JOINT ESTIMATION OF BREEDING VALUES AND HETEROGENEOUS VARIANCES BY MIXED MODEL EQUATIONS IN LARGE NATIONAL DATA SETS

T.H.E. Meuwissen[1] and G. de Jong[2]

[1]ID-DLO, Box 501, 3700 AM Zeist, The Netherlands
[2]NRS, Box 454, 6088 AL Arnhem, The Netherlands

ABSTRACT

A breeding value evaluation method that corrects for phenotypic heterogeneous variances is presented. The method is computationally simple and feasible for large scale data sets. It involves iteration on the common animal model equations corrected for heterogeneity and estimation of heterogeneity factors from a set of mixed model equations.

Variance correction factors are determined with full account of fixed effects, e.g., breeds, which can increase within herd-year-parity variances. Further, they account for all selection, which mainly affects variances of records of later parities.

INTRODUCTION

Numerous studies indicate heterogeneity of within herd variances of milk production data (Everett et al., 1982; Hill et al., 1983; Lofgren et al., 1985; Meinert et al., 1988). Breeding value evaluation methods that account for this heterogeneity of variances seem therefore pertinent. Although not accounting for heterogeneity of variances decreases rates of genetic gain only marginally (Meuwissen and Van der Werf, 1993), a simple correction method reduced biases of breeding values by about 25% (Van der Werf et al., 1994).

Hill (1984) suggested scaling of records by a posterior estimate of the phenotypic variance, which involved regression of the estimated standard deviation towards a mean value. This scaling with phenotypic variances assumed homogeneity of heritability. A similar procedure could be applied to account for heterogeneity of heritability, but sampling errors of estimates of within herd heritability are so large that regressed estimates will differ little from the mean heritability (Visscher and Hill, 1992).

Wiggans and VanRaden (1991) implemented a heterogeneity of variance correction in the US dairy cattle breeding value evaluation system. Because of the large size of the data set, phenotypic variances were simply estimated by: $(y_i'y_i-(1'y_i)^2/n_i)/(n_i-1)$, where y_i = vector of records in herd-year-parity i, 1 = vector of ones, and n_i = number of records. These estimates were regressed towards a year-region-parity mean, where also information of adjacent years of the herd were used. They neglected variance due to breeds (or phantom groups), covariances due to genetic relationships and variance reduction due to selection.

If the estimation of within herd-year-parity variances neglects breed effects, these will inflate the estimate of the variance. Subsequent correction for heterogeneous variances will thus unjustly reduce differences between breeds. Hence, superior breeds will be underestimated and inferior ones overestimated. This results in reduced selection of animals from superior breeds, which may decrease rates of gain substantially.

A substantial proportion of the cows will be culled on first lactation records, such that cows with later lactations will show reduced variance. Hence, in herd-years-parity classes of later parities, variances are expected to be reduced by selection. However, simple methods, that correct for heterogeneity, will inflate the variance within later herd-year-parities to that of unselected records. Accounting for variance reduction due to selection seems therefore pertinent.

The aim of this paper is to present a heterogeneity of phenotypic variances correction method, that accounts for breed or genetic group effects and variance reduction due to selection, and that can be implemented in large scale breeding value evaluation methods. We intend to implement the method in the Dutch breeding value evaluation system.

METHODS

Models

The data are modeled for herd-year-parity i by:

$$y_i = (X_i b + Z_i u + e_i) \exp(\tfrac{1}{2}\gamma_i) \qquad [1]$$

where : $\mathbf{b}$ = fixed effect vector (genetic groups and herd-year-seasons); $\mathbf{u}$ = vector of breeding values; $\mathbf{e}_i$ = vector of environmental effects; $\mathbf{X}_i$ and $\mathbf{Z}_i$ are design matrices for fixed effects and breeding values within herd-year-parity i. $\text{Var}(\mathbf{u}) = \mathbf{A}\sigma_a^2$, where $\mathbf{A}$ is the additive genetic relationship matrix and σ_a^2 is the additive genetic variance. $\text{Var}(\mathbf{e}_i) = \mathbf{I}\sigma_e^2$, where σ_e^2 = the residual variance. The values of σ_a^2 and σ_e^2 are assumed known, which implies a constant heritability across herd-year-parities.

