Modèle hétéroscédastique pour l'évaluation génétique des animaux d'élevage




Christèle Robert-Granié, UMR1388 GenPhySE, Toulouse
Journée Applibugs, 11 juin 2014

ALIMENTATION
AGRICULTURE
ENVIRONNEMENT

INRA

Evaluation Génétique des Animaux

Améliorer les populations d'animaux domestiques

↪ Sélectionner comme reproducteurs les meilleurs sujets selon un ou plusieurs caractères

↪ Evaluation de la valeur génétique transmissible

Modèle Polygénique infinitésimal

La valeur phénotypique (*P*) s'exprime comme : $P = G + E$

G : valeur génétique (effet moyen du génotype sur un caractère donné)
E : effet de l'environnement

En l'absence d'interaction génotype x environnement, on a :

$$Var(P) = Var(G) + Var(E) \qquad Cov(G,E) = 0$$

$$G = A + D + I$$

A : valeur génétique additive
D : valeur de dominance
I : valeur de l'épistasie

Seule A se transmet de parent aux descendants

Modèle Polygénique infinitésimal

Valeur génétique additive d'un descendant connaissant celles de ses parents

Aléa de méiose : d'un descendant à l'autre, ce ne sont pas exactement les mêmes gènes qui sont transmis par les parents E(w) = 0

$$A_{desc} = \frac{1}{2} A_{père} + \frac{1}{2} A_{mère} + w$$

Le père (la mère) transmet la moitié de ses gènes, la moitié de sa valeur génétique additive

Le sélectionneur s'intéresse à A

Un caractère quantitatif est gouverné par un grand nb de loci à effets individuels faibles et indépendants.
La somme des effet moyens des gènes suit une distribution normale

Objectif de l'amélioration génétique des animaux

Sélectionner au sein d'une population, c'est mettre à la reproduction les « meilleures » femelles avec les « meilleurs » mâles.

Les « meilleur(e)s » sont ceux (celles) qui portent les meilleurs gènes, c'est-à-dire ceux (celles) qui ont les plus fortes valeurs génétiques additives (*A*) pour les caractères que l'on souhaite améliorer.

Les « meilleur(e)s » sont ceux (celles) qui transmettront la plus forte supériorité.

→ **estimer / prédire** les valeurs génétiques additives (*A*) .

Informations utilisées pour l'évaluation Génétique

- Les données (y) : performances ou phénotypes
- Pedigree et généalogie des animaux (individu, père, mère)
- Les effets de milieu identifiés (ex: troupeau, année, saison, âge, …)

- Des effets de milieu non identifiés → résidus d'un modèle
- La valeur génétique additive → inconnue

Performances = valeur génétique additive + milieux identifiés + milieux non identifiés

$\mathbf{y}$ $\quad \mathbf{u} \sim N(0, \mathbf{A}\sigma_u^2)$ $\quad \boldsymbol{\beta}$ $\quad \mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2)$

A : matrice de parenté

Perf. du candidat

Perf. des ascendants, descendants, collatéraux

Modèle Statistique: Modèle linéaire mixte

$$y = X\beta + Zu + Wp + e$$

$u \sim N(0, \sigma_u^2 \mathbf{A})$ $\quad p \sim N(0, \sigma_p^2 I)$ $\quad e \sim N(0, \sigma_e^2 I)$

Valeur génétique additive

Var. génétique additive

Var. env. permanent

Var. environnementale

Modèle animal

Permet une estimation des effets fixes et prédiction des valeurs génétiques en une étape (BLUP)

Permet une prise en compte des effets de la sélection passée

Tient compte des accouplements raisonnés

Modèle Statistique: Modèle linéaire mixte

$$y = X\beta + Zu + Wp + e$$

$u \sim N(0, \sigma_u^2 \mathbf{A})$ $\quad p \sim N(0, \sigma_p^2 I)$ $\quad e \sim N(0, \sigma_e^2 I)$

Valeur génétique additive

Var. génétique additive

Var. env. permanent

Var. environnementale

Estimation des effets fixes et prédiction des valeurs génétiques → BLUP

Estimation des composantes de la variance → REML

Henderson (1963), Patterson et Thomson (1971)

