Reprinted From
STATISTICAL GENETICS AND PLANT BREEDING
NAS–NRC 982 1963

Selection Index and Expected Genetic Advance

C. R. HENDERSON
Department of Animal Husbandry, Cornell University, Ithaca, New York

THE selection index employed in plant and animal breeding refers usually to a linear combination of observations that is used to compute, for each individual available for choice, a criterion for selection. We shall call the mathematical description of this linear function the selection index, I, and a numerical value actually computed by an index from the observations on a particular individual, the selection criterion. For example, suppose that the records available on each of several dairy sires are y_1 = mean of 10 progeny, y_2 = dam's record. Then the index might be something like

$$I = .77(y_1 - \mu_1) + .08(y_2 - \mu_2).$$

If, for a particular sire, $y_1 = 450$, $y_2 = 500$, $\mu_1 = 460$, $\mu_2 = 480$, the selection criterion for this sire would be

$$.77(450 - 460) + .08(500 - 480) = -6.1$$

The selection index can be used for several different purposes, e.g.,

1. Selection on a single trait using information on the individual and certain of its relatives (5).
2. Selection on two or more traits using records made by the individual (3).
3. Selection on two or more traits using records on the individual and its relatives.
4. Selection of line-crosses using data in addition to that on the specific cross (4).

The first application of the selection index to plant breeding was by Smith (7), and the first to animals by Hazel (3). An excellent brief description of the method was given by Comstock (2). Cochran (1) presented many of the mathematical and statistical problems encountered in constructing indexes.

The foregoing publications and all others on the subject, so far as I am aware, have justified the procedure only for the case in which the information available on each candidate for selection is the same. More precisely, the N records and the underlying genetic value available on each individual are a random sample from some known (N + 1)-variate population. In actual practice, at least in animal breeding, this is seldom true. Rather, choices must be made among animals with different amounts of information. It does turn out, as will be shown in this paper, that the selection index procedure is in fact valid for the latter case.

SELECTION INDEX FOR THE EQUAL INFORMATION CASE

N records, say $y_1, \ldots, y_N$, are available on each candidate for selection. The breeding value of this individual is denoted by T. We shall deliberately not define breeding value at this time, but will do so later in the paper. $y_1, \ldots, y_N$, T are assumed to have an $(N + 1)$-variate normal distribution with variance-covariance matrix

$$\begin{bmatrix} C_{11} & C_{12} & \ldots & C_{1N} & t_1 \\ C_{12} & C_{22} & \ldots & C_{2N} & t_2 \\ \vdots & & & & \vdots \\ C_{1N} & C_{2N} & \ldots & C_{NN} & t_N \\ t_1 & t_2 & \ldots & t_N & g \end{bmatrix}$$

or in matrix notation,

$$\begin{bmatrix} & \cdot & \\ C & \cdot & t \\ \cdot & \cdot & \cdot \\ t' & \cdot & g \\ & \cdot & \end{bmatrix},$$ where C is an $N \times N$, non-singular matrix, t is an $N \times 1$ vector, and g is a scalar. The y's have means $\mu_1, \ldots, \mu_N$.

Construction of the Index

An index is wanted of the following form

$$I = b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N).$$

Of all such linear functions which one is "best" in some sense? To answer this question we must define what we mean by best. A logical criterion would be that one which in the long run maximizes genetic progress, see, for example, Lush (6). Now the expected value of any particular T selected on the basis of such an index is

$$E(T|I) = \mu_T + b_{TI}(I - \mu_I)$$

$$= \mu_T + \frac{\sigma_{TI}}{\sigma^2_I}(I - \mu_I).$$

This is the well known formula for the regression of one variable on a second variable in the bivariate normal distribution. This is not true for other distributions, but may be a suitable approximation. Then the mean of the T's in a selected group is

$$E(\bar{T}|\bar{I}) = \mu_T + \frac{\sigma_{TI}}{\sigma^2_I}(\bar{I} - \mu_I).$$

If selection is strictly according to the index, $\bar{I} - \mu_I$ is equal to $\frac{z}{q}\sigma_I$, where z is the

ordinate of the unit normal distribution at the point of truncation, and q is the fraction of indexed individuals that is selected. Thus, the expected genetic progress in one cycle of selection on an index is

$$\frac{\sigma_{TI}}{\sigma^2_I}\frac{z}{q} - \sigma_I, \text{ which can be re-written as}$$

$$r_{TI}\frac{z}{q} - \sigma_T.$$

Since for any given population and intensity of selection $\frac{z}{q}\sigma_T$ is constant, the *b's* of the index should be chosen so as to maximize r_{TI}. Differentiating $\log r_{TI} = \log \sigma_{TI} - \frac{1}{2}\log \sigma^2_T - \frac{1}{2}\log \sigma^2_I$ with respect to $b_1, \ldots, b_N$, equating the partial derivatives to zero, and noting that

$$\sigma_{TI} = b_1\sigma_{y_1T} + \ldots + b_N\sigma_{yNT}$$
$$\text{and } \sigma^2_I = b^2_1\sigma^2_{y1} + 2b_1b_2\sigma_{y1y2} + \ldots + b^2_N\sigma^2_{yN}$$

the following equations in the *b's* are obtained:

$$b_1\sigma^2_{y1} + b_2\sigma_{y1y2} + \ldots + b_N\sigma_{y1yN} = \sigma_{y1T}\frac{\sigma^2_I}{\sigma_{TI}}$$

$$b_1\sigma_{y1y2} + b_2\sigma^2_{y2} + \ldots + b_N\sigma_{y2yN} = \sigma_{y2T}\frac{\sigma^2_I}{\sigma_{TI}}$$

etc.

Since the magnitude of σ_I^2/σ_{TI} does not affect the proportionality of the *b's*, it has no effect on r_{TI} and can be chosen arbitrarily. For convenience let us choose the value, 1. Thus we have the above equations with σ_I^2/σ_{TI} deleted. In matrix notation the equations now are

$$Cb = t, \qquad (1)$$

where *b* is the $N \times 1$ vector, $b_1, b_2, \ldots, b_N$, C is the variance-covariance matrix of the *y's*, and *t* is the vector of σ_{yT}'s. Note that these index equations are exactly like "normal" equations of multiple regression except that population variances and covariances appear in place of sample sums of squares and cross-products.

Expected Genetic Progress

With the *b's* determined, the expected genetic progress in one cycle of selection by truncation of a set of selection criteria can be computed from

$$r_{TI}\sigma_T\frac{z}{q}.$$

r_{TI} can be calculated conveniently by noting that

$$r^2_{TI} = \frac{(\sigma_{TI})^2}{\sigma^2_I \, \sigma^2_T} = \frac{\sigma_{TI}}{\sigma^2_T} \left(\text{since } \frac{\sigma_{TI}}{\sigma^2_I} = 1\right)$$

$$= \frac{b_1\sigma_{y_1T} + \ldots + b_N\sigma_{y_NT}}{\sigma^2_T}.$$

Also, we note that the expected value of a particular T, given the selection criterion, I_o, is

$$E(T|I_0) = \mu_T + \frac{\sigma_{TI}}{\sigma^2_I}(I_0 - \mu_I)$$

$$= \mu_T + I_0, \text{ since } \sigma_{TI}/\sigma^2_I = 1 \text{ and } \mu_I = 0.$$

Other Properties of the Selection Criterion

The selection criterion computed by the selection index has other properties of interest in addition to maximization of r_{TI} and of expected genetic progress.

