Chapter 4

Selection Index

So far we have seen how to predict breeding values based on a single own-performance record (3.3.1), based on repeated records (3.3.2) and based on progeny records (3.4). For real livestock breeding populations, these three procedures are not flexible enough, because we want to predict breeding values for a given selection candidate based on all available information. In the past, two different methods were developed which fulfill the requirement of considering all available information in the prediction of breeding values. Theses methods are

1. Selection Index Theory ((Hazel, 1943) and (Hazel and Lush, 1943)) and
2. Best Linear Unbiased Prediction (BLUP) ((Henderson, 1973) and (Henderson, 1975))

Both methods are based on the same genetic model. The main difference between the two methods consists in the way how they correct for identifiable systematic environmental effects. We start with a treatment of selection index theory. In chapter 5, the BLUP-based methods will be introduced.

4.1 Introduction

In principle, prediction of breeding values aims at assessing the genetic potential of a selection candidate that is due to additive gene effects based on all available information, such that the correlation between true and predicted breeding value is maximal. Because, we want to do this for a large number of selection candidates, we can formulate our aim in a more general way. For a given population, we want to predict breeding values for all animals in the population using all available information, such that the correlation between true and predicted breeding values are maximized. An alternative objective for the prediction to the maximization of the correlation between true and predicted breeding values is the minimization of the mean squared error of the prediction. The description of the aims of our procedure to predict breeding values shows that we are dealing with two different concepts of breeding value.

1. **True breeding value** which corresponds to the sum of all additive gene-effects
2. **Predicted breeding value** which is a function of the phenotypic observations (y) that is determined by statistical methods. As a prediction it is always associated with a certain error which we want to be minimal.

The prediction of breeding values has three different objectives.

1. Selection candidates are ranked according to the predicted breeding values. Hence, it provides a criterion for selecting parents out of a pool of selection candidates
2. Predicted breeding values are used to assess the response to selection and is important for planning a breeding program
3. Predicted breeding values are one criterion that affect the price of breeding animals and the price of seamen.

The definition 2.1 of the term breeding value has several problems when it comes to its potential usefulness for predicting breeding values.

- It is impossible to generate an infinite number of progeny before having a reliable prediction of the breeding value
- Due to the above mentioned objectives, we want to have a prediction of the breeding value available as early as possible.
- The predicted breeding value should be as accurate as possible

To address these issues, the above mentioned methods were developed. We start with the method of the selection index.

4.2 Selection Index Method

The selection index is a method to predict the breeding value of an animal (i) by using all available information on the animal and on its relatives. The result of the selection index method is an assignment of a numerical value (I) to each animal. All animals in the population can then be ranked according to their index value. The ranking according to the index value can be used as selection criterion. In principle the index I is defined as linear combination of all available information. This can be written as

$$I = \hat{a_i} = b_1 y_1 + b_2 y_2 + \cdots + b_n y_n = b^T y \tag{4.1}$$

where b is a vector of index weights and y is a vector of information sources. Here we assume that all values in y are corrected for appropriate mean levels. The resulting index value I in (4.1) is used as the predicted breeding value $\hat{a_i}$. From a statistical point of view equation (4.1) corresponds to a multiple linear regression. The vector of index weights b are understood as partial regression coefficients.

4.3 Aggregate Genotype

In most practical livestock breeding scenarios, we want to improve a population at the genetic level with respect to more than one trait or characteristic, simultaneously. This requires a procedure that enables us to combine the breeding values of several trait into one selection criterion. This criterion is called the **aggregate genotype** H. It is defined as

$$H = w_1 a_1 + w_2 a_2 + \cdots + w_m a_m = w^T a \tag{4.2}$$

where a corresponds to the vector of true breeding values and w is a vector of economic values. The economic value w_k for a given trait k is defined as the marginal change in profit caused by a small change in the population mean (μ_k) of the trait k. At this point, we are not describing how the economic values w_k are derived, but we consider them to be known. For the construction of the selection index, we are using the general form of the aggregate genotype H. Once the selection index is constructed, we can go back to the simple scenario of considering just one trait which reduces the aggregate genotype H to the true breeding value a of the single trait.