Modèle linéaire mixte et variances hétérogènes

$$y = X\beta + Zu + Wp + e$$

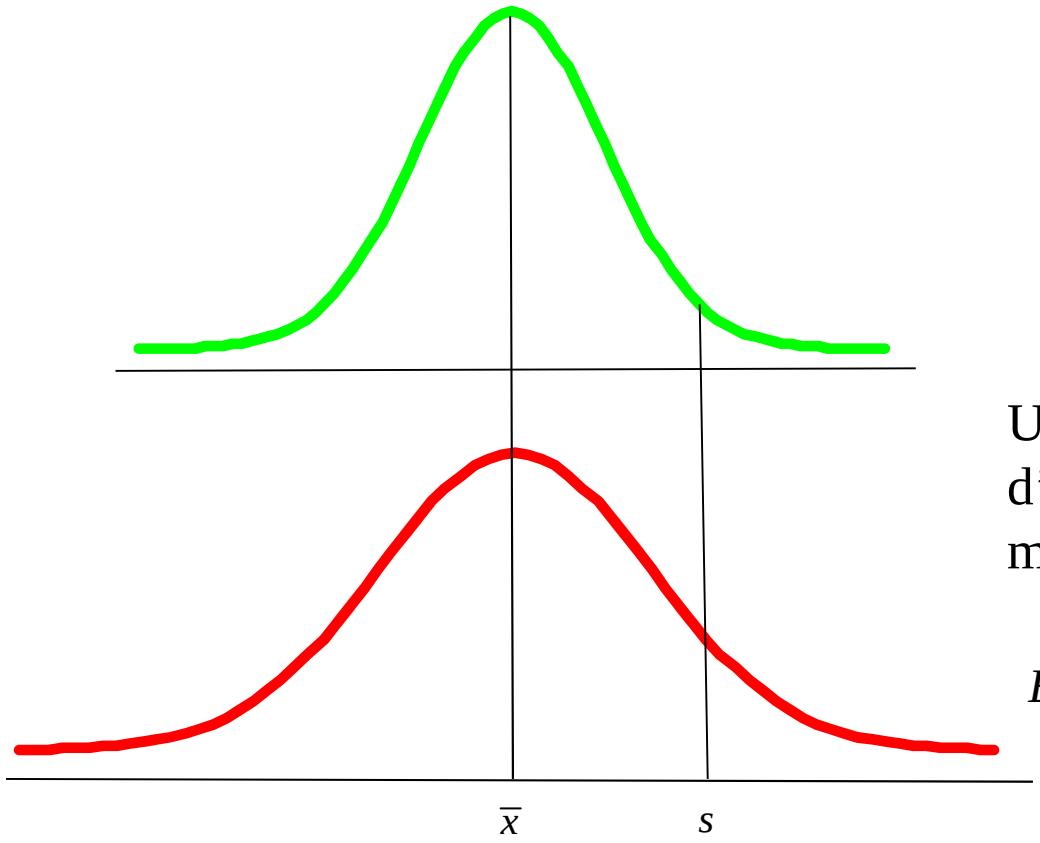
Mise en évidence d'hétérogénéité des variances résiduelles et génétiques :

- Production laitière entre troupeaux chez les bovins laitiers
- Caractères de morphologie, Performance de croissance, etc…

Plusieurs raisons :

- Conditions d'élevage hétérogènes (alimentation, logement)
- Stratégies différentes de sélection selon les élevages
- Traitements préférentiels de certains groupes d'animaux
- Existence d'interaction génotype x milieu

Conséquences de la non prise en compte des variances hétérogènes sur la sélection



Une plus grande proportion d'animaux est sélectionnée dans les milieux les plus variables.

Everett et al (1982), Hill (1984), etc …

Modélisation de l'hétéroscédasticité
Approche monofactorielle

Décrite par une matrice de variance-covariance résiduelle diagonale par blocs, chaque bloc correspond à un niveau de facteur de variation.

$$Var(e) = \begin{pmatrix} \sigma^2_{e_1} & & & \\ & \ldots & & \\ & & \sigma^2_{e_i} & \\ & & & \ldots \\ & & & \sigma^2_{e_n} \end{pmatrix}$$

n cellules
Beaucoup trop de paramètres ...

