SIMULATION STUDY ON LINEAR MIXED MODELS WITH CONTAMINATED NORMAL DISTRIBUTION IN ANIMAL BREEDING

I.G. Pereira¹, H.N. Oliveira² and G.J.M. Rosa³

¹Animal Science Department, University of Lavras - MG, Brazil
²Department of Animal Breeding and Nutrition, São Paulo State University - SP, Brazil
³Department of Biostatistics, São Paulo State University - SP, Brazil

INTRODUCTION
The methodology of linear mixed models is widely used in genetics and animal breeding. In general, the normal distribution is assumed both for residuals and random effects in these models, which makes inferences very sensitive to the presence of discrepant values in observations (Rogers and Tukey, 1972). An alternative in this sense refers to models with thick-tailed distributions, which has proved quite effective in robust estimation and of easy implementation in a Bayesian context (Rosa, 1999). However, the studies about the use of these distributions for estimation of (co)variance components and genetic parameters and prediction of genetic values of animals are few (Strandén and Gianola, 1999). This work compares estimates of variance components and predicted genetic values obtained by Gaussian and robust linear mixed models, in data simulated with different levels of discrepant values in observations.

MATERIAL AND METHODS
Fifty data files with 1,000 animals each, distributed in five generations of 200 animals without selection, were simulated. All the animals presented records for three hypothetical productive traits (FV, FV1 and FV2). For the 200 animals of the first generation, genetic and residual values were generated, from normal distributions with mean zero and variances equal to 36.0 and 64.0, respectively. The animals were randomly distributed into two levels of fixed effects and with probability determined by $\phi$ (0.20 or 0.40), they could be allocated to a contaminant or non-contaminant population. The values of the characteristics were obtained from the summation of the genetic value, the systematic (fixed) effect and the residual of each animal. The phenotypic value FV stood for the population with normal distribution with no contaminants and the other two, identified as FV1 and FV2 were generated in a similar manner to the former except that the residual value of each animal allocated to the group of contaminants was divided by the square root of a certain value $\tau$ (between 0.0625 and 0.25) before being summed to the others. For FV1, the value of $\tau$ was always greater than for FV2. For the animals of the other generations, the procedure used was similar, except that the genetic value of the animals was generated as the summation of the means of the parents’ genetic values plus the Mendelian sample variation.

Thus, for each population three data sets were simulated. The first one had normal distribution without any contaminants, and the other two presented a proportion of contaminants given by $\phi$, with residual variance equal to that of the original population divided by $\tau$. The presence of
those individuals with discrepant observations characterizes a new distribution with leptokurtic shape, relative to the contaminated normal distribution. Five replications for each of the ten different sets of parameters were generated totaling 50 populations.

The fixed effects and the direct and residual (random) genetic effects were taken into account in the models. For Gaussian models, analyses both in the context of maximum likelihood and of Bayesian Inference (BI) were performed. For these analyses, the softwares MTDFREML (Boldman et al., 1995) and MTGSAM (Van Tassel and Van Vleck, 1995) were used, respectively. The total of samples generated in the analysis by MTGSAM was 250,000. As a robust alternative to the Gaussian model, the contaminated normal distribution was used for the residuals. A computer program suited to those analyses was adapted from a Fortran 77 program developed by Dr. Daniel Sorensen for Gaussian analysis.

RESULTS AND DISCUSSION
Table 1 presents the average of the estimates of the parameters of contaminated normal distribution ($\tau$ and $\phi$) over the five replicates for a particular set of parameters. The estimates of $\tau$ for FV1 and FV2 were quite close to the real ones used in the simulation. In some of the 10 sets for FV where no contamination takes place and therefore, one should estimate $\tau=1$ and/or $\phi=0$, some intermediate means appear. The explanation for that variation is that when the estimates of $\tau$ were close to one, the value of $\phi$ is unimportant, as the residual values were the same into the two populations. On the other hand, when the estimates of $\phi$ were close to zero, the value of $\tau$ does not matter, as there were no contaminants. However, as these two scenarios could alternate for each of the five replicates, the average estimates of $\tau$ and $\phi$ should be different from the expected ($\tau=1$ and $\phi=0$). So, although the mean of the two parameters in all the replicates assumed intermediate values, there were no adjustment problems.

It is seem that the residual variance estimate obtained by REML presents a higher value than that used in the simulation, in a growing manner in terms of the contamination degree, increasing from FV to FV2 about 28% (Table 1). Similar trend was observed in the posterior mean by IB in Gaussian model. Those estimates are, in fact, estimates of average values of the residual variance of the two subpopulations, which made up the sample. Therefore, these models do not distinguish the two sub-populations on the data sets. On the other hand, the estimates of genetic variance (GV) obtained by REML and by IB in Gaussian model were less influenced, which might be expected since the heterogeneity occurred only on the residual variance. By comparing estimates of variance components obtained by REML and IB in data simulated in the absence of contamination, Van Tassel et al. (1995) and Van Tassel and Van Vleck (1996) found similar estimates. Nevertheless, Wright et al. (2000) verified that the estimative of GV obtained by REML differed greatly from the posterior mean by IB.

