Modélisation des données longitudinales: Modèles linéaires mixte (2/2)

A. Latouche

Notions abordées

- Estimation dans le modèle linéaire mixte
- Prédiction
- Exercices

- Formulation pour 1 variable explicative

$$Y_{ij} = (\alpha + u_{0i}) + (\beta + u_{1i})X_{ij} + \epsilon_{ij},$$

where

- u_{0i} is the random intercept: modelling baseline individual heterogeneity (*e.g* difference in cognitive ability at birth)
- u_{1i} is the random slope: modelling individual heterogeneity in the X-Y relationship (*e.g*. differences in the age-related evolution of cognitive ability)
- Main assumption: both u_{0i} and u_{1i} are assumed Gaussian centered on 0 with variance σ_0, σ_1 respectively.

Ecriture Matricielle du modèle linéaire Mixte

$$Y = X\beta + ZA + U$$

- Y réponse dans R^n
- X matrice $n \times p$
- β vecteur des p effets fixes
- Z est une matrice (connue) de taille $n \times q$
- A est un vecteur Gaussien de R^q :

Il vient

- $E(Y) = X\beta$
- $Var(Y) = V = Var(ZA) + Var(U) = \sum_{k=1} \sigma_k^2 Z_k Z_k' + \sigma^2 I$

donc

$$Y \sim N(X\beta, V)$$

Ecriture matricielle : dimension des composantes du modèle

- n patients voient q médecins et lors des visites une variables quantitative continue est mesurée (score de mobilité Y), p vecteurs des effets fixes (variables Age (en années, status marital, sexe)
- Les données ne sont pas (forcément) équilibrées : Tous les médecins ne voient pas le même nombre de patients

$$\overbrace{\mathbf{y}}^{\text{n x 1}} = \overbrace{\underbrace{\mathbf{X}}_{\text{n x p}} \underbrace{\beta}_{\text{p x 1}}}^{\text{n x 1}} + \overbrace{\underbrace{\mathbf{Z}}_{\text{n x q}} \underbrace{\gamma}_{\text{q x 1}}}^{\text{n x 1}} + \overbrace{\varepsilon}^{\text{n x 1}}$$

$$n = \sum_{j}^{q} n_j$$

Estimation des effets fixes β

$$Y = X\beta + ZA + U$$

Si les données sont équilibrées

$$\hat{\beta} = (X'X)^{-1}X'Y$$

il n'est pas nécessaire d'estimer la matrice de variance-covariance

Estimation des effets fixes β cas déséquilibré

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Y$$

où $\hat{V} = \sum_{k=1} \hat{\sigma}_k^2 Z_k Z_k' + \hat{\sigma}^2 \mathbf{I}$

Il faut donc estimer les termes de la matrice d'effets aléatoires avant d'estimer les effets fixes

Estimation de V par maximum de vraisemblance

$Y \sim N(X\beta, V)$

Densité

$$pr(Y \in \Omega) = \int_{\Omega} f(y)dy$$

$$f(y) = \frac{1}{(2\pi)^{n/2} det(V)^{1/2}} \exp\left[-\frac{1}{2}(y - X\ \beta)' V^{-1}(y - X\ \beta)\right]$$

La log–vraisemblance du modèle mixte vaut

$$l(y, \beta, V) = cst - \frac{1}{2}\log det(V) - \frac{1}{2}(y - X\ \beta)' V^{-1}(y - X\ \beta)$$

Estimation par ML (suite)

$$l(y, \beta, V) = cst - \frac{1}{2} \log det(V) - \frac{1}{2}(y - X\ \beta)' V^{-1}(y - X\ \beta)$$

Il vient

$$\frac{\partial l}{\partial \beta} = X'\ V^{-1}\ y - X'\ V^{-1}\ X\ \beta$$

Soit p équations

D'autre part

$$\frac{\partial V}{\partial \sigma_k^2} = Z_k Z_k'$$

Estimation par ML (suite)

$$l(y, \beta, V) = cst - \frac{1}{2} \log det(V) - \frac{1}{2}(y - X\ \beta)' V^{-1} (y - X\ \beta)$$

$$\frac{\partial l}{\partial \sigma_k^2} = -\frac{1}{2} tr(V^{-1}\ Z_k\ Z_k') + \frac{1}{2}(y - X\ \beta)' V^{-1}\ Z_k\ Z_k'\ V^{-1}(y - X\ \beta)$$

soit $K+1$ équations

Au total $n + K + 1$ équations non linéaires à résoudre

On obtient

$$\begin{aligned} \hat{\beta} &= (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} Y \\ tr(V^{-1}\ Z_k\ Z_k') &= (y - X\ \beta)' V^{-1}\ Z_k\ Z_k'\ V^{-1}(y - X\ \beta) \end{aligned}$$

$\Rightarrow$ Ces estimateurs sont biaisés

Estimation par REML

(Restricted/Residual Maximum Likelihood)
La positivité des σ_k n'est pas prise en compte par l'estimateur du ML.
La méthode REML, sépare l'estimation des β et des σ.
Elle ne présente pas de biais systématique

Predictions individuelles

On considère le modèle suivant

$$Y_i = X_{ik}\beta + A_i + Z_{ij}$$

Notation : $Y_i = (Y_{i1} \dots Y_{iN_i})$ le vecteur des observations de l'individu i

- Que valent $[A_i|Y_i]$?
- Comme A_i et Y_k sont indépendants quand $i \neq k$, $[A_i|\mathbf{Y}] = [A_i|Y_i]$.
- $\text{var}(A_i) = \sigma^2$ et $\text{var}(Y_i) = V_0$
- $\text{cov}(A_i, Y_{ij}) = \text{cov}(A_i, A_i + Z_{ij}) = \sigma^2$
- $\text{E}(A_i) = 0$, $\text{E}(Y_i) = X_i\beta$.
- La variance conditionnelle de dépend pas des observations

Prediction d'une nouvelle réponse

- On mesure une nouvelle valeur X_{ik} pour l'individu i
- Peut-on prédire la valeur de Y_{ik} ?