The fixed plus random effects within herd-year-parity i are scaled by a factor $\exp(\frac{1}{2}\gamma_i)$ to obtain the records, which resembles the multiplicative mixed model of Kachman and Everett (1993). The variance of a record in herd-year-parity i is $(\sigma_a^2+\sigma_e^2)*\exp(\gamma_i)$. The exponential of γ_i is taken to ensure that this variance is positive for every estimate of γ_i. Furthermore, variances of estimates of variance tend to increase with their size, i.e. $\text{Var}(\hat{\sigma}^2) = 2\sigma^4/(n-1)$, where n = the number of records. A log transformation renders this variance approximately constant: $\text{Var}(\ln(\hat{\sigma}^2)) \approx 2/(n-1)$. Hence, $\ln(\hat{\sigma}^2) = \ln(\exp(\gamma_i)) = \gamma_i$ has approximately constant error variance, which is desirable when γ_i analysed by a statistical model.

The following linear model for γ_i is assumed:

$\gamma_i = \mathbf{S}_i\boldsymbol{\beta}$,

where $\boldsymbol{\beta}$ = vector with effects on γ_i; $\mathbf{S}_i'$ = design vector.

In the following section, the effects $\boldsymbol{\beta}$ are assumed fixed for simplicity. After this, it will be indicated that $\boldsymbol{\beta}$ should be analyzed as random and the analysis will be extended to random $\boldsymbol{\beta}$. The actual effects that are involved in $\boldsymbol{\beta}$ are not important for the fixed $\boldsymbol{\beta}$ section and will be explained in the section with random $\boldsymbol{\beta}$.

Estimation of γ_i with fixed $\boldsymbol{\beta}$

The derivation of the estimators of γ_i follows that of Foulley et al. (1992) and San Cristobal et al. (1993). We want to maximize the log likelihood of the data for $\boldsymbol{\beta}$:

$$\ln p(\mathbf{y},\mathbf{u}|\boldsymbol{\beta}) = \text{const} - \tfrac{1}{2}\Sigma n_i\mathbf{S}_i\boldsymbol{\beta} - \tfrac{1}{2}\Sigma\mathbf{e}_i'\mathbf{e}_i, \qquad [2]$$

where const does not depend on $\boldsymbol{\beta}$, $\mathbf{e}_i = \mathbf{y}_i\exp(-\frac{1}{2}\mathbf{S}_i\boldsymbol{\beta}) - \mathbf{X}_i\mathbf{b} - \mathbf{Z}_i\mathbf{u})$ (see [1]), n_i = the number of records in herd-year-parity i, and summation is over herd-year-parities. The derivative of [2] to $\boldsymbol{\beta}$ is:

$$-\tfrac{1}{2}\Sigma n_i\mathbf{S}_i + \tfrac{1}{2}\Sigma\mathbf{y}_i'\mathbf{e}_i\exp(-\tfrac{1}{2}\mathbf{S}_i\boldsymbol{\beta})\mathbf{S}_i = \Sigma\mathbf{S}_i\mathbf{z}_i = \mathbf{S}\mathbf{z}, \qquad [3]$$

where $z_i = [y_i'e_i\exp(-\frac{1}{2}S_i\beta)-n_i]/2$. Since, e and, equivalently, u are unknown, we have to take expectations over $u|y,\beta$ of [3] to obtain the derivative of $\ln p(y|\beta)$ (see Foulley et al., 1992):

$$d \ln p(y|\beta) /d\beta = S\hat{z}, \quad [4]$$

where $\hat{z}_i = [y_i'\hat{e}_i\exp(-\frac{1}{2}S_i\beta)-n_i]/2$.