Average estimates of heritability ($h^2$) obtained by REML and by IB in Gaussian model were much lower than those obtained by the robust model, mainly for FV2 (Table 1). This was expected since $h^2$ presented by the robust model is concerned only with the non-contaminant population. The $R^2$ of the regressions of the predicted genetic values on the actual ones were very similar for estimations either by REML or by IB, under the Gaussian model (Table 1).
Table 1. Comparison of pure Gaussian and Robust approaches to linear mixed models: for 10 groups of parameters ($\phi = 0.2$; $\tau_1 = 0.25$ and $\tau_2 = 0.0625$, respectively for FV1 and FV2), with genetic additive variance and residual variance means equal to 36.0 and 64.0, respectively

<table>
<thead>
<tr>
<th>Model/Method</th>
<th>R²</th>
<th>Trait</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>MSE$^c$</th>
<th>$\sigma_a^2$</th>
<th>$\sigma_e^2$</th>
<th>$\sigma_a^2/\tau$</th>
<th>$h^2$</th>
<th>$h^2^*$</th>
<th>$\tau$</th>
<th>$\phi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>G/REML</td>
<td></td>
<td>FV</td>
<td>0.39</td>
<td>0.58</td>
<td>0.49</td>
<td>17.0029</td>
<td>31.67</td>
<td>66.35</td>
<td>—</td>
<td>0.32</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G/REML</td>
<td></td>
<td>FV1</td>
<td>0.28</td>
<td>0.51</td>
<td>0.40</td>
<td>19.7917</td>
<td>30.95</td>
<td>105.39</td>
<td>—</td>
<td>0.23</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G/REML</td>
<td></td>
<td>FV2</td>
<td>0.12</td>
<td>0.36</td>
<td>0.26</td>
<td>24.6615</td>
<td>33.38</td>
<td>258.62</td>
<td>—</td>
<td>0.12</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G/Bayes</td>
<td></td>
<td>FV</td>
<td>0.39</td>
<td>0.58</td>
<td>0.48</td>
<td>17.0354</td>
<td>32.87</td>
<td>65.82</td>
<td>—</td>
<td>0.33</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G/Bayes</td>
<td></td>
<td>FV1</td>
<td>0.28</td>
<td>0.51</td>
<td>0.40</td>
<td>19.8042</td>
<td>33.36</td>
<td>104.61</td>
<td>—</td>
<td>0.24</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>G/Bayes</td>
<td></td>
<td>FV2</td>
<td>0.13</td>
<td>0.36</td>
<td>0.26</td>
<td>24.6552</td>
<td>36.23</td>
<td>257.24</td>
<td>—</td>
<td>0.12</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>R/Bayes</td>
<td></td>
<td>FV</td>
<td>0.39</td>
<td>0.58</td>
<td>0.49</td>
<td>16.9714</td>
<td>33.34</td>
<td>63.31</td>
<td>401.71</td>
<td>0.35</td>
<td>0.08</td>
<td>0.1576</td>
<td>0.0494</td>
</tr>
<tr>
<td>R/Bayes</td>
<td></td>
<td>FV1</td>
<td>0.29</td>
<td>0.53</td>
<td>0.42</td>
<td>19.2320</td>
<td>33.75</td>
<td>79.94</td>
<td>348.02</td>
<td>0.30</td>
<td>0.09</td>
<td>0.2297</td>
<td>0.1352</td>
</tr>
<tr>
<td>R/Bayes</td>
<td></td>
<td>FV2</td>
<td>0.27</td>
<td>0.50</td>
<td>0.38</td>
<td>20.5507</td>
<td>28.88</td>
<td>65.01</td>
<td>910.50</td>
<td>0.30</td>
<td>0.03</td>
<td>0.0714</td>
<td>0.2311</td>
</tr>
</tbody>
</table>

$^a$ - G/REML = Gaussian/Restricted Maximum Likelihood; G/Bayes = Gaussian/Bayesian Inference; R/Bayes = Robust (Contaminated Normal) / Bayesian Inference.

$^b$ - VF = trait without contamination; VF1 = trait with contamination given by $\tau_1$; VF2 = trait with contamination given by $\tau_2$.

$^c$ - MSE = Mean square error of the regression of the true breeding values genetic on predicted genetic values.

$^d$ - $\sigma_a^2$ = additive genetic variance; $\sigma_e^2$ = Residual variance of the without contamination population; $\sigma_a^2/\tau$ = Residual variance of the contamination population $h^2$ = Heritability of the without contamination population; $h^2^*$ = Heritability of the contamination population; $\tau$ e $\phi$ = Parameters of the Contaminated Normal.

Compared to the robust model, the results are similar for FV but for both FV1 and FV2, the robust one presented higher $R^2$. These results stress the superiority of the robust model in predicting the genetic value for conditions in which there are contaminant individuals, with no important differences if under normality. The $R^2$ decreases as the ratio of contaminants increase and as well in terms of the value of $\tau$ and $\phi$. Therefore, the greater the percentage of contaminants and the more different the residual variances of the subpopulations, the worse will be the prediction of the genetic value. Greater mean squares of the residuals for FV1 and FV2 under the Gaussian model compared with those of the robust are observed, which denotes poorer adjustment quality of the Gaussian relative to the robust models. Strandén (1996) observed similar results.
CONCLUSION
Under the conditions of the present work, Gaussian and robust models present similar results for data with normal distribution. However, in the situations with heterogeneity of the residual variance, the robust model employing the contaminated normal distribution yields more reliable inferences, both for variance component estimation and for prediction of genetic merits.

REFERENCES