4.4 Theory of Index Construction

The term *index construction* stands for the computation of the vector of index weights b for a given set of information sources and a given aggregate genotype. Independently from the available information sources, the following parameters must be known

- heritabilities and phenotypic standard deviations for the traits in the aggregate genotype and for the traits in the index.
- phenotypic correlations between the traits in the index
- genetic correlations between the traits in the index and the traits in the aggregate genotype
- genetic correlations between the traits in the aggregate genotype
- economic values for the traits in the aggregate genotype

The objective of the index construction is to maximize the correlation r_{HI} between the index I and the aggregate genotype H. Because the index I corresponds to a multiple linear regression, the mean squared error between aggregate genotype and index is to be minimized. From this it follows that

$$E(H - I)^2 \rightarrow \ \min \tag{4.3}$$

The solution to the index construction objective in equation (4.3) leads to the so-called index normal equations which have the following form.

$$Pb = Gw \tag{4.4}$$

where P is the phenotypic variance-covariance matrix between all traits in the index, G is the genetic variance-covariance matrix between the traits in the aggregate genotype and in the index and w is a vector of known economic values. Solving for the vector of unknown index weights b leads to

$$b = P^{-1}Gw \tag{4.5}$$

The accuracy of the index is assessed by the correlation r_{HI} between the index I and the aggregate genotype H. The higher this correlation, the better the approximation of H by I. The correlation r_{HI} can be computed as shown in (4.6). The terms for $cov(H, I)$, σ_H and σ_I are taken from (4.23) and for b we insert the solution taken from (4.5).

$$\begin{aligned}
r_{HI} &= \frac{cov(H,I)}{\sigma_H \sigma_I} \\
&= \frac{w^T * G^T * b}{\sqrt{(w^T * C * w) * (b^T * P * b)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * ((P^{-1} * G * w)^T * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * G * w)}} \\
&= \sqrt{\frac{w^T * G^T * P^{-1} * G * w}{w^T * C * w}} \\
&= \frac{\sigma_I}{\sigma_H}
\end{aligned} \tag{4.6}$$

The response to selection R which results from applying a selection scheme according to the index I per generation is computed as

$$
\begin{aligned}
R &= i * r_{HI} * \sigma_H \\
&= i * \frac{\sigma_I}{\sigma_H} * \sigma_H \\
&= i * \sigma_I
\end{aligned} \tag{4.7}
$$

where i is the selection intensity.

4.5 Example of Index with Own Performance

The simplest case of an index I is the one where the aggregate genotype H consists of one trait and the index I contains a single own performance record of the same trait. This is equivalent to using the index I to predicting the breeding value a of an animal based on own phenotypic own performance record y. Hence we can set

$$
H = a \qquad \text{and} \qquad I = by^*
$$

During the index construction, we have assumed the information in the index to be corrected for the appropriate population mean μ. For our example here, we can set $y^* = y - \mu$. To determine the unknown index weight b which is on our example just a single number, we have to specify P, G and w. Because, we are looking at just one trait, the vector of economic values w is set to one. The matrix P was defined to be the variance-covariance matrix between the traits in the index. As the index I contains just one phenotypic record, then P corresponds to the phenotypic variance σ_y^2 of our trait of interest. The matrix G was defined to be the genetic variance-covariance matrix between the traits in the aggregate genotype and the traits in the index. In our example we have just one trait which is the same in H and in I, hence G corresponds to the additive genetic variance σ_a^2. In summary, we have found that

$$
\begin{aligned}
P &= \sigma_y^2 \\
G &= \sigma_a^2 \\
w &= 1
\end{aligned} \tag{4.8}
$$

Inserting the terms of (4.8) into equation (4.5) to compute the index weight b results in

$$
\begin{aligned}
b &= P^{-1} * G * w \\
&= \sigma_y^{-2} * \sigma_a^2 * 1 \\
&= \frac{\sigma_a^2}{\sigma_y^2} = h^2
\end{aligned} \tag{4.9}
$$

Using the index weight b found in (4.9) to compute the index I, we get

$$
\begin{aligned}
I &= by^* \\
&= h^2(y - \mu) \\
&= \hat{a_i}
\end{aligned} \tag{4.10}
$$

The index value I that we obtained in (4.10) corresponds to the predicted breeding value for a given trait of an animal i based on an own performance phenotypic record of animal i in the respective trait. Comparing the predicted breeding value obtained in (4.10) using selection index theory to the result obtained from the regression approach in (3.5) shows that they are identical.

