Chapter 6

Prediction Of Breeding Values

The breeding value u_i of animal i cannot be observed or measured. It has to be estimated from phenotypic observations using the genetic model shown in (6.1) and in .

$$y_i = \mu + u_i + d_i + i_i + e_i \tag{6.1}$$

where y_i is the observation of animal i, μ corresponds to a general population mean of the observed phenotypic values, u_i is the breeding value of animal i, d_i is the dominance deviation and i_i the epistasis effect. The random error term is symbolized with e_i. For the prediction of breeding values, dominance deviations and epistasis effects are not considered to be important. Therefore they are ignored. This leads to a simpler model shown in (6.2).

$$y_i = \mu + u_i + e_i \tag{6.2}$$

The expected values and the variance-covariance matrix of the model shown in (6.2) are specified as

$$\begin{aligned} E\begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} &= \begin{bmatrix} \mu \\ 0 \\ 0 \end{bmatrix} \\ var\begin{bmatrix} y_i \\ u_i \\ e_i \end{bmatrix} &= \begin{bmatrix} \sigma_y^2 & \sigma_u^2 & \sigma_e^2 \\ \sigma_u^2 & \sigma_u^2 & 0 \\ \sigma_e^2 & 0 & \sigma_e^2 \end{bmatrix} \end{aligned} \tag{6.3}$$

6.1 Genetic Properties of Predicted Breeding Values

As already mentioned, breeding values cannot be observed but must be estimated or predicted from phenotypic data. Prediction or estimation of an unknown parameter using statistical modelling expresses the predicted or estimated quantity as a mathematical function of the observed data. The question is how this function should look like and what properties the predicted breeding values should fulfill. In the case of livestock breeding, one objective for the predicted breeding values is that the response to selection is maximized.

[Henderson, 1963] found that the improvement of an offspring generation compared to the parent generation can be maximized when parents are selected based on the conditional expected value ($E(u|y)$) of the true breeding value u given the observed phenotypic values y. Under the assumption of multivariate normality for u and y, the conditional expected value ($E(u|y)$) can be written as

$$E(u|y) = E(u) + cov(u, y^T) * var(y)^{-1} * (y - E(y)) \tag{6.4}$$

Applying the expression in (6.4) for the prediction of breeding values, we can use the property that the breeding value was defined as deviation from the general mean which means that the expected value $E(u)$ of the true breeding value u is $E(u) = 0$.

$$\begin{aligned} \hat{u} = E(u|y) &= 0 + cov(u, y^T) * var(y)^{-1} * (y - E(y)) \\ &= cov(u, y^T) * var(y)^{-1} * (y - \mu) \end{aligned} \tag{6.5}$$

Equation (6.5) shows that the predicted breeding value consists of two parts.

1. $(y - \mu)$ shows that the observed phenotypic values are corrected for the fixed non-genetic environmental effects represented by μ.
2. $cov(u, y^T) * var(y)^{-1}$ corresponds to a weighting factor with which the corrected phenotypic values are multiplied. The weighting factor depends on population parameters such as the heritability and the genetic correlation.

Based on the definition of the predicted breeding value $\hat{u}$ given in (6.5), the following properties can be derived.

6.1.1 Unbiasedness

The expected value ($E(\hat{u})$) of the predicted breeding value $\hat{u}$ can be computed as:

$$\begin{aligned} E(\hat{u}) &= E(cov(u, y^T) * var(y)^{-1} * (y - E(y))) \\ &= cov(u, y^T) * var(y)^{-1} * E(y - E(y)) \\ &= cov(u, y^T) * var(y)^{-1} * (E(y) - E(y)) = 0 \end{aligned} \tag{6.6}$$

Because we have already specified that $E(u) = 0$, it follows that $E(\hat{u}) = E(u) = 0$. This means that $\hat{u}$ is an unbiased estimator of u.

