Linear Mixed Models used in Quantitative Genomics

Stefan McKinnon Høj-Edwards & Peter Sørensen

2022-03-28

Contents

1 Linear mixed models **1**

2 Linear mixed model for marker effects (M-BLUP) **2**

2.1 Assumptions 2

2.2 Covariances 3

3 Linear mixed model for individual effects (G-BLUP) **3**

3.1 Assumptions 3

3.2 Covariances 3

4 Proof of equivalence of M-BLUP and G-BLUP **4**

5 Some notes on calculation of W **5**

6 Expansion of G-BLUP and M-BLUP to handle multiple marker sets **6**

1 Linear mixed models

Linear mixed effect models (LMMs) are widely used in genetics and application includes single marker association analysis, estimation of genetic variance and heritability, and prediction of genetic predisposition or disease risk. In this chapter we will start with a general introduction to the linear mixed models to establish the notation and assumptions used throughout the notes. Two models that are used to model the relationship between genotype and phenotype are introduced, but the application and usage is reserved for chapter **??**. We will show that the two models are equivalent and then expand the models to contain two (or more) random effects (section 6), followed by a comparison of the models used in these notes (section **??**).

The two models are M-BLUP, which models the *marker effects* of each observed genetic marker, and G-BLUP, which models the *genetic values* of an individual. 'BLUP' in this context is an abbreviation for Best Linear Unbiased Predictor, and the solutions to these two BLUP models are shown in the 'BLUP' chapter of these notes.

It is important here to understand the distinction between what is known as the 'true model' and an 'instrumental model'. The true model is what generated the data; in the context of these notes, this would be the biological machinery of genes being expressed and that ultimately produce the observed phenotypes.

As we are still trying to infer the workings of this complex biological machinery, we instead refer to an instrumental model that reflects our understanding and what we are trying to compute.

For the following, we refer to the number of genotyped individuals as q, number of observations as n, and number of markers as m. To introduce the two models, we are starting with simplified versions of the models, assuming one observation per individual. This removes the necessity for the Z matrix, which will be introduced in chapter **??**. We also assume that fixed effects have been reduced to a single intercept, i.e. $X\beta = \mu$.

2 Linear mixed model for marker effects (M-BLUP)

The M-BLUP linear mixed model is a simple starting point, assuming the linear combination of marker effects for each individual.

$$\mathbf{y} = \mu + \mathbf{W}\mathbf{b} + \mathbf{e} \tag{1}$$

where

- $\mathbf{y}$ is the n-length vector of observations and is the *linear combination* of the *random variables* $\mathbf{b}$ and $\mathbf{e}$ plus
- μ the intercept.
- $\mathbf{W}$ is the $n \times m$ genotype matrix[1],
- $\mathbf{b}$ is the m-length vector of marker effects, and
- $\mathbf{e}$ is the n-length residual vector.

The genotype matrix $\mathbf{W}$ is the (scaled and centred) marker matrix, linking each of the q individuals to the genotype at each of the m loci. See section 5 for a discussion of this matrix.

2.1 Assumptions

The M-BLUP assumes a priori that the *marker effects* $\mathbf{b}$ are uncorrelated, i.e. $\mathbf{b} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_b^2\right)$, where σ_b^2 is the variance component and $\mathbf{I}$ is a $m \times m$ identity matrix. That is, the marker effects are assumed to be independently sampled from a normal distribution with mean 0 and variance σ_b^2. Do not confuse this assumption with that the markers themselves might or might not be correlated!

Residuals are assumed uncorrelated, but might be weighted, i.e. $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{D}\sigma_e^2\right)$ where $\mathbf{D}$ is a $n \times n$ diagonal matrix and each diagonal element may take a value corresponding to the uncertainty of the observation. The residuals and marker effects are uncorrelated, as displayed in (3c).

Expectations of $\mathbf{y}$ are $\mathrm{E}(\mathbf{y}|\mu) = \mu$, as the expectation of the random variables are 0.

