A General Method to Audit BLUP Programs

***M. Wensch-Dorendorf*[1]*, J. Wensch*[2]*, H.H. Swalve*[1]**

[1]*Institute of Animal Breeding and Husbandry, Martin-Luther-University, Halle-Wittenberg*

[2]*Institute of Mathematics, University Potsdam*

Introduction

International genetic evaluations as carried out by Interbull rely on the results of national evaluations undertaken by numerous computing centres throughout the world. The BLUP method has become the standard methodology for the estimation of breeding values. Worldwide, a large number of BLUP computer programs, each one usually tailored towards the needs of an individual country, are in use. It is of vital interest for Interbull as well as all countries participating in MACE evaluations that national evaluations reach a high quality level. This quality is a demand at all tiers of the evaluation procedure, starting at the quality of data entering from recording systems and stretching over data editing and preparation, the BLUP run itself to post-processing of the results.

Aim of the present project is the development of a method to generate data, i.e. the simulation of breeding values and phenotypes such that the BLUP procedure used is able to estimate the simulated breeding values numerically exact and not only asymptotically. The method should be viewed as an aid for the development and further refinement of BLUP programs

Material and Methods

Multiple trait model

As usual, a general model for phenotypes listed in vector y is

$$y = X\beta + Zu + e \qquad (1)$$

where $u \sim N(0, G)$, $e \sim N(0, R)$, $G = G_0 \otimes A$, $R = R_0 \otimes I$. A is the numerator relationship matrix, G_0 and R_0 are the variance/covariance matrixes of genetic and residual effects. X and Z are incidence matrices pertaining to fixed and random effects, respectively, and u denotes a vector of genetic effects (breeding values) which are unknown and are to be predicted. The vector of unknown fixed effects is given as ß, and e is the vector of residual effects.

Following Henderson (1973), under the assumption that residual effects among observations (animals) are uncorrelated, ß and u can be estimated using the Mixed Model Equation (MME) given as

$$\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{pmatrix}$$

Estimation method

From theory, it is well known that a solution to the MME maximises the Likelihood-function for model (1)

$$L(u,e) = c \exp(-u^TG^{-1}u/2) \exp(-e^TR^{-1}e/2)$$

c is a normalization factor depending on G and R. The Maximisation of L(u,e) is equivalent to the minimisation problem

$$e^T R^{-1} e + u^T G^{-1} u \to \text{Min} .$$

Setting

$$r = \begin{pmatrix} R^{-1/2}e \\ G^{-1/2}u \end{pmatrix}$$

we end up with the Least Squares Problem minimise $\|r\|^2_2 = e^T R^{-1} e + u^T G^{-1} u$ where

$$r = \underbrace{\begin{pmatrix} R^{-1/2}y \\ 0 \end{pmatrix}}_{b} - \underbrace{\begin{pmatrix} R^{-1/2}X & R^{-1/2}Z \\ 0 & -G^{-1/2} \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} \beta \\ u \end{pmatrix}}_{x}$$

is obtained by substituting $e = y - X\beta - Zu$ from model (1).

$\|.\|_2$ is the euklidian norm. $R^{-1/2}$ is the unique symmetric positive definite matrix B with $BB = R^{-1}$.

Using the definitions of A, b and x as given above, the analogy of MME and the normal equation can be shown. A general Least Squares Problem has the formulation: find the vector x with

$$||Ax - b||_2 \rightarrow \mathrm{Min}_x \quad (2)$$

It is well known that x is a solution of (2) if and only if the normal equation

$A^TAx = A^Tb$ is fulfilled which is equivalent to $A^Tr = 0$ for the residual
$r = Ax-b$. By substituting the definitions of A, b and x in the normal equation, the MME can be represented by:

$$\begin{aligned} A^TAx &= \begin{pmatrix} X^TR^{-1/2} & 0 \\ Z^TR^{-1/2} & -G^{-1/2} \end{pmatrix} \begin{pmatrix} R^{-1/2}X & R^{-1/2}Z \\ 0 & -G^{-1/2} \end{pmatrix} x \\ &= \begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + G^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ u \end{pmatrix} \\ A^Tb &= \begin{pmatrix} X^TR^{-1/2} & 0 \\ Z^TR^{-1/2} & -G^{-1/2} \end{pmatrix} \begin{pmatrix} R^{-1/2}y \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} X^TR^{-1}y \\ Z^TR^{-1}y \end{pmatrix} \end{aligned}$$

