Comparison of two models for estimation of variance components in a sample of Spanish Holstein Friesians

By J. P. Gutiérrez, J. Cañon, and M. Rico

Introduction

Most research in animal breeding methodology has been undertaken in developed countries. In these, methodology has been applied to large data sets, more or less balanced and connected, and containing full genealogy. However, the situation in many other countries is really far from such type of data sets. Thus, the efficiency of some methods should be tested, before introducing more expensive techniques which could be unnecessary.

Henderson (1975) suggested three possible ways for comparing alternative methods. The first one is based on analytic methods, which require complicated mathematical manipulations. The second possibility is to check methods on simulated data, where true values are known (Liu and McAllister 1984; van der Werf and de Boer 1989). In this case, conclusions are dependent on the simulated population. Finally, the most extended way, is to apply them to the same real data set (Dempster et al. 1983; van Raden and Jung 1988). If results are different it is not possible to say which method is better.

The objective of this research was to develop a simulation method which allows comparison of variance component estimation methods using different models in a particular population.

Materials and methods

Data

A sample of 5,496 Spanish Holstein Friesian lactations, covering the period 1980 to 1988 were used. Records belonged to 3,725 cows, daughters of 324 sires and distributed in 354 herds. The pedigree file contained 4,946 animals.

Unbalancedness, shown in Figures 1, 2 and 3, and Table 1, can be summarized as follows:

- More than one half of the herds were represented with 5 or less cows. Less than 5% had more than 35 (Figure 1).
- 66.5% of sires had between 5 and 10 daughters and only 1% of sires had more than 50 (Figure 2).

Fig. 1. Distribution of herds according to the number of cows

U.S. Copyright Clearance Center Code Statement: 0931-2668/94/1102-0169 $2.50/0
**Fig. 2.** Distribution of sires according to the number of daughters

**Fig. 3.** Distribution of cows according to the number of records

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cows with only the sire known</td>
<td>2412</td>
<td></td>
</tr>
<tr>
<td>Cows with both parents known</td>
<td>936</td>
<td>85 %</td>
</tr>
<tr>
<td></td>
<td>310</td>
<td>15 %</td>
</tr>
</tbody>
</table>
- For 85% of the cows only the sire was known (Table 1).
- 65% of the cows had only one lactation (Figure 3).

Models

The models tested were:
- Animal model
- Sire and cow nested within sire model

The animal model was:

\[ Y_{ij} = \mu + RA_i + a_i + p_i + e_{ij} \]

with:

\[
\begin{align*}
E(Y_{ij}) &= \mu + RA_i \\
E(a_i) &= E(p_i) = E(e_{ij}) = 0 \\
\text{Var} \begin{bmatrix}
  a \\
  p \\
  e
\end{bmatrix} &= \begin{bmatrix}
  \sigma^2_a & 0 & 0 \\
  0 & \sigma^2_p & 0 \\
  0 & 0 & \sigma^2_e
\end{bmatrix}
\end{align*}
\]

where:
- \( Y_{ij} \) is the \( k \)th record of the \( j \)th cow in the \( i \)th herd-year subclass,
- \( \mu \) the overall mean,
- \( RA_i \) the \( i \)th fixed herd-year of calving subclass,
- \( a_i \) the \( j \)th level of the random additive genetic effect,
- \( p_i \) the \( j \)th level of the random permanent environmental effect of the \( j \)th cow
- \( e_{ij} \) the \( k \)th level of random residual effect,
- \( A \) the numerator relationship matrix,
- \( \sigma^2_a \), \( \sigma^2_p \), and \( \sigma^2_e \), the genetic, permanent environmental and residual random effects variances, respectively.

