Génét. Sél. Evol., 1983, **15** (1), 91-98

Selection on selected records

B. GOFFINET

I.N.R.A., Laboratoire de Biométrie,
Centre de Recherches de Toulouse, chemin de Borde-Rouge,
F 31320 Castanet-Tolosan

Summary

The problem of selecting individuals according to their additive genetic values and of estimating those values, is considered. It is assumed that the selection is based on a vector of observations made on a group of individuals which were themselves selected according to a certain vector of observations.

An optimal selection rule applicable irrespective of the distribution of the random variable involved in the setting is derived. In particular, it is shown that the restrictions regarding the use of the BLUP (Best Linear Unbiased Predictor) pointed out by HENDERSON, can be relaxed.

Key-words : Selection, mixed models, BLUP.

Résumé

Sélection sur données issues de sélection

On considère le problème de la sélection d'individus pour leurs valeurs génétiques additives et de l'estimation de ces valeurs. La sélection est basée sur un vecteur d'observations faites sur un ensemble d'individus eux-mêmes issus d'une sélection sur un certain vecteur d'observations.

On obtient une règle optimale de sélection applicable quelle que soit la distribution des variables aléatoires de l'expérience. En particulier, on montre que les contraintes d'utilisation du BLUP (meilleur prédicteur linéaire sans biais) proposées par Henderson, peuvent être atténuées.

Mots-clés : Sélection, modèle mixte, BLUP.

I. Introduction

Animal and plant breeders are often faced with the problem of choosing items, e.g. sires or varieties, among a set of available candidates. Generally, selection is based on a vector of observations made on these or other items which were themselves selected according to another vector of observations. Therefore, it is important to develop a selection rule that is optimal in some sense. HENDERSON (1973, 1975), in a multi-

variate normal setting, showed that if certain conditions related to fixed parameters in a linear model describing the observations are met, then the best linear unbiased predictor (BLUP) eliminates the bias resulting from the previous selection, and retains its properties.

The objective of this article is to derive an optimal selection rule applicable irrespective of the distribution of the random variables involved in the setting. In particular, it is shown that the restrictions regarding the use of BLUP pointed out by HENDERSON can be relaxed. As the problem of best estimating the merit of the candidates for selection, e.g. sires, is closely related to the development of an optimal selection rule, this is also adressed here.

II. Setting an optimality criteria

To illustrate, consider two sires with one progeny each. A variable Y is measured in these two progeny and we assume the model :

$$Y_{ij} = s_i + e_{ij} \tag{1}$$

where s_i is the genetic value of sire i (i = 1, 2) and e_{ij} represents variability about it.

Thus we have :

$$Y_{11} = s_1 + e_{11} \text{ sire 1, progeny 1} \tag{2}$$

$$Y_{21} = s_2 + e_{21} \text{ sire 2, progeny 1} \tag{3}$$

On the basis of the first progeny, one of the sires, say sire *i*, seems more promising, so Y is measured on a second progeny and we have :

$$Y_{i2} = s_i + e_{i2} \text{ sire i, progeny 2} \tag{4}$$

The problem is to estimate s_1 and s_2 and to select one of the two males to be kept as a breeder.

Let $s' = [s_1\ s_2]$ be the vector of genetic values. Optimality is achieved by finding indicator variables F_1 and F_2 such that :

$$\Omega = E\ (F_1\ s_1 + F_2\ s_2) \tag{5}$$

is maximum ; the variables F_1 and F_2 depend on the data. As, in general, a fixed number of sires is to be selected — one in the case of the example — we can take :

$$F_1 + F_2 = \text{Constant}, \tag{6}$$

COCHRAN (1951), studied :

$$E\ (F_1 + F_2) = \text{Constant}, \tag{7}$$

so this less restrictive constraint will not be considered here. Further, we define as best estimator of s_i, the function of the data $\tilde{s}_i$ which minimizes average squared risk :

$$\Omega_i = E\ (s_i - \tilde{s}_i)^2\ ;\ i = 1, 2 \tag{8}$$

We also consider another random variable :

$$N = h\ (Y_{11}, Y_{21})$$

which is a function of the values of the first progeny of the two sires. This variable takes the value 1 or 2, depending on which of the two sires was considered more promising and so measured on a second progeny.

Let us consider now the case where the variable Y is measured on a second progeny of the sire 1 whatever the values taken by Y_{11} and Y_{21}.

The measured variable is now $\dot{Y}_{12}$ and the restriction of $\dot{Y}_{12}$ to $N = 1$, is Y_{12} (we define also $\dot{Y}_{22}$ with the same manner).

