Published on *Plant Breeding E-Learning in Africa* (https://pbea.agron.iastate.edu)

Multiple Trait Selection

By William Beavis, Kendall Lamkey (ISU)

Objectives

- Explain the role of selection for multiple traits on genetic gain
- Learn methods for selecting multiple traits
- Learn how to evaluate the efficiency and effectiveness of selection on multiple traits

Multiple Trait Selection Strategies

Multistage selection: Selection for different traits at different stages during cultivar development.

Tandem selection: Selection for one trait until that trait is improved, then for a second; etc. until finally each has been improved to the desired level.

Independent culling levels: A certain level of merit is established for each trait and all individuals below that level are discarded regardless of values for other traits.

Index selection: Select for all traits simultaneously by using some index of net merit.

Index Selection

Index Selection - Definition and Objective

An **index** is the best linear prediction of an individual's or line's breeding value and takes the form of the multiple regression of breeding value on all sources of information.

The objective of an index is to find the linear combination of phenotypic values that maximizes the expected gain, or equivalently that maximizes the correlation between the index value and true worth (breeding value).

Index Selection Theory

Define

$$P_i = G_i + E_i(i = 1, ..., n)$$

where

P_i is the observed value of attribute i for an individual or line
G_i is the average of phenotypic values over a population of environments
E_i is the non-genotypic contributions from environments.

Note: Genotype x environment interactions are permitted, assuming genotypes and environments are associated entirely at random; such interactions are incorporated into E_i. If GxE are not random, then see Cooper and DeLacy (1994)

Assume G_i is composed entirely of additive effects of genes (breeding values). Define the genotypic economic value of an individual as:

$$H = \sum_{i=1}^{n} a_i G_i$$

where, the a_i are known relative economic values. Assume the quantities P_i and H are such that the regression of H on P_i is linear. Selection will then be based on the linear function.

$$I = \sum_{i=1}^{n} b_i P_i = \sum_{i=1}^{n} b_i (G_i + E_i)$$

Assumptions

Assume an equal amount of information on all individuals to be evaluated and selected. Assume that the distributions of P_i, G_i, and E_i are unknown, but that the mean and covariances are known.

Then,

$$P_i\tilde{N}ID(0,^{P}\sigma_{ii})(i=1,...,n))(G_i\tilde{N}ID(0,^{G}\sigma_{ii})(i=1,...,n))(E_i\tilde{N}ID(0,^{E}\sigma_{ii})(i=1,...,n))(Cov(G_i,E_i)=\ 0\ (i=1,...,n))(Cov(P_i,G_i)=\ {}^{G}\sigma_{ii}(i=1,...,n))(Cov(P_i,E_i)=0\ \ (i=1,...,n))(Cov(P_i,P_j)=\ {}^{P}\sigma_{ij}\ (i=j=1,...,n))(Cov(G_i,G_j)=\ {}^{G}\sigma_{ij}\ (i=j=1,...,n))(Cov(E_i,E_j)=\ {}^{E}\sigma_{ij}\ (i=j=1,...,n)$$

Mean and Covariance of H and I

With these assumptions we can derive the mean and covariance of H and I.

$E(H) = \sum_{i=1}^{n} a_i E(G_i) = 0$	$V(H) = \sigma_H^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \ {}^{G}\sigma_{ij}$
$E(I) = \sum_{i=1}^{n} b_i E(P_i) = 0$	$V(I) = \sigma_I^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} b_i b_j \ {}^{P}\sigma_{ij}$
$Cov(I, H) = \sigma_{IH} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \ {}^{G}\sigma_{ij}$	$Cov(G_i, I) = \sigma_{GiL} = \sum_{j} b_j^G \sigma_{ij}$

The objective of a selection index is to use some linear combination of trait values (I) to predict true genetic worth (H).

This can be accomplished by:

- maximizing expected genetic gain.
- maximizing the correlation of the sample index (I) with true worth (H).
- maximizing the probability of correct selection.
- minimizing the $E(I\text{-}H)^2$.

Williams (1962) showed that maximizing the correlation between I and H also maximizes the expected genetic gain and the probability of correct selection.

Derivation of the Optimum Index

Derivation of the Optimum Index

Maximizing correlation of I with H:

$$r_{IH} = \frac{Cov(I,H)}{\sqrt{V(I)V(H)}} = \frac{\sum_i \sum_j a_i b_j \, {}^G\sigma_{ij}}{\sqrt{(\sum_i \sum_j b_i b_j \, {}^G\sigma_{ij})(\sum_i \sum_j a_i a_j \, {}^G\sigma_{ij})}}$$

It can be shown that maximixing r_{IH} is equivalent to maximizing $\log(r_{IH})$.

$$\log r_{IH} = \log(\sum_i \sum_j a_i b_j \, {}^G\sigma_{ij}) - \frac{1}{2}\log(\sum_i \sum_j b_i b_j \, {}^P\sigma_{ij}) - \frac{1}{2}\log(\sum_i \sum_j a_i a_j \, {}^G\sigma_{ij})$$

Using least squares and differentiating with respect to b_j we get:

Invalid Equation

Rewriting the Normal Equations

These equations are called the normal equations (unrelated to normal distribution) and constitute n equations in n unknowns.