In order to apply the Newton-Raphson algorithm to find the maximum of the likelihood, we need to take second derivatives of [3]:

$$d^2 \ln p(y,u|\beta) /d\beta^2 = -\frac{1}{2}y_i'y_i\exp(-S_i\beta)S_i'S_i + \frac{1}{4}y_i'(X_ib+Z_iu)\exp(-\frac{1}{2}S_i\beta)S_i'S_i$$
$$= S'W_1S, \quad [5]$$

where $W_1 = \text{diag}\{\frac{1}{4}y_i'(X_ib+Z_iu)\exp(-\frac{1}{2}S_i\beta)-\frac{1}{2}y_i'y_i\exp(-S_i\beta)\}$. The derivative of [4] is (Foulley et al., 1992):

$$d^2 \ln p(y|\beta) /d\beta^2 = E_{u|y,\beta}(d^2 \ln p(y,u|\beta) /d\beta^2) + \text{Var}_{u|y,\beta}(d \ln p(y,u|\beta) /d\beta)$$
$$= S'\hat{W}_1S + S'\hat{W}_2S = -S'\hat{W}S,$$

where $\hat{W}_1 = \text{diag}\{\frac{1}{4}y_i'(X_i\hat{b}+Z_i\hat{u})\exp(-\frac{1}{2}S_i\beta)-\frac{1}{2}y_i'y_i\exp(-S_i\beta)\}$

$\hat{W}_2 = \text{diag}\{\frac{1}{4}y_i'T_iCT_i'y_i\exp(-S_i\beta)\}$,

with $T_i = [X_i'Z_i']'$ and C = the inverse of the coefficient matrix of the animal model equations (see [7]); and $\hat{W} = -(\hat{W}_1 + \hat{W}_2)$, which is diagonal. Note that the term $CT_i'y_i$ equals the solution of the animal model equations if only data of herd-year-parity i are available. In large data sets with many herd-year-parities, computation of every $CT_i'y_i$ may not be feasible, but it can be approximated by ignoring genetic relationships across herds.

The Newton-Raphson algorithm becomes now:

$$S'\hat{W}S\beta^{[q+1]} = S'(z + \hat{W}S\beta^{[q]}), \quad [6]$$

where q denotes the iteration number. Note that equations [6] are similar to generalized linear model equations. Each iteration on [6] requires solutions of b and u of the animal model equations:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z+\lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y_c \\ Z'y_c \end{bmatrix}, \quad [7]$$

where $y_c = y\exp(-\frac{1}{2}S_i\beta^{[q]})$.

In this section, all effects on γ_i, i.e. β, were assumed fixed. But, since herd-year-parities may be small, there may be little information on β, in particular, if the model is $\gamma_i = \beta_i$, i.e. there is a fixed effect for every γ_i. More reliable estimates of β can be obtained by regressing every γ_i back to a mean, i.e. β is assumed random (Hill, 1984). Further, by imposing a correlation structure on the random effects β, information on

adjacent herd-year-parity classes could increase the accuracy of the estimate of β_i. This correlation structure will be imposed by a autoregressive model, as explained in the next section.

An autoregressive model for γ_i with random $\boldsymbol{\beta}$

Our derivation is again similar to that of Foulley et al. (1992) and San Cristobal et al. (1993), which used a Bayesian approach. $\boldsymbol{\beta}$ contains fixed effects $\boldsymbol{\beta_1}$, i.e. the mean towards the γ_i are regressed, and random effects $\boldsymbol{\beta_2}$, i.e. the individual effect of herd-year-parity i on γ_i. As fixed effects on γ_i we have chosen the overal mean and the regression factor of γ_i on the average production level in herd-year-parity i, which adds up to the mean towards γ_i is regressed. There is one random effect β_{2i} for every γ_i. Hence,

$\gamma_i = \mathbf{S_{1i}}\boldsymbol{\beta_1} + \mathbf{S_{2i}}\boldsymbol{\beta_2}$,

where $\mathbf{S_{1i}} = [1, p_i]$, with p_i = mean production of herd-year-parity i; and

$\mathbf{S_{2i}}$ = has a one at position i and zeros elsewhere.

