Chapter 5

Variance Components Estimation

In applied prediction of breeding values using BLUP animal models, variance components for all random effects are required as input. These variance components must be estimated from the data. In more detail, given the assumed linear mixed effects model

$$y = Xb + Zu + e \tag{5.1}$$

where y is a vector of length N with observations, b is an unknown vector of length p of fixed effects, u is an unknown vector of length q of random breeding values and e is an unknown vector of length N of random residuals. The matrices X and Z are known design matrices linking the corresponding effects to the observations. As part of the model specification, the variances of the random effects are defined as

$$\begin{aligned} var(u) &= A\sigma_u^2 \\ var(e) &= I\sigma_e^2 \end{aligned} \tag{5.2}$$

In (5.2) σ_u^2 and σ_e^2 are the unknown variance components that must be estimated from the data. The matrix A is the numerator relationship matrix that can be constructed based on the pedigree and I is the identity matrix. The material presented in this chapter is based on [Essl, 1987] and [Searle et al., 1992] and it shows different methods how variance components for different models can be estimated.

5.1 Estimation Of Genetic Components

For each trait that should be considered in an aggregate genotype, the first thing to be analysed is whether the observed variability in the phenotypic values of the trait can be partly explained by a genetic component. Because only traits with a detectable genetic component can be used for improving a population on the genetic level. The genetic component quantifies the part of the phenotypic variability which is passed from parents to offspring. Hence from a livestock breeding point of view, the ratio between the genetic variability (quantified by σ_u^2) and the phenotypic variability (measured by σ_p^2) is important and is termed as **heritability** (h^2).

$$h^2 = \frac{\sigma_u^2}{\sigma_p^2} \tag{5.3}$$

One first method that we want to introduce is based on the very well-known statistical technique called **analysis of variance** (ANOVA). ANOVA is shown in the next subsection for a simple application of estimating repeatability for a dataset with repeated observations of the same trait. Later this can be generalized to the estimation of the variability due to genetic components.

5.2 Estimation Of Repeatability

The term **repeatability** indicates how similar repeated measurements of the same quantity are. For example, if we measure the same trait on any given animal multiple times, the measurements are expected to vary. But because the measurements are done on the same animal, the variability is probably smaller compared to measurements from different animals. This phenomenon can be quantified by a ratio of variance components which is called repeatability.

The computation of the repeatability is shown using the following example dataset from 10 randomly selected bulls. From each bull the shoulder height is measured three times.

Table 5.1: Repeated Measurements of Shoulder Height in cm

Bull	M1	M2	M3
1	135	136	134
2	129	130	128
3	135	133	136
4	127	127	125
5	126	129	129
6	128	129	128
7	127	132	130

8	129	128	125
9	126	125	127
10	132	131	134

Now we want to check whether the measurements for the same bull have a smaller variability compared to measurements from different bulls. We first create a plot which might already give us some indications.

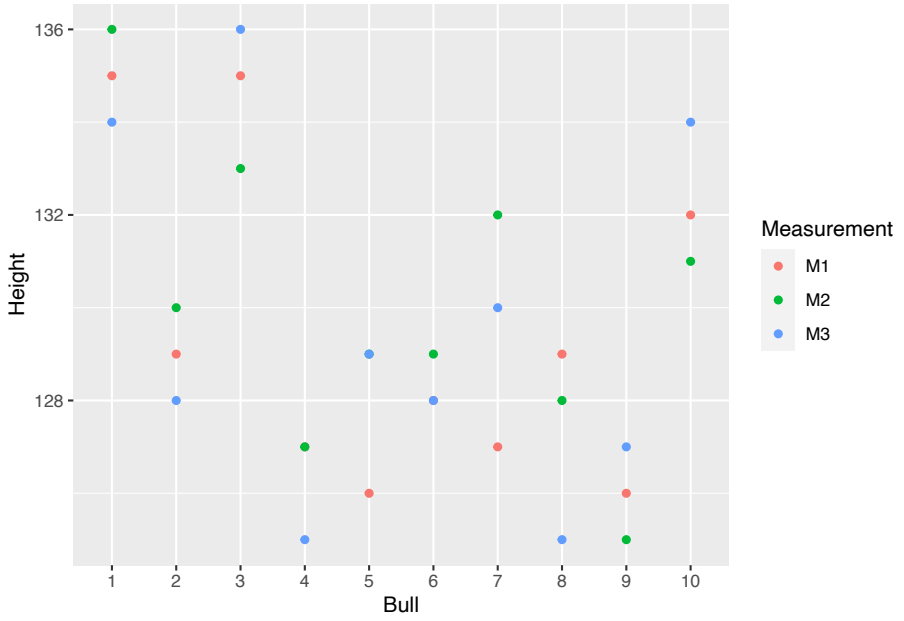

Figure 5.1: Repeated Measurements of Shoulder Height for Ten Bulls

From Figure 5.1 alone, it is difficult to say whether measurements for the same animal are more similar than measurements from different animals. We use the following model to provide a quantitative answer for the previously formulated question.