Variance is given as

$$\begin{aligned}
\mathrm{Var}(\mathbf{y}|\mu) &= \mathrm{Var}(\mathbf{W}\mathbf{b} + e|\mu) && \text{(2a)}\\
&= \mathrm{Var}(\mathbf{W}\mathbf{b}) + \mathrm{Var}(\mathbf{e}) + \mathrm{Cov}(\mathbf{W}\mathbf{b}, \mathbf{e}) + \mathrm{Cov}(\mathbf{e}, \mathbf{W}\mathbf{b}) && \text{(2b)}\\
&= \mathbf{W}\mathbf{W}'\,\mathrm{Var}(\mathbf{b}) + \mathrm{Var}(\mathbf{e}) + 0 + 0 && \text{(2c)}\\
&= \mathbf{W}\mathbf{W}'\mathbf{I}\sigma_b^2 + \mathbf{D}\sigma_e^2 && \text{(2d)}
\end{aligned}$$

Step 2c is possible because the two random variables are uncorrelated.

[1] Generally, $\mathbf{W}$ has q rows, one row for each individual.

2.2 Covariances

Covariances can be expressed as

$$\mathrm{Cov}(\mathbf{y}, \mathbf{b}) = \mathbf{W}\sigma_b^2 \tag{3a}$$
$$\mathrm{Cov}(\mathbf{y}, \mathbf{e}) = \mathbf{D}\sigma_e^2 \tag{3b}$$

which can be summarised as

$$\mathrm{Var}\begin{pmatrix}\mathbf{b}\\ \mathbf{e}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{I}\sigma_b^2 & \mathbf{0} & \mathbf{W}'\sigma_b^2\\ \mathbf{0} & \mathbf{D}\sigma_e^2 & \mathbf{D}\sigma_e^2\\ \mathbf{W}\sigma_b^2 & \mathbf{D}\sigma_e^2 & \mathbf{W}\mathbf{W}'\mathbf{I}\sigma_b^2 + \mathbf{D}\sigma_e^2\end{pmatrix}. \tag{3c}$$

3 Linear mixed model for individual effects (G-BLUP)

This model is similar to M-BLUP, although we are instead modelling the *genetic values* (sometimes referred to as Genomic Estimated Breeding Values; GEBV) instead of the marker effects. We assume the connection between marker effects and genetic values is $\mathbf{Wb} = \mathbf{g}$. There are some computational advantages for using G-BLUP when $n \ll m$, but this will be covered later.

The G-BLUP model can be written as

$$\mathbf{y} = \mu + \mathbf{g} + \mathbf{e} \tag{4}$$

where $\mathbf{y}$, μ and $\mathbf{e}$ are as for (1), and $\mathbf{g}$ is the q-length vector of genetic values.

3.1 Assumptions

In G-BLUP we make no (explicit) assumptions on the marker effects, instead we assume that the genetic values are correlated by the relationship between the individuals, i.e. by pedigree or genetic similarity, such that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2)$, where $\mathbf{G}$ is the $q \times q$ genomic relationship matrix; see remarks in section~5.

For the observations ($\mathbf{y}$) and residuals ($\mathbf{e}$), the assumptions are as in section~2.1. Variance is given as

$$\begin{aligned}\mathrm{Var}(\mathbf{y}|\mu) &= \mathrm{Var}(\mathbf{g} + e|\mu) && (5a)\\ &= \mathrm{Var}(\mathbf{g}) + \mathrm{Var}(\mathbf{e}) + \mathrm{Cov}(\mathbf{g}, \mathbf{e}) + \mathrm{Cov}(\mathbf{e}, \mathbf{g}) && (5b)\\ &= \mathbf{G}\sigma_g^2 + \mathbf{D}\sigma_e^2. && (5c)\end{aligned}$$

As in M-BLUP, step 5c is possible due to the two random variables are assumed uncorrelated.

