A BAYESIAN INTERPRETATION OF THE GENETIC SELECTION INDEX

by

Daniel L. Solomon

BU-362-M April, 1971

ABSTRACT

The general formulation proposed by Henderson (1963), of the genetic selection index model is shown to have a Bayesian interpretation in which the distribution associated with genetic values is treated as a prior distribution. A Bayes rule is constructed for the index in the case in which the expected values of records are unknown.

Biometrics Unit Mimeo Series, Cornell University, Ithaca, N.Y. 14850

A BAYESIAN INTERPRETATION OF THE GENETIC SELECTION INDEX

by

Daniel L. Solomon

BU-362-M April, 1971

In the usual formulation of the genetic selection index problem (see for example Comstock (1948)) one supposes that for each candidate for selection, observations $Y_1, Y_2, \ldots, Y_n$ are available on phenotypes corresponding to N traits of interest. It is further supposed that each phenotype is related to an unobservable genotype through the linear model $Y_i = \mu_i + u_i + e_i$ where μ_i is a constant, u_i is the genetic value corresponding to the genotype for the i^{th} trait and e_i is environmental "noise." That is $\underline{Y} = \underline{\mu} + \underline{u} + \underline{e}$ where each component is an N dimensional column vector. Now if $\underline{v} = (v_1, v_2, \ldots, v_n)'$ is an N-vector of constants representing the relative economic values of the N traits, then one wishes to construct an index, I, (a function of $\underline{Y}$) to use in selection for the "aggregate genetic value" $T = \underline{v}'\underline{u}$.

The usual approach is to assume that $\underline{u}$ and $\underline{e}$ are independent N-variate normal random variables, say $\underline{u} \sim N(\underline{0}, \underline{G})$ independent of $\underline{e} \sim N(\underline{0}, \underline{E})$ where $\underline{G}$ and $\underline{E}$ are positive definite and symmetric matrices of order N . Then one requires the index I to be a scalar valued linear function, $I = \underline{b}'(\underline{Y} - \underline{\mu})$ and determines the vector $\underline{b}$ to maximize the correlation between I and T . The result is $I = \underline{v}'\underline{G}\underline{P}^{-1}(\underline{Y} - \underline{\mu})$, where $\underline{P} = \underline{G} + \underline{E}$. This index has a number of desirable properties (see Henderson (1963)), and it will be here demonstrated that it also has a Bayesian interpretation.

In a Bayesian context, with distribution assumptions as above, the distribution for $\underline{u} \sim N(\underline{0}, \underline{G})$ is viewed as the prior distribution, and then the likelihood (distribution of $\underline{Y}$ given $\underline{u}$) is $N(\underline{\mu} + \underline{u}, \underline{E})$. If we seek a Bayes rule, I, for $T = \underline{v}'\underline{u}$ and are operating under quadratic loss, $(I - T)^2$, then the Bayes rule is the mean of the posterior distribution of T given $\underline{Y}$. i.e.,

$$I = E(T|\underline{Y}) = \underline{v}'E(\underline{u}|\underline{Y}) = \underline{v}'\underline{G}\underline{P}^{-1}(\underline{Y} - \underline{\mu})$$

as before. Note that for the Bayesian model, it is not necessary to assume that the index is linear.

A more general formulation of the problem (see Henderson (1963)) is to suppose that

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{Z}\underline{u} + \underline{e}$$

where $\underline{Y}$ is an (observable) N-variate random variable, $\underline{X}$ is a known $N \times p$ matrix of rank $p \leq N$, $\underline{\beta}$ is a p-vector of parameters, $\underline{Z}$ is a known $N \times r$ matrix of rank r, $\underline{u}$ is an (unobservable) r-variate normal random variable with mean vector $\underline{0}$ and positive definite covariance matrix $\underline{G}$, $\underline{e}$ is an N-variate normal random variable with mean vector $\underline{0}$ and positive definite covariance matrix $\underline{E}$, and $\underline{u}$ and $\underline{e}$ are independent.

Thus, here, the N phenotypes depend on N linear functions of r genetic values. This model reduces to that described above if we let $r = N$, $\underline{X}\underline{\beta} = \underline{\mu}$ and $\underline{Z} = \underline{I}_N$, the identity matrix of order N . Again, if we seek an index to select for $T = \underline{v}'\underline{u}$ (where now $\underline{v}$ is an r-vector) the usual result has a Bayesian interpretation. Thus the prior distribution for $\underline{u}$ is r-variate $N(\underline{0}, \underline{G})$ and the likelihood of $\underline{Y}$ given $\underline{u}$ is N-variate $N(\underline{X}\underline{\beta} + \underline{Z}\underline{u}, \underline{E})$ so that with quadratic loss, the Bayes rule is

$$I = E(T|\underline{Y}) = \underline{v}'E(\underline{u}|\underline{Y}) = \underline{v}'\underline{G}\underline{Z}'\underline{A}^{-1}(\underline{Y} - \underline{X}\underline{\beta}),$$

where $\underline{A} = \underline{Z}\underline{G}\underline{Z}' + \underline{E}$ is the marginal covariance matrix of $\underline{Y}$. Again linearity obtains for the Bayes rule but is assumed in the usual approach. Note that the economic weighting $T = \underline{v}'\underline{u}$ is not the only way to make use of the genetic values. We may calculate a Bayes rule, $\hat{\underline{u}}$, for the whole vector, $\underline{u}$. If the loss function is $(\hat{\underline{u}} - \underline{u})'\underline{K}(\hat{\underline{u}} - \underline{u})$ for any positive definite matrix $\underline{K}$ of order r, then the result is simply

$$\hat{u} = E(u|Y) = GZ'A^{-1}(Y - X\beta), \tag{1}$$

and does not depend on K .

