THE UTILIZATION OF SELECTION INDEX IN ANIMAL BREEDING

L'utilisation de l'index de sélection dans les programmes d'élevage

Das Selektionsindexbenützung in Tierzucht

K. RÖNNINGEN

INTRODUCTION

The value of an animal is almost always affected by several traits. Therefore the breeder has to consider different characteristics when choosing the most valuable animals as parents for the next generation. Further, the amount of information may also vary from one individual to another. There are many ways of selecting animals for breeding, but the most efficient method is that which usually results in maximum genetic improvement per unit of time.

In 1942 HAZEL & LUSH suggested three methods of selection for net merit. The three well-known methods are designated as the «tandem» method, the method of «independent culling levels» and the «total score» method. The first consists of selecting for one trait at a time until that is improved, then for a second trait, and so on until finally each trait has been improved to the desired level. The second method requires the establishing of a certain lower level of merit for each trait, below which all individuals are discarded, regardless of the superiority or inferiority of their other traits. The «total score» method (also called the selection index method) consists of selecting for all the traits simultaneously by using some index of net merit constructed, as expressed by HAZEL & LUSH (1942), «by adding into one figure the credits and penalties given each animal according to the degree of its superiority or inferiority in each trait». The conclusion drawn from the study of HAZEL & LUSH (1942) and YOUNG (1961) is that the index method is never less efficient than the other two methods, but the use of an independent culling level permits earlier selection without waiting until all traits are recorded. Therefore it would seem warranted to discuss the selection index in greater detail.

* Department of Animal Breeding, Agricultural College, S-750 07 Uppsala 7, Sweden.
For the sake of completeness it might be mentioned that Aplanalp (1972) suggested the use of «Selection of extremes». By this method individuals are ranked for each trait separately and proportions of the highest ranking individuals in each trait are selected to give the desired number of parents.

Short description of the selection index theory and the properties associated with the selection criterion computed by the selection index.

The first application of the selection index to plant breeding was that of Smith (1936), while the first application to animals was that of Hazel (1943). Since that time several papers considering the selection index theory can be found in the literature (for instance Comstock, 1948; Cochran, 1951; Legates & Lush, 1954; Henderson, 1952; Young, 1961; Young, 1964; Young & Weiler, 1960. See Rønningen (1972) for a more complete review). Henderson (1963) discussed the theory and the use of the selection index in an excellent way. He emphasized that the selection index can be used for several purposes, e.g.:

* Selection based on a single trait, using information on the individual and certain of its relatives.

** Selection on two or more traits using records derived from the individual.

*** Selection on two or more traits using records from the individual and its relatives.

**** Selection of line-crosses using data in addition to that on specific cross.

Briefly, the index approach may be summarized in the following way. Let:

\[ Y' = Y, \ldots, Y_m \] be a vector of additive genetic values for the \( m \) traits included in the true breeding value.

\[ v' = v, \ldots, v_n \] be a vector of relative economic values of these traits.

\[ X' = X, \ldots, X_n \] be a vector of phenotypic measures for the \( n \) variables or sources of information to be included in the index. Different phenotypic measures and different sources of information can also be included, simultaneously.

\[ b' = b, \ldots, b_n \] be a vector of factors (partial regression coefficients) to be used in the index.

\( P \) be an \( n \times n \) matrix of phenotypic covariances between the \( X \).

\( G \) be an \( n \times m \) matrix of genetic covariances between the \( n \) variables in \( X \) and the \( m \) traits in \( Y \).

\( C \) be an \( m \times m \) matrix of genetic covariances between the \( m \) traits in \( Y \).

The index is now:

\[ I = b'X \]

and its variance becomes

\[ \sigma^2_I = b'Pb \]

The true breeding value is

\[ T = v'Y \]

and its variance is

\[ \sigma^2_T = v'Cv \]
The general index equations become

\[ Pb = Gv \]

which may be solved to give

\[ b = P^{-1}Gv \]

The covariance between index and true breeding value is

\[ \sigma_{ti} = b'Gv = b'Pb \]

Note that this verifies that the variance of the index and the covariance of index and true breeding value are equal.

