

## Chapter 8

# Multiple Traits

So far we have seen how to predict breeding values using the BLUP animal model. These breeding values were always only for one trait. From a statistical point of view, experts are calling such analyses **univariate** evaluations. In reality, livestock breeders want to improve their animals in a population with respect to several traits simultaneously. If there are genetic relationships (measured by genetic correlations) between traits, univariate predictions of breeding values do not make optimal use of the available information. This effect is stronger, if certain traits can only be observed in animals of one sex. According to (Mrode, 2005) who cites (Schaeffer, 1984) and (Thompson and Meyer, 1986), the increased accuracy of the evaluations is one of the main advantage of multivariate BLUP analyses. Predictions of breeding values for several traits in a single evaluation is called **multivariate** prediction of breeding values. Such multivariate analyses can be implemented in different ways such as

- combining different corrected information sources in a multivariate selection index (no longer used).
- multivariate prediction of breeding values using BLUP animal model (method of choice).
- combining predicted breeding values from univariate analyses.

Before the introduction of the BLUP animal model, breeding values were estimated using a method that is called **selection index** method. A brief review about selection index theory is given in section 8.3. While selection indices are no longer used to estimate breeding values, selection index theory is still used to predict the aggregate genotype, as will be shown later in this chapter. Before that, we start with an introduction to multivariate BLUP methods.

## 8.1 Multivariate Predictions Of Breeding Values Using BLUP

The prediction of breeding values using a multivariate BLUP model is the optimal prediction procedure. It has advantages, if

- some traits have lower heritability than others
- environmental correlations exist between traits measured on the same animal
- some traits are available only a subset of all animals
- some traits were used for a first round of selection

In principle, a multivariate analysis can be thought of as several univariate analyses which are stacked one on top of the other. Let us assume that we have two traits. For the first trait, we define the model

$$y_1 = X_1\beta_1 + Z_1u_1 + e_1$$

Similarly for the second trait, we define the model

$$y_2 = X_2\beta_2 + Z_2u_2 + e_2$$

If we group the data by traits, then we can write the multivariate model as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}$$

The genetic variance-covariance matrix  $G_0$  for the two traits has the following structure.

$$G_0 = \begin{bmatrix} \sigma_{g_1}^2 & \sigma_{g_1, g_2} \\ \sigma_{g_1, g_2} & \sigma_{g_2}^2 \end{bmatrix} = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix}$$

The inverse  $G_0^{-1}$  of  $G_0$  can be formulated as

$$G_0^{-1} = \begin{bmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{bmatrix}$$

For the random residual effects, the variance-covariance matrix  $R_0$  for the two traits can be written as

$$R_0 = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix}$$

### 8.1. MULTIVARIATE PREDICTIONS OF BREEDING VALUES USING BLUP89

Also the inverse  $R_0^{-1}$  can be written as

$$R_0^{-1} = \begin{bmatrix} r^{11} & r^{12} \\ r^{21} & r^{22} \end{bmatrix}$$

The variance-covariance matrix for the complete set of true breeding values and for all random residuals can be written as

$$\text{var}(u) = \text{var} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A \\ g_{21}A & g_{22}A \end{bmatrix} = G_0 \otimes A = G$$

where  $A$  is the numerator relationship matrix and  $\otimes$  denotes the so-called **Kroneckerproduct** of two matrices. The variance-covariance matrix  $R$  for the random residuals is given by

$$R = \text{var}(e) = \text{var} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} r_{11}I_n & r_{12}I_n \\ r_{21}I_n & r_{22}I_n \end{bmatrix} = R_0 \otimes I_n$$

where  $n$  corresponds to the number of animals in the pedigree. The covariances between all elements of  $a$  and  $e$  are 0. This is denoted by combining both random vectors into a single vector and writing down the variance-covariance matrix of the combined vector as

$$\text{var} \begin{bmatrix} u \\ e \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix} = \text{var} \begin{bmatrix} u_1 \\ u_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} g_{11}A & g_{12}A & 0 & 0 \\ g_{21}A & g_{22}A & 0 & 0 \\ 0 & 0 & r_{11}I_n & r_{12}I_n \\ 0 & 0 & r_{21}I_n & r_{22}I_n \end{bmatrix}$$

The solutions to get estimates of fixed effects and to get predictions for breeding values are obtained from the solutions of mixed model equations. These have exactly the same structure as in the univariate case.

