Introduction sur la régression dite "aléatoire" et son utilisation dans l'amélioration génétique des animaux domestiques

J.L. Gourdine

INRA, Unité de Recherches Zootechniques, Domaine Duclos, 97170 Petit-Bourg, Guadeloupe.

Séminaire AOC du 08/12/2005, Université Antilles-Guyane , Fouillole, Guadeloupe

Amélioration génétique ???

L'ensemble des techniques utilisées pour modifier le potentiel héréditaire des animaux.

Objectif : fournir à l'éleveur un animal qui réponde à ses souhaits en recherchant une adéquation optimale entre le matériel génétique et les conditions du milieu.

Amélioration génétique ???

Avant :

Méthodes empiriques

Maintenant :

Programmation

Avant : « Puisque leurs mères, qui sont demi-sœurs, sont bonne productrices, je garderais certains de ces veaux pour en faire des reproducteurs.»

Maintenant : « A partir de l'évaluation génétique basée sur la résolution des équations du modèle mixte, les meilleures valeurs génétiques sont estimées pour les veaux 5 et 18. »

Régression aléatoire ???

Random Regression Models (RRM)
(Henderson, 1982 ; Laird et Ware, 1982)

↗ **Etude et utilisation des RRM au cours de ces 8 dernières années :**

7th World Congress on Genetics Applied to Livestock Production, 2002, Montpellier :

24 papiers sur RRM

Contexte économique

↗ Pop. humaine ➡ ↗ Demande alimentaire (animale)

⬇

↗ Prod. Animale en tenant compte des contraintes (perception sociale, pollution, santé humaine)

Contexte économique

Programmes avec minimisant les coûts et optimisant un système de prod. animale écologique et durable.

Collecte de l'information :

données de type longitudinal

(ex. poids vif; prod. laitière, etc …)

Contexte économique

Outils statistiques particuliers pour traiter ce type de données et estimer la valeur génétique

(Valeur génétique : quantification des effets moyens des gènes intervenant pour le caractère étudié et transmissibles à la descendance.)

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

Y_{ijk}: k^{ème} performance de l'individu i réalisée dans les conditions du milieu j

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

m_j: somme des effets du milieu identifiés à laquelle est soumise la performance (ex. moyenne générale, saison) : EFFETS FIXES

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

a_i: valeur génétique de l'animal i (écart à la moyenne)

Déterminisme polygénique : nombreux gènes à effets faibles a_{il} sur la performance de l'animal i

$$a_i = \sum_l a_{il}$$

$a_i \sim N(0, \sigma^2_a)$ (loi des grands nombres)

Ex. : Pour deux individus i et m demi-frères, on a : $Cov(a_i a_m) = \sigma^2_a/4$

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

a_i: valeur génétique de l'animal i (écart à la moyenne)

L'ensemble des effets génétiques $\{a_i\}_i \sim N_N (0 , A\sigma^2_a)$

A : matrice de parenté

⬇

Information avant observations = lien de parenté :

a_i est un effet aléatoire du modèle

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

e_{ijk}: résiduelle. EFFETS ALEATOIRES

$e_{ijk} \sim N(0, \sigma^2_e)$

Modèle génétique de base

$$Y_{ijk} = m_j + a_i + e_{ijk}$$

Y_{ijk}: k[ème] performance

m_j: effets du milieu EFFETS FIXES

a_i: valeur génétique. EFFETS ALEATOIRES

e_{ijk}: résiduelle. EFFETS ALEATOIRES

MODELE MIXTE

Structure de base d'un RRM

$$Y_{it} = F + g(t) + r(a,x,k_A)_i + r(p_e,x,k_R) + e_{it}$$

Y_{it}: observation sur animal i à l'instant t

F : ensemble des effets fixes indépendants de t

g(t) : fonction ajustant la trajectoire phénotypique de l'individu « moyen » de la population

Structure de base d'un RRM

$$Y_{it} = F + g(t) + r(a,x,k_A)_i + r(p_e,x,k_R)_i + e_{it}$$

$r(a,x,k_A)$: fonction de régression aléatoire.