6.1.2 Variance

The variance $var(\hat{u})$ is the same as the covariance $cov(u, \hat{u})$ between the true and predicted breeding value.

$$\begin{aligned} var(\hat{u}) &= var(cov(u, y^T) * var(y)^{-1} * (y - E(y))) \\ &= cov(u, y^T) * var(y)^{-1} * var(y - E(y)) * var(y)^{-1} * cov(y, u^T) \\ &= cov(u, y^T) * var(y)^{-1} * cov(y, u^T) \\ cov(u, \hat{u}) &= cov(u, (cov(u, y^T) * var(y)^{-1} * (y - E(y)))^T) \\ &= cov(u, (y - E(y))^T) * var(y)^{-1} * cov(y, u^T) \\ &= cov(u, y^T) * var(y)^{-1} * cov(y, u^T) = var(\hat{u}) \end{aligned} \tag{6.7}$$

6.1.3 Accuracy

The accuracy of a predicted breeding value is measured by the correlation $r_{u,\hat{u}}$ between true and estimated breeding value. This correlation is computed as

$$r_{u,\hat{u}} = \frac{cov(u, \hat{u})}{\sqrt{var(u) * var(\hat{u})}} = \sqrt{\frac{var(\hat{u})}{var(u)}} \tag{6.8}$$

Alternatively, the reliability B which corresponds to the square of the accuracy is often specified when predicted breeding values are published.

6.1.4 Conditional Density

In some cases, e.g., for specifying confidence intervals of true breeding values, it might be interesting to have a look at the conditional density $f(u|\hat{u})$. This density is a multivariate normal density with expected value $E(u|\hat{u})$ and variance $var(u|\hat{u})$. These values can be computed based on the theory of conditional multivariate normal densities.

$$
\begin{aligned}
E(u|\hat{u}) &= E(u) + cov(u, \hat{u}^T) * var(\hat{u})^{-1} * (\hat{u} - E(\hat{u})) = \hat{u} \\
var(u|\hat{u}) &= var(u) - cov(u, \hat{u}^T) * var(\hat{u})^{-1} * cov(\hat{u}, u^T) \\
&= var(u) * \left[1 - \frac{cov(u, \hat{u}^T)^2}{var(u) * var(\hat{u})}\right] \\
&= var(u) * \left[1 - r^2_{u,\hat{u}}\right] \qquad (6.9)
\end{aligned}
$$

6.1.5 Prediction Error Variance (PEV)

Because every prediction is associated with an error, the same is true for the predicted breeding values $\hat{u}$. The variability of the error for the predicted breeding values are quantified by the prediction error variance (PEV). This is computed as

$$
\begin{aligned}
var(u - \hat{u}) &= var(u) - 2cov(u, \hat{u}) + var(\hat{u}) = var(u) - var(\hat{u}) \\
&= var(u) * \left[1 - \frac{var(\hat{u})}{var(u)}\right] \\
&= var(u) * \left[1 - r^2_{u,\hat{u}}\right] \qquad (6.10)
\end{aligned}
$$

As shown in subsection 6.1.6, the standard error of prediction (SEP) can be a useful quantity. SEP corresponds just to the square root of PEV. Hence

$$
\begin{aligned}
SEP(\hat{u}) = \sqrt{var(u - \hat{u})} &= \sqrt{var(u) * \left[1 - r^2_{u,\hat{u}}\right]} \\
&= \sqrt{1 - r^2_{u,\hat{u}}} * \sigma_u \qquad (6.11)
\end{aligned}
$$

with $\sigma_u = \sqrt{var(u)}$.

6.1.6 Confidence Intervals

The confidence interval (CI) must always be associated with a certain error-level which is usually denoted by α. Typical values of α might be 0.05 or 0.01. For $\alpha = 0.05$, we can specify the 95%-confidence interval. Within this interval around the expected value, the enclosed surface by the density curve corresponds to 0.95.

