A Multiplicative Random Regression Model for Test-Day Data with Heterogeneous Variance

M. Lidauer and E. A. Mäntysaari
Agrifood Research Finland MTT, Animal Production Research, FIN-31600 Jokioinen, Finland

1. Introduction

Finnish genetic evaluation of dairy cattle is based on test-day (TD) yields, applying a multiple-trait random regression TD model. The model considers the first lactation milk, protein, and fat yield, plus the later lactation milk, protein, and fat yield as different traits, where all the later lactations are modelled as repeated observations (Lidauer et al., 2000). Currently, yields are scaled to correct for differences in variance among later lactations. Heterogeneous variance among herd-test-days (HTD) and time is ignored.

Heterogeneous variance deteriorates BLUP properties of solutions of breeding values, which reduces reliability of ranking of breeding candidates. Accuracy of bull dam selection would benefit most from correcting heterogeneous variance properly. Accounting for heterogeneous variance by fitting multiplicative mixed model equations (Kachman and Everett, 1993) was found most appealing for several reasons as given in Meuwissen et al. (1996). It accounts for genetic differences among breeds, which is important under Finnish circumstances where herds typically consist of different breeds, and the breeds are evaluated simultaneously. Further, it assumes a homogeneous heritability among herds; it accounts for reduced variance in later lactations due to selection; and it avoids selection that favours animals from inferior breeds and from herds with closely related animals (Meuwissen et al., 1996).

Objective of this work was to elaborate, whether it is feasible to correct for heterogeneous variance in large TD data by applying a multiplicative mixed models approach (Meuwissen et al., 1996) for a multiple-trait random regression TD model.

2. Material and methods

2.1 Models

Random regression test-day model
The original model was a multiple-trait multiple-lactation reduced rank random regression TD model, having different model equations for first and later lactations (Lidauer et al., 2000). A TD yield, made on days in milk *d*, was modelled for first lactation as,

$$y_{Fijklmnoq} = AGE_{Fi} + DCC_{Fj} + YM_{Fkl} + HY_{Fmk} + \sum_{r=1}^{5} \phi(d)_r \, CYSP_{Fnr} + htd_{Fmkl} + \sum_{r=1}^{6} s(d)_{Fr} \, a_{or} + \sum_{r=1}^{6} t(d)_{Fr} \, p_{or} + e_{Fijklmnoq},$$

and for later lactations as,

$$y_{Lijklmnopq} = AGE_{Li} + DCC_{Lj} + YM_{Lkl} + HY_{Lmk} + \sum_{r=1}^{5} \phi(d)_r \, CYSP_{Lmr} + htd_{Lmkl} + \sum_{r=1}^{6} s(d)_{Lr} \, a_{o(r+6)} + \sum_{r=1}^{6} t(d)_{Lr} \, p_{o(r+6)} + \sum_{r=1}^{6} t(d)_{Lr} \, w_{opr} + e_{Lijklmnopq},$$

where trait F is first lactation milk, protein, or fat yield *(i.e., F=1,2,3)* and trait L is later lactation milk, protein, or fat yield *(L=4,5,6)*. Thus, there were 6 traits in the statistical model. The model for TD yields of later lactations described observations of different lactations as repeated observations.

The fixed effects were age at calving (*AGE*), days carried calf (*DCC*), year × month of production (*YM*), herd × year of production (*HY*), and a regression function of *d*, nested within calving year × calving season × parity (*CYSP*) (see Lidauer et al., 2000).

The random effects were herd-test-day (*htd*), regression coefficients for the breeding values a_{or} (r=1,..,12), the non-genetic lactation curves within first lactation and across later lactations p_{or} (r=1,..,12), corresponding non-genetic lactation curves within later lactations w_{or} (r=1,..,6), and measurement error $e_{.ijklmnopq}$, respectively.

Multiplicative random regression test-day model
Effects in the multiplicative random regression TD model were identical with those in the original model. However, data were assumed to be homogeneous within strata and heterogeneous across strata. TD observations for the same trait, which belonged to the same year × month × parity class i and the same HTD class j represented a stratum. TD-yields that belonged to a first lactation stratum of trait F, or to a later lactation stratum of trait L were modelled as:

$$\mathbf{y}_{Fij}\lambda_{Fij} = \mathbf{X}_{Fij}\mathbf{b}_F + \mathbf{C}_{Fij}\mathbf{h}_F + \mathbf{Z}_{Fij}\mathbf{a}_F + \mathbf{W}_{Fij}\mathbf{p}_F + \mathbf{e}_{Fij}, \text{ or}$$
$$\mathbf{y}_{Lij}\lambda_{Lij} = \mathbf{X}_{Lij}\mathbf{b}_L + \mathbf{C}_{Lij}\mathbf{h}_L + \mathbf{Z}_{Lij}\mathbf{a}_L + \mathbf{W}_{Lij}\mathbf{p}_L + \mathbf{L}_{Lij}\mathbf{w}_L + \mathbf{e}_{Lij},$$

