Best Linear Unbiased Prediction used in Quantitative Genomics

Stefan McKinnon Høj-Edwards & Peter Sørensen

2022-03-14

Contents

1 Introduction

Breeding value estimation in animal and plant breeding programmes are nowadays based on the BLUP (Best Linear Unbiased Prediction) method. The estimation of breeding values based on multiple sources of informations must correct for the redundancy between them (e.g., the redundant information provided by parents and grandparents). Moreover, they need to be adjusted for average effects in the populations, "fixed effects". So far we have referred to that fixed effect as the population mean and we have assumed this asjustment μ to be known. Indeed, we defined the true breeding value a and the non-identifiable environmental effects e as deviations from a common mean, the average effect of all fixed genetic and environmental factors captured by the population mean μ. But this is only true in a single idealized population where all selection candidates are kept in the same environment in which they deliver their performances at the same time. In practice the phenotypic records often need to be adjusted for systematic (fixed) environmental effects, such as age, parity, litter size, days open, sex, herd, year, season, management, etc. Several of those effects fluctuate very little over time, so accurate estimates of their effect may be obtained from previous ("historical") sets

of data. Effects of factors like herd, year, season, and management fluctuate more and are therefore best estimated directly from the data to be used in the genetic evaluations.

Compared to the idealized cases described in the previous section, a practical breeding scenario poses two problems: accounting for heteregeneous sources of genetic information (different types of relatives); and adjusting for fixed effects in the breeding population(s) (fixed environmental or genetic effects). The BLUP solution to these problems was presented by Charles R. Henderson in several publications (e.g. Henderson1973a) and Henderson1975). 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 model. Here, mixed refers to the presence of two types of effects: fixed effects (identifiable effects from environmental or genetic factors) and random effects (non-identifiiable effects from segregating genetic factors and fluctuating environmental conditions). The methodology developed by Henderson is called **BLUP** and the properties of this methodology are directly incorporated into its name:

- **B** stands for **best** which means that the correlation between the true (a) and the predicted breeding value ($\hat{a}$) is maximal or the prediction error variance ($var(a - \hat{a})$) 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 approaches are widely used in genetic evaluations, 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 also the efficient algorithms developed by Henderson for computing predicted breeding values, even in very large breeding populations. The theoretical foundations, and the development of efficient algorithms, together with the availability of large computational resources at a very low price, have made BLUP the de-facto standard for breeding value estimation.

This chapter focuses on the usage and application of linear mixed models, as they are applied in these notes. The first section applies standard probability theory and linear algebra to predict genetic values and marker effects, assuming the variances are known. Because the solutions provided in section 2 are not always computational feasible, the Mixed Models Equations have been developed, which are described in section 3. These solutions however also require that the variances are known, so estimating these with Maximum Likelihood and Restricted Maximum Likelihood are outlined in Chapter **??**.Finally, we will return to predicting genetic values and estimating a model's predictive ability in section **??**.

For the following chapter, we will extend the simple model notation of M-BLUP and G-BLUP to a general form that allows multiple observations per individual, hence the matrix $\mathbf{Z}$ emerges; an $n \times q$ incidence matrix relating observations to individuals. We also expand the intercept μ to $\mathbf{X}\boldsymbol{\beta}$, the $n \times p$ incidence matrix relating observations to the p-length vector of fixed effects, $\boldsymbol{\beta}$. Written in their entirety, the models become:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{ZWb} + \mathbf{e} \tag{1}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zg} + \mathbf{e}. \tag{2}$$

We are using the terms 'Estimating' and 'Predicting' almost interchangeable, and the terms might led to some confusion as to whether they are merely synonyms or intone a more subtle difference. It appears that it has become common practice to 'estimate' fixed effects and 'predict' random effects [Robinson, 1991], but, as it usually is with field specific jargon, it depends on the school of thought, e.g. classical statistics or Bayesian inference. Another view on the difference is that we predict future performances or values yet to be observed, while we estimate values (e.g. variance components) inherent in the data.

2 Predicting genetic values and marker effects

Predicting genetic values has generally been the most important task for animal geneticists as they are used by farmers to select animals for reproduction. Originally, genetic values have been calculated by using the expected relationships among individuals by using the pedigree based relationship matrix A. Nowadays, the advent of genome-wide genotyping at cheaper cost is revolutionizing the field by allowing the construction of the relationship matrix taking into account the realized relationships among individuals by using SNP marker data: the genetic relationship matrix $\mathbf{G}$.

For this section, we will cover three scenarios, all based on the M-BLUP and G-BLUP models, and the assumptions in chapter **??**. To solve the conditional expectations, we will be relying on standard probability theory and linear algebra. These tools are briefly described in the 'Stat boxes'.

The first scenario is predicting genetic values based on the observed values, i.e. finding the solution to $\mathrm{E}[\mathbf{g}|\mathbf{y}]$. In the second scenario, we will predict the genetic values after predicting the marker effects, i.e. $\mathrm{E}[\mathbf{g}|\mathbf{b}]$. The third scenario is the reverse of the second, predicting marker effects from predicted genetic values, i.e. $\mathrm{E}[\mathbf{b}|\mathbf{g}]$.

2.1 Predicting genetic values from observations: $\mathrm{E}[\mathbf{g}|\mathbf{y}]$

Here, we want to predict the genetic values, or put in another way, predict the effects that can be attributed to the genetic relationship after correcting for fixed effects.

We will in the following show that

$$\mathrm{E}\left[\mathbf{g}|\mathbf{y}\right] = 0 + \mathbf{G}\mathbf{Z}'\sigma_g^2(\mathbf{Z}\mathbf{G}\mathbf{Z}'\sigma_g^2 + \mathbf{D}\sigma_e^2)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}$$

by applying stat box 1.

