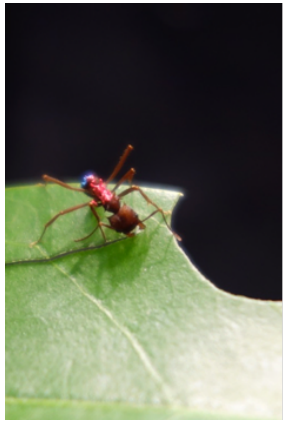
MoDoBio Modélisation des Données en Biologie

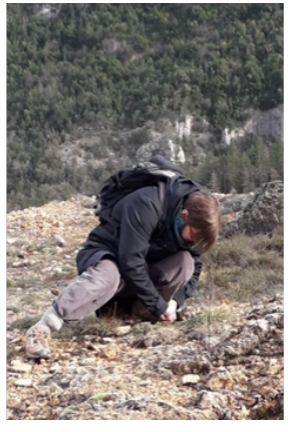
L3 Sciences de la Vie, parcours Biologie-Ecologie

Elodie Brunel MA et Céline Devaux BE

Université de Montpellier – Faculté des Sciences

2024-2025






ANalysis of CoVAriance

(Comparer des pentes entre groupes)

Objectifs d'une ANCOVA (à 1 facteur)

L'ANOVA (ANalysis of COVAriance) est une méthode d'analyse permettant d'étudier la dépendance d'une variable **quantitative** à une variable **qualitative** et une variable **quantitative**.

Exemple : effet de l'herbivorie sur la croissance d'*Ipomopsis* dans le parc de l'Imperial College

Variables mesurées

- le poids (en mg) de fruits produits
- la taille (en cm) initiale de la racine
- la présence de lapins (exclusion)

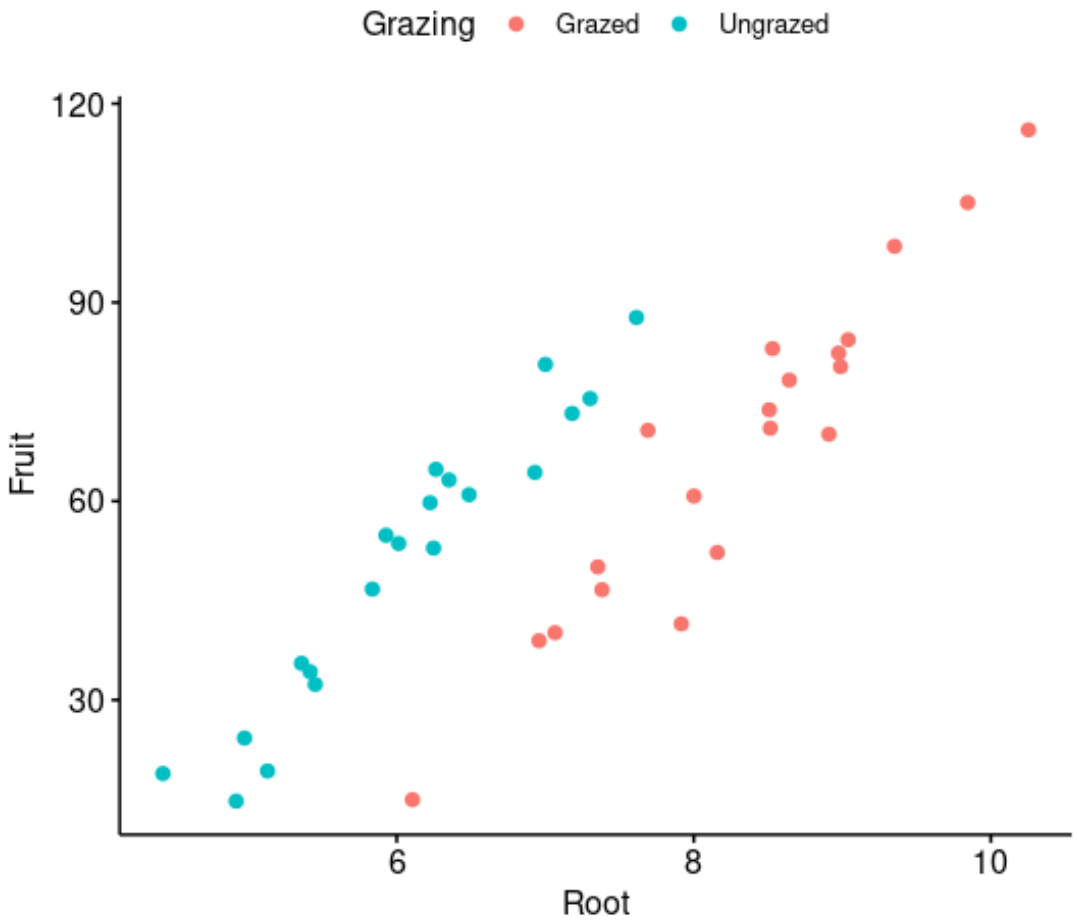
Objectifs d'une ANCOVA (à 1 facteur)

question biologique est-ce que l'herbivorie modifie l'effet de la taille initiale sur la fitness des plantes ?

question statistique est-ce que la régression de la fitness sur la taille dépend de l'herbivorie ?