- a : effets génétiques additifs
- x : vecteur des covariables de temps
- k_A : ordre de la fonction de régression

$r(a,x,k_A)_i = a_{i1} \times x_{t1} + a_{i2} \times x_{t1} + \ldots + a_{ikA} \times x_{tkA}$ avec :

- a_{ij} : coefficients de régression aléatoire pour les effets génétiques additifs

Structure de base d'un RRM

$$Y_{it} = F + g(t) + r(a,x,k_A)_i + r(p_e,x,k_R) + e_{it}$$

$r(p_e,x,k_R)$: fonction de régression aléatoire.

- p_e : effets d'environnement permanent
- x : vecteur des covariables de temps
- k_R : ordre de la fonction de régression

Structure de base d'un RRM

$$Y_{it} = F + g(t) + r(a,x,k_A)_i + r(p_e,x,k_R) + e_{it}$$

Objectif des fonctions de régression aléatoire :

- **modéliser les écarts individuels (dû aux effets aléatoires) à la trajectoire phénotypique moyenne**
- **fournir une description du potentiel génétique de l'animal sur la période considérée**

Structure de base d'un RRM

Modélisation des écarts individuels :

utilisation des polynômes orthogonaux de Legendre

(standardisés à l'intervalle [- 1 ; + 1]) comme covariables de temps (Kirkpatrick et coll., 1990)

car :

- **faciles à calculer**
- **Réduction des corrélations des paramètres à estimer => facilite la convergence des procédures d'estimation**

Structure de base d'un RRM

Définition d'une famille de polynômes orthogonaux :

Soit $\Im$ l'ensemble des fonctions continues sur $J = [a;b]$

La famille $\{P_n\}_{n \in N}$ de polynômes orthogonaux vérifient :

$$\begin{cases} \forall k \neq l, \int_a^b P_k(t)P_l(t)dt = 0 \\ \forall k \in N, \int_a^b P_k^2(t)dt = 1 \end{cases}$$

On ramène à [-1;1] par changm^t de variable affine

$$w = \frac{2(t - t_{min})}{(t_{max} - t_{min})} - 1$$

Structure de base d'un RRM

Polynômes de Legendre :

Formule générale :

$\forall k \in N, \forall t \in \Re,$

$$\varphi_k(t) = \left(\frac{2k+1}{2}\right)^{\frac{1}{2}} \frac{1}{2^k k!} \frac{d^k}{dt^k}\left[\left(t^2 - 1\right)^k\right]$$

$\forall k \in N, \forall t \in \Re,$

$$\varphi_0(t) = \sqrt{\frac{1}{2}} ; \varphi_1(t) = \sqrt{\frac{3}{2}} \times t ; \varphi_2(t) = \frac{1}{2}\sqrt{\frac{5}{2}} \times \left(3t^2 - 1\right)$$

Structure de base d'un RRM

Réécriture du modèle RRM:

$$Y_{it} = F + g(t) + r(a,x,k_A)_i + r(p_e,x,k_R) + e_{it}$$

$$y_{it} = F + g(t) + \sum_{m=0}^{k_A-1} \alpha_{im}\varphi_m\left(t_{ij}^*\right) + \sum_{m=0}^{k_R-1} \gamma_{im}\varphi_m\left(t_{ij}^*\right) + \varepsilon_{ij}$$

α_{im} : coefficients de régression aléatoire pour les effets génétiques additifs

γ_{im} : coefficients de régression aléatoire pour les effets d'environnement permanent

$\varphi_m(t^*_{ij})$: $m^{\text{ème}}$ polynôme de Legendre pour le temps t^*_{ij} standardisé ($t^*_{ij} \in [-1 ; +1]$)

Structure de base d'un RRM

Notation Matricielle RRM:

$$Y = Xb + Z_1 a + Z_2 p + e$$

Y : vecteur de N observations

b: vecteur des effets fixes; X matrice d'incidence

a : vecteur des k_A coefficients de régression aléatoire pour les effets génétiques additifs; Z_1 : matrice d'incidence

p : vecteur des k_R coefficients de régression aléatoire pour les effets d'environnement permanent; Z_2 : matrice d'incidence