Modélisation de l'hétéroscédasticité : modèle dit « structural »

$$\ln \sigma^2_{e_i} = h'_{e_i} \delta_e$$

$$\ln \sigma^2_{e_i} = h'_{e_i} \delta_e + q'_{e_i} v_e \quad avec \quad v_e \sim N(0, \sigma^2_{v_e} \Sigma)$$

Effets fixes

(facteurs et/ou covariables)

Effet aléatoire

Foulley et al (1990)

Exemple Bovins laitiers : Facteurs d'hétérogénéité

- **l'année de production** : le niveau de production a considérablement augmenté dans le temps, une augmentation de la variabilité a été constaté au cours des 30 dernières années, l'héritabilité restant quant à elle à peu près constante.

- **la région de production** : les régions extensives, herbagères ou montagneuses bénéficient d'une variabilité phénotypique moindre que les régions plus intensives.

- **la conduite du troupeau** : source d'hétérogénéité importante et dont l'origine est très variable : le niveau de production moyen, le pourcentage de lactations courtes terminées, l'état sanitaire du troupeau, l'existence de lots conduits de façon différente.

Modèle sur les variances

Le principal facteur de variation de la variance résiduelle est la combinaison **troupeau année**. Mais le nombre de données par troupeau année est faible et ne permet donc pas une bonne estimation de la variance résiduelle.

Nous avons choisi d'utiliser deux sources d'information :
- la variabilité résiduelle estimée dans **la région et l'année** considérées (servira de moyenne de référence)
- la variabilité résiduelle estimée dans le même **troupeau** les années précédentes et suivantes (permettra d'améliorer la précision de l'estimation de la variabilité résiduelle)

- <u>Le modèle retenu sur la variance résiduelle</u>:

$$\ln \sigma^2_{e_i} = h'_{e_i}\delta_e + q'_{e_i}v_e \quad avec \quad v_e \sim N(0, \Sigma\sigma^2_{v_e})$$

Effet fixe région-année

Effet aléatoire troupeau-année

Modèle sur les variances

Variance résiduelle : les effets troupeau-année supposés corrélés intra-troupeau suivant un processus autorégressif.

Ex: la variance des effets troupeau-année, pour un troupeau donné suivi sur une période de 4 années :

$$\Sigma\sigma^2_{v_e} = \sigma^2_{v_e}\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

Variances génétiques et d'environnement permanent :

Si héritabilité et répétabilité constantes et connues

$$\sigma^2_{u_i} = t_1^2\sigma^2_{e_i} \qquad t_1^2 = \frac{h^2}{(1-r)}$$

$$\sigma^2_{p_i} = t_2^2\sigma^2_{e_i} \qquad t_2^2 = \frac{r-h^2}{(1-r)}$$

Modèle d'évaluation chez les bovins laitiers

$$y_i = X_i\beta + t_1\sigma_{e_i} Z_i u^* + t_2\sigma_{e_i} W_i p^* + e_i$$

$$u^* \sim N(0, A) \qquad p^* \sim N(0, I_q) \qquad e_i \sim N(0, I_{n_i}\sigma^2_{e_i})$$

Effet fixe région-année — Effet aléatoire troupeau-année

$$\ln\ \sigma^2_{e_i} = h'_{e_i}\delta_e\ +\ q'_{e_i}v_e\ \ avec\ \ v_e \sim N(0, \Sigma\sigma^2_{v_e})$$

AR(1)

$$\sigma^2_{u_i} = t_1^2\sigma^2_{e_i}$$
$$\sigma^2_{p_i} = t_2^2\sigma^2_{e_i}$$

Héritabilité et Répétabilité constantes

$\gamma = (\delta_e, t_1, t_2, \sigma^2_{v_e}, \rho)$: paramètres à estimer → EM-ML

Robert-Granié et al (1999)