3.2 Covariances

Covariances for G-BLUP can be expressed as

$$\mathrm{Cov}(\mathbf{y}, \mathbf{g}) = \mathbf{G}\sigma_g^2 \tag{6a}$$
$$\mathrm{Cov}(\mathbf{y}, \mathbf{e}) = \mathbf{D}\sigma_e^2 \tag{6b}$$

and the entire model can be summarised as

$$\mathrm{Var}\begin{pmatrix}\mathbf{g}\\ \mathbf{e}\\ \mathbf{y}\end{pmatrix} = \begin{pmatrix}\mathbf{G}\sigma_g^2 & \mathbf{0} & \mathbf{G}'\sigma_g^2\\ \mathbf{0} & \mathbf{D}\sigma_e^2 & \mathbf{D}\sigma_e^2\\ \mathbf{G}\sigma_g^2 & \mathbf{D}\sigma_e^2 & \mathbf{G}\sigma_g^2 + \mathbf{D}\sigma_e^2\end{pmatrix}. \tag{6c}$$

4 Proof of equivalence of M-BLUP and G-BLUP

To prove that G-BLUP and M-BLUP are equivalent, we must show that the expectation and variance of the models are identical. In the two sections above, the expectations are shown to be $\mathrm{E}(\mathbf{y}|\mu) = \mu$, and the variances are derived in (2d) and (5c). We must therefore show the following equivalence:

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

Assuming that $\mathbf{G}$ is calculated[2] as $\frac{\mathbf{W}\mathbf{W}'}{m}$, cancelling the component from the residual, and dropping the identity, we get

$$m \cdot \sigma_b^2 = \sigma_g^2 \tag{7}$$

Alas, we are left to show that the two variance components differ by a factor of m. Assuming that the columns of genotype matrix W has been centred and scaled [VanRaden, 2008], we can assume

$$\mathbf{W} \sim (0, 1)$$

That is, that $\mathbf{W}$ is sampled from an unknown distribution with mean zero and variance 1. The genetic value for the ith individual is the linear combination of the marker effects, i.e.

$$g_i = \mathbf{w}_i\mathbf{b} = \sum_j^m w_{ij} b_j \tag{8}$$

where $\mathbf{w}_i$ is the ithrow vector of $\mathbf{W}$ and j is the jthlocus. Thus the conditional expectation of g_i given $\mathbf{W}$ is

$$\mathrm{E}(g_i|\mathbf{w}_i) = \mathbf{w}_i\,\mathrm{E}(\mathbf{b}) = \mathbf{w}_i \cdot 0 = 0 \tag{9}$$

and the conditional variance is

$$\begin{aligned}\mathrm{Var}(g_i|\mathbf{w}_i) &= \mathbf{w}_i\,\mathrm{Var}(\mathbf{b})\mathbf{w}_i' \\ &= \mathbf{w}_i\mathbf{w}_i'\sigma_b^2 \\ &= \left[\sum_j^m w_{ij}\right]\sigma_b^2\end{aligned}$$

[2]The assumptions and implications of this are too cumbersome to discuss here, so the reader is directed to section 5 for a discussion on the calculation of $\mathbf{W}$ and $\mathbf{G}$.

Hence, the marginal variance of g_i is:

$$\begin{aligned}
\mathrm{Var}(g_i) &= \mathrm{Var}_{\mathbf{w}_i}[\mathrm{E}(g_i|\mathbf{w}_i)] + \mathrm{E}_{\mathbf{w}_i}[\mathrm{Var}(g_i|\mathbf{w}_i)] && \text{(10a)}\\
&= 0 + \mathrm{E}_{\mathbf{w}_i}[\mathrm{Var}(g_i|\mathbf{w}_i)] && \text{(10b)}\\
&= \mathrm{E}_{\mathbf{w}_i}\left[\sigma_b^2 \sum_j^m w_{ij}^2\right] && \text{(10c)}\\
&= \sigma_b^2 \sum_j^m \mathrm{E}(w_{ij}^2) && \text{(10d)}\\
&= \sigma_b^2 \sum_j^m \left[\mathrm{Var}(w_{ij}) + [\mathrm{E}(w_{ij})]^2\right] && \text{(10e)}\\
&= \sigma_b^2 \sum_j^m [1+0] && \text{(10f)}\\
\sigma_g^2 &= \sigma_b^2 m && \text{(10g)}
\end{aligned}$$

Expanding the expectation of a squared variable in step (10d) is done by the rule displayed in the stat box 6. The notation of $\mathrm{E}_{\mathbf{w}_i}$ and $\mathrm{Var}_{\mathbf{w}_i}$ indicates the expectation and variance are with respect to the *distribution* of $\mathbf{w}_i$.

