Chapter 4

Best Linear Unbiased Prediction (BLUP)

The prediction of breeding values requires to correct the information sources for an appropriate comparison value. So far we have referred to that comparison value as the population mean and we have assumed this correction value to be known. In reality, the computation of these comparison values is a difficult problem. This problem is one of the reasons that nowadays the predictions of all breeding values are based on a method that is called **BLUP**. In this chapter, we first want to have a closer look at the problem of computing these correction factors with which the information sources must be adjusted. After that, the BLUP method will be introduced.

4.1 Problem of Correction

In theory, the population mean is the ideal correction value for all information sources. From our standard model we can derive

$$y = \mu + u + e \qquad \rightarrow \qquad \bar{y} = \bar{\mu} + \bar{u} + \bar{e} = \mu \tag{4.1}$$

Because, we defined the true breeding value u and the non-identifiable environmental effects e as deviations from a common mean, the average effect of all identifiable environmental components is captured by the population mean μ. But this is only true in an idealized population where all selection candidates are kept in the same environment and where they deliver their performances at the same time. In real world scenarios, this is unrealistic, because e.g. own performance values and progeny performances cannot be delivered at the same time. Furthermore, selection candidates are kept in different herds in different

environments. All these factors do have an influence on the performance of the recorded animals and hence on the predicted breeding values. But good methods for predicting breeding values should be able to correct for such environmental influences. If that is not the case, environmental factors will **bias** the predicted breeding values. To avoid such biases, performance records were subdivided into environmental classes. In dairy cattle such classes were formed based on herds, calving year, calving season and age at first calving. In pigs, performance records might be divided into herds, years and fattening batches. From now on, we call the combination of these environmental effects on the performance records as **identifiable systematic fixed effects**. For the prediction of breeding values, we assume that these fixed effects in a given comparison class have all the same influence on the performance of the animals that are in the same class. Hence if we group all animals who show the same levels of all fixed effects into one comparison class, any biases from the identifiable environment can be avoided.

The more environmental factors can be considered in forming the comparison classes, the better we can correct our performance records for the environmental effects. But when the number of environmental factors increases the number of animals per comparison class decreases. From the statistical point of view, the small number of observations in comparison classes reduce the accuracy with which the environmental fixed effects can be estimated. With smaller comparison groups, the risk that the average breeding value of animals in such a comparison is not zero increases. In case the average breeding value in a comparison group is not zero, predicted breeding values show a deviation which is called **bias**. The occurrence of bias can be shown as follows. Let us assume the average performance of all animals in a comparison group (CG) to be $\bar{y}_{CG}$:

$$\bar{y}_{CG} = \mu + \bar{u}_{CG} + \bar{e}_{CG} \tag{4.2}$$

In case the average breeding value $\bar{u}_{CG}$ is zero, the population mean μ measures the average identifiable environment effect. If $\bar{u}_{CG}$ is not zero, then the predicted breeding value $\hat{u}_i$ using an older method called selection index, the index value I corresponds to

$$\begin{aligned} I &= b(y_i - (\mu + \bar{u}_{CG})) \\ &= b(y_i - \mu) - b\bar{u}_{CG} \\ &= \hat{u}_i - b\bar{u}_{CG} \end{aligned} \tag{4.3}$$

The first term in the result of (4.3) corresponds to the predicted breeding value where the second term measures the **bias**. This depends on the average breeding values of the animals of the comparison group. If the average breeding value of all animals in the comparison group is zero, then the predicted breeding value from (4.3) is unbiased. Because we have to know the breeding values

of the animals in the comparison group to get an unbiased prediction of the breeding value for a given animal and the breeding values of the animals in the comparison group must also be predicted, this consists of a "chicken-and-egg" problem which cannot be solved.

The solution to this was presented by Charles R. Henderson in several publications ((Henderson, 1973)) and (Henderson, 1975)). The key idea behind the solution is to estimate the identifiable environmental factors as fixed effects and to predict the breeding values as random effects simultaneously in a linear mixed effects model. The properties of the methodology developed by Henderson are similar to those of the selection index method. But the main advantage of Henderson's methodologies is that phenotypic records do not need to be corrected before breeding values can be predicted. But the effects of the identifiable environmental factors are also a result which come out of the analysis. The methodology developed by Henderson is called **BLUP** and the properties of this methodology are directly incorporated into the name where

- **B** stands for **best** which means that the correlation between the true (u) and the predicted breeding value ($\hat{u}$) is maximal or the prediction error variance ($var(u - \hat{u})$) is minimal.
- **L** stands for **linear** which means the predicted breeding values are linear functions of the observations (y)
- **U** stands for **unbiased** which means that the expected values of the predicted breeding values are equal to the true breeding values
- **P** stands for **prediction**

BLUP based approaches have found widespread usage in genetic evaluations. They are used for both traditional predictions of breeding values and also for predicting genomic breeding values. The popularity of BLUP is not only due to the theoretical foundations behind BLUP, but Henderson has also developed efficient algorithms to be able to compute predicted breeding values for very large livestock breeding populations. The theoretic foundations, the development of efficient algorithms together with the availability of large computational resources at a very low price have made BLUP to become the de-facto standard methodology for predicting breeding values.