Estimations des effets fixes et aléatoires
Cas hétérogène

$$\ln \ \sigma^2_{e_i} = h^{'}_{e_i}\delta_e$$

$$\sigma^2_{u_i} = t_1^2\sigma^2_{e_i} \text{ et } \sigma^2_{p_i} = t_2^2\sigma^2_{e_i}$$

$$u^* \sim N(0, A)$$
$$p^* \sim N(0, I_q)$$
$$e_i \sim N(0, I_{n_i}\sigma^2_{e_i})$$

$$y_i = X_i\beta + t_1\sigma_{e_i}Z_iu^* + t_2\sigma_{e_i}W_ip^* + e_i$$

$$\begin{bmatrix} \sum_{i=1}^{p} X_i^{'}X_i\sigma_{e_i}^{-2} & \sum_{i=1}^{p} X_i^{'}Z_it_1\sigma_{e_i}^{-1} & \sum_{i=1}^{p} X_i^{'}W_it_2\sigma_{e_i}^{-1} \\ \sum_{i=1}^{p} Z_i^{'}X_it_1\sigma_{e_i}^{-1} & \sum_{i=1}^{p} Z_i^{'}Z_it_1^2 + A^{-1} & \sum_{i=1}^{p} Z_i^{'}W_it_1t_2 \\ \sum_{i=1}^{p} W_i^{'}X_it_2\sigma_{e_i}^{-1} & \sum_{i=1}^{p} W_i^{'}Z_it_1t_2 & \sum_{i=1}^{p} W_i^{'}W_it_2^2 + I_q \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u}^* \\ \hat{p}^* \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{p} X_i^{'}y_i\sigma_{e_i}^{-2} \\ \sum_{i=1}^{p} Z_i^{'}y_it_1\sigma_{e_i}^{-1} \\ \sum_{i=1}^{p} W_i^{'}y_it_2\sigma_{e_i}^{-1} \end{bmatrix}$$

Conséquences sur les index laitiers

Mâle : Forte corrélation entre les index homogène vs hétérogène (0.99)

Peu de modification dans le classement des meilleurs taureaux

les bons index ↘ dans les troupeaux à forte variabilité et ↗ dans les troupeaux à faible variabilité

Femelle: Les index sont beaucoup plus affectés, d 'autant plus que la répétabilité entre années du troupeau est élevée ; les apparentées directes sont souvent dans le même troupeau

Corrélation entre les index homogène vs hétérogène (0.6 à 0.7)

Reclassement important dans le top

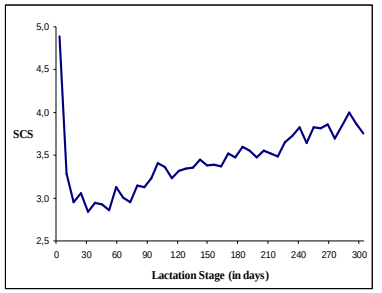
Le mode de recrutement de l'élite change avec le modèle hétérogène

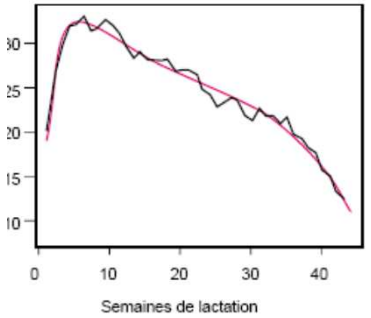
En modèle homogène, seuls les troupeaux à forte variabilité contribuaient à l'élite

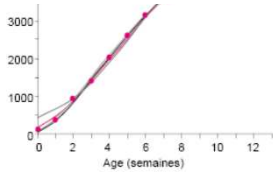
Modélisation des données longitudinales

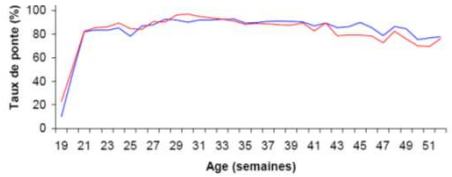
Comment affiner la modélisation de phénotypes répétés au cours du temps ?