The accuracy r_{HI} of the predicted breeding value $(\hat{a_i})$ using selection index theory is computed as shown in (4.6)

$$\begin{aligned} r_{HI} &= \frac{\sigma_I}{\sigma_H} \\ &= \frac{b\sigma_y}{\sigma_a} \\ &= \frac{h^2\sigma_y}{\sigma_a} \\ &= h \end{aligned} \tag{4.11}$$

Similarly to the predicted breeding value, the accuracy r_{HI} that results from selection index theory is identical to what was found using the regression approach.

4.6 Example with Progeny Records

The prediction of breeding values for a given animal i based on progeny records is very common in livestock breeding. Examples are dairy cattle where bulls are evaluated based on lactation records of daughters. Similarly for beef cattle or pigs where sires are evaluated based on carcass performance of their progeny. For a very long time this has been the standard method to predict breeding values to select parents in a breeding program. First we assume that the progeny of animal i are all half-sibs. Before, we can use the performance records of the progeny to predict breeding values for the parents, we have to correct them with the appropriate mean performance. After the correction the progeny performance values are averaged for a given parent. These mean performance values for a given parent i are called $\bar{y_i}$ and are used to predict the breeding values. Hence our index I for a given animal i is defined as

$$I = b\bar{y_i} \tag{4.12}$$

Because, we are only looking at a single trait, the aggregate genotype H corresponds to the single true breeding value a of this trait and the economic weight w is 1. Now we are ready to set up the index normal equations. In general these equations have the form

$$Pb = Gw \tag{4.13}$$

where P corresponds to the variance-covariance matrix of the information sources in the index. Our index I as defined in (4.12) contains just one source of information, namely the average $\bar{y_i}$ of the progeny performance values of animal i. In general the phenotypic variance of the mean $\bar{y}$ of n progeny performance values corresponds to

$$\sigma_{\bar{y}}^2 = \frac{1 + (n-1)t}{n}\sigma_y^2 \tag{4.14}$$

For our case with the progeny records, t takes the value of $\frac{1}{4}h^2$. For more details on how to compute $\sigma_{\bar{y}}^2$, see section 4.8. Hence the matrix P reduces to a single number

$$P = \sigma_{\bar{y}}^2 = \frac{1 + (n-1)h^2/4}{n}\sigma_y^2 \tag{4.15}$$

The matrix G in (4.13) is the genetic covariance matrix between the traits in H and the information sources in I. In our current example $G = cov(a_i, \bar{y_i}) = \frac{1}{2}\sigma_a^2$. For more details on how to compute G, see section 4.8.2. Now that we have all the components of (4.13), we can insert them and solve for b.

$$\begin{aligned}\frac{1 + (n-1)h^2/4}{n}\sigma_y^2 * b &= \frac{1}{2}\sigma_a^2 \\ b &= \frac{2nh^2}{4 + (n-1)h^2} \\ &= \frac{2n}{n+k} \end{aligned} \tag{4.16}$$

where $k = \frac{4-h^2}{h^2}$.