5 Some notes on calculation of W

The genotype matrix $\mathbf{W}$ can be designed in several ways. The overall assumption is that it links the q genotyped individuals to the m markers (after quality control, etc.). It is therefore an $q \times m$ matrix, where each row corresponds to an individual and each column to a marker.

The most basic matrix is the *marker count matrix* $\mathbf{M}$ or *minimal allele count matrix*, where each element in $\mathbf{M}$ takes the value 0, 1 or 2 (in diploid species). Legarra and Misztal [2008] use an allele coding of -1, 0 and 1, and in this case, diagonals of $\mathbf{MM}'$ count the number of homozygous loci per individual, and diagonals of $\mathbf{M}'\mathbf{M}$ count the homozygous individuals per locus [VanRaden, 2008].

Choice of allele coding should not affect the statistical inference of variance, but it can affect the reliabilities of estimated breeding values [**?**]. Besides the papers mentioned, **?** also has a useful discussion of the coding and the consequences in e.g. estimating breeding values.

Here, unless otherwise noted, each column vector ($\mathbf{w}_i$) of the genotype matrix $\mathbf{W}$ is defined as

$$\mathbf{w}_i = \frac{\mathbf{m}_i - \mathrm{Mean}(\mathbf{m}_i)}{\sqrt{\mathrm{Var}(\mathbf{m}_i)}} \tag{11a}$$

or equivalently

$$\mathbf{w}_i = \frac{\mathbf{m}_i - 2p_i}{\sqrt{2p_i(1-p_i)}} \tag{11b}$$

where p_i is the allele frequency of the ith marker[3].

This calculation *scales and centres* each column of the resulting $\mathbf{W}$ matrix to expectation equal 0 and variance equal 1.

Setting the mean equal to 0 is done in order to give more emphasis to the rare alleles which are thought to have a larger impact on most of the traits of interest. Another consequence is that the resulting genomic

[3]The allele frequency is preferable from an unselected base population [VanRaden, 2008].

relationship matrix better describes family relationships, as the rare alleles usually only exist within closely related individuals. The resulting genomic relationship matrix $\mathbf{G}$ is calculated as in VanRaden [2008]:

$$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m} \tag{12}$$

Another method widely used to calculate the genomic relationship matrix is explained in VanRaden [2008]. Instead of centring and scaling each column of $\mathbf{M}$ to obtain $\mathbf{W}$, the scaling is done on the relationship matrix using the total variance, ν_p.:

$$\mathbf{w}_i^* = \mathbf{m}_i - 2p_i \tag{13}$$

$$\mathbf{G}^* = \frac{\mathbf{W}^*\mathbf{W}^{*\prime}}{\nu_p} \tag{14}$$

where $\nu_p = 2\sum_i p_i(1-p_i)$. The resulting $\mathbf{G}^*$ matrix has been scaled by the total variance at all loci (ν_p) rather than scaling each locus (i.e. column) by their own variance.

The last method produces a genomic relationship matrix having the same scale as the Numerator Relationship Matrix, $\mathbf{A}$, calculated using the pedigree. This means that the genomic inbreeding coefficient for an individual i can be calculated as $\mathbf{G}_{ii}^* - 1$, and the genomic relationship among two individuals j and k as $\mathbf{G}_{jk}^*/(\sqrt{\mathbf{G}_{jj}^*} + \sqrt{\mathbf{G}_{kk}^*})$ [VanRaden, 2008]. Further, Ober et al. [2012] showed that $\mathrm{E}(\mathbf{G}^*) = \mathbf{A}.\%$, confirming all the properties listed above.