- Evolution des comptages de cellules somatiques dans le lait chez les bovins
- Cinétiques de débits de traite en ovins
- Carrière de production de semence chez les ovins
- Courbes de croissance
- Taux de ponte
- etc …





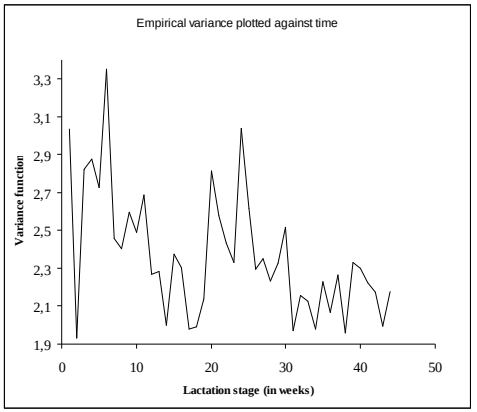




Modélisation des données longitudinales

Comment affiner la modélisation de phénotypes répétés au cours du temps ?

Evolution de la variance



Comment modéliser les corrélations entre mesures répétées et des variabilités hétérogènes ?

Comment choisir le meilleur modèle ?

Modélisation des données longitudinales

Dans le cadre du modèle linéaire mixte

- Polynôme fractionnaire
- Modèle structural sur les variances
- Processus autorégressif

Estimation des paramètres
EM-REML

- Test robuste des effets fixes

Liang et Zeger (1986)

Estimateurs robustes

-Test contraint sur les variances

Self et Liang (1987)

Test du rapport de vraisemblance

Polynômes fractionnaires

An extension of conventional polynomials with real power terms (Royston and Altman, 1994) :

$$\phi_m(t;\xi,p)=\sum_{j=0}^{m}\xi_j H_j(t)$$

$$H_j(t)=\begin{cases} t^{(p_j)} & \text{if } p_j \neq p_{j-1}, \\ H_{j-1}(t)\ln(t) & \text{if } p_j = p_{j-1}. \end{cases}$$

$$t^{(p_j)}=\begin{cases} t^{p_j} & \text{if } p_j \neq 0 \\ \ln(t) & \text{if } p_j = 0 \end{cases}$$

$$H_0(t)=1,\ p_0=0$$

t : positive real covariate

m : positive integer (degree)

p={p_j} : vector of ordered real powers (j=1,…,m)

ξ={ξ_j} : vector of real coefficients

Un exemple de polynôme fractionnaire

A fractional polynomial of degree m=3 with powers p={0,0,0.5} :

$$\phi_3\left(t;\boldsymbol{\xi},\mathbf{p}=\{0,0,1/2\}\right)=\xi_0+\xi_1\ln(t)+\xi_2[\ln(t)]^2+\xi_3 t^{0.5}$$

We need to determine :

