Theor Appl Genet (1986) 72:822–825

Optimal properties of the conditional mean as a selection criterion

R. L. Fernando and D. Gianola

Department of Animal Sciences, University of Illinois, Urbana, IL 61801, USA

Received February 10, 1986; Accepted April 14, 1986
Communicated by L. D. Van Vleck

Summary. Rules for selection that maximize the expected merit of selected candidates are discussed. When the proportion selected is constant, selection based on conditional means of merit given the observations is optimum in the above sense, regardless of the distribution. This does not hold if the proportion selected is random. When the expected value of the observations is a linear function of a set of unknown parameters, selection can be based on a vector of "corrected" records, **w**. It is shown that under normality, the conditional mean of merit given **w** is the best linear unbiased predictor (BLUP), provided that the expected value of the merit function is the same in all candidates. A Bayesian argument is given to justify the use of BLUP as a selection rule when the expected merit differs from candidate to candidate.

Key words: Optimum rules for selection – BLUP (best linear unbiased predictor)

Introduction

Evaluation of candidates for selective breeding is required in animal and plant improvement programs. However, the variable to be improved, e.g., breeding value, cannot be observed or measured directly. Thus, the evaluation must be based on measurements which are observable and statistically related to breeding value.

Rules for ranking candidates for selection have been discussed by Cochran (1951); Henderson (1963, 1973, 1975, 1977, 1984); Bulmer (1980) and Portnoy (1982). Cochran (1951) considered the case where each candidate had a set of measurements (y_i) correlated with its unobservable breeding value (T_i). Assuming that (T_i, y_i) are identically and independently distributed, he showed that selection of candidates with the largest $E(T_i \mid y_i)$, maximizes the expected value of the T_i's in the selected individuals in the class of rules that select the same expected proportion of candidates. Bulmer (1980) showed that selection based on $E(T_i \mid y_i)$ also maximizes the mean of the selected candidates when a fixed number of individuals is to be kept. Although Bulmer (1980) assumed that each condidate had the same amount of information, this author stated that the above result would also hold with unequal information or when the proportion selected varies about a fixed expected value.

The vector $E(\mathbf{T} \mid \mathbf{y})$ often cannot be calculated because necessary parameters such as $E(\mathbf{y})$, where **y** is a vector including all records, are not known. For the case where $E(\mathbf{y})$ is unknown but the variance-covariance structure is known, Bulmer (1980) suggested to base selection on $E(\mathbf{T} \mid \mathbf{w})$, where

$$\mathbf{w} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}},$$

and $\mathbf{X}\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $E(\mathbf{y})$. Goffinet (1983) showed that selection upon $E(\mathbf{T} \mid \mathbf{w})$ maximizes the mean of a fixed number of selected individuals in the class of translation invariant functions. The objective of this paper is to further examine properties of conditional means as rules for selection.

Selection based on conditional means

Consider a random process generating an $m \times 1$ unobservable vector **T** and an $n \times 1$ observable vector **y**. In animal breeding, **T** may represent breeding values of animals for some trait, or functions of breeding values for several traits; the vector **y** typically contains phenotypic measurements taken on the animals to be evaluated or on their relatives.

Selection of a fixed number of candidates

The objective is to select $k < m$ members of T_1, T_2, ..., T_m with a rule based on **y** such that

$$E\left[\frac{1}{k}\sum_{i \in S(y)} T_i\right] \qquad (1)$$

is maximum, where S (**y**) is the set of k indices of the selected T_i's given **y**. Selecting the k candidates with the largest

$$\hat{T}_i = E(T_i \mid \mathbf{y})$$

meets the above requirement (Goffinet 1983). In order to show this, let

$$T_{S(y)} = \frac{1}{k} \sum_{i \in S(y)} T_i , \tag{2}$$

and write

$$E[T_{S(y)}] = E_y \{E[T_{S(y)} \mid \mathbf{y}]\} . \tag{3}$$

Now, for each **y**

$$E[T_{S(y)} \mid \mathbf{y}] = \frac{1}{k} \sum_{i \in S(y)} \hat{T}_i \tag{4}$$

is maximized by defining S (**y**) as the set of k indices of the largest $\hat{T}_i$'s. Therefore, (3) is maximum when for each realization of **y**, (4) is also maximum. This is accomplished when those candidates with the largest $E(T_i \mid \mathbf{y})$ are selected. Note that the above proof does not require independence of candidates, equality of information or linearity of the conditional mean. Thus, when the number of individuals to be selected is fixed, selection upon the conditional mean is optimum regardless of the form of the joint distribution of **T** and **y**.

