Le modèle linéaire mixte : L2M

... ou "modèle linéaire à effets aléatoires", "modèle à composantes de la variance".

▷ Écriture du modèle

$$Y = X\beta + ZU + \epsilon \;=\; X\beta + \sum_{k=1}^{K} Z_k U_k + \epsilon$$

- $\epsilon \sim \mathcal{N}_{\mathbb{R}^N}(0, R = \theta_0 V_0)$, V_0 est une matrice connue.

- $U_k \sim \mathcal{N}_{\mathbb{R}^{q_k}}(0, \theta_k G_k)$ pour tout $k = 1, \ldots, K$, les G_k sont des matrices connues.
$U_1, U_2, \ldots, U_k$ sont k effets aléatoires non observés, ils sont indépendants entre eux et indépendants de ϵ.

$U \sim \mathcal{N}_{\mathbb{R}^q}(0, G_\theta)$ où $q = \sum_{k=1}^{K} q_k$

▷ Propriétés du modèle (1)

- conditionnellement aux effets aléatoires

$$\begin{aligned} E(Y|U) &= X\beta + ZU \\ Var(Y|U) &= R = \theta_0 V_0 \end{aligned}$$

- marginalement

$$\begin{aligned} E(Y) &= X\beta \\ Var(Y) &= \Gamma_\theta = R + ZG_\theta Z' = \sum_{k=0}^{K} \theta_k V_k \end{aligned}$$

où $V_k = Z_k G_k Z_k'$

les θ_k ($\in \mathbb{R}^+$) sont appelés composantes de la variance.

▷ Propriétés du modèle (2)

- $Y|U \sim \mathcal{N}_N\left(X\beta + ZU, R\right)$
- $Y \sim \mathcal{N}_N\left(X\beta, \Gamma_\theta\right)$
- $\begin{pmatrix} Y \\ U \end{pmatrix} \sim \mathcal{N}_{N+q}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} R + ZG_\theta Z' & ZG_\theta \\ G_\theta Z' & G_\theta \end{pmatrix}\right)$
- $U|Y \sim \mathcal{N}_q\left(G_\theta Z' \Gamma_\theta^{-1}(y - X\beta), G_\theta - G_\theta Z' \Gamma_\theta^{-1} Z G_\theta\right)$.

▷ Interprétation du modèle

• décomposition des erreurs du modèle

on extrait des erreurs une part de variabilité des données que l'on modélise comme provenant d'un effet aléatoire identifié

• décomposition de la partie explicative du modèle

on ajoute aux effets fixes des effets aléatoires que l'on n'observe pas directement et dont les niveaux réalisés au sein des données ne sont considérés que comme des tirages aléatoires d'une famille de valeurs

⇒ intégration de plusieurs sources d'aléa et précision sur la structure de dépendance des données

▷ **Exemple** : Suivi de croissance de pommiers

On mesure pendant 4 années consécutives la taille de la pousse annuelle de pommiers issus de 2 variétés différentes et soumis à 4 conditions différentes de croissance. Pour chaque condition, 10 arbres de chaque variété ont été mesurés.

Dans cette situation :
• *variété* et *condition d'expérience* sont 2 facteurs à effet fixe pour lesquels il nous intéresse de mesurer l'effet de chacun de leur niveau. Question : est-ce qu'une variété a une pousse plus rapide qu'une autre ? est-ce que telle condition expérimentale entraîne une pousse plus importante ?

• sur chaque arbre, on répète l'observation 4 fois, ces 4 données vont être soumises à un effet propre à l'arbre, qui les rend dépendantes entre elles mais dont la mesure ne nous intéresse pas spécifiquement. Le niveau de chaque arbre est introduit dans le modèle comme une réalisation aléatoire extraite parmi tous les effets d'arbre possibles.

▷ Effet fixe | Effet aléatoire

$$Y = \underbrace{X\beta}_{\text{partie effets fixes}} + \underbrace{ZU}_{\text{partie effets aléatoires}} + \epsilon$$

Effet fixe

- nombre "fini et réduit" de niveaux possibles
- intérêt dans chaque niveau du facteur

Effet aléatoire

- nombre "infini ou très important" de niveaux possibles
- intérêt dans la variabilité de ces niveaux

▷ Système direct des équations ML

Rappel : $l(\beta,\theta) = -\frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|\Gamma_\theta|) - \frac{1}{2}(y - X\beta)'\,\Gamma_\theta^{-1}\,(y - X\beta)$

En dérivant par rapport à β et par rapport à chaque θ_j composante de la variance, on obtient le système :

$$\left\{\begin{array}{lcll} X'\,{\Gamma_\theta}^{-1}\,X\,\beta & = & X'\,\Gamma_\theta^{-1}\,y & \\ \mathrm{tr}({\Gamma_\theta}^{-1}V_j) & = & y'\,P_\theta\,V_j\,P_\theta\,y & \qquad j = 0,...,K \end{array}\right.$$

$\Rightarrow$ système d'équations non linéaires en θ (avec contrainte) dont la résolution directe n'est pas aisée.