respectively. The vectors $\mathbf{b}_{.}$, $\mathbf{h}_{.}$, $\mathbf{a}_{.}$, $\mathbf{p}_{.}$, $\mathbf{w}_{.}$, and $\mathbf{e}_{.}$ contained fixed effects, random HTD, genetic animal, non-genetic animal across all lactations and within lactation, non-genetic animal within later lactations, and residual effects of trait F or L, respectively. The matrices $\mathbf{X}_{.ij}$, $\mathbf{C}_{.ij}$, $\mathbf{Z}_{.ij}$, $\mathbf{W}_{.ij}$, and $\mathbf{L}_{.ij}$ were design matrices related to observations in stratum ij, and $\lambda_{.ij}$ was a multiplicative adjustment factor for all observations on a particular trait of stratum ij, and was calculated as $\lambda_{.ij} = e^{-0.5(\beta_{.i}+\beta_{.j})}$.

The β-values are estimates for the heteroskedasticity in the TD data and were estimated simultaneously while solving the model for breeding values. For each of the traits the same model for estimating the heteroskedasticity was defined:

$$s_{ij} = \mu + \beta_{YMP\_i} + \beta_{htd\_j} + \varepsilon_{ij},$$

where s_{ij} was an estimate for stratum ij; μ was the overall mean; β_{YMP_i} was a fixed year × month × parity classification; β_{htd_j} was a random HTD classification; and ε_{ij} was the residual. For random β_{htd}-effects a within herd correlation structure according to a first order autoregressive model (Wade and Quaas, 1993) was assumed.

The pseudo-observation s_{ij} was estimated for each trait as given in Meuwissen et al. (1996). At iteration round q:

$$s_{ij} = \beta^{[q]}_{YMP\_i} + \beta^{[q]}_{htd\_j} + \left(w^{[q]}_{ij}\right)^{-1} z^{[q]}_{ij},$$

where $w^{[q]}_{ij} = \dfrac{\hat{\mathbf{y}}'^{[q]}_{ij}\hat{\mathbf{y}}^{[q]}_{ij}}{4\sigma^2_e} + \dfrac{1}{2n_{ij}}$ is the variance of $z^{[q]}_{ij}$, and $z^{[q]}_{ij} = \dfrac{\mathbf{y}'_{ij}\lambda^{[q]}_{ij}\left(\mathbf{y}_{ij}\lambda^{[q]}_{ij} - \hat{\mathbf{y}}^{[q]}_{ij}\right)}{2\sigma^2_e} - \dfrac{n_{ij}}{2}$;

$\mathbf{y}_{ij}$ is the vector of observations in stratum ij; $\hat{\mathbf{y}}^{[q]}_{ij}$ is the vector of expectations for $\mathbf{y}_{ij}\lambda^{[q]}_{ij}$ in round q; $\lambda^{[q]}_{ij}$ is the multiplicative adjustment factor for stratum ij at round q; σ^2_e is the residual variance of the particular trait; and n_{ij} is the number of observations in stratum ij.

When solving for β-effects, s_{ij} was weighted by w_{ij}, resulting mixed model equations of the form: $\left[\mathbf{S}'\mathbf{W}^{[q]}\mathbf{S} + \Delta\right]\boldsymbol{\beta}^{[q+1]} = \mathbf{S}'\mathbf{W}^{[q]}\mathbf{s}$, where $\mathbf{s}$ was the vector of s_{ij}'s; $\boldsymbol{\beta}$ contained the β-effects; $\mathbf{S}$ was the corresponding design matrix; $\mathbf{W}$ was diagonal with w_{ij}'s on the diagonal; and Δ had the form:

$$\Delta = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}^{-1} \end{bmatrix} 1/\sigma^2_{\beta htd},$$

where **H** was the block diagonal matrix of a first order autoregessive process for the random β_{htd}-effects. The correlation between consecutive HTDs within herd (ρ_{htd}) and the variance $\sigma^2_{\beta htd}$ were estimated via derivative free REML using a sample of the data, which included 600 herds.

Two modifications were made on the approach of Meuwissen et al. (1996). First, variance components for the random regression TD model were not re-estimated. Second, Meuwissen et al. (1996) considered a single trait model. Generalization of the method to multiple traits would require the covariances between traits. This would be a grand methodological and computational challenge. For simplicity, the adjustment factors were estimated independently across traits. To guarantee the same mean for all traits for the adjustment factors, the overall mean for the β-estimates was excluded when calculating $\lambda_{\cdot ij}$.

