J.-L. FOULLEY et E. MANFREDI

INRA Station de Génétique quantitative et appliquée 78352 Jouy-en-Josas Cedex

L'évaluation des reproducteurs

L'évaluation génétique des reproducteurs pour des caractères à seuil

Résumé. *Cet article rappelle les principales caractéristiques du modèle à seuils de Sewall Wright applicable aux variables discrètes binaires et polytomiques ordonnées ainsi que ses principaux domaines d'application notamment en génétique et sélection animale. En prenant l'exemple d'un caractère dichotomique, on montre que l'analyse statistique de ces caractères rentre dans le cadre de la théorie du modèle linéaire généralisé de Mc Cullagh et Nelder. On mentionne ensuite l'approche bayésienne de Gianola et Foulley d'évaluation des reproducteurs. Diverses extensions sont enfin discutées.*

L'évaluation génétique des reproducteurs repose actuellement sur le BLUP ("Best linear unbiased prediction", Henderson 1984) pour les paramètres de position et le REML ("Restricted maximum likelihood, Patterson et Thompson 1971) pour les paramètres de dispersion. Ces méthodes statistiques se justifient pleinement dans le cadre du modèle linéaire gaussien.

Dans le cas de variables discrètes (tableau 1), l'application directe ou après aménagement du BLUP pose de sérieuses difficultés conceptuelles liées à la dépendance entre la fréquence et la variance des caractéristiques discrètes étudiées (Foulley 1987). Quant aux algorithmes de calcul du REML, l'application de ceux-ci aux variables discrètes ne répond qu'à des motifs d'opportunité calculatoire. Dans l'esprit de l'analyse de variance figurent les méthodes inférentielles de Taguchi qui sont très usitées dans l'industrie mais peu connues en sélection et qui s'avèrent en tout état de cause très critiquables d'un point de vue théorique. Par ailleurs, l'analyse des données fournit tout une gamme d'outils intéressants pour le traitement statistique des données catégorielles qui sont particulièrement adaptés à une approche statistique descriptive et exploratoire mais qui se révèlent plus difficiles à exploiter dans une optique inférentielle comme c'est le cas en génétique et sélection.

Tableau 1. *Typologie sommaire des variables discrètes.*

Variable	Expression	Exemple
Binaire	0-1	mort-vivant
Polytomique ordinale	1-2-3	facile, assité, difficile
Polytomique nominale	A-B-C	pointage
Comptage	0-1-2-3	nombre d'ovules
Classement	1er, 2ème, nème	résultats de course

Le modèle "béta-binomial" des données ou son pendant "Dirichlet-multinomial" pour plusieurs catégories, offre un cadre conceptuel plus rigoureux et intéressant vis-à-vis de l'inférence statistique ; il autorise en particulier le développement d'estimateurs du maximum de vraisemblance qui sont étroitement apparentés au BLUP (Foulley 1987). Malheureusement, ce modèle n'est pas généralisable à une situation plus complexe que celle d'un modèle aléatoire à un seul facteur.

L'analyse génétique des caractères discontinus n'a eu de cesse de préoccuper les chercheurs depuis les origines de la génétique. L'expression discrète des phénotypes incline naturellement à une approche factorielle du déterminisme génétique avec, toutefois, de sérieuses difficultés d'ajustement du modèle aux observations à moins d'un recours à des concepts "ad hoc" tels que celui de pénétrance et d'expressivité variable par exemple. De même, l'étude de la transmission du caractère d'une génération à la suivante ne peut plus s'appréhender simplement comme en présence d'un caractère continu, par les techniques classiques de régression et de corrélation. Il faut alors analyser des tables de contingence par des indices d'association spécifiques de telles structures.