Truncation selection

The scheme considered by Cochran (1951), is one of truncation selection. Here, the proportion of candidates selected is not constant across realizations of the random process, but a truncation point corresponding to a given expected proportion of selected candidates can be chosen. Bulmer (1980, 1982) asserted that selection based on $E(\mathbf{T} \mid \mathbf{y})$ would be optimal in the sense of (1) under a truncation selection scheme. However, this is not so as demonstrated by the following hypothetical example.

Let **y** be a discrete random vector taking one of two mutually exclusive and exhaustive states, with $P(\mathbf{y} = \mathbf{y}_1) = P(\mathbf{y} = \mathbf{y}_2) = 0.5$. Further, let the joint distribution of **T** (a 3×1 vector) and **y** be such that

$$E(\mathbf{T} \mid \mathbf{y}) = \begin{bmatrix} 25 \\ 10 \\ 10 \end{bmatrix}, \quad E(\mathbf{T} \mid \mathbf{y}) = \begin{bmatrix} 10 \\ 9.99 \\ 9.99 \end{bmatrix}.$$

Suppose now that truncation selection is practiced on $E(T_i \mid \mathbf{y})$ such that the proportion selected is on average 2/3. Then, the truncation point is 10 units and the expected value of the selected T's is

$$E(T_S) = 1/3 [E(T_1 \mid \mathbf{y}_1) + E(T_2 \mid \mathbf{y}_1) + E(T_3 \mid \mathbf{y}_1)] \cdot P(\mathbf{y} = \mathbf{y}_1) + E(T_1 \mid \mathbf{y}_2) \cdot P(\mathbf{y} = \mathbf{y}_2) = 12.5 .$$

Now, if individual 1 is selected when $\mathbf{y} = \mathbf{y}_1$, and individuals 1, 2 and 3 are selected when $\mathbf{y} = \mathbf{y}_2$, the proportion selected is on the average 2/3, and

$$E(T_s) = E(T_1 \mid \mathbf{y}_1) \cdot P(\mathbf{y} = \mathbf{y}_1) + 1/3 [E(T_1 \mid \mathbf{y}_2) + E(T_2 \mid \mathbf{y}_2) + E(T_3 \mid \mathbf{y}_2)] \cdot P(\mathbf{y} = \mathbf{y}_2) = 17.50 .$$

Clearly, truncation selection upon conditional means does not always maximize the expected merit of the selected individuals when the proportion selected is not constant. In fact, Cochran's (1951) result on the optimality of truncation selection based on conditional means depends on certain restrictions on distributional assumptions, i.e., the joint distribution of T_i and $\mathbf{y}_i$ must be the same in all candidates (Henderson 1973, 1977).

Bulmer (1980) stated that selection based on the conditional means would be more "efficient" in a truncation scheme than when a fixed number of candidates is selected. This would be so because more candidates could be selected, e.g., in "good" years than in "bad" years. The following example illustrates that this is not always true. As before, let **y** be a discrete vector with $P(\mathbf{y} = \mathbf{y}_1) = P(\mathbf{y} = \mathbf{y}_2) = 0.5$. Also, take

$$E(\mathbf{T} \mid \mathbf{y}_1) = \begin{bmatrix} 15 \\ 15 \\ 10 \end{bmatrix} \quad \text{and} \quad E(\mathbf{T} \mid \mathbf{y}_2) = \begin{bmatrix} 10 \\ 9.99 \\ 9 \end{bmatrix}.$$

If, on average, a proportion equal to 2/3 is to be selected, the truncation point must be 10. Hence, individuals 1, 2 and 3 would be selected when $\mathbf{y} = \mathbf{y}_1$, but only 1 would be kept when $\mathbf{y} = \mathbf{y}_2$. The expected value of the merit of the candidates selected under truncation selection is:

$$E(T_S) = \sum_{i=1}^{3} [E(T_i \mid \mathbf{y}_1)/3] \cdot P(\mathbf{y} = \mathbf{y}_1) + E(T_1 \mid \mathbf{y}_2) \cdot P(\mathbf{y} = \mathbf{y}_2) = 11.67 .$$

