Chapter 6

Additional Aspects of BLUP

This chapter introduces interesting additional aspects and special properties of BLUP-based predicted breeding values. As we have seen in chapter 4, predicted breeding values which result from solving Henderson's mixed model equations are predictions and these predictions always depend on some assumptions. These assumptions are more or less valid depending on the dataset that is analysed to produce the results. Furthermore, predicted breeding values are a function of recorded data and such data is never perfect. Therefore, we need a measure to quantify how good our predictions are. Such a measure is the **accurracy** of the predicted breeding values.

One of the reasons, BLUP is nowadays the method of choice for predicting breeding values is the fact that in the BLUP animal model all available information is used. This property can be shown by decomposing the predicted breeding values from an animal model.

6.1 Accurracy

The accuracy for a BLUP-based animal model is no longer as easily derived as with the prediction of breeding values based on own-performance or progeny records. The animal model is a linear mixed effects model containing fixed and random effects. Due to the properties of BLUP-based methods, the estimates of the fixed effects and the prediction of the random effects have minimum error variance. For the fixed effects, this error variance can be computed as

$$var(\beta - \hat{\beta}) = var(\hat{\beta})$$

because fixed effects β do not have any variance. For the random effects u the prediction error variance (PEV) does not simplify to the variance of the predicted effects $\hat{u}$. Random effects by their nature do have a certain variance which is part of the model specification. For a BLUP animal model the variance of the random effects u correspond to $var(u) = A * \sigma_u^2$. Any meaningful prediction $\hat{u}$ of a random effect u should also satisfy that the variance $var(\hat{u})$ predicts $var(u)$ as closely as possible. Following this argument $var(\hat{u})$ cannot correspond to the prediction error variance. The prediction error variance $PEV(\hat{u})$ is computed as

$$PEV(\hat{u}) = var(u - \hat{u}) = var(u) + var(\hat{u}) - 2 * cov(u, \hat{u}) = var(u) - var(\hat{u})$$

Henderson found that $PEV(\hat{u})$ depends on the inverse of the coefficient matrix in the mixed model equations.

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + U^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} C^{11} & C^{12} \\ C^{21} & C^{22} \end{bmatrix}$$

We can state that

$$PEV(\hat{u}) = var(u - \hat{u}) = var(u) - var(\hat{u}) = C^{22} \tag{6.1}$$

For a single animal i, the prediction error variance is $PEV(\hat{u}_i) = C_{ii}^{22}$ where C_{ii}^{22} is the i-th diagonal element in the matrix C^{22}. The accuracy of $\hat{u}_i$ is measured by the squared correlation $r_{u,\hat{u}}^2$ between true and predicted breeding value. This correlation is defined as

$$r_{u,\hat{u}} = \frac{cov(u_i, \hat{u}_i)}{\sqrt{var(u_i) * var(\hat{u}_i)}} = \frac{var(\hat{u}_i)}{\sqrt{var(u_i) * var(\hat{u}_i)}} = \sqrt{\frac{var(\hat{u}_i)}{var(u_i)}} \tag{6.2}$$

Combining equations (6.2) and (6.1) by solving both for $var(\hat{u}_i)$ leads to

$$\begin{aligned} var(\hat{u}_i) &= var(u_i) - C_{ii}^{22} \\ var(\hat{u}_i) &= r_{u,\hat{u}}^2 * var(u_i) \\ PEV(\hat{u}_i) = C_{ii}^{22} &= var(u_i) - r_{u,\hat{u}}^2 * var(u_i) = (1 - r_{u,\hat{u}}^2) * var(u_i) \end{aligned} \tag{6.3}$$

Solving equation (6.3) for $r_{u,\hat{u}}^2$ which is the measure commonly used to assign a certain level of accuracy to the predicted breeding value $\hat{u}_i$ of a given animal i.

$$r_{u,\hat{u}}^2 = 1 - \frac{C_{ii}^{22}}{var(u_i)} = 1 - \frac{PEV(\hat{u}_i)}{var(u_i)} \tag{6.4}$$

From equation (6.4) it becomes clear that the smaller $PEV(\hat{u}_i)$ is the higher the accuracy $r^2_{u,\hat{u}}$ is. In the limit where $PEV(\hat{u}_i)$ tends to 0, the accuracy will tend to 1. Based on the definition of $PEV(\hat{u}_i)$ in (6.1), it can be seen that $PEV(\hat{u}_i)$ tends to 0, if $var(\hat{u}_i)$ tends towards $var(u_i)$. That means the better the variance $var(\hat{u}_i)$ of the predicted breeding values $\hat{u}_i$ approximates the variance $var(u_i)$, the smaller the value for $PEV(\hat{u}_i)$ and the higher the accuracy $r^2_{u,\hat{u}}$ of the predicted breeding value $\hat{u}_i$ will be. On the other hand, if $var(\hat{u}_i)$ tends to 0 which means the prediction of $var(u_i)$ by $var(\hat{u}_i)$ is very poor, $PEV(\hat{u}_i)$ tends to $var(u_i)$ and the accuracy $r^2_{u,\hat{u}}$ tends to its minimum which is 0.

