

Introduction aux MLMs et estimation de (co)variances génétiques

Jemay Salomon

UMR GQE Le Moulon

Université Paris–Saclay, INRAE, CNRS, AgroParisTech

Outline

Modèle linéaire simple

- Notations

- Vraisemblance

- Estimation de β

- Estimation de σ_e^2

Modèle linéaire mixte

- Notations

- Vraisemblance

- Estimation de β et prediction de \mathbf{u}

- Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Notations

$$y = X\beta + \epsilon$$

- y : vecteur de dimension n contenant les observations (réponses)
- X : matrice d'incidence de dimension $n \times p$ des variables explicatives (prédicteurs)
- β : vecteur de dimension p contenant les paramètres correspondant aux effets des variables explicatives sur la moyenne des observations
- ϵ : vecteur de dimension n contenant les erreurs modélisées par des variables aléatoires
- n : nombre d'observations

Notations

- i : indice indiquant la i -ème observation, donc $i \in \{1, \dots, n\}$
- p : nombre de variables explicatives; on suppose $n > p$ (pas toujours le cas!)
- j : indice indiquant la j -ème variable explicative, donc $j \in \{1, \dots, p\}$
- R : matrice de dimension $n \times n$ de variance-covariance des erreurs (supposée définie positive, donc inversible); $R = \sigma_e^2 I_n$ où I_n correspond à la matrice identité $n \times n$

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Vraisemblance

- ▶ données : $\mathcal{D} = \{y \mid X\}$
- ▶ paramètres : $\Theta = \{\beta, \sigma_e^2\}$

$$y = X\beta + \epsilon \quad \text{avec} \quad \epsilon \stackrel{\text{iid}}{\sim} \mathcal{N}_{\mathbb{R}^n}(0, \sigma_e^2 I_n)$$

$$\Leftrightarrow y \mid X, \beta, \sigma_e^2 \sim \mathcal{N}(X\beta, \sigma_e^2 I_n)$$

L'espérance et la variance-covariance des observations sont :

- ▶ $\mathbb{E}[y \mid X, \beta] = X\beta$
- ▶ $\text{Cov}[y \mid \sigma_e^2] = \sigma_e^2 I_n$

Vraisemblance

Les observations étant modélisées comme indépendantes conditionnellement aux prédicteurs, on peut utiliser leur produit :

$$\mathcal{L}(\Theta; \mathcal{D}) = f(y \mid X, \beta, \sigma_e^2) = \prod_{i=1}^n f(y_i \mid X_{i\cdot}, \beta, \sigma_e^2)$$

où $X_{i\cdot}$ étant la i -ème ligne de X .

En pratique, on utilise la **log-vraisemblance** ℓ , et le produit se transforme en somme :

$$\begin{aligned}\ell(\Theta; \mathcal{D}) &= \sum_{i=1}^n \log f(y_i \mid x_i, \Theta) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma_e^2) - \frac{1}{2\sigma_e^2} (y - X\beta)^\top (y - X\beta)\end{aligned}$$

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Estimation de β

Distance à minimiser

La "longueur" du vecteur d'erreurs $\epsilon = y - X\beta$.

Méthode des moindres carrés ordinaires (OLS)

Identifier $\hat{\beta}$ qui minimise la somme des carrés des erreurs (ESS) :

$$\hat{\beta}_{\text{OLS}} = \arg \min_{\beta} \text{ESS}$$

$$\text{ESS} = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - X_{i\cdot}\beta)^2$$

Estimation de β

Forme matricielle

$$\text{ESS} = (y - X\beta)^\top (y - X\beta) = \|y - X\beta\|_2^2$$

Dérivées

- ▶ Dérivée première : $\frac{d\text{ESS}}{d\beta}(\beta) = -2X^\top (y - X\beta)$
- ▶ Dérivée seconde : $\frac{d^2\text{ESS}}{d\beta^2}(\beta) = 2X^\top X$

Convexité

La dérivée seconde étant positive, l'ESS est convexe : il existe un minimum global.

Estimation de β

Équations normales de Gauss

Annulation de la dérivée première :

$$\frac{d\text{ESS}}{d\beta}(\hat{\beta}) = 0 \quad \Leftrightarrow \quad X^{\top} X \hat{\beta} = X^{\top} y$$

Estimation de β

$$\hat{\beta} = (X^{\top} X)^{-} X^{\top} y$$

où $-$ désigne l'inverse généralisée.

Grandeurs dérivées

- ▶ Valeurs ajustées : $\hat{y} = X\hat{\beta}$
- ▶ Résidus : $\hat{\epsilon} = y - X\hat{\beta} = y - \hat{y}$

Estimation de β

Projection orthogonale

$\hat{\beta}$ minimise la distance entre y et $C(X) = X\beta$: projection orthogonale de y sur $C(X)$.

$$X^{\top}(y - X\beta) = 0$$

Matrice de projection (hat matrix) : $P = X(X^{\top}X)^{-1}X^{\top}$

Maximum de vraisemblance

Sous l'hypothèse de normalité, on retrouve les mêmes équations :

$$\frac{\partial \ell}{\partial \beta}(\hat{\beta}) = 0 \quad \Leftrightarrow \quad (X^{\top}X)\hat{\beta} = X^{\top}y$$

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Estimation de σ_e^2

Somme des carrés résiduelle (RSS)

$$\text{RSS} = \sum_i \hat{\epsilon}_i^2 = (y - \hat{y})^\top (y - \hat{y}) = y^\top (I_n - P)y$$

avec $P = X(X^\top X)^{-1}X^\top$ (matrice de projection, idempotente et symétrique).