Our starting point is the general G-BLUP equation written as a set of linear equations for $\mathbf{y}$ and $\mathbf{g}$:

$$\underbrace{\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\end{pmatrix}}_{y^*} = \underbrace{\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}}_{c} + \underbrace{\begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}}_{A}\underbrace{\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix}}_{x} \tag{4a}$$

and $\mathbf{g}$ and $\mathbf{e}$

$$\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix} \sim N\left[\underbrace{\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}}_{m}, \underbrace{\begin{pmatrix}\mathbf{G}\sigma_g^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\sigma_e^2\end{pmatrix}}_{V}\right] \tag{4b}$$

where the top line in (4a) corresponds to the conditional expectation of y, the bottom line corresponds to the marginal expectation of g, and where the notations below both equations corresponds exactly to those in stat box 1.

We can now derive the joint distribution of G-BLUP by applying stat box 1:

$$\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\end{pmatrix} \sim N\left[\begin{pmatrix}\mu\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{Z}\mathbf{G}\mathbf{Z}'\sigma_g^2 + \mathbf{D}_n\sigma_e^2 & \mathbf{Z}\mathbf{G}\sigma_g^2\\ \mathbf{G}\mathbf{Z}'\sigma_g^2 & \mathbf{G}\sigma_g^2\end{pmatrix}\right]. \tag{5}$$

The proof is followed, first for the joint expectation of (4a):

$$
\begin{aligned}
\mathrm{E}\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\end{pmatrix}\right] &= \mathrm{E}\left[\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix}\right] \\
&= \mathrm{E}\left[\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}\right] + \begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\mathrm{E}\left[\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix}\right] \\
&= \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\mathrm{E}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}\right] \\
&= \begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{0}\end{pmatrix}
\end{aligned}
$$

The variance is derived as:

$$
\begin{aligned}
\mathrm{Var}\left[\begin{pmatrix}\mathbf{y}\\ \mathbf{g}\end{pmatrix}\right] &= \mathrm{Var}\left[\begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix}\right] \\
&= \begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\mathrm{Var}\left[\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\end{pmatrix}\right]\begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}' \\
&= \begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{G}\sigma_g^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{D}\sigma_e^2\end{pmatrix}\begin{pmatrix}\mathbf{Z} & \mathbf{I}\\ \mathbf{I} & \mathbf{0}\end{pmatrix}' \\
&= \begin{pmatrix}\mathbf{Z}\mathbf{G}\mathbf{Z}'\sigma_g^2 + \mathbf{D}\sigma_e^2 & \mathbf{Z}\mathbf{G}\sigma_g^2\\ \mathbf{G}\mathbf{Z}'\sigma_g^2 & \mathbf{G}\sigma_g^2\end{pmatrix}
\end{aligned}
$$

By applying stat box 3, the conditional expectation of $\mathbf{g}$ given $\mathbf{y}$ is therefore

$$
\mathrm{E}\left[\mathbf{g}|\mathbf{y}\right] = 0 + \mathbf{G}\mathbf{Z}'\sigma_g^2(\mathbf{Z}\mathbf{G}\mathbf{Z}'\sigma_g^2 + \mathbf{D}\sigma_e^2)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3}
$$

which is the Best Linear Unbiased Predictor of $\mathbf{g}$. 'Best' because it has the minimum mean squared error among the possible predictors,'Linear' because it is a linear function of the data, 'Unbiased' because the average value of the predictor is equal to the one of the quantity being predicted.

The construct $\mathbf{Z}\mathbf{G}\mathbf{Z}'$ expands the genomic relationship matrix from $q \times q$ to $n \times n$ and, if there are multiple observations per individual, the result is rank deficient (see (**??**), p. **??**). So when the data set to be analysed contains a lot of observations it might not possible to calculate the inverse in (3), partly due to the structure, partly due to the sheer size. The Mixed Model Equations (MME), described in section 3, can therefore be used as the dimensions of the matrices are in the order of the sum of number of fixed effects and genetic values (or marker effects). Furthermore, the matrices needed inverted in MME can be sparse so more effective inversion algorithms can be applied, except in animal breeding where the genomic relationship matrix usually are dense – but the inverse of these can usually be made while constructing the ordinary matrix.

2.2 Predicting g from b: E[g|b]

Having predicted values of marker effects ($\hat{\mathbf{b}}$), we can predict genetic values based on the linear combination:

$$
\hat{\mathbf{g}} = \mathbf{W}\hat{\mathbf{b}} \tag{6a}
$$

or

$$
\hat{g}_i = \sum_{j=1}^{m} w_{i,j}\hat{b}_j. \tag{6b}
$$

2.3 Predicting b from g: E[b|g]

This is the reverse of the above, but as W is not necessarily a square and invertible matrix, the solution is not straightforward. The solution is

$$\mathrm{E}\left[\mathbf{b}|\mathbf{g}\right] = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\mathbf{g} \tag{7}$$

which will be derived in the following. We could do as for E[**g**|**y**], but we will here use another approach that is a more general solution.

To achieve the goal, it is useful to write down the joint distribution of **b**, **g** and **y**. The starting point to build this joint distribution is to keep in mind the assumptions of M-BLUP and G-BLUP that we have already stated in section **??**. To make things clear they are summarised in table 1 for the general form.