The correlation between index and true breeding value is therefore

\[ r_{ti} = \frac{\sigma_{ti}}{\sigma_i \sigma_T} = \frac{b'Pb}{\sqrt{v'Cy}} \]

Further, the genetic progress from selection based on the index may be described as follows:

The average genetic superiority of the individuals selected over the mean of all candidates for selection is:

\[ \Delta T = r_{ti}i\sigma_T \]

\[ i \] is the selection differential on a standardized normal distribution.

The gain in aggregate genotype arising from one round of selection on the index can therefore be expressed in the following way,

\[ \Delta T = r_{ti}i\sigma_T = \frac{\sigma_{ti}}{\sigma_T} i\sigma_T = \sqrt{\frac{\sigma_{ti}}{\sigma_T}} i = \sqrt{b'Pb} \]

Henderson (1963) has also given a method for constructing the selection index where the covariances between the true breeding value and the sources of information are used in the index equation.

As concerns the selection index with restriction, reference should be made to Kempthorne & Nordskog (1959), Wilton & Van Vleck (1968, 1969), Tallis (1962), Cunningham & Gedrem (1970), Cunningham et al. (1970), Rao (1962, 1963) and James (1968). Further, Henderson (1952, 1963) has considered the use of a selection index in the selection of line-crosses using data in addition to that on the specific cross.

The criterion to be used in selection could also have been developed from Henderson's mixed model (1959, 1963). If least squares equations are set up under the assumption that the random elements of the model, with the exception of \( e \) \( y = X\beta + Z\mu + e \), are fixed and the inverse of the variance—covariance matrix of the random elements is then added, it can be solved directly for the maximum likelihood estimators of the fixed elements of the linear model and for the criterion to be used in selection.

Which properties are associated with the selection criterion computed by the selection index? They are:
a) The correlation between $T$ and $I$ is maximized. This property does not require the multivariate normal distribution (Kempthorne, 1957; Henderson, 1963).
b) The expected squared difference between $I$ and $T$ [i.e. $E ((I-T)^2)$] is minimal among all linear functions of the general form of the selection index. This property does not require the multivariate normal distribution (Tallis, 1962; Henderson, 1963).
c) The probability of selecting one of the largest sample values of total merit by selecting the largest value of the index criteria is maximal (Williams, 1962).
d) The probability of selecting the higher merit of any two individuals is maximal (Henderson, 1963).
e) The genetic progress in any one-round selection by the index is maximal (Henderson, 1963).

The properties in c), d) and e) require the multivariate normal distribution.

Some comments on the selection index

As other reporters are going to consider the principles for evaluating dairy bulls and cows in greater detail I will limit my consideration to some topics which may be of interest when applying the method of selection index.

1. Defining the aggregate genotype

Several questions can be asked with respect to the definition of the aggregate genotype. Some considerations may therefore be made.
a) All traits of economic importance must be included in the aggregate genotype (Gjedrem, 1971). Thus any trait of economic importance should be considered even though it is not recorded. An interesting study was carried out by Van Vleck (1970) where each trait is assumed to have one direct and one maternal additive genetic component of economic importance.

If the breeding program involves specialized sire and dam lines the conclusion may be different (Smith, 1964).

b) Efforts should be made to estimate the economic values as precisely as possible. On the other hand the efficiency of the index does not suffer much from relatively large errors in the economic values (Pease et al., 1967; Rønningen, 1971). Pease et al. (1967) found that an error of ±50% in the economic weight for any one trait reduced the efficiency of the index by less than 2%. Altogether, seven traits were considered in swine.

c) Different ways of defining the aggregate genotype with respect to milk yield, fat percentage and milk fat yield may be found in cattle. In many countries the economic value of a unit of fat percentage varies with the level of production. Therefore Syrstad (1966) developed an index based on milk and fat yield rather than on milk yield and fat percentage. Wilton & Van Vleck (1968, 1969) used another approach in developing sire and cow indexes. They defined the aggregate genotype as follows:

$$T = (\mu_M + T_m) [a_1 + a_2 (\mu_F + T_f)]$$

where

$\mu_M$ and $\mu_F$ are the population means for milk yield and fat percentage.
and $T_f$ are the genotypic deviations from population means for milk yield and fat percentage.

$$a_i = v_i - v_j B_r$$

where $B_r$ is the base test; $v_i$ is the value per kilogram of milk at the base test; $v_j$ is the value of each kilogram of milk with each change of 0.1% test.

$$a_i = v_j$$

It may be mentioned here that the substitution method, that is, where sub-indexes for milk yield and fat percentage are developed by the use of information on both milk yield and fat percentage, and the sub-indexes replace milk yield and fat percentage in the merit equation, gives the same result as the conventional method, when linear or quadratic models of total merit are studied (Wilton et al., 1968; Rönningen, 1971). The advantage of using the substitution method is that the economic weights can easily be varied.