Espérance de RSS

$$\begin{aligned}\mathbb{E}[\text{RSS}] &= \mathbb{E}[\text{tr}[y^\top (I_n - P)y]] \\ &= \text{tr}[(I_n - P)\mathbb{E}[yy^\top]] \\ &= \text{tr}[(I_n - P)(X\beta\beta^\top X^\top + \sigma_e^2 I_n)] \\ &= \sigma_e^2 \text{tr}[(I_n - P)] \\ &= \sigma_e^2 (n - r(X))\end{aligned}$$

Estimation de σ_e^2

Estimateur sans biais (OLS)

$$S_{\text{OLS}}^2 = \frac{\text{RSS}}{n - r(X)}$$

où $r(X) = p$ quand X est de plein rang.

Estimation de σ_e^2

Estimateur du maximum de vraisemblance

$$\frac{\partial \ell}{\partial \sigma_e^2}(\hat{\sigma}_e^2) = 0 \quad \Leftrightarrow \quad \hat{\sigma}_{e\text{ML}}^2 = \frac{1}{n}(y - X\hat{\beta})^\top (y - X\hat{\beta}) = \frac{\text{RSS}}{n}$$

Biais de l'estimateur ML

$$\mathbb{E}[\hat{\sigma}_{e\text{ML}}^2] = \frac{n - r(X)}{n} \sigma_e^2$$

L'estimateur ML est **biaisé** (sous-estime σ_e^2).

Estimation de σ_e^2

Interprétation

- ▶ **Estimateur OLS** : sans biais, divise par $n - r(X)$ (degrés de liberté)
- ▶ **Estimateur ML** : biaisé, divise par n (néglige l'incertitude sur $\hat{\beta}$)
- ▶ Le biais vient de la perte de degrés de liberté due à l'estimation de β

En pratique

On utilise généralement l'estimateur OLS : $S^2 = \frac{RSS}{n-p}$ (quand X est de plein rang).

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Notations

$$y = X\beta + Zu + \epsilon$$

Notations complémentaires:

- ▶ q : nombre de variables aléatoires pour structurer la variance-covariance des observations, avec $n > q$
- ▶ k : indice indiquant la k -ème variable aléatoire, donc $k \in \{1, \dots, q\}$
- ▶ Z : matrice d'incidence de dimension $n \times q$ reliant les y_i aux u_k
- ▶ G : matrice de variance-covariance de dimension $q \times q$ du vecteur u , telle que $G = \sigma_u^2 A$ où A est connue et définie positive
- ▶ φ : vecteur des composantes de la variance, ici égal à $(\sigma_u^2, \sigma_e^2)^\top$
- ▶ H : matrice de variance-covariance de dimension $n \times n$ dépendant de φ

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Vraisemblance

- ▶ données : $\mathcal{D} = \{y \mid X, Z\}$
- ▶ paramètres : $\Theta = \{\beta, \sigma_u^2, \sigma_e^2\}$

Vraisemblance :

$$y = X\beta + Zu + \epsilon$$

avec $u \sim \mathcal{N}(0, G)$, $\epsilon \stackrel{\text{iid}}{\sim} \mathcal{N}(0, R)$ et $\text{Cov}[u, \epsilon] = 0$

De plus, on a ici :

- ▶ $G = \sigma_u^2 A$
- ▶ $R = \sigma_e^2 V$

Vraisemblance du modèle mixte (suite)

De manière générale, on peut donc écrire :

$$y \mid \beta, u, R \sim \mathcal{N}(X\beta + Zu, R)$$

Après intégration des u_k (on dit aussi qu'elles ont été "marginalisées"), on obtient :

$$y \mid \beta, G, R \sim \mathcal{N}(X\beta, ZGZ^\top + R)$$

L'espérance et la variance-covariance des observations sont bien des fonctions linéaires de paramètres :

- ▶ $\mathbb{E}[y \mid \beta] = X\beta$
- ▶ $\text{Cov}[y \mid \varphi] := H = ZGZ^\top + R$
(égale ici à $\sigma_u^2 ZAZ^\top + \sigma_e^2 V$)

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Estimation de β et prédiction de u

Deux approches

- ▶ **Paradigme fréquentiste :**
 - ▶ $\hat{\beta}$: BLUE (Best Linear Unbiased Estimator)
 - ▶ \hat{u} : BLUP (Best Linear Unbiased Predictor)
- ▶ **Paradigme bayésien :** mêmes formules, interprétation différente

