ORIGINAL ARTICLE

The Unbiased Estimation of Variance Components by a Simple Method

Hiroaki Iwaisaki

Department of Animal Science, Faculty of Agriculture,
Niigata University, Niigata 950-21, Japan

SUMMARY

A set of quadratic forms are proposed, which contain an approximate solution vector and the right-hand side vector in the mixed model equations after absorption of fixed effects on one and another side, respectively. The approximate solutions are based on utilizing information included in the offdiagonal elements as well as in the diagonal elements of the coefficient matrix. The quadratics are translation invariant and whose matrices are not symmetric. Using the quadratics, a method for unbiased estimation of variance components in the mixed linear model is described as a simple approximation to the minimum variance quadratic unbiased estimation. The current approach does not require the calculation of a generalized inverse of the coefficient matrix as do not other approximate approaches. The relative ability of the present method to eliminate bias due to a culling type of selection is examined by Monte Carlo simulation.

1. Introduction

MIVQUE (MINimum Variance Quadratic Unbiased Estimation) of variance components was developed independently by Rao (1971) and LaMotte (1970, 1971). This method is one of those that have attracted the interest of animal breeders because of its relatively desirable properties when observations are normally distributed. For example, MIVQUE estimators have been shown to be free of bias from some forms of selection for animals (e.g., Sorensen and Kennedy, 1984). The computational requirements for MIVQUE, however, are relatively large. LaMotte's MIVQUE algorithm requires the calculation of the inverse of the large variance-covariance matrix of observations. Henderson's (1984a) alternative algorithm still involves the calculation of a generalized inverse of the coefficient matrix of Henderson's (1973, 1975) mixed model equations. Hence, estimation with MIVQUE can be impossible for large data sets.

Recently, researchers have been prolific in finding quadratics resembling those used in MIVQUE and having expectations easier to compute. Examples are Henderson's (1980, 1984b) simple method-D2, Henderson's simple method-D1 (Hudson and Van Vleck, 1982), and Iwaisaki's (1992a) approximate MIVQUE (which will be denoted as HSM-D2, HSM-D1, and IAM-M1, respectively). With a reasonably good choice of the prior information about variance ratios, the estimators by the HSM-D2 method have smaller variances than those by Henderson's (1953) Method 3 (Henderson, 1980; Dempfle et al., 1983). The IAM-M1 method yields estimates less biased by the culling type of selection than the other approximate methods. However, IAM-M1 requires a larger computational burden.

In this context, the objectives of this study were to propose another set of quadratic forms and to describe an approach as a simple approximation

Key words: Approximation; MIVQUE; Mixed linear model; Quadratic form;
Unbalanced data; Variance component.

to MIVQUE of variance components for unbalanced, mixed model situations.

2. Basic mixed linear model

Let

$$y = Xb + \sum_{i=1}^s Z_i u_i + e$$

denote the basic mixed linear model for a variance component estimation where y is a $N \times 1$ vector of observations; b is a $p \times 1$ vector of unknown fixed effects; u_i is a $q_i \times 1$ vector of the effects of the i^{th} random factor with $u' = (u'_1 \ u'_2 \ \dots \ u'_s)$; e is a $N \times 1$ vector of random residuals; and X and Z_i are $N \times p$ and $N \times q_i$ known incidence matrices, respectively.

Assume that $u_1, u_2, \dots, u_s$, and e are mutually independent and

$$E[u_i] = E[e] = 0, \text{Var}(u_i) = I\sigma_i^2 \text{ and } \text{Var}(e) = I\sigma_0^2. \quad (*)$$