2. **Unknown means**

Let us consider the following index: $I = \sum_{i=1}^{n} \Sigma b_i (X^*_i - \mu_i)$. A reasonable question is then what would be the best thing to do if the $\mu$'s are not known. As pointed out by Henderson (1963), arbitrary values of the $\mu$'s can be used in cases where the information is the same from one individual to another. This is not true with unequal information. One way to solve the problem in the case of unequal information is to set up the index: $I = \sum_{i=1}^{n} X^*_i$, and then maximize the correlation between the true breeding value and the index, subject to the condition that $E(I) = 0$. A second logical approach to the problem of unknown $\mu$'s is to use their estimates in the regular index.

3. **Selection index in data bank**

One way to make the breeding scheme as efficient as possible is to use all the available information to rank the animals according to their breeding values. In short, this is also the principle of the data bank, i.e. accumulating, storing and applying the information at hand at any given time. The theory of data bank is rather, old, but its use has become more justified as the computing facilities improve. In the last years, therefore, several studies on the use of data bank have appeared in the literature (cf. Skjervold, 1968; Skjervold, 1970; Gjedrem, 1969; Syrstad, 1971; Danell, 1972; Moen, 1972; Lindhe et al., 1973; Fimland, 1973).

As shown in Figure 1 there are several sources of information on an animal. In practice, the question is not to use only one of these sources, but to combine all the available sources of information in an optimal way. If we consider the sequence in cattle breeding the first estimate on the breeding value of an animal may be based on the index values of sire and dam. The next index value of the proband (bull calf) will be estimated at the stage where the performance test for meat characteristics is ready (dual-purpose cattle). This index should be based
on the index values of the parents and the performance testing result of the proband and his sibs (cf. Fimland, 1973). When the daughters have had their first lactation (or part of it) a more complete index for the bull can be calculated. An example of the use of data bank in cattle breeding is given in Figure 2.

In the data bank, every individual has, of course, its own account where the information is recorded. If we consider a cow, the information may be:

- Identification
- Date, year and month of birth
- Number of recording years
- The yield, fertility characteristics, etc.
- Body size
- Herd averages
- Identification of sire and dam
- The index value
- By using this account a list of the elite dams may be developed.

It would be too comprehensive to go into the various aspects of the data bank in detail here. In the following I will therefore consider only two aspects: the use of sub-indexes which, finally, are combined to form a final index. Let us consider a simple example. Growth rate \( (X_1) \) and feed efficiency \( (X_2) \) are recorded for young bulls. We can then build up a sub-index: \( I_i = b_1X_1 + b_2X_2 \); and \( I_i \) is stored. As the bulls get older, information on milk yield \( (X_3) \) together with milk charac-

![Diagram of Different sources of information on an individual.](image-url)
teristics \((X)\) of their daughters becomes available, and another sub-index can be presented: \(I_2 = b_3 X_3 + b_4 X_4\), and \(I_2\) is stored. Thus only two sources of information on the parents and information on performance testing of the proband and his sibs can be presented:

\[
I_2 = b_3 X_3 + b_4 X_4
\]

and \(I_2\) is stored. Thus only two sources of information instead of four are stored. The «final» index is:

\[
I = c_1 I_1 + c_2 I_2
\]

All the traits are of economic importance. To what extent this method functions satisfactorily depends partly on the amount of information from the various sources. It seems that the sub-index method is a good approximation (Moen, 1972; Rønningen, 1972). Thus, Moen (1972) found a loss of precision of only 0.75% when using the sub-index method in the ranking of boars.

The effect of selection on variances and covariances was considered by Cochran (1951). If selection is based on \(I\), then the variances and covariances to be used in the estimation of \(b_3\) and \(b_4\) are reduced (see also Rønningen, 1972). It turns out, however, that the variances and covariances are reduced to the same extent and the partial regression coefficients are thus not affected. The correlation between the index and the aggregate genotype is affected.

