Adjustment for Heterogeneous Genetic and Non-Genetic (Co)Variance Structures in Test-Day Models Using a Transformation on Random Regression Effect Regressors

N. Gengler [1,2] and G. R. Wiggans [3]
[1]*National Fund for Scientific Research, B-1000 Brussels, Belgium*
[2]*Animal Science Unit, Gembloux Agricultural University, B-5030 Gembloux, Belgium*
[3]*Animal Improvement Programs Laboratory, Agricultural Research Service,*
USDA, Beltsville, MD 20705-2350, USA

Abstract

A method of accounting for differences in variation in components of test-day milk production records was developed. This method could improve accuracy of genetic evaluations. A random regression model is used to analyze data, and then a transformation is applied to the random regression coefficients. When evaluations with and without this transformation were compared, rankings of animals were similar but some re-rankings occurred. A correlation analysis showed a reduction in effect of production level on variation in regression coefficients, which indicated the success of the transformation. The method suggests a number of innovative solutions to problems related to heterogeneous variance structures.

Introduction

Accounting for heterogeneity of (co)variances among test-day observations is an important component of test-day model developments. The fact that heterogeneity exists in both variances and covariances makes this issue complex to address. Both pre-adjustment and integrated methods have been considered. The objective of this paper is to demonstrate that a feature of random regression models (RRM) or random coefficient models can be used to adjust for heterogeneous (co)variance structures among observations.

Materials and Methods

(Co)variance structures among test-day yields. Consider the following RRM:

$$\mathbf{y} = \mathbf{Ff} + \sum_i \mathbf{Q}_i \mathbf{u}_i + \mathbf{e},$$

where $\mathbf{y}$ = vector of records, $\mathbf{f}$ = vector of fixed effects, $\mathbf{F}$ = incidence matrix linking $\mathbf{y}$ and $\mathbf{f}$, $\mathbf{Q}_i$ = matrix of regressors, $\mathbf{u}_i$ = vector of random effects i, and $\mathbf{e}$ = vector of residuals. The test-day record is considered to be nested in a given lactation of a given animal. The (co)variances among observations for this lactation and animal then would be:

$$\mathrm{Var}(\mathbf{y}) = \sum_i \mathbf{Q}_i \mathrm{Var}(\mathbf{u}_i) \mathbf{Q}_i' + \mathrm{Var}(\mathbf{e}),$$

which can be rewritten as:

$$\mathrm{Var}(\mathbf{y}) = \sum_i \mathbf{Q}_i \mathbf{G}_i \mathbf{Q}_i' + \mathbf{R}.$$

At this stage, the matrices of regressors can be used to generate (co)variance structures by modeling the (co)variances as functions of regression variables:

$$\mathrm{Var}(\mathbf{y}_j) = \sum_i \mathbf{Q}_{ij} \mathbf{G}_{ij} \mathbf{Q}_{ij}' + \mathbf{R}_j,$$

where $\mathbf{G}_{ij}$ are (co)variance matrices of effect i in environment j. If the transformation matrix ($\mathbf{T}$) renders $\mathbf{G}_{ij}$ independent of the heterogeneity strata ($\mathbf{G}_{0i} = \mathbf{T}_{ij} \mathbf{G}_{ij} \mathbf{T}_{ij}'$), then

$$\mathrm{Var}(\mathbf{y}_j) = \sum_i \mathbf{Q}_{ij} (\mathbf{T}_{ij})^{-1} \mathbf{G}_{0i} (\mathbf{T}_{ij}')^{-1} \mathbf{Q}_{ij}' + \mathbf{R}_j.$$

Conceptually this is a simple transformation $\mathbf{T}_{ij}^*$ of regressors, where the matrix of coefficients is "bent" through $\mathbf{Q}_{ij}^* = \mathbf{Q}_{ij} \mathbf{T}_{ij}^* = \mathbf{Q}_{ij} (\mathbf{T}_{ij})^{-1}$ to replace $\mathbf{G}_{ij}$, which differ by environment with matrix $\mathbf{G}_{0i}$ and this for every effect i. This process allows modeling heterogeneous (co)variance structures

easily for both non-genetic and genetic effects. For the latter, the underlying assumption is that genetic correlations between environments are unity for every transformed regression. Although several possibilities exist for $\mathbf{T}$, the obvious candidate is the inverse of the lower Cholesky decomposition because $\mathbf{G}_{0i}$ then becomes an identity matrix.