- the best value of m (degree)
- the best value of the power vector p

Polynôme fractionnaire appliqué aux données de SCS

Scores de Cellule Somatique

Polynôme classique

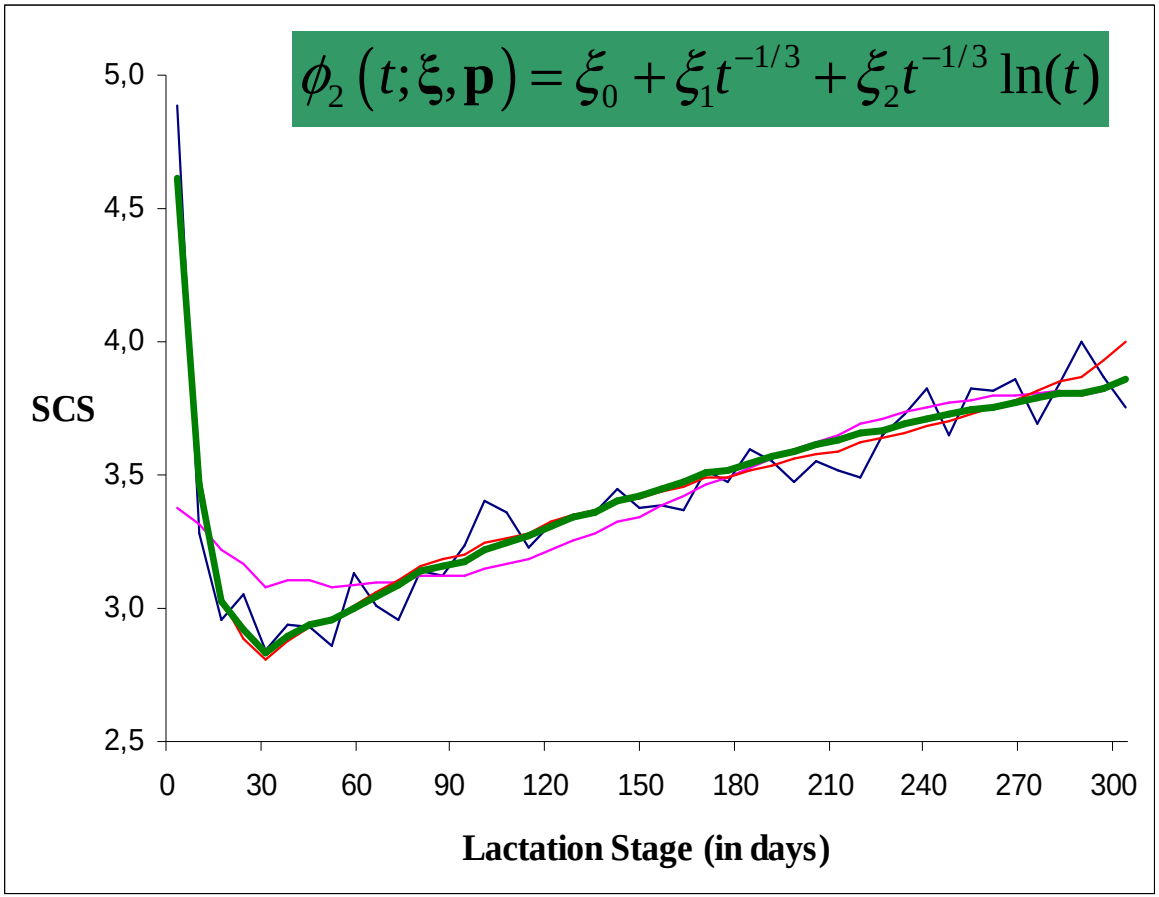
m=3 and p=(1,2,3)

Fonction de Ali & Schaeffer

m=4 and p=(0,0,1,2)

Polynôme fractionnaire

m=2 and p=(-1/3,-1/3)



Robert-Granié et al (2004)

Modélisation des données longitudinales

Dans le cadre du modèle non linéaire mixte homoscédastique

$$y_{ij} = f(z_{ij}, \beta, \phi_i) + \sigma\varepsilon_{ij}^* \text{ avec } \phi_i \sim N(A_i\mu, \Gamma) \text{ et } \varepsilon_{ij}^* \sim N(0,1)$$

- Méthode exacte : SAEM-MCMC (Kuhn et Lavielle, 2004)

- Mise en œuvre de l'algorithme et implémentation de critères

  Paramètres dans l'algorithme de Métropolis-Hastings
  Critère de lissage
  Critère d'arrêt

Thèse Mylène DUVAL

Modèle non linéaire mixte hétéroscédastique

$$y_{ij} = f(z_{ij}, \beta, \phi_i) + g(w_{ij}, \delta, \psi_{ij})\varepsilon_{ij}^*$$

Effets fixes

Effets aléatoires

$$\phi_i \sim N(A_i\mu, \Gamma) \text{ et } \varepsilon_{ij}^* \sim N(0,1)$$

Modèle structural

$$\log(g(w_{ij}, \delta, \psi_{ij})^2) = w_{ij}^{'}\delta + q_{ij}^{'}\nu \text{ avec } \nu \sim N(0, G_\nu)$$