Herd-year-parity effects are assumed to be correlated within herd-parities according to an autoregressive model, i.e., if a herd-parity contains 4 herd-year-parities, the variance of the herd-year-parity effects pertaining to that herd is (Wade and Quaas, 1993):

$$\sigma_{hyp}^2 \; * \begin{bmatrix} 1 & a & a^2 & a^3 \\ a & 1 & a & a^2 \\ a^2 & a & 1 & a \\ a^3 & a^2 & a & 1 \end{bmatrix}, \qquad [8]$$

where a = the correlation between consecutive years within a herd. Let H denote the block diagonal variance-covariance of all herd-year-parity effects, which contains a block as [8] for each herd-parity.

Since the variance of the herd-year-parity effects is finite, there is a prior distribution for these effects. As we shall see later, only the first two moments of this prior distribution will enter the estimating equations, hence, we may use the normal distribution as a prior:

$$\ln p(\boldsymbol{\beta}) = \text{const} - \tfrac{1}{2}\boldsymbol{\beta}'\Lambda\boldsymbol{\beta}, \qquad [9]$$

where $\Lambda = \begin{bmatrix} 0 & 0 \\ 0 & H^{-1} \end{bmatrix}$. The inverse of H is easily obtained from Wade and Quaas (1993).

A Bayesian approach maximizes the log of the posterior density, which is the sum of [2] and [9]. The derivative of [9] to $\boldsymbol{\beta}$ is: $-\Lambda\boldsymbol{\beta}$, and the

derivative of the log posterior density is thus: $\mathbf{Sz} - \Lambda\boldsymbol{\beta}$. The second derivative of [9] to $\boldsymbol{\beta}$ is $-\Lambda$, hence, the second derivative of the log posterior density is: $-\mathbf{S}'\hat{\mathbf{W}}\mathbf{S} - \Lambda$. The Newton-Raphson algorithm becomes now:

$$[\mathbf{S}'\hat{\mathbf{W}}\mathbf{S} + \Lambda](\boldsymbol{\beta}^{[q+1]} - \boldsymbol{\beta}^{[q]}) = \mathbf{S}'\mathbf{z} - \Lambda\boldsymbol{\beta}^{[q]}, \qquad [10]$$

which replaces [6]. These equations [10] are used in combination with [7] to obtain estimates of breeding values $\mathbf{u}$, which are from [7].

Note that [10] resembles a mixed model, with $\mathbf{H}$ being the variance of the fixed effects. Hence, estimation of breeding values requires iteration on the common animal model with correction for heterogeneity [7] and a mixed model for the estimation heterogeity factors [10]. The estimation of the parameters of the auto-regressive model, a and σ_{hyp}^2, can be estimated as Wade et al. (1993).

DISCUSSION AND CONCLUSIONS

The multiplicative mixed model that was used here was proposed by Kachman and Everett (1993). Mostly, heterogeneity of variance models only scale the random effects (animal and error), but fixed effects are also scaled in multiplicative models. Fixed effects that are across homogeneous variance classes, e.g., breed effects, will have a different effect depending on the estimate of the scaling factor. Hence, an assumption underlying these multiplicative models is that the best breed performs better in the more variable herd.

The present model differs only slightly from that of Kachman and Everett (1993). They modeled the scaling factor directly instead of γ_i and used an inverted chi-squared prior distribution for the scaling factors. These slight differences resulted in equations that did not resemble mixed model equations and seemed more complicated.

The computationally most demanding step is the calculation of breeding values for each herd-year-parity given the data of that herd-year-parity: $\mathbf{CT}_i'\mathbf{y}_i$. Strictly, this should account for all genetic relationships. Fortunately, only breeding values for the animals within herd-year-parity i are needed, since the term $\mathbf{y}_i'\mathbf{T}_i$ in $\mathbf{y}_i'\mathbf{T}_i\mathbf{C}\mathbf{T}_i'\mathbf{y}_i$ has only non-zero elements for these animals. The latter is in contrast with the method of Everett and Kachman (1993), which needs these breeding value estimates for all animals.