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}, Z = \begin{bmatrix} Z_1 & 0 \\ 0 & Z_2 \end{bmatrix}, \hat{u} = \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \end{bmatrix}$$

Based on the specification of the variance-covariance matrices described earlier, we get  $R^{-1} = R_0^{-1} \otimes I_n$  and  $G^{-1} = G_0^{-1} \otimes A^{-1}$ .

## 8.2 Multitrait Selection

Now that we have predicted breeding values for a collection of traits available, the question is how to integrate these predicted breeding values into a consistent selection criterion. Selection index theory provides a tool to optimally combine different sources of information in order to approximate the aggregate genotype  $H$ . In section 8.2.3, we will return to this topic once again. Although the selection index provides an ideal framework to combine estimated breeding values into an optimal selection criterion, we are going to describe to other selection procedures that are commonly used in practical livestock breeding. The two procedures are

1. Tandem-Selection and
2. Selection based on independent selection thresholds.

The description of these procedures aims at showing the negative consequences that results from the use of these sub-optimal selection strategies.

### 8.2.1 Tandem-Selection

The term **Tandem-Selection** stands for the strategy of always improving the population with respect to just one trait. Once the breeding goal for this trait is reached the population is improved with respect to a different trait. This sequence of single-trait improvements is continued until the breeding goal for all traits is reached.

The problem with Tandem-Selection is that while improving the population for a given trait, the population can only realize correlated selection responses for all other traits. These correlated selection responses might be very small or can even be negative which causes the time that it takes to reach the breeding goal for all relevant traits to be very long.

### 8.2.2 Selection Based On Independent Selection Thresholds

This method was used before the selection index was discovered. This selection procedure is very easy to apply. First selection thresholds are defined for all traits. In the next steps, all animals which are above the thresholds for all traits are selected as parents of the next generation. With this method, selection responses for all traits can be obtained in the early generations after the implementation of this selection strategy.

We are using the following example to show how selection based on independent selection thresholds is used. For reasons of simplicity, we restrict ourselves on two traits. But the results can be generalized without any problems.

### 8.2.2.1 Example: Selection On Independent Thresholds

In a dairy cattle population, breeders want to improve milk yield and protein content. We assume the following selection thresholds for the two traits of interest during first lactation