1. $E(I - T)^2$ is minimum among all linear functions of the general form of the selection index. That is, the average value of the squared deviations of criteria from true breeding values is minimum. This is easy to prove by minimizing, for variations in b,

$$E(I - T)^2 = E[b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N) - T]^2$$

$$= b_1{}^2\sigma^2_{y_1} + 2b_1b_2\sigma_{y_1y_2} + \ldots + b^2_N\sigma^2_{y_N} - b_1\sigma_{y_1T} - \ldots$$

$$- b_N\sigma_{y_NT} + \sigma^2_T.$$

When this expression is differentiated with respect to b's and the partial derivatives are equated to zero, the equations of (1) are obtained. Note that this property does not require the multivariate normal distribution, nor does the property maximization of r_{TI}. If the value of $E(I - T)^2$ is wanted for a particular index, it can be computed either by

$$\sigma^2_T - \sigma_{TI} = \sigma^2_T - (b_1\sigma_{y_1T} + \ldots + b_N\sigma_{y_NT}) \text{ or by}$$

$$\sigma^2_T(1 - r^2_{TI}).$$

A proof of these computing formulas is,

$$E(I - T)^2 = \sigma^2_I - 2\sigma_{TI} + \sigma^2_T$$

$$= \sigma^2_T - \sigma_{TI}, \text{ since } \sigma^2_I = \sigma_{TI}$$

$$= \sigma^2_T\left(1 - \frac{\sigma_{TI}}{\sigma^2_T}\right)$$

$$= \sigma^2_T\left(1 - \frac{(\sigma_{TI})^2}{\sigma^2_T \, \sigma^2_I}\right) \text{since } \frac{\sigma_{TI}}{\sigma^2_I} = 1$$

$$= \sigma^2_T(1 - r^2_{TI}).$$

It is also of interest to note that

$$\sigma^2_I = r^2_{TI}\sigma^2_I$$

The proof of this is,

$$\sigma^2_I = \frac{(\sigma^2_I)^2}{\sigma^2_I} = \frac{(\sigma_{TI})^2}{\sigma^2_I} = \frac{(\sigma_{TI})^2}{\sigma^2_T \sigma^2_I} \sigma^2_T = r^2_{TI}\sigma^2_T.$$

2. $E(T|y_1,\ldots,y_N)$ = the selection criterion in the multivariate normal case. This comes directly from the well known result concerning the mean of a conditional distribution in the multivariate normal distribution. Thus, the average value of *T's* associated with a given set of *y's* is equal to

$$\mu_T + b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N),$$

where the *b's* are exactly those of the selection index. Accordingly, we can state that the selection index procedure takes as the selection criterion the average value of all *T's* that are associated with *y's* equal to those on the individual that is a candidate for selection. Of course, this subset of *T's* shows variation, but less than the variation of *T's* in the entire population. From multivariate normal theory, this variance is

$$\sigma^2_T(1 - r_{TI}^2).$$

3. The probability of selecting the higher of a pair of *T's* is maximized. The proof of this is presented in the next section of this paper.

Unknown Means

What if the μ'*s* are not known? In the equal information case any arbitrary values can be used, for it can be seen that

$$\begin{aligned} I &= b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N) \\ &= b_1y_1 + \ldots + b_Ny_N - (b_1\mu_1 + \ldots + b_N\mu_N). \end{aligned}$$

Notice that the same function of the μ'*s* appears in each selection criterion and consequently has no effect on ranking. This is not the case when the information is different from one individual to another.

SELECTION INDEX FOR THE UNEQUAL INFORMATION CASE

When two individuals have different information available for evaluating their breeding values, it is clear that different indexes are required. But then there is more than one r_{TI}, and it is obvious that the justification of the selection index method described in the preceding section no longer is valid. For example, suppose selection is from two kinds, A and B. All individuals in the A group have the same kind of information, and an index say I_A is used to discriminate among them; similarly for the B group, I_B is used for discrimination. Then the expected progress through selection on the basis of these two indexes is

$$(N_A r_{TI_A} z_A + N_B r_{TI_B} z_B)/(q_A N_A + q_B N_B),$$

where N_A and N_B are the numbers of individuals available for selection in the two groups, $q_A N_A + q_B N_B$ is the number of individuals required to be selected, and z_a and z_b are ordinates of the unit normal distribution at the point of truncation. Maximization of this expression appears difficult since two sets of *b's*, q_A, and q_B must be determined. The difficulties multiply rapidly as the number of

different groups increases. Strangely enough this problem seems not to have been considered in previous discussions of selection.

Maximizing Probability of Selecting the Better of Two Individuals

The problem created by unequal information in the individuals considered for selection can be solved by finding a selection criterion which will maximize the probability of selecting the better of any two individuals. This method should then certainly maximize genetic progress. Suppose we have a set of records $y_1, \ldots, y_N$ available for choosing between individuals A and B with breeding values T_A and T_B. For example, y_1 might be the record on A, y_2 the record on the dam of A, and $y_3, \ldots, y_{12}$ the records on 10 progeny of B. The variance-covariance matrix of the y's is as before, C. The covariance between T_A and the y's is the vector, t_A and between T_B and the y's is t_B. T_A and T_B are assumed to have the same mean and can have any variance-covariance matrix we choose. These variables and the y's are assumed to follow the multivariate normal distribution. We want two indexes, one to compute a selection criterion for A and the second to compute a criterion for B.

$$I_A = b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N),$$
$$I_B = b_1^*(y_1 - \mu_1) + \ldots + b_N^*(y_N - \mu_N).$$

Note that the same set of records is used for the two indexes, but some of the b's and b^*'s may be zero.

In order to maximize the probability of selecting the better of two T's the following probabilities must be as large as possible.

$$P(I_A - I_B > O \mid T_A - T_B > O),$$
$$P(I_A - I_B < O \mid T_A - T_B < O).$$

Now for any fixed value of $T_A - T_B$, say k, the distribution of $I_A - I_B$, is normal with mean

$$\mu_{I_A} - \mu_{I_B} + b_{I_DT_D}(k - \mu_{T_A} + \mu_{T_B}),$$

where $I_D = I_A - I_B$ and $T_D = T_A - T_B$. This mean then simplifies to $b_{I_DT_D}k$, since $\mu_{I_A} = \mu_{I_B} = 0$ and $\mu_{T_A} = \mu_{T_B}$. The variance of this conditional distribution is

$$(1 - r^2_{I_DT_D})\,\sigma^2_{I_D}.$$

The probabilities above can be maximized if we maximize the ratio of the mean to the standard deviation when k is positive and minimize this ratio when k is negative. Both of these can be accomplished if we maximize the ratio of $b_{I_DT_D}$ to the standard deviation, that is,

$$\frac{b_{I_DT_D}}{\sqrt{(1 - r^2_{I_DT_D})\sigma^2_{I_D}}}$$
$$= \frac{\sigma_{I_DT_D}}{\sigma_{I_D}\,\sigma^2_{T_D}} \Big/ \sqrt{(1 - r^2_{I_DT_D})}$$
$$= \frac{1}{\sigma_{T_D}} \frac{r_{I_DT_D}}{\sqrt{1 - r^2_{I_DT_D}}}. \qquad (1a)$$