1 / Le modèle à seuils

L'idée d'une susceptibilité normale sous-jacente à l'expression du caractère s'est fait jour et s'est développée peu à peu dans l'esprit des chercheurs pour pallier toutes ces difficultés. Pearson apparaît comme un pionnier dans ce domaine ; fort de sa maîtrise de la distribution multinormale, il introduit le concept de corrélation tétrachorique entre variables discrètes pour quantifier les ressemblances entre apparentés en terme classique de corrélation. Wright (1934)

introduit le modèle à seuils pour rendre compte de l'écart à des proportions mendéliennes monofactorielles dans l'analyse de l'hérédité du nombre de doigts du membre postérieur lors de croisements entre lignées de cobaye (tableau 2).

Tableau 2. *Rappel des travaux de Wright (1934) sur des lignées consanguines de cobaye différent par le nombre de doigts.*

	Phénotypes[1]		Génotypes		
	{+}	{d}	++	+d	dd
Lignée 2	1	0	1	0	0
Lignée D	0	1	0	0	1
F_1	1	0	0	1	0
F_2	3/4	1/4	1/4	1/2	1/4
$R = F_1 \times D$	1/2	1/2	0	1/2	1/2
{+}R x D	23%[2] (1/2)[3]	(1/2)[3]	0	1/2	1/2
{d}R x D	16%[4] (0)[3]	(1)[3]	0	0	1

(1) {+} 3 doigts ; {d} 4 doigts
(2) valeur observée sur 186 individus
(3) proportions attendues sous l'hypothèse d'un gène récessif (d) autosomal responsable du doigt surnuméraire
(4) valeur observée sur 119 individus.

Le formalisme du modèle à seuils est en fait très simple, notamment pour un caractère tout-ou-rien comme le rappelle le développement suivant. Pour des raisons de simplicité d'exposition, nous limiterons cette présentation au modèle à seuils relatif à des réponses dichotomiques dit "threshold dichotomy distribution" dans la terminologie du généticien Wright ou "probit normal binomial distribution" dans celle des statisticiens sachant que ce modèle s'étend très aisèment au cas polytomique ordinal. Désignons par X la variable aléatoire relative au phénotype sous-jacent d'un individu d'une population donnée; on suppose que X est distribuée sur une échelle continue sous-jacente munie d'un seuil τ, suivant une loi normale $N(\mu,\sigma^2)$, de moyenne μ et de variance σ^2 ; dans ces conditions, la probabilité qu'un individu tiré au hasard dans la population présente un des phénotypes tout-ou-rien (Y = 1 par exemple) est donnée par :

$$\Pi = \int_{\tau}^{+\infty} (2\pi)^{-1/2} \exp[-(t-\mu)^2 / 2\sigma^2]\, dt; \quad (1)$$

Après le changement de variable $t^* = (t - \mu) / \sigma$, cette probabilité s'exprime à partir de la fonction de répartition $\Phi(.)$ de la loi normale par :

$$\Pi = \Phi[(\mu - \tau) / \sigma] \quad (2)$$

avec pour argument, l'écart standardisé de la moyenne de la population au seuil.

La non linéarité de la relation entre expressions binaire et sous-jacente se manifeste également au niveau des valeurs génétiques définies sur ces deux échelles . En effet, si l'on suppose un déterminisme génétique sous-jacent purement additif, on peut écrire $X = \mu + a + e$ où $a \sim N(0, \sigma_a^2)$ et $e \sim N(0, \sigma_e^2)$ désignent les effets génétiques et de milieu respectivement; la valeur génétique sur l'échelle binaire (g) correspond par définition au phénotype moyen des individus ayant tous la même valeur génétique (sous-jacente) soit $g = Pr(X \geq \tau \mid \mu, a)$. Si l'on place l'origine au seuil ($\tau = 0$), g s'exprime par $g = \Phi[(\mu + a) / \sigma_e]$. On peut alors calculer aisément les moments de cette variable aléatoire, ce qui permet d'expliciter les relations entre paramètres génétiques sur les deux échelles. Comme l'ont montré Foulley et Im (1989), cette variable a pour espérance :

