Le modèle de base de la sélection: justification et limites

Publications G Lefort-INRA/DGA

Lauvergne J.-J., Lefort G., 1973. Nouvelle méthode pour analyser le comportement et la fréquence des gènes récessifs à effets visibles dans les populations bovines. *CR Acad. Sci. Paris*, Série D, 277, 2793-2796

Matheron G., Poujardieu B., Lefort G., 1974. Un modèle d'estimation des paramètres génétiques en présence d'effets directs et maternels chez le lapin. *1er congrès mondial de génétique appliquée à l'élevage*, Madrid, 7-11 oct 1974, vol III (Symposia), 447-454

Lefort G., Ollivier L., Sellier P.,1975. Analyse du comportement et de la fréquence des genes à effets visibles dans des fratries de germains et de demi-germains. *Ann Génét Sél Anim.*, 7, 365-377

Foulley J.-L., Lefort G., 1978. Méthodes d'estimation des effets directs et maternels en sélection animale. *Ann. Génét. Sél. Anim.*, 19, 475-496.

Références

-Présentation au séminaire de Méribel du Département d'Amélioration des Plantes (22-27 mars 1979)

-Diffusé dans le Tocsin du Radiateur, 1979, vol 1 (INRA-Publication, M Rives, Editeur)

-Publié dans « Biométrie et Génétique » 1980,
JM Legay et al, Ed, 1-14

Plan

Introduction

Fixe ou aléatoire ? A cette question relative à la nature d'un paramètre, la valeur d'un géniteur par exemple, le généticien répond souvent de manière intuitive sans que les raisons logiques de son choix apparaissent clairement ; parfois même, il considère un paramètre comme aléatoire dans un premier calcul et comme fixe dans un second et ne paraît pas gêné par cette incohérence. Pourtant l'interprétation des résultats et en particulier le classement des reproducteurs dépendent du choix qui est fait.

Introduction

Je voudrais ici apporter des éléments de réponse et montrer comment la conception bayesienne de la statistique permet de justifier dans de nombreux cas la démarche habituelle des généticiens et peut conduire à l'améliorer. A titre d'exemple, je considérerai le cas d'une sélection sur descendance et plus précisément le testage des taureaux Charolais (cet exemple a le tort d'être relatif à des animaux mais l'avantage de rester simple sans approximations abusives). Un seul caractère sera pris en compte, le gain moyen quotidien des fils pendant une période

1. Pratique courante/1.1 Le Modèle et l'indexation…

$$(1)\ Y_{ij} = \mu + A_i + B_{ij};\ 1 \le i \le I\ \ ;\ \ 1 \le j \le J_i$$

où les aléatoires A_i et B_{ij} sont indépendantes centrées de variance γ_A pour les A_i et γ pour les B_{ij}.

1. Pratique courante/1.1 Le Modèle et l'indexation…

$$Y_{ij} = \mu + A_i + B_{ij} \qquad 1 \le i \le I \quad 1 \le j \le J_i$$

Si μ, γ_A, γ étaient connus, l'estimateur MSE de A_i défini par la régression de A_i sur l'ensemble des Y_{ij}

$$\hat{A}_i = \frac{\gamma_A}{\gamma_A + (\gamma / J_i)} (y_{i.} - \mu)$$

1. Pratique courante/1.1 Le Modèle/Remarques

$$\hat{\mu} = \bar{y}_{..} \qquad I(J-1)\hat{\gamma} = \sum_{ij} (\bar{y}_{ij} - \bar{y}_{i.})^2$$

$$(I - 1)(\hat{\gamma}_A + \hat{\gamma}/J) = \sum_{i} (\bar{y}_{i.} - \bar{y}_{..})^2$$

1. Pratique courante/1.1 Le Modèle/Remarques

Tableau: Estimation des composantes de la variance

Source	DL	SC	CM	E(CM)
Entre	$I-1$	$SC_E = J\sum_{i=1}^{I}\left(y_{i.}-y_{..}\right)^2$	$SC_E/(I-1)$	$\gamma + J\gamma_A$
Intra	$I(J-1)$	$SC_I = \sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij}-y_{i.}\right)^2$	$SC_I/[I(J-1)]$	γ