4.2 Numeric Example

We want to use a concrete numeric example of a small population to explain how breeding values are predicted using the BLUP methodology. The phenotypic observations consist of measurements of the trait **weaning weight** in beef cattle. Table 4.1 gives an overview of the dataset.

Table 4.1: Example Data Set for Weaning Weight in Beef Cattle

Animal	Sire	Dam	Herd	Weaning Weight
12	1	4	1	2.61
13	1	4	1	2.31
14	1	5	1	2.44
15	1	5	1	2.41
16	1	6	2	2.51
17	1	6	2	2.55
18	1	7	2	2.14
19	1	7	2	2.61
20	2	8	1	2.34
21	2	8	1	1.99
22	2	9	1	3.10
23	2	9	1	2.81
24	2	10	2	2.14
25	2	10	2	2.41
26	3	11	2	2.54
27	3	11	2	3.16

We assume the phenotypic variance (σ_p^2) to be 0.1014 and the heritability (h^2) corresponds to 0.25.

4.3 Linear Mixed Effects Model

A simple linear model contains fixed effects such as *herd* or *sex* of an animal and tries to explain the observations as linear functions of such effects. Because the effects considered in a model cannot account for all influences of a given set of observations, every model must have a random residual component. If a linear model contains besides the residuals any additional random effects, then this model is called a **mixed linear effects model**.

4.3.1 Fixed Versus Random Effects

Unfortunately, there is no unique and generally accepted definition of which effects should be fixed and which should be random. There are generally accepted guidelines of how to classify effects as fixed or as random. Table 4.2 lists a few criteria that might be helpful.

Table 4.2: Classification Factors of Fixed and Random Effects

fixed effect	random effects
classes can be defined exactly	realized value come from an underlying distribution
the value of a class does not have an apriori expected value	each realization is unique
values are exactly estimable	observations are influenced by the variance of the random effect
the expected value of a class effect is of primary interest	main interest is on the variance not on the expected value
fixed effects can be corrected for	

Certain factors such as herd, sex, breed or feeding regimes can be classified unambiguously as fixed effects. On the other hand breeding values are always random effects. Because, we know that breeding values have an expected value of 0 and have a certain variance, they must be modeled as random effects where these properties can be integrated into the model. Furthermore, each animal has a different realization of a breeding value. Exceptions are mono-clonal twins and clones.

From a practical point of view, the software program that is used to analyse the data has also an influence on whether a certain effect is treated as fixed or as random. If a certain effect has very many levels such as herds, then it is sometimes better for the analysis to treat such an effect as random.

4.3.2 Model Specification

In a linear mixed effects model a single observation y_{ijk} is decomposed according to equation (4.4)

$$y_{ijk} = \beta_i + u_j + e_{ijk} \tag{4.4}$$

where β_i stands for the $i{-}^{th}$ level of a fixed effect, u_j is the $j-th$ realization of the random effect u and e_{ijk} is the residual effect of the $k{-}^{th}$ observation}. Because, we do not want to model just one observation, but we want to include all observations of a complete population, it is helpful to convert the model in (4.4) into matrix-vector notation. This is shown in equation (4.5)

$$y = X\beta + Zu + e \tag{4.5}$$

where

y	vector of length n of all observations
β	vector of length p of all fixed effects
X	$n \times p$ design matrix linking the fixed effects to the observations
u	vector of length n_u of random effects
Z	$n \times n_u$ design matrix linking random effect to the observations
e	vector of length n of random residual effects.

Furthermore, we assume the following relations for the expected values and for the variances. As already mentioned the random effects are defined as deviations and hence their expected value is set to zero.

$$E(u) = 0 \qquad \text{and} \qquad E(e) = 0 \tag{4.6}$$

From this it follows that $E(y) = X\beta$. The variance-covariance matrices for the random effects are set to

$$var(u) = G \qquad \text{and} \qquad var(e) = R \tag{4.7}$$

Under the assumption that $cov(u, e^T) = 0$, we can compute $var(y) = Z * var(u) * Z^T + var(e) = ZGZ^T + R = V$.

In model (4.5) the vectors β and u are unknown. The solution of the model (4.5) for the unknowns β and u leads to estimates $\hat{\beta}$ for the fixed effects β and for predicted random effects $\hat{u}$. Unlike with the selection index, with BLUP, we do not have to correct the observations before predicting random effects.