The sire model included: overall mean (\( \mu \)), fixed herd-year effect (\( RA_i \)), random sire effect (\( s \)), random cow effect (\( v \)) and random residual effect (\( c \)):

\[ y_{ijkl} = \mu + RA_i + s_i + v_k + e_{ijkl} \]

with expected values and variance-covariance matrix:

\[
\begin{align*}
E(Y_{ijkl}) &= \mu + RA_i \\
E(s_i) &= E(v_k) = E(e_{ijkl}) = 0 \\
\text{Var} \begin{bmatrix}
  s \\
  v \\
  e
\end{bmatrix} &= \begin{bmatrix}
  \sigma^2_s & 0 & 0 \\
  0 & \sigma^2_v & 0 \\
  0 & 0 & \sigma^2_e
\end{bmatrix}
\end{align*}
\]

where:
- \( Y_{ijkl} \) is the \( k \)th record of the \( k \)th cow, sired by the \( j \)th bull, performing in the \( i \)th herd-year
- \( \sigma^2_s \), \( \sigma^2_v \) and \( \sigma^2_e \) are the sire, cow and residual random effects variances.

REML procedures were used with both of the models (Meyer 1987a, 1987b, 1988).

The animal model used sparse matrix solvers (Bolma and Van Vleck 1991).

All the runs were carried out on the IBM mainframe model 4381.
Simulation method

Milk and fat yields were simulated three times, using prior estimates of genetic additive, permanent and residual variances, whilst keeping the actual mating scheme, so that the simulation mimicked the actual population.

The simulation method was a modification of Rönningen (1974), allowing fixed effects. Each record in the population was replaced by the addition of the effects in the following animal model:

\[ Y_{ik} = RA_i + x_i + p_j + e_{ik} \]  \hspace{1cm} [1]

where:

- \( RA_i \) is the \( i \)th herd-year fixed effect subclass,
- \( x_i \) the additive genetic value of the \( j \)th cow,
- \( p_j \) the \( j \)th level of the environmental permanent effect,
- \( e_{ik} \) the \( k \)th residual effect.

The fixed effect value included in the model was the least squared estimation of such effect in the actual sample. All the repetitions were also done without the fixed effect in order to know its influence in the estimation. This case, model used both for the simulation and for the analysis did not include the herd-year effect.

Variance components were estimated in the actual population using a sire and cow nested model which included the birthyear of the sire as a genetic group effect. Resultant heritabilities and repeatabilities were 0.21 and 0.40 respectively for milk yield and 0.19 and 0.36 for fat yield. These values were used for simulating the random part of the model.

Genetic values in the base population were obtained by the product of the genetic standard deviation \( \sigma_a \) and \( z_n \), a random standard normal value:

\[ a_i = z_n \sigma_a \]

Genetic values for animals with known parents were obtained by:

\[ a_i = \frac{1}{2}a_j + \frac{1}{2}a_k + a_i \]

where:

- \( j \) and \( k \) are the parents of \( i \),
- \( a_j, a_k \), and \( a_i \), the genetic values for \( i, j \) and \( k \) individuals,
- \( \Phi_i \) the deviation produced by the Mendelian effect in the \( i \)th individual.

The permanent environmental \( p_j \) and residual \( e_{ik} \) effects were obtained as:

\[ p_j = z_n \Phi_p \]
\[ e_{ik} = z_n \sigma_e \]

where \( z_n \) and \( z_k \) are random standard normal values.

Then we can keep the actual structure of the sample: the same number of daughters per sire, in the same herds, at the same years and with the same number of records per cow.

Results and discussion

Table 2 shows the differences between simulated and obtained values, expressed as a percentage of expected values. Results are displayed for both models, with and without the inclusion of the fixed effect, for the three repetitions, and only for the additive genetic and environmental permanent variances. Results were in general very far from expected values.

Laboratory experiments are designed looking for the optimal structure in order to achieve some purpose (Hill 1970). However, the structure of field data sets can greatly influence the quality of the estimates.
Table 2. Differences between expected and obtained values, expressed as a percentage of the expected values