- milk yield: 6900 kg
- protein content: 3.5 %

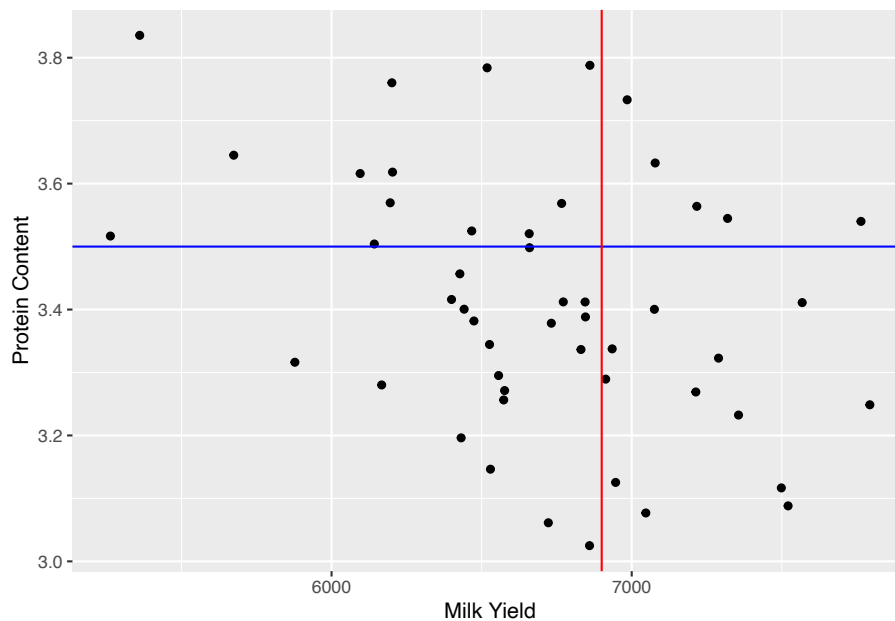


Figure 8.1: Milk Yield and Protein Content For Dairy Cows

Figure 8.1 shows the performance data for a herd of dairy cows. The selection thresholds (red line for milk yield and blue line for protein content) divide the diagram into four quadrants. None of the cows in the lower right quadrant does meet any of the selection criterion imposed by the thresholds. The cows in the upper left quadrant fulfill the requirements for protein content and the cows in the lower right quadrant fulfill the requirements for milk yield. Only the cows in the upper right quadrant fulfill the requirements for both traits.

The disadvantage of this selection strategy becomes apparent with the cows in the upper left and in the lower right quadrant. They are culled and thereby not considered as parents of the next generation even though, they have good performances in one of the traits. From a statistical genetics point of view there are three problems associated with a selection strategy that is based on independent selection thresholds

1. livestock breeders tend to put the thresholds for all traits in the range of positive predicted breeding values. This leads to an exclusion of very many animals and a dramatic reduction in genetic variability
2. genetic relationships between traits are completely ignored. These relationships must be considered when defining selection thresholds. Otherwise the expected genetic gain will not be as expected.
3. differences in the economic relevance of the different traits are completely ignored. Putting the threshold in all traits into the range of positive predicted breeding values leads to a high emphasis on traits with a high heritability. Traits with lower heritability will have only very small selection responses.

### 8.2.3 Selection Index

In section 8.2, we have already briefly described how we can use selection index theory to approximate the aggregate genotype  $H$  in an optimal way. Just as a reminder, the aggregate genotype  $H$  combines all economically relevant traits into a single value using a linear function of the true breeding values  $u$  and taking the economic values  $w$  as weighting factors. Given that  $H$  corresponds to the linear function

$$H = w^T u$$

and we want to approximate  $H$  by an Index  $I$  which is a linear function of all predicted breeding values  $\hat{u}$ , we can write

$$I = b^T \hat{u}$$

where  $b$  is a vector of unknown index weights. The vector  $b$  is determined using the optimality condition of minimum prediction error variance which results in

$$b = P^{-1} G w \tag{8.1}$$

where  $P$  is the variance-covariance matrix between all information sources and  $G$  is the covariance matrix between the information sources and the traits in the aggregate genotype. In case where the traits in the aggregate genotype  $H$  and in the index  $I$  are the same, the matrices  $P$  and  $G$  are defined as

$$P = \text{var}(\hat{u})$$

and

$$G = \text{cov}(u, \hat{u})$$

For predicted breeding values using BLUP, it can be shown that  $cov(u, \hat{u}) = var(\hat{u})$  and therefore  $P = G$ . Using that equality in equation (8.1), we get

$$b = w$$

which means that the vector  $b$  of index weights corresponds to the vector of economic values  $w$ .

The use of the selection index theory to combine predicted breeding values as information sources to approximate the aggregate genotype has the following advantages

- genetic relationships (correlations) between traits in the aggregate genotype are considered correctly
- relationships between information sources in the index are considered correctly
- information from auxiliary traits can be used
- differences in economic relevance of different traits are considered correctly
- expected selection responses can be estimated and thereby quantified

Despite all these advantages, index selection alone is very rarely used in practical livestock breeding. The reason for this is that every population has a few traits that are difficult to associate with an economic value or for some traits it is difficult to come up with genetic parameters. As a consequence of that, in practical livestock breeding we will always find a mix of index selection and a variety of independent selection thresholds.