In the context of predicted breeding values, the confidence interval for a specified error-level α of the true breeding value u given a predicted breeding value $\hat{u}$ might be of interest for certain quantification of risks when using parents with predicted breeding values with different accuracy values.

To get the desired confidence interval, we have a look at the conditional density described in 6.1.4. An example of such a density is shown in the diagram below.

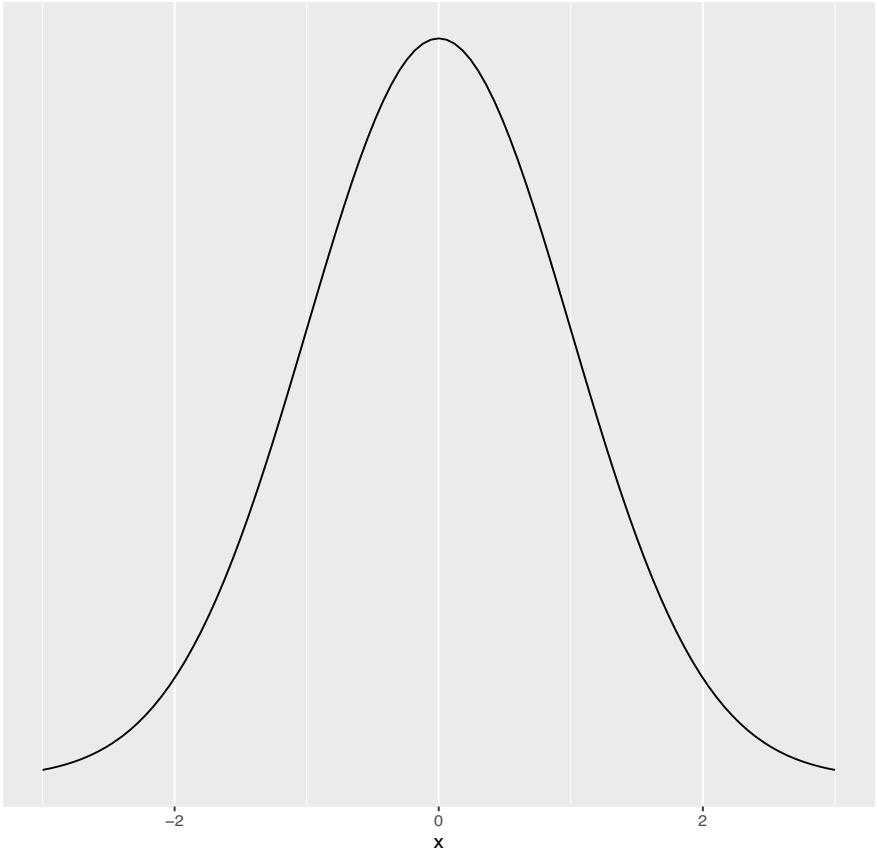

Figure 6.1: Conditional Density with Confidence Interval

When comparing (6.9) and (6.11), it follows that the standard deviation of the conditional density that determines the CI is given by SEP.

The confidence interval CI at a given error-level (as in Figure 6.1 we assume $\alpha = 0.05$) is defined by the lower limit l and the upper limit m. From the plot in Figure 6.1, the limits are defined as

$$\begin{aligned} l &= \hat{u} - z * SEP \\ m &= \hat{u} + z * SEP \end{aligned} \tag{6.12}$$

where z is the quantile value such that the surface under the curve corresponds to $1-\alpha$. For $\alpha = 0.05$, the value of z corresponds to 1.96. The values for z are obtained via the R-function `qnorm()`.

6.2 Best Linear Unbiased Prediction

The abbreviation `BLUP` stands for **B**est **L**inear **U**nbiased **P**rediction and summarizes statistical properties of a certain class of predictors. [Henderson, 1975] showed that the predicted breeding value as defined in (6.5) has the BLUP properties. BLUP-based procedures are the de-facto standard methods to predict breeding values. These procedures use linear mixed effects models to simultaneously estimate fixed effects and predict breeding values. Linear models that contain besides the random error other random effects are called linear mixed effects model. The classification into fixed and random effects is not based on a universally accepted definition. For some effects with distinct and a finite number of classes it is easy to classify them as fixed effects. On the other hand breeding values of animals are always modeled as random effects.