6.2 Confidence Intervals of Predicted Breeding Values

The prediction error variance (PEV) determines the confidence interval of the predicted breeding values. The square root of PEV corresponds to the standard error of prediction (SEP).

$$SEP(\hat{u}_i) = \sqrt{PEV(\hat{u}_i)} = \sqrt{(1 - r^2_{u,\hat{u}}) * var(u_i)}$$

Assuming the predicted breeding values $\hat{u}$ follow a normal distribution and SEP gives a measure of how much the predictions vary. For a given error probability (α) the confidence interval can be derived for probability of $1 - \alpha$. For a given genetic standard deviation σ_u of 12, an error probability of $\alpha = 0.05$ and range of accuracies, the width of the confidence intervals can be computed. The results of these interval widths are shown in Table 6.1.

Table 6.1: Widths of Confidence Intervals for Given Accuracies

Accurracy	Interval Width
0.40	36.44
0.50	33.26
0.60	29.75
0.70	25.76
0.80	21.04
0.90	14.88
0.95	10.52
0.99	4.70

For a given predicted breeding value of 100 and an accuracy of 0.99 the confidence interval is 100 ± 2.35. The same confidence interval is also shown in Figure 6.1.

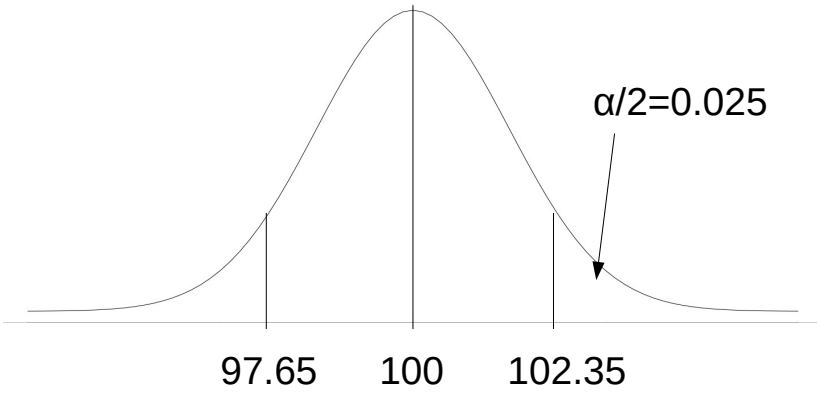

Figure 6.1: Confidence Interval of Predicted Breeding Value

6.3 Relevance of Accurracies

The relevance that is assigned to the accuracies of the predicted breeding values depends on the livestock species and also on the individual breeder. The assessment of the importance of the accuracies is not always easy and is different whether we are looking at a single animal or whether we are looking at a population.

Predicted breeding values are unbiased, hence low accuracies are not considered to be something "bad". For single animals with predicted breeding values with low accuracies, their predicted breeding value is expected to change more. But the change of the predicted breeding values can be in both directions. Because most breeders want to avoid negative changes, high accuracies are taken to be important.

6.4 Response to Selection

The classical definition of accuracy as described above is the correlation $r_{u_i,\hat{u}_i}$ for a single animal i across conceptual repeated sampling. This correlation is a measure of the expected change of a predicted breeding value for animal i with increasing information. Together with the link of this correlation to the standard error of prediction (SEP) of the predicted breeding value $\hat{u}_i$, the quantity $r_{u_i,\hat{u}_i}$ can also be used to make statements about the potential risk of producing offspring with undesired characteristics, when using animal i as a parent.

Accuracies are also important to predict genetic progress in a selection scheme. This use applies only to large unrelated populations and was suitable for selection programs that were based on selection index procedure for determining parents of a future generation. However for a joint analysis of a complete population, the relevant measure according to [Bijma, 2012] is the correlation between true and predicted breeding values in the selection candidates. This correlation is a property of a population and not of a single individual. More details on how to estimate this "population accuracy" which should be used in the prediction of selection response is described in [Legarra and Reverter, 2018].