Under the assumptions (*), the expectation and dispersion matrix of y are

$$E[y] = Xb \text{ and } \text{Var}(y) = \sum_{i=1}^s Z_i Z_i' \sigma_i^2 + I\sigma_0^2,$$

respectively. The mixed model equations (Henderson, 1973, 1975) are then

$$\begin{bmatrix} X'X & X'Z_1 & \dots & X'Z_s \\ Z_1'X & Z_1'Z_1 + Ik_1 & \dots & Z_1'Z_s \\ \vdots & \vdots & \ddots & \vdots \\ Z_s'X & Z_s'Z_1 & \dots & Z_s'Z_s + Ik_s \end{bmatrix} \begin{bmatrix} b^o \\ \hat{u}_1 \\ \vdots \\ \hat{u}_s \end{bmatrix} = \begin{bmatrix} X'y \\ Z_1'y \\ \vdots \\ Z_s'y \end{bmatrix}$$

with $k_i = \tilde{\sigma}_0^2 / \tilde{\sigma}_i^2$ which is the prior variance ratio.

Absorbing fixed effects from the equations gives

$$\begin{bmatrix} Z_1'MZ_1 + Ik_1 & Z_1'MZ_2 & \dots & Z_1'MZ_s \\ Z_2'MZ_1 & Z_2'MZ_2 + Ik_2 & \dots & Z_2'MZ_s \\ \vdots & \vdots & \ddots & \vdots \\ Z_s'MZ_1 & Z_s'MZ_2 & \dots & Z_s'MZ_s + Ik_s \end{bmatrix} \begin{bmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \vdots \\ \hat{u}_s \end{bmatrix} = \begin{bmatrix} Z_1'My \\ Z_2'My \\ \vdots \\ Z_s'My \end{bmatrix}$$

where $M = I - X(X'X)^{-1}X'$ with $MX = 0$ and $MM = M$.

3. Proposed method

The IAM-M1 method utilizes approximate solutions from the mixed model equations after absorbing fixed effects. The vector of approximate solutions is given as

$$\bar{u}_i = D_i P_i y$$

where $D_i = \{\text{Diag}(P_i Z_i + I_{k_i})\}^{-1}$,

$$P_i = [2Z'_1 - (Z'_1 M Z_1 + I_{k_1}) \{\text{Diag}(Z'_1 M Z_1 + I_{k_1})\}^{-1} Z'_1] M \text{ for } s = 1,$$

$$Z'_i [I - M Z_{(i)} [\text{Diag} \{ \sum_{\substack{j=1 \\ j \neq i}}^s (Z'_j M Z_j + I_{k_j}) \}]^{-1} Z'_{(i)}] M \text{ for } s \geq 2,$$

with $Z_{(i)}$ standing for the matrix in which Z_i is deleted from $Z = (Z_1 Z_2 \dots Z_s)$, and Σ^+ is the direct sum operator. These solutions for $s = 1$ are based on approximately solving the mixed model equations with a modification of further absorbing the equations for the corresponding $q_1 - 1$ levels into the equation for the f^{th} level ($f = 1, 2, \dots, q_1$), in which $\{\text{Diag}(Z'_{(f)} M Z_{(f)} + I_{k_1})\}^{-1}$ is used in place of $(Z'_{(f)} M Z_{(f)} + I_{k_1})^{-1}$, where $Z_{(f)}$ stands for the $N \times (q_1 - 1)$ matrix with the f^{th} column deleted from Z_1 . Note that D_1 used is a form to give lower bounds for the diagonal elements of the inverse of the coefficient matrix which is superior to $\{\text{Diag}(Z'_1 M Z_1 + I_{k_1})\}^{-1}$. For $s \geq 2$, the similar idea is applied in a more simplified way.

Then, using the same vector, the set of quadratic forms proposed in this paper are

$$y' Q_i y = \bar{u}'_i Z'_i M y \text{ and } y' Q_0 y = y' M y.$$

It is satisfied that $Q_i X = 0$ and $Q_i M = Q_i$. Note that Q_i is not symmetric. The expectations of these quadratics become

$$E(y' Q_i y) = \sum_{j=1}^s \text{tr}(Z'_j P_i D_i Z'_i M Z_j) \sigma_j^2 + \text{tr}(Z'_i P_i D_i) \sigma_0^2$$

and

$$E(y' Q_0 y) = \sum_{j=1}^s \text{tr}(Z'_j M Z_j) \sigma_j^2 + \{N - r(X)\} \sigma_0^2,$$

where $r(X)$ is the rank of X .