$$\Pi = E(g) = \Phi[\mu / (\sigma_a^2 + \sigma_e^2)^{1/2}] \quad (3)$$

et pour variance :

$$\sigma_g^2 = \Phi_2(\tilde{\mu},\tilde{\mu};h^2) - \Phi^2(\tilde{\mu}) \quad (4)$$

où $\tilde{\mu} = \mu / (\sigma_a^2 + \sigma_e^2)^{1/2}$, $h^2 = \sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$

est l'héritabilité du caractère sur l'échelle sous-jacente et $\Phi_2(\alpha,\beta;\rho)$ est la fonction de répartition de la loi binormale réduite d'arguments α,β et de corrélation ρ. L'expression classique

$$\sigma_g^2 = h^2 \phi^2(\tilde{\mu}) \quad (5)$$

donnée par Robertson (1950) et Dempster et Lerner (1950) où $\Phi(.)$ est la densité de la loi normale réduite, correspond en fait à une approximation au premier ordre de la formule précédente au voisinage de $h^2 = 0$ (Foulley et Im 1989).

D'un point de vue génétique, l'hypothèse de normalité sur un continuum sous-jacent s'accorde bien avec celle d'un déterminisme polygénique classiquement adoptée dans l'étude des caractères quantitatifs. L'analyse génétique des caractères discrets à seuils s'intègre donc naturellement dans le cadre habituel de la génétique quantitative et de ses concepts. Il en résulte une cohérence de l'analyse, en particulier dans l'étude d'un mélange de caractères discrets et continus (Foulley *et al* 1983) et dans celle de caractères à hérédité mixte impliquant un gène majeur et des polygènes (Foulley et Elsen 1988). Le caractère attractif de ce modèle s'est concrétisé par de nombreuses applications dans divers secteurs tels que par exemple les suivants :

- sensibilité aux maladies et anomalies congénitales chez l'homme (Falconer, 1965; Curnow et Smith, 1975) comme chez l'animal (cf. par exemple le syndrome dit "des pattes écartées" chez le porc) ;

- déterminisme génétique et environnemental du sexe (cf. une application chez certains poissons et la tortue) ;

- caractéristiques de reproduction et d'adaptation en zootechnie telles que la fertilité, les difficultés de vêlage des bovins, la taille de portée et la survie des agneaux, la gémellité des bovins, la morphologie des pieds et la qualité de la laine chez le mouton.

2 / Le modèle statistique

D'un point de vue statistique, le modèle à seuils est un cas particulier de la théorie des modèles linéaires généralisés (Mc Cullagh et Nelder 1989) puisque dans son développement le plus simple d'une variable binaire, il s'explicite grâce à une fonction de lien "probit" $\Phi^{-1}(\Pi)$. De ce fait, le modèle à seuils peut être abordé dans un cadre statistique très riche qui ouvre sur des applications dépassant largement le domaine de la génétique humaine et de la sélection animale pour s'étendre par exemple à la neurophysio-

logie et la séismologie, à la théorie des sondages, à la psychologie, aux sciences sociales et à l'économétrie (Foulley et Manfredi 1991).

Les modèles utilisés en sélection animale sont classiquement des modèles mixtes des facteurs de variation (Henderson 1984) impliquant d'une part des effets fixes relatifs à des facteurs environnementaux (année, saison, élevage, type de conduite) et de niveau génétique des populations (effet "groupe") et, d'autre part, des effets aléatoires correspondant aux individus candidats à la sélection et retenus (effet "père" ou effet " animal" par exemple). De plus, à des fins de sélection, l'inférence statistique porte à la fois sur l'estimation de certains effets fixes et sur la prédiction d'effets aléatoires. Il y a là une originalité qui n'a pas toujours été prise en compte par la statistique générale et qui a motivé un intérêt et des développements statistiques spécifiques de la part des généticiens quantitatifs.

