1 Draft for Master Thesis Project: "Generalized Linear Mixed Models In Genetic Evaluations"

Selection of livestock animals based on predicted breeding values has led to tremendous improvements of product quality and of production efficiency in livestock production. The breeding value of an animal is an unobservable quantity which assesses the value of the genetic potential of a given animal. Because, breeding values cannot be observed, they must be predicted based on phenotypic observations of animals. The prediction of breeding values is based on Best Linear Unbiased Predictions (BLUP) of random effects in a linear mixed effects model. The linear mixed effects model explains observations as a linear function of fixed effects and random breeding values. Estimates of fixed effects and predictions of random breeding values are solutions of the so-called mixed model equations which are proposed by (Henderson 1982). The quality of the ranking of the predicted breeding values is assessed by comparing average phenotypic performances of offspring from top-ranking and bottom-ranking parents. Average phenotypic performances of offspring from top-ranking parents must be significantly higher than average phenotypic performances of offspring from bottom-ranking parents.

Strictly speaking, the solutions of the mixed model equations are only valid, if the observations and the breeding values both follow multivariate normal distributions. There are references such as (Negussie, Strandén, and Mäntysaari 2008) which show that the predicted breeding values are also valid as ranking criterion for parents when phenotypes are not normally distributed. Such references are used as justification to use the linear mixed effects model framework together with the solutions from the mixed model equations for traits that do clearly not follow a normal distribution. Extreme cases of such data are observations that show a binary distribution. From a livestock breeding point of view, it is not clear what the distribution of the predicted breeding values should look like and how such an extreme distribution of observations affect the ranking of breeding animals according to the predicted breeding values.

Recently, breeding organisations are more interested in the potential of improving their populations with respect to health traits or with respect to animal behavior. Examples of such traits for which Qualitas recently developed a routine evaluations process are

- Twin and multiple birth in cattle
- Early-life calf survival in dairy cattle
- Carcass conformation in beef cattle

The results of the above three evaluations showed some problems that were previously not found when using the same procedure for other traits.

1. Goodness of fit of the mixed linear effect model as assessed by criteria such as AIC and BIC was very low.
2. The standard deviation of the predicted breeding values was very low
3. As a consequence of 2, the standard errors of prediction for the predicted breeding values were very high.
4. The ranking of the animals according to the predicted breeding value has a low quality as measured by the top-bottom comparisons.

The aim of this project is to assess the benefit of using different types of models such as the generalized linear mixed model for genetic evaluations of trait response variables showing binary or categorical distributions. There are a few studies such as (Hoeschele et al. 1986), (Tempelman 1998), (Koenig et al. 2005) or (Vazquez et al. 2009) that have already tried such models. To the best of our knowledge no routine evaluations have been implemented using the class of generalized linear mixed models. Important pre-requisites for the implementation of a routine evaluation is to be able to completely characterize the properties of a genetic evaluation using generalized linear mixed model. Furthermore, it is also important to evaluate different possible software solutions for a routine evaluation pipeline.

The benefits of using generalized linear mixed models is evaluated using a simulated dataset which has approximately the same structure as the datasets that are analysed during the routine evaluations. In a second phase of the project, datasets that come from real-world genetic evaluations should als be evaluated.

2 Abstract

Predicted breeding values are the most important selection criterion in modern animal breeding programs. The strict implementation of breeding programs in which parents are selected based on their predicted breeding values has lead to the tremendous improvement of product quality and of production efficiency in livestock production. This system of using predicted breeding values to change the animals in a certain livestock population with respect to certain traits works best when the trait values follow a normal distribution. It is not clear what happens when this assumption about the trait distribution is not met. There are a few references that the selection system is invariant to the trait distribution. But recent practical results still point to potential problems which might be related to the distribution of the trait values. The aim of this project is to investigate the influence of the trait distribution, to propose new models and to evaluate the benefit of this new models.

3 Terminology

The **phenotype** of an animal or a **phenotypic observation** of an animal is an observation or a measurement of a given trait of interest.

A **gene** is a location somewhere in the genome which occurs in different variants and is responsible for the expression of some characteristic or trait. The different genetic variants at a given location across the whole population are called **alleles**. The different variants that occur at a single location are called a **genotype**. Genotypes can only be observed or measured partially, but the complete genotypic information of an animal is generally considered to be unknown or unobservable.

4 Background

4.1 Livestock Breeding

In livestock breeding parents are selected from a population such that their offspring has an expected phenotypic performance that is closer to a pre-defined breeding goal compared to the parent generation. The selection of parent animals cannot be done based on phenotypic observations because parents do not pass phenotypes to their offspring, but a random sample of their alleles. Hence it is reasonable to select parents based on the value of the alleles that are passed to the offspring. This value is called the **breeding value**. Because the complete genotypic information of an animal cannot be observed, the breeding value can also not be observed or not measured. But breeding values can be estimated or predicted based on observed phenotypic data.

4.2 Prediction[1] of Breeding Values

The prediction of breeding values happens at the interface of the two scientific disciplines

- quantitative genetics and
- statistics.

[1] In livestock breeding literature the term `prediction` is used for random effects and `estimation` is used for fixed effects.