The (co)variances are easily derived from this:

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b},\mathbf{b}) &= \mathrm{Var}(\mathbf{b}) = \mathbf{I}\sigma_b^2 \\
\mathrm{Cov}(\mathbf{b},\mathbf{g}) &= \mathrm{Cov}(\mathbf{b},\mathbf{W}\mathbf{b}) = \mathrm{Cov}(\mathbf{b},\mathbf{b})\mathbf{W}' = \mathbf{W}'\sigma_b^2 \\
\mathrm{Cov}(\mathbf{b},\mathbf{y}) &= \mathrm{Cov}(\mathbf{b},\mathbf{Z}\mathbf{W}\mathbf{b}+\mathbf{e}) = \mathrm{Cov}(\mathbf{b},\mathbf{Z}\mathbf{W}\mathbf{b}) \\
&= \mathrm{Cov}(\mathbf{b},\mathbf{b})(\mathbf{Z}\mathbf{W})' = \mathbf{W}'\mathbf{Z}'\sigma_b^2 \\
\mathrm{Cov}(\mathbf{g},\mathbf{b}) &= \mathrm{Cov}(\mathbf{W}\mathbf{b},\mathbf{b}) = \mathbf{W}\,\mathrm{Cov}(\mathbf{b},\mathbf{b}) = \mathbf{W}\sigma_b^2 \\
\mathrm{Cov}(\mathbf{g},\mathbf{g}) &= \mathrm{Cov}(\mathbf{W}\mathbf{b},\mathbf{W}\mathbf{b}) = \mathbf{W}\,\mathrm{Cov}(\mathbf{b},\mathbf{b})\mathbf{W}' \\
&= \mathbf{W}\mathbf{W}'\sigma_b^2 \\
\mathrm{Cov}(\mathbf{g},\mathbf{y}) &= \mathrm{Cov}(\mathbf{y},\mathbf{g})' = (\mathbf{Z}\mathbf{G})'\sigma_g^2 = \mathbf{W}\mathbf{W}'\mathbf{Z}'\sigma_b^2 \\
\mathrm{Cov}(\mathbf{y},\mathbf{b}) &= \mathrm{Cov}(\mathbf{Z}\mathbf{W}\mathbf{b}+\mathbf{e},\mathbf{b}) = \mathrm{Cov}(\mathbf{Z}\mathbf{W}\mathbf{b},\mathbf{b}) = \mathbf{Z}\mathbf{W}\sigma_b^2 \\
\mathrm{Cov}(\mathbf{y},\mathbf{g}) &= \mathbf{Z}\mathbf{G}\sigma_g^2 = \mathbf{Z}\mathbf{W}\mathbf{W}'\sigma_b^2 \\
\mathrm{Cov}(\mathbf{y},\mathbf{y}) &= \mathrm{Var}(\mathbf{y}) = \mathrm{Var}(\mathbf{Z}\mathbf{W}\mathbf{b}+\mathbf{e}) = \mathbf{Z}\mathbf{W}\mathbf{W}'\mathbf{Z}'\sigma_b^2 + \mathbf{D}\sigma_e^2
\end{aligned}$$

Having now all the necessary information we can build the joint distribution of **b**, **g** and **y** by hand:

$$\begin{pmatrix}\mathbf{b}\\ \mathbf{g}\\ \mathbf{y}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mu\end{pmatrix}, \begin{pmatrix}\mathbf{I}\sigma_b^2 & \mathbf{W}'\sigma_b^2 & \mathbf{W}'\mathbf{Z}'\sigma_b^2 \\ \mathbf{W}\sigma_b^2 & \mathbf{W}\mathbf{W}'\sigma_b^2 & \mathbf{W}\mathbf{W}'\mathbf{Z}'\sigma_b^2 \\ \mathbf{Z}\mathbf{W}\sigma_b^2 & \mathbf{Z}\mathbf{W}\mathbf{W}'\sigma_b^2 & \mathbf{Z}\mathbf{W}\mathbf{W}'\mathbf{Z}'\sigma_b^2 + \mathbf{D}\sigma_e^2\end{pmatrix}\right] \tag{8}$$

Dividing σ_b^2 out of the matrix, we get the result

$$\begin{pmatrix}\mathbf{b}\\ \mathbf{g}\\ \mathbf{y}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mu\end{pmatrix}, \begin{pmatrix}\mathbf{I} & \mathbf{W}' & \mathbf{W}'\mathbf{Z}' \\ \mathbf{W} & \mathbf{W}\mathbf{W}' & \mathbf{W}\mathbf{W}'\mathbf{Z}' \\ \mathbf{Z}\mathbf{W} & \mathbf{Z}\mathbf{W}\mathbf{W}' & \mathbf{Z}\mathbf{W}\mathbf{W}'\mathbf{Z}' + \mathbf{D}\lambda\end{pmatrix}\right] \tag{9}$$

Table 1: Summary of assumptions for the BLUPs.

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{W}\mathbf{b} + \mathbf{e}$ (**??**)	$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$
$\mathbf{g} = \mathbf{W}\mathbf{b}$	$\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$
	$\mathbf{b} \sim N(0, \mathbf{I}\sigma_b^2)$
	$\mathbf{e} \sim N(0, \mathbf{D}\sigma_e^2)$
$\mathrm{Cov}(\mathbf{b},\mathbf{e}) = 0$, $\mathrm{Cov}(\mathbf{g},\mathbf{e}) = 0$, $\mathrm{Cov}(\mathbf{y},\mathbf{g}) = \mathbf{Z}\mathbf{G}\sigma_g^2$	

where $\lambda = \frac{\sigma_e^2}{\sigma_b^2}$.

This is a degenerated distribution, as the two left columns in the bottom matrix are linear dependent by multiplying with $\mathbf{W}$, i.e. the variance-covariance matrix is not full rank [1].

By applying stat box 3 on $\mathbf{b}$ and $\mathbf{g}$, we get the conditional expectation of $\mathbf{b}$ given $\mathbf{g}$:

$$\mathrm{E}[\mathbf{b}|\mathbf{g}] = 0 + \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\mathbf{g} \tag{10}$$

However, the inverse of $\mathbf{W}\mathbf{W}'$ does not exist, as $\mathbf{W}$ is not full rank, due to the definition of $\mathbf{W}$. The inverse of $\mathbf{W}\mathbf{W}'$ exists only if $\mathbf{W}\mathbf{W}'$ is non-singular, which might be possible if the genotypes are in the original allele coding (i.e. 0,1,2). This problem can be easily circumvented by conditioning on $\mathbf{y}$ rather than on $\mathbf{g}$, leading to:

$$\mathrm{E}[\mathbf{b}|\mathbf{y}] = 0 + \mathbf{W}'\mathbf{Z}'(\mathbf{Z}\mathbf{W}\mathbf{W}'\mathbf{Z}' + \mathbf{D}\lambda)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{11}$$

a solution similar to (3).