### 8.3 Review On Selection Index Theory

Before the introduction of the BLUP animal model ((Henderson, 1973b) and (Henderson, 1975)), breeding values were estimated using **Selection Index Theory** ((Hazel, 1943) and (Hazel and Lush, 1942)). Both methods - selection index and BLUP - 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.

#### 8.3.1 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 semen.

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

- 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.

### 8.3.2 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

$$I = \hat{u}_i = b_1y_1 + b_2y_2 + \dots + b_ny_n = b^T y \quad (8.2)$$

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 (8.2) is used as the predicted breeding value  $\hat{u}_i$ . From a statistical point of view equation (8.2) corresponds to a multiple linear regression. The vector of index weights  $b$  are understood as partial regression coefficients.

### 8.3.3 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

$$H = w_1u_1 + w_2u_2 + \dots + w_mu_m = w^T u \quad (8.3)$$

where  $u$  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 ( $\mu_k$ ) 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  $u$  of the single trait.

### 8.3.4 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

$$E(H - I)^2 \rightarrow \min \quad (8.4)$$

The solution to the index construction objective in equation (8.4) leads to the so-called index normal equations which have the following form.

$$Pb = Gw \quad (8.5)$$

where  $P$  is the variance-covariance matrix between all information sources 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

$$b = P^{-1}Gw \quad (8.6)$$

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 (8.7). The terms for  $cov(H, I)$ ,  $\sigma_H$  and  $\sigma_I$  are taken from (8.24) and for  $b$  we insert the solution taken from (8.6).

$$\begin{aligned}
r_{HI} &= \frac{\text{cov}(H, I)}{\sigma_H \sigma_I} \\
&= \frac{w^T * G^T * b}{\sqrt{(w^T * C * w) * (b^T * P * b)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * ((P^{-1} * G * w)^T * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * P * P^{-1} * G * w)}} \\
&= \frac{w^T * G^T * P^{-1} * G * w}{\sqrt{(w^T * C * w) * (w^T * G^T * P^{-1} * G * w)}} \\
&= \sqrt{\frac{w^T * G^T * P^{-1} * G * w}{w^T * C * w}} \\
&= \frac{\sigma_I}{\sigma_H}
\end{aligned} \tag{8.7}$$

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

$$\begin{aligned}
R &= i * r_{HI} * \sigma_H \\
&= i * \frac{\sigma_I}{\sigma_H} * \sigma_H \\
&= i * \sigma_I
\end{aligned} \tag{8.8}$$

where  $i$  is the selection intensity.

### 8.3.5 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  $u$  of an animal based on own phenotypic own performance record  $y$ . Hence we can set

$$H = u \quad \text{and} \quad I = by^*$$

During the index construction, we have assumed the information in the index to be corrected for the appropriate population mean  $\mu$ . 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  $\sigma_y^2$  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  $\sigma_u^2$ . In summary, we have found that

$$\begin{aligned} P &= \sigma_y^2 \\ G &= \sigma_u^2 \\ w &= 1 \end{aligned} \tag{8.9}$$

Inserting the terms of (8.9) into equation (8.6) to compute the index weight  $b$  results in

$$\begin{aligned} b &= P^{-1} * G * w \\ &= \sigma_y^{-2} * \sigma_u^2 * 1 \\ &= \frac{\sigma_u^2}{\sigma_y^2} = h^2 \end{aligned} \tag{8.10}$$

Using the index weight  $b$  found in (8.10) to compute the index  $I$ , we get

$$\begin{aligned} I &= by^* \\ &= h^2(y - \mu) \\ &= \hat{u}_i \end{aligned} \tag{8.11}$$

The index value  $I$  that we obtained in (8.11) 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 (8.11) 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 ( $\hat{u}_i$ ) using selection index theory is computed as shown in (8.7)

$$\begin{aligned}
r_{HI} &= \frac{\sigma_I}{\sigma_H} \\
&= \frac{b\sigma_y}{\sigma_u} \\
&= \frac{h^2\sigma_y}{\sigma_u} \\
&= h
\end{aligned} \tag{8.12}$$

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.

