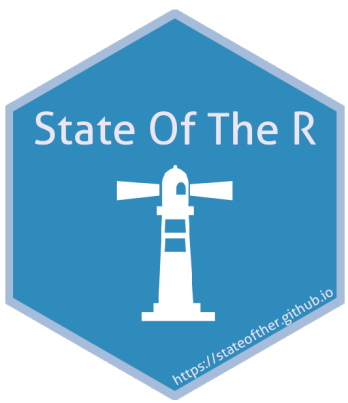
# Statistical modelling for biological data with R

## Day 5 - Mixed Model

Marie-Pierre Etienne

https://marieetienne.github.io

Novembre 2019




# Plan

# Plan

# Plan

# Etude de la consommation de 5 voitures

**Situation 1** 5 amis veulent comparer la consommation de leur voitures respectives. Ils répètent 3 fois la procédure suivante. Une voiture roule 200km après avoir rempli son réservoir et on constate ensuite combien elle a consommé.

Le modèle

$$Y_{ik} = \mu + \alpha_i + E_{ik}$$

$$E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Question : les 5 voitures ont elles la même consommation ?

On s'intéresse à ces 5 voitures en particulier

# Etude de la consommation de 5 voitures

**Situation 2** A la sortie d'une usine, 5 voitures sont choisies au hasard pour étudier leur consommation. On répète 3 fois la procédure suivante. Une voiture roule 200km après avoir rempli son réservoir et on constate ensuite combien elle a consommé.

Le modèle

$$Y_{ik} = \mu + A_i + E_{ik}$$

$$A_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_A^2) \qquad \text{et} \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Question : Les voitures produites par l'usine ont-elles la même consommation ?

On ne s'intéresse pas à ces 5 voitures en particulier, mais on veut généraliser à l'ensemble de l'usine. On auraut pu choisir 5 vautres voitures. Le choix de la voiture parmi toute les voitures possibles est aléatoire.

# Plan

# Plan

# Héritabilité - 1

On se demande si un trait est fortement héritable ou non : les individus issus d'un même ascendant sont-ils plus semblables que ceux issus d'ascendants différents ?

**Données** $I$ pères $i = 1, 2, \ldots, I$ chacun ayant eu $n_i$ descendants numérotés $k = 1, 2, \ldots, n_i$. $Y_{ij}$ est la valeur du trait pour le $k$-ème descendant du $i$-ème père.

**Modèle**

$$Y_{ik} = \mu + A_i + E_{ik}$$

$$A_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_A^2) \qquad \text{et} \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

La variance vaut

$$V(Y_{ik}) = \sigma_A^2 + \sigma^2,$$

et la covariance

$$cov(Y_{ik}, Y_{i'k'}) = 0, \qquad cov(Y_{ik}, Y_{ik'}) = \sigma_A^2$$

# Héritabilité - 2

**Modèle**

$$Y_{ik} = \mu + A_i + E_{ik}$$

$$A_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_A^2) \qquad \text{et} \qquad E_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Composante de la variance $\sigma_A^2$ et $\sigma^2$.

Ecriture matricielle :

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{Z}A + \boldsymbol{E}$$

$$Var(Y) = \sigma_A^2 \boldsymbol{Z}\boldsymbol{Z}' + \sigma^2 I_n$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{O} & \cdots & \boldsymbol{O} \\ \boldsymbol{O} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{O} \\ \boldsymbol{O} & \cdots & \boldsymbol{O} & \boldsymbol{R} \end{bmatrix} \qquad \text{où} \qquad \boldsymbol{R} = \begin{bmatrix} \sigma^2 + \sigma_A^2 & \sigma_A^2 & \cdots & \sigma_A^2 \\ \sigma_A^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_A^2 \\ \sigma_A^2 & \cdots & \sigma_A^2 & \sigma^2 + \sigma_A^2 \end{bmatrix}$$

# Plan

# Mesures répétées

Effet d'un régime sur la prise de poids d'un animal au cours du temps. Plusieurs animaux (indicés par $j$) reçoivent chaque régime (noté $i$) et un animal ne reçoit qu'un régime au cours de l'expérience.

On mesure le poids, noté $Y_{ijt}$, de chaque animal au bout de $t$ semaines ($t = 1, \ldots, T$). On parle donc de mesures répétées, au cours du temps sur un même individu. De telles mesures sont aussi fréquemment appelées **données longitudinales**.

**Modèle** la dépendance dépend du temps séparant 2 mesures.

$$E(Y_{ijt}) = \mu + \alpha_i + \gamma_t + (\alpha\gamma)_{it}$$

**Structure de dépendance**

$$Var(Y_{ijt}) = \sigma^2$$

$$Cov(Y_{ijt}, Y_{i'j't'}) = \begin{cases} \sigma^2 \rho^{|t-t'|} & \text{si } (i,j) = (i',j'), \\ 0 & \text{sinon.} \end{cases} .$$

# Plan

# Version générique du modèle

**Modèle**

$$Y_{ik} = \boldsymbol{X}\theta + \boldsymbol{E}$$

$$\boldsymbol{E} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$$

# Plan

# Des idées sur l'estimation des paramètres

- Cas $\boldsymbol{\Sigma}$ connu (Moindres carrés généralisé)

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{Y};\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}).$$

- Cas $\boldsymbol{\Sigma}$ inconnu
  - Estimation par maximum de vraisemblance : pas d'estimateur explicite en général, biais dans la variance
  - Maximum de vraisemblance restreint : On cherche $T$ tel que $T\boldsymbol{X} = 0$ et donc $T\boldsymbol{Y} = T\mathbb{E}$ ne dépend plus de $\theta$ et $var(T\boldsymbol{E}) = Tvar(\boldsymbol{E})T'$.