Table 1: Two forms of genotype matrix $\mathbf{W}$.

Variant	Calculation	Relationship	Properties
$\mathbf{W}$	$\mathbf{w}_i = \frac{\mathbf{m}_i - 2p_i}{\sqrt{2p_i(1-p_i)}}$	$\mathbf{G} = \frac{\mathbf{W}\mathbf{W}'}{m}$	$\mathbf{W} \sim (0,1)$
$\mathbf{W}^*$	$\mathbf{w}_i^* = \mathbf{m}_i - 2p_i$	$\mathbf{G}^* = \frac{\mathbf{W}^*\mathbf{W}^{*\prime}}{\nu_p}$	$\mathbf{W} \sim (0,?)$

$\nu_p = 2\sum_i p_i(1-p_i)$

We note here that when utilising $\mathbf{W}$ to construct the genomic relationship matrix, the markers with high heterozygosity are weighted more than rare alleles due to the scaling by $\sqrt{2p_i(1-p_i)}$. %With the alternative genotype matrix, W^*, all markers are weighted equally.

As noted in the beginning of this section, the allele coding should not affect the estimated variance [**?**]. The scaling, however, will. Furthermore, the M-BLUP model assumes a priori that the estimated marker effects are sampled from the same distribution. If we had any notion that some of the markers might be more informative towards a complex trait, we have two choices: we can specify a weight for each marker based on e.g. results from a GWAS, or, if we want to isolate the contribution from a set of markers, we can extend the BLUP models to having two random effects. The latter is described in the following section.

6 Expansion of G-BLUP and M-BLUP to handle multiple marker sets

Here, we will expand M-BLUP and G-BLUP to having two random genetic effects, indexed by 1 and 2. Assumptions for these models are generally as above, and in the following example we have split the markers into two groups 1 and 2. We start with the equations:

$$\mathbf{y} = \mu + \mathbf{W}_1\mathbf{b}_1 + \mathbf{W}_2\mathbf{b}_2 + \mathbf{e} \tag{15a}$$
$$\mathbf{y} = \mu + \mathbf{g}_1 + \mathbf{g}_2 + \mathbf{e} \tag{15b}$$

where the variables are as in (1) and (4), but indexed for marker group 1 and 2.

We can summarise the distributions of these as

$$\begin{pmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \mathbf{e}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{I}_{m_1}\sigma^2_{b_1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_{m_2}\sigma^2_{b_2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{D}_n\sigma^2_e\end{pmatrix}\right] \tag{16a}$$

for M-BLUP and for G-BLUP as

$$\begin{pmatrix}\mathbf{g}_1\\ \mathbf{g}_2\\ \mathbf{e}\end{pmatrix} \sim N\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{G}_1\sigma^2_{g_1} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}_2\sigma^2_{g_2} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{D}_n\sigma^2_e\end{pmatrix}\right] \tag{16b}$$

where G_1 and G_2, respectively, are genetic relationship matrices constructed on the subsets of markers, respectively, instead of all markers, and are a priori assumed uncorrelated.

The keen observer will notice that these are very similar to those in (3c) and (6c), although here, the two random variables $\mathbf{b}_1$ and $\mathbf{b}_2$ (and $\mathbf{g}_1$ and $\mathbf{g}_2$) are assumed completely independent, an assumption that might be disputed.

Stat 1: Marginalising conditional expectations

$$\begin{aligned} E(X) &= E_Y\left[E(X|Y)\right] \\ Var(X) &= E(X^2) - \left[E(X)\right]^2 \\ &\Rightarrow E(X^2) = Var(X) + \left(E(X)\right)^2 \end{aligned}$$

References

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.

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.

P.M. VanRaden. Efficient methods to compute genomic predictions. *J. Dairy Sci.*, 91(11):4414–23, November 2008. ISSN 1525-3198. doi: 10.3168/jds.2007-0980. URL http://www.ncbi.nlm.nih.gov/pubmed/18946147;http://www.journalofdairyscience.org/article/S0022-0302(08)70990-1/abstract.