Residual condition

Now a constraint can be derived such that the MME are exactly fulfilled. If vectors ß, u and e are given such that $A^Tr = 0$ and the y-values are derived using model (1), the MME are fulfilled exactly. The term $A^Tr = 0$ gives:

$$0 = A^Tr = \begin{pmatrix} X^TR^{-1/2} & 0 \\ Z^TR^{-1/2} & -G^{-1/2} \end{pmatrix} \begin{pmatrix} R^{-1/2}e \\ G^{-1/2}u \end{pmatrix}$$

This leads to the following residual conditions:

$$\begin{aligned} 0 &= X^T R^{-1} e && (3) \\ 0 &= Z^T R^{-1} e - G^{-1}u && (4) \\ u &= GZ^T R^{-1} e && (4') \end{aligned}$$

With model equation (1) and the residual condition (3), and (4), Henderson's MME are fulfilled exact. The vector of fixed effects ß can be chosen arbitrarily since the residual conditions are independent of this vector. However, the residual condition (4) depends on A^{-1} since $G^{-1} = G^{-1}_0 \otimes A^{-1}$. A linear system of equations has to be solved to fulfil residual condition (3) and (4).

Model (1) is very simple. In addition to the residual effect e, only one random effect is in the model. Following, an extension of the model equation for further random effects is derived. Let's assume a further random effect is added to model (1), denoted by w with associated incidence matrix Z_w. Model (1) becomes:

$$y = X\beta + Zu + Z_w w + e \qquad (1')$$

with $w \sim N(0, W)$, $W = W_0 \otimes M$. The residual vector r to be minimised becomes:

$$r = \begin{pmatrix} R^{-1/2}e \\ G^{-1/2}u \\ W^{-1/2}w \end{pmatrix}$$

and by using $e = y - X\beta - Zu - Z_w w$ the resulting matrices A, x and b become:

$$A = \begin{pmatrix} R^{-1/2}X & R^{-1/2}Z & R^{-1/2}Z_w \\ 0 & -G^{-1/2} & 0 \\ 0 & 0 & -W^{-1/2} \end{pmatrix}$$

$$x = \begin{pmatrix} \beta \\ u \\ w \end{pmatrix}, \ b = \begin{pmatrix} R^{-1/2}y \\ 0 \\ 0 \end{pmatrix}$$

Again using $A^Tr = 0$ as an additional residual condition leads to:

$$0 = Z^T_w R^{-1} e - W^{-1} w \qquad (5)$$

or

$$w = WZ^T_w R^{-1} e \quad (5')$$

Analogously to condition (4') which represented the relation between residual effect e and random effect u, condition (5') gives the relation between residual effect e and the new random effect w. It is trivial to see that the extension made here is not limited to the inclusion of one additional random effect. Rather, further random effects may be included and by this most scenarios in animal breeding can be represented (random regression, maternal effects, dominance, gametic effects, …) .

The residual condition (3) as a stand-alone condition is equivalent to (1) without any random effect except e.

If condition (3) and analogously conditions (4), and (5) are fulfilled (and possibly even more conditions as necessary), the MME are fulfilled exactly if we use model (1) or extensions like in (1') to generate data sets y.

Implementation

To fulfil conditions (3) and (4), two approaches have been proposed up to now. The first approach is based on an idea by Thompson (1997). Thompson's idea for a single trait scenario can be characterized as follows:

$$\begin{aligned} 0 &= \sum_{i=1}^{B} u_i \\ u_o &= \frac{1}{2}(u_s + u_d) \\ e &= RZ(Z^T Z)^{-1} G^{-1} u \end{aligned}$$

with B = number of base animals (animal 1,…, B are base animals), s = sire, d = dam and o = offspring. Furthermore, Thompson assumed one fixed effect class for base animals. With u and e as given above the residual conditions (3) and (4) are fulfilled exactly.