The joint distribution found in (8) could also have been derived from a set of linear equations, in a similar manner to section 2.1, by starting from

$$\begin{pmatrix}\mathbf{b}\\ \mathbf{g}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{X}\boldsymbol{\beta}\end{pmatrix} + \begin{pmatrix}\mathbf{I} & \mathbf{0}\\ \mathbf{W} & \mathbf{0}\\ \mathbf{Z}\mathbf{W} & \mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{b}\\ \mathbf{e}\end{pmatrix}$$

.

3 Mixed Model Equations

The mixed model equations (MME) were first introduced by Henderson in 1949. This set of equations is used to jointly obtain estimates of fixed and random effects while avoiding the need of $(\mathbf{G}\sigma_g^2 + \mathbf{D}\sigma_e^2)^{-1}$, which is difficult to compute when the data set is large (computational time is proportional to n^3). A fundamental property of the estimates of random and fixed effects obtained is that they are, respectively, the BLUP and the BLUE (i.e. same properties as BLUP but for fixed effects; 'E' stands for Estimator). The proof of this was published by Goldberger [1962] and Henderson [1963] (with the help of Searle). This discovery has revolutionized the field of animal breeding since it has allowed to estimate breeding values in relatively short time, providing that variance components are known (if they are not, they can be estimated from the data as explained in section **??**).

3.1 MME for G-BLUP

The derivation for MME below follows the approach by Henderson et al. [1959]. It involves the maximization of the joint density of $\mathbf{y}$ and $\mathbf{g}$ with respect to $\boldsymbol{\beta}$ and $\mathbf{g}$, assuming a multivariate normal distribution and known variances. This joint density can be written as:

[1] A degenerate distribution will only return a single value; a normal distribution with zero variance – at first an as absurd concept as a black hole – has its entire mass in a single point, and will only return a single value. Writing the probability distribution for a matrix normal distribution, it requires the inversion and determinant of the variance-covariance matrix. Hence, a non-invertible variance-covariance matrix will give a normal distribution with zero variance, thus a degenerated distribution.

$$
\begin{aligned}
p(\mathbf{y},\mathbf{g}) &= p(\mathbf{y}|\mathbf{g})p(\mathbf{g}) \\
&\propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g})\right] \\
&\quad \exp\left[-\frac{1}{2}\mathbf{g}'\mathbf{F}^{-1}\mathbf{g}\right] \qquad (12)
\end{aligned}
$$

where $\mathbf{R} = \mathbf{D}\sigma_e^2$, and $\mathbf{F} = \mathbf{G}\sigma_g^2$.

Transferring (12) into log-space, it becomes:

$$
\begin{aligned}
\ln p(\mathbf{y},\mathbf{g}) = \mathit{constant} - \qquad (13) \\
\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g})'\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g}) - \frac{1}{2}\mathbf{g}'\mathbf{F}^{-1}\mathbf{g}
\end{aligned}
$$

To maximize this function, the derivatives with respect to $\mathbf{b}$ and $\mathbf{g}$ have to be taken:

$$\frac{\partial \ln p(\mathbf{y},\mathbf{g})}{\partial \boldsymbol{\beta}} = (-\mathbf{X}')[-\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g})] \qquad (14a)$$

$$\frac{\partial \ln p(\mathbf{y},\mathbf{g})}{\partial \mathbf{g}} = (-\mathbf{Z}')[-\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{g})] + (-\mathbf{F}^{-1}\mathbf{g}) \qquad (14b)$$

Setting (14a) and (14b) equal to 0 gives the estimates of $\boldsymbol{\beta}$ and $\mathbf{g}$, expressed by:

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{g}} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \qquad (15a)$$

and

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{g}} + \mathbf{F}^{-1}\hat{\mathbf{g}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \qquad (15b)$$

These equations can then be rearranged in the traditional mixed model notation:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{F}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad (16)$$

Assuming $\mathbf{R}$ and $\mathbf{F}$ to be non-singular, $\mathbf{R}^{-1}$ can be factored out from both sides as it is an identity matrix times a scalar. This leads to:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\mathbf{G}^{-1}\lambda_g \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad (17)$$

where $\lambda_g = \sigma_e^2/\sigma_g^2$ which determines to what extent estimates of genetic values are regressed back towards the mean, i.e. how much contribution from the genetic relationship will we allow to be included.

3.2 MME for M-BLUP

The derivation of the MME for M-BLUP follows that for G-BLUP, resulting in

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{W} \\ \mathbf{W}'\mathbf{Z}'\mathbf{X} & \mathbf{W}'\mathbf{Z}'\mathbf{Z}\mathbf{W}+\mathbf{I}\lambda_b \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{b}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{Z}'\mathbf{y} \end{bmatrix} \qquad (18)$$

where $\lambda_b = \sigma_e^2/\sigma_b^2$.

3.3 Solving the Mixed Models Equations

This is straightforward. The MME is a linear set of equations consisting of (left to right), the *kernel* or *coefficient matrix*, vector of unknowns to be estimated/predicted, and on the right hand side, a matrix of known values. The solution requires the inversion of the kernel, which can be done by any of the large number of standard matrix inversion approaches. However easy this might seem, solving MME have become even easier with methods directed at solving systems of linear equations, such as the MME. These methods include the Cholesky decomposition or Gauss-Seidel method, that involve decomposing the coefficient matrix, or matrix-free variants that can save considerable computation time, by relying on functions of the coefficient matrix instead of whole-matrix operations [Legarra and Misztal, 2008].