It is difficult to specify the probability law of Y_{12}, but the two joint laws :

$$(Y_{11}, Y_{21}, s, \dot{Y}_{i2}) \qquad i = 1, 2 \tag{9}$$

can be considered know.

The estimator $\tilde{s}_i$ of s_i which minimizes Ω_i must also minimize :

$$E\left((s_i - \tilde{s}_i)^2 \mid Y_{11} = y_{11}, Y_{21} = y_{21}, N = n, Y_{n2} = y_{n2}\right) \tag{10}$$

So, we get $\tilde{s}_i$ which minimizes Ω_i in the case where we observe n :

$$\begin{aligned} \tilde{s}_i &= E\left(s_i \mid Y_{11} = y_{11}, Y_{21} = y_{21}, N = n, Y_{n2} = y_{n2}\right) \\ &= E\left(s_i \mid Y_{11} = y_{11}, Y_{21} = y_{21}, N = n, \dot{Y}_{n2} = y_{n2}\right) \end{aligned} \tag{11}$$

As the value of $N = h(Y_{11}, Y_{21})$ is known once Y_{11} and Y_{21} are realized, we get :

$$\tilde{s}_i = E\left(s_i \mid Y_{11} = y_{11}, Y_{21} = y_{21}, \dot{Y}_{n2} = y_{n2}\right) \text{ for } N = n \tag{12}$$

Note that when s_i, Y_{11}, Y_{21}, $\dot{Y}_{n2}$ are tetravariate normal, (12) yields the best linear predictor of s_i from Y_{11}, Y_{21} and $\dot{Y}_{n2}$.

From (5) and (6), the optimal selection policy is similarly obtained by maximizing :

$$E\left(F_1 s_1 + F_2 s_2 \mid Y_{11} = y_{11}, Y_{21} = y_{21}, \dot{Y}_{n2} = y_{n2}\right) \text{ for } N = n \tag{13}$$

subject to $F_1 + F_2 = 1$, to observe (6). If sire 1 is selected, $F_1 = 1$ and $F_2 = 0$, and (13) becomes :

$$E\left(s_1 \mid Y_{11} = y_{11}, Y_{21} = y_{21}, \dot{Y}_{n2} = y_{n2}\right) = \tilde{s}_1 \text{ for } N = n \tag{14}$$

and likewise $\tilde{s}_2$ if sire 2 is selected. Therefore, to maximise (13), we order the sires on the basis of the values of $\tilde{s}_1$ and $\tilde{s}_2$ (equation 12) and choose the individual with the largest $\tilde{s}_i$.

III. General case with known arbitrary density

In general, there is a first stage in which q_o candidates, e.g. sires, have data represented by a vector Y_o, containing information on one or several variables. For example, Y_o may represent progeny records on body weight and conformation score at weaning in beef cattle. The vector of genetic values is s and it may include the « merit » for one or more traits, or functions thereof.

In the second stage, N experiment plans are possible. To the experiment plan n, corresponds the random vector Y_n. The vector Y_n that will be measured in the second stage depends on the realization of the random variable :

$$N = h(Y_o, \varepsilon) \tag{15}$$

where ε represents independent externalities such as random deaths of sires. The variate N can take values from 1 to N, and associated with each value of N there is a different configuration of the second stage setting. Further, Y_n, will comprise data from q_n sires. While in general $q_n < q_o$, this is not necessarily so as all sires may be kept for the second stage but allowed to reproduce at different rates.

As in II, we define $\dot{Y}_1, \dot{Y}_2, \ldots, \dot{Y}_N$. $\dot{Y}_n$ corresponds to the random vector measured on the experiment plan n if this plan n was used whatever the value of N (e.g. if there was not preselection).

The restriction of $\dot{Y}_n$ to $N = n$ is Y_n.

The N joint probability laws :

$$L(Y_o, s, \dot{Y}_n) \qquad n = 1, \ldots, N$$

are assumed known.

Similarly to (11) and (12), the best estimator of s is :

$$\begin{aligned} \tilde{s} &= E(s \mid Y_o = y_o, N = n, Y_n = y_n) \\ &= E(s \mid Y_o = y_o, N = n, \dot{Y}_n = y_n) \end{aligned} \tag{17}$$

Since $(Y_o, \dot{Y}_n, s)$ and ε are independent, and since N is a function of Y_o and ε, (17) can be written as :

$$\tilde{s} = E(s \mid Y_o = y_o, \dot{Y}_n = y_n) \text{ for } N = n \tag{18}$$

As in (13), the optimal selection policy results from ranking the sires on the basis of the values of (18) and then choosing those with the largest values.