With this the predicted breeding value $\hat{a_i}$ for animal i based on the average progeny performance values using the index approach corresponds to

$$\hat{a_i} = I = b * (\bar{y_i} - \mu) = \frac{2n}{n+k} * (\bar{y_i} - \mu) \tag{4.17}$$

The accuracy for the predicted breeding value in (4.17) is

$$r_{HI} = \sqrt{\frac{n}{n+k}} \tag{4.18}$$

4.7 Appendix: Derivation of Index Normal Equations

In this section we want to show how to derive the index normal equations from the objective criterion in the index construction procedure. The objective criterion was formulated in equation (4.3) as

$$\Psi = E(H-I)^2 \rightarrow \ \min \tag{4.19}$$

The derivation starts by inserting the definitions of H and I into (4.19).

$$\begin{aligned}\Psi = E(H-I)^2 &= E(H^2 - 2 * H * I + I^2) \\ &= E(H^2) - 2 * E(H * I) + E(I^2)\end{aligned} \tag{4.20}$$

Both the expected value $E(H)$ of the aggregate genotype H and the expected value $E(I)$ of the index are both 0. This can be seen by the following expansion

$$E(H) = E(w^Ta) = w^T * E(a) = w^T * 0 = 0 \tag{4.21}$$

because the breeding values a are defined as deviations, there expected value $E(a)$ is always 0. Similarly for the index I, we mentioned that the components in the vector y denoting the information sources that enter the index I are corrected by suitable population means. Due to this correction, we can state that $E(y) = 0$ and thereby $E(I) = 0$. Using these results on the expected values of H and I, we can further develop (4.20)

$$\begin{aligned}\Psi &= var(H) - 2 * cov(H, I) + var(I) \\ &= var(w^Ta) - 2 * cov(w^Ta, b^Ty) + var(b^Ty) \\ &= w^T var(a) w - 2 * w^T cov(a, y^T) b + b^T var(y) b \\ &= w^T C w - 2 * w^T G^T b + b^T P b \end{aligned} \tag{4.22}$$

where C is the variance-covariance matrix of the true breeding values of the traits in the aggregated genotype, G^T is the genetic variance-covariance matrix between the traits in the aggregate genotype and the traits in the index and P is the phenotypic variance-covariance matrix between the traits in the index. Hence we can state

$$\begin{aligned} var(H) &= w^T * C * w \\ cov(H, I) &= w^T * G^T * b \\ var(I) &= b^T * P * b \end{aligned} \tag{4.23}$$

In the objective criterion in (4.19), we stated that Ψ should be minimized. This is done by computing the derivative of Ψ with respect to the vector b. The solution vector b that sets that derivative to 0 corresponds to the solution that we are looking for. The derivative of Ψ with respect to the vector b is also called the gradient and can be computed as

$$\frac{\partial \Psi}{\partial b} = 0 - 2 * w^T * G^T + 2b^TP \tag{4.24}$$

Setting (4.24) to 0 leads to

$$\begin{aligned} 0 &= -2 * w^T * G^T + 2b^TP \\ w^TG^T &= b^TP \\ Pb &= Gw \end{aligned} \tag{4.25}$$

The last line in (4.25) follows by transposing both sides of the second last line and because P is symmetric, $P^T = P$. As a result we obtain the index normal equations which can be solved for the unknown vector b by pre-multiplying both sides with the inversion matrix P^{-1} of P.

$$b = P^{-1}Gw \tag{4.26}$$

Because P is a variance-covariance matrix, it is guaranteed to be positive definite and its inverse P^{-1} does exist.

4.8 Appendix: Derivation of the Index Components for the Example of the Mean Progeny Performance

4.8.1 Variance of Mean Progeny Performance

The mean performance values of a group of progeny for a given parent has the following structure

$$\bar{y_i} = \frac{1}{n}\sum_{k=1}^{n} y_{i,k} \tag{4.27}$$

where y_k is the corrected performance value of progeny k of animal i. Each y_k can be decomposed into

$$\begin{aligned} y_{i,k} &= a_k + e_k \\ &= \frac{1}{2}a_i + \frac{1}{2}a_{d,k} + m_k + e_k \end{aligned} \tag{4.28}$$