# Prédictions des effets aléatoires

$$\boldsymbol{Y} = \boldsymbol{X}\theta + \boldsymbol{Z}A + \boldsymbol{E}$$

On veut prédire les effets aléatoires $A$, pour ceci on forme le BLUP (Best Linear Unbiased Predictor).

$\boldsymbol{\Sigma} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{S}$, en notant $\boldsymbol{D} = Var(\boldsymbol{U})$ et $\boldsymbol{S} = Var(\boldsymbol{E})$.

La loi conditionnelle du vecteur de $\boldsymbol{U}$

$$E(\boldsymbol{U}|\boldsymbol{Y}) = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\theta}).$$

# Plan

# Test sur des composantes de la variance

En général assez difficile. Si la taille de l'échantillon on peut utiliser un test asymptotique de rapport de vraisemblance

On peut approcher la loi asymptotique par une loi du $\chi^2(DDL)$, DDL étant le nombre de contraintes permettant de passer du modèle complet au modèle restreint.

# Test sur les effets fixes

On peut utiliser le test du maximum de vraisemblance. Il existe des situations spécifiques dans lesquelles des tests exacts existent (voir Daudin (2015), chapitre modèle mixte).

# Plan

# Plan

# Les samares du frêne

Cette étude est issue des travaux de thèse de S. Brachet Brachet (1999) et d'O. Ronce Ronce (1999).

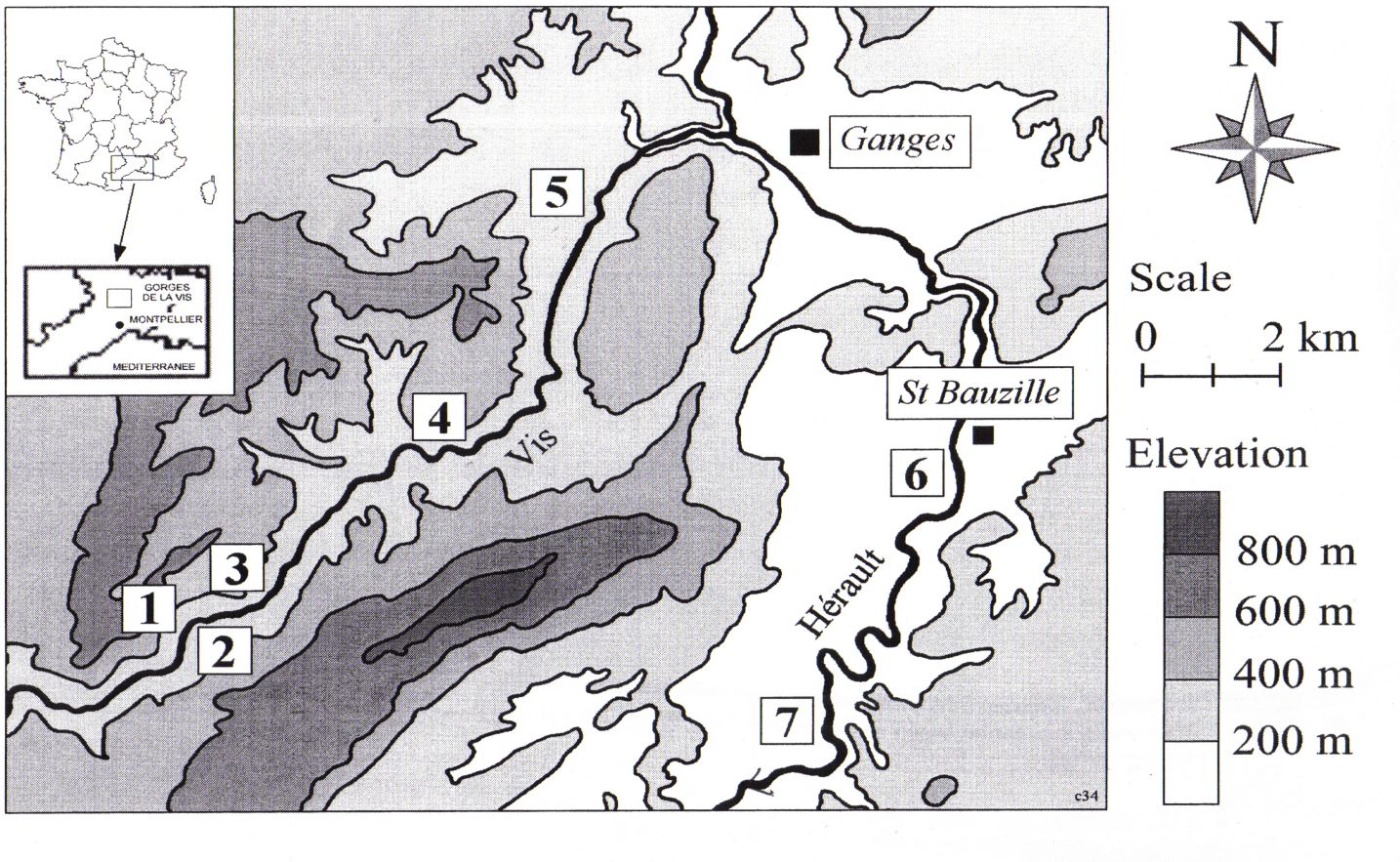



Figure 1: Samares

# Les samares du frêne - présentation données

- 7 sites numérotés de l'amont vers l'aval et sont situés respectivement à 0.2, 1.0, 6.3, 11.3, 22.1 et 30.7 km du premier site en suivant le cours de la rivière.
- Dans chaque site, on a échantillonné entre 7 et 29 arbres, soit $m = 118$ arbres au total.
- Sur chaque de ces arbres, on a prélevé entre 20 et 40 samares, soit $n = 2420$ au total.
- Par samare, on a ensuite mesuré le poids (en grammes) et la surface (en $\text{cm}^2$) de chaque samares et calculé l'indice de dispersion

$$Y = 100 \times \text{surface} \,/\, \text{poids}.$$

**Questions**

- Existe-t-til une grande variabilité de l'indice de dispersion entre les arbres et
- Les sites les plus récemment colonisés sont-ils peuplés d'arbres produisant des samares plus dispersives ou moins dispersives.