Estimation de β et prédiction de u

Log-densité conjointe

$$\begin{aligned}\log f(y, u \mid \beta, G, R) &= \log f(y \mid \beta, u, R) + \log f(u \mid G) \\ &= -\frac{1}{2} \left[\begin{aligned} &n \log 2\pi + \log |R| + (y - X\beta - Zu)^\top R^{-1} (y - X\beta - Zu) \\ &+ q \log 2\pi + \log |G| + u^\top G^{-1} u \end{aligned} \right]\end{aligned}$$

Estimation de β et prediction de u

Estimateur BLUE de β

$$\hat{\beta} = (X^{\top} H^{-1} X)^{-1} X^{\top} H^{-1} y$$

Moindres carrés généralisés

Prédicteur BLUP de u

$$\hat{u} = GZ^{\top} H^{-1} (y - X\hat{\beta})$$

Outline

Modèle linéaire simple

Notations

Vraisemblance

Estimation de β

Estimation de σ_e^2

Modèle linéaire mixte

Notations

Vraisemblance

Estimation de β et prediction de \mathbf{u}

Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Estimation de σ_u^2 et σ_e^2

Problème du maximum de vraisemblance (ML)

- ▶ Comme pour le modèle linéaire classique, ML produit des estimateurs **biaisés** de σ_u^2 et σ_e^2
- ▶ Nécessité d'une méthode spécifique

Maximum de vraisemblance restreinte (ReML)

Méthode développée pour estimer les composantes de la variance :

- ▶ Décompose la vraisemblance en deux parties
- ▶ Une partie ne dépend que des variables aléatoires u sans β

Principe du ReML

Élimination de β

- ▶ On cherche des vecteurs tels que $v^\top X = 0$
- ▶ Il existe $n - r(X)$ vecteurs linéairement indépendants
- ▶ Exemple : $S = I_n - X(X^\top X)^{-1}X^\top$ vérifie $SX = 0$

Vraisemblance restreinte

$$K^\top y \sim \mathcal{N}(0, K^\top H(\varphi)K)$$

avec $K^\top K = I_n$ et $K^\top X = 0$

Procédure

1. Maximiser la vraisemblance restreinte pour obtenir $\hat{\varphi}$
2. Calculer $\hat{\beta}|\hat{\varphi}$ (BLUE empirique)
3. Calculer $\hat{u}|\hat{\varphi}$ (BLUP empirique)

Algorithme EM

Principe général

Algorithme pour modèles avec "données manquantes" (variables latentes) :

- ▶ **Données manquantes** : $z = (\beta^\top, u^\top)^\top$
- ▶ **Données complètes** : $x = (y^\top, z^\top)^\top$

Fonction Q

$$Q(\varphi; \varphi^{(t)}) = \mathbb{E}_{z|y, \varphi^{(t)}}[\ell(\varphi; x)]$$

Espérance de la log-vraisemblance complète

Algorithme EM

Étapes de l'algorithme

1. **E (Expectation)** : Calculer $Q(\varphi; \varphi^{(t)})$
2. **M (Maximization)** : Maximiser Q par rapport à φ :

$$\varphi^{(t+1)} = \arg \max_{\varphi} Q(\varphi; \varphi^{(t)})$$

Vraisemblance complète du modèle mixte

$$L(\varphi; \mathbf{x}) \propto f(\mathbf{y} | \boldsymbol{\beta}, \mathbf{u}, \sigma_e^2) \times f(\mathbf{u} | \sigma_u^2)$$

$$\ell(\sigma_e^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_e^2 - \frac{\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}}{2\sigma_e^2}$$

$$\ell(\sigma_u^2) = -\frac{q}{2} \log 2\pi - \frac{q}{2} \log \sigma_u^2 - \frac{\mathbf{u}^\top \mathbf{u}}{2\sigma_u^2}$$

Références

- ▶ **Dagnelie (2012)** : *Principes d'expérimentation: planification des expériences et analyses de leurs résultats*
- ▶ **Dempfle (1977)** : *Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens*
- ▶ **Foulley (2002)** : *Méthodes du maximum de vraisemblance en modèle linéaire mixte*
- ▶ **Foulley (2002)** : *Algorithme EM : théorie et application au modèle mixte*
- ▶ **Robert (2001)** : *L'analyse statistique bayésienne*

Références

- ▶ **Gumedze et Dunne (2011)** : *Parameter estimation and inference in the linear mixed model*
- ▶ **Henderson (1950)** : *Estimation of genetic parameters*
- ▶ **Henderson et al. (1959)** : *The Estimation of Environmental and Genetic Trends from Records Subject to Culling*
- ▶ **Verbyla (1990)** : *A Conditional Derivation of Residual Maximum Likelihood*
- ▶ **Wand (2002)** : *Vector differential calculus in statistics*

Outline

Modèle linéaire simple

- Notations

- Vraisemblance

- Estimation de β

- Estimation de σ_e^2

Modèle linéaire mixte

- Notations

- Vraisemblance

- Estimation de β et prediction de \mathbf{u}

- Estimation de σ_u^2, σ_e^2

Activité - 45-60 min

Activité

- Code R