MIXED LINEAR MODELS WITH AN AUTOREGRESSIVE ERROR STRUCTURE

K. M. Wade, R. L. Quaas and L. D. Van Vleck
Department of Animal Science, Cornell University, Ithaca, NY 14853, USA

SUMMARY

Models utilizing a first order autoregressive – AR(1) – error structure were examined with a view to implementation in sire evaluation. The model comprised a fixed effect, a random [AR(1) covariance structure] effect within the fixed effect, and a residual. There were, therefore, three parameters to be estimated; the variance components associated with the random and the residual effects, and the correlation coefficient (ρ). Restricted Maximum Likelihood was used as the method of estimation and achieved via an Expectation Maximization algorithm. In the case of the latter, updates for the two variance components were achieved by closed form estimation, based on the assumption that ρ was known. The update of ρ, however, could not be found in this manner and was achieved by Fisher scoring, using the updated variance component associated with the random effect. For the (co)variance matrix associated with the random effect, direct methods were derived for finding elements of its inverse, and of the derivative of that inverse with respect to ρ. Simulation was used to test those methods, and the procedure was found to be robust as long as ρ was not close to $\pm$ unity. These results, along with some preliminary investigations of field data, suggest that the procedures developed may be beneficial in the modelling of nonzero covariances among records of animals in the same contemporary group.

INTRODUCTION

The treatment of contemporary group (CG) as fixed has been the subject of much discussion in the past, but has traditionally been included as fixed in order to correct for the possible bias due to non-random use of sires across herds (see for example Chauhan, 1987; Henderson, 1975; Meyer, 1987; Preisinger *et al.*, 1986; Van Vleck, 1987). This becomes apparent if breeders fail to use a random sample of the merit in bulls available for breeding. There is also the problem of preferential treatment, as alluded to by Meyer (1987), which will cause an association between sire and herd. Thus, the fitting of CG as a fixed effect in the modelling of dairy/beef traits has been justified on the basis that this interaction does exist and needs to be accounted for in some way.

Henderson (1975), however, showed that in genetic evaluations where herds are treated as random rather than fixed, the prediction error variance of differences in genetic estimates is smaller. Schaeffer (personal communication) has also demonstrated that treatment of HYS as fixed rather than random may bias the estimates of the genetic variance upwards; this is so because the denominator of the estimator should not include the rank of the incidence matrix, associated with the HYS, along with the rank of that incidence matrix associated with the random effect whose variance is being estimated. A more pressing argument for the treatment of CG as random is advanced when discussing the concept of effective number. Except in the cases of large herds, quite a lot of information is lost when HYS is treated as fixed (Brotherstone, Hill and Thompson, 1989). This problem of effective number takes on a special significance when dealing with countries/evaluations with small herd sizes (Preisinger *et al.*, 1986; Chauhan, 1988). Three options are available at this point; one can retain the idea of treating CG as fixed and accept the fact that there will be a loss of records *or* one can relax the environmental group size to, for example, herd-years or herds. The third option involves considering CG as a random effect. Meyer (1987) pointed out that "ignoring the environmental covariance between cows in the same sub-class in evaluating sires will result in overestimates of the reliability of sire proofs, especially if the sire x environment sub-classes are large". This study used an AR(1) to model the (co)variance among records in the same environment. With the derivation of computationally feasible techniques, the problem of treating CG as random, without the assumption of zero covariance, might be tackled in a relatively simple manner.

MATERIALS AND METHODS

The model of interest in this study was: $y_{ijk} = \beta_i + t_{j(i)} + e_{ijk}$, where y is the ijkth observation, β is the ith fixed effect, t is the jth random effect within the ith fixed effect and e is ijkth random residual. In matrix notation the model can be written as: $\mathbf{y = X\beta + Qt + e}$, where **X** and **Q** are model matrices which relate observations to their respective $\boldsymbol{\beta}$ and **t**. The assumptions for this model are as shown below:

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{t}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{X\beta}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \text{ and var}\begin{pmatrix}\mathbf{y}\\ \mathbf{t}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{V} & \mathbf{QS} & \mathbf{R}\\ \mathbf{SQ'} & \mathbf{S} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{pmatrix},$$

where $\mathbf{V = QSQ' + R}$, $\mathbf{S = H}\sigma_t^2$ and $\mathbf{R = I}\sigma_e^2$.