The single trait case was implemented by Täubert *et al.* (2002). The resulting data equation for y values for the multiple trait case has been shown in Wensch-Dorendorf *et al.* (2005). Solving for u, e and y is very cheap and hence an implementation is simple. A reason for the simplicity is that for any trait in the model

$$A^{-1}u = (u_b^T, 0^T)^T$$

is valid which means that it is unnecessary to set up A^{-1} . Vector u contains all genetic effects for the respective trait and u_b the genetic effects for the base animals ($u_b^T = (u_1,\ldots,u_B)$). The vector 0^T is a nullvector with dimension equal to the number of none base animals.

A proposal for a second approach was given by Leclerc and Ducrocq (2006). The idea of this approach is to fulfil condition (3) first and than solve for other random effects by using (4'), (5') etc. . Leclerc and Ducrocq (2006) give a proposal for e to fulfil condition (3) by using the normal equation:

$$\begin{aligned} &\text{set :} \quad e = X\hat{\beta} - e^* \\ &\text{solve :} \, X^T R^{-1}\hat{\beta} = X^T R^{-1} e^* \end{aligned}$$

where e* is arbitrary. To fulfil condition (3), other proposals for e are possible, for instance:

$$\begin{aligned} &e = R(X(X^T X)^{-1} X^T - I)e^*, \\ &(X^T X \text{ regular}) \end{aligned}$$

or to express e by using LQ decomposition of X^T (or QR)

$$\begin{aligned} &\text{solve } X^T = LQ = \begin{pmatrix} L_k & 0 \end{pmatrix} Q, \\ &\text{choose: } e = RQ^T \begin{pmatrix} 0 \\ e^* \end{pmatrix}, \text{ since} \\ &\underbrace{\begin{pmatrix} L_k & 0 \end{pmatrix} Q}_{X^T} R^{-1} \underbrace{RQ^T \begin{pmatrix} 0 \\ e^* \end{pmatrix}}_{e} = \begin{pmatrix} L_k & 0 \end{pmatrix} \begin{pmatrix} 0 \\ e^* \end{pmatrix} = 0 \end{aligned}$$

where e* is also arbitrary, k=Rank(X) and L_k is regular.

All these approaches and special proposals fulfil condition (3) and (4) in such a way that an audit of BLUP-Programs is possible.

Results

In our project as a first step the programs PEST (Groeneveld, 1990) and three BLUPF90 variants (Misztal, 1999, Ducrocq *et al.,* 2003) have been tested. Work on a case study (up to 10 Mio animals, 3 traits with 2 random effects + fixed regression + rest) is in progress.

The yet preliminary results from the comparison of true (simulated) and estimated breeding values point out the following:

- The correlation between simulated BV and EBV is very high (0.999 to 1.0) for all programs
- However, minor changes in rank when inspecting top 100 lists can be observed.

Summary

With the procedure presented here, an effective and general method to audit BLUP computer programs is introduced. This method can be of help when developing and validating a BLUP program. Work on the comparison of several well known programs is under way.

Preliminary results indicate that minor rank changes may occur although the correlation between true (simulated) and estimated breeding values may be close to unity.

Literature

Ducrocq, V. & Druet, T. 2003. Advances in computing strategies for the solution of huge mixed model equations. *54th Annual Meeting of the EAAP,* Rome, Italy.

Leclerc, H. & Ducrocq, V. 2006. Exact validation of genetic evaluation software. Interbull meeting 2006, Kuopio. *Interbull Bulletin 35,* 168-171.

Groeneveld, E. 1990. *PEST User's Manual.*

Henderson, C.R. 1973. Sire evaluation and genetic trends. *Proc. Anim. Breed. and Genet. Symp. in Honour of Dr. J.L. Lush*, ASA und ADSA, Champaign, Illinois.

Misztal, I. 1999. Complex Models, More Data: Simpler Programming? *Interbull Bulletin 20,* 32-41.

Täubert, H., Swalve, H.H. & Simianer, H. 2002. The Interbull Audit Project Part II: Development of a Program for Auditing Breeding Value Estimation Programs. *Interbull Bulletin 29,* 165-167.

Thompson, R. 1997. Generating data to check mixed model equations. *Personal communication.*

Wensch-Dorendorf, M., Swalve, H.H. & Wensch, J. 2005. Simulation of multiple trait data for testing breeding value estimation programs. *56th Annual Meeting of the EAAP,* Book of Abstracts No. 11, Session 17, Poster 36, Uppsala, Sweden, p. 207.