If all candidates for selection provide the same information, i.e., values of the same random variable Y, above, and selection is based on ranking by the index, then this ranking does not depend on the value of β . That is, the difference in values of the index when applied to two individuals does not depend on β . Thus, in this case, β need not be known. If however, β must be estimated, Henderson (1963) replaces β in the index, by its maximum likelihood estimator $\hat{\beta} = (X'A^{-1}X)^{-1}X'A^{-1}Y$. Notice that this can lead to difficulties for some pathological models. That is, with β replaced by $\hat{\beta}$, the index becomes

$$GZ'A^{-1}(Y - X\hat{\beta}) = GZ'A^{-1}\left[I_N - X(X'A^{-1}X)^{-1}X'A^{-1}\right]Y ,$$

and if the model happens to have $Z' = BX'$ for some $r \times p$ matrix B, then the index is zero for all Y .

Now, if β is unknown, then to be consistent with the Bayesian approach, one is required to have a prior distribution for β . Thus, with $\theta' \equiv (\beta' \; u')$ and $W \equiv (X \; Z)$, we have $Y = W\theta + e$, where as before, $e \sim N(0, E)$, independent of θ . The prior distribution for θ is taken to be (p+r)-variate normal with mean vector $\theta_o = (\beta_o' \; 0')'$ and positive definite covariance matrix

$$G^* = \begin{bmatrix} G_\beta & G_{\beta u} \\ G_{\beta u}' & G_u \end{bmatrix} .$$

The likelihood is then N-variate normal, $N(\underline{W}\underline{\theta}, \underline{E})$, and the Bayes rule for $\underline{\theta}$ with respect to the quadratic loss $(\tilde{\underline{\theta}} - \underline{\theta})'\underline{K}(\tilde{\underline{\theta}} - \theta)$ (where $\underline{K}$ is any positive definite matrix of order p+r) is the posterior mean of $\underline{\theta}$. That is, the Bayes rule is

$$\tilde{\underline{\theta}}(\underline{Y}) = E(\underline{\theta}|\underline{Y}) = \underline{\theta}_o + \underline{G}^*\underline{W}'(\underline{W}\underline{G}^*\underline{W}' + \underline{E})^{-1}(\underline{Y} - \underline{W}\underline{\theta}_o) .$$

(Note that the (p+r) × N matrix of covariances between components of $\underline{\theta}$ and $\underline{Y}$ is $\underline{G}^*\underline{W}'$.)

Our interest is only in the last r rows of $\tilde{\underline{\theta}}(\underline{Y})$, namely

$$\tilde{\underline{u}}(\underline{Y}) = (\underline{G}'_{\beta u}\underline{X}' + \underline{G}_u\underline{Z}')(\underline{W}\underline{G}^*\underline{W}' + \underline{E})^{-1}(\underline{Y} - \underline{X}\underline{\beta}_o) .$$

If $(\underline{Z}\underline{G}'_{\beta u}\underline{X}' + \underline{A})$ is non-singular, where $\underline{A} = \underline{Z}\underline{G}_u\underline{Z}' + \underline{E}$ as above, this can be written

$$\tilde{\underline{u}}(\underline{Y}) = (\underline{G}'_{\beta u}\underline{X}' + \underline{G}_u\underline{Z}')(\underline{Z}\underline{G}'_{\beta u}\underline{X}' + \underline{A})^{-1}(\underline{Y} - \underline{X}\tilde{\underline{\beta}})$$

where $\tilde{\underline{\beta}}$ is the Bayes rule for $\underline{\beta}$, i.e., the first p rows of $\tilde{\underline{\theta}}(\underline{Y})$.

If $\underline{\beta}$ and $\underline{u}$ are a-priori independent so that $\underline{G}_{\beta u} = \underline{0}$, (and thus $(\underline{Z}\underline{G}'_{\beta u} + \underline{A}) = \underline{A}$ is non-singular), then the Bayes rule for $\underline{u}$ reduces to

$$\tilde{\underline{u}}_I(\underline{Y}) = \underline{G}_u\underline{Z}'\underline{A}^{-1}(\underline{Y} - \underline{X}\tilde{\underline{\beta}}), \qquad (2)$$

and

$$\underline{X}\tilde{\underline{\beta}} = \left[\underline{X}\underline{G}_\beta\underline{X}'(\underline{X}\underline{G}_\beta\underline{X}' + \underline{A})^{-1}\right]\underline{Y} + \left[\underline{I}_N - \underline{X}\underline{G}_\beta\underline{X}'(\underline{X}\underline{G}_\beta\underline{X}' + \underline{A})^{-1}\right](\underline{X}\underline{\beta}_o) ,$$

is a matrix weighted average of the observation, $\underline{Y}$ and the a-priori mean $X\underline{\beta}_0$. Thus if under the prior distribution, $\underline{\beta}$ and $\underline{u}$ are independent, the Bayes rule (2) for $\underline{u}$ is the usual index (1) with $\underline{\beta}$ replaced by its Bayes rule $\tilde{\underline{\beta}}$ rather than its maximum likelihood estimator $\hat{\underline{\beta}}$.

REFERENCES

Comstock, R. E. [1948]. Statistics in animal breeding research. Proc. Auburn Conf. on Stat. Applied to Research. Stat. Lab., Auburn, 74-76.

Henderson, C. R. [1963]. Selection index and expected genetic advance. In Statistical Genetics and Plant Breeding. NAS-NRC 982, 141-63.