# Descrption des données

```
samares <- read.table("https://marieetienne.github.io/datasets/Sama

samares %>% mutate(Site = as_factor(Site)) -> samares
```

# Le plan d'expérience

```
samares %>% group_by(Site) %>% summarize(n = n())
```

```
## # A tibble: 7 x 2
##   Site      n
##   <fct> <int>
## 1 1       280
## 2 2       300
## 3 3       200
## 4 4       580
## 5 5       580
## 6 6       300
## 7 7       140
```

```
samares %>% group_by(Site, Arbre) %>% summarize(n = n())
```

```
## # A tibble: 119 x 3
## # Groups:   Site [7]
##    Site  Arbre     n
##    <fct> <int> <int>
##  1 1         1    20
```

# Creation de la variable dispersion

```
samares %>%  mutate(Disp=Surface/Poids, lDisp=log10(Surface/Poids))
```

# Plan

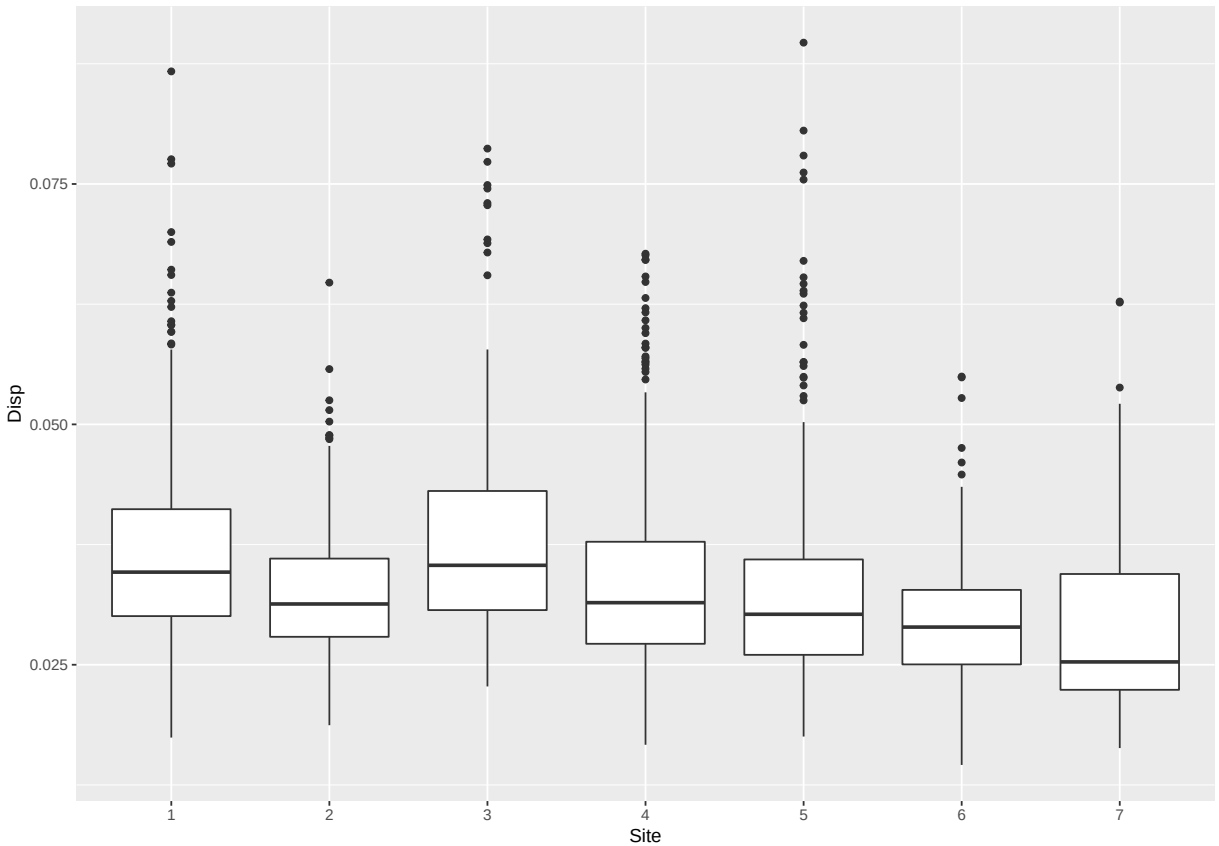
# Par Site

```
samares %>%
  group_by(Site) %>%
  summarise(m_disp = mean(Disp))
```

```
## # A tibble: 7 x 2
##   Site  m_disp
##   <fct>  <dbl>
## 1 1     0.0369
## 2 2     0.0326
## 3 3     0.0381
## 4 4     0.0334
## 5 5     0.0321
## 6 6     0.0291
## 7 7     0.0290
```