The results generalize to a k-stage selection setting. If $Y^{[k]}_{n_k}$ $(n_k = 1, \ldots, N_k)$ indicates the vector that will be measured in the k^{th} stage $(k = 1, \ldots, K)$ following preselection, then :

$$\tilde{s} = E(s \mid Y_o = y_o, \dot{Y}^{[1]}_{n_1} = y^{[1]}_{n_1}, \ldots, \dot{Y}^{[K]}_{n_k} = y^{[K]}_{n_k}) \tag{19}$$

if we define $\dot{Y}^{[k]}_{n_k}$ as before, gives the best estimator of merit, and ranking with (19) optimizes the selection program. Note that in the multivariate normal case (18) and (19) give the best linear predictor, or classical selection index in certain settings (Smith, 1936 ; Hazel, 1943). (This is correct despite the fact that the random variable $Y^{[k]}_i$, restricted to the case where they are in fact observed, don't have a normal distribution.)

IV. Case with unknown first moments

Often the expectations of the random variables $Y_o, \dot{Y}_1, \ldots, \dot{Y}_N$ are unknown, but one assumes a linear model :

$$\begin{aligned} E(Y_o) &= A_o\beta_o \\ E(\dot{Y}_n) &= A_n\beta_n \qquad n = 1, \ldots, N \end{aligned}$$

where $A_o, A_1, \ldots, A_n$ are the known matrices of the indicators and $\beta_o, \beta_1, \ldots, \beta_N$ are the unknown vectors of the fixed effects. The vectors $\beta_o, \beta_1, \ldots, \beta_N$ might

have values in common, for example in the case where Y_o and $\dot{Y}_n$ represent the same trait measured for different individuals. In general one can write :

$$E(Y_o) = A_o\beta$$
$$E(\dot{Y}_n) = A_n\beta \qquad n = 1, \ldots, N$$

The N joint probability laws :

$$(Y_o - E(Y_o),\ s,\ \dot{Y}_n - E(\dot{Y}_n)) \qquad n = 1, \ldots, N$$

will be assumed known. The class of estimators (or criteria of selection) $\hat{s}$ will be restricted to the class of functions which are invariant under translation, i.e. functions that satisfy :

$$\hat{s} = f[y_o,\ n,\ y_n] = f[y_o + A_o\beta,\ n,\ y_n + A_n\beta]\ \forall\beta \tag{20}$$

Under this restriction, the estimators (or criteria of selection) $\hat{s}$ take the same values as vector β moves.

Let :

$$A_{on} = \begin{pmatrix} A_o \\ A_n \end{pmatrix}$$

and let P_{on} be a projector onto the orthogonal to the space spanned by the columns of A_{on}. Let :

$$V_n = \text{var}\begin{pmatrix} Y_o \\ Y_n \end{pmatrix}.$$

We may chose :

$$P_{on} = I - A_{on}(A'_{on}\, V_n^{-1}\, A_{on})^{-1}\, A'_{on}\, V_n^{-1}$$

Note that P_{on} eliminates fixed effects and retains the most information.

The set E_1 of functions $f(y_o,\ n,\ y_n)$ which satisfies (20) is the same as the set E_2 of functions of the form :

$$g\left(P_{on}\begin{pmatrix} y_o \\ y_n \end{pmatrix},\ n\right)$$

where g is any function.

Proof :

- $E_2 \subset E_1$

$$g\left(P_{on}\begin{pmatrix} y_o \\ y_n \end{pmatrix},\ n\right) = g\left(P_{on}\left(\begin{pmatrix} y_o \\ y_n \end{pmatrix} + A_{on}\beta\right),\ n\right)$$

- $E_1 \subset E_2$

$$\text{left f be invariant and } P_{on}\begin{pmatrix} y_{o1} \\ y_{n1} \end{pmatrix} = P_{on}\begin{pmatrix} y_{o2} \\ y_{n2} \end{pmatrix} \rightarrow$$
$$\begin{pmatrix} y_{o1} \\ y_{n1} \end{pmatrix} = \begin{pmatrix} y_{o2} \\ y_{n2} \end{pmatrix} + A_{on}\beta \rightarrow f\left(\begin{pmatrix} y_{o1} \\ y_{n1} \end{pmatrix},\ n\right) = f\left(\begin{pmatrix} y_{o2} \\ y_{n2} \end{pmatrix},\ n\right)$$

The different projections of $(Y_o,\ \dot{Y}_n)$ have expectations which are equal to zero, and therefore known.