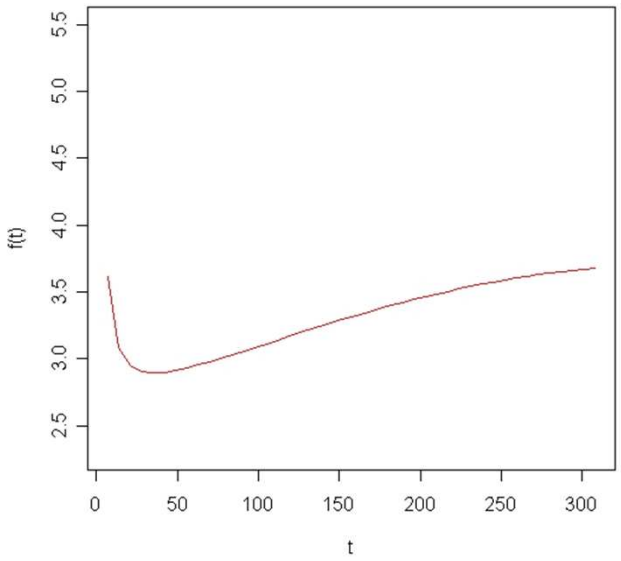
Liaison moyenne-variance

$$g(z_{ij}, \beta, \phi_i)^2 = rf(z_{ij}, \beta, \phi_i)^p$$

Modéliser les profils moyens et individuels des scores de cellules somatiques

Fonction moyenne

- Morant et Gnanasakthy (1989)



$$f(t_{ij},\phi_1,\phi_2,\phi_3,\phi_4) = \phi_1 \exp\left(\phi_2 t_{ij} + \phi_3 {t_{ij}}^2 + \phi_4 / t_{ij}\right)$$

ϕ_1 associé à la valeur initiale

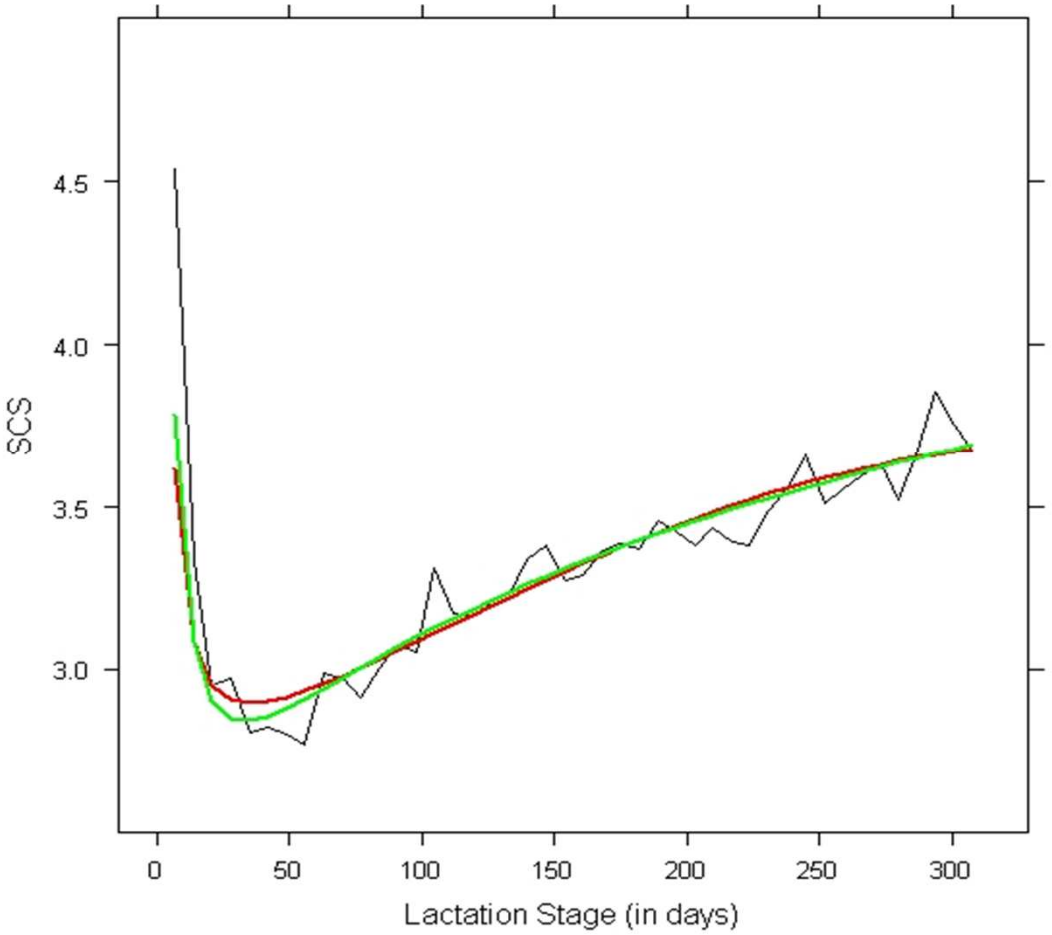
ϕ_2 associé à la pente de la courbe en fin de lactation