The kernel on the left hand side determines the size of the equation to be solved. For G-BLUP (17), the kernel is a $(p+q) \times (p+q)$ matrix, or $O(q^2)$ using the Big O notation[2], assuming that p, the number of fixed effects, is much smaller than q.

For M-BLUP however, the size of the kernel is $O(m^2)$. This implies that if $q \ll m$, it might be computational beneficial to use G-BLUP instead of M-BLUP as the order of the set of equations to be solved would be much smaller for G-BLUP than M-BLUP.

3.4 MME for two random marker or genetic factors

Here, we will briefly display the MME for G-BLUP and M-BLUP with two random factors, as those given in section **??**, page **??**.

The MME for both G-BLUP and M-BLUP are presented in (19) and (**??**):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}_1^{-1}\lambda_{g_1} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}_2^{-1}\lambda_{g_2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{g}}_1 \\ \hat{\mathbf{g}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \end{bmatrix} \tag{19}$$

where $\lambda_{g_1} = \sigma_e^2/\sigma_{g_1}^2$ and $\lambda_{b_2} = \sigma_e^2/\sigma_{b_2}^2$.

For M-BLUP, we have

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{W}_1 & \mathbf{X}'\mathbf{Z}\mathbf{W}_2 \\ \mathbf{W}_1'\mathbf{Z}'\mathbf{X} & \mathbf{W}_1'\mathbf{Z}'\mathbf{Z}\mathbf{W}_1 + \mathbf{I}\lambda_{b_1} & \mathbf{W}_1'\mathbf{Z}'\mathbf{Z}\mathbf{W}_2 \\ \mathbf{W}_2'\mathbf{Z}'\mathbf{X} & \mathbf{W}_2'\mathbf{Z}'\mathbf{Z}\mathbf{W}_1 & \mathbf{W}_2'\mathbf{Z}'\mathbf{Z}\mathbf{W}_2 + \mathbf{I}\lambda_{b_2} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{Z}'\mathbf{y} \\ \mathbf{W}_2'\mathbf{Z}'\mathbf{y} \end{bmatrix}$$

where $\lambda_{b_1} = \sigma_e^2/\sigma_{b_1}^2$ and $\lambda_{b_2} = \sigma_e^2/\sigma_{b_2}^2$.

3.5 Some useful notations and properties of MME

We now consider the variances and standard errors of the predicted effects. These differ from the variance components (σ_g^2, σ_e^2, etc.) as giving the uncertainty of the predicted values of $\hat{\mathbf{b}}$, $\hat{\mathbf{g}}$, and $\hat{\boldsymbol{\beta}}$. In other words, the variance components are scalars and properties of the model, while the following are vectors (variance-covariance matrices) that detail the uncertainty (correlation) of the realised values. To express these vectors, some useful notations of the coefficient matrix are presented.

[2]The Big O notation describes the growth rate of a function; here that the size of the kernel quadruples if q doubles.

Recall the MME from section 3, page 6 for G-BLUP:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + F^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \tag{16}$$

where $R = D_n\sigma_e^2$ and $F = G\sigma_g^2$. Denoting the coefficient matrix as $\mathbf{C}$, it and its inverse can be written as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{\beta\beta} & \mathbf{C}_{\beta g} \\ \mathbf{C}_{g\beta} & \mathbf{C}_{gg} \end{bmatrix} \quad \text{and} \quad \mathbf{C}^{-1} = \begin{bmatrix} \mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta g} \\ \mathbf{C}^{g\beta} & \mathbf{C}^{gg} \end{bmatrix}. \tag{20}$$

Incidentally, $\text{Var}\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{g}} - \mathbf{g} \end{pmatrix} = \mathbf{C}^{-1}$. Following from this, the covariance matrix for the estimated fixed effects are given by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{C}^{\beta\beta} \tag{21}$$

and the standard errors are simply the square roots of the diagonal elements. The covariance matrix for the predicted genetic values is

$$\text{Var}(\hat{\mathbf{g}}) = \mathbf{G} - \mathbf{C}^{gg} \tag{22}$$

but it is more viable to consider the prediction errors $\hat{\mathbf{g}} - \mathbf{g}$, as the variance of this 'includes variance from both the prediction error and the random effects $\mathbf{g}$ themselves' [Lynch and Walsh, 1998, p. 754];

$$\text{Var}(\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{C}^{gg}. \tag{23}$$

We will also introduce the *hat matrix*[3], which is the matrix that transforms the $\mathbf{y}$ into $\hat{\mathbf{y}}$, i.e.

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \tag{24}$$

For G-BLUP, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{g}}$, $\hat{\mathbf{y}}$ can be expressed as either the marginal or the conditional predictions [Orenti et al., 2012]:

$$\hat{\mathbf{y}}_M = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{25a}$$

$$\hat{\mathbf{y}}_C = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\mathbf{g}} \tag{25b}$$

with corresponding hat matrices

$$\mathbf{H}_M = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \tag{26a}$$

$$\mathbf{H}_C = \mathbf{I} - \mathbf{V}^{-1} + \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}. \tag{26b}$$

We can also derive the two *hat like* matrixes, $\mathbf{H}_\beta$ and $\mathbf{H}_g$, such that $\hat{\mathbf{y}} = \mathbf{H}_\beta\mathbf{y} + \mathbf{H}_g\mathbf{y}$. These will be useful in later chapters for the sums of squares.