Prediction d'une nouvelle réponse

- On mesure une nouvelle valeur X_{ik} pour l'individu i
- Peut-on prédire la valeur de Y_{ik} ?
- Il faut déterminer $[X_{ik}\beta + A_i + Z_{ij}|Y_i]$

Prediction d'une nouvelle réponse

- On mesure une nouvelle valeur X_{ik} pour l'individu i
- Peut-on prédire la valeur de Y_{ik} ?
- Il faut déterminer $[X_{ik}\beta + A_i + Z_{ij}|Y_i]$
- On se donne X_{ik}, et on suppose que β est connu

Prediction d'une nouvelle réponse

- On mesure une nouvelle valeur X_{ik} pour l'individu i
- Peut-on prédire la valeur de Y_{ik} ?
- Il faut déterminer $[X_{ik}\beta + A_i + Z_{ij}|Y_i]$
- On se donne X_{ik}, et on suppose que β est connu
- Alors $\mathrm{E}(Y_{ik}|Y_i) = X_{ik}\beta + \mathrm{E}(A_i|Y_i)$
- et

Prediction d'une nouvelle réponse

- On mesure une nouvelle valeur X_{ik} pour l'individu i
- Peut-on prédire la valeur de Y_{ik} ?
- Il faut déterminer $[X_{ik}\beta + A_i + Z_{ij}|Y_i]$
- On se donne X_{ik}, et on suppose que β est connu
- Alors $\mathrm{E}(Y_{ik}|Y_i) = X_{ik}\beta + \mathrm{E}(A_i|Y_i)$
- et $\mathrm{var}(Y_{ik}|Y_i) = \mathrm{var}(A_i|Y_i) + \tau^2$

Exercice 1

Vous observez des données longitudinales (y_{ij}, x_{ij}) où les y sont les réponses et les x_{ij} sont des variables explicatives (scalaires) pour la personnes i. ...,n
On suppose qu'on peut modéliser les y_{ij} par la régression

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$, $\varepsilon_{ij} \sim N(0, \tau^2)$

1. Quelle méthode proposez vous pour estimer β_c

2. Quelle méthode proposez vous pour estimer β_L?

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?
   Effectuer une régression linéaire simple de la réponse au temps initial sur le predicteur au temps initial (baseline) soit

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?
   Effectuer une régression linéaire simple de la réponse au temps initial sur le predicteur au temps initial (baseline) soit

$$y_{i1} = x_{i1}\beta_c + \varepsilon_{i1}$$

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?
   Effectuer une régression linéaire simple de la réponse au temps initial sur le predicteur au temps initial (baseline) soit

$$y_{i1} = x_{i1}\beta_c + \varepsilon_{i1}$$

2. Quelle méthode proposez vous pour estimer β_L ?

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?
   Effectuer une régression linéaire simple de la réponse au temps initial sur le predicteur au temps initial (baseline) soit

$$y_{i1} = x_{i1}\beta_c + \varepsilon_{i1}$$

2. Quelle méthode proposez vous pour estimer β_L ?
   Effectuer une régression linéaire simple de $y_{ij} - y_{i1}$ sur $x_{ij} - x_{i1}$ soit

Exercice 1

$$y_{ij} = x_{i1}\beta_c + (x_{ij} - x_{i1})\beta_L + \varepsilon_{ij}$$

pour $i = 1, \ldots, m$ et $j = 1, \ldots, n$

1. Quelle méthode proposez vous pour estimer β_c ?
   Effectuer une régression linéaire simple de la réponse au temps initial sur le predicteur au temps initial (baseline) soit

$$y_{i1} = x_{i1}\beta_c + \varepsilon_{i1}$$

2. Quelle méthode proposez vous pour estimer β_L ?
   Effectuer une régression linéaire simple de $y_{ij} - y_{i1}$ sur $x_{ij} - x_{i1}$ soit

$$y_{ij} - y_{i1} = \beta_L(x_{ij} - x_{i1}) + (\varepsilon_{ij} - \varepsilon_{i1})$$

Exercice2

Contexte : Taux de corticoïdes Y_{ij} chez le sportif i à la mesure j.
On a estimé les paramètres du modèle suivant $Y_{ij} = \mu + A_i + \varepsilon_{ij}$.
On supposera que : $A_i \sim N(0, \sigma_A^2)$, $\varepsilon_{ij} \sim N(0, \sigma^2)$ et A_i et ε_{ij} sont indépendants.

$\hat{\mu}$	$\hat{\sigma}_A$	$\hat{\sigma}$
99.91	10.50	0.98

et

id	(Intercept)
1	-9.29
2	-19.87
3	0.47
4	0.36

Table : Prédictions des effets aléatoires chez 4 individus

- Calculer un intervalle de prédiction (à 95 %) du taux de corticoïde chez l'individu 1 et 3

Exercice2

La loi de la réponse conditionnellement à l'effet aléatoire est
$[Y|A_i]$ $N(\mu + A_i, \sigma^2)$
on en déduit l'intervalle de prédiction $\mu + A_i \pm 1.96\ \sigma$ en substituant ces valeurs par leurs estimations on trouve les 2 intervalles