They can be rewritten as follows:

$$\sum_i b_i\ {}^P\sigma_{ij} = \sum_i a_i\ {}^G\sigma_{ij}(\frac{\sigma_I^2}{\sigma_{IH}})\ j = (1, ..., n)$$

Because we are only interested in relative values of b_i the constant term can be dropped resulting in the following equations:

$$\sum_i b_i\ {}^P\sigma_{ij} = \sum_i a_i\ {}^G\sigma_{ij}\ j = (1, ..., n)$$

Further Calculations

Consider these equations for 2 traits (n=2):

$$b_1 \ {}^{P}\sigma_{11} + b_2 \ {}^{P}\sigma_{21} = a_1 \ {}^{G}\sigma_{11} + a_2 \ {}^{G}\sigma_{21} \ (j = 1))(b_1 \ {}^{P}\sigma_{12} + b_2 \ {}^{P}\sigma_{22} = a_1 \ {}^{G}\sigma_{12} + a_2 \ {}^{G}\sigma_{22} \ (j = 2)$$

Solving we get:

Invalid Equation

Minimizing E(I-H)^2

The same equations can be derived by minimizing $E(I\text{-}H)^2$,

$$E(I-H)^2 = E[(\sum_i b_i P_i) - (\sum_i a_i G_i)]^2)(E(\sum_i b_i P_i)^2 - 2E[(\sum_i b_i P_i)(\sum_i a_i G_i)] + E(\sum_i a_i G_i)^2)(\sum_i \sum_j b_i b_j\ {}^P\sigma_{ij} - 2\sum_i \sum_j a_i b_j\ {}^G\sigma_{ij} + \sum_i \sum_j a_i a_j\ {}^G\sigma_{ij}$$

because,

$$E(\sum_i b_i P_i)^2 = V(I))(E[(\sum_i b_i P_i)(\sum_i a_i G_i)] = \mathrm{Cov}(I,H))(E(\sum_i a_i G_i)^2 = V(H)$$

Applying least squares:

$$\frac{\partial E(I-H)^2}{\partial_{ij}} = 2\sum_i b_i\ {}^P\sigma_{ij} - 2\sum_i a_i\ {}^G\sigma_{ij} = 0\ (j = 1, ..., n)$$

Dividing through by 2 and rearranging we get the normal equations:

$$(\sum_i b_i\ {}^P\sigma_{ij} = \sum_i a_i\ {}^G\sigma_{ij}\quad j = (1, ..., n)$$

which are identical to the equations presented previously.

Expected Genetic Gains

Expected Genetic Gains

To derive the expected genetic gains we need to make assumptions about the distributions of P_i, G_i and E_i.

Assume:

1. P_i, G_i and E_i are distributed normally with the mean and covariance structure given earlier.
2. Truncation selection on I.

Then,

$$\Delta H = E(\bar{H}_S - \bar{H}) = \beta_{HI}(\bar{I}_S - \bar{I})$$

where β_{HI} is the regression coefficient of H on I, which gives the mean value of H or any value of I. This is the standard way to calculate predicted gains from univariate selection. See, for example, Empig et al. (1972).

Truncation Selection

The situation with regard to truncation selection is based on the following:

where,
$\bar{I}$ is mean value of index in the population;
c is truncation point;
z is height of ordinate of the standard normal curve at the truncation point c;
P is the proportion of the population selected; and
$\bar{I}_s$ is the mean of the selected individuals.

Then,

$$f(I) = \frac{1}{\sigma_I \sqrt{2\pi}} e^{-[\frac{(I-\bar{I})^2}{2\sigma^2}]}$$

defines the frequency f(I) of individuals with index value I.

Selection Relationships

The proportion saved is related to the truncation point by

$$S = \int_c^\infty f(I)\partial I$$

and, the mean value of the selected group is:

$$\bar{I_S} = \frac{1}{S}\int_c^\infty If(I)\partial I$$

The selection differential (D) is given by:

$$D = (\bar{I_S} - \bar{I}) = \frac{1}{S}[\int_c^\infty If(I)\partial I - \bar{I}] = \frac{1}{S}\int_c^\infty (I - \bar{I})f(I)\partial I$$

Standardized Regression Coefficient

Let,

Invalid Equation

Then,

Invalid Equation

where, z is the height of the ordinate at the truncation point.