$$y_{ij} = \mu + t_i + \epsilon_{ij} \tag{5.4}$$

where

y_{ij}	measurement j of animal i
μ	expected value of y
t_i	deviation of y_{ij} from μ attributed to animal i
ϵ_{ij}	measurement error

5.2.1 Estimation

Given the definition of t_i and ϵ_{ij} as random effects, the following relationships hold

- $E(t_i) = 0$
- $\sigma_t^2 = E(t_i^2)$: variance component of total variance (σ_y^2) which can be attributed to the t-effects
- $E(\epsilon_{ij}) = 0$
- $\sigma_\epsilon^2 = E(\epsilon_{ij}^2)$: variance component attributed to ϵ-effects
- $\sigma_y^2 = \sigma_t^2 + \sigma_\epsilon^2$

The repeatability w is defined as the following ratio between variance components

$$w = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\epsilon^2} \tag{5.5}$$

The variance components σ_t^2 and σ_ϵ^2 are estimated using an analysis of variance. The result of such an analysis is shown in the following table.

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Bull         9  286.7   31.85   13.85 8.74e-07 ***
## Residuals   20   46.0    2.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the theory of analysis of variance the expected values of the mean sum of squares can be equated to the following variance components.

Effect	$E(MeanSq)$
Bull	$\sigma_\epsilon^2 + n * \sigma_t^2$
Error	σ_ϵ^2
Total	$\sigma_\epsilon^2 + \frac{N-n}{N-1} * \sigma_t^2$

where n is the number of measurement per bull and N is the total number of measurements.

The numeric values of the compute `Mean Sq` values are now taken as estimates for the respective variance components. Therefore

$$\hat{\sigma}_\epsilon^2 = 2.3$$

and

$$\hat{\sigma}_t^2 = \frac{31.85 - 2.3}{3} = 9.85$$

The estimated repeatability can now be computed as

$$\widehat{w} = \frac{\widehat{\sigma}_t^2}{\widehat{\sigma}_t^2 + \widehat{\sigma}_\epsilon^2} = 0.81$$

5.3 Estimation Of Sire Variance

The technique of estimating variance components using ANOVA can also be applied to a data set where offspring performance records are grouped by their sires using a sire model. From the statistical point of view a sire model is a linear mixed effects model for each observation, the effect of the sire is expressed by a random effect. In matrix vector notation this model can be written as

$$y = Xb + Zs + e \tag{5.6}$$

where y is a vector of length N of observations, b is a vector of length p of fixed effects, s is a vector of length r with random sire effects and e is a vector of length N of random error terms. The matrices X and Z are incidence matrices for b and s, linking the respective effects to the observations. An example of such a data set is used in Problem 1 of Exercise 3.

The variance component σ_s^2 for the random sire component s is estimated the same way as shown in subsection 5.2 using an ANOVA table. For the sire model the ANOVA table has the following structure

Effect	Degrees of Freedom	Sum Sq	Mean Sq	$E(Mean\ Sq)$
Sire $(s\|b)$	$r-1$	$SSQ(s\|b)$	$SSQ(s\|b)/(r-1)$	$\sigma_e^2 + k * \sigma_s^2$
Residual (e)	$N-r$	$SSQ(e)$	$SSQ(e)/(N-r)$	σ_e^2

where

$$SSQ(s|b) = SSQ(sb) - SSQ(b)$$

$$SSQ(sb) = \sum_{i=1}^{r} \left[\left(\sum_{j=1}^{n_i} y_{ij} \right)^2 / n_i \right]$$

$$SSQ(b) = \left(\sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij} \right)^2 / N$$

$$SSQ(e) = SSQ(y) - SSQ(sb)$$

$$SSQ(y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} y_{ij}^2$$

$$k = \frac{1}{r-1}\left[N - \frac{\sum_{i=1}^{r} n_i^2}{N}\right]$$

with r the number of sires and n_i the number of progeny for sire i.

The numeric computation of estimating σ_s^2 and σ_e^2 is the topic of Problem 1 of Exercise 3. The dataset that is used in Exercise 3 is a simplified version where only certain genetic relationships occur and where the number of environmental effects are kept at a very low number. To address the higher complexity of real-world datasets obtained in the field, other methods have been developed. Furthermore the ANOVA-based techniques when applied to real data can produce negative estimates for variance components. Because variance components are on a quadratic scale, they cannot be negative and from negative variances, the standard deviations are not defined in the scope of real numbers. Hence negative variance component estimates are outside of the parameter domain.