2.2 Solving strategy

Both, the random regression TD model and the model for β-effects were solved by preconditioned conjugated gradient method. The following iterative scheme was found best in respect to time used to attain the convergence of $\lambda_{\cdot ij}$ factors:

For iteration cycle q=0 start with the initialization:
1) perform 10 rounds of iterations to solve for breeding values
2) set **β=0**; and perform 50 rounds of iterations to solve for β-effects
3) calculate the multiplicative adjustment factors $\lambda_{\cdot ij}$

For iteration cycles q=1,2,…
1) calculate $\mathbf{y}_{ij}\,\lambda_{\cdot ij}$
2) perform 3 rounds[§] of iterations to solve for breeding values
3) calculate s_{ij} and w_{ij} values
4) perform 20 rounds of iterations to solve for β-effects
5) calculate the multiplicative adjustment factors $\lambda_{\cdot ij}$
6) if convergence of $\lambda_{\cdot ij}$ values, exit cycle

Finnish the iteration of breeding values until converged

[§] perform 20 rounds of iterations for q = 20, 40, and 60.

The multiplicative adjustment factors were considered converged when $\left(\lambda^{[q]} - \lambda^{[q-1]}\right)^T \left(\lambda^{[q]} - \lambda^{[q-1]}\right) / \lambda^{[q]T} \lambda^{[q]}$ was smaller than 10^{-7}.

2.3 Test data

The data for testing included all Finnish TD yields from all lactations of cows that calved for their first time after 1987. TD yields were included until June 2000. There were 1.09 million cows, which had on average 9.4 [14.3], 4.4 [6.7], and 4.4 [6.7] TD observations on milk, protein, and fat yield in first and [later] lactations, respectively, giving in total 26.3 million TD records. Pedigree data comprised of 1.57 million animals of three breeds (Ayrshire, Holstein-Friesian and Finncattle).

3. Results

Four estimation cycles were performed to obtain parameters for the autoregressive process. Then the ratio of $\sigma^2_{\beta htd}$ over σ^2_ε was 0.08 [0.16], 0.17 [0.16], and 0.13 [0.15] for first [later] lactation milk, protein, and fat yield, respectively. Estimates for the ρ_{htd} were 0.95 [0.96], 0.82 [0.82], and 0.81 [0.80] for first [later] lactation milk, protein, and fat yield, respectively. A ρ_{htd} of 0.85 was chosen for setting up **H**.

There were 61 million unknowns and 17 million adjustment factors to be estimated. Convergence of the adjustment factors was reached after 54 main cycles. Total computing time was about twice of the time needed for solving the original model.

The method removed heterogeneous variance among HTDs, later lactations, and time. Standard deviation (SD) of the milk yield was 20% larger in herds with best cows than in average herds, but after adjustment for heterogeneous variance it was about same (Table 1). After adjustment, average intra HTD variance of residuals was similar between small and large herds, and among later lactations (Table 2). Further, the multiplicative model removed heterogeneous variance among years, but not among seasons (Figure 1).

Accounting for heterogeneous variance had little effect on estimated breeding values (EBV) of bulls, but large effect on EBVs of cows. The correlation between EBVs from both models were between 0.994 and 0.997 for active bulls (born within 1991 to 1993 and having at least 60 daughters with records) and between 0.983 and 0.987 for cows born in 1996, but correlations were between 0.73 and 0.91 for the best 1000 cows. Only 340 cows remained in the group of best 500 cows when accounting for heterogeneous variance. The genetic SD of EBVs from the multiplicative model were between 13% and 19% smaller for active bulls and between 17% and 21% smaller for cows born in 1996. Genetic trend over time showed between 21% and 25% smaller yearly progress for bulls and between 10% and 23% less for cows when applying the multiplicative model.

4. Conclusions

Applying multiplicative mixed models (Meuwissen et al., 1996), to account for heterogeneous variance, is feasible for large TD models. Results from this study were consistent with findings presented in the literature (Meuwissen et al., 1996; Robert-Granié et al., 1999). However, additional work is needed to elaborate whether the presented method is valid for the multiple-trait model. Further, work should also concentrate on the robustness of the presented method when more data accumulates.

Acknowledgement

We thank the Finnish Animal Breeding Association (FABA) for providing the data and the Finnish Agricultural Data Processing Centre for supplying the computing platform.

Reference

Kachman, S.D. & Everett, R.W. 1993. A multiplicative mixed model when the variances are heterogeneous. *J. Dairy Sci. 76*, 859-867.