The variance (σ_y^2) of a single phenotypic observation ($y_{i,k}$) of progeny k of parent i can be computed as

$$\begin{aligned} \sigma_y^2 = var(y_{i,k}) &= var(\frac{1}{2}a_i + \frac{1}{2}a_{d,k} + m_k + e_k) \\ &= var(\frac{1}{2}a_i) + var(\frac{1}{2}a_{d,k}) + var(m_k) + var(e_k) \\ &= \frac{1}{4}var(a) + \frac{1}{4}var(a_{d,k}) + var(m_k) + var(e_k) \\ &= \frac{1}{4}\sigma_a^2 + \frac{1}{4}var(a_{d,k}) + var(m_k) + var(e_k) \end{aligned} \tag{4.29}$$

In (4.29) we have assumed that all the pairwise covariances between the terms are 0. We define the intra-class correlation t which is the part of the total variance which is attributed to the permanent effect in the single performance records.

$$t = \frac{1/4\sigma_a^2}{\sigma_y^2} = \frac{1}{4}h^2 \tag{4.30}$$

Inserting the decomposition of (4.28) into (4.27) leads to

$$\begin{aligned} \bar{y_i} &= \frac{1}{n}\sum_{k=1}^{n} y_{i,k} \\ &= \frac{1}{n}\sum_{k=1}^{n}(\frac{1}{2}a_i + \frac{1}{2}a_{d,k} + m_k + e_k) \\ &= \frac{1}{2}a_i + \frac{1}{n}\sum_{k=1}^{n}\frac{1}{2}a_{d,k} + \frac{1}{n}\sum_{k=1}^{n} m_k + \frac{1}{n}\sum_{k=1}^{n} e_k \end{aligned} \tag{4.31}$$

Taking the variance on both sides of (4.31) leads to our final result the variance ($\sigma_{\bar{y}}^2$) of the mean progeny performance.

$$
\begin{aligned}
\sigma_{\bar{y}}^2 = var(\bar{y_i}) &= var(\frac{1}{2}a_i + \frac{1}{n}\sum_{k=1}^{n}\frac{1}{2}a_{d,k} + \frac{1}{n}\sum_{k=1}^{n}m_k + \frac{1}{n}\sum_{k=1}^{n}e_k) \\
&= var(\frac{1}{2}a_i) + var(\frac{1}{n}\sum_{k=1}^{n}\frac{1}{2}a_{d,k}) + var(\frac{1}{n}\sum_{k=1}^{n}m_k) + var(\frac{1}{n}\sum_{k=1}^{n}e_k) \\
&= \frac{1}{4}\sigma_a^2 + \frac{1}{4n}var(a_{d,k}) + \frac{1}{n}var(m_k) + \frac{1}{n}var(e_k) \\
&= \frac{1}{4}\sigma_a^2 + \frac{1}{n}\left(\frac{1}{4}var(a_{d,k}) + var(m_k) + var(e_k)\right) \\
&= t * \sigma_y^2 + \frac{1}{n}(1-t) * \sigma_y^2 \\
&= \frac{n * t + 1 - t}{n} * \sigma_y^2 \\
&= \frac{1 + (n-1)t}{n} * \sigma_y^2 \qquad (4.32)
\end{aligned}
$$

Because, we saw earlier that $t = h^2/4$, we can insert that into (4.32) which brings us to the final result

$$
\sigma_{\bar{y}}^2 = \frac{1 + (n-1)h^2/4}{n} * \sigma_y^2 \qquad (4.33)
$$

4.8.2 Covariance between True Breeding Value and Mean Progeny Performance

The set-up of the index normal equations requires the matrix G which corresponds to the genetic covariance between the trait in the aggregate genotype and the information sources in the index. For the example with the mean progeny performance values, the matrix G is defined as

$$
\begin{aligned}
G = cov(a_i, \bar{y_i}) &= cov(a_i, \frac{1}{n}\sum_{k=1}^{n}y_{i,k}) \\
&= cov\left(a_i, \frac{1}{2}a_i + \frac{1}{n}\sum_{k=1}^{n}\left[\frac{1}{2}a_{d,k} + m_k + e_k\right]\right) \\
&= cov(a_i, \frac{1}{2}a_i) \\
&= \frac{1}{2}\sigma_a^2 \qquad (4.34)
\end{aligned}
$$

In (4.34), we have used that the covariance between a_i and all other components of $y_{i,k}$, except a_i is 0.