Since $\frac{1}{\sigma_{T_D}}$ is constant, maximization of (1a) is certainly accomplished by maximizing $r_{I_DT_D}$. But since $I_D = I_A - I_B$ is

$$I_D = (b_1 - b_1^*)y_1 + \ldots + (b_N - b_N^*)y_N$$
$$= \text{say } \beta_1 y_1 + \ldots + \beta_N y_N,$$

it is necessary now simply to solve the usual index equations (1) of the form

$$\beta_1\sigma^2_{y_1} + \beta_2\sigma_{y_1y_2} + \ldots + \beta_N\sigma_{y_1y_N} = \sigma_{y_1T_D} = \sigma_{y_1T_A} - \sigma_{y_1T_B},$$

etc.,
or in matrix notation,

$$C\,\beta = t_A - t_B \text{ since } \sigma_{yT_D} = \sigma_{yT_A} - \sigma_{yT_B} = t_A - t_B.$$

Then, $\beta = C^{-1}(t_A - t_B)$

$$= C^{-1}t_A - C^{-1}t_B. \qquad (2)$$

Now, suppose we compute separate indexes for evaluating A and B as though A were to be ranked relative only to others with the same information and B relative to others with the same information, but different from A's. Using equation (1), we have

$$C\,b_A = t_A \text{ or}$$
$$b_A = C^{-1}t_A, \text{ and}$$
$$C\,b_B = t_B \text{ or}$$
$$b_B = C^{-1}t_B.$$

Now note that,

$$b_A - b_B = C^{-1}t_A - C^{-1}t_B,$$

which is exactly the same as β, see (2). Thus, we have proved that the usual selection index criteria are best for ranking regardless of unequal information.

Unknown Means

It was shown in an earlier section that lack of information concerning the μ's has no effect on ranking when all individuals have the same information. This is not true, however, with unequal information. In the case above, involving A and B,

$$I_A - I_B = (b_1 - b_1^*)(y_1 - \mu_1) + \ldots + (b_N - b_N^*)(y_N - \mu_N).$$

Clearly this difference, which we use in choosing between A and B, contains a function of the μ's, and if the μ's are unknown, the difference cannot be computed. One way out of this difficulty is to let

$$I = b_1y_1 + \ldots + b_Ny_N \text{ rather than}$$
$$b_1(y_1 - \mu_1) + \ldots + b_N(y_N - \mu_N),$$

and then to maximize r_{TI} subject to the condition that $E(I) = O$. To illustrate, suppose y_1, y_2, y_3 are assumed to have a common mean, μ and we want an index,

$$I = b_1y_1 + b_2y_2 + b_3y_3,$$

subject to E(I) = O. Now,

$$E(I) = E(b_1y_1 + b_2y_2 + b_3y_3)$$
$$= (b_1 + b_2 + b_3)\,\mu.$$

Consequently, E(I) = 0 if $b_1 + b_2 + b_3$ is required to equal 0. This condition must therefore be imposed on the selection index equations. Suppose the usual equations are

$$\begin{aligned} 20\,b_1 + \quad b_2 + \; 2\,b_3 &= 5 \\ b_1 + 25\,b_2 + \; 3\,b_3 &= 2 \\ 2\,b_1 + \; 3\,b_2 + 30\,b_3 &= 1. \end{aligned}$$

By augmenting these equations with a Lagrange multiplier, a, as follows, maximization of r_{TI} subject to $b_1 + b_2 + b_3 = 0$ is accomplished.

$$\begin{aligned} 20\,b_1 + \quad b_2 + \; 2\,b_3 + a &= 5 \\ b_1 + 25\,b_2 + \; 3\,b_3 + a &= 2 \\ 2\,b_1 + \; 3\,b_2 + 30\,b_3 + a &= 1 \\ b_1 + \quad b_2 + \quad b_3 \quad\; &= 0. \end{aligned}$$

The solution to these equations is $b_1 = .1077$, $b_2 = .0367$, $b_3 = -.0710$, a = 3.0241. This is in contrast to the following solution when μ is known, $b_1 = .2455$, $b_2 = .0690$, $b_3 = .0101$.

A second logical approach to the problem of unknown μ's is to use their estimates in the regular index. In the above example, the index would be,

$$I = .2455(y_1 - \hat{\mu}) + .0690(y_2 - \hat{\mu}) + .0101(y_3 - \hat{\mu}).$$

Now it turns out that if the estimators used are those obtained by maximum likelihood from the *y's* that were employed in the index, the index is actually the same as that derived by requiring E(I) = 0. Let us illustrate in the above example. The maximum likelihood (m.l.) estimator of μ is $k_1y_1 + k_2y_2 + k_3y_3$, where the *k's* are the solution to the following equations:

$$\begin{aligned} 20\,k_1 + \quad k_2 + \; 2\,k_3 + a &= 0 \\ k_1 + 25\,k_2 + \; 3\,k_3 + a &= 0 \\ 2\,k_1 + \; 3\,k_2 + 30\,k_3 + a &= 0 \\ k_1 + \quad k_2 + \quad k_3 \quad\; &= 1. \end{aligned}$$

The solution is $k_1 = .4246$, $k_2 = .3257$, $k_3 = .2497$, a = −9.3169.

Then, $I = .2455(y_1 - \hat{\mu}) + .0690(y_2 - \hat{\mu}) + .0101(y_3 - \hat{\mu})$

$$= .2455\,y_1 + .0690\,y_2 + .0101\,y_3 - .3246\,\hat{\mu}$$
$$= .2455\,y_1 + .0690\,y_2 + .0101\,y_3 - .3246(.4246\,y_1 + .3257\,y_2 + .2497\,y_3)$$
$$= .1077\,y_1 - .0367\,y_2 - .0710\,y_3,$$

which is exactly the same as the index which requires E(I) = 0.

A general proof of the equivalence of these methods follows: The records available for evaluating an individual are the elements of an N $\times$ 1 vector, y, with variance-covariance matrix, C, and means $X\beta$, where X is a known $N \times p$ matrix and β is an unknown p $\times$ 1 vector. The covariance between T and y is the $N \times 1$ vector, t.

Then the usual selection index is

$b'(y - X\beta) = t'C^{-1}(y - X\beta)$, and if the m.l. estimators of β are substituted for β it becomes

$$t'C^{-1}(y - X\hat{\beta}).$$

The m.l. estimator is $\hat{\beta} = Ly$, where L, a p $\times$ N matrix, is the solution to

$$CL' + XA = 0$$
$$X'L' = I,$$

where A is a p^2 Lagrange multiplier, and I is a p^2 identity matrix (not the selection index). Solving these equations,

$$L' = C^{-1}X(X'C^{-1}X)^{-1}.$$

Therefore, $\hat{\beta} = Ly = (X'C^{-1}X)^{-1}X'C^{-1}y$.

Then the index $= t'C^{-1}[y - X(X'C^{-1}X)^{-1}X'C^{-1}y]$

$$= t'C^{-1}[I - X(X'C^{-1}X)^{-1}X'C^{-1}]y. \quad (3)$$

In the second method r_{TI} is maximized, subject to E(I) = 0. In this case b is the solution to the following equations

$$Cb + Xa = d$$
$$X'b = 0,$$

where a is a p $\times$ 1 Lagrange multiplier, and 0 is a p $\times$ 1 null vector. Solving these equations,

$$b = [I - C^{-1}X(X'C^{-1}X)^{-1}X']C^{-1}t,$$

and the selection index $= b'X$

$$= t'C^{-1}[I - X(X'C^{-1}X)^{-1}X'C^{-1}]y,$$

as in (3), thus completing the proof.