4. Variances and covariances

In the literature many studies deal with the efficiency of the selection index when unreliable variances and covariances are used. Thus Williams et al. (1961) considered the following four indexes:

- **Index I** for ranking and selecting bull calves
- **Index II** for selecting bull to be progeny tested
- **Index III** for selecting bull sires and cow sires

**Figure 2.**—An example of the use of data bank for selecting bulls in cattle breeding.
The *optimal* index which was developed on the basis of the population parameters.

The *estimated* index which was developed on the basis of estimates of variances and covariances.

The *reduced* index where variables having a weak correlation with the true genetic values were ignored.

The *base* index where the phenotypic values were weighted by the economic values.

**Williams et al. (1961)** concluded that the *base* index or the *reduced* index should be preferred if the variances and covariances are estimated on small materials. As expected, the *optimal* index was the most efficient.

**Heidhues & Henderson (1962)** discussed the use of the estimated, base, heritability, and optimal indexes. In the heritability index the phenotypic values were weighted by the product of heritability estimate and economic value. They concluded that the efficiency of the three former reached 70-90% that of the optimal index. Further, they recommend the use of heritability index or base index in cases where the estimates are unreliable.

**Harris (1964)** also emphasized the importance of estimating variances and covariances on a large data set. There was a tendency to overestimate the genetic progress from selection for a particular calculated index when using a limited amount of data. The volume of data necessary must be in excess of 1000 observations for paternal half-sib estimation for two trait indexes when the true heritabilities are around 0.2.

According to **Pease et al. (1967)** errors in the heritability estimates reduce the efficiency of the selection index to a lesser extent than do errors in the correlations.

**Mao (1971)** also found the true genetic gain to be over-estimated when parameter estimates are used. He also recommended the use of base index in those cases where the number of observations available for estimation are small. An interesting part of Mao's study is his method of dealing with unreasonable estimates. The following modifications were tried:

\[
\begin{align*}
\text{If } & \hat{\sigma}_{T_i}^2 \leq 0, \text{ set } \sigma_{T_i}^2 = 0.005 \\
\text{If } & \hat{\sigma}_{X_i}^2 < \hat{\sigma}_{T_i}^2, \text{ set } \sigma_{X_i}^2 = \sigma_{T_i}^2 \\
\text{If } & |\hat{\sigma}_{T_i T_i'}| > \hat{\sigma}_{T_i} \hat{\sigma}_{T_i'}, \text{ set } |\hat{\sigma}_{T_i T_i'}| = \hat{\sigma}_{T_i} \hat{\sigma}_{T_i'} \\
\end{align*}
\]

retaining the original sign for \(\hat{\sigma}_{T_i T_i'}\).

Such modifications involve problems, in that the mathematical properties of those estimates are unknown. **Mao (1971)** concluded that modification processes of unreasonable sample estimates lead to a more effective index construction and gain prediction in small samples, and no notable effects are observed when sample size becomes large. According to **Searle (1970)** the probability of obtaining a negative estimate of the sire component of variance is close to 10% when ten sires and five offspring from each sire are available.
5. Distribution

As pointed out earlier, three of the optimum properties require the assumption of a multivariate normal distribution of $Y$ and $X$. Little work has been carried out on the problem of non-normality in the context of selection index theory, except for that of Cochran (1951). To quote Cochran (1951): «In general, a linear index will not be the best index, and predictions of the expected gain in $y$, based on normal theory, are likely to be in error. Unfortunately, it cannot be taken for granted that a moderate departure from normality will have little effect. This may be so if selection is not intense and $y$ has only a small correlation with the $x$’s, so that progress is slow. But in intense selection the gains depend primarily on the shapes of the tails of frequency distributions. As is well known, a frequency curve which looks quite similar to the normal curve may differ greatly in its tail. A combination of theoretical investigations with sampling experiments on natural populations is suggested». Quite a few other papers (e.g. Mather, 1949; Feller, 1957) have discussed the distribution of genotypic values of quantitative traits. These papers generally substantiate the usual assumption of normality by an argument based on the central limit theorem. This fact may also explain the neglect shown for the problem of non-normality up to date. Heidhues (1961) and Glodek (1971) assumed that the normal distribution was a reasonable approximation for traits which are affected by many genes, each with a small effect.

