A NOTE ON THE GENETIC SELECTION INDEX

Rama C. Nair

BU-539-M

November 1974

Abstract

A new genetic selection index is proposed, when $\underline{\beta}$ is unknown, in the model $\underline{Y} = X\underline{\beta} + Z\underline{u} + \underline{e}$, $E\begin{pmatrix}\underline{e}\\ \underline{u}\end{pmatrix} = 0$, $V\begin{pmatrix}\underline{e}\\ \underline{u}\end{pmatrix} = \begin{pmatrix}E & 0\\ 0 & G\end{pmatrix}$, and its properties examined.

A NOTE ON THE GENETIC SELECTION INDEX

Rama C. Nair

BU-539-M November 1974

1. Introduction

Consider the genetic model (Henderson [1963])

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + Z\underset{\sim}{u} + \underset{\sim}{e} \;; \quad u \sim N(\underset{\sim}{0}, G), \;\; e \sim N(\underset{\sim}{0}, E)$$
$$\mathrm{Cov}(\underset{\sim}{u}, \underset{\sim}{e}) = \underset{\sim}{0} \;\; .$$

where $\underset{\sim}{Y}_{n\times 1}$ is an observable random vector, $\underset{\sim}{\beta}_{p\times 1}$ is a fixed vector of parameters and $\underset{\sim}{u}_{r\times 1}$ is a random vector of genetic parameters. $X_{n\times p}$ and $Z_{n\times r}$ are known coefficient matrices.

In the selection index problem we observe the vector $\underset{\sim}{Y}$, which are (usually) the observations on phenotypic values (corresponding to various traits) on a number of candidates for selection; $\underset{\sim}{u}$ is the vector of genotypic values (usually a.g.v.) and as such nonobservable; $\underset{\sim}{\beta}$ is a vector of fixed parameters. The problem is to obtain an index based on the observations $\underset{\sim}{Y}$ to rank the candidates in order of predicted genetic values, the objective being to select the best candidates to use in breeding programs to increase the expected phenotypic value. One way to do this is to obtain an index that has maximum correlation with the genetic values to be predicted.

In the classical selection index problem, a selection index $I(\underset{\sim}{Y})$ is chosen so as to minimise the prediction error. The best predictor in the sense of minimising mean squared error of prediction is given by $E(\underset{\sim}{u}|\underset{\sim}{y})$ (Rao [1965]).

Solomon [1971] considers a Bayesian approach to the problem and obtains that the Bayes procedure, assuming quadratic loss and normal prior (i.e.

$\binom{\beta}{u} \sim N(\mu, V)$), is the same as the classical selection index, assuming normality for u. He also asserts that the Bayes procedure is linear in the observations, without restricting selection indices to be linear, whereas classical selection indices are not. This assertion is not true. The linearity of the Bayes procedure is inherent in the assumption of a multivariate normal prior for $\binom{\beta}{u}$. Even in the classical case, if we assume that $u \sim N(\lambda, G)$, <u>we do get</u> the best predictor to be linear in the observations (<u>without restricting attention to linear indices</u>),

$$\text{i.e.} \qquad \text{Best Predictor} = E(u|Y) = GZ'(ZGZ' + E)^{-1}(Y - X\beta) \quad .$$

In contrast, in the Bayesian approach, if we assume a non-normal prior for β and a normal prior for u, we do not obtain a linear index. Indeed, it is quite tedious to obtain (and sometimes impossible to evaluate analytically) a Bayes procedure $I(Y)$, when we assume non-normal priors for β. By assuming normality, we are restricting the index, and the classical selection index is much easier to compute.

Moreover, when we know very little about β (or completely unknown), it is not very appealing to put a normal prior on β. Of course, one might argue that a normal prior with large variance might be appropriate. Or Jeffrey's non-informative prior could be used, with the result of complicating the computations. (The latter should not be a problem in this computer age!) However, it is quite disturbing (at least to this author) to use a continuous prior when we have point priors. Accordingly, the substitution of β by its m.l.e. $\hat{\beta}$ would be more appropriate. A new procedure is given in the next section, in which we first estimate β and then construct a predictor for u based on $Y - X\hat{\beta}$.

2. A new procedure

Suppose, in the model $\underset{\sim}{Y} = X\underset{\sim}{\beta} + Z\underset{\sim}{u} + \underset{\sim}{e}$; $E(\underset{\sim}{e}) = E(\underset{\sim}{u}) = 0$; $V\begin{pmatrix}\underset{\sim}{e}\\ \underset{\sim}{u}\end{pmatrix} = \begin{pmatrix}E & 0\\ 0 & G\end{pmatrix}$; β is unknown. Then Henderson replaces β by $\hat{\beta}$ (the m.l.e. of β) in the selection index, $I(\underset{\sim}{Y}) = GZ'(ZGZ' + E)^{-1}(\underset{\sim}{Y} - X\underset{\sim}{\beta})$ and Solomon replaces β by its Bayes estimator, assuming β has a normal prior. An objection to the Henderson procedure is that it is ad hoc. We shall now construct a procedure in which we shall replace $\underset{\sim}{\beta}$ by its GLS estimate. The GLS estimate for β is

$$\underset{\sim}{\hat{\beta}} = (X'A^{-1}X)^{-1}(X'A^{-1}\underset{\sim}{Y})$$

where $A = ZGZ' + E$.