SETTING UP SELECTION INDEX EQUATIONS FOR ONE TRAIT

It is apparent from the preceding sections that the selection index method has very desirable properties at least in the multivariate normal distribution. But it must also be recognized that, strictly speaking, these properties exist only when the necessary population variances and covariances are known. Of course, the C matrix, the variance-covariance matrix of y's, can be estimated directly from an adequately large sample from the population of y's. In contrast, the covariance between T and the y's cannot always be estimated directly since T is sometimes unobservable. Therefore, quantitative genetic theory is then invoked to infer the value of such covariances. Also, on some occasions the elements of C are inferred from a combination of data and theory, if data alone are inadequate.

Coefficients of Left Hand Sides of Index Equations

Ideally one should like to have a very large sample from the N-variate population represented by the y's. Then the variance-covariance matrix can be estimated accurately enough that there need be no concern about the consequences of using an estimate of C rather than parameter values.

Computing C when all genetic variation is additive. In animal breeding the elements of C are sometimes estimated under the assumption that the model underlying the record on the *ith* animal is

$$y_i = \mu_i + g_i + e_i, \tag{4}$$

and that on the *jth* animal is

$$y_j = \mu_j + g_j + e_j,$$

where μ_i and μ_j are fixed, g_i and g_j are additive genetic values of the two individuals, and e_i and e_j represent all other causes of variation. It is assumed that g_i, g_j, e_i, e_j, follow a multivariate distribution with all covariances zero except that between g_i and g_j, which is stated to be a $a_{ij}\sigma^2_g$, where a_{ij} is the numerator of Wright's (8) coefficient of inbreeding and σ^2_g is the population additive genetic variance (the initial population in case there has been inbreeding). The variance of y_i is assumed to be $\sigma^2_e+(1+F_i)\sigma^2_g$, where σ^2_e is the variance of e in the original population, and F_i is the inbreeding coefficient of the *ith* individual. These assumptions imply:

1. No selection since the period defining the initial population.
2. All genetic variance is additively genetic.
3. No covariance between additive genetic values and environmental values and no covariance between environmental values.

Then the C matrix for computing b's to use with single records on N individuals is

$$\begin{pmatrix} \sigma^2_y + F_1\sigma^2_g & a_{12}\sigma^2_g & \cdots & a_{1N}\sigma^2_g \\ a_{12}\sigma^2_g & \sigma^2_y + F_2\sigma^2_g & \cdots & a_{2N}\sigma^2_g \\ \vdots & \vdots & & \vdots \\ a_{1N}\sigma^2_g & a_{2N}\sigma^2_g & & \sigma^2_y + F_N\sigma^2_g \end{pmatrix}, \tag{5}$$

where $\sigma_y^2 = \sigma_g^2 + \sigma_e^2$ = variance of records in the initial population. It is sometimes convenient to write this matrix as

$$\sigma^2_y \begin{pmatrix} 1 + F_1h^2 & a_{12}h^2 & \cdots & a_{1N}h^2 \\ a_{12}h^2 & 1 + F_2h^2 & \cdots & a_{2N}h^2 \\ \vdots & \vdots & & \vdots \end{pmatrix}, \tag{6}$$

where h^2 = heritability in the narrow sense = σ^2_g/σ^2_y.

More than one record per individual. In animal breeding applications two or more records on the same trait of an animal are sometimes used in selection. Let us assume as an approximation that the correlation between two records on an animal is $(r + Fh^2)/(1 + Fh^2)$, where r is the correlation in the initial population between records on the same animal. This implies a model

$$y_{ij} = \mu + p_i + g_i + \acute{e}_{ij},$$

where $p_i + e'_i = e_i$ of the model in (4); p_i is permanent to the individual, its variance is σ^2_p, and it is not affected by inbreeding. All elements of the model are uncorrelated. Then,

$$r = (\sigma^2_g + \sigma^2_p)/\sigma^2_y.$$

Under these assumptions and when the y's refer to the means of n_1 records in the first individual, n_2 in the second, etc., the *ith* diagonal element of (6) is modified to

$$\frac{1 + (n_i - 1)r}{n_i} + F_i h^2, \text{ etc.} \tag{7}$$

When $n = 1$, the diagonal element simplifies to $1 + Fh^2$, as it should.

Using group means. Oftentimes we wish to use the mean of some group, such as a set of progeny or of sibs in the selection index. Under the same assumptions as already stated in this section, the diagonal of (6) corresponding to any group, say the *ith* is

$$\left[\frac{1 + (n_i - 1)r}{n_i} + F_i h^2 + (p_i - 1)a_{ii'}h^2\right]/p_i, \tag{8}$$

where n_i is the number of records on each member of the group,
p_i is the number of individuals in the group,
F_i is the inbreeding coefficient of each member of the group, and
$a_{ii'}$ is the intra-group numerator relationship.

The off-diagonal elements of (6) remain the same as though there were only one member in the group. This, of course implies, that every member of a group has the same relationship to any other individual whose record is used in the selection index. Note that when $p_i = 1$, the expression in (8) reduces to (7), and when $n_i = 1$ reduces to

$$[1 + F_i h^2 + (p_i - 1)a_{ii'}h^2]/p_i.$$

Use of Genetic Variance Components

In a population with no inbreeding and with the environment contributing nothing to covariance between records on different individuals it is easy to express covariances between relatives' records in terms of Wright's coefficient of relationship, dominance relationship, and components of genetic variance. These genetic components are,

1. Additive: variance due to single gene effects.
2. Dominance: additional variance due to allelic gene pairs.
3. Additive × additive: additional variance due to non-allelic gene pairs.
4. Additive × dominance: additional variance due to a single gene and an allelic gene pair,

and so on.

In general, let σ^2_{ij} refer to the variance due to the interaction of i non-

allelic genes and j allelic gene pairs. Given that there are q loci which contribute to the genetic variance of a trait, the total variance is

$$\sum_{i=0}^{q}\sum_{j=0}^{q}\sigma^2_{ij}, \qquad 1 \leq i+j \leq q.$$

Then, the covariance between two related individuals, is

$$\sum_{i=0}^{q}\sum_{j=0}^{q} a^i d^j \sigma^2_{ij}, \qquad 1 \leq i+j \leq q, \tag{9}$$

where a is the Wright coefficient of relationship between i and j, and d is the dominance relationship between them. The dominance relationship is computed as follows for individuals A and B.

$$A\begin{cases}C\\D\end{cases} \qquad B\begin{cases}E\\F\end{cases}$$

$$d_{AB} = \frac{1}{4}[a_{CE}a_{DF} + a_{CF}a_{DE}]. \tag{9}$$

To illustrate (9), a and d for non-inbred full sibs are $\frac{1}{2}$ and $\frac{1}{4}$, respectively. Thus, the genetic contribution to their covariance is

$$\frac{1}{4}\sigma^2_{01} + \frac{1}{16}\sigma^2_{02} + \frac{1}{64}\sigma^2_{03} + \ldots$$

$$+\frac{1}{2}\sigma^2_{10} + \frac{1}{8}\sigma^2_{11} + \frac{1}{32}\sigma^2_{12} + \ldots$$

$$+\frac{1}{4}\sigma^2_{20} + \frac{1}{16}\sigma^2_{21} + \frac{1}{64}\sigma^2_{22} + \ldots$$

etc.