Typically this is represented as

$$\frac{S}{\sigma_I} = \frac{z}{P} = i$$

where, k is the standardized regression coefficient.

Expected Gain

Expected gain can then be represented as:

Invalid Equation

From the normal equations derived earlier we have:

$$\sum_i b_i \, {}^P\sigma_{ij} = \sum_i a_i \, {}^G\sigma_{ij} \qquad j = (1, ..., n)$$

Substituting we get,

Invalid Equation

Predicted Gain

However, the predicted gain is more useful when written in terms of the correlation between H and I designated as r_{HI}.

Invalid Equation

In the selection index literature, r_{HI} is called the accuracy of selection, because it is a measure of how well the index (I) measures the true worth (H).

Alternative selection indices (I) can be compared using r_{HI} as long as the selection goal (H) remains the same for each of the indices.

Expected Genetic Gains for Each Trait

Let ΔG_i be the expected genetic gain in trait i when selection is on I.

Invalid Equation

This index requires that you know the true values of the population parameters. However, estimates of the population parameters are often substituted for the true values and the resulting index is called the **estimated index** or the **Smith-Hazel index**.

Matrix Representation

Matrix Representation of Selection Indices

With this notation the normal equations can be written as:

$$Pb = Ga)and(b = P^{-1}Ga$$

Some Results

$\sigma_I^2 = \underline{b}'P\underline{b})(\sigma_H^2 = \underline{a}'G\underline{a})(\sigma_{IH} = b'Ga = b'Pb = \sigma_I^2)(\mathrm{Cov}(G_i, I) = \sigma_{GiI} = b'G_i,)whereGiistheithrowofG.$

Invalid Equation

Construction of a Selection index

Optimum Index

$$P\underline{b} = G\underline{a})(\underline{b} = P^{-1}G\underline{a}$$

where *P*, *G* and *a* are known without error.

$$I = b'\underline{x})(H = \underline{a'}\ \underline{y}$$

Smith-Hazel Index

Same as the optimum index, only we use estimates of P, G, and a, designated as $\hat{P}$, $\hat{G}$ and $\hat{a}$, respectively.

$$\hat{P}\underline{\hat{b}} = \hat{G}\underline{\hat{a}})(\underline{\hat{b}} = \hat{P}^{-1}\hat{G}\hat{a})(\hat{I} = \hat{b}'\underline{x})(H = \underline{\hat{a}'}\underline{y}$$

Base Index

The Base Index was apparently first suggested by Brim et al. (1959) and named the base index by Williams (1962). The base index is constructed simply by allowing b = a. Therefore,

$$I = \underline{a'x})(H = \underline{a'}\underline{y}$$

Some results for this index include:

$$\sigma_I^2 = \underline{a}'P\underline{a})(\sigma_H^2 = \underline{a'a})(\mathrm{Cov}(I,H) = \underline{a}'G\underline{a})(r_{IH} = \frac{\underline{a}'G\underline{a}}{\sqrt{(a'Pa)(a'Ga)}} = \sqrt{\frac{\underline{a}'G\underline{a}}{\underline{a}'P\underline{a}}}$$

The foremost attribute of this index is its simplicity of construction and of interpretation. Also, this index does not require estimation of genetic parameters.

Multiplicative Index

The multiplicative index was first proposed by Elston (1963). This index is also sometimes called the weight-free index, because it does not require the specification of index weights or economic values.

The general form of this index is:

$$I = (X_1 - k_1)(X_2 - k_2) \cdots (X_n - k_n)$$

where, k_1 is the minimum value of trait X_1 set by the breeder.

In addition to being weight-free, this index also does not require estimation of genetic or phenotypic parameters. Because this is a curvilinear index, theory is not available to predict gains. Baker (1974) found that this index can be approximated by using a linear index, where the weights are the reciprocals of the phenotypic standard deviations of the traits in the index. This essentially amounts to an index with equal weighting per phenotypic standard deviation. Approximate predicted gains can then be obtained for this index using the Smith-Hazel index theory.

Desired Gain Index

The desired gain index was suggested by Pesek and Baker (1969). This index allows the breeder to specify a vector of desired gains, $\underline{q}$ and then substitute this into the predicted gain equation and solve for $\underline{b}$ the index weights. The solution for $\underline{b}$ is:

$$\underline{b} = G^{-1}\underline{q}$$

where $\underline{q}$ is the vector of desired gains and $I = \underline{b'x}$, as before.

This index was proposed to eliminate the need to specify economic weights. However, in practice there are some difficulties with the index in specifying the vector of desired gains.

This index will result in maximum gains in each trait according to the relative importance assigned by the breeder in specifying the desired gains.

Predicted gains can be obtained by substituting the vector of index weights into the conventional Smith-Hazel predicted gain equations.