Structure de base d'un RRM

$$\mathrm{Var}\begin{pmatrix} a \\ p \\ e \end{pmatrix} = \begin{pmatrix} A \otimes G & 0 & 0 \\ 0 & I \otimes P & 0 \\ 0 & 0 & R \end{pmatrix}$$

A : matrice de parenté

G : matrice de variance-covariance pour les effets génétiques additifs

P : matrice de variance-covariance pour les effets d'environnement permanent

I : matrice identité

R : matrice de variance- covariance résiduelle

$\otimes$: produit de Kronecker ou produit tensoriel

Structure de base d'un RRM

Rappel : ⊗ produit de Kronecker ou produit tensoriel

Soient deux matrices

$$A_{m \times n} = (a_{ij}) \text{ et } B_{p \times r} = (b_{ij}),$$

on a :

$$A \otimes B_{mp \times nr} = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nn}B \end{pmatrix}$$

Structure de base d'un RRM

Estimation des paramètres b, a et p :

équations du modèle mixte

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z_1 & X'R^{-1}Z_2 \\ Z_1'R^{-1}X & Z_1'R^{-1}Z_1 + A^{-1} \otimes G^{-1} & Z_1'R^{-1}Z_2 \\ Z_2'R^{-1}X & Z_2'R^{-1}Z_1 & Z_2'R^{-1}Z_2 + I \otimes P^{-1} \end{pmatrix} \times \begin{pmatrix} \hat{b} \\ \hat{a} \\ \hat{p} \end{pmatrix}$$

$$= \begin{pmatrix} X'R^{-1}y \\ Z_1'R^{-1}y \\ Z_2'R^{-1}y \end{pmatrix}$$

Structure de base d'un RRM

Propriétés statistiques :

- **si G, P et R sont connues alors :**

  $\hat{\mathrm{b}}$: meilleur estimateur sans biais de b (BLUE)

  $\hat{\mathrm{a}}$ et $\hat{\mathrm{p}}$: meilleurs prédictions sans biais de a et p (BLUP)

Pour estimer G, P et R (méthode REML) :

- **hypothèse de normalité des variables aléatoires**
- **EBLUE (empirical best linear unbiased estimator)**
- **EBLUP (empirical best linear unbiased predictor)**

Exemple :
Étude de la variabilité génétique de la croissance de la vache de race Créole
(Gourdine, Menendez-Buxadera et Naves, 2002)

Structure des données :

- 227 vaches dont 117 mères
- critère de sélection : PV18
- mesures de la naissance jusqu'à 10 ans

Question : la sélection peut-elle se faire plus tôt ?

Aide à la réponse : utilisation des RRM et résolution des équations du modèle mixte

Soit y_{ij} le $j^{ème}$ poids vif mesuré sur la vache i à l'âge t_{ij}

$$Y_{ij} = \text{Effets fixes } (F_i + \Sigma_m (\beta_m t_{ij}^m)_{m=0,\ldots 3})$$

$$+ \text{ Effets aléatoires } (g(t_{ij}) + r(t_{ij}) + e_{ij})$$

F : effet du numéro de portée de la mère sur la croissance de l'animal

$\Sigma(\beta_m t_{ij}^m)$: relation âge-poids de la vache « moyenne »