If the proportion selected is fixed at 2/3, 1 and 2 would always be selected and

$$E(T_S) = 1/2 [E(T_1 \mid \mathbf{y}_1) + E(T_2 \mid \mathbf{y}_1)] \cdot P(\mathbf{y} = \mathbf{y}_1) + 1/2 [E(T_1 \mid \mathbf{y}_2) + E(T_2 \mid \mathbf{y}_2)] \cdot P(\mathbf{y} = \mathbf{y}_2) = 12.50 .$$

The above shows that the mean of the selected candidates can be larger when a fixed proportion is selected than under truncation selection.

Selection with first moments unknown

Use of the conditional mean as a selection rule requires knowledge of the conditional distribution of **T** given **y**.

For example, under multivariate normality, $E(\mathbf{T})$, $E(\mathbf{y})$ and the variance-covariance matrix of $\mathbf{T}$ and $\mathbf{y}$ are required in order to calculate $E(\mathbf{T} \mid \mathbf{y})$. In animal breeding applications (Henderson 1973, 1975), there are situations in which the needed variances and covariances are known, but the first moments are unknown. Suppose

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}, \tag{5}$$

where, without loss of generality, $\mathbf{X}$ has full-column rank r. Following Bulmer (1982), calculate

$$\mathbf{w} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{6}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$, and $\mathbf{V} = \text{var}(\mathbf{y})$. Under multivariate normality, the distribution of $\mathbf{w}$ no longer depends on $\boldsymbol{\beta}$, so candidates could be ranked according to the magnitude of

$$E(\mathbf{T} \mid \mathbf{w}) \tag{7}$$

which, in the sense used in the preceding section, is optimal among rules based on $\mathbf{w}$. The vector of "corrected" observations in (6) can be written as

$$\mathbf{w} = \mathbf{P}\mathbf{y}. \tag{8}$$

Where $\mathbf{P} = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]$ is an $n \times n$ matrix of rank $n - r$. Goffinet (1983) stated that $\mathbf{P}$ retains "the most information", and that (7) is optimal in the class of translation invariant rules. However, this result is useful only when $E(\mathbf{T} \mid \mathbf{w})$ does not depend on $\boldsymbol{\beta}$.

It is shown below that the conditional mean of $\mathbf{T}$ given any other set of $n - r$ linearly independent translation invariant linear functions of the data is identical to (7). Consider a matrix $\mathbf{P}^*$ such that

$$\mathbf{P}^*\mathbf{X} = \mathbf{0} \tag{9a}$$

and

$$\text{rank}(\mathbf{P}^*) = n - r. \tag{9b}$$

This implies that the rows of $\mathbf{P}^*$ span the null space of $\mathbf{X}$. Therefore, the rows of any two matrices, say $\mathbf{P}_1^*$ and $\mathbf{P}_2^*$, satisfying (9a) and (9b), are linear combinations of the rows of the other. Thus, the elements of $\mathbf{w}_1^* = \mathbf{P}_1^*\mathbf{y}$ can be written as linear combinations of $\mathbf{w}_2^* = \mathbf{P}_2^*\mathbf{y}$, and vice versa. Therefore, $E(\mathbf{T} \mid \mathbf{P}^*\mathbf{y})$ is invariant to the choice of $\mathbf{P}^*$. Because $\mathbf{P}$ of (8) satisfies (9a) and (9b)

$$E(\mathbf{T} \mid \mathbf{w}) = E(\mathbf{T} \mid \mathbf{w}^*) \tag{10}$$

for $\mathbf{w}^* = \mathbf{P}^*\mathbf{y}$.