<table>
<thead>
<tr>
<th></th>
<th>MILK</th>
<th></th>
<th></th>
<th>FAT</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sigma^2_{E}$ NFE*</td>
<td>FE NFE</td>
<td>$\sigma^2_{E}$ NFE</td>
<td>FE NFE</td>
<td>$\sigma^2_{E}$ NFE</td>
<td>FE NFE</td>
</tr>
<tr>
<td>Sire</td>
<td>26</td>
<td>13</td>
<td>18</td>
<td>30</td>
<td>31</td>
<td>84</td>
</tr>
<tr>
<td>Model</td>
<td>10</td>
<td>2</td>
<td>32</td>
<td>47</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>Average</td>
<td>-11</td>
<td>-2</td>
<td>59</td>
<td>-8</td>
<td>65</td>
<td>19</td>
</tr>
<tr>
<td>Animal</td>
<td>28</td>
<td>1</td>
<td>12</td>
<td>-22</td>
<td>-42</td>
<td>-30</td>
</tr>
<tr>
<td>Model</td>
<td>18</td>
<td>12</td>
<td>25</td>
<td>-27</td>
<td>19</td>
<td>-29</td>
</tr>
<tr>
<td>Average</td>
<td>-7</td>
<td>2</td>
<td>40</td>
<td>-12</td>
<td>33</td>
<td>17</td>
</tr>
</tbody>
</table>

FE = Fixed effects included; NFE = Fixed effects not included.

Phenotypic variance is directly calculated from the data. Residual variance is also obtained in a direct way (Harville 1977). The quality of the estimates for the other two variance components, genetic and permanent variances, depends on the data structure.

Genetic variance estimates could have been biased by the different number of animals per sire (Figure 2). There are a few sires with more than the others. However, this disequilibrium influence is corrected when all the animals are connected through the numerator relationship matrix. On the other hand, and looking at Table 1, 85% of the cows had one parent missing in their pedigrees. Jensen et al. (1990) showed that unbiased estimates were achieved only when a complete numerator relationship matrix was included.

Permanent environmental variance estimates were more seriously affected by the unbalanced scheme. 65% of the cows had only the first record (Figure 3). This estimation would be impossible to carry out if all the animals had produced only once.

Therefore, unbalancedness and missing information could have been the causes of the differences found between obtained and expected values.

The fixed effect, just as it was considered in the simulation, increased phenotypic variance. It will be noted that both genetic and permanent estimates were normally increased when the fixed effect was included (Table 2). Differences between obtained and expected values were also smaller when the fixed effect was ignored.

Although the low number of replications does not lead to a definitive conclusion, the results show no advantages of the animal model procedure over the cheaper (in terms of number of equations) sire model for these data, suggesting that the latter may be preferred when data are characterized by a small number of individuals, little pedigree information and highly unbalanced distribution of effects.

Acknowledgements

Authors are extremely grateful to Keith G. Boldman who modified Meyer program and kindly provided it for this research.

We also appreciate comments by anonymous referee.

Summary

Two models for estimating genetic parameters in animal populations were compared. The cheaper, in terms of computing costs, was based on a sire model and the most expensive on an animal model. In order to check the accuracy of both methods they were applied to the best available sample of Spanish Friesian cattle.

Milk production of each cow, as well as fat contents were simulated using prior estimates of genetic
additive, permanent environment and residual variances, keeping the actual mating scheme, so that the simulation mimicked the actual population structure.

Results, under the above premises, show no advantages of the more costly procedure over the cheaper one, suggesting that a sire model based estimating procedure for genetic parameters may be preferred when a small number of individuals, little pedigree information and highly disequilibrated distribution of effects characterize the data.

Zusammenfassung

Vergleich zweier Methoden der Varianzkomponentenschätzung beim Spanischen Holstein


Unter den obenannten Voraussetzungen zeigen die Ergebnisse keinen Vorteil der rechnerisch aufwendigeren Prozedur gegenüber der einfachen. Die Schlußfolgerung lautet, daß das Vatermodell unter den Bedingungen der kleinen Tierzahlen, wenig Pedigree-Informationen und einer stark unbalancierten Verteilung der Tiere auf die einzelnen Effektklassen zu bevorzugen ist.

References


MERTEN, K., 1988b: DFREML2 Programs to estimate variance components for individual animal models by restricted maximum likelihood. User notes. Univ. of Edinburgh.


Authors' address: Dr. J. P. Gutiérrez, Faculty of Veterinary Medicine, University Complutense, E-28040 Madrid, Spain.