1. Pratique courante/1.2. Discussion du modèle

1.2. Discussion du modèle

Le $j^{ème}$ descendant du $i^{ème}$ taureau est le résultat du croisement de ce taureau et d'une vache prise au hasard dans la population (ou dans une sous population assez importante) : ce choix aléatoire de la mère et la variabilité phénotypique justifient que le caractère y_{ij} soit considéré comme la valeur d'une aléatoire et se traduit par l'introduction de B_{ij}. Par

1. Pratique courante/1.2. Discussion du modèle

contre, il paraît difficile de justifier le caractère aléatoire de la valeur A_i du taureau i : chacun de ces taureaux est un animal bien déterminé et le but poursuivi est précisément de choisir les meilleurs reproducteurs parmi les individus testés. On a souvent prétendu qu'on pouvait considérer les I taureaux comme un échantillon aléatoire de la population des mâles. Mais cet argument ne résiste pas à l'examen : les taureaux testés sont présélectionnés sur des critères permettant d'espérer qu'ils seront satisfaisants et ne représentent qu'eux-mêmes.

1. Pratique courante/1.2. Discussion du modèle

Le modèle d'interprétation des résultats y_{ij} est donc en bonne logique :

$$(2) \quad Y_{ij} = \mu_i + B_{ij}$$

où la loi des B_{ij} est la même que dans le modèle (1) ; par contre le paramètre μ_i, valeur du taureau i, est un nombre réel inconnu mais certain (fixe). Il y a là une contradiction avec le modèle utilisé par les généticiens et les conséquences en sont graves si on utilise pour le modèle (2) les estimateurs classiques des moindres carrés, soit :

$$\hat{\mu}_i = \bar{y}_{i.}$$

En effet, le classement des reproducteurs peut différer beaucoup dans les deux traitements si les effectifs J_i sont très inégaux comme le montre l'exemple numérique suivant :

1. Pratique courante/1.2. Discussion du modèle

$\bar{\gamma}_A/\hat{\gamma} = 0,1$ $\quad \bar{y}_{1.} - \hat{\mu} = 0,77$ $\quad \bar{y}_{2.} - \hat{\mu} = 1$

$J_1 = 100$ $\quad J_2 = 10$

Dans ce cas

$\bar{\bar{A}}_1 = \frac{0,1}{0,11} \times 0,77 = 0,7$ $\quad \bar{\bar{A}}_2 = \frac{0,1}{0,2} \times 1 = 0,5$

et donc

$\bar{\mu}_1 < \bar{\mu}_2$ $\quad \bar{\bar{A}}_1 > \bar{\bar{A}}_2$

2. Les bases de la stat. bayésienne/2.1. Mesure de la vraisemblance sur l'ensemble des paramètres.

L'approche bayesienne complète de ce modèle en probabilisant l'ensemble Θ des paramètres θ ; cette loi de probabilité définit la plus ou moins grande vraisemblance accordée par l'expérimentateur aux diverses valeurs possibles en fonctions de ses connaissances (expériences antérieures, bibliographie,...) ; si par exemple, pour le paramètre réel θ_1 la probabilité de l'intervalle $[u_1,v_1]$ est double de celle de l'intervalle $[u_2,v_2]$, cela signifie que, compte tenu de l'information disponible sur ce paramètre, l'expérimentateur pense qu'il y a

2. Les bases de la stat. bayésienne/2.1. Mesure de la vraisemblance sur l'ensemble des paramètres/suite

$[u_1,v_1]$ contre $[u_2,v_2]$). Cette loi de probabilité est donc un résumé de la connaissance plus ou moins précise qu'a le spécialiste des valeurs des paramètres ; cette loi sera d'autant plus dispersée que les paramètres sont moins bien connus. Le cas où l'information a priori est négligeable s'obtient comme cas limite en faisant tendre la variance de la loi vers $+\infty$.

Dans l'interprétation de résultats expérimentaux, on distingue :

- la loi a priori qui définit l'information avant réalisation de l'expérience,
- la loi a posteriori qui combine l'information a priori et celle apportée par les résultats de l'expérience.

2. Les bases de la stat. bayésienne/2.2. Interprétation des résultats d'une expérience

statistique bayesienne) ; la densité a posteriori de θ est :

$$p_1(\theta/\tilde{y}) = p_0(\theta)\, p(\tilde{y}/\theta) \,/\, \int p_0(\theta)\, p(\tilde{y}/\theta)\, d\theta$$

(l'intégrale est p-dimensionnelle si $\theta=\theta_1,\ldots,\theta_p$).