# Par Site

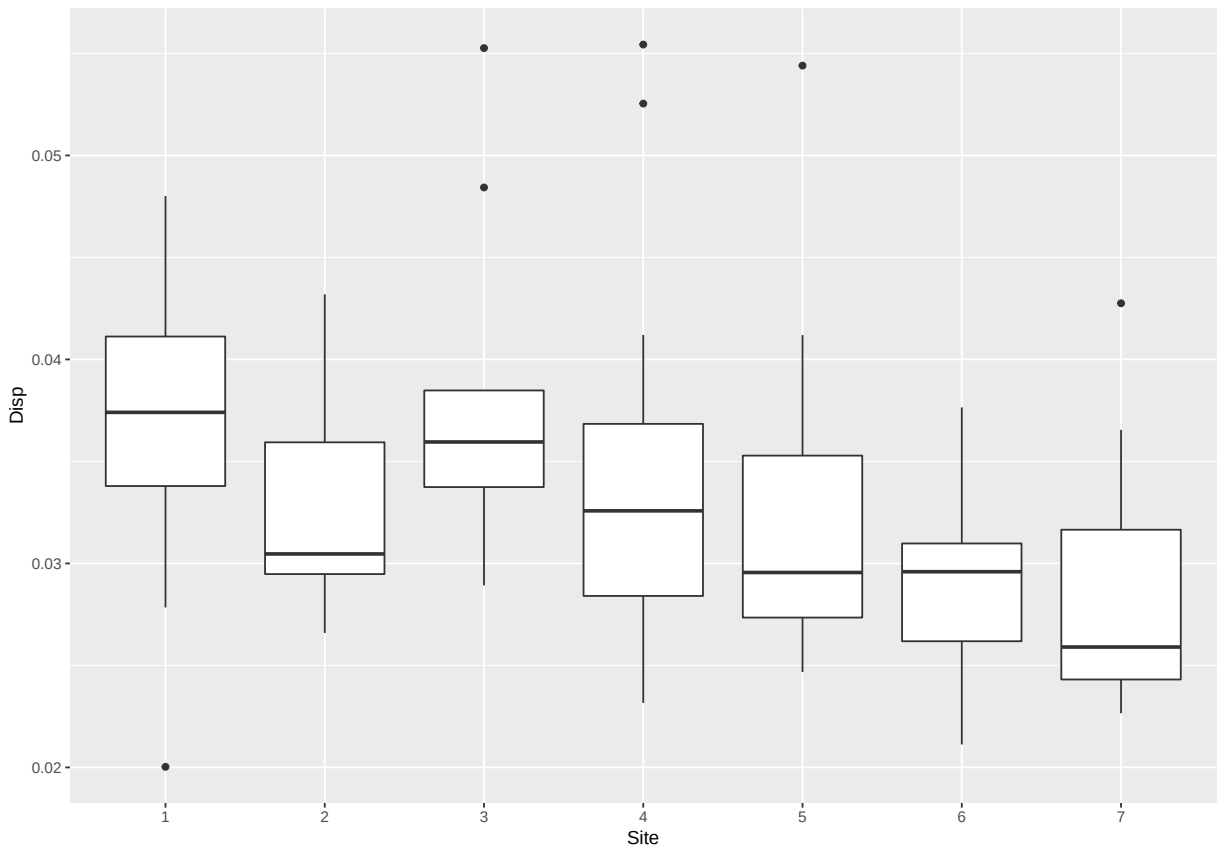
```
samares %>%
  ggplot() +
  geom_boxplot(aes(x= Site, y = Disp))
```




##Analyse a l'echelle de l'arbre

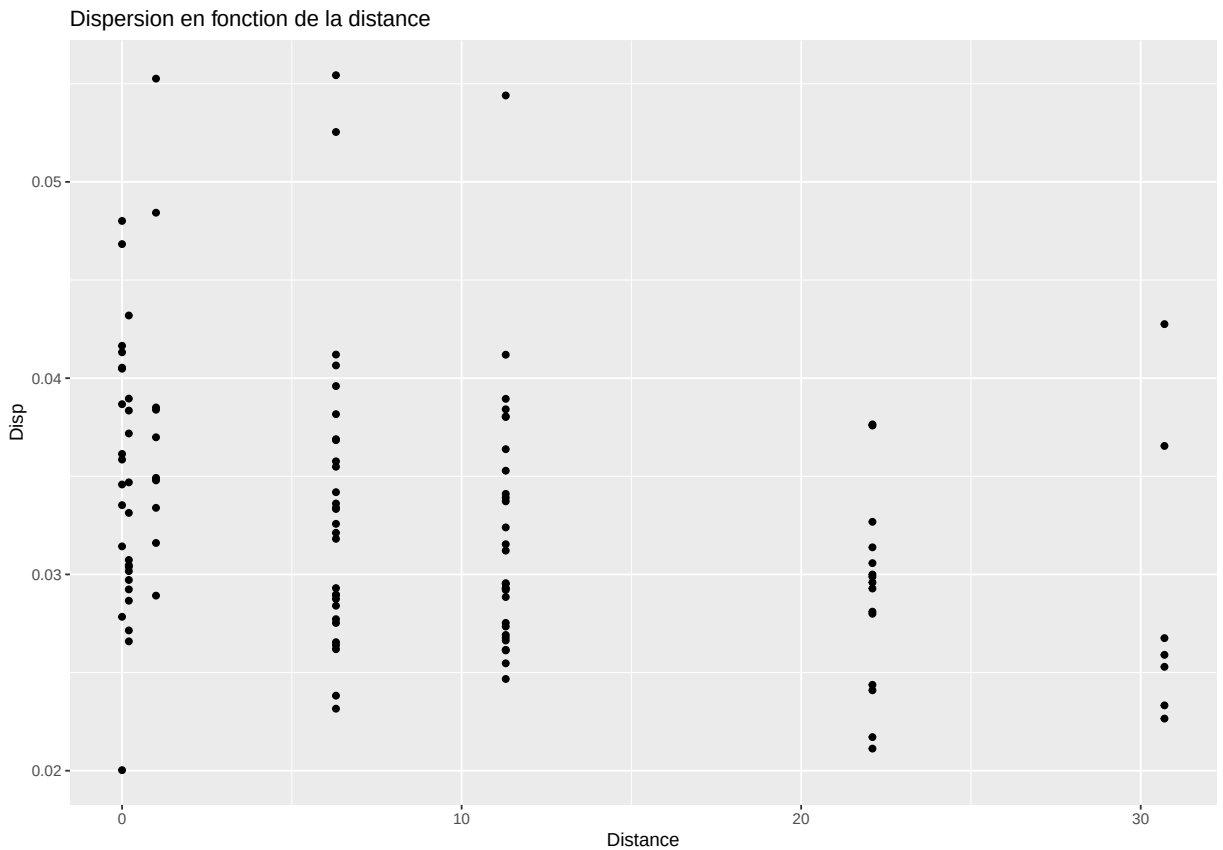
# Creation du jeu de donnees a l'echelle de l'arbre

```
samares %>%
  group_by(Site, NomSite, Arbre) %>%
  summarise_all(mean) -> samares_arbre
samares_arbre %>%   ggplot() +
  geom_boxplot(aes(x= Site, y = Disp))
```