5.4 Development Of Further Methods

In this subsection, we focus on methods which are still used today. The currently used methods for variance components estimation are either based on Likelihood approaches or are the result of a so-called Bayesian procedure.

5.4.1 Maximum Likelihood

The first maximum likelihood approach to estimate variance components for linear mixed effects models was developed by [Hartley and Rao, 1967]. As the term `maximum likelihood` implies it, the presented method is based on the likelihood L where L is defined as

$$L(\theta) = f(y|\theta) \tag{5.7}$$

where θ is the vector of all unknown parameters to be estimated. For the linear mixed effect model

$$y = Xb + Zu + e$$

and under the assumption of the data being normally distributed, [Hartley and Rao, 1967] specify L as

$$L(\theta) = (2\pi)^{-1/2n}\sigma^{-n}|H|^{-1/2} * exp\left\{-\frac{1}{2\sigma^2}(y - Xb)^T H^{-1}(y - Xb)\right\} \tag{5.8}$$

where $var(y) = H\sigma^2 = Z^T GZ + R$ with $var(u) = I\sigma_u^2$ and $var(e) = R = I\sigma^2$. The maximum likelihoods for σ_u^2 and σ^2 are the values that maximize the function likelihood function L. It has to be noted that in (5.8) not only the variance components, but also the fixed effects b are unknown. These must also be estimated from the data.

The maximization of L is done by taking the partial derivatives of $\lambda = logL$ with respect to all unknown parameters. Then these partial derivatives are set to 0 and the resulting solutions are taken as maximum likelihood estimates.

The problem with the just described maximum likelihood approach is that the unknown fixed effects b have to be estimated at the same time. As a consequence of that the maximum likelihood estimates of the variance components depend on b. This is considered as an undesirable property. The solution for this problem was developed by [Patterson and Thompson, 1971] and is called **Restricted Maximum Likelihood** (REML). In REML the observations y are transformed as Sy and Qy with the following properties

(i) The matrix S has rank $n - t$ and the matrix Q has rank t
(ii) The result of the two transformations are independent, that means $cov(Sy, Qy) = 0$ which is met when $SHQ^T = 0$
(iii) The matrix S is chosen such that $E(Sy) = 0$ which means $SX = 0$
(iv) The matrix QX is of rank t, so that every linear function of the elements of Qy estimate a linear function of b.

From (i) and (ii) it follows that the likelihood L of y is the product of the likelihoods of Sy and Qy that means

$$\lambda = \lambda' + \lambda''$$

Suitable matrices S and Q are given by

$$S = I - X(X^T X)^{-1} X^T$$

and

$$Q = X^T H^{-1}.$$

With these transformations, the variance components σ^2 and σ_u^2 can be estimated by maximizing λ' which is the logarithm of the likelihood of Sy and is independent of any influence of the fixed effects b. Based on this property, REML is the de-facto standard for variance components estimation in applied livestock breeding. The R-packages `lme4` or `pedigreemm` can be used to get estimates for variance components using either Maximum Likelihood (ML) or REML.

5.5 Bayesian Procedures

Theoretical foundations for using Bayesian methods in animal breeding were laid by [Gianola and Fernando, 1986]. These foundations spanned more than just the topic of variance components. A detailed implementation scheme for a mixed linear effects model using Gibbs sampling for datasets originating in the area of livestock breeding was first described by [Wang et al., 1993].

5.5.1 The Gibbs Sampler For The Gaussian Mixed Linear Model

This subsection summarises the most important results of [Wang et al., 1993].

5.5.1.1 Model

The univariate mixed linear effects model with a vector b of p fixed effects and a vector u of q random breeding values is considered.

$$y = Xb + Zu + e$$

where y is a vector of length n containing the data. The vector e (length n) is a vector of random residuals. The matrices X $(n \times p)$ and Z $(n \times q)$ are incidence matrices linking fixed effects and random effects to observations.

The conditional distribution that generates the data is given by

$$y|b, u, \sigma_e^2 \sim \mathcal{N}(Xb + Zu, I\sigma_e^2)$$

where I is a $n \times n$ Identity matrix and σ_e^2 is the variance of the random residuals.