Cette loi a posteriori contient toute l'information sur θ ; mais il sera souvent nécessaire de condenser cette information et de donner par exemple :

- une valeur numérique $\tilde{\theta}$, l'estimation de θ (si on veut minimiser l'erreur quadratique moyenne, on prendra par $\tilde{\theta}$ l'espérance de la loi a posteriori) ;
- un indicateur de dispersion, la variance ou l'écart-type de la loi a posteriori.

2. Les bases de la stat. bayésienne/2.3. Démarche Bayésienne empirique

2.3. Démarche Bayesienne empirique

Dans certains cas, la loi a priori sur Θ dépend elle aussi de paramètres indéterminés, $\eta=(\eta_1,\ldots,\eta_q)$. Ces paramètres figurent donc dans la loi conjointe de Y et $\underline{\theta}$ et dans la loi marginale de Y qui s'en déduit. La démarche bayesienne empirique consiste à faire l'étude de la loi a posteriori comme si les paramètres $\eta_1,\ldots,\eta_q$ étaient connus et à les remplacer ensuite par des estimations obtenues à partir des résultats expérimentaux (le modèle utilisé est défini par la loi marginale de Y).
Au contraire dans une démarche strictement bayesienne on définira une loi a priori sur l'ensemble H des paramètres $\eta_1,\ldots,\eta_q$ et on pourra utiliser la méthode 2.2.

3. Etude bayésienne empirique du modèle de la sélection/3.1. loi a priori associée aux espérances μ_i

3.1. Loi a priori associée aux espèrances μ_i

Dans le cas du bétail charolais élevé en petites exploitations avec un faible taux d'insémination artificielle, les taureaux soumis au testage sont en général inconnus (en particulier le père n'a pas été testé). La seule information qu'on possède est donc que ces taureaux proviennent d'élevages du même type et ont été présélectionnés sur les mêmes critères.

3. Etude bayésienne empirique du modèle de la sélection/3.1. loi a priori associée aux espérances μ_i/suite

1. L'information a priori sur ces taureaux est donc invariante par une quelconque permutation, ce qui implique que les aléatoires $\underline{\mu}_i$ ont toutes même loi. L'espérance et la variance de cette loi seront notées μ et γ_A

2. La connaissance de l'écart $\mu_h - \mu$ c'est-à-dire de la supériorité du taureau h sur la moyenne attendue du groupe n'apporte aucune information sur l'écart $\mu_i - \mu$ relatif au taureau i, ce qui implique l'indépendance a priori des $\underline{\mu}_i$

Les aléatoires $\underline{\mu}_i$ sont indépendantes et ont toutes même loi normale $N(\mu, \gamma_A)$

Il est commode d'introduire les aléatoires centrées $\underline{\alpha}_i = \underline{\mu}_i - \mu$ qui sont indépendantes normales centrées de variance γ_A.

Dans la loi a priori des $\underline{\mu}_i$ subsistent deux paramètres inconnus μ et γ_A ; le modèle contient un troisième paramètre inconnu la variance γ des B_{ij}.

3. Etude bayésienne empirique du modèle de la sélection/3.2. Estimation des μ_i, « indexation »

L'étude mathématique est très simple si les paramètres μ, γ_A et γ sont connus ; en effet dans ce cas, la loi a posteriori de $\underline{\alpha}_i$ (sachant $Y=\bar{y}$) est une loi normale définie par son espérance :

$$\bar{A}_i = \frac{\gamma_A}{\gamma_A+\gamma/J_i}(\bar{y}_{i.} - \mu)$$

et sa variance

$$\delta_i = \gamma_A \frac{\gamma/J_i}{\gamma_A+\gamma/J_i}$$

En remplaçant dans l'expression de $\hat{A}_i$ les paramètres inconnus par leurs estimations on obtient l'index du taureau i :

$$\hat{A}_i = \frac{\bar{\gamma}_A}{\bar{\gamma}_A+\bar{\gamma}/J_i}(\bar{y}_{i.} - \hat{\mu})$$

3. Etude bayésienne empirique du modèle de la sélection/3.2. Estimation des μ_i, « indexation »

$$\hat{A}_i = \underbrace{\frac{\gamma_A}{\gamma_A + (\gamma / J_i)}}_{CD_i = \frac{J_i}{J_i + \lambda} \text{ où } \lambda = \gamma / \gamma_A} (y_{i.} - \mu) \Leftrightarrow \boxed{\hat{A}_i = \frac{\frac{y_{i.} - \mu}{\gamma / J_i} + \frac{0}{\gamma_A}}{\frac{1}{\gamma / J_i} + \frac{1}{\gamma_A}}}$$

$$\delta_i = \underbrace{\gamma_A \frac{\gamma / J_i}{\gamma_A + (\gamma / J_i)}}_{(1 - CD_i)\gamma_A} \Leftrightarrow \boxed{\frac{1}{\delta_i} = \frac{1}{\gamma / J_i} + \frac{1}{\gamma_A}}$$