Multivariate normality

In many applications, the joint distribution of $\mathbf{T}$ and $\mathbf{w}$ of (6) can be approximated reasonably by a multivariate normal distribution. Clearly, $E(\mathbf{w}) = \mathbf{0}$, and let $E(\mathbf{T}) = \mathbf{0}$, $\text{Cov}(\mathbf{T}, \mathbf{w}') = \mathbf{C}'$ and $\text{Var}(\mathbf{w}) = \mathbf{W}$. With $\mathbf{W}$ nonsingular, the best rule would be to select candidates corresponding to the largest elements of

$$\hat{\mathbf{T}} = E(\mathbf{T} \mid \mathbf{w}) = \mathbf{C}'\mathbf{W}^{-1}\mathbf{w}.$$

However, when $\mathbf{w}$ is calculated as in (6) or (8), the matrix $\mathbf{W}$ is singular. In this situation, Gianola and Goffinet (1982) suggested that

$$\mathbf{T}^* = \mathbf{C}'\mathbf{W}^-\mathbf{w} \tag{11}$$

may be used to rank candidates, where $\mathbf{W}^-$ is a generalized inverse of $\mathbf{W}$. They showed that this leads to the best linear unbiased predictor (BLUP) of $\mathbf{T}$ (Henderson 1973). Further, when $\mathbf{T}$ and $\mathbf{w}$ have a singular multivariate normal distribution, then

$$\mathbf{T}^* = E(\mathbf{T} \mid \mathbf{w}). \tag{12}$$

This can be demonstrated by writing

$$\mathbf{T} = \mathbf{C}'\mathbf{W}^-\mathbf{w} + (\mathbf{T} - \mathbf{C}'\mathbf{W}^-\mathbf{w}) \tag{13}$$

so

$$\begin{aligned} E(\mathbf{T} \mid \mathbf{w}) &= E(\mathbf{C}'\mathbf{W}^-\mathbf{w} \mid \mathbf{w}) + E[(\mathbf{T} - \mathbf{C}'\mathbf{W}^-\mathbf{w}) \mid \mathbf{w}] \\ &= \mathbf{T}^* + E[(\mathbf{T} - \mathbf{C}'\mathbf{W}^-\mathbf{w}) \mid \mathbf{w}]. \end{aligned}$$

Therefore, for (12) to be true, it suffices to show that $E[(\mathbf{T} - \mathbf{C}'\mathbf{W}^-\mathbf{w}) \mid \mathbf{w}] = \mathbf{0}$. A sufficient condition for this to hold is

$$\text{Cov}[(\mathbf{T} - \mathbf{C}'\mathbf{W}^-\mathbf{w}), \mathbf{w}'] = \mathbf{C}' - \mathbf{C}'\mathbf{W}^-\mathbf{W} = \mathbf{0},$$

or, equivalently, because $\mathbf{W}$ is symmetric, $\mathbf{C} = \mathbf{W}(\mathbf{W}^-)'\mathbf{C}$. Now, $\mathbf{C}$ can be written as $\mathbf{C} = \mathbf{W}\mathbf{L}$ (Rao 1973; Fernando 1984), so,

$$\mathbf{W}(\mathbf{W}^-)'\mathbf{C} = \mathbf{W}(\mathbf{W}^-)'\mathbf{W}\mathbf{L} = \mathbf{C}$$

because $\mathbf{W}(\mathbf{W}^-)'\mathbf{W} = \mathbf{W}$. This result is invariant to the generalized inverse used. Therefore, (12) is true.

Because $E(\mathbf{T} \mid \mathbf{w}) = E(\mathbf{T} \mid \mathbf{w}^*)$

$$E(\mathbf{T} \mid \mathbf{w}^*) = \text{BLUP}(\mathbf{T}). \tag{15}$$

For example, if

$$\mathbf{w}^* = [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{y},$$

which implies that $\mathbf{y}$ is "corrected for the fixed effects by ordinary least squares, then

$$E\{\mathbf{T} \mid [\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}']\mathbf{y}\} = \text{BLUP}(\mathbf{T}).$$