Little progress has been made in estimating these genetic components, but if good estimates were available and if environmental covariances could be eliminated, the problem of setting up C for calculation of indexes would be completely solved for non-inbred populations. Apparently gene frequencies are required to determine the contribution of many of the components to covariance between relatives in inbred populations and, of course, these frequencies are not available for genes affecting most traits of economic importance.

Right Hand Side of Index Equations

The right hand sides of the equations are $\sigma_{y_1T}, \ldots, \sigma_{y_NT} = t$ and depend obviously on our definition of T. Three different definitions seem logical in animal breeding applications when selection is for the individual:

1. Future production of the individual.
2. Production of progeny of the individual.
3. Production of descendants of the individual.

(In plant breeding, selection is often among lines or line-crosses. We shall discuss our definition of T for these cases in a later section).

Future production on the individual. If T = future production and if it is assumed that all records on the individual have correlation, r (= repeatability), with each other, $\sigma_{yT} = r\sigma^2_y$ in a non-inbred population. If serial correlations exist, a different σ_{yT} must be assumed for first with second records as compared to first with third, etc. In any case σ_{yT} is always a covariance between actual records, and consequently the problem of setting up the right hand side of the index equations is exactly the same as that for the coefficient matrix on the left.

Progeny production. If selection for production of progeny is the main concern of the breeder, the covariances between y and T are simply covariances between records on particular relatives. For example, suppose y_1 is a record on the dam of the individual considered for selection, and y_2 is the mean of paternal sibs of the individual. Then,

σ_{y_1T} = covariance between grandam's and grandprogeny's records.
σ_{y_2T} = covariance between "half-aunt" and niece.

Descendants' production. If selection is for descendants, this is almost equivalent to selection for additive genetic value, for note that in a non-inbred population the covariances between an individual's record and its descendants' records are

$$\text{Progeny: } \frac{1}{2}\sigma^2_{10} + \frac{1}{4}\sigma^2_{20} + \frac{1}{8}\sigma^2_{30} + \ldots$$

$$\text{Grand progeny: } \frac{1}{4}\sigma^2_{10} + \frac{1}{16}\sigma^2_{20} + \frac{1}{64}\sigma^2_{30} + \ldots$$

$$\text{Great grand progeny: } \frac{1}{8}\sigma^2_{10} + \frac{1}{64}\sigma^2_{20} + \frac{1}{512}\sigma^2_{30} + \ldots$$

$$\text{Descendant } n \text{ generations removed: } \frac{1}{2^n}\sigma^2_{10} + \frac{1}{2^{2n}}\sigma^2_{20} + \ldots + \frac{1}{2^{in}}\sigma^2_{i0} + \ldots$$

Thus, it is obvious that after very few generations, the coefficient of σ^2_{10} is overwhelmingly large as compared to any of the other components. Consequently, we should be primarily concerned with additive genetic value, that is we can let T = additive genetic value. Then σ_{y_iT} is simply $a_{i\alpha}\sigma^2_{10}$, where $a_{i\alpha}$ is the relationship between the animal with the *ith* record and α, the animal being evaluated. Further, we note that the value chosen for σ^2_{10}, appearing as it does in all right hand members, does not affect ranking, and consequently is not needed to maximize progress through selection. If, however, we wish to estimate how much progress will, in fact, be made we do need to know either σ^2_{10} or h^2.

If we use $a_{i\alpha}\sigma^2_{10}$ as right hand sides of equations in conjunction with left hand coefficients of the form in (6), we can then divide both sides of the equations by σ^2_y and obtain selection index equations requiring knowledge only of relationships, inbreeding coefficients, h^2, and if repeated records are used, r.

Then, r_{TI} has a simple computing form,

$$r_{TI} = \sqrt{\frac{b_1\sigma_{X_1T} + \ldots + b_N\sigma_{X_NT}}{\sigma^2_T}}$$

$$= \sqrt{\frac{b_1a_{1_\alpha}\sigma^2_{10} + \ldots + b_Na_{N_\alpha}\sigma^2_{10}}{\sigma^2_{10}}}$$

$$= \sqrt{b_1a_{1_\alpha} + \ldots + b_Na_{N_\alpha}} \quad .$$

Let us illustrate these last simple procedures. We wish to construct an index based on the individual's record, y_1, and a record on each of the parents, y_2, y_3. Then the equations to be solved for b's, using the simplifying assumptions are

$$\begin{bmatrix} 1 & \frac{1}{2}-h^2 & \frac{1}{2}-h^2 \\ \frac{1}{2}-h^2 & 1 & 0 \\ \frac{1}{2}-h^2 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} h^2 \\ \frac{1}{2}-h^2 \\ \frac{1}{2}-h^2 \end{bmatrix}.$$

The solution is $b_1 = h^2(2 - h^2)/(2 - h^4)$,

$$b_2 = b_3 = h^2(1 - h^2)/(2 - h^4), \text{ and}$$

$$r_{TI} = \sqrt{b_1(1) + b_2\left(\frac{1}{2}\right) + b_3\left(\frac{1}{2}\right)}$$

$$= \sqrt{\frac{h^2(3 - 2h^2)}{2 - h^4}} \quad .$$

As a second illustration, suppose we wish to select sires on the basis of the mean of p half-sib progeny. Then the index equations are

$$\frac{1 + (p-1)\frac{1}{4}h^2}{p}\, b = \frac{1}{2}h^2$$

$$b = \frac{2ph^2}{4 + (p-1)h^2},$$

$$\text{and } r_{TI} = \sqrt{\frac{ph^2}{4 + (p-1)h^2}}.$$

Alternative Computational Procedures

An interesting and sometimes useful variation on the selection index method is the following,

$$I = \gamma_1 \sigma_{y_1T} + \ldots + \gamma_N \sigma_{y_NT},$$

where γ's are the solution to the following equations,

$$\sigma^2_{1y}\gamma_1 + \sigma_{y_1y_2}\gamma_2 + \ldots + \sigma_{y_1y_N}\gamma_N = y_1 - \mu_1$$
$$\sigma_{y_1y_2}\gamma_1 + \sigma^2_{y_2}\gamma_2 + \ldots + \sigma_{y_2y_N}\gamma_N = y_2 - \mu_2,$$

It is seen that this procedure simply interchanges $(y_i - \mu_i)$ and σ_{y_iT} as compared to the conventional procedure. The advantage of this method is that if we wish to evaluate several individuals from the same set of records, we need to solve only one set of equations, for note that the right hand members are $y - \mu$, and these remain the same for all individuals to be evaluated from that set of records. In contrast, the usual method has on the right hand side σ_{y_iT}, which changes from one individual to the next, as T changes.

The proof of the identity of the two methods is very simple. In the usual method,

$$I = b'(y - \mu),$$

where b is the solution to

$$Cb = t, \text{ or}$$
$$b = C^{-1}t.$$

$$\text{Therefore, } I = (C^{-1}t)'(y - \mu) \qquad (10)$$
$$= t'C^{-1}(y - \mu).$$

In the new method

$$I = \gamma't,$$

where γ is the solution to

$$C\gamma = y - \mu.$$

$$\text{Therefore, } I = [C^{-1}(y - \mu)]'t$$
$$= (y - \mu)'C^{-1}t$$
$$= t'C^{-1}(y - \mu).$$

This is the same as (10) since a scaler is equal to its transpose.