Résolution itérative du système équivalent :

$$\left\{\begin{array}{rcl} X'\,{\Gamma_\theta}^{-1}\,X\,\beta & = & X'\,\Gamma_\theta^{-1}\,y \\ \Big(\mathrm{tr}({\Gamma_\theta}^{-1}V_j{\Gamma_\theta}^{-1}V_k)\Big)_{j,k=0,...,K} \begin{pmatrix}\theta_0\\ \vdots \\ \theta_K\end{pmatrix} & = & \Big(y'\,P_\theta\,V_j\,P_\theta\,y\Big)_{j=0,...,K} \end{array}\right.$$

$\Rightarrow$ inversion de Γ_θ reste difficile

▷ Équations de Henderson

Origine : dans ses travaux en sélection animale dans les années 60, intérêt particulier pour la prédiction de U

Objectif double :
- prédiction BLUP (Best Linear Unbiased Predictor) de U
- estimation BLUE (Best Linear Unbiased Estimator) de β

$\Rightarrow$ estimation des composantes de la variance comme sous-produit

Outil : la vraisemblance "jointe"

$$\begin{aligned} l(\beta, U; y, U) \;=\; & -\tfrac{N+q}{2}\ln(2\pi) - \tfrac{1}{2}\ln(|R|) - \tfrac{1}{2}\ln(|G_\theta|) \\ & -\tfrac{1}{2}\{(y - X\beta - ZU)'R^{-1}(y - X\beta - ZU) + U'G_\theta^{-1}U\} \end{aligned}$$

Par maximisation en β et U, on obtient le système :

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G_\theta^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ U \end{pmatrix} = \begin{pmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{pmatrix}.$$

appelées *équations du modèle mixte*
$\Rightarrow$ nécessitent l'inversion de R et G_θ (souvent diagonales) et du système de taille $p + q$.

Ce système est équivalent à :

$$\begin{cases} X'\Gamma_\theta^{-1}X\,\beta &= X'\Gamma_\theta^{-1}\,y \\ U &= G_\theta Z'\Gamma_\theta^{-1}\,(y - X\beta) \;=\; E(U|y)\,. \end{cases}$$

Remarque : sans la présence de G_θ^{-1}, estimation du modèle à effet fixe pour U.

▷ Estimation des composantes de la variance

Harville en 77 propose d'utiliser les solutions du système de Henderson pour construire les estimateurs ML ou REML de θ dans un schéma itératif :

ML
$$\left[\begin{array}{rcl} \theta_j^{(m+1)} & = & \dfrac{U_j'^{(m)}\, G_j^{-1}\, U_j^{(m)}}{q_j - \dfrac{\mathrm{tr}(C_{jj}^{*(m)})}{\theta_j^{(m)}}} \\ \theta_0^{(m+1)} & = & \dfrac{(y - X\beta^{(m)} - ZU^{(m)})'\, V_0^{-1}\, (y - X\beta^{(m)} - ZU^{(m)})}{N - \sum_{j=1}^{K}(q_j - \dfrac{\mathrm{tr}(G_j^{-1}C_{jj}^{*(m)})}{\theta_j^{(m)}})} \end{array}\right.$$

où C^* : inverse de la matrice formée par les q dernières lignes et colonnes de la matrice des coefficients du système de Henderson

C_{jj}^* : $j^{\text{ème}}$ sous matrice de C^*, correspondant au $j^{\text{ème}}$ effet aléatoire.

REML

$$\left[\begin{array}{rcl} \theta_j^{(m+1)} & = & \dfrac{U_j'^{(m)}\, G_j^{-1}\, U_j^{(m)}}{q_j - \dfrac{\mathrm{tr}(C_{jj}^{(m)})}{\theta_j^{(m)}}} \\ \theta_0^{(m+1)} & = & \dfrac{(y - X\beta^{(m)} - ZU^{(m)})'\, V_0^{-1}\, (y - X\beta^{(m)} - ZU^{(m)})}{N - rang(X) - \sum_{j=1}^{K}(q_j - \dfrac{\mathrm{tr}(G_j^{-1}C_{jj})}{\theta_j^{(m)}})} \end{array}\right. \qquad (1)$$

où C : est la matrice formée des q dernières lignes et colonnes de l'inverse de la matrice des coefficients du système de Henderson

C_{jj} : $j^{\text{ème}}$ sous matrice de C, correspondant au $j^{\text{ème}}$ effet aléatoire.

▷ Algorithme EM

- données *observées* : y
  données *manquantes* : u
  données *complètes* : $x = (y', u')'$

• on cherche des statistiques exhaustives, fonction des données complètes $t(x)$ qui permettent de réaliser l'estimation des paramètres inconnus $\gamma = (\beta, \theta)$

• algorithme : à chaque itération $[t]$ pour une valeur courante des paramètres $\gamma^{[t]}$, il se décompose en 2 étapes :
- étape E : calculer l'espérance conditionnelle de $t(x)$ sachant les données observées y et la valeur $\gamma^{[t]}$.
- étape M : maximiser la vraisemblance des données complètes en remplaçant les statistiques exhaustives par l'espérance conditionnelle et obtenir ainsi $\gamma^{[t+1]}$.