6. Regressed unweighted least squares versus weighted least squares methods.

Harvey (1971) considered the use of direct maximum likelihood or weighted least squares estimation procedures to adjust for fixed effects and obtain estimates for sire effects, in comparison with the regressed unweighted least squares method. Briefly, the former consists, computationally, of adding the ratio between the error component ($\sigma^2_e$, known) and the sire component ($\sigma^2_s$, known) to each diagonal element of the least squares coefficient matrix in each of the sire equations. The regressed unweighted least squares method consists of estimating sire effects ($s$) adjusted for herd effects, by regressing the unweighted least squares estimates of the $s$ as follows:

$$s_i = \frac{\sigma^2_e}{\sigma^2_e + c_{ij}\sigma^2_s} \hat{s}_i$$

$c_{ij} =$ the inverse diagonal element associated with the $j^{th}$ sire constant.

Harvey (1971) concluded from his empirical study that the regressed unweighted least squares procedure should be preferred if the correlation between herd and sire effects is as large as 0.25. When the correlation is 0.42-0.45 the direct maximum likelihood seems to be about 4 to 8% less accurate than the regressed unweighted least squares procedure. Further, a correlation among herd and sire effects caused both the estimate of the sire variance component and the intra-class correlation estimate to be biased downwards when method 3 of Henderson was used to estimate the variance components.
Fimland (1972) has given a procedure for testing the hypothesis of equal expectations for inter- and intra block (herd) estimates. He pointed out that a combination of the intra- and inter-block estimators (i.e. weighted least squares) should not be performed if they have different expectations.

In this review the study of Henderson (1971) should also be mentioned. Frequently, milk production records are available for the progeny of several sires which are distributed among herds in different periods of the time. The progeny group size may also vary among sires. The problem then is to choose some computationally feasible prediction method which has optimum properties. Henderson (1971) considered cases where herds and sires are fixed, herds are fixed and sires are random, and where both herds and sires are random in the prediction. The variances of the differences (prediction error variances) among sires were smallest in the latter case (herds and sires are random) and largest in the former (herds and sires are fixed).

Henderson (1971) also developed a method for eliminating bias due to selection. By this method it can be decided if, for example, sires 1 and 2 have progeny in treatment 2, because their progeny gave more favourable results in treatment 1 than did the progeny of sire 3. Even the assumption of zero expectation of the error element in a model can be tested.

7. Probabilities deriving from sire evaluation models

In developing breeding plans it may be of interest to state some probabilities as to the breeding value of sires. The questions which are frequently asked are the following: Given an index value of a bull, what is the probability that the true breeding value exceeds a specified value? What is the probability that the genetic level of a group of bulls exceeds the genetic level of another group of bulls, given the difference between the index values? These problems are discussed by Allaire (1970) and Rønningen (1972) and reference may be made to these studies.

SUMMARY

In the present study the selection index theory is described. The properties of the selection criterion computed by the selection index are summarized. Furthermore, some comments are made on the definition of the aggregate genotype. It is emphasized that the selection index theory is of great importance in the use of the data bank. In particular, the possibility of building up sub-indexes should be utilized. If the estimates of the parameters to be used in developing the selection index are found in small samples it may be wise to apply a modified selection index. The regressed unweighted least squares and the weighted least squares methods are also considered.

When developing breeding programmes, it is also recommended that probabilities with respect to the breeding value of an animal or groups of animals are estimated.

RESUME

La théorie de l'index de sélection est décrite. Les propriétés de la critère de sélection, calculée par l'index de sélection, sont résumés. D'ailleurs, la définition
du génotype composé est critiqué. La théorie de l'index de sélection est très importante à l'application d'un «data bank». Surtout, il faut utiliser la possibilité de construire des sub-indexes. Quand les estimateurs des paramètres, à l'aide desquels l'index de sélection est obtenu, viennent d'un échantillon limité, il est préféré d'appliquer un index de sélection différent. Les méthodes des moindres carrés non pondérés et des moindres carrés pondérés sont évaluées.

En developpant un programme d'élevage il est aussi recommandable d'estimer des probabilités de la valeur d'élevage d'un animal et d'un groupe d'animaux.

**ZUSAMMENFASSUNG**


Es ist empfehlenswert beim Aufbau eines Züchtungsprogrammes die Wahrscheinlichkeiten der Zuchtungswert eines Tieres oder einer Gruppe von Tieren zu schätzen.

**LITERATURE**