Since samples are independent of each other, the mixed model equations can be

expressed on a sample basis as follows:

$$\begin{pmatrix} \mathbf{Q}_i'\mathbf{Q}_i + \mathbf{H}_i^{-1}\alpha & \mathbf{Q}_i'\mathbf{1}_i \\ \mathbf{1}_i'\mathbf{Q}_i & \mathbf{1}_i'\mathbf{1}_i \end{pmatrix} \begin{pmatrix} \hat{\mathbf{t}}_i \\ \hat{\beta}_i \end{pmatrix} = \begin{pmatrix} \mathbf{Q}_i'\mathbf{y}_i \\ \mathbf{1}_i'\mathbf{y}_i \end{pmatrix}, \text{ where } \alpha = \sigma_e^2 / \sigma_t^2.$$

This leads to a system of equations of order the number of time-periods represented in a sample plus one (for the sample effect). The ordering, such that time-periods precede the sample effect, has implications with regard to computing. REML was chosen as the method of estimation and achieved using an Expectation Maximization algorithm. The Q function (using the notation of Dempster *et al.*, 1977) is shown below, after accounted for constants.

$$= \sum_{i=1}^{s} \left\{ -\left[q_i \ln \sigma_t^2 - \ln |\mathbf{H}_i^{-1}| \right] - \frac{1}{\sigma_t^2} \left[\text{tr}(\mathbf{H}_i^{-1}\mathbf{E}_i^{(t)}) \right] - n_i \ln \sigma_e^2 - \frac{1}{\sigma_e^2} \left[\text{tr}(\mathbf{E}_i^{(e)}) \right] \right\}$$

where s is the number of samples, q_i is the number of time-periods in a sample, $\mathbf{H}_i\sigma_t^2$ is the variance of $\mathbf{t}_i$, n_i is the total number of observations in a sample and $\mathbf{E}_i^{(t)} = [\hat{\mathbf{t}}_i\hat{\mathbf{t}}_i' + \text{var}(\hat{\mathbf{t}}_i - \mathbf{t}_i)]$ and $\mathbf{E}_i^{(e)} = [\hat{\mathbf{e}}_i\hat{\mathbf{e}}_i' + \text{var}(\hat{\mathbf{e}}_i - \mathbf{e}_i)]$. Having taken derivatives of this likelihood with respect to the parameters, set those results equal to zero and solved for the parameters, the updates for both the variance components are found as shown below:

$$\sigma_t^{2^{k+1}} = \left[\frac{\sum_{i=1}^{s} \text{tr}(\mathbf{H}_i^{-1}\mathbf{E}_i^{(t)})}{q} \right]^k \text{ and } \sigma_e^{2^{k+1}} = \left[\frac{\sum_{i=1}^{s} \text{tr}(\mathbf{E}_i^{(e)})}{N} \right]^k.$$

In the case of ρ, a closed form solution was not possible. The update of the correlation coefficient was, therefore, achieved via Fisher scoring as follows:

$$\rho^{k+1} = \rho^k - E\left(\frac{\partial^2 Q}{\partial \rho^2} \right)^{-1} \left(\frac{\partial Q}{\partial \rho} \right).$$

Simulation was performed 1) to help verify the methodology and 2) to assess the efficiency of the methods in reaching the desired degree of convergence. Convergence was defined in all studies as that round of iteration when consecutive estimates of *all* parameters differed by less than 10^{-5}.

RESULTS AND DISCUSSION

The results from simulation are shown in Table 1 and the methods derived seem efficient; they may be useful for the proposed application (CG in sire

evaluation) as well as others. It is too early to conclude that an AR(1) *is* or *is not* appropriate for an application like treatment of CG as random in sire evaluation; the primary aim of this work was to derive methods for an efficient implementation of such an error structure and to begin investigation into the proposed application. An initial study of field data (milk and fat production), wherein month was considered random AR(1) within fixed herd-year, yielded a correlation coefficient between months in the same herd-year of .8 for both traits. The effort involved in incorporating a structure like the one discussed here is trivial, and provides a method for easily accommodating small covariances between months that are far apart.

Table 1 Simulation Results †

Starting values			Rounds to	Estimates of parameters		
ρ	σ_t^2	σ_e^2	convergence	$\hat{\rho}$	$\hat{\sigma}_t^2$	$\hat{\sigma}_e^2$
0.7	4	20	41	0.71	3.83	20.10
−0.7	20	4	50	0.71	3.83	20.10
0.01	400	400	49	0.71	3.83	20.10
10^{-8}	4	20	45	0.71	3.83	20.10

† True values for σ_t^2 were 4 (units)^2, 20 (units)^2 for σ_e^2 and 0.7 for ρ. Numbers of observations per sample and per time period were assigned randomly yielding a total of 21,815 observations contained in 2321 year-months.

REFERENCES

Brotherstone, S., Hill, W. G. and Thompson, R. 1989. Animal Production 48: 283.
Chauhan, V. P. S. 1987. Livestock Production Science 16: 117.
Chauhan, V. P. S. 1988. J. Anim. Breed. Genet. 105: 294.
Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. J. Roy. Stat. Soc. Ser. B. 39: 1.
Henderson, C. R. 1975. Biometrics 31: 423.
Meyer, K. 1987. Livestock Production Science 17: 95.
Preisinger, R., Claus, J. and Kalm, E. 1986. 3rd Congr. Gen. Appl. Liv. Prod. 12: 409.
Van Vleck, L. D. 1987. J. Dairy Science 70: 2456.