Consider $\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}$ and we shall construct a selection index based on $\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}$ which will minimise the error of $\hat{I} = \underset{\sim}{b}'(\underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}})$ as a predictor of $\underset{\sim}{u}$. Let

$$\underset{\sim}{v} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}} = (\underset{\sim}{Y} - X(X'A^{-1}X)^{-1}X'A^{-1}\underset{\sim}{Y}) .$$

Then

$$\begin{pmatrix}\underset{\sim}{u}\\ \underset{\sim}{v}\end{pmatrix} \sim N\left[\begin{pmatrix}\underset{\sim}{0}\\ \underset{\sim}{0}\end{pmatrix}, \begin{pmatrix}G & GZ'P'\\ P'ZG & PAP'\end{pmatrix}\right]$$

where

$$P = I - X(X'A^{-1}X)^{-1}X'A^{-1} .$$

Notice that

$$P = PP$$

$$\text{and } r(P) = \text{tr}P = n - s ,$$

where $s = r(X)$

$\therefore$ P is singular and so is PAP' .

Now, the 'best' predictor $\hat{u}$ of $\underset{\sim}{u}$ based on $\underset{\sim}{v} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}$, in the sense of minimum mean squared error (equivalent to a loss function $(\underset{\sim}{\hat{u}} - \underset{\sim}{u})'K(\underset{\sim}{\hat{u}} - \underset{\sim}{u})$ for any positive definite matrix K) is given by

$$\begin{aligned}\hat{\underline{u}} = I(\underline{Y}) &= E(\underline{u}|\underline{v}) \\ &= GZ'P'\{P(ZGZ' + R)P'\}^{-}\underline{v} \\ &= GZ'P'(PAP')^{-}P\underline{Y}, \text{ since } \underline{v} = P\underline{Y} \quad .\end{aligned}$$

Notice that A is of full rank,

$$r(PAP') = r(P) = n - s \quad .$$

$\therefore$ $P'(PAP')^{-}P$ is unique for <u>any</u> generalised inverse $(PAP')^{-}$ of (PAP'). (Rao & Mitra [1971]). Thus the selection index $\hat{\underline{u}}$ is unique, no matter what generalised inverse we choose.

3. <u>Some properties of $\hat{\underline{u}}$</u>

1) $E(\hat{\underline{u}}) = GZ'P'(PAP')^{-}E(P\underline{Y}) = \underline{0} = E(\underline{u})$

2) $V(\hat{\underline{u}}) = GZ'P'(PAP')^{-}PAP'(PAP')^{-}PZG = GZ'P'(PAP')^{-}PZG$
since $P'(PAP')^{-}PAP' = P'$ (Rao & Mitra [1971].)

3) $Cov(\underline{u},\hat{\underline{u}}) = Cov(\hat{\underline{u}},\underline{u}) = GZ'P'(PAP')^{-}PZG = V(\hat{\underline{u}})$

4) $V(\hat{\underline{u}} - \underline{u}) = V(\underline{u}) - V(\hat{\underline{u}})$ by (3)
$= G - GZ'P'(PAP')^{-}PZG$

5) $Cov(\hat{\underline{u}},K\hat{\underline{\beta}}) = \underline{0}$ for any estimable function $K\underline{\beta}$
$K\underline{\beta}$ estimable $\Rightarrow K = LX$
$\therefore Cov(\hat{\underline{u}},K\hat{\underline{\beta}}) = GZ'P'(PAP')^{-}P\, Cov(\underline{Y},X\hat{\underline{\beta}})L$
$= \underline{0}$

6) The correlation of $\hat{u}_i$ with u_i is maximised.

The proof is easy, considering the fact that $\hat{u}_i = E(u_i|f(\underline{y}))$.

In conclusion we note that the same properties hold when A is not of full rank, by replacing A^{-1} by A^{-}, a generalized inverse of A . Also, the procedure can be extended easily to the case $V\begin{pmatrix} \underline{e} \\ \underline{u} \end{pmatrix} = \begin{pmatrix} E & C \\ C & G \end{pmatrix}$, $C \neq 0$.

References

1. Henderson, C. R. [1963]. Selection index and expected genetic advance. In Stat. Genetics and Plant Breeding, NAS NRC-982.

2. Rao, C. R. [1965]. Linear Statistical Inference and its Applications. John Wiley & Sons.

3. Rao, C. R. and Mitra, S. K. [1971]. Generalized Inverses of Matrices. John Wiley & Sons.

4. Solomon, D. L. [1971]. A Bayesian interpretation of the genetic selection index, BU-362-M in the Biometrics Unit Mimeo Series, Cornell University.

5. Solomon, D. L. [1974]. Personal communication.