Discussion

Selection based on $E(\mathbf{T} \mid \mathbf{y})$ has the appealing property of maximizing the mean of the selected candidates when a fixed number of individuals is to be selected. However, ordering candidates with $E(\mathbf{T} \mid \mathbf{y})$ does not, in general, maximize the probability of correctly ordering elements of $\mathbf{T}$ (Portnoy 1982). For example,

consider the discrete unobservable random variables T_1 and T_2 with Prob $(T_1 = -1, T_2 = 0) = 0.9$, and Prob $(T_1 = 100, T_2 = 1) = 0.1$. If we want to select one of the two elements such that the expected value of the selected one is maximum, we would select T_1 because $E(T_1) = 9.1$ and $E(T_2) = 0.1$. However,

$$\text{Prob}(T_1 < T_2) = 0.9 > \text{Prob}(T_2 < T_1) = 0.1 .$$

The linear transformation $\mathbf{w}$ in (8) also arises in restricted maximum likelihood estimation (REML) of variance components (Patterson and Thompson 1971; Searle 1979), where inferences on these parameters are based on the likelihood of $\mathbf{w}$ under normality. Harville (1977) pointed out that the likelihood functions of linear transformations in the class of $\mathbf{w}$ differ only by an additive constant. Thus, inferences on variances are invariant to the linear transformation used. Similarly, it can be shown, under normality, that the joint distributions of $\mathbf{T}$ and of linear transformations $\mathbf{w}$ differ only by a constant. Hence, inferences based on the joint (or conditional) distribution of $\mathbf{T}$ and $\mathbf{w}$ are invariant to the particular linear transformation used.

Suppose $\mathbf{T} = \mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'\mathbf{u}$ is a linear merit function. Under multivariate normality, and using a Bayesian argument with a flat prior for $\boldsymbol{\beta}$,

$$E(\mathbf{T} \mid \mathbf{y}) = \text{BLUP}(\mathbf{T}) = \mathbf{K}'\boldsymbol{\beta} + \mathbf{M}'E(\mathbf{u} \mid \mathbf{w})$$

where $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$ (Dempfle 1977; Gianola and Fernando 1986). This justifies the use of BLUP even when the elements of $E(\mathbf{T})$ are unknown.

Another complication arises when variances and covariances are unknown. Gianola and Fernando (1986) suggested that given certain conditions, e.g., a "peaked" likelihood for the unknown variances and covariances, using REML estimates in place of the true values would give a reasonable approximation to $E(\mathbf{T} \mid \mathbf{y})$ under normality. In general, however, the problem of finding an optimum rule for selection, when $E(\mathbf{T})$, $\mathbf{C}$ and $\mathbf{V}$ are unknown remains to be solved.

References

Bulmer MG (1980) The mathematical theory of quantitative genetics. Clarendon Press, Oxford

Bulmer MG (1982) Sire evaluation with best linear unbiased predictors. Biometrics 38:1085–1088

Cochran WG (1951) Improvement by meams of selection. In: Proc 2nd Berkeley Symp Math Stat Prob, pp 449–470

Dempfle L (1977) Relation entre BLUP (Best linear unbiased prediction) et estimateurs bayesiens. Ann Genet Sel Anim 9:27–32

Fernando RL (1984) Selection and assortative mating. PhD Thesis, University of Illinois

Gianola D, Fernando RL (1986) Bayesian methods in animal breeding theory. J Anim Sci 63:217–244

Gianola D, Goffinet B (1982) Sire evaluation with best linear unbiased predictors. Biometrics 38:1085–1088

Goffinet B (1983) Selection on selected records. Genet Sel Evol 15:91–98

Harville DA (1977) Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338

Henderson CR (1963) Selection index and expected genetic advance. In: Hanson WD, Robinson HF (eds) Statistical genetics and plant breeding. NAS-NRC 982, Washington DC, pp 141–163

Henderson CR (1973) Sire evaluation and genetic trends. In: Proc Anim Breed Genet Symp in Honor of Dr Jay L Lush. Am Soc Anim Sci and Am Dairy Sci Assoc, Champaign Ill, pp 10–41

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423–449

Henderson CR (1977) Prediction of future records. In: Pollak E, Kempthorne O, Bailey TB Jr (eds) Proc Int Conf Quant Genet. Iowa State University Press, Ames, Iowa, pp 615–638

Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Guelph, Ontario, Canada

Patterson HD, Thompson R (1971) Recovery of interblock information when block sizes are unequal. Biometrika 58: 545–554

Portnoy S (1982) Maximizing the probability of correctly ordering random variables using linear predictors. J Multivar Anal 12:256–269

Rao CR (1973) Linear statistical inference and its applications. Wiley and Sons, New York

Searle SR (1979) Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology. Cornell University Ithaca, New York