Lidauer, M., Mäntysaari, E.A., Strandén, I. & Pösö, J. 2000. Multiple-trait random regression test-day model for all lactations. *Interbull Bulletin 25*, 81-86.

Meuwissen, T.H.E., De Jong, G. & Engel, B. 1996. Joint estimation of breeding values and heterogeneous variances of large data files. *J. Dairy Sci. 79*, 310-316.

Robert-Granié, C., Bonaïti, B., Boichard, D. & Barbat, A. 1999. Accounting for variance heterogeneity in French dairy cattle genetic evaluation. *Livest. Prod. Sci. 60*, 343-357.

Wade, K.M. & Quaas, R.L. 1993. Solutions to a system of equations involving a first-order autoregressive process. *J. Dairy Sci. 76*, 3026-3032.

Table 1. Mean and standard deviation (SD) of a sample of later lactation milk yields (kg) without (y) and with adjustment (y_{adj}) for heterogeneous variance, by different groups of herds. The sample included all milk yields recorded in 1998

		All herds		Herds of best 500 cows	
		y	y_{adj}	y	y_{adj}
Number of herds		17215	17215	383	416
Size of contemporary group	Mean	8.0	8.0	11.2	10.8
	SD	3.9	3.9	5.2	4.8
Average milk yield of contemporary group	Mean	21.8	20.4	25.5	20.3
	SD	3.4	1.6	3.3	1.4
SD of milk yield in contemporary group	Mean	6.5	5.7	7.8	5.8
	SD	1.3	0.7	1.1	0.5

Table 2. Average intra herd-test-day variance of residuals for milk yield (kg^2) of herds with an average milk yield between 24.0 and 25.0 kg, by different model, herd size, and lactation

		Lactation				
	Herd size	1	2	3	4	5
Original	5-12	3.02	5.61	5.57	5.26	5.22
	16-55	3.27	6.03	5.76	5.76	5.10
With adjustment	5-12	2.00	3.20	3.46	3.56	3.67
	16-55	2.03	3.17	3.36	3.36	3.56

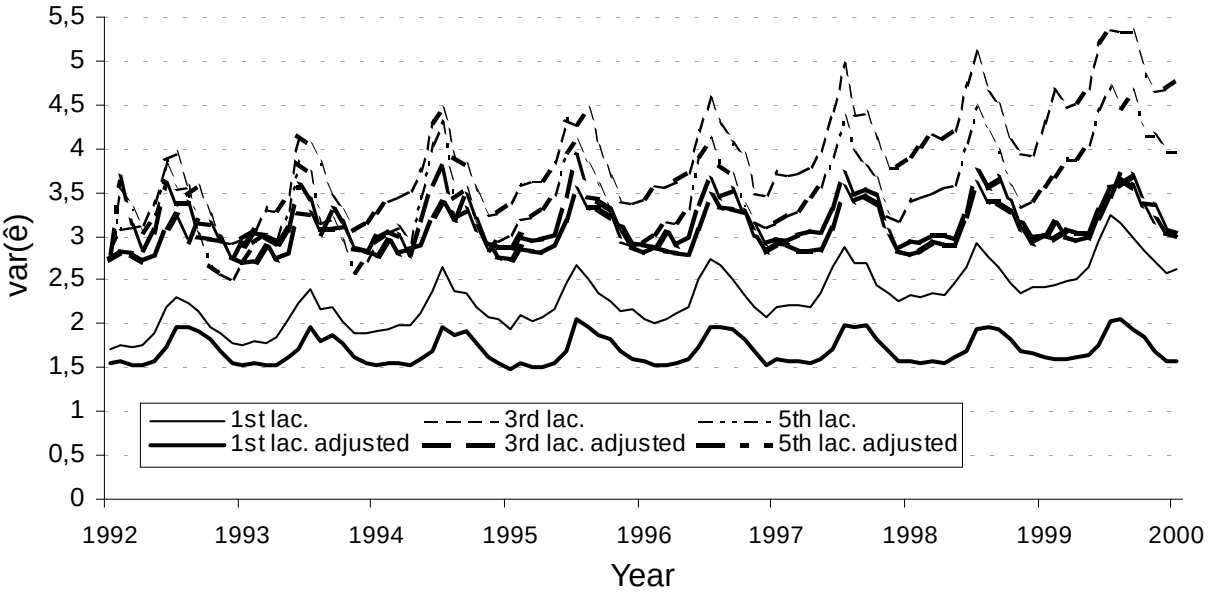

Fig 1. Comparison of residual variance (var(ê)) for milk yield (kg^2) among time between a model without and with adjustment for heterogeneous variance.