ϕ_3 associé à la variation de la pente

ϕ_4 associé à la pente de la courbe en début de lactation

Modéliser les profils moyens et individuels des scores de cellules somatiques

Fonction moyenne



Profils moyens

Modèle non linéaire de Morant et Gnanasakthy (1989)

BIC=17327

Modèle linéaire (Robert-Granié et al., 2004)

BIC=17668

Structures de variance résiduelle

- La relation moyenne-variance

$$\sigma_{ij}^{\ 2} = \delta_1 f_{ij}^{\ \delta_2}$$

- Un modèle linéaire fixe sur la log-variance

$$\log(\sigma_{ij}^{\ 2}) = \delta_0 + \delta_1 t_{ij} + \delta_2 t_{ij}^{\ 2}$$

- Un modèle linéaire mixte sur la log-variance

$$\log(\sigma_{ij}^{\ 2}) = \delta_{0i} + \delta_{1i} t_{ij} + \delta_{2i} t_{ij}^{\ 2}$$

$$\text{où } (\delta_{0i}, \delta_{1i}, \delta_{2i}) \sim N\left((\delta_0, \delta_1, \delta_2), \Delta\right)$$

Résultats

	Critère BIC	
	Modèle linéaire	Modèle non linéaire
$\sigma_{ij}^2 = \sigma^2 \quad \forall i,j$	17572 (-0)	17208 (-0)
$\sigma_{ij}^2 = \delta_1 f_{ij}^{\delta_2}$	17457 (-115)	17107 (-101)
$\log(\sigma_{ij}^2) = \delta_0 + \delta_1 t_{ij} + \delta_2 t_{ij}^2$	17388 (-184)	16897 (-311)
$\log(\sigma_{ij}^2) = \delta_{0i} + \delta_{1i} t_{ij} + \delta_{2i} t_{ij}^2$	15292 (-2280)	15086 (-2122)

Conclusion : meilleur modèle

$$\log(\sigma_{ij}^{\,2}) = \delta_{0i} + \delta_{1i} t_{ij} + \delta_{2i} t_{ij}^{\,2}$$

$(\delta_{0i}, \delta_{1i}, \delta_{2i})$ aléatoires et corrélés

$$y_{ij} = \phi_1 \exp\left(\phi_2 t_{ij} + \phi_3 t_{ij}^{\,2} + \phi_4 / t_{ij}\right) + \sigma_{ij}\varepsilon_{ij} \; *$$

ϕ_1, ϕ_2, ϕ_3, et ϕ_4 sont aléatoires et corrélés

ϕ_2, ϕ_3 dépendent de la saison de vêlage de la vache

Modèle de Morant et Gnanasakthy (1989)

Question appliquée liée à la génétique animale

Identifier des facteurs de variations
Etude de la variabilité génétique des caractères
Evaluation génétique des reproducteurs

Modélisation la plus appropriée

Etude statistique
Modélisation
Méthode d'estimation
Tests

Mise en oeuvre du ou des modèles en routine

Christe
Journé

Course

« Basic Statistical Methods for Longitudinal Data Analysis”
(BSMLDA2002)

Sunday, August 18, 2002, Montpellier, France
(8:45 AM to 4:30 PM)

Instructors :

Jean-Louis Foulley and Christèle Robert-Granié
INRA, Animal Genetics, France

Les enseignements à l’ENSAI Rennes

Séminaire METHODES MCMC ET APPROCHES CONNEXES La Londe Les Maures en 2005

Programmer en APL, toute une histoire !!!

```
      ∇DET[⎕]∇
     ∇ Z←DET A;B;P;I
[1]    I←⎕IO
[2]    Z←1
[3]   L:P←(|A[;I])⍳⌈/|A[;I]
[4]    →(P=I)/LL
[5]    A[I,P;]←A[P,I;]
[6]    Z←-Z
[7]   LL:Z←Z×B←A[I;I]
[8]    →(0 1 ∨.=Z,1↑⍴A)/0
[9]    A←1 1 ↓A-(A[;I]⌽B)∘.×A[I;]
[10]   →L
[11]  ⍝EVALUATES A DETERMINANT
     ∇
```