- tester l'effet de l'herbivorie
- en contrôlant pour la taille initiale
- tester l'interaction taille initiale / herbivorie

Objectifs d'une ANCOVA (à 1 facteur)



les relations semblent linéaires pour les deux niveaux d'herbivorie

Le modèle statistique

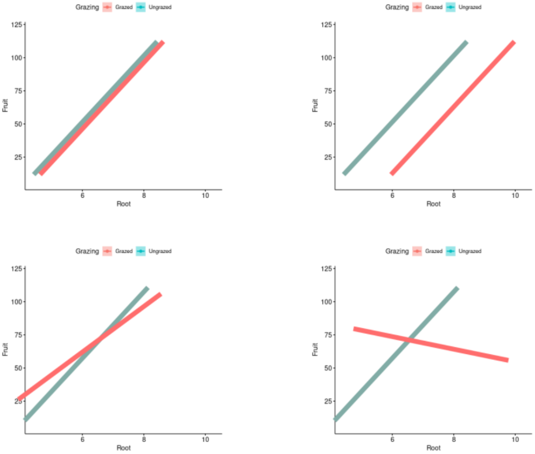
Soient Y et X deux variables quantitatives (le poids des fruits, en *mg*, et la taille initiale de la racine, en *cm*) et un facteur avec 2 modalités ($J = 2$ groupes) : `"avec lapins"` (`"grazed"`), `"sans lapins"` (`"ungrazed"`)

Y est une variable aléatoire qui dépend de x et du groupe j par

$$Y_{ij} = \beta_0 + \alpha_j + \beta_j X_{ij} + \epsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2_\epsilon)$$

- les paramètres α_j, β_0 et β_j sont inconnus
- la variance σ^2_ϵ est inconnue
- $\mathbb{E}(Y_{ij}) = \beta_0 + \alpha_j + \beta_j x_{ij} + \mathbb{E}(\epsilon_{ij}) = \beta_0 + \alpha_j + \beta_j x_{ij}$
- $\mathrm{Var}(Y_{ij}) = \mathrm{Var}(\epsilon_{ij}) = \sigma^2_\epsilon$

Les modèles statistiques possibles



Le modèle statistique complet

$$Y_{ij} = \beta_0 + \alpha_j + \beta_j X_{ij} + \epsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

avec

- i : l'individu au sein du groupe = herbivorie
- j : l'indice du groupe = herbivorie
- Y_{ij} est l'observation de Y de l'individu i dans le groupe j
- β_0 est l'ordonnée à l'origine pour le groupe $j = 1$
- α_j est l'effet additif du groupe j sur la moyenne globale
- β_j est la pente de droite de régression pour le groupe j
- ε_{ij} : erreur aléatoire inobservée

Les erreurs ε_{ij} sont supposées indépendantes, de loi gaussienne, centrées et de même variance σ_ε^2.

Le modèle statistique : autres écritures possibles

$$Y_{ij} = \beta_0 + \alpha_j + \beta_j X_{ij} + \epsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$
$$Y_{ij} = \mu + \alpha_j + \beta_j (x_{ij} - \overline{x_j}) + \varepsilon_{ij}$$

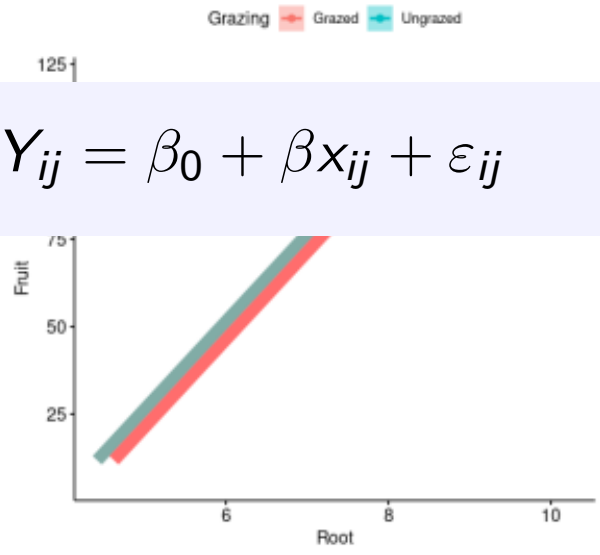
$$Y_{ij} = \beta_0 + \alpha_j + \beta_j X_{ij} + \epsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$
$$Y_{ij} = \beta_0 + \alpha_j + \beta x_{ij} + \delta_j x_{ij} + \varepsilon_{ij}$$

$$Y_{ij} = \mu + \alpha_j + \beta_j (x_{ij} - \overline{x_j}) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$$
$$Y_{ij} = \mu + \alpha_j + \beta (x_{ij} - \overline{x}) + \delta_j (x_{ij} - \overline{x_j}) + \varepsilon_{ij}$$