This simple generalization of the standardization of random effects approach is used in France (Robert-Granié *et al.*, 1999). Here the order of random regressions can be chosen so that the first transformed regression is defined as the standardized constant term. Also, heterogeneity in $\mathbf{G}_{ij}$ can be modeled by modeling $\mathbf{T}_{ij}^{*}$. Therefore, if $\mathbf{T}_{ij}^{*}$ is assumed to be dependent on a continuous variate (e.g., production level within heterogeneity strata), the coefficient of that matrix can be modeled as a function of that variate. Updating of those matrices during solution of the mixed-model equations also would be possible.

Heterogeneous error variances can be modified in a similar fashion by replacing $\mathbf{Q}$ with an identity matrix. Mixed-model equations then are adjusted by imposing weighting according to assumed residual variance of a given record.

Example computations. As a demonstration, a general data set was used to compute variance components in four sub-populations differentiated by production level. Herds could change production level after 2 yr. The 222,679 first-lactation test-day yields from herds in New York, Wisconsin, and California were grouped into four subsets of between 55,604 and 55,685 records. Yields were pre-adjusted additively for age and lactation stage. Adjustment factors were those of Bormann *et al.* (2002). Difference between means of cows in the highest and lowest production subsets was 14.1 kg (37.4 *vs.* 23.3).

(Co)variance components were estimated using EM-REML and the following random regression model:

$$\mathbf{y} = \mathbf{X}\mathbf{t} + \mathbf{Q}_{\mathbf{h}}\mathbf{h} + \mathbf{Q}_{\mathbf{a}}\mathbf{a} + \mathbf{Q}_{\mathbf{p}}\mathbf{p} + \mathbf{e}\,,$$

where $\mathbf{y}$ = vector of test-day records for milk yields; $\mathbf{t}$ = vector of fixed herd, test-day, and milking-frequency class effects; $\mathbf{h}$ = vector of random herd and time (2 yr of calvings) effects; $\mathbf{p}$ = vector of random permanent environmental (PE) effects; $\mathbf{a}$ = vector of animal effects (breeding

values); $\mathbf{e}$ = residual effect; $\mathbf{X}$ = incidence matrix linking $\mathbf{y}$ and $\mathbf{t}$; $\mathbf{Q}$ = matrix of constant, linear, and quadratic modified Legendre polynomials: $I_0 = 1$, $I_1 = 3^{0.5}x$, and $I_2 = (5/4)^{0.5}(3x^2 - 1)$, where $x = -1 + 2[(DIM - 1)/(365 - 1)]$; $\mathbf{Q}_{\mathbf{h}}$, $\mathbf{Q}_{\mathbf{a}}$, and $\mathbf{Q}_{\mathbf{p}}$ = matrices of regressors (r^0, r^1, and r^2) linking $\mathbf{y}$ with $\mathbf{h}$, $\mathbf{a}$, and $\mathbf{p}$. The herd-time effect was introduced because earlier studies on the same data set showed that the portion of total variance explained by that effect was not negligible (Gengler & Wiggans, 2001).

Estimated (co)variance components were transformed into lower Cholesky triangular matrices: $\mathbf{L}_{ij} = \text{chol}(\mathbf{G}_{ij})$. Every coefficient k of the matrices was modeled as a constant, linear, and quadratic function of standardized production level p:

$$l_{ijk} = \alpha_{0ik} + \alpha_{1ik}p_j + \alpha_{2ik}p_j^2\,.$$

Data were analyzed with and without adjustment for heterogeneous variance (HV). Regular mixed-model equations using the mean coefficients ($\bar{l}_{ijk} = \alpha_{0ik}$) provided solutions without HV adjustment. Mixed-model equations with transformed regressions, where $\mathbf{T}$ were from equations using standardized mean herd test-day production over 2 yr, provided solutions with HV adjustment.

Rankings were computed for cows and for sires with ≥ 10 daughters from both unadjusted and HV-adjusted results. Estimated breeding value (EBV) was expressed on a 305-d lactation basis; EBV from the model including transformation were back-transformed to a mean scale. For cows, the same was done for the sum of EBV and PE effects. Second, variances of solutions for genetic and PE effects were computed in every herd, test-day, and milking-frequency class and compared with mean milk production level of that class. If adjustment was successful, correlation between those variances and mean production level should be reduced.

Results and Discussion

(Co)variance components for low-, medium-, and high-producing herds. Raw (co)variances were estimated and modeled. Only mean and extreme values for p (-1, 0, 1) are reported. Figures 1, 2 and 3 show heritabilities and relative vari-

ances for PE and common herd environment, respectively.