Restricted Selection Index

Restricted selection indices were first derived by Kempthorne and Nordskog (1959). Since then, various restricted indices have been derived by Cunningham et al. (1970), and James (1968). See Lin (1978) for a complete list. Basically, restricted selection indices involve holding the genetic gains in one or more traits to a constant or zero, while changing the means of other traits in the desired direction. The basic method is to impose the restriction on the index equations that $Cov(G_i, I) = 0$.

The simplest procedure to accomplish this was given perhaps by Cunningham et al. (1970). Their method involved solving the following set of equations.

$$\begin{pmatrix} P & G_i \\ G_i & O \end{pmatrix} \begin{pmatrix} \underline{b} \\ \underline{b_d} \end{pmatrix} = \begin{pmatrix} G \\ O \end{pmatrix} [a]) where, (b = [I{-}P^{-1}G_i(G_i'P^{-1}G_i)^{-1}G_i']P^{-1}G\underline{a}$$

and the dummy variable *b_d* is given by:

$$\underline{b_d} = (G_i'P^{-1}G_i)^{-1}G_i'P^{-1}G\underline{a}$$

$\underline{b}$ is the vector of index weights to use in the index equation as $I = \underline{b'x}$, as before. The dummy variable is not use in the index equation.

This method has the interesting consequences that the value obtained for the dummy variable is the negative of the economic weight needed to produce zero change in that trait in an unrestricted selection index.

Rank Summation Index

The rank summation index was first suggested by Mulamba and Mock (1978). Basically this index involves obtaining the ranks of each of the traits to be included in the index and then calculating the index by summing up the trait ranks.

$$I = \sum_{i=1}^{n} \mathrm{rank}(X_i).$$

The primary advantages of this index are that genetic parameters need not be calculated, it transforms the data so that the variances for each trait are identical, and it does not require the specification of economic weights, although they can be used.

As with the multiplicative index, predicted gains cannot be calculated for this index. However, Crosbie et al. (1980) found that the same prediction equation used for the multiplicative index provides a reasonably good approximation of the predicted gains for the rank summation index.

Selection Index Efficiency

Methods to Compare Selection Index Efficiency

Cunningham (1969) provided a method for comparing the relative efficiencies of selection indices. He was primarily interested in deleting traits from the index that their relative contribution to the gain in the true worth (H) could be determined. Dropping traits from the index means that fewer genetic parameters need to be estimated providing a considerable savings in cost.

Define the index containing all the traits of interest as the original index and define the index with one trait dropped out as the i^{th} reduced index. Then the efficiency of the ith reduced index relative to that of the original index is the ratio of their standard deviations. Cunningham showed this to be:

$$\sqrt{\frac{\underline{b}'P\underline{b}-\frac{b_i^2}{W_{ii}}}{\underline{b}'P\underline{b}}}$$

where, b_i is the i^{th} weighting factor in the original index, and W_{ii} is the corresponding diagonal element in the inverse of P. A more usual procedure is to compare the gain for the i^{th}trait when selection is on *I*, relative to single trait selection for the i^{th} trait.

Effect of Correlations on Index Weights

To determine the effects of correlation on index weights we need to derive the index equations $P\underline{b} = G\underline{a}$ in terms of genetic and phenotypic correlation coefficients.

The derivation follows. Let,

Invalid Equation

Derivation

$$P^* = \begin{pmatrix} 1 & {}^{P}r_{12} & \cdots & {}^{P}r_{1n} \\ {}^{P}r_{21} & 1 & \cdots & {}^{P}r_{2n} \\ \vdots & \vdots & & \vdots \\ {}^{P}r_{n_1} & {}^{P}r_{n_2} & \cdots & 1 \end{pmatrix})(G^* = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix})(\underline{\hat{b}}^* = (P^*)^{-1}G^*\underline{a}$$

For 2 traits (n=2),

$$\underline{\hat{b}}^* = \begin{pmatrix} a_1({h_1}^2 - {}^{P}r_{12}h_{12}) + a_2(h_{12} - {}^{P}r_{12}{h_2}^2) \\ a_1(h_{12} - {}^{P}r_{12}{h_1}^2) + a_2({h_2}^2 - {}^{P}r_{12}h_{12}) \end{pmatrix} \left(\frac{1}{1-({}^{P}r_{12})^2}\right)$$

When correlations are zero: ${}^{P}r_{12} = {}^{G}r_{12} = 0$,

Invalid Equation

Smith et al. (1981) have shown that when | r | < 0.30, the use of the above index is nearly as efficient as using the Smith-Hazel index.

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Quantitative Genetics Multiple Trait Selection Author: William Beavis, and Kendall Lamkey (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Beavis, W. and K. Lamkey. 2016. Multiple Trait Selection. *In* Quantitative Genetics, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.

Source URL: https://pbea.agron.iastate.edu/course-materials/quantitative-genetics/multiple-trait-selection?cover=1