conclusions tirées des analyses.

5.6 Conclusion pour la question 5

1. **Relevé (`releve`)** :
   - Le relevé est confirmé comme le facteur explicatif principal de la présence de cavités basses. Les effets significatifs observés dans les deux modèles soulignent une forte **hétérogénéité spatiale**, avec des différences marquées entre plusieurs sites, notamment BLO_12, BLO_17, BLO_4 et BLO_9.
2. **Altitude (`alti`)** :
   - L'altitude ne montre aucun effet significatif dans les deux modèles. Son rôle dans la probabilité de présence de cavités basses semble négligeable dans ce contexte.
3. **Diamètre (`DBH`)** :
   - L'ajout de `DBH` n'apporte aucune amélioration statistique ou prédictive au modèle. Les tests montrent qu'il n'a pas d'impact détectable sur la présence de cavités basses, ce qui est confirmé par l'augmentation de l'AIC et l'absence de réduction notable de la deviance résiduelle.
4. **Validité des hypothèses** :
   - Les vérifications réalisées avec **DHARMa** confirment que le modèle respecte les hypothèses essentielles de distribution uniforme, de dispersion et d'indépendance des résidus. Cela valide la robustesse et la fiabilité des conclusions tirées.

Synthèse :
L'analyse met en évidence que le relevé est le principal facteur influençant la présence de cavités basses, tandis que ni l'altitude ni le diamètre des arbres ne contribuent significativement. Ces résultats suggèrent que des facteurs environnementaux ou spatiaux propres à chaque site sont prépondérants et mériteraient une exploration plus approfondie.