If the μ's are unknown we can substitute their m.l. estimates in the right hand sides of these new equations or we can obtain identical results by letting the index $= \gamma't$, where γ is the solution to

$$C\gamma + X\alpha = y,$$
$$X'\gamma = 0,$$

and α is a $p \times 1$ Lagrange multiplier. The solution to γ is

$$[I - C^{-1}X(X'C^{-1}X)^{-1}X']C^{-1}y,$$

and the index is then $\gamma't = t'\gamma$

$$= t'[I - C^{-1}X(X'C^{-1}X)^{-1}X']C^{-1}y$$
$$= t'C^{-1}[I - X(X'C^{-1}X')^{-1}X'C^{-1}]y$$

which is the same as (3), the procedure described for maximizing r_{TI} subject to $E(I) = 0$.

Another interesting procedure, an expansion of which is useful in problems involving line crosses and in cases with unknown μ's, will now be described. Let $y_i = \mu + g_i + e_i \qquad i = 1, \ldots, N$

We wish to rank according to g's, their variance-covariance matrix being G. The variance-covariance matrix of e's is E, and g's and e's are uncorrelated. Consequently, the variance-covariance matrix of y's is $(G + E)$, and the covariance between y and g is G. Now it can be shown that the criteria for selection, say $v_1, \ldots, v_N$ = the vector v, are the solutions to

$$(I + EG^{-1})v = y - \mu \text{ or}$$
$$v = (I + EG^{-1})^{-1}(y - \mu). \tag{11}$$

To prove that this solution is identical to the conventional one we note that the criteria in the ordinary index procedure are

$B'y$, where B, an $N \times N$ matrix, is the solution to

$(G + E)B = G$, or

$B = (G + E)^{-1}G$.

Therefore, the criteria $= G'(G + E)^{-1}(y - \mu)$

$= G(G + E)^{-1}(y - \mu)$, since G is symmetric.

$= [(G + E)G^{-1}]^{-1}(y - \mu)$

$= (I + EG^{-1})^{-1}(y - \mu)$

$= v$ shown in (11).

When $\mu = X\beta$ is unknown, the following procedure yields simultaneously the m.l. estimator of β and selection criteria based on maximizing r_{TI} subject to $E(I) = 0$. Also, the procedure is equivalent to substituting $\hat{\beta}$ = m.l. estimator for β in the usual index equation.

$$X'E^{-1}X\hat{\beta} + X'E^{-1}v = X'E^{-1}y$$
$$E^{-1}X\hat{\beta} + (E^{-1} + G^{-1})v = E^{-1}y. \tag{12}$$

The last of these equations can be written

$$X\hat{\beta} + (I + EG^{-1})v = y \text{ or}$$

$v = (I + EG^{-1})^{-1}(y - X\hat{\beta})$, where $\hat{\beta}$ is some estimate of β. This is the same v as above when $\hat{\beta}$ is substituted for β. To prove that $\hat{\beta}$ is the m.l. estimator of β, we note that the m.l. estimator of β is the solution to

$$X'(G + E)^{-1}X\hat{\beta} = X'(G + E)^{-1}y \text{ or}$$
$$\hat{\beta} = [X'(G + E)^{-1}X]^{-1}(G + E)^{-1}y. \tag{13}$$

When we eliminate v from (12), the following equations result

$$X'WX\hat{\beta} = X'Wy, \text{ where}$$
$$W = E^{-1} - E^{-1}(I + EG^{-1})^{-1}$$
$$= E^{-1} - [E + EG^{-1}E]^{-1}.$$

Consequently we can show that the solution to $\hat{\beta}$ in (12) is m.l. if we prove that $W = (G + E)^{-1}$, or that $(G + E)W = I$

$$(G + E)W = (G + E)[E^{-1} - (E + EG^{-1}E)^{-1}]$$
$$= GE^{-1} + I - (G + E)(E + EG^{-1}E)^{-1}$$
$$= GE^{-1} + I - GE^{-1} = I, \text{ thus completing the proof.}$$

In many applications of the above method the e's are uncorrelated and have common variance σ^2_e. That is, $E = \sigma^2_e I$ and $E^{-1} = \frac{1}{\sigma^2_e} I$.

Consequently, by multiplying each equation of (12) by σ^2_e we obtain

$$X'X\beta + X'v = X'y$$
$$X\beta + (I + \sigma^2_e G^{-1})v = y.$$

To illustrate, let y_1 = the record on individual, and y_2 and y_3 = records on parents. The mean of each y is μ. The model is the simple one of (4). Then,

$$X' = (1 \quad 1 \quad 1),$$
$$\sigma^2_e = (1 - h^2)\sigma^2_y,$$

$$G = h^2\sigma^2_y \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 0 & 1 \end{bmatrix}, \text{ and } G^{-1} = \frac{1}{2h^2\sigma^2_y} \begin{bmatrix} 4 & -2 & -2 \\ -2 & 3 & 1 \\ -2 & 1 & 3 \end{bmatrix}.$$

Then, the equations to be solved to evaluate these three individuals are

$$\begin{bmatrix} 3 & 1 & 1 & 1 \\ 1 & \frac{4-2h^2}{2h^2} & \frac{-2(1-h^2)}{2h^2} & \frac{-2(1-h^2)}{2h^2} \\ 1 & \frac{-2(1-h^2)}{2h^2} & \frac{3-h^2}{2h^2} & \frac{1-h^2}{2h^2} \\ 1 & \frac{-2(1-h^2)}{2h^2} & \frac{1-h^2}{2h^2} & \frac{3-h^2}{2h^2} \end{bmatrix} \cdot \begin{bmatrix} \mu \\ v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} y. \\ y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

SELECTION INDEX FOR MORE THAN ONE TRAIT

The application of the selection index to selection for more than one trait requires only a simple extension of the principles described for one trait selection. In fact, if we define T properly, the techniques are exactly the same as in single trait selection. Suppose it is desired to select for breeding value with respect to s different traits and we denote the breeding values of these traits by $T_1, T_2, \ldots, T_s$. The records available for use in selection may be phenotypic observations on some or all of these traits in the candidates for selection or in their relatives.

One possibility for using a selection index on these several traits would be to construct selection indexes for computing a separate criterion for each trait on each individual and then to select on trait one only in the first generation, trait two in the second, and so on. This is called "tandem" selection. A second possibility would be to compute criteria as in tandem selection and then to select only those with all criteria equal to or higher than chosen minima. This is called selection by "independent culling levels." If, however, it is possible to assign to the traits relative economic values for increases of one unit, breeding value can then be defined as a weighted function of breeding values for the various traits. Thus, if the relative values are $v_1, v_2, \ldots, v_s$, the breeding value is defined as

$$T = v_1T_1 + \ldots + v_sT_s.$$

Employing this definition of T, the selection index equations, from the procedure of (1), have left members = C = the variance-covariance matrix of y's, and and the right members are elements of the N × 1 vector,

$$t = (\sigma_{y_1T}\ \sigma_{y_2T}\ \ldots \sigma_{y_NT})',$$

where $\sigma_{y_iT} = v_1\sigma_{y_iT_1} + \ldots + v_s\sigma_{y_iT_s}$.

Let t_1 = elements of vector of $\sigma_{y_iT_1}$,
t_2, = elements of vector of $\sigma_{y_iT_2}$,
etc.