The N joint probability laws :

$$L\left(s,\ P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix}\right)$$

are then also known. Now, the best estimator (and best selection criteria) $\tilde{s}$ is, analogously to the previous case,

$$\tilde{s} = E\left(s \mid P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix} = z_n,\ N = h\ (y_o,\ \varepsilon) = n\right) \text{ with } z_n = P_{on}\begin{pmatrix} y_o \\ y_n \end{pmatrix}$$

However, if no restrictions are placed on the class of functions h, it is not possible to obtain a simple result which is independent of h. One possible constraint that can be imposed is that the function h be invariant under translation, i.e. that :

$$h\ (Y_o,\ \varepsilon) = h\ (Y_o + A_o\ \beta,\ \varepsilon)$$

Let P_o be a projector on the orthogonal of the space spanned by A_o.

Using the same arguments as for f, the invariant functions h must be of the form $\ell\ [P_o\ (Y_o),\ \varepsilon]$. The significance of the proposed constraint can be seen as follows : consider those linear combinations of observations that eliminate the fixed effects, and then any function, linear or non linear, of these linear combinations. The result is a selection criterion, based on the first variable, which is invariant under translation. This then is a generalization of the form proposed by HENDERSON (1973), which is limited to linear functions of the linear combinations.

The estimator $\tilde{s}$ which minimizes Ω_i within the class of estimators invariant under translation of the fixed parameters (or which maximizes Ω within the same class) is then :

$$\tilde{s} = E\left(s \mid P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix} = z_n,\ N = \ell\left[P_o\ (Y_o),\ \varepsilon\right] = n\right)$$

As a function of $(Y_o,\ \dot{Y}_n)$, $P_o\ (Y_o)$ is invariant, and therefore a function of the maximum invariant $P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix}$. Thus one obtains :

$$\tilde{s} = E\left(s \mid P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix} = z_n\right) \text{ for } N = n$$

In the case of multinormality, every unbiased linear estimator of s is a linear function of :

$$Z_n = P_{on}\begin{pmatrix} Y_o \\ \dot{Y}_n \end{pmatrix}.$$

Inside these estimators, the conditional expectation minimizes the average square risk. So, $\tilde{s}$ is the BLUP.

V. Conclusions

Results presented in this paper may have interesting applications.

Let us, for instance, consider the case of individuals selected on a quantitative trait (such as growth characteristics of males recorded in performance-test stations) and, thereafter evaluated for a categorical trait (progeny test for prolificacy on daughter

groups). Proofs are given in this paper that evaluation and selection according to the second trait will not be biased if :

i) all information related to the 2 sets of records is used ;

ii) the first selection is made according to an invariant criterion (with respect to all environmental effects affecting performance test data) as the BLUP.

For all results supplied here, the joint probability law of the random variables defined in the experiment must be known. In the opposite case, when the variance-covariance matrix is replaced by an estimate, the properties of the corresponding estimators $\hat{S}$ remain unknown.

When expectations of the predictor random variables are unknown, consideration is restricted to estimators which are translation invariant for fixed effects. As a matter of fact, it corresponds to a generalization of Henderson's results. This restriction is not necessary, but in the general case, the derivation of optimal estimators is too complicated.

In addition, it was assumed throughout this study that a fixed number of sires was selected. If an optimal selection policy with fixed expectation of the number of selected sires was applied, it would be necessary to know the distribution law of the random variable $N = h(Y_0, \varepsilon)$ and therefore to exactly know how the selection at the first stage was carried out.

Received 23 september 1982.
Accepted 14 december 1982.

Acknowledgements

The author wishes to thank J.-L. FOULLEY and D. GIANOLA for helpful comments.

References

COCHRAN W.G., 1951. Improving by means of selection. *Proc. second Berkeley Symposium,* 449-470.

HAZEL L.N., 1943. The genetic basis for constructing selection indexes. *Genetics,* **28,** 476-490.

HENDERSON C.R., 1963. Selection index and expected genetic advance. *In : Statistical Genetics and Plant Breeding,* N.A.S.-N.C.R., 141.

HENDERSON C.R., 1975. Best Linear Unbiased Estimation and prediction under a selection model. *Biometrics,* **31,** 423-447.

POLLACK E.J., QUAAS R.L., 1981. Monte Carlo study of genetic evaluating using sequentially selected records. *J. anim. Sci.,* **52,** 257-264.

RAO C.R., 1963. Problems of selection involving programming techniques. *Proc. IBM scientific computing symposium ;* Statistics IBM Data Processing Division, White Plains, New York, 29-51.

SMITH H.F., 1936. A discriminant function for plant selection. *Ann. Eugen.,* **7,** 240-250.