Hence, the estimators by the current approach (which will be denoted as IAM-M3) are obtained as

$$\hat{\alpha} = H^{-1} \beta$$

where $\alpha' = (\sigma_0^2 \sigma_1^2 \sigma_2^2 \dots \sigma_s^2)$, $\beta' = (y' Q_0 y \ y' Q_1 y \dots y' Q_s y)$, and H is the appropriate matrix whose elements are the coefficients of σ_j^2 ($j = 1, 2, \dots, s$) and σ_0^2 in the expectations of the quadratics. These estimators are translation invariant, quadratic, and unbiased.

4. Simulation

The proposed method was compared to MIVQUE, HSM-D1, HSM-D2, and IAM-M1 methods by Monte Carlo simulation. A mixed model data structure was simulated assuming a model containing herd, time-period, sire, and residual effects. No interactive effects were assumed. Whereas herd and time-period effects were regarded as fixed, sire and residual effects were assumed as random. Two types of data sets were created, that is, unselected data not containing selection and selected data containing a culling type of selection on sires. Sire and residual

variances (denoted as σ_s^2 and σ_e^2 , respectively) used to generate data were 1 and 15, respectively, which corresponds to heritability ($4\sigma_s^2/(\sigma_s^2 + \sigma_e^2)$) of .25. Sire and residual values were assumed to follow normal distributions with zero means and variances given. Herd and time-period effects were generated as normal random variables and then treated as fixed effects in the analyses. The proportions of variations due to herd and time-period effects to total variation were set to 30% and 20%, respectively. The numbers of levels for herd and time-period were ten and two, respectively. A total of thirty sires were generated. In the first period, with unselected and selected data, progeny of each sire were assigned at random to herds. With unselected data, only 10 sires chosen randomly had an additional progeny in the second period that was randomly distributed across herds. With selected data, only top 10 sires selected being based on their true values had an additional progeny in the second period. The numbers of progeny per sire in the first and second periods varied randomly ranging from 10 to 50 and from 40 to 80, respectively. In this simulation, the prior information about the error to sire variance ratio was assumed to be true. Five hundred replicates were conducted with different random seed numbers for each of the two cases of unselected and selected data, and the five methods were applied to each data set.

The estimated biases, variances, and mean squared errors (MSEs) of the estimators by the five methods are shown in Tables 1 and 2 for unselected and selected data, respectively. With unselected data, the biases for σ_s^2 and σ_e^2

Table 1
Estimated biases, variances, and MSEs with unselected data^{a)}

Variance component		Method				
		IAM-M3	MIVQUE	HSM-D1	HSM-D2	IAM-M1
σ_s^2	Bias	.029	.027	.028	.023	.026
	Variance	.169	.144	.172	.146	.144
	MSE	.170	.145	.173	.147	.145
σ_e^2	Bias	.042	.042	.043	.047	.045
	Variance	.321	.316	.322	.319	.319
	MSE	.323	.318	.323	.321	.321

^{a)}The true values of σ_s^2 and σ_e^2 were 1 and 15, respectively.

Table 2
Estimated biases, variances, and MSEs with selected data^{a)}

Variance component		Method				
		IAM-M3	MIVQUE	HSM-D1	HSM-D2	IAM-M1
σ_s^2	Bias	.376	.120	.444	.383	.219
	Variance	.067	.115	.058	.064	.091
	MSE	.208	.129	.255	.211	.139
σ_e^2	Bias	.132	.044	.195	.139	-.015
	Variance	.301	.301	.304	.314	.312
	MSE	.318	.303	.342	.333	.312

^{a)}The true values of σ_s^2 and σ_e^2 were 1 and 15, respectively.

were obviously negligible for all methods so that the estimated variances paralleled MSEs very closely, and MSEs for the five methods were quite similar. When applied to the