6.3 The Linear Mixed Effects Model

In general the linear mixed effects model has the following structure

$$y = Xb + Zu + e \tag{6.13}$$

where

y	vector of length n with observations
b	vector of length p with fixed effects
u	vector of length q with random breeding values
e	vector of length n with random error terms
X	$n \times p$ incidence matrix linking fixed effects to observations
Z	$n \times q$ incidence matrix linking random breeding values to observations

With a linear mixed effects model, we also have to specify the expected values and the variance-covariance matrices for the random components in the model. We start with the expected values in vector-notation.

$$E \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \end{bmatrix} \tag{6.14}$$

The variance-covariance matrices can be written as

$$var \begin{bmatrix} y \\ u \\ e \end{bmatrix} = \begin{bmatrix} ZGZ^T + R & ZG & 0 \\ GZ^T & G & 0 \\ 0 & 0 & R \end{bmatrix} \tag{6.15}$$

6.4 Solutions To The Linear Mixed Effects Model

In the linear mixed effects model (6.13) the vectors b and u are unknown and are of primary interest when using the model in livestock breeding applications. Applying the BLUP properties to the model (6.13) leads to the following solutions for b and u.

$$\hat{u} = GZ^T V^{-1}(y - X\hat{b}) \tag{6.16}$$

where V stands for the variance covariance matrix $var(y)$ of the phenotypic observations. From (6.15), we can see that V can be computed as $V = ZGZ^T + R$. The vector $\hat{b}$ stands for the best linear unbiased estimate of the fixed effects b which are given by

$$\hat{b} = (X^T V^{-1} X)^- X^T V^{-1} y \tag{6.17}$$

where $(X^T V^{-1} X)^-$ stands for a generalized inverse of $(X^T V^{-1} X)$.

Both solutions in (6.16) and (6.17) contain the matrix V^{-1} which has the dimensions $n \times n$ where n stands for the number of observations. In practical breeding programs the number of observations is frequently in the order of 10^6 which makes the computation of V^{-1} practically impossible. CR Henderson and his team found that the solution of the so-called mixed model equations lead to equivalent solutions for $\hat{b}$ and $\hat{u}$. The mixed model equations are given by

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{6.18}$$

The application of the linear mixed effects model (6.13) in Livestock Breeding have led to two specially named models depending on what was used as random effects. These models are

1. **sire model**: for each observation the influence of the animals sire was used as random effect.
2. **animal model**: the influence of the genetic component of each animal was used as random effect.

6.5 Sire Model

In the sire model the effect of the sire is taken as a random effect. This leads to the following model

$$y = Xb + Zs + e \tag{6.19}$$

where s is a vector of length q_s with all sire effects. All other components are the same as specified for the model (6.13). The variance-covariance matrix $var(s)$ of all sire effects corresponds to

$$var(s) = A_s * \sigma_s^2 \tag{6.20}$$

where A_s is the numerator relationship matrix only considering relations on the male side of the pedigree. The variance component σ_s^2 is obtained from the variance components estimation using a sire model and it corresponds to $0.25\sigma_a^2$ where σ_a^2 is the genetic additive variance.

The sire model was developed for livestock populations with large half-sib families where a given sire has a large number of offspring. As a result all sires get predicted breeding values. All female animals are not considered in the evaluation and do not get predicted breeding values.

The usage of sire models for predicting breeding values was mainly at the beginning of the introduction of the BLUP methodology. It was motivated by the lower requirements of computing power.