4.2.1 Quantitative Genetics

The genetic model which is a result from quantitative genetics establishes the connection between observed phenotypic value (p) and unobservable genotypic value (g). In a very simple form the genetic model can be stated as

$$p = g + e \tag{1}$$

where e denotes the value of the non-genetic components which are sometimes also referred to as environmental component values. The environmental component values can be divided into systematic factors (β) that can be observed or measured and a non-observable random error term (ϵ). The genotypic value can be split into an additive part g_A, a dominance part g_D and an epistatic part g_I, depending on the mode of inheritance of a given trait. More formally, we can write

$$g = g_A + g_D + g_I \tag{2}$$

For our purposes, only the additive component (g_A) is important and we set that to be the breeding value which is called b. The two other genetic parts will be grouped together with the random error term into a new residual term called ϵ^*. After this re-arrangement, the genetic model becomes

$$p = \beta + b + \epsilon^* \tag{3}$$

4.2.2 Statistics

From a statistics point of view, the genetic model given in (3) looks a lot like a linear mixed effects model. Usually the symbol for the phenotypic observation p in (3) is replaced by the variable y. Combining all observations from a given population into a single dataset and writing the model in matrix-vector notation results in the following statistical model.

$$y = X\beta + Zb + \epsilon \tag{4}$$

where

- y is a vector of length n containing random phenotypic observations,
- β is a vector of length p containing fixed effects
- b is a vector of length q containing random breeding value and
- ϵ is a vector of length n containing random error terms.

The matrices X $(n \times p)$ and Z $(n \times q)$ are design matrices linking fixed effects and random breeding values to observations, respectively.

Unfortunately, the notation used in the livestock breeding literature very often does not distinguish between random variables and observed values. Making that distinction according to the notation used in (Bates et al. 2015), we get the vector-valued random response variable represented by the symbol $\mathcal{Y}$ with observed values y. The conditional distribution of $\mathcal{Y}$ given $\mathcal{B} = b$ has the form

$$(\mathcal{Y}|\mathcal{B} = b) \sim \mathcal{N}(X\beta + Zb, V) \tag{5}$$

The matrices X and Z are as defined above. The unconditional distribution of $\mathcal{B}$ is a multivariate normal distribution with variance-covariance matrix Σ.

$$\mathcal{B} \sim \mathcal{N}(0, \Sigma) \tag{6}$$

In the case where $\mathcal{B}$ represents breeding values, the variance-covariance matrix Σ has the form

$$\Sigma = A * \sigma_a^2 \tag{7}$$

where A is the known numerator-relationship matrix which contains the ancestral relationships between the animals in the dataset. The variance component σ_a^2 stands for the genetic additive variance.

As a consequence of the above assumptions, the random errors $\mathcal{E}$ follow a multivariate normal distribution with

$$\mathcal{E} \sim \mathcal{N}(0, R) \tag{8}$$

The variance-covariance matrix R is usually assumed to have a diagonal structure given by $R = I_n \sigma_\epsilon^2$ where I_n is the $n \times n$ Identity matrix.

4.2.3 Solutions

The fixed effects are estimated using the Best Linear Unbiased Estimate (BLUE). For the random breeding values, the so-called Best Linear Unbiased Predictor (BLUP) which is also known under the name `conditional mode` is used. The expression for both terms contain the inverse V^{-1} of the variance-covariance matrix V. In practical livestock breeding scenarios this matrix is very big and its inverse cannot be computed.

Henderson (Henderson 1982) found that the solution to the mixed model equations have the same properties. These equations have the following structure

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{9}$$

Resulting from the solution of these equations, $\hat{u}$ is the BLUP of u and $\hat{\beta}$ is the BLUE of β. But it is important to notice that these equivalence is only valid under the distributional assumptions described in the previous subsection.

References

Bates, Douglas, Martin Mächler, Bolker Benjamin M., and Steven C. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1). https://doi.org/10.18637/jss.v067.i01.

Henderson, Charles R. 1982. "Analysis of Covariance in the Mixed Model: Higher-Level, Nonhomogeneous, and Random Regressions." *Biometrics* 38 (3): 623–40. https://doi.org/doi:10.2307/2530044.

Hoeschele, Ina, J. L. Foulley, J. J. Colleau, and D. Gianola. 1986. "Genetic evaluation for multiple binary responses." *Genetics Selection Evolution1* 18 (3): 299–320. https://doi.org/DOI:10.1186/1297-9686-18-3-299.

Koenig, S., A. R. Sharifi, H. Wentrot, D. Landmann, M. Eise, and H. Simianer. 2005. "Genetic parameters of claw and foot disorders estimated with logistic models." *Journal of Dairy Science* 88 (9). Elsevier: 3316–25. https://doi.org/10.3168/jds.S0022-0302(05)73015-0.

Negussie, Enyew, Ismo Strandén, and Esa A. Mäntysaari. 2008. "Genetic analysis of liability to clinical mastitis, with somatic cell score and production traits using bivariate threshold-linear and linear-linear models." *Livestock Science* 117 (1): 52–59. https://doi.org/10.1016/j.livsci.2007.11.009.

Tempelman, Robert J. 1998. "Generalized Linear Mixed Models in Dairy Cattle Breeding." *Journal of Dairy Science* 81 (5): 1428–44. https://doi.org/10.3168/jds.S0022-0302(98)75707-8.

Vazquez, A I, D Gianola, D Bates, K A Weigel, and B Heringstad. 2009. "Assessment of Poisson, logit, and linear models for genetic analysis of clinical mastitis in Norwegian Red cows." *Journal of Dairy Science* 92 (2). Elsevier: 739–48. https://doi.org/10.3168/jds.2008-1325.