

Statistical modelling for biological data with R

Day 5 - Mixed Model

Marie-Pierre Etienne

<https://marieetienne.github.io>

Novembre 2019



Plan

① Introduction

② Exemple

Plan

① Introduction

② Exemple

Plan

1 Introduction

Deux situations, même données, deux modèles

Mesure de l'héritabilité d'un caractère

Merures répétées

Estimation

Tests

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

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

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

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

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

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

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 aurait pu choisir 5 autres voitures. Le choix de la voiture parmi toutes les voitures possibles est aléatoire.

Plan

1 Introduction

Deux situations, même données, deux modèles

Mesure de l'héritabilité d'un caractère

Merures répétées

Estimation

Tests

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

Modèle

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

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

La variance vaut

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

et la covariance

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

Héritabilité - 2

Modèle

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

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

Composante de la variance σ_A^2 et σ^2 .

Ecriture matricielle :

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

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

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

Plan

1 Introduction

Deux situations, même données, deux modèles

Mesure de l'héritabilité d'un caractère

Merures répétées

Estimation

Tests

Mesures répétées

Effet d'un régime sur la prise de poids d'un animal au cours du temps. Plusieurs animaux (indiqués 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 ($t = 1, \dots, T$). 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

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

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

Version générique du modèle

Modèle

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

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

Plan

1 Introduction

Deux situations, même données, deux modèles

Mesure de l'héritabilité d'un caractère

Merures répétées

Estimation

Tests

Des idées sur l'estimation des paramètres

- Cas Σ connu (Moindres carrés généralisé)

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

- Cas Σ 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 $T\mathbf{X} = 0$ et donc $T\mathbf{Y} = T\mathbb{E}$ ne dépend plus de $\boldsymbol{\theta}$ et $\text{var}(T\mathbf{E}) = T\text{var}(\mathbf{E})T'$.

Prédictions des effets aléatoires

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

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

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

La loi conditionnelle du vecteur de \mathbf{U}

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

Plan

1 Introduction

Deux situations, même données, deux modèles

Mesure de l'héritabilité d'un caractère

Merures répétées

Estimation

Tests

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

1 Introduction

2 Exemple

Les samares du frêne

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

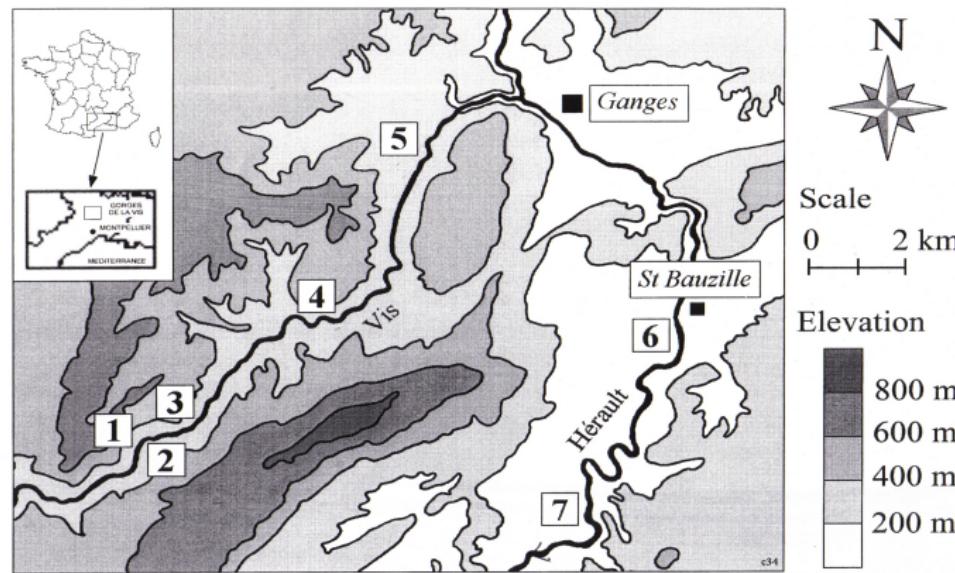


Figure 1: Samares

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 cm^2) de chaque samares et calculé l'indice de dispersion

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

Questions

- Existe-t-il 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.

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.