Soit y_{jr} la variable aléatoire binaire (0,1) relative à la rème observation de la classe (j = 1,2,...,J). Conformément à la théorie du modèle linéaire généralisé (McCullagh et Nelder 1989), la probabilité de réponse $\Pi_j = Pr(y_{jr} = 1 | \theta)$ d'une observation de la classe j est transformée par la fonction de lien probit de façon à obtenir un prédicteur linéaire des variables explicatives

$$\Pi_j = \Phi(\eta_j) \quad (6a)$$

$$\eta_j = \Phi^{-1}(\Pi_j) = x_j'\beta + z_j'u = t_j'\theta \quad (6b)$$

où $\theta = (\beta',u')'$ est un vecteur regroupant les vecteurs des effets fixes (β) et aléatoires

$[u \sim N(0,\Sigma_u)]$ et $t_j' = (x_j',z_j')$

est le vecteur ligne d'incidence correspondant.

Différentes méthodes statistiques sont envisageables pour estimer les paramètres de position (θ) et de dispersion (Σ_u) de ce modèle. Foulley et Manfredi (1991) distinguent les trois grandes méthodes suivantes : 1) l'approche linéaire de Grizzle, Starmer et Koch et son extension bayésienne au modèle mixte par Foulley et Im (1989) ; 2) l'approche du modèle linéaire généralisé et de la quasi-vraisemblance de Gilmour, Anderson et Rae (1985-87) et enfin 3) la méthode bayésienne du mode conjoint a posteriori (MAP) de Gianola et Foulley (1983) et Harville et Mee (1984).

De nombreuses applications de la méthodologie GF-HM ont vu le jour en sélection animale. Un des domaines priviligiés de celles-ci concernent l'évaluation génétique des taureaux sur les difficultés de vêlage (expérience pilote "Maine Anjou", cf tableau 3).

3 / Extension à d'autres situations

La méthode GF-HM se généralise très bien à des polytomies ordonnées (Gianola et Foulley 1983). Pour les C catégories ordonnées délimitées par les seuils $(\tau_1,\tau_2,...,\tau_c,...,\tau_{c-1})$, on peut écrire sachant le paramètre (θ) et avec la convention

$(\tau_0 = -\infty; \tau_c = +\infty)$

$$\begin{cases} \Pi_{jc} = \Phi(\tau_c - \eta_j) - \Phi(\tau_{c-1} - \eta_j) \\ \eta_j = t_j'\theta \end{cases} \quad (7)$$

Maintes extensions ont été effectuées dans le cadre de l'approche bayésienne, notamment dans les situations suivantes : réponses binaires multiples complètes ou avec information manquante (Foulley 1987) ; mélange de variables binaires et continues (Foulley *et al* 1983) ; mélanges de variables binaires et de Poisson (Foulley *et al* 1987). L'approche a été également étendue à des situations d'assignation incertaine des observations à certains facteurs de variation tels que le génotype majeur (Foulley et Elsen 1988) ou la paternité (Foulley 1987).

Cette approche du modèle linéaire généralisé peut être étendue à d'autres distributions de variables discrètes intervenant en sélection animale telle que la distribution de Poisson (taux d'ovulation, prolificité) comme l'ont montré Foulley *et al* (1987).

Identification		Effet "Père de veau"					Effet "Père de vache"		
Nom	Numéro	Nbre de. veaux	% observé	% vêlages "génisses"	Indice % ajusté(2)	Indice standard(3)	Nbre de petits fils	Indice % ajusté(2)	Indice standard(3)
Bison	53 86 122 529	126	93,7	12	84,0	0,52			
Arondi	53 85 124 269	133	85,7	8	76,1	-1,10			
Upont	53 83 122 061	91	93,4	12	85,0	0,74	33	81,7	-0,01
Tartuffe	53 82 124 858	120	90,8	9	81,2	-0,10	25	83,4	0,49
Aricot	53 85 123 140	223	93,3	34	87,5	1,40	15	82,8	0,30
Avril	49 85 122 504	274	88,3	3	76,0	-1,12	23	82,4	0,19
Telo	49 82 124 614	199	91,0	15	83,6	0,43	41	83,1	0,40
Rinola	53 80 126 125	479	90,6	58	87,9	1,51	41	85,0	0,97
Pingouin	53 79 126 131	1032	85,8	67	83,4	0,37	120	81,9	0,06
Raisin	85 80 128 099	723	91,1	9	81,1	-0,12	114	84,1	0,69
Lascar	53 75 123 925	2100	90,0	43	85,7	0,91	576	80,7	-0,27
Liran	53 75 123 998	1717	92,1	5	80,5	-0,26	574	82,3	0,18