Heritabilities were substantially higher for high-producing herds Medium-producing herds were intermediate for heritability, which was only partially reflected in PE, however, as only high-producing herds differed substantially.

Common herd-time environment did not show such a clear pattern. Low-producing herds showed a higher level of variance at the beginning of lactation, medium-producing herds at the end of lactation, and high-producing herds in the middle of lactation.

As shown in Figure 4, the pattern of phenotypic variance over a lactation was not consistent across herd production levels. In low-producing herds, variances were nearly constant. In medium-producing herds, variances increased slightly in later lactation, and in high-producing herds, the rate of this variance increase was substantially greater. Therefore, phenotypic variance correction should allow different corrections according to lactation stage.
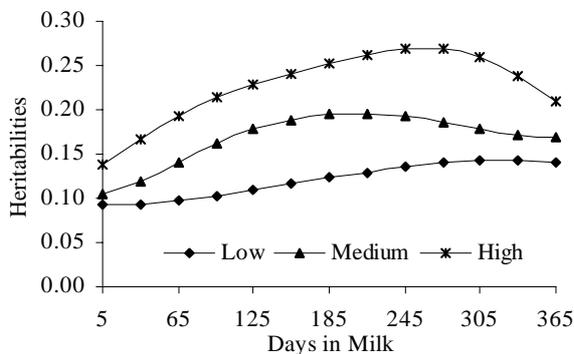

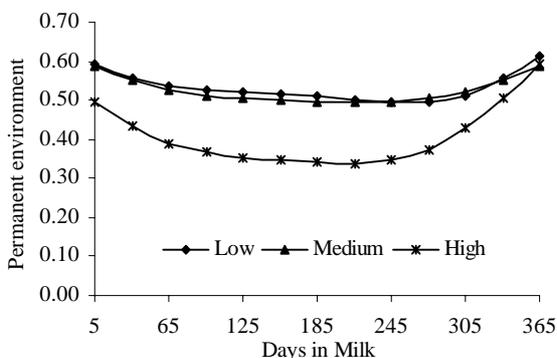**Figure 1.** Heritabilities in high-, medium-, and low-producing herds.

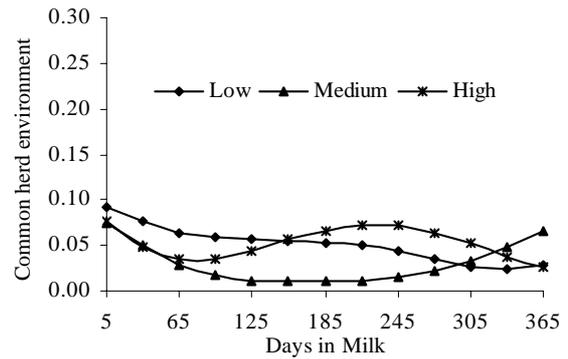**Figure 2.** Relative PE variances in high-, medium-, and low-producing herds.

Figure 3. Relative common herd-time environmental variances in high-, medium-, and low-producing herds.

Although phenotypic correlations were remarkably stable, genetic correlations decreased with production level, especially for very low-producing herds.. Figures 5 and 6 show this for the correlation of day 5 with other days in milk. This finding could have a consequence on animal ranking because assuming correlations that are too high might produce some problems in early evaluations of dairy sires if a bull is tested in low-producing herds and has only daughters in early lactation.
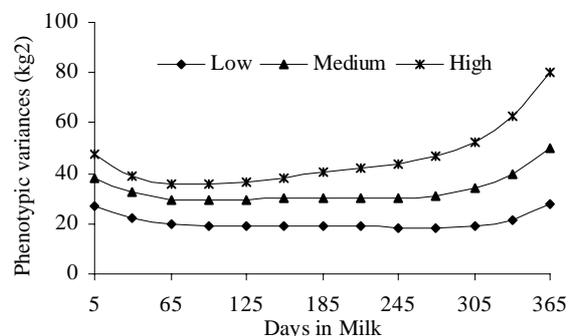

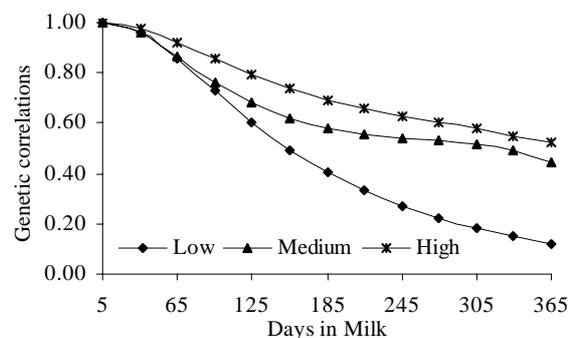**Figure 4.** Absolute phenotypic variances in high-, medium-, and low-producing herds.