### 8.3.6 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

$$I = b\bar{y}_i \tag{8.13}$$

Because, we are only looking at a single trait, the aggregate genotype  $H$  corresponds to the single true breeding value  $u$  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

$$Pb = Gw \tag{8.14}$$

where  $P$  corresponds to the variance-covariance matrix of the information sources in the index. Our index  $I$  as defined in (8.13) 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

$$\sigma_y^2 = \frac{1 + (n-1)t}{n} \sigma_y^2 \quad (8.15)$$

For our case with the progeny records,  $t$  takes the value of  $\frac{1}{4}h^2$ . For more details on how to compute  $\sigma_y^2$ , see section 8.3.8. Hence the matrix  $P$  reduces to a single number

$$P = \sigma_y^2 = \frac{1 + (n-1)h^2/4}{n} \sigma_y^2 \quad (8.16)$$

The matrix  $G$  in (8.14) is the genetic covariance matrix between the traits in  $H$  and the information sources in  $I$ . In our current example  $G = \text{cov}(u_i, \bar{y}_i) = \frac{1}{2}\sigma_u^2$ . For more details on how to compute  $G$ , see section 8.3.8.2. Now that we have all the components of (8.14), we can insert them and solve for  $b$ .

$$\begin{aligned} \frac{1 + (n-1)h^2/4}{n} \sigma_y^2 * b &= \frac{1}{2} \sigma_u^2 \\ b &= \frac{2nh^2}{4 + (n-1)h^2} \\ &= \frac{2n}{n+k} \end{aligned} \quad (8.17)$$

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

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

$$\hat{u}_i = I = b * (\bar{y}_i - \mu) = \frac{2n}{n+k} * (\bar{y}_i - \mu) \quad (8.18)$$

The accuracy for the predicted breeding value in (8.18) is

$$r_{HI} = \sqrt{\frac{n}{n+k}} \quad (8.19)$$

### 8.3.7 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 (8.4) as

$$\Psi = E(H - I)^2 \rightarrow \min \quad (8.20)$$

The derivation starts by inserting the definitions of  $H$  and  $I$  into (8.20).

$$\begin{aligned}\Psi &= E(H - I)^2 = E(H^2 - 2 * H * I + I^2) \\ &= E(H^2) - 2 * E(H * I) + E(I^2)\end{aligned}\quad (8.21)$$

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

$$E(H) = E(w^T a) = w^T * E(u) = w^T * 0 = 0 \quad (8.22)$$

because the breeding values  $u$  are defined as deviations, their expected value  $E(u)$  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 (8.21)

$$\begin{aligned}\Psi &= \text{var}(H) - 2 * \text{cov}(H, I) + \text{var}(I) \\ &= \text{var}(w^T u) - 2 * \text{cov}(w^T u, b^T y) + \text{var}(b^T y) \\ &= w^T \text{var}(u) w - 2 * w^T \text{cov}(u, y^T) b + b^T \text{var}(y) b \\ &= w^T C w - 2 * w^T G^T b + b^T P b\end{aligned}\quad (8.23)$$

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

$$\begin{aligned}\text{var}(H) &= w^T * C * w \\ \text{cov}(H, I) &= w^T * G^T * b \\ \text{var}(I) &= b^T * P * b\end{aligned}\quad (8.24)$$

In the objective criterion in (8.20), we stated that  $\Psi$  should be minimized. This is done by computing the derivative of  $\Psi$  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  $\Psi$  with respect to the vector  $b$  is also called the gradient and can be computed as

$$\frac{\partial \Psi}{\partial b} = 0 - 2 * w^T * G^T + 2b^T P \quad (8.25)$$

Setting (8.25) to 0 leads to

$$\begin{aligned} 0 &= -2 * w^T * G^T + 2b^T P \\ w^T G^T &= b^T P \\ Pb &= Gw \end{aligned} \tag{8.26}$$