In general, the following dependencies of a desired increase in population accuracy to other parameters in the breeding program can be established.

- generation intervals increase, because we need to wait for more progeny to deliver a performance record
- more progeny per selection candidate must be tested, hence the number of selection candidates and the selection intensities decrease
- costs for testing animals increase.

For livestock species such as cattle and horses, breeders usually assign too much relevance to accuracies. In general selection response could be increased by lowering the generation interval and increasing the selection intensities and thereby accepting lower levels of accuracies.

6.5 Decomposition of Predicted Breeding Value

The mixed model equations as they are shown in (4.10) can be written in the following abbreviated form

$$M * s = r$$

where

M	coefficient matrix
s	vector of unknowns
r	vector of right-hand sides

The vector s of unknowns in the mixed model equations consists of the vector $\hat{\beta}$ of estimates of fixed effects and the vector $\hat{u}$ of predicted breeding values, which means

$$s = \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix}$$

Because the vector $\hat{\beta}$ has length p, the first p components in s correspond to estimates of fixed effects. The remaining q components of s correspond to the q predicted breeding values of vector $\hat{u}$. Let us assume that we want to have a closer look at how the predicted breeding value $\hat{u}_i$ of the animal at position i in the vector $\hat{u}$. The component $\hat{u}_i$ can be found on position $p+i$ in the vector s. As a consequence of that the $(p+i)$-th line in M contains the coefficients that are relevant for the computation of the predicted breeding value $\hat{u}_i$. These coefficients determine what type of information is used to compute $\hat{u}_i$. In what follows, we describe how these coefficients are determined.

For the decomposition, we are using a simpler model which is shown in (6.5)

$$y_i = \mu + u_i + e_i \tag{6.5}$$

where

y_i	Observation for animal i
u_i	breeding value of animal i with a variance of $(1 + F_i)\sigma_u^2$
e_i	random residual effect with variance σ_e^2
μ	single fixed effect

The above defined model is used to analyse a dataset in which all animals have an observation. Animal i has parents s and d and n progeny k_j (with $j = 1, \ldots, n$) and n mates l_j (with $j = 1, \ldots, n$). From this it follows that progeny k_j has parents i and l_j.

For this simple model (6.5) the mixed model equations also have a reduced complexity. Because, we only have one fixed effect which is present in all observations the matrix X has just one column of all ones. Because all animals have an observation, the matrix Z corresponds to the identity matrix.

Taking into account Henderson's rule for setting up A^{-1} directly, the equation for observation y_i which corresponds to the $(i+1)$-th[1] equation in our mixed effects model.

$$y_i = \hat{\mu} + \left[1 + \alpha\delta^{(i)} + \frac{\alpha}{4}\sum_{j=1}^{n}\delta^{(k_j)}\right]\hat{u}_i - \frac{\alpha}{2}\delta^{(i)}\hat{u}_s - \frac{\alpha}{2}\delta^{(i)}\hat{u}_d - \frac{\alpha}{2}\sum_{j=1}^{n}\delta^{(k_j)}\hat{u}_{k_j} + \frac{\alpha}{4}\sum_{j=1}^{n}\delta^{(k_j)}\hat{u}_{l_j} \tag{6.6}$$

where α ration between variance components σ_e^2/σ_u^2
$\delta^{(j)}$ contribution for animal j to A^{-1}

Solving (6.6) for $\hat{u}_i$ leads to

$$\hat{u}_i = \frac{1}{1 + \alpha\delta^{(i)} + \frac{\alpha}{4}\sum_{j=1}^{n}\delta^{(k_j)}}\left[y_i - \hat{\mu} + \frac{\alpha}{2}\left\{\delta^{(i)}(\hat{u}_s + \hat{u}_d) + \sum_{j=1}^{n}\delta^{(k_j)}(\hat{u}_{k_j} - \frac{1}{2}\hat{u}_{l_j})\right\}\right] \tag{6.7}$$

From the decomposition in (6.7), we can see that the predicted breeding value $\hat{u}_i$ consists of the following components

- Predicted breeding values $\hat{u}_s$ and $\hat{u}_d$ of parents s and d of i
- Own performance y_i of i
- Predicted breeding values $\hat{u}_{k_j}$ and $\hat{u}_{l_j}$ of progeny k_j and mates l_j

An explicit example of a decomposition in (6.7) will be used as an exercise problem.

[1] For the general case, this would be $(p+i)$-th equation. In the simple example, we have $p = 1$.