TD: Estimation de (co)variances génétiques

Jemay Salomon

UMR GQE Le Moulon
Université Paris–Saclay, INRAE, CNRS, AgroParisTech

Outline

Outline

Outline

Notations

$$\mathrm{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- y : vecteur de dimension n contenant les observations (réponses)
- X : matrice d'incidence de dimension $n \times p$ des variables explicatives (prédicteurs)
- $\boldsymbol{\beta}$: vecteur de dimension p contenant les paramètres correspondant aux effets des variables explicatives sur la moyenne des observations
- $\boldsymbol{\epsilon}$: 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 $i \in \{1, \ldots, n\}$
- p : nombre de variables explicatives; on suppose $n > p$ (pas toujours le cas!)
- j : indice indiquant la j-ème variable explicative, donc $j \in \{1, \ldots, p\}$
- R : matrice de dimension $n \times n$ de variance-covariance des erreurs (supposée définie positive, donc inversible); $R = \sigma^2 I_n$ où I_n correspond à la matrice identité $n \times n$

Outline

Vraisemblance

- **données** : $\mathcal{D} = \{\mathrm{y} \mid X\}$
- **paramètres** : $\Theta = \{\boldsymbol{\beta}, \sigma^2\}$

$$\mathrm{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{avec} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$$

$$\Leftrightarrow \quad \mathrm{y} \mid X, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$$

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

- $\mathbb{E}[\mathrm{y} \mid X, \boldsymbol{\beta}] = X\boldsymbol{\beta}$
- $\mathrm{Cov}[\mathrm{y} \mid \sigma^2] = \sigma^2 I_n$

Vraisemblance

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

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

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 :

$$\ell(\Theta; \mathcal{D}) = \sum_{i=1}^{n} \log f(y_i \mid \mathrm{x}_i, \Theta)$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathrm{y} - X\boldsymbol{\beta})^\top(\mathrm{y} - X\boldsymbol{\beta})$$

Outline

Estimation de β

Distance à minimiser

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

Méthode des moindres carrés ordinaires (OLS)

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

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

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

Estimation de β

Forme matricielle

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

Dérivées

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

Convexité

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

Estimation de $\boldsymbol{\beta}$

Équations normales de Gauss

Annulation de la dérivée première :

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

Estimation de $\boldsymbol{\beta}$

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

où $^-$ désigne l'inverse généralisée.

Grandeurs dérivées

- **Valeurs ajustées** : $\hat{\mathrm{y}} = X\hat{\boldsymbol{\beta}}$
- **Résidus** : $\hat{\boldsymbol{\epsilon}} = \mathrm{y} - X\hat{\boldsymbol{\beta}} = \mathrm{y} - \hat{\mathrm{y}}$

Estimation de β

Projection orthogonale

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

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

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

Maximum de vraisemblance

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

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

Outline

Estimation de σ^2

Somme des carrés résiduelle (RSS)

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

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

Espérance de RSS

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

Estimation de σ^2

Estimateur sans biais (OLS)

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

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

Estimation de σ^2

Estimateur du maximum de vraisemblance

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

Biais de l'estimateur ML

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

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

Estimation de σ^2

Interprétation

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

En pratique

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

Outline

Outline

Notations

$$\mathrm{y} = X\boldsymbol{\beta} + Zu + \boldsymbol{\epsilon}$$

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
- $\boldsymbol{\varphi}$: vecteur des composantes de la variance, ici égal à $(\sigma_u^2, \sigma^2)^\top$
- H : matrice de variance-covariance de dimension $n \times n$ dépendant de $\boldsymbol{\varphi}$

Outline

Vraisemblance

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

Vraisemblance :

$$\mathrm{y} = X\boldsymbol{\beta} + Z\mathrm{u} + \boldsymbol{\epsilon}$$

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

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 :

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

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

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

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

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

Outline

Estimation de β et prédiction de u

Deux approches

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

Estimation de $\boldsymbol{\beta}$ et prédiction de u

Log-densité conjointe

$$\log f(\mathrm{y}, \mathrm{u} \mid \boldsymbol{\beta}, G, R) = \log f(\mathrm{y} \mid \boldsymbol{\beta}, \mathrm{u}, R) + \log f(\mathrm{u} \mid G)$$

$$= -\frac{1}{2}\left[\begin{array}{l} n \log 2\pi + \log |R| + (\mathrm{y} - X\boldsymbol{\beta} - Z\mathrm{u})^\top R^{-1} (\mathrm{y} - X\boldsymbol{\beta} - Z\mathrm{u}) \\ +q \log 2\pi + \log |G| + \mathrm{u}^\top G^{-1} \mathrm{u} \end{array}\right]$$

Estimation de β et prediction de u

Estimateur BLUE de $\boldsymbol{\beta}$

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

Moindres carrés généralisés

Prédicteur BLUP de u

$$\hat{\mathrm{u}} = G Z^\top H^{-1} (\mathrm{y} - X \hat{\boldsymbol{\beta}})$$

Outline

Estimation de σ_u^2 et σ^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 $\boldsymbol{\beta}$

Principe du ReML

Élimination de $\boldsymbol{\beta}$

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

Vraisemblance restreinte

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

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

Procédure

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

Algorithme EM

Principe général

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

- **Données manquantes** : $\mathrm{z} = (\boldsymbol{\beta}^\top, \mathrm{u}^\top)^\top$
- **Données complètes** : $\mathrm{x} = (\mathrm{y}^\top, \mathrm{z}^\top)^\top$

Fonction Q

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

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

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

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

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

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

Activité

- Code R