Figure 5. Genetic correlations between day 5 and other days in milk in high-, medium-, and low-producing herds.
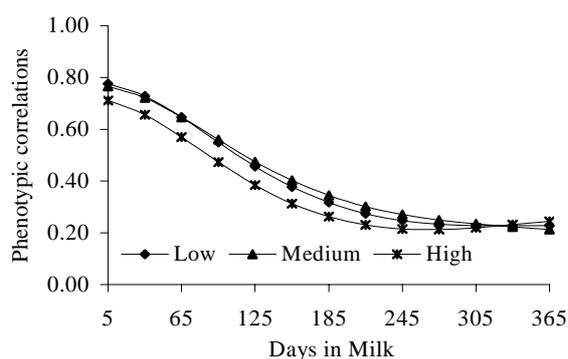
Figure 6. Phenotypic correlations between day 5 and other days in milk in high-, medium-, and low-producing herds.

Comparison of rankings with and without HV adjustment. Rank correlations for EBV were over 0.99 and over 0.98 for PE. However, some re-ranking still occurred as shown in Tables 1 and 2 for the top 10 cows and sires with ≥10 daughters with records, respectively. The greatest re-ranking occurred for EBV + PE. Although not totally free from re-ranking, EBV were quite stable. One bull with 211 daughters gained 10 positions, which is a large increase for an animal with that many daughters.

Table 1. Comparison of rankings for 305-d lactation EBV and PE effects with and without HV adjustment for top 10 cows.

Ranking with HV adjustment	Ranking without HV adjustment	
	EBV	EBV + PE
1	1	5
2	3	4
3	2	7
4	4	9
5	5	3
6	8	1
7	6	10
8	15	24
9	11	2
10	25	50

Table 2. Comparison of rankings for 305-d lactation EBV with and without HV adjustment for top 10 bulls with ≥10 daughters with records.

EBV ranking		
With HV Adjustment	Without HV adjustment	Daughters (no.)
1	1	54
2	2	67
3	3	159
4	14	14
5	6	141
6	16	211
7	4	10
8	11	10
9	7	10
10	9	12

Comparison of intra-herd, test-day, and milking-frequency class variances with and without HV adjustment. Table 3 shows that variances of genetic and PE regressions solutions within herd, test-day, and milking-frequency classes were less correlated with class mean yields with HV adjustment. However, the reduction in correlation was less for genetic than for PE regressions, for which correlations became nearly 0. The most likely explanation is that the assumption of perfect genetic correlations across environments reduces effects of HV adjustment, which also could explain the low level of re-rankings.

Table 3. Correlations of Legendre coefficient variances with milk yield levels with and without HV adjustment.

HV Adjustment	Genetic regression			PE regression		
	r^0	r^1	r^2	r^0	r^1	r^2
No	0.47	0.57	0.56	0.48	0.56	0.54
Yes	0.42	0.40	0.39	−0.02	0.12	0.10

Conclusions

An alternative way to address the issue of heterogeneity of test-day yield (co)variances is to transform regressors for random regression effects. Example computations to demonstrate the method showed that some animal reranking occurs because of the effect of this transformation on genetic and PE effects. In addition, correlations of intra-herd test-day frequency class variances for genetic and PE regressions with production level were reduced with HV adjustment.

Implications

The presented HV adjustment suggests innovative solutions for a number of problems. First, the general concept can be used in both pre-adjustment (single transformation of regressors before data analysis) and integrated (transformation during analysis and updated) evaluations systems. Because every regression of every test-day yield of a given cow can be adjusted, extreme flexibility can be achieved. For example, for multi-breed evaluation, differences in (co)variance structures among breeds can be accommodated and even crossbreds included by interpolation based on proportion of genes. This benefit could be especially important if breeds are to be evaluated together because of their simultaneous presence in contemporary groups or the existence of crossbreds (e.g., Jerseys and Holsteins in the United States and dual-purpose Belgian Blues and Holsteins in Belgium).

Acknowledgments

References

Bormann, J., Wiggans, G.R., Druet, T. & Gengler, N. 2002. Estimating effects of permanent environment, lactation stage, age, and pregnancy on test-day yield. *J. Dairy Sci. 85* (Jan), online.

Gengler, N. & Wiggans, G.R. 2001. Variance of effects of lactation stage within herd by herd yield. *J. Dairy Sci. 84 (Suppl. 1),* 216 (abstr. 896).

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.