5.5.1.2 Prior Distributions

In a Bayesian analysis all unknowns must be assigned a prior distribution. In our case of the mixed linear effects model the unknowns are b, u, σ_e^2 and σ_u^2. Usually flat priors are assumed for the fixed effects b. Hence

$$p(b) \propto c$$

where c is a constant that does not depend on b. Further, the random effect u are assumed to follow a normal distribution, i.e.,

$$u|G, \sigma_u^2 \sim N(0, G * \sigma_u^2)$$

where σ_u^2 is the variance of the prior distribution of u and G is a known matrix. In the case of livestock breeding, G is the additive numerator relationship matrix.

The priors of the variance components σ_e^2 and σ_u^2 were assumed to be independent scaled inverted chi-square (χ^2) distributions such that

$$p(\sigma_e^2|\nu_e, s_e^2) \propto (\sigma_e^2)^{-\nu_e/2-1} exp(-\frac{1}{2}\nu_e s_e^2/\sigma_e^2)$$

and

$$p(\sigma_u^2|\nu_u, s_u^2) \propto (\sigma_u^2)^{-\nu_u/2-1} exp(-\frac{1}{2}\nu_u s_u^2/\sigma_u^2)$$

The quantities ν_e, ν_s, s_e^2 and s_u^2 are called hyper-parameters and must either be assumed based on experience from previous analyses or based on reasonable assumptions.

5.5.1.3 Joint Posterior Density

First, we have to introduce some additional notation. Let

$$\theta^T = (b^T, u^T) = (\theta_1, \theta_2, ..., \theta_N)$$

with $N = p+q$. The vector θ without the i^{th} element θ_i is denoted by θ_{-i} where

$$\theta_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_N)$$

Further, let

$$s^T = (s_u^2, s_e^2)$$

and

$$\nu^T = (\nu_u, \nu_e)$$

The joint posterior distribution can be written as

$$p(\theta, \sigma_u^2, \sigma_e^2|y, s, \nu) \propto p(\theta) * p(\sigma_u^2|\nu_u, s_u^2) * p(\sigma_e^2|\nu_e, s_e^2) * p(y|\theta, \sigma_e^2)$$

The above determined distributions can now be plugged into the joint posterior. The result of this is not shown here. The joint posterior is then used to determine the fully conditional densities which are the building blocks of the Gibbs sampler.

5.5.1.4 Fully Conditional Posterior Densities

Fully conditional densities for each of the unknown components are determined from the joint posterior distribution by regarding all other components as known. Let the matrix $W = w_{ij}$ with $i, j = 1, ..., N$ and the vector $r = r_i$ with $i = 1, ...N$ be the coefficient matrix and the right-hand side of the mixed model equations respectively, then the conditional density of each element θ_i in the vector θ follows a normal distribution with

$$\theta_i | y, \theta_{-i}, \sigma_u^2, \sigma_e^2, s, \nu \sim \mathcal{N}(\tilde{\theta}_i, \tilde{v}_i)$$

where $\tilde{\theta}_i = (r_i - \sum_{j=1, j \neq i}^{N} w_{ij}\theta_j)/w_{ii}$ and $\tilde{v}_i = \sigma_e^2/w_{ii}$.

The conditional posterior density of σ_e^2 corresponds to

$$\sigma_e^2 | y, \theta, \sigma_u^2, s, \nu \sim \tilde{\nu_e} \tilde{s_e}^2 \chi_{\tilde{\nu_e}}^{-2}$$

which corresponds to a scaled inverted chi-square distribution. The parameters of the above distribution are defined as

$$\tilde{\nu_e} = n + \nu_e$$

and

$$\tilde{s_e}^2 = \left[(y - Xb - Zu)^T (y - Xb - Zu) + \nu_e s_e^2 \right] / \tilde{\nu_e}$$

Analogously, the conditional posterior density of σ_u^2 can be derived as

$$\sigma_u^2 | y, \theta, \sigma_e^2, s, \nu \sim \tilde{\nu_u} \tilde{s_u}^2 \chi_{\tilde{\nu_u}}^{-2}$$

with

$$\tilde{\nu_u} = q + \nu_u$$

and

$$\tilde{s_u}^2 = \left[u^T G^{-1} u + \nu_u s_u^2 \right] / \tilde{\nu_u}$$

5.5.1.5 Implementation of the Gibbs Sampler

The above specified fully conditional posterior densities are used to draw in turn for every unknown component a random number from the respective specified conditional posterior distribution. The sampled random numbers are stored for making inferences about the unknown components in the statistical model. One example inference consists of the Bayesian estimate of a given unknown. The

Bayesian estimate is computed from the means of the random samples that were drawn using the Gibbs sampler.

5.5.2 Practical Consideration

Because the availability of widely used and tested software implementing Bayesian procedures is limited, these procedures are not used in practical livestock breeding.