Then, the right hand side of the selection index equations are

$$t = v_1t_1 + \ldots + v_st_s.$$

Consequently, the index equations are

$$Cb = t \text{ and}$$
$$b = C^{-1}t$$
$$= C^{-1}v_1t_1 + \ldots + C^{-1}v_st_s,$$

and the selection index is

$$b'y = v_1t_1'C^{-1}y + \ldots + v_st_s'C^{-1}y. \qquad (15)$$

An alternative procedure that leads to exactly the same result is to construct separate indexes for each trait and then to weight either these indexes or the sets of s criteria by the economic values, that is,

$$I = v_1I_1 + \ldots + v_sI_s.$$

The proof of the equivalence of these methods follows.

$I_1 = b_1'y$, where $b_1 = C^{-1}t_1$,
$I_2 = b_2'y$, where $b_2 = C^{-1}t_2$,
etc. Then,

$$I = v_1b_1'y + \ldots + v_sb_s'y$$
$$= v_1t_1'C^{-1}y + \ldots + v_st_s'C^{-1}y, \text{ which is the same as (15).}$$

This latter method has the distinct advantage that changes in relative economic values with time or differences from one location to another do not require construction of new indexes. For example, an extension worker who is

asked to advise dairymen on selection for both type and production realizes that the value of type relative to milk production is great for the breeder who capitalizes on show ring winnings by selling breeding stock but is of little or no value to the dairyman who sells only cull cows. The extension worker can, however, give this advice to all,

1. Evaluate animals for milk production with an index
   $I_m = b_1y_1 + \ldots + b_Ny_N$.
2. Evaluate the same animals for type with another index
   $I_t = \beta_1y_1 + \ldots + \beta_Ny_N$.
3. Weight the above two criteria computed for each animal by v_m and v_t.

The dairyman must decide for himself what values to use for v_m and v_t.

SELECTION OF LINES AND LINE CROSSES

The selection index method need not be restricted to selection of individuals, for exactly the same principles can be applied to discriminating among lines, line-crosses, or other genetic groups.

Selection of Groups for Top-Crossing

A certain number of genetic groups, inbred lines for example, are to be selected for top-crossing on some specified population. A test is performed in which q individuals are selected at random from the *ith* group and n_{ij} top-cross progeny of the *jth* individual from the *ith* group are observed. The following model is assumed:

$$y_{ijk} = g_i + p_{ij} + e_{ijk}, \tag{16}$$

g, p, and e are normally, independently distributed with means 0 and variances σ^2_g, σ^2_p, σ^2_e. We wish to maximize progress in $\bar{g}$ by using an index of the form,

$$I_i = b_{i1}\bar{y}_{i1.} + b_{i2}\bar{y}_{i2.} + \ldots$$

The C matrix has according to the model (16), the following elements:

$$\text{diagonals} = \sigma^2_g + \frac{1}{n^2_{i.}} \sum_j n_{ij}{}^2\sigma^2_p + \frac{1}{n_{i.}}\sigma^2_e$$

$$\text{off diagonals} = \sigma^2_g.$$

The right hand sides are all σ^2_g.

Selection of Single Crosses

A random sample of lines from some population is chosen for producing some or all of the possible single crosses. A random sample of n_{ij} progeny from the cross of line i by line j is observed. On the basis of these results a certain number of crosses is chosen for further testing or for commercial production. A simple criterion is the line cross mean, but if n_{ij} is small, this clearly is not a very accurate method. It seems logical to suppose that a better criterion could be

found by using also the mean of the reciprocal cross and the data from all other crosses in which either of the parental lines appears.

A simple model that is appropriate for some species is

$$y_{ijk} = g_i + g_j + s_{ij} + e_{ijk}.$$

The elements of this model are normally and independently distributed with means 0 and variances σ^2_g, σ^2_s, σ^2_e. It is assumed that reciprocal crosses are equal, except for sampling. Consequently $s_{ij} = s_{ji}$. The model also assumes either that the lines are homozygous or that only one progeny per parent is tested. The model can be expanded to incorporate less restrictive assumptions, but it suffices to illustrate the principles of index selection of crosses.

Selection for general combining ability. By definition, general combining ability refers to the relative value of the *g's*. Consequently $T_i = g_e$. A simple indexing procedure to evaluate the αth line is $I = b_\alpha \bar{y}_\alpha$ where $\bar{y}_\alpha$ is the mean of all observations on the αth line, and

$$b_\alpha = \sigma^2_g/\sigma^2_{\bar{y}\alpha},$$
$$\sigma^2_{\bar{y}\alpha} = \sigma^2_g + (\sigma^2_g + \sigma^2_s)[\sum_{j\neq\alpha} (n_{\alpha j} + n_{j\alpha})^2]/(n_{\alpha.} + n_{.\alpha})^2$$
$$+ \sigma^2_e/(n_{\alpha.} + n_{.\alpha}).$$

If subclass numbers are unequal, a better index can be constructed by utilizing the data on all crosses rather than just those having the α line as a parent. Now the index is

$$I_\alpha = \sum_{i<j} b_{ij}\bar{y}_{ij}, \text{ where}$$
$$\bar{y}_{ij} = (y_{ij.} + y_{ji.})/(n_{ij} + n_{ji}).$$

To compute these *b*'s we use equations (1) where

$$\text{Diagonal element of C} = 2\sigma^2_g + \sigma^2_s + \sigma^2_e/(n_{ij} + n_{ji}), \tag{17}$$
$$\text{Off-diagonal elements of C having one subscript in common} = \sigma^2_g,$$
$$\text{Off-diagonal elements of C having no subscript in common} = 0, \text{ and}$$
$$\text{Right hand members} = \text{covariance between } \bar{y}_{ij} \text{ and } g_\alpha$$
$$= \sigma^2_g \text{ if one subscript of } \bar{y}_{ij} \text{ is } \alpha$$
$$= 0 \text{ if neither subscript is } \alpha.$$

Selection for single cross performance. In this case T is the value of a single cross, which for the cross of α by γ is

$$g_\alpha + g_\gamma + s_{\alpha\gamma}.$$

A variety of procedures all leading to the same result can be used. The problem is quite analogous to selection for more than one trait since breeding value in the single cross is a linear function of underlying random variables (*g's* and *s*) while that for multiple trait selection is a linear function of breeding values for the several traits.

One method is to use the index,

$$\sum_{i<j} b_{ij}\bar{y}_{ij},$$

where $\bar{y}_{ij} = (y_{ij.} + y_{ji.})/(n_{ij} + n_{ji})$. Then the C matrix is the same as described above, (17). The covariances for the right hand side of equations (1) when the cross is, say $\alpha \times \gamma$, are

Covariance with $\bar{y}_{\alpha\beta}$: $2\sigma^2_g + \sigma^2_s$,
with $\bar{y}_{\alpha j}$, $\bar{y}_{i\alpha}$, $\bar{y}_{\gamma j}$, $\bar{y}_{i\gamma}$: σ^2_g, and
with all other $\bar{y}_{ij}$: 0.

This method is tedious since it requires as many solutions to the index equations as there are crosses to be evaluated. Consequently it is desirable to use instead the method described in an earlier section, in which y's and σ_{yT}'s are interchanged.

SOLUTION TO THE SELECTION INDEX USING LEAST SQUARES EQUATIONS THAT ARE APPROPRIATELY MODIFIED

Let the linear model for y, and N × 1 vector of observations be,

$$y = X\beta + Zu + e \tag{18}$$

X is a known N × p matrix of rank p.