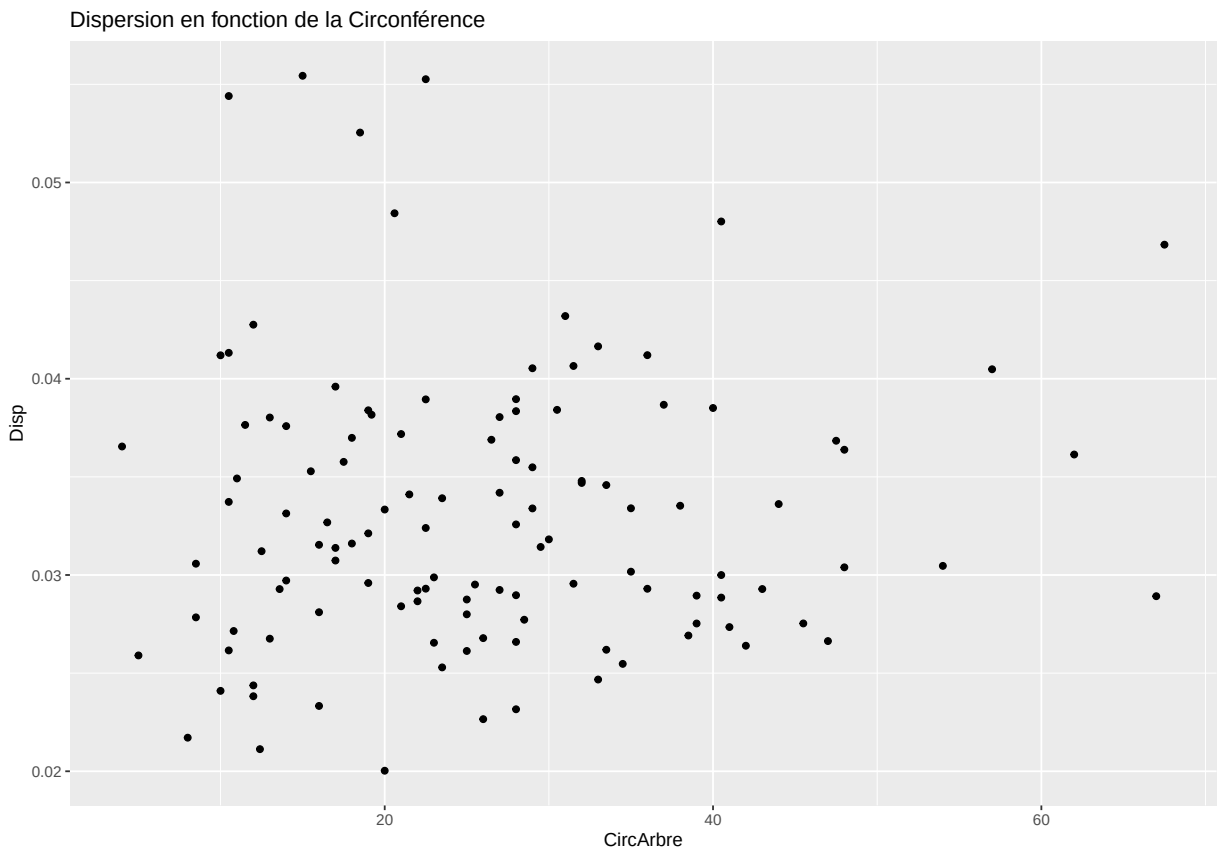
# Creation du jeu de donnees a l'echelle de l'arbre

```
samares_arbre %>% ggplot() +
  geom_point(aes(x= Distance, y = Disp)) +
  ggtitle(label = 'Dispersion en fonction de la distance')
```



# Creation du jeu de donnees a l'echelle de l'arbre

```
samares_arbre %>%
  ggplot() +
  geom_point(aes(x= CircArbre, y = Disp)) +
  ggtitle(label = 'Dispersion en fonction de la Circonférence')
```



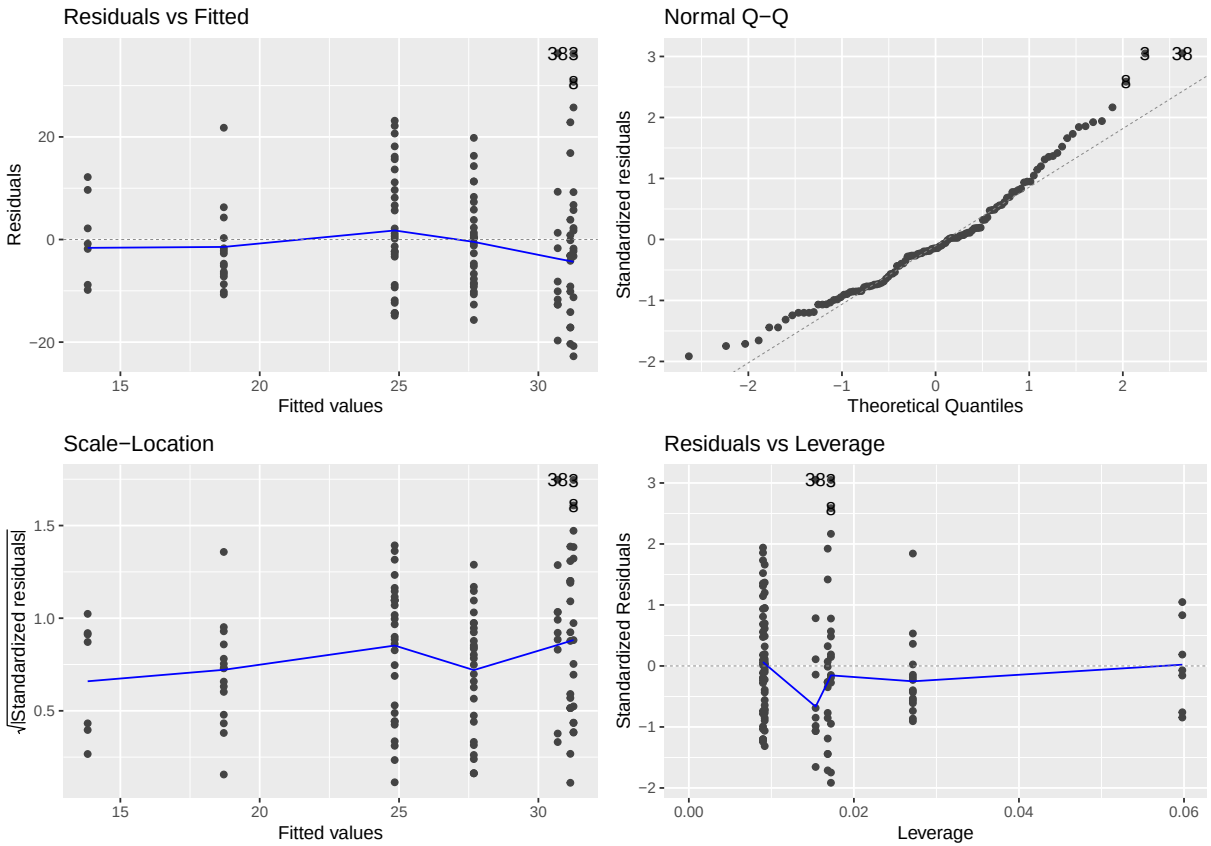
# Y a -t-il des différences d'age entre les peuplement

```
circ.lm <- lm( CircArbre~ Distance, data = samares_arbre)
anova(circ.lm)

## Analysis of Variance Table
##
## Response: CircArbre
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Distance     1  2957.4 2957.39  20.583 1.392e-05 ***
## Residuals  117 16810.9  143.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Y a -t-il des différences d'age entre les peuplement - diagnostic

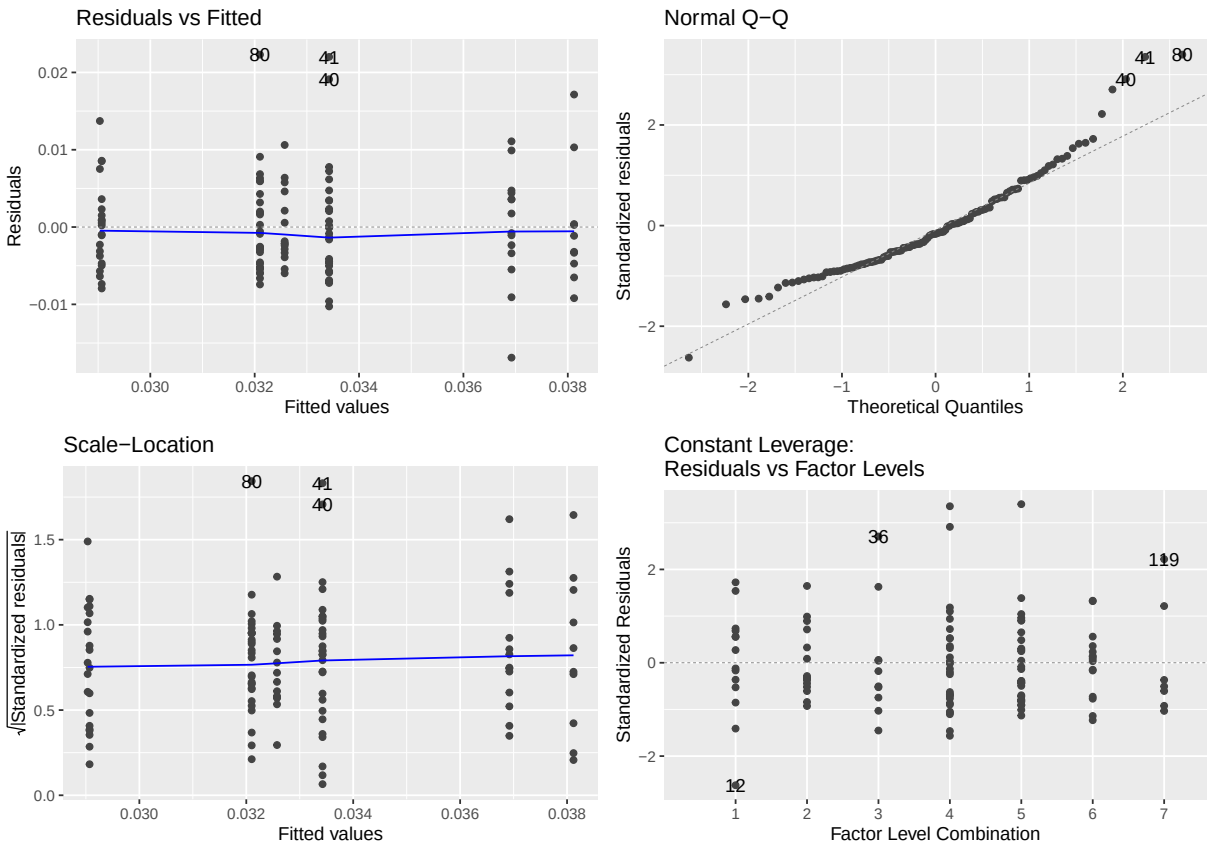
```
library(ggfortify)
autoplot(circ.lm)
```




```
disp.lm <- lm( Disp~ Site, data = samares_arbre)
```

# Anova 1 : modele dispersion en fonction du site - Diagnostic

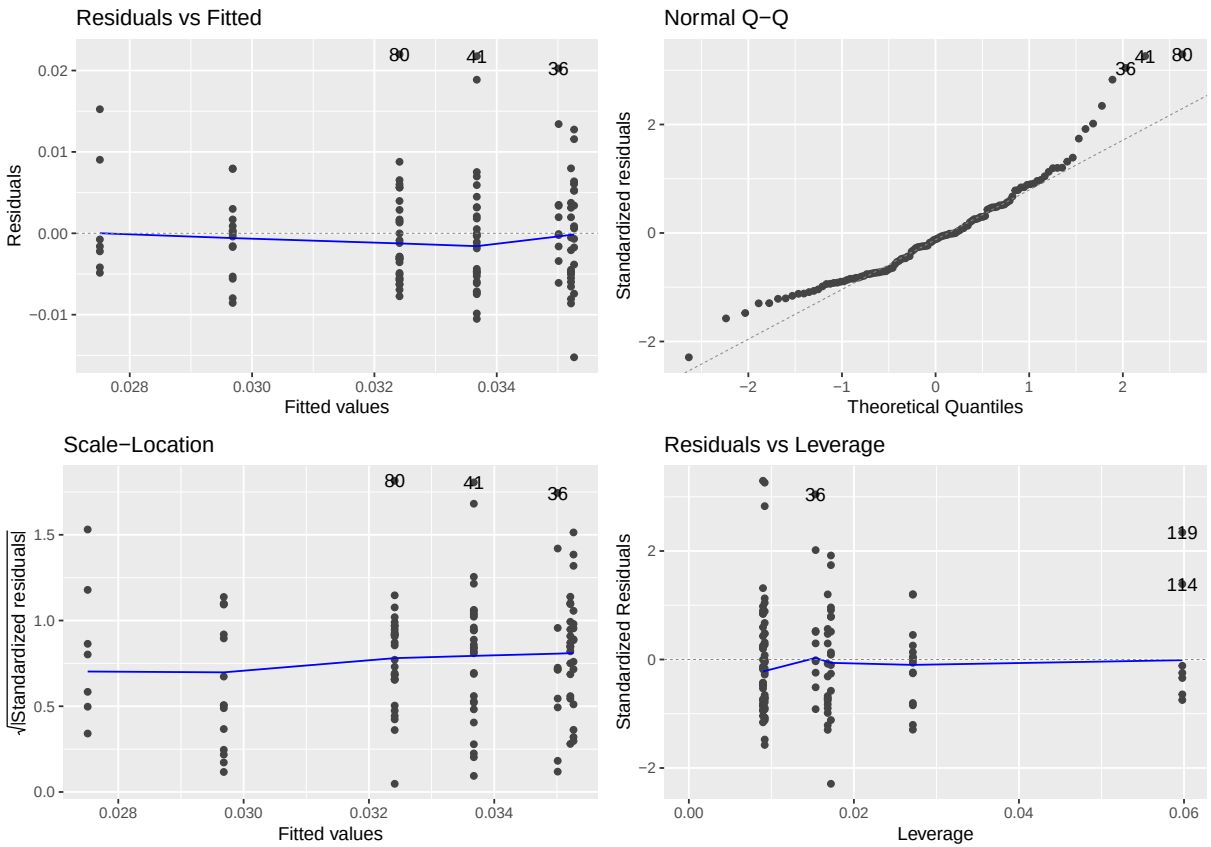
```
autoplot(disp.lm)
```




```
disp.lm2 <- lm( Disp ~ Distance, data=samares_arbre)
anova(disp.lm2)
```

# Regression : dispersion en fonction de distance - Diagnostic

```
autoplot(disp.lm2)
```



# Plan

# Plan

# Plan

# Plan

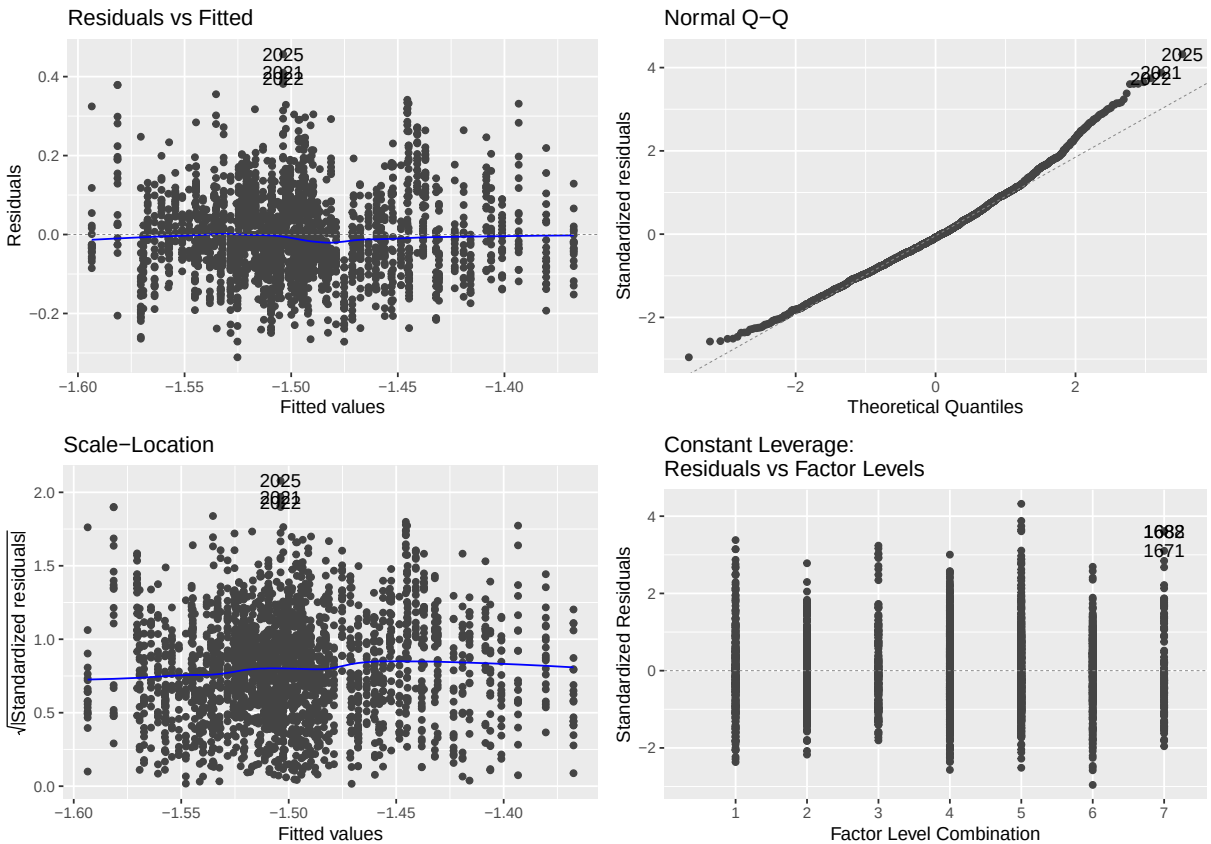
# Anova 2 facteurs fixes

```
ldisp.lm4 <- lm( lDisp ~ Site + Site:Arbre, data=samares)
anova(ldisp.lm4)

## Analysis of Variance Table
##
## Response: lDisp
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Site          6  2.7417 0.45696  40.828 < 2.2e-16 ***
## Site:Arbre    7  1.6695 0.23851  21.310 < 2.2e-16 ***
## Residuals  2366 26.4806 0.01119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Anova 2 facteurs fixes - Diagnostic

```
autoplot(ldisp.lm4)
```



# Modèle mixte

```
library(lmerTest)
#modele mixte : effet Arbre aleatoire
ldisp.lmm <- lmer(lDisp ~ Site + (1|Site:Arbre),data=samares)
ldisp.lmm2 <- lmer(lDisp ~ Site*CircArbre+(1|Site:Arbre),data=samar

anova(ldisp.lmm, type = 'II')

## Type II Analysis of Variance Table with Satterthwaite's method
##       Sum Sq  Mean Sq NumDF DenDF F value   Pr(>F)
## Site 0.11803 0.019672     6   112  3.3981 0.004065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ranef(ldisp.lmm)

## $`Site:Arbre`
##       (Intercept)
## 1:1   0.051716021
## 1:2   0.044221850
## 1:3   0.099022078
```

# Références

Brachet, S. (1999). *La dispersion : Déterminisme et conséquences. Approche théorique et experimentale chez le frêne* (PhD thesis). ENGREF. Retrieved from http://www.theses.fr/2000ENGR0058

Daudin, J.-J. (2015). *Le modèle linéaire et ses extensions - Modèle linéaire général, modèle linéaire généralisé, modèle mixte, plans d'expériences (Niveau C)* (p. 336 p.). Edition Ellipses.

Ronce, O. (1999). *Histoires de vie dans un habitat fragmenté : Étude théorique de l'évolution de la dispersion et d'autres traits* (PhD thesis). Montpellier II.