Neglection of more distant genetic relationships, e.g., those across herds, results probably in minor biases of phenotypic variances, because the product of the heritability and the genetic relationship will be small. Therefore, we will implement an C_i, that accounts for relationships within a herd, including those through the sires of the cows in that herd, but more distant relationships will be neglected.

Weigel and Gianola (1993) propose a computationally simple Bayesian method, which could be applied to national evaluations. But, as in the approach of Wiggans and VanRaden (1991), herd-year-parities are assumed independent. Problems with breed effects across herd-year-parities and variance reductions due to selection are ignored.

A model that corrects for phenotypic heterogeneity of variances is presented. When covariances due to distant genetic relationships are neglected in estimating within herd-year-parity variances, the model could be included in national breeding value estimation procedures. The main advantages of the presented model are that estimates of within herd-year-parity variances account for genetic group effects and are corrected for variance reductions due to selection. The derivation of the method followed Bayesian ideas, but the resulting sets of mixed model equations [7] and [10] suggests that a frequentists interpretation is possible (see Gianola et al., 1992).

REFERENCES

Everett, R.W., J.F. Keown, and J.F. Taylor, 1982. The problem of heterogeneous within herd error variances when identifying elite cows. J. Dairy Sci 65 (Suppl. 1): 100.

Foulley, J.L., M. San Cristobal, D. Gianola and S. Im, 1992.Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. Comput. Stat. Data. Anal. 13: 291-421.

Gianola, D., J.L. Foulley, R.L. Fernando, C.R. Henderson, and K.A. Weigel, 1992. Estimation of heterogeneous variances using empirical Bayes methods: theoretical considerations. J. Dairy Sci. 75: 2805-2823.

Hill, W.G., 1984. On selection among groups with heterogeneous variance. Anim. Prod. 39: 473-477.

Hill, W.G., M.R. Edwards, M.-K.A. Ahmed and R. Thompson, 1983.

Heritability of milk yield and composition at different levels and variability of production. Anim. Prod. 36: 59-68.

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

Lofgren, D.L., W.E. Vinson, R.E. Pearson, and R.L. Powell, 1985. Heritability of milk yield at different herd means and variance for production. J. Dairy Sci. 68: 2737-2739.

Meinert, T.R., R.E. Pearson, W.E. Vinson, and B.G. Cassell, 1988. Prediction of daughter's performance from dam's cow index adjusted for within herd variance. J. Dairy Sci. 71: 2220-2231.

Meuwissen, T.H.E., and J.H.J. van der Werf, 1993. Impact of heterogeneous within herd variances on dairy cattle breeding schemes: a simulation study. Livest. Prod. Sci. 33: 31-41.

San Cristobal, M., J.L. Foulley and E. Manfredi, 1993. Inference about multiplicative heteroskedastic components of variance in a mixed linear Guassian model with an application to beef cattle breeding. Genet. sel. Evol. 35: 3-30.

Van der Werf, J.H.J., T.H.E. Meuwissen, and G. de Jong, 1994. Effects of correction for heterogeneity of variance on bias and accuracy of breeding value estimation in Dutch dairy cattle. J. Dairy Sci. (accepted).

Visscher, P.M. and W.G. Hill, 1992. Heterogeneity of variance and dairy cattle breeding. Anim. Prod. 55: 321-329.

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

Wade, K.M., R.L. Quaas, and L.D. VanVleck, 1993. Estimation of the parameters involved in a first order autoregressive process. for contemporary groups. J. Dairy Sci. 76: 3033-3040.

Weigel, K.A., and D. Gianola, 1993. A computationally simple Bayesian method for estimation of heterogeneous variances. J. Dairy Sci. 76: 1455-1465.

Wiggans, G.R., and P.M. VanRaden, 1991. Method and effect od adjustment for heterogeneous variance. J. Dairy Sci. 74: 4350-4357.