[3] It puts the hat on $\mathbf{y}$.

$$\mathbf{Zg} = \mathbf{Z}\begin{bmatrix}\mathbf{C}^{g\beta} & \mathbf{C}^{gg}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1}\end{bmatrix}\mathbf{y} = \mathbf{H}_g\mathbf{y}$$
$$\mathbf{H}_g = \mathbf{Z}\begin{bmatrix}\mathbf{C}^{g\beta} & \mathbf{C}^{gg}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1}\end{bmatrix} \tag{27}$$

and similarly for $\mathbf{H}_\beta$:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\begin{bmatrix}\mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta g}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1}\end{bmatrix}\mathbf{y} = \mathbf{H}_\beta\mathbf{y}$$
$$\mathbf{H}_\beta = \mathbf{X}\begin{bmatrix}\mathbf{C}^{\beta\beta} & \mathbf{C}^{\beta g}\end{bmatrix}\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1} \\ \mathbf{Z}'\mathbf{R}^{-1}\end{bmatrix}. \tag{28}$$

4 Cross validation

So far we have been discussing model fit, predicted marker effects and genetic values, and estimated variance components for evaluating a model. But a model may be used for more than trying to understand how the biological machinery produces the observed phenotypes. In animal breeding and personalised medicine, predicting future outcomes is highly valuable, and thus evaluating a models predictive ability has its merit. The predictive ability stems from the models' ability to generalise to another dataset, a feature that is penalized if the model has been overfitted.

Cross-validation is one method for estimating the predictive ability of a model.The data is split into a training set and a validation set, the model is fitted to the training set, and then used to predict the observations of the validation set. The observations of the validation set are then compared to the predicted observations, and this is the basis of the predictive ability.

A typical approach is the K-fold cross-validation, where the data is split into K sets, and each is in turn used as the validation set while the rest are used to fit the model. The predictive ability is then the mean of the K comparisons.Typical choices for K are 5 or 10 [Legarra et al., 2008, Hastie et al., 2009]. Other approaches exists such as `Leave-one-out' or`Repeated random sub-sampling validation', where the first corresponds to K-fold with K equals to number of data points. For the latter, the data points for the validation set are sampled at randomly for each 'fold', thus the validation subsets may overlap. For now, we will focus on the K-fold approach.

How to split the data set needs to be considered. If the dataset is homogeneous, samples may be divided at random, as seen in A) in figure 1. However, the division should always reflect how the resulting model is intended to be used, but this can introduce different sources of error. In animal breeding, the data might span several generations and/or multiple herds, in which case the dataset is no longer homogeneous and one has to be attentive to this. So to the use of cross validation and prediction, we will quickly find ourselves with an extrapolation which comes with an inherent cost, see C) in figure 1. We might also attempt to do cross-validation across multiple breeds, a situation akin to D).

As we will see below, the prediction ability of the models depend on the size of the variance components and the relationship between the two groups.

However, many other strategies have been tested and used to divide the data set. For example, Luan et al. [2009] divided the data set according to the year of birth of progeny and chose the animals for TD in number equal to the number of sires selected that particular year. By contrast, Legarra et al. [2008] sampled both within and across families. There is still uncertainty whether it is necessary to have close relatives and close population LD among individuals in TD and VD; what is sure is that there must be genomic relationships among the two sets of individuals. Legarra et al. [2008] obtained better predictive ability with close relatives

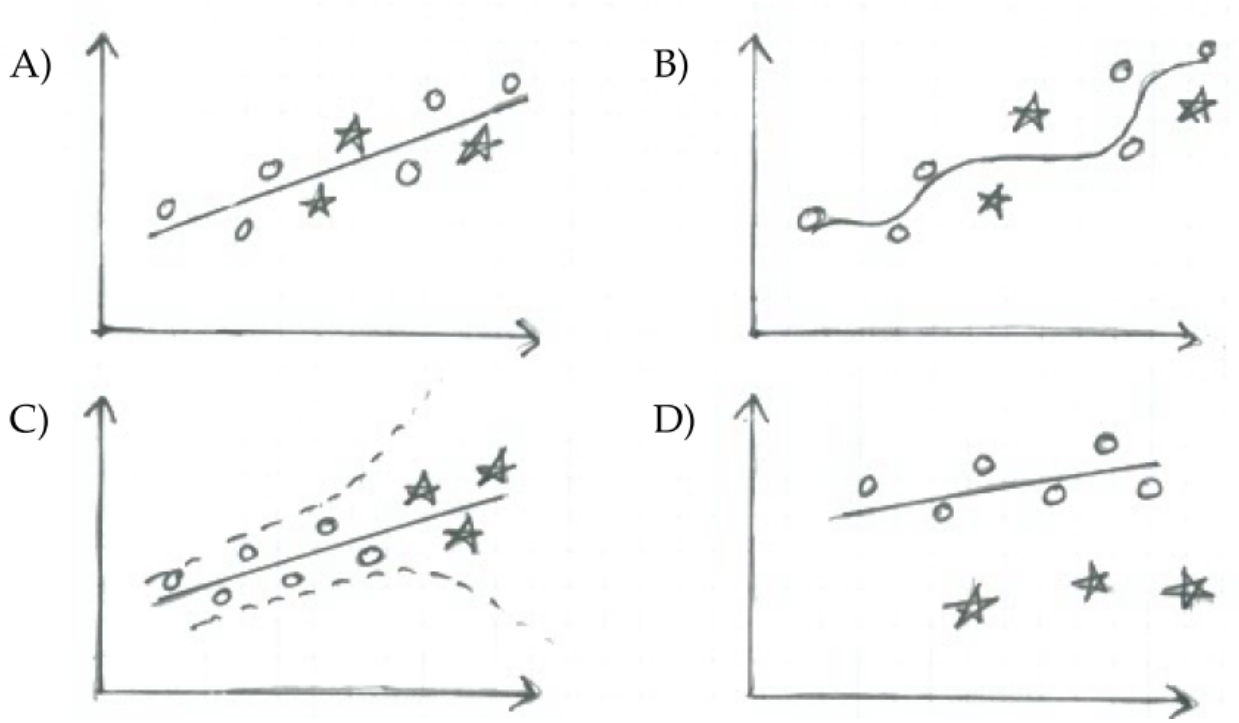

Figure 1: Illustration of splitting data for cross-valdiation. Circles are used as training set, stars as validation set. A) Simplistic 3-fold cross-validation where data set is homogeneous. B) Overfitted model on training set following bad prediction of validation set. C) Extrapolating from training set, resulting in increased confidence interval. D) Data not homogeneous and/or splitting data set based on a prior screen