• Statistiques exhaustives

si les U_k étaient observées, vraisemblance complétée

$$\begin{aligned} l(\beta,\theta;y,u) &= -\tfrac{N+q}{2}\ln(2\pi) - \tfrac{1}{2}\ln(|R|) - \tfrac{1}{2}\ln(|G_\theta|) \\ &\quad -\tfrac{1}{2}\{(y-X\beta-ZU)'R^{-1}(y-X\beta-ZU) + U'G_\theta^{-1}U\} \\ &= -\tfrac{N+q}{2}\ln(2\pi) - \tfrac{1}{2}\ln(|R|) - \tfrac{1}{2}\textstyle\sum_{k=1}^{K}\ln(\theta_k^{q_k}|G_k|) \\ &\quad -\tfrac{1}{2}\{(y-X\beta-ZU)'R^{-1}(y-X\beta-ZU) + \textstyle\sum_{k=1}^{K}\theta_k^{-1}U_k'G_k^{-1}U_k\} \end{aligned}$$

les estimations des paramètres seraient obtenues par :

$$\widehat{\beta} = (X'V_0^{-1}X)^{-1}X'V_0^{-1}(y-\textstyle\sum_{k=1}^{K}Z_kU_k)$$

$$\widehat{\theta_k} = \frac{U_k'G_k^{-1}U_k}{q_k}$$

$$\widehat{\theta_0} = \frac{(y-X\beta-ZU)'V_0^{-1}(y-X\beta-ZU)}{N}$$

les statistiques exhaustives sont : $U_k'G_k^{-1}U_k$, $y-\sum_{k=1}^{K}Z_kU_k$ et $(y-X\beta-ZU)'V_0^{-1}(y-X\beta-ZU)$

- Espérances conditionnelles a posteriori :

$$E(U_k' G_k^{-1} U_k | y) = \theta_k^2 (y - X\beta)' \Gamma_\theta^{-1} V_k \Gamma_\theta^{-1} (y - X\beta) + q_k \theta_k - \theta_k^2 \mathrm{tr}(\Gamma_\theta^{-1} V_k)$$

$$E(y - \textstyle\sum_{k=1}^{K} Z_k U_k | y) = X\beta + \theta_0 V_0 \Gamma_\theta^{-1} (y - X\beta)$$

$$\begin{aligned} &E((y - X\beta - ZU)' V_0^{-1} (y - X\beta - ZU) | y) \\ &\quad = \theta_0^2 (y - X\beta)' \Gamma_\theta^{-1} V_0 \Gamma_\theta^{-1} (y - X\beta) + \theta_0 N - \theta_0^2 \mathrm{tr}(\Gamma_\theta^{-1} V_0) \end{aligned}$$

D'où l'algorithme EM pour ML:

$$\begin{cases} q_k\theta_k^{[t+1]} &= \theta_k^{2[t]}(y-X\beta^{[t]})'\Gamma_{\theta^{[t]}}^{-1}V_k\Gamma_{\theta^{[t]}}^{-1}(y-X\beta^{[t]}) + q_k\theta_k^{[t]} - \theta_k^{2[t]}\mathrm{tr}(\Gamma_{\theta^{[t]}}^{-1}V_k) \\ X\beta^{[t+1]} &= (X'V_0^{-1}X)^{-1}X'V_0^{-1}[X\beta^{[t]} + \theta_0^{[t]}V_0\Gamma_{\theta^{[t]}}^{-1}(y-X\beta^{[t]})] \\ N\theta_0^{[t+1]} &= \theta_0^{2[t]}(y-X\beta^{[t]})'\Gamma_{\theta^{[t]}}^{-1}V_0\Gamma_{\theta^{[t]}}^{-1}(y-X\beta^{[t]}) + \theta_0^{[t]}N - \theta_0^{2[t]}\mathrm{tr}(\Gamma_{\theta^{[t]}}^{-1}V_0) \end{cases}$$

Remarque : la 1^{re} partie de l'algorithme peut aussi s'écrire :

$$\left\{ q_k\theta_k^{[t+1]} = \theta_k^{2[t]}y'P_{\theta^{[t]}}V_kP_{\theta^{[t]}}y + q_k\theta_k^{[t]} - \theta_k^{2[t]}\mathrm{tr}(\Gamma_{\theta^{[t]}}^{-1}V_k) \right.$$

et en remplaçant $\Gamma_{\theta^{[t]}}^{-1}$ par $P_{\theta^{[t]}}$ on obtient l'algorithme EM pour REML.

Vers les GL2M ...

ou "modèle linéaire généralisé mixte"

- loi autre que la loi gaussienne dans la famille exponentielle

hypothèse posée sur la loi conditionnelle de Y sachant U

- prédicteur linéaire : $\eta = X\beta + ZU$

- $E(Y|U) = g^{-1}(\eta)$
  $Var(Y|U)$ est fonction de l'espérance conditionnelle

$\Rightarrow$ Problème : loi marginale de Y ???