4. Discussion/4.1. Limite du modèle

4.1. Limites du modèle et aménagements

La validité du modèle est conditionnée par le choix correct de la loi a priori. L'hypothèse de base est que l'information a priori est invariante par une permutation des taureaux ; or cette hypothèse n'est pas toujours vérifiée et on peut en donner quelques exemples classiques.

1. Si deux des taureaux testés 1 et 2 par exemple sont apparentés, on peut encore admettre que les aléatoires $\underline{\alpha}_i$ ont toutes même loi marginale ; mais $\underline{\alpha}_1$ et $\underline{\alpha}_2$ ne sont plus indépendantes

4. Discussion/4.1. Limite du modèle et aménagements

2. Si le père du taureau 1 a été testé, il n'y a plus permutabilité de l'information sur le taureau 1 et de l'information sur les autres taureaux et il faut tenir compte des résultats du père qui sera indexé par 0 dans la définition de la loi a priori de $\underline{\alpha}_1$.

Par exemple, on pourra considérer que, avant testage, l'information sur le père était équivalente à celle relative aux autres taureaux, introduire entre $\underline{\alpha}_0$ et $\underline{\alpha}_1$ une covariance $\gamma_A/2$ (suggérée par la covariance père fils) et prendre pour loi a priori de $\underline{\alpha}_1$ la loi conditionnelle de $\underline{\alpha}_1$ sachant $Y_{0.} = \tilde{y}_{0.}$ (moyenne des résultats des fils du taureau 0).

4. Discussion/4.2. Propositions

4.2. Propositions

La démarche bayesienne empirique ne me paraît pas totalement satisfaisante ; la cohérence de la méthode implique une étude complètement bayesienne d'autant plus justifiée qu'on dispose d'une information non négligeable sur les paramètres γ et γ_A.

1. Le paramètre γ est la variance intrafamilles de demi-frères ; on a donc des idées assez précises sur sa moyenne et sa dispersion et on peut lui associer une mesure a priori bien définie (classiquement on pose que γ^{-1} a une loi $\lambda\chi^2_\nu$ dépendant des deux paramètres λ et ν qu'on fixe en choisissant moyenne et dispersion).

4. Discussion/4.2. Propositions

il suffit donc de prendre pour μ une loi a priori correspondant à une information négligeable.

2. Le rapport $4\gamma_A / (\gamma_A+\gamma)$ serait l'héritabilité du caractère si les taureaux étaient pris au hasard dans la population ; il est donc compris entre 0 et 1 et sans doute voisin de l'héritabilité mais un peu inférieur (on peut penser que la variance γ_A entre taureaux présélectionnés est inférieure à la variance dans la population) ; on en déduit aisément la loi a priori du rapport γ_A/γ.

$$h^2 = 4\gamma_A/(\gamma_A+\gamma);\ \lambda=\gamma/\gamma_A \Rightarrow \lambda=(4/h^2)-1$$

$$h^2 \sim \pi(.) \Rightarrow \lambda \sim \omega(.)$$

4. Discussion/4.2. Propositions

L'étude précédente permet de définir complètement la loi a priori des paramètres $\underline{\mu}_i$ (ou $\underline{\alpha}_i$) et γ. Dans une première étape discutée en 3.1, on laisse dans cette loi trois paramètres inconnus μ, γ_A, γ, ou, de manière équivalente, on a donné la loi conditionnelle connaissant μ, γ_A, γ ; dans une seconde étape on vient de définir la loi de $\underline{\mu}$, $\underline{\gamma}_A$, $\underline{\gamma}$, ce qui (probabilités composées) détermine complètement la loi conjointe de $\underline{\mu}_1,\ldots,\underline{\mu}_I$, $\underline{\mu}$, $\underline{\gamma}_A$, $\underline{\gamma}$ et donc la loi conjointe de $\underline{\mu}_1,\ldots,\underline{\mu}_I$, $\underline{\gamma}$ marginale de la précédente.