6.6 Animal Model

The so-called animal model uses the breeding value of each animal as a random effect. In matrix-vector notation the animal model has the following structure.

$$y = Xb + Za + e \tag{6.21}$$

where a is a vector of length q_a containing the breeding values of all animals in the pedigree. All other components are the same as specified for the model (6.13). The variance-covariance matrix $var(a)$ of the breeding values is defined as

$$var(a) = A\sigma_a^2 \tag{6.22}$$

where A is the numerator relationship matrix containing all identity by descent relationships in a pedigree. The matrix A is symmetric. The off-diagonal elements $(A)_{ij}$ in row i and column j contain the probability of animals i and j sharing alleles that are identical by descent. Identity by descent of two alleles means that two alleles can be traced back in the pedigree to a common ancestor. As an example the value of $(A)_{ij}$ between two half-sibs i and j is $1/4$ and between two full-sibs it is $1/2$. The diagonal elements $(A)_{ii}$ of matrix A corresponds to $1 + F_i$ where F_i is the inbreeding coefficient of animal i. The inbreeding coefficient corresponds to $1/2$ of the relationship coefficient between parents s and d of i.

6.7 Genomic BLUP

With the rapid increase of animals with genotypic information in the form of SNP-genotypes, this information can be used to predict genomic breeding values. BLUP-based methods can be used to predict genomic breeding values based on genomic information. The genomic BLUP models can have two different forms.

1. marker-effect models where each SNP-locus is taken as a random effect.
2. breeding value based models where genomic breeding values are included as random effects.

As long as the number of genotyped animals is lower than the number of genotypes determined per animal, then model (2) has advantages over model (1) with respect to requirements of computing power. But as the number of genotyped animals increases very rapidly, model (1) might be favorable over model (2).

At this point it is important to note that whether we use genomic information or not, the goal of the prediction of breeding values remains still the same. We want to get a prediction of the genetic potential of each animal. This prediction must be as accurate as possible. The realized values of the predictions are used to rank the animals. Based on these rankings the selection decisions are taken to determine the parents of the future generations of livestock animals.

In GBLUP genomic breeding values are directly predicted by including the genomic breeding values u which corresponds to the sum of all SNP-allele effects directly as a random effect in the model.

$$y = X\beta + Wu + e \tag{6.23}$$

where W is the design matrix linking genomic breeding values to observations. The mixed model equations are the defined as

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} W \\ W^T R^{-1} X & W^T R^{-1} W + G^{-1} * \lambda \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ W^T R^{-1} y \end{bmatrix} \tag{6.24}$$

where G is defined as the genomic relationship matrix and λ corresponds to the ratio of the residual variance divided by the total genetic variance explained by all SNP-loci. We can see that the GBLUP model looks very similar to the animal model, except that the covariances between random effects in the animal model are based on the numerator relationship matrix and in GBLUP they are modeled via the genomic relationship matrix G. This means in the animal model the covariance between random breeding values is based on the concept of common ancestry and identity-by-descent. This is replaced in GBLUP by the concept of sharing the same alleles based on identity-by-state which is assumed to be the cause of the covariance between random genomic breeding values.

6.8 How Does GBLUP Work

The genomic relationship matrix G allows to predict genomic breeding values for animals with SNP-Genotypes without any observation in the dataset. This fact is the basis of the large benefit of genomic selection. As soon as a young animal is born, its SNP genotypes can be determined and a genomic breeding value can be predicted. This genomic breeding value is much more accurate then the traditional breeding value based only on ancestral information.

The breeding value model is a linear mixed effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (6.25). In this form the Inverse G^{-1} of G and the vector $\hat{g}$ of predicted genotypic breeding values are split into one part corresponding to the animals with observations and a second part for the animals without phenotypic information.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{6.25}$$

The matrix $G^{(11)}$ denotes the part of G^{-1} corresponding to the animals with phenotypic observations. Similarly, $G^{(22)}$ stands for the part of the animals without genotypic observations. The matrices $G^{(12)}$ and $G^{(21)}$ are the parts of G^{-1} which link the two groups of animals. The same partitioning holds for the vector of predicted breeding values. The vector $\hat{g}_1$ contains the predicted breeding values for the animals with observations and the vector $\hat{g}_2$ contains the predicted breeding values of all animals without phenotypic observations.