The last line in (8.26) 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  $P$ .

$$b = P^{-1}Gw \tag{8.27}$$

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

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

#### 8.3.8.1 Variance of Mean Progeny Performance

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

$$\bar{y}_i = \frac{1}{n} \sum_{k=1}^n y_{i,k} \tag{8.28}$$

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

$$\begin{aligned} y_{i,k} &= u_k + e_k \\ &= \frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k \end{aligned} \tag{8.29}$$

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



$$\begin{aligned}
\sigma_y^2 = \text{var}(y_{i,k}) &= \text{var}\left(\frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k\right) \\
&= \text{var}\left(\frac{1}{2}u_i\right) + \text{var}\left(\frac{1}{2}u_{d,k}\right) + \text{var}(m_k) + \text{var}(e_k) \\
&= \frac{1}{4}\text{var}(u) + \frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \tag{8.30}
\end{aligned}$$

In (8.30) 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.

$$t = \frac{1/4\sigma_u^2}{\sigma_y^2} = \frac{1}{4}h^2 \tag{8.31}$$

Inserting the decomposition of (8.29) into (8.28) leads to

$$\begin{aligned}
\bar{y}_i &= \frac{1}{n} \sum_{k=1}^n y_{i,k} \\
&= \frac{1}{n} \sum_{k=1}^n \left(\frac{1}{2}u_i + \frac{1}{2}u_{d,k} + m_k + e_k\right) \\
&= \frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k} + \frac{1}{n} \sum_{k=1}^n m_k + \frac{1}{n} \sum_{k=1}^n e_k \tag{8.32}
\end{aligned}$$

Taking the variance on both sides of (8.32) leads to our final result the variance ( $\sigma_{\bar{y}}^2$ ) of the mean progeny performance.

$$\begin{aligned}
\sigma_y^2 = \text{var}(\bar{y}_i) &= \text{var}\left(\frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k} + \frac{1}{n} \sum_{k=1}^n m_k + \frac{1}{n} \sum_{k=1}^n e_k\right) \\
&= \text{var}\left(\frac{1}{2}u_i\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n \frac{1}{2}u_{d,k}\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n m_k\right) + \text{var}\left(\frac{1}{n} \sum_{k=1}^n e_k\right) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{4n}\text{var}(u_{d,k}) + \frac{1}{n}\text{var}(m_k) + \frac{1}{n}\text{var}(e_k) \\
&= \frac{1}{4}\sigma_u^2 + \frac{1}{n} \left( \frac{1}{4}\text{var}(u_{d,k}) + \text{var}(m_k) + \text{var}(e_k) \right) \\
&= t * \sigma_y^2 + \frac{1}{n}(1-t) * \sigma_y^2 \\
&= \frac{n * t + 1 - t}{n} * \sigma_y^2 \\
&= \frac{1 + (n-1)t}{n} * \sigma_y^2
\end{aligned} \tag{8.33}$$

Because, we saw earlier that  $t = h^2/4$ , we can insert that into (8.33) which brings us to the final result

$$\sigma_y^2 = \frac{1 + (n-1)h^2/4}{n} * \sigma_y^2 \tag{8.34}$$

### 8.3.8.2 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

$$\begin{aligned}
G = \text{cov}(u_i, \bar{y}_i) &= \text{cov}\left(u_i, \frac{1}{n} \sum_{k=1}^n y_{i,k}\right) \\
&= \text{cov}\left(u_i, \frac{1}{2}u_i + \frac{1}{n} \sum_{k=1}^n \left[\frac{1}{2}u_{d,k} + m_k + e_k\right]\right) \\
&= \text{cov}\left(u_i, \frac{1}{2}u_i\right) \\
&= \frac{1}{2}\sigma_u^2
\end{aligned} \tag{8.35}$$

In (8.35), we have used that the covariance between  $u_i$  and all other components of  $y_{i,k}$ , except  $u_i$  is 0.