whereas Ober et al. [2012] and Habier et al. [2007] found sufficient to good predictive abilities with distant relatives (even inbred lines in the case of Ober's paper) showing that long range LD could be useful as well. Further, this is also influenced by the method used to estimate parameters - G-BLUP makes more use of the genetic relationships while BayesB uses more the population LD.

4.1 Prediction in cross-validation: Individuals without phenotype

We will now cover the cross-validation for G-BLUP. We will denote the observations in the validation set as $\mathbf{y}_2$ and those in the training set as $\mathbf{y}_1$. Even though we mask the observations in $\mathbf{y}_2$ as unknown or missing, the genetic relationship for these observations are still known. The challenge now is to estimate $\mathbf{y}_2$ based on the training data set. For this, we derive the joint distribution of two G-BLUP models:

$$\begin{pmatrix}\mathbf{y}_2\\ \mathbf{y}_1\end{pmatrix} = \begin{pmatrix}\mathbf{X}_2\boldsymbol{\beta}_2\\ \mathbf{X}_1\boldsymbol{\beta}_1\end{pmatrix} + \begin{pmatrix}\mathbf{g}_2\\ \mathbf{g}_1\end{pmatrix} + \begin{pmatrix}\mathbf{e}_2\\ \mathbf{e}_1\end{pmatrix} \tag{29}$$

in the same manner as we did in section 2.3. Assumptions are as noted in section **??**, but we annotate $\mathbf{G}$ as:

$$\begin{pmatrix}\mathbf{g}_2\\ \mathbf{g}_1\end{pmatrix} \sim N\left[\mathbf{0}, \begin{pmatrix}\mathbf{G}_{22} & \mathbf{G}_{21}\\ \mathbf{G}_{12} & \mathbf{G}_{11}\end{pmatrix}\sigma_g^2\right] \tag{30}$$

We emphasise that the variance components here are the same for the training and validation set! The variance components are typically estimated from training set.

The joint distribution requires the marginal expectation of $\left(\begin{smallmatrix}\mathbf{y}_2\\ \mathbf{y}_1\end{smallmatrix}\right)$, as well as the marginal variances for $\mathbf{y}_2$ and $\mathbf{y}_1$ and the covariances between these two. The marginal expectation is, as in 'ordinary' G-BLUP, the fixed

effects, $\binom{\mathbf{X}_2\boldsymbol{\beta}_2}{\mathbf{X}_1\boldsymbol{\beta}_1}$. The marginal variances are easily derived from (**??**) and applying the indices for the training and validation set. The covariance will be derived as follows:

$$\begin{aligned}
\text{Cov}(\mathbf{y}_2, \mathbf{y}_1) &= \text{Cov}(\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{g}_2 + \mathbf{e}_2, \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{g}_1 + \mathbf{e}_1) && \text{(31a)}\\
&= \text{Cov}(\mathbf{g}_2 + \mathbf{e}_2, \mathbf{g}_1 + \mathbf{e}_1) && \text{(31b)}\\
&= \text{Cov}(\mathbf{g}_2, \mathbf{g}_1) + \text{Cov}(\mathbf{g}_2, \mathbf{e}_1) + \\
&\quad \text{Cov}(\mathbf{e}_2, \mathbf{g}_1) + \text{Cov}(\mathbf{e}_2, \mathbf{e}_1) && \text{(31c)}\\
&= \text{Cov}(\mathbf{g}_2, \mathbf{g}_1) = \mathbf{G}_{21}\sigma_g^2 && \text{(31d)}
\end{aligned}$$

Step (31c) is achieved by applying the last property in stat box 4 and the last three terms are set to zero as all residuals are independent and as well are genetic values and residuals.

With this in place, we can set up the join distribution as

$$\begin{bmatrix}\mathbf{y}_2\\ \mathbf{y}_1\end{bmatrix} \sim N\left[\begin{pmatrix}\mathbf{X}_2\boldsymbol{\beta}_2\\ \mathbf{X}_1\boldsymbol{\beta}_1\end{pmatrix}, \quad \begin{pmatrix}\mathbf{G}_{22}\sigma_g^2 + \mathbf{D}_n\sigma_e^2 & \mathbf{G}_{21}\sigma_g^2\\ \mathbf{G}_{12}\sigma_g^2 & \mathbf{G}_{11}\sigma_g^2 + \mathbf{D}_n\sigma_e^2\end{pmatrix}\right] \tag{32}$$

Hence, the expectation of $\mathbf{y}_2$ given $\mathbf{y}_1$ is, by applying stat box 3:

$$\text{E}[\mathbf{y}_2|\mathbf{y}_1] = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{G}_{21}\sigma_g^2[\mathbf{G}_{11}\sigma_g^2 + \mathbf{D}_n\sigma_e^2]^{-1}(\mathbf{y}_1 - \mathbf{X}_1\boldsymbol{\beta}_1) \tag{33}$$

This emphasises the importance of the relationship between the two groups and the variance component σ_g^2. If there is no or low relationship between the training and validation set, $\mathbf{G}_{21}$ is close to zero and the prediction of values of $\mathbf{y}_2$ is entirely controlled by the fixed effects. Likewise, if the genetic model cannot account for any of the observed variance, i.e. the variance component is close to zero, predicting with said genetic model is ... pointless.