Based on the last line of (6.25) the predicted breeding values $\hat{g}_2$ of all animals without phenotypic observations can be computed from the predicted breeding values $\hat{g}_1$ from the animals with observations.

$$\hat{g}_2 = -\left(G^{22}\right)^{-1} G^{21}\hat{g}_1 \tag{6.26}$$

Equation (6.26) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations. The genomic regression is responsible why genomic selection has lead to a change in paradigmn in cattle breeding with a dramatic reduction of the generation interval in the breeding program. In pig breeding genomic breeding values based on equation (6.26) are used to make a selection decision between litter mates and thereby determine which of the piglets are kept and which are sent to the testing station.

Bibliography

E W Brascamp. Selection indices with constraints. *Anim. Breed. Abstr*, 52 (9):645–654, 1984. URL http://scholar.google.com/scholar?hl=en{&}btnG=Search{&}q=intitle:Selection+indices+with+constraints{#}0.

Bundesrat. Verordnung über die Tierzucht, 2012. URL https://www.admin.ch/opc/de/classified-compilation/19880226/201510010000/814.011.pdf.

Alois Essl. *Statistische Mehoden in der Tierproduktion*. Österreichischer Agrarverlag, Wien, 1987. ISBN 3-7040-0859-1.

Daniel Gianola and Rohan L Fernando. Bayesian Methods in Animal Breeding Theory. *Journal of Animal Science*, 63(1):217–244, 1986. ISSN 0021-8812. doi: 10.2527/jas1986.631217x. URL https://doi.org/10.2527/jas1986.631217x.

J P Gibson and B W Kennedy. The use of constrained selection indexes in breeding for economic merit. *Theoretical and Applied Genetics*, 80(6):801–805, 1990. doi: https://doi.org/10.1007/BF00224195. URL https://link.springer.com/content/pdf/10.1007{%}2FBF00224195.pdf.

Kay-Uwe Götz. Quantitative Genetik und Zuchtplanung, 2014.

Author H O Hartley and J N K Rao. Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model. *Biometrika*, 54(1/2):93–108, 1967. URL https://www.jstor.org/stable/2333854.

C R Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423—-449, 1975.

C.R. Henderson. Selection index and expected geentic advance. In W.D. Hanson and H.F. Robinson, editors, *Statistical Genetics and Plant Breed- ing*, pages 141–163. Washington, DC, publicatio edition, 1963.

Bernard Lehmann, Eva Reinhard, Adrian Aebi, Christian Hofer, Andrea Leute, and Dominique Kohli. Strategie Tierzucht 2030, 2018. URL https://www.newsd.admin.ch/newsd/message/attachments/52496.pdf.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554, 1971. ISSN 00063444. doi: 10.1093/biomet/58.3.545.

F. Phocas, C. Bloch, P. Chapelle, F. Bécherel, G. Renand, and F. Ménissier. Developing a breeding objective for a French purebred beef cattle selection programme. *Livestock Production Science*, 57(1):49–65, 1998. ISSN 03016226. doi: 10.1016/S0301-6226(98)00157-2. URL https://www.sciencedirect.com/science/article/abs/pii/S0301622698001572.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance Components*. Hoboken, NJ, 1992. ISBN ISBN-I 3 978-0-470-00959-8.

CS Wang, JJ Rutledge, and D Gianola. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. *Genetics Selection Evolution*, 26(2):91, 1993. ISSN 1297-9686. doi: 10.1186/1297-9686-26-2-91.

Ping Yin and Xitao Fan. Estimating r2 shrinkage in multiple regression: A comparison of different analytical methods. *Journal of Experimental Education*, 69(2):203–224, 2001. ISSN 19400683. doi: 10.1080/00220970109600656.