Tableau 3. *Classement de taureaux Maine Anjou utilisés en insémination artificielle sur les facilités de vêlage (notes 1 et 2) effets "pères de veau et de vache" (Inra(1) - Upra-Union Maine Anjou : résultats 1991).*

(1) Méthodologie de Gianola et Foulley (1983) ; algorithme de Manfredi ; calculs de Manfredi et San Cristobal
(2) Exprimé en fréquence de vêlages faciles (conditions 1 et 2) pour des vêlages de ***génisses***
*(3) En unité d'****écart type de valeur génétique*** *transmise.*

Références bibliographiques

Curnow R, Smith C.,1975. Multifactorial models for familial diseases in man. J R Statist Soc A 138, 131-169.

Dempster E. R.Lerner I.M., 1950. Heritability of threshold characters. Genetics 35: 212-236.

Falconer DS.,1965. The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann. Hum. Genet ., 29, 51-76.

Foulley J. L., 1987. Méthodes d'évaluation des reproducteurs pour des caractères discrets à déterminisme polygénique en sélection animale. Thèse d'Etat, Université de Paris-Sud-Orsay.

Foulley J.L., Elsen J.M., 1988. Posterior probability of the sire's genotype at a major locus based on progeny test results for discrete characters. Génét. Sél. Evol., 20, 227-238.

Foulley J.L., Manfredi E., 1991. Approaches statistiques de l'évaluation génétique des reproducteurs pour des caractères binaires à seuils. Genet. Sel. Evol., 23, 309-338.

Foulley J. L., Im S., 1989. Probability statements about the transmitting ability of progeny-tested sires for an all-or-none trait with an application to twinning in cattle. Genet. Sel. Evol., 21, 247-267.

Foulley J.L., Gianola D., Thompson R., 1983. Prediction of genetic merit from data on binary and quantitative variates with an application to calving difficulty birth weight and pelvic opening. Génét. Sél. Evol., 15, 401-424.

Foulley J.L., Gianola D., Im S., 1987. Genetic evaluation for traits distributed as Poisson-binomial with reference to reproductive traits. Theor. Appl. Genet., 73, 870-877.

Gianola D., Foulley J.L., 1983. Sire evaluation for ordered categorical data with a threshold model. Génét. Sél. Evol., 15, 201-224.

Gilmour A. , Anderson R. D., Rae A., 1985. The analysis of binomial data by a generalized linear mixed model. Biometrika, 72, 593-599.

Harville D.A., Mee R.W., 1984. A mixed model procedure for analyzing ordered categorical data. Biometrics , 40, 393-408.

Henderson C. R., 1984. Applications of linear models in animal breeding, University of Guelph, Guelph.

McCullagh P., Nelder J.,1989. Generalized linear models, 2nd ed. Chapman & Hall, London.

Patterson H. D., Thompson R.,1971. Recovery of interblock information when block sizes are unequal. Biometrika , 58, 545-558.

Robertson A., 1950. Proof that the additive heritability on the p scale is given by the expression $\bar{z}^2 h_x^2 / \bar{p}\bar{q}$. Genetics, 35, 234-236.

Wright S., 1934. The results of crosses between inbred strains of guinea pigs differing in number of digits. Genetics 19, 537-551.