β is an unknown p × 1 vector.

Z is a known N × r matrix of rank r.

u is an r × 1 vector having a multivariate normal distribution with means = O, and variance-covariance matrix = D, which is a non-singular, r^2 matrix.

e is an N × 1 vector having a multivariate normal distribution with means = O and variance-covariance matrix = R, which is a non-singular, N^2 matrix.

u and e are independently distributed.

We wish to estimate β by m.l. and to use these estimators, $\hat{\beta}$, in selection indexes of the form,

$$\hat{u} = B'(y - X\hat{\beta})$$

$\hat{u}$ is an r × 1 vector corresponding to u, but this does not necessarily imply that $\hat{u}$ is an estimator of u. Rather it is a set of criteria for selection.

B is an N × r matrix computed according to the principle of selection index construction.

According to the model, (18), the variance-covariance matrix of y is $A = R + ZDZ$, and the covariance between y and u is ZD, an N × r matrix. Consequently, the index equations are,

$$AB = ZD \text{ and}$$
$$B = A^{-1}ZD.$$

Therefore, $\hat{u} = DZ'A^{-1}(y - X\hat{\beta})$. (19)

The m.l. estimator of β can be found by solving the following equations

$$X'A^{-1}X\hat{\beta} = X'A^{-1}y \text{ or}$$
$$\hat{\beta} = (X'A^{-1}X)^{-1}X'A^{-1}y. \tag{20}$$

An alternative procedure that is often much easier requires setting up least squares equations to solve for β and u as though u were fixed and then adding D^{-1} to the lower r^2 submatrix of coefficients. The following equations result. This method was suggested by Henderson (4) in 1952.

$$\begin{aligned} X'R^{-1}X\tilde{\beta} + X'R^{-1}Z\tilde{u} &= X'R^{-1}y \\ Z'R^{-1}X\tilde{\beta} + (Z'R^{-1}Z + D^{-1})\tilde{u} &= Z'R^{-1}y \end{aligned} \qquad (21)$$

We must now prove that $\tilde{\beta} = \hat{\beta}$ of (20) and that $\tilde{u} = \hat{u}$ of (19). To prove the former, we note that since in (21)

$$\tilde{u} = (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}(y - X\tilde{\beta}), \qquad (22)$$

equation (21) can be reduced to

$$X'W^{-1}X\tilde{\beta} = X'W^{-1}y, \text{ where}$$
$$W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}.$$

Therefore, if $W = A^{-1}$, $\tilde{\beta} = \hat{\beta}$. We show that this is true by proving $AW = I$.

$$\begin{aligned} AW &= (R + ZDZ')[R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}] \\ &= I + ZDZ'R^{-1} - Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1} \\ &\quad - ZDZ'R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1} \\ &= I + ZDZ'R^{-1} - Z(I + DZ'R^{-1}Z)(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1} \\ &= I + ZDZ'R^{-1} - ZD(D^{-1} + Z'R^{-1}Z)(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1} \\ &= I + ZDZ'R^{-1} - ZDZ'R^{-1} \\ &= I. \end{aligned}$$

In order to show that $\tilde{u} = \hat{u}$ we prove the following,

$$\begin{aligned} \tilde{u} &= (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}(y - X\tilde{\beta}), \text{ from (22).} \\ &= (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}AA^{-1}(y - X\tilde{\beta}) \\ &= (Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}(ZDZ' + R)A^{-1}(y - X\tilde{\beta}) \\ &= (Z'R^{-1}Z + D^{-1})^{-1}(Z'R^{-1}ZDZ' + Z')A^{-1}(y - X\tilde{\beta}) \\ &= (Z'R^{-1}Z + D^{-1})^{-1}(Z'R^{-1}Z + D^{-1})DZ'A^{-1}(y - X\tilde{\beta}) \\ &= DZ'A^{-1}(y - X\tilde{\beta}) \\ &= \hat{u} \text{ of (19).} \end{aligned}$$

Thus, we have proved that if least squares equations are set up under the assumption that the random elements of the model, except for e, are fixed and then add the inverse of the variance-covariance matrix of the random elements, we can solve directly for the m.l. estimators of the fixed elements of the linear model and for criteria to use in selection. In many problems this method has distinct computational advantages over the conventional selection index method and over the usual m.l. estimation (weighted least squares) of the fixed elements of the linear model.

In most applications R is diagonal or better yet is $\sigma_e^2 I$, which greatly simplifies setting up (21). Also in some cases D also is diagonal, in the single cross example above, for instance. But if D is a large non-diagonal matrix, its inversion can be avoided if the following equations are written,

$$\begin{aligned} X'R^{-1}X\tilde{\beta} + X'R^{-1}ZD\tilde{v} &= X'R^{-1}y, \\ DZ'R^{-1}X\tilde{\beta} + (DZ'R^{-1}ZD + D)\tilde{v} &= DZ'R^{-1}y. \end{aligned}$$

Then, $\tilde{\beta}$ has the same value as in (21), and $\tilde{u} = D\tilde{v}$ has the same value as $\tilde{u}$ in (21). The proof of this is

1. Substitute $D^{-1}\tilde{u}$ for $\tilde{v}$ in (22).
2. Pre-multiply the last equation of (22) by D^{-1}.
3. Note that the resulting equations are identical to (21).

It is interesting to note that the lower r^2 submatrix of the inverse of the coefficients of the left side of (21) is the variance-covariance matrix of the deviation of $\hat{u}$'s from their respective u's. That is,

$$E(\hat{u} - u)(\hat{u} - u)'.$$

CONSEQUENCES OF USING PARAMETER ESTIMATES AND ASSUMING NORMALITY

Some of the unsolved problems of index selection are:

1. What are the consequences of non-normality on the efficiency of a selection index constructed as though y and T have the multivariate normal distribution when they actually have some other distribution?
2. What are the consequences of using variance and covariance estimates in place of parameter values on (a) the effectiveness of selection and (b) on prediction of genetic advance?
3. How should indexes be constructed to maximize genetic progress when either or both of the assumptions, normality and known parameters, do not hold?

The use of electronic computers, which are becoming increasingly available to plant and animal breeders, for sampling investigations of these problems appears promising. Work along these lines is in progress at Iowa State University, Cornell University, and probably elsewhere.

REFERENCES

1. Cochran, W. G., 1951, Improvement by means of selection. Proc. Second Berkeley Symp. on Math. Stat. and Prob. 449–470.
2. Comstock, R. E., 1948, Statistics in animal breeding research. Proc. Auburn Conf. on Stat. Applied to Res. Stat. Lab., Auburn.
3. Hazel, L. N., 1943, Genetic basis for selection indices. *Genetics.* **28:** 476–490.
4. Henderson, C. R., 1952, Specific and general combining ability. Heterosis Ames, Iowa State College Press.
5. Legates, J. E. and Lush, J. L., 1954, A selection index for fat production in dairy cattle utilizing the fat yields of the cow and her close relatives. *Jour. Dairy Sci.* **37:** 744–753.
6. Lush, J. L., 1948, The genetics of populations. Mimeo. Iowa State University.
7. Smith, F. H., 1936, A discriminant function for plant selection. *Ann. of Eugenics.* **7:** 240–250.
8. Wright, S., 1922, Coefficients of inbreeding and relationship. *Am. Nat.* **56:** 330–338.