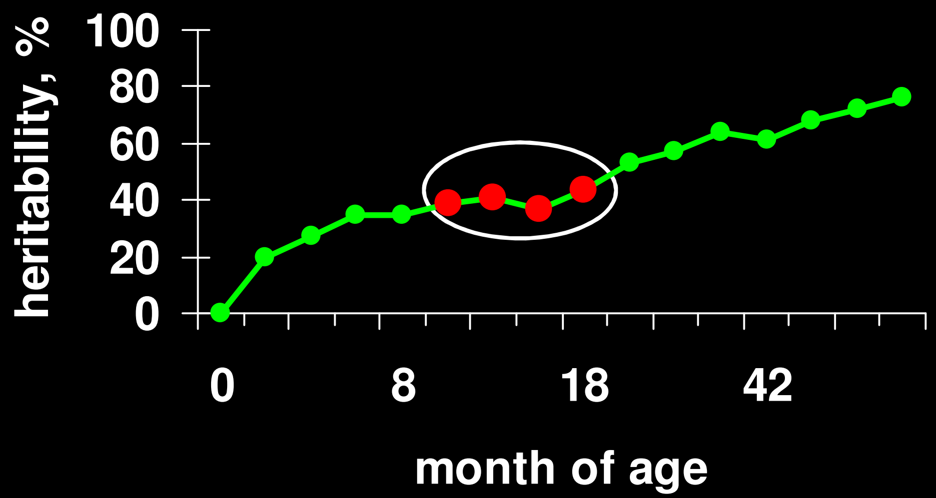
g : fonction de régression aléatoire pour les effets génétiques additifs
r : fonction de régression aléatoire pour les effets d'environnement permanent
e : bruit de fond

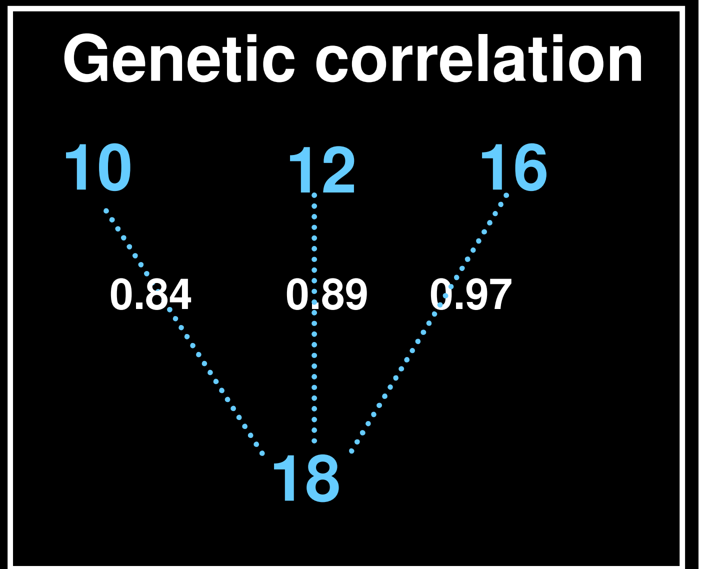
Est-il possible de sélectionner plus tôt?

Estimation de la partie aléatoire :

écarts individuels à la trajectoire moyenne

$$\text{Héritabilité} = \frac{\text{Variation génétique additive}}{\text{Variation phénotypique totale}}$$





Conclusion

RRM : méthode standard en analyse génétique quantitative (données longitudinales).

Mais : ce n'est pas une fin :

- **Système d'évaluation en constante évolution**
- **Recherches dynamiques (fonction Splines, Bayesien estimation, RRM multivariate)**

Quelques références bibliographiques

• Henderson C.R. (1982): *Analysis of covariance in the mixed model : higher level, nonhomogeneous and random regressions.* Biometrics 38, 623-640

• Laird N.M and Ware J.H (1982) : *Random effects models for longitudinal data.* Biometrics 38, 963-974

• Gilmour AR, Thompson R, Cullis BR (1995) : *Average Information REML, an efficient algorithm for variance parameters estimation in linear mixed models*. Biometrics 51, 1440-1450

• Menéndez-Buxadera A. (2004) : *Basic principles and importance of the utilization of random regression models in animal breeding programs.* CRAG : 46 pp.

• Fischer T.M, Gilmour A.R, van der Werf J.H.J. *Computing approximate standard errors for genetic parameters derived from random regression models fitted by average information REML.* Genetics Selection Evolution. 36, 363-369.

• Meyer K. (2005) : *Advances in methodology for random regression analyses.* Australian Journal of Experimental Agriculture 45, 847–858