Stat 1: Joint distribution of linear equations

Given a vector of multivariate normal random variables, $x \sim N(m, V)$, and a set of linearly independent linear functions, A, then $y^* = c + Ax$ is a vector of multivariate normal random variables, where c is a vector of constants. The joint distribution of y^* is

$$y^* \sim N[Am + c, AVA'].$$

Example: Using M-BLUP in (**??**), $y^* = y$, and we can set $A = (W\ I_n)$, $x = \binom{b}{e}$, $c = 1_n\mu$, and $m = 0$. The variance-covariance matrix for b and e is found by equations (**??**), giving $V = \left(\begin{smallmatrix} I_m\sigma_b^2 & 0\\ 0 & D_n\sigma_e^2\end{smallmatrix}\right)$. Thus, the joint distribution from above becomes:

$$\begin{aligned}
y &\sim N\left[0 + I_n\mu,\ \begin{pmatrix}W & I_n\end{pmatrix}\begin{pmatrix}I_n\sigma_g^2 & 0\\ 0 & D_n\sigma_e^2\end{pmatrix}\begin{pmatrix}W'\\ I_n\end{pmatrix}\right]\\
&= N\left[I_n\mu,\ WW'\sigma_g^2 + D_n\sigma_e^2\right]
\end{aligned}$$

which is identical to what we found in (**??**), section **??**.

References

Arthur S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *J. Am. Stat. Assoc.*, 57(298):369–375, 1962.

D. Habier, R. L. Fernando, and J. C. M. Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007. doi: 10.1534/genetics.107.081190. URL http://www.genetics.org/content/177/4/2389.abstract.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2009. ISBN 978-0-387-84857-0. doi: 10.1007/b94608.

C. R. Henderson. Estimation of changes in herd environment. *J. Dairy Sci.*, 32(8):706 (Abstract), 1949. doi: 10.3168/jds.S0022-0302(49)92104-9.

C. R. Henderson, Oscar Kempthorne, S. R. Searle, and C. N. von Krosigk. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, 15(2):192–218, 1959. doi: 10.2307/2527669. URL http://www.jstor.org/stable/10.2307/2527669.

Charles R. Henderson. Selection index and expected genetic advance. In Warren Durward Hanson and Harold Frank Robinson, editors, *Statistical Genetics and Plant Breeding*, volume 982, chapter 141-163, page 623. National Academy of Sciences – National Research Council, Washington, DC, 1963.

A. Legarra and I. Misztal. Technical note: Computing strategies in genome-wide selection. *J. Dairy Sci.*, 91(1):360–6, January 2008. ISSN 1525-3198. doi: 10.3168/jds.2007-0403. URL http://www.ncbi.nlm.nih.gov/pubmed/18096959.

Andrés Legarra, Christèle Robert-Granié, Eduardo Manfredi, and Jean-Michel Elsen. Performance of genomic selection in mice. *Genetics*, 180(1):611–8, September 2008. ISSN 0016-6731. doi: 10.1534/genetics.108.088575. URL http://www.genetics.org/content/180/1/611.long.

Tu Luan, John A. Woolliams, Sigbjørn Lien, Matthew Kent, Morten Svendsen, and Theo H. E. Meuwissen. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics*, 183:1119–1126, 2009. ISSN 0016-6731. doi: 10.1534/genetics.109.107391.

Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, USA, 1998. ISBN 0-87893-481-2.

Ulrike Ober, Julien F Ayroles, Eric A Stone, Stephen Richards, Dianhui Zhu, Richard A Gibbs, Christian Stricker, Daniel Gianola, Martin Schlather, Trudy F C Mackay, and Henner Simianer. Using whole-genome sequence data to predict quantitative trait phenotypes in *Drosophila melanogaster*. *PLoS Genet.*, 8(5):e1002685, January 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002685. URL http://dx.plos.org/10.1371/journal.pgen.1002685.

Annalisa Orenti, Giuseppe Marano, Patrizia Boracchi, and Ettore Marubini. Pinpointing outliers in experimental data: the Hat matrix in Anova for fixed and mixed effects models, 2012. ISSN 1723-7815. URL http://ijphjournal.it/article/view/8663.

Stat 2: A joint distribution

We take notice in standard probability theory, that

$$P(X,Y) = P(X|Y)P(Y)$$

hence the conditional expectation of X on Y is dependent on the marginal distribution of Y and the joint distribution of both variables.

Stat 3: Conditional expectation of a bivariate normal distribution

Given:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}\right]$$

Conditional expectation is

$$\mathrm{E}[X_1|X_2] = \mu_1 + V_{12}V_{22}^{-1}(X_2 - \mu_2)$$

Stat 4: Properties of covariances

If x, y, w and v are random variables, and a, b, c and d are non-random (i.e. constant), the following applies:

$$\begin{aligned} \mathrm{Cov}(x, a) &= 0 \\ \mathrm{Cov}(x, y) &= \mathrm{Cov}(y, x) \\ \mathrm{Cov}(ax, by) &= ab\,\mathrm{Cov}(x, y) \\ \mathrm{Cov}(a + x, b + y) &= \mathrm{Cov}(x, y) \\ \mathrm{Cov}(ax, +by, cw + dv) &= ac\,\mathrm{Cov}(x, w) + ad\,\mathrm{Cov}(x, v) + \\ &\quad bc\,\mathrm{Cov}(y, w) + bd\,\mathrm{Cov}(y, v) \end{aligned}$$

G. K. Robinson. That BLUP is a Good Thing: The Estimation of Random Effects. *Stat. Sci.*, 6(1):15–32, February 1991. ISSN 2168-8745. doi: 10.1214/ss/1177011926. URL http://projecteuclid.org/euclid.ss/1177011926.