4.3.3 The Solution

An outline of how to derive the BLUP solutions for $\hat{\beta}$ and $\hat{u}$ will be given in an Appendix. The details of this derivation are not important. Therefore, we are presenting here directly the result which are

$$\hat{u} = GZ^T V^{-1}(y - X\hat{\beta}) \tag{4.8}$$

We call $\hat{u}$ the best linear unbiased prediction of u or shorter $\hat{u} = BLUP(u)$. For $\hat{\beta}$, we insert the generalized least squares estimator (GLS) which corresponds to

$$\hat{\beta} = (X^T V^{-1} X)^- X^T V^{-1} y \tag{4.9}$$

The matrix $(X^T V^{-1} X)^-$ denotes the generalized inverse of the matrix $(X^T V^{-1} X)$. The generalized inverse K^- can be replaced with the simple inverse K^{-1}, whenever the columns of matrix K are linearly independent[1].

[1] For our examples that are shown here, we can always use the simple inverse.

Analogously to $\hat{u}$, $\hat{\beta}$ is called the best linear unbiased estimator of the fixed effects β. In short, we can state $\hat{\beta} = BLUE(\beta)$.

4.3.4 Mixed Model Equations

The solutions shown in (4.8) for $\hat{u}$ and in (4.9) for $\hat{\beta}$ are not suitable for practical purposes. Both solutions contain the inverse V^{-1} of matrix V. The matrix V corresponds to the variance-covariance matrix of all observations y. The inverse matrix V^{-1} is not easy to compute and furthermore procedures to invert general matrices are computationally expensive and are prone to rounding errors. In one of his many papers, Henderson has shown that the results for $\hat{u}$ and $\hat{\beta}$ are the same when solving the following system of equations simultaneously.

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{4.10}$$

The above shown equations are called **mixed model equations** (MME). They do no longer contain the inverse V^1 and hence these MME are much simpler to solve. The MME contain the inverses R^{-1} and G^{-1}, but we will see with concrete examples that they are much easier to invert. As a consequence, whenever we have to predict breeding values using BLUP, we will use the mixed model equations shown in (4.10).

4.4 Sire Model

The application of the linear mixed effects model from (4.5) to the numerical example in table 4.1. As random effects u we are taking the father s of each animal i with an observation. As fixed effects β we are using the herd effect. When fathers are modeled as random effects, then we call this model a **sire model**. Setting up a sire model for the data in table 4.1 looks as follows

$$\begin{bmatrix} 2.61 \\ 2.31 \\ 2.44 \\ 2.41 \\ 2.51 \\ 2.55 \\ 2.14 \\ 2.61 \\ 2.34 \\ 1.99 \\ 3.1 \\ 2.81 \\ 2.14 \\ 2.41 \\ 2.54 \\ 3.16 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ s_3 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \\ e_{11} \\ e_{12} \\ e_{13} \\ e_{14} \\ e_{15} \\ e_{16} \end{bmatrix}$$

Besides the equation for the sire model we also have to specify the expected values and the variances of all random components. To be able to distinguish the sire model from the general linear mixed effects model, we usually call the random sire effect s and no longer u. The expected values for the random variables were already stated when discussing the general linear mixed effects model in section 4.3.2. Hence

$$E(s) = 0 \qquad \text{and} \qquad E(e) = 0 \qquad \rightarrow \qquad E(y) = X\beta \tag{4.11}$$

For the variances there are a few simplifications that we can use in our sire model. The covariance between the random effects s and e are assumed to be 0. The covariances among the single residual effects are also assumed to be 0. Hence, the variance-covariance matrix of the residual effects are $var(e) = I * \sigma_e^2$. The variance of the sire effects s is

$$var(s) = A_s * \sigma_s^2 = G$$

where A_s is the additive genetic relationship matrix between the sires. We will be deriving the matrix A_s in a later chapter. Because our sires are not related, we can say that $A_s = I$ and hence

$$G = I * \frac{\sigma_u^2}{4}$$

Now we are ready to set up the mixed model equations from (4.10) for the sire model. The computation of the numerical solutions from the mixed model equations will be the topic of an exercise.

4.5 Animal Model

The mixed model equations are a universal tool to find BLUPs of random effects and BLUEs of fixed effect simultaneously. On the other hand it is not satisfactory that with the sire model only sires obtain predicted breeding values. All information that is known about the mothers was completely ignored when we specified the sire model. A better approach would be to combine all available information from a given population. This can be done by replacing in the sire model the random sire effects by random animals effects. As a result each animal in the dataset receives a random effect which models its breeding value. This type of model is called an **animal model**. Because the animal model has the breeding values of all animals as random effects, they are often referred to with the variable or the vector a[2] and no longer s as in the sire model. The variance-covariance matrix ($var(a)$) between all animal effects is proportional to the additive genetic relationship matrix A among all animals. We will see in a later chapter how to compute the matrix A.

[2]This is not the same as the genotypic value in a single locus model.