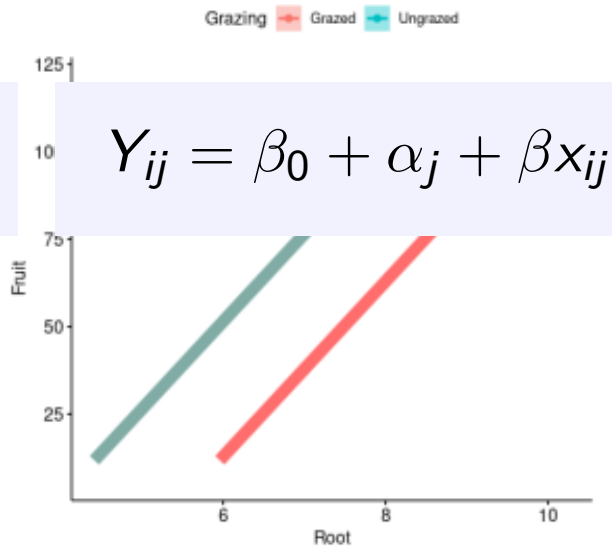
Le modèle statistique : autres écritures possibles

- Y_{ij} est l'observation de Y de l'individu i dans le groupe j
- μ est la moyenne globale pour Y
- $\overline{x_j}$ est la moyenne de X pour le groupe j
- α_j est l'effet additif du groupe j sur la moyenne globale
- β est la pente (globale) de la droite de régression
- δ_j est l'effet additif du groupe j sur la pente globale
- ε_{ij} : erreur aléatoire inobservée

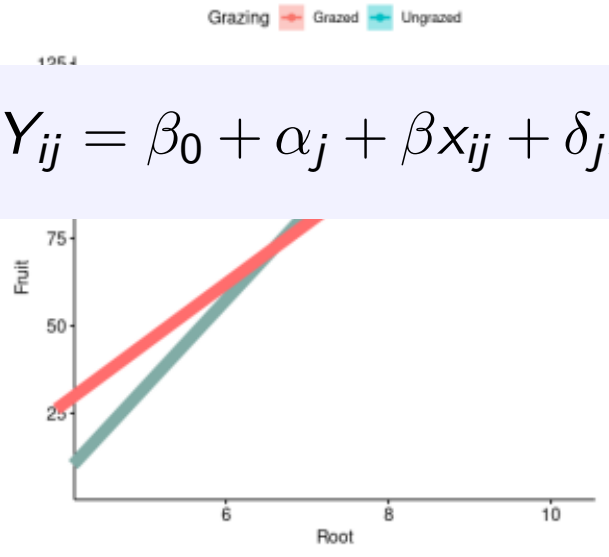
$$Y_{ij} = \beta_0 + \beta x_{ij} + \varepsilon_{ij}$$



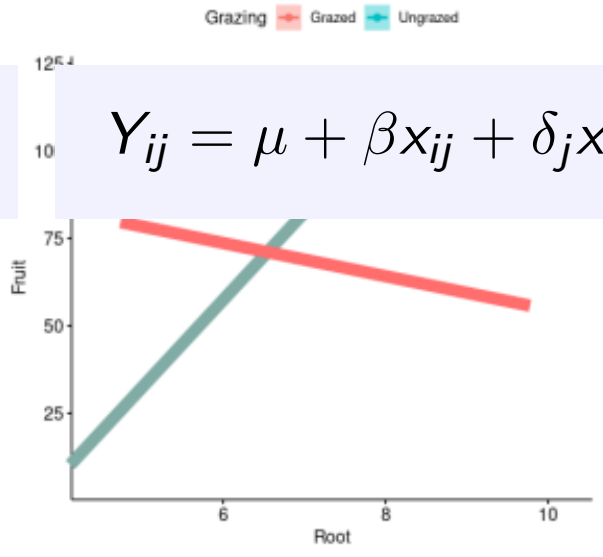
$$Y_{ij} = \beta_0 + \alpha_j + \beta x_{ij} + \varepsilon_{ij}$$



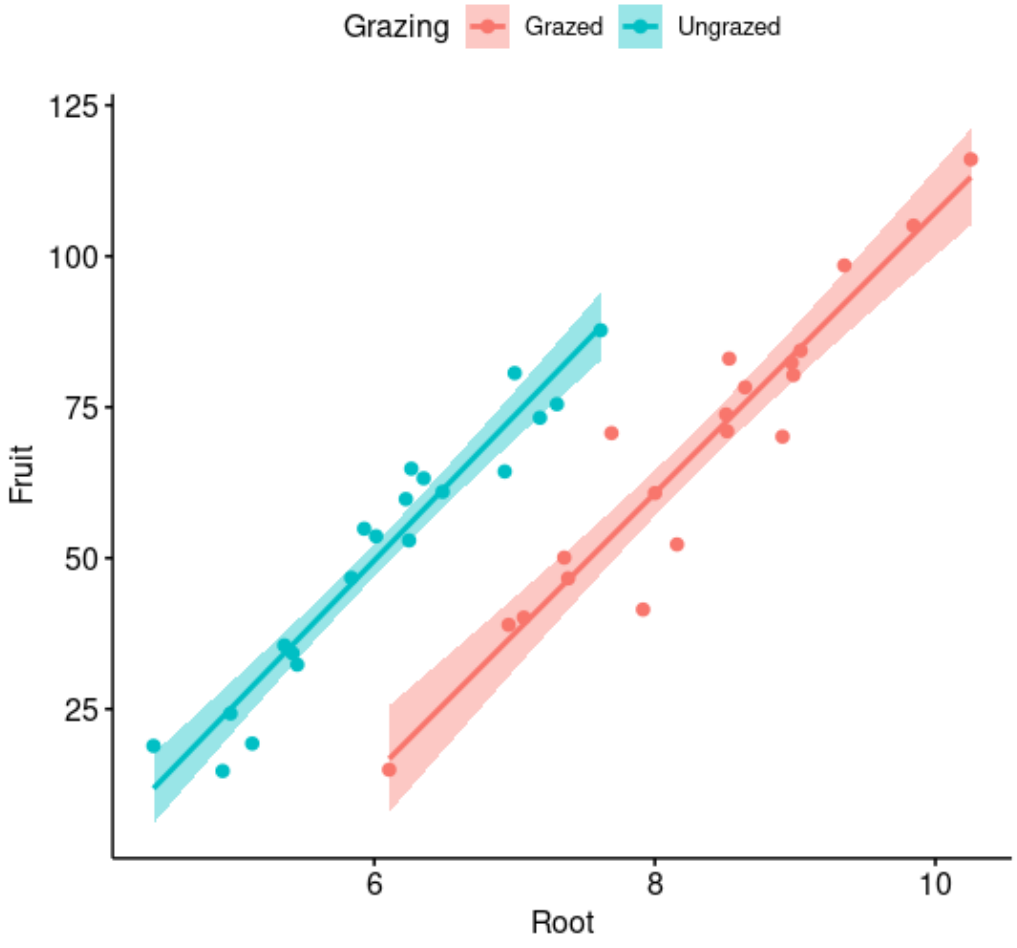
$$Y_{ij} = \beta_0 + \alpha_j + \beta x_{ij} + \delta_j x_{ij} + \varepsilon_{ij}$$



$$Y_{ij} = \mu + \beta x_{ij} + \delta_j x_{ij} + \varepsilon_{ij}$$



Le modèle statistique



on peut estimer une droite de régression dans chaque groupe
mais comment calculer les SST, SSM et SSR ?

Décomposition de la variabilité

Théorème

$$\underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y})^2}_{SST} = \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\hat{y}_{ij}-\bar{y})^2}_{SSM} + \underbrace{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\hat{y}_{ij})^2}_{SSR}$$

variabilité totale de Y | variabilité expliquée | variabilité résiduelle inexpliquée

Mise en oeuvre

On sait calculer le *SSM* du facteur herbivorie

On sait calculer le *SSM* pour chaque droite de régression

On sait calculer le *SSM* d'une droite de régression commune ? ?

globale	avec lapins	sans lapin
$\overline{y} = 59.41$	$\overline{y_g} = 67.94$	$\overline{y_u} = 50.88$
	$\alpha_g = 8.53$	$\alpha_u = -8.53$

avec $n_g = n_u = 20$

$$SSM_H = \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\hat{y} - \overline{y})^2 = \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\alpha_j)^2$$

$$= n(\alpha_g^2 + \alpha_u^2) = 2910.436$$

Mise en oeuvre

On sait calculer le *SSM* du facteur herbivorie
On sait calculer le *SSM* pour chaque droite de régression
On sait calculer le *SSM* d'une droite de régression commune

avec `lapins("grazed")`	sans `lapin ("ungrazed")`
$S_{yy,g} = 11837.79$	$S_{yy,u} = 8995.606$
$S_{xx,g} = 19.911$	$S_{xx,u} = 14.58677$
$S_{xy,g} = 462.7415$	$S_{xy,u} = 350.0302$
$SSM_g = 10754.29$	$SSM_u = 8399.466$
$SSR_g = 1083.509$	$SSR_u = 596.1403$

avec $SSM_{regression} = \frac{S_{xy}^2}{S_{xx}}$ (pour chaque droite)

Mise en oeuvre

On sait calculer le *SSM* du facteur herbivorie
On sait estimer les paramètres chaque droite de régression
On sait calculer le *SSM* d'une droite de régression commune

avec lapins	sans lapin
$\hat{\beta}_g = \frac{S_{xy,g}}{S_{xx,g}}$	$\hat{\beta}_u = \frac{S_{xy,u}}{S_{xx,u}}$
$\hat{\beta}_g = 23.40$	$\hat{\beta}_u = 23.996$
$\hat{\beta}_{0,g} = -126.50$	$\hat{\beta}_{0,u} = -94.36$

Mise en oeuvre

On sait calculer le SS_M du facteur herbivorie
On sait calculer le SS_M pour chaque droite de régression
On sait calculer le SS_M d'une droite de régression commune

avec lapins	sans lapin	complet
$S_{yy,g} = 11837.79$	$S_{yy,u} = 8995.606$	$S_{yy,g+u} = 20833.4$
$S_{xx,g} = 19.911$	$S_{xx,u} = 14.58677$	$S_{xx,g+u} = 34.498$
$S_{xy,g} = 462.7415$	$S_{xy,u} = 350.0302$	$S_{xy,g+u} = 812.7717$
$SSM_g = 10754.29$	$SSM_u = 8399.466$	$SSM_{complet} = 19153.75$
		$= SSM_g + SSM_u$

Mise en oeuvre

On sait calculer le *SSM* du facteur herbivorie
On sait calculer le *SSM* pour chaque droite de régression
On sait calculer le *SSM* d'une droite de régression commune ? ?

avec lapins	sans lapin	
$S_{yy,g} = 11837.79$	$S_{yy,u} = 8995.606$	
$S_{xx,g} = 19.911$	$S_{xx,u} = 14.58677$	$S_{xx,g+u} = 34.498$
$S_{xy,g} = 462.7415$	$S_{xy,u} = 350.0302$	$S_{xy,g+u} = 812.7717$
$SSM_g = 10754.29$	$SSM_u = 8399.466$	

$$SSM_{additif} = \frac{S^2_{xy,g+u}}{S_{xx,g+u}}$$
$$SSM_{additif} = 19148.94$$

Mise en oeuvre

On sait calculer le *SSM* du facteur herbivorie
On sait calculer le *SSM* pour chaque droite de régression
On sait calculer le *SSM* d'une droite de régression commune
On en déduit le *SSM* de l'interaction

$$SSM_g = 10754.29$$
$$SSM_u = 8399.466$$
$$SSM_{complet} = SSM_g + SSM_u = 19153.75$$
$$SSM_{additif} = 19148.94$$
$$SSM_{Hx} = 19153.75 - 19148.94$$
$$SSM_{Hx} = 4.81$$
$$SSR = 1679.6$$

Test statistique "ANOVA" de Fisher

Sous $\mathcal{H}_0$:

$$F = \frac{SSM/(df_M)}{SSR/(df_R)} \sim F_{df_M, df_R}$$

	Df	Sum of Sq.	Mean Sq.	F value	p value
grazing	1	2910.4	2910.4	62.4	2.62×10^{-9}
root	1	19148.9	19148.9	410.4	$<2 \times 10^{-16}$
graz :root	1	4.81	4.81	0.10	0.75
residuals	36	1679.6	46.7		

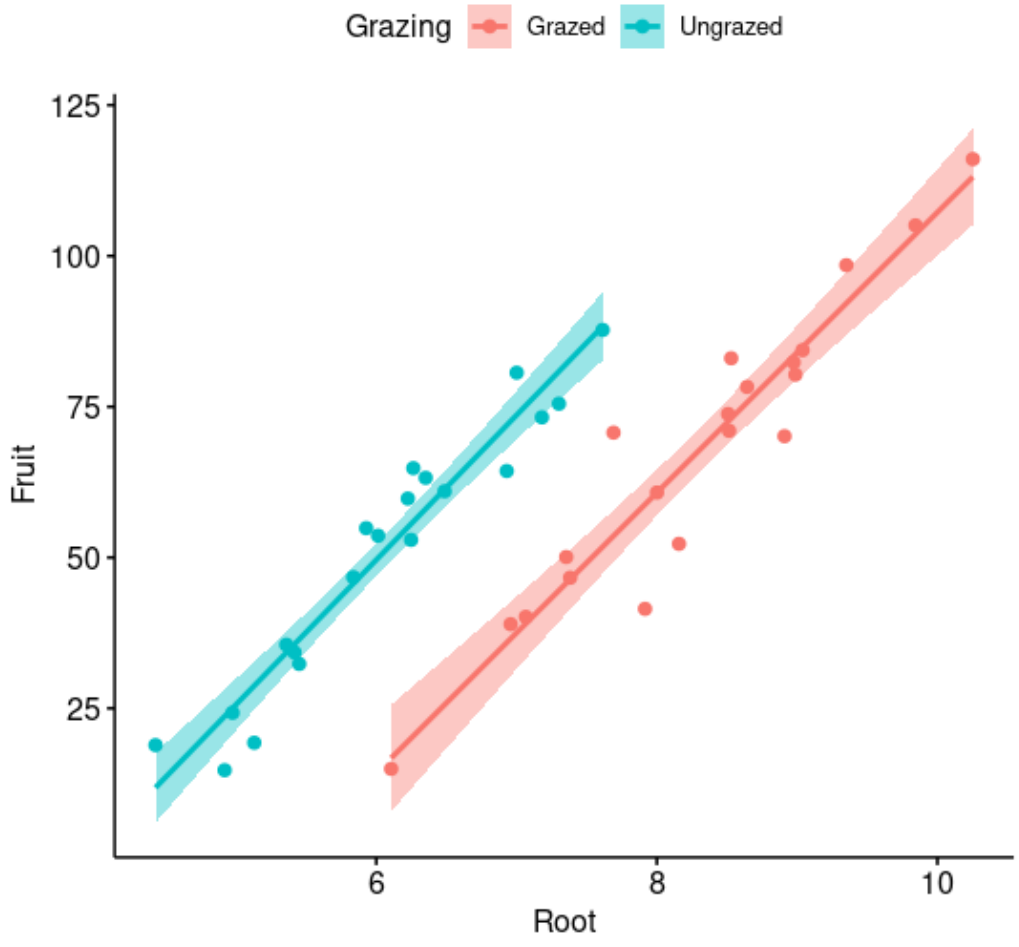
Sélection de modèle

la relation entre la fitness et la taille initiale n'est pas affectée par la présence de lapins.

	Df	Sum of Sq.	Mean Sq.	F value	p value
grazing	1	2910.4	2910.4	63.9	1.4×10^{-9}
root	1	19148.9	19148.9	420.6	$<2 \times 10^{-16}$
residuals	37	1684.4	**45.5**		

- l'herbivorie affecte la fitness
- la taille initiale affecte la fitness

Mais : problème expérimental



plus grandes plantes dans les cages d'exclusion

Sélection de modèle

	Df	Sum of Sq.	Mean Sq.	F value	p value
grazing	1	2910.4	2910.4	63.9	1.4×10^{-9}
root	1	19148.9	19148.9	420.6	$<2 \times 10^{-16}$
residuals	37	1684.4	45.5		

	Df	Sum of Sq.	Mean Sq.	F value	p value
root	1	16795.0	16795.0	368.91	$<2 \times 10^{-16}$
grazing	1	5264.4	5264.4	115.63	6.1×10^{-13}
residuals	37	1684.4	45.5		

une manière plus conservative (p-value plus grande)

interprétation de la taille des effets

$$Y_{ij} = \underbrace{\beta_0 + \alpha_j}_{\text{ordonnée à l'origine}} + \beta\, x_{ij} + \varepsilon_{ij}$$

$$\hat{\beta} = \frac{S_{xy,g+u}}{S_{xx,g+u}} = \frac{812.77}{34.498} = 23.560$$

$$\hat{\beta}_0 = \hat{\beta}_{0,g} = \overline{y}_g - \hat{\beta}\,\overline{x}_g = 67.94 - 23.56 \times 8.31 = -127.83$$

$$\hat{\beta}_{0,u} = \overline{y}_u - \hat{\beta}\,\overline{x}_u = 50.88 - 23.56 \times 6.05 = -91.73$$

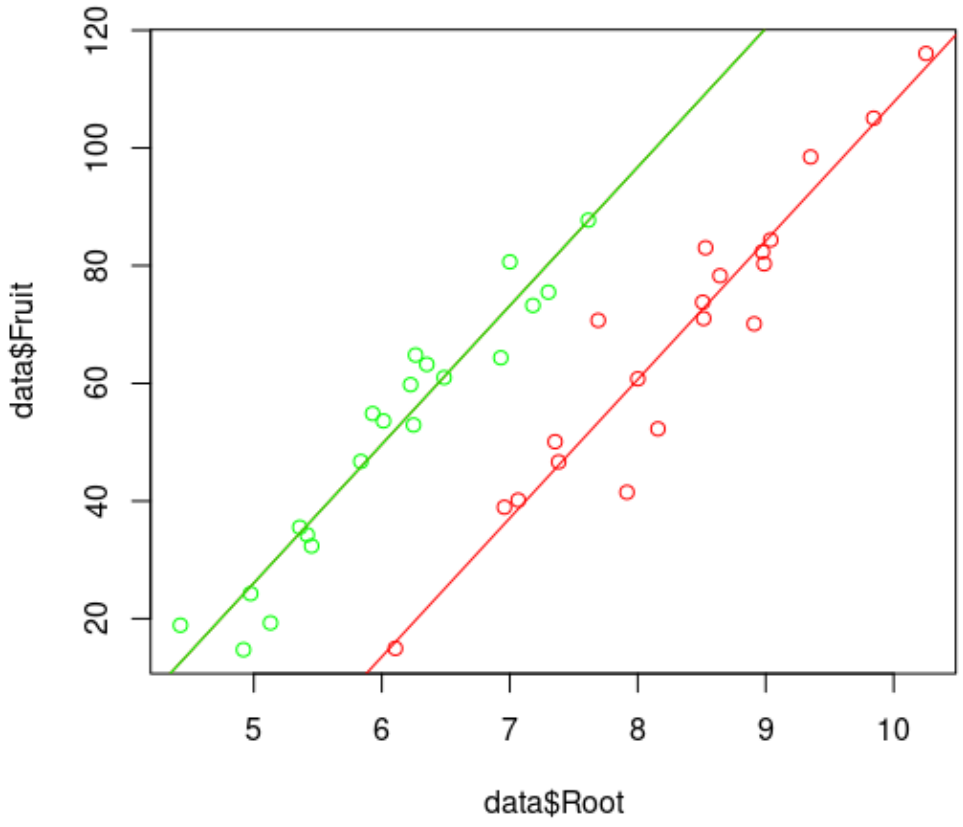
$$\hat{\alpha}_u = -91.73 + 127.83 = 36.10$$

interprétation de la taille des effects

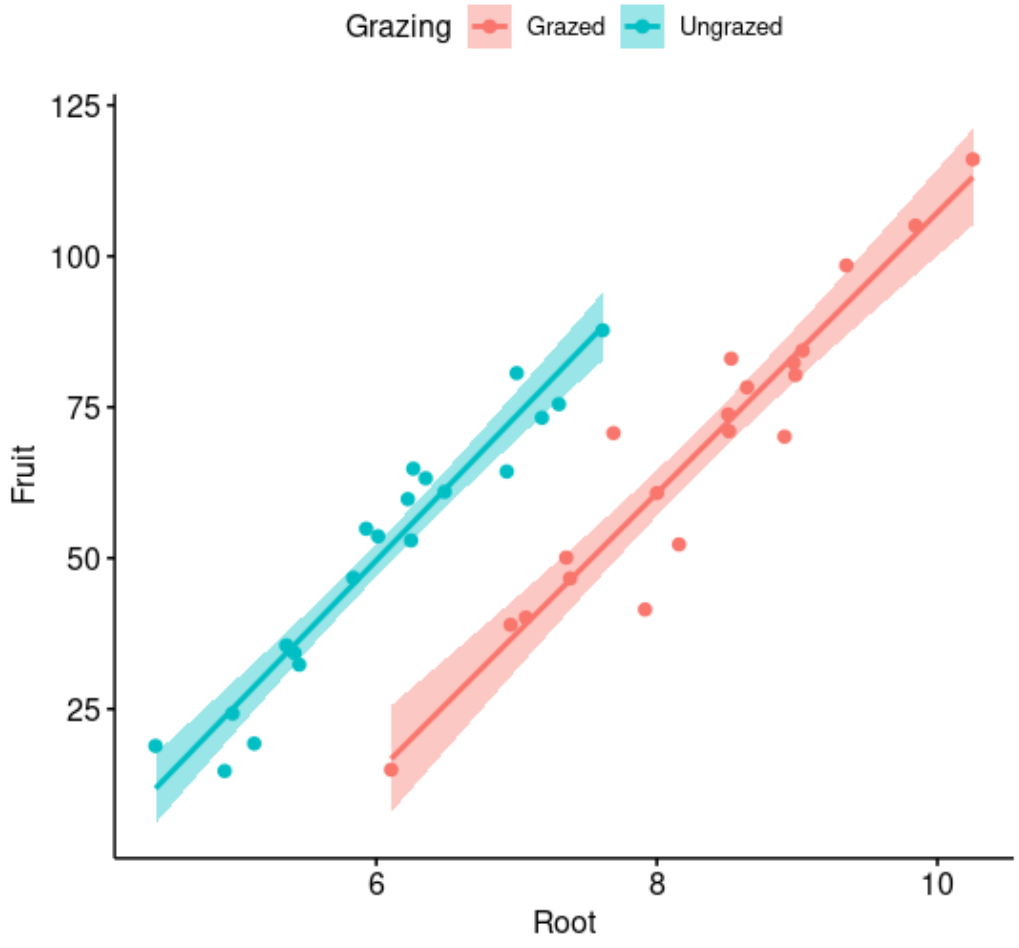
Coefficients :

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-127.829	9.664	-13.23	1.35×10^{-15}
Root	23.560	1.149	20.51	$<2 \times 10^{-16}$
GrazingUngrazed	36.103	3.357	10.75	6.11×10^{-13}

Visualisation des données et du modèle



Intervalles de confiance de la réponse moyenne

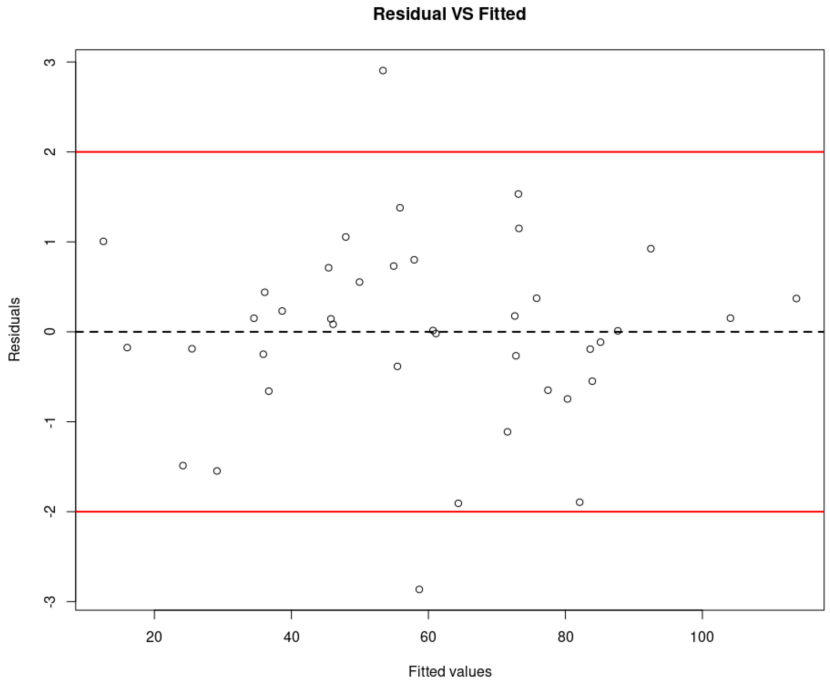


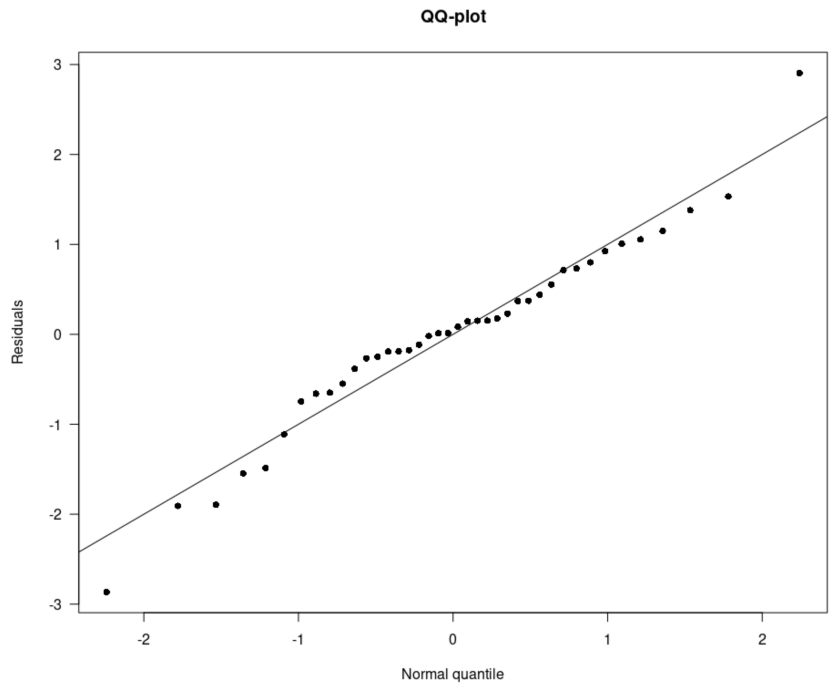
Vérification des hypothèses de validité de l'ANCOVA

- Les erreurs suivent d'une loi gaussienne.
  On peut faire un test de Shapiro (si le niveau de réplication est grand) et une inspection graphique.
- Les variances des erreurs sont égales dans chaque groupe et selon X.
  On peut faire un test d'homogénéité des variances à condition que l'hypothèse de normalité soit satisfaite.
- Les erreurs sont indépendantes.
  On peut faire un test de Durbin Watson et une inspection graphique.

Vérification des hypothèses de validité de l'ANCOVA

- Les erreurs suivent d'une loi gaussienne.
- Les variances des erreurs sont égales dans chaque groupe et selon X.
- Les erreurs sont indépendantes.





Vérification des hypothèses de validité de l'ANOVA

```
Shapiro-Wilk normality test
data:  residuals(model)
W = 0.97358, p-value = 0.4637
--------------------------------
studentized Breusch-Pagan test
data:  model
BP = 1.7063, df = 2, p-value = 0.4261
--------------------------------
Durbin-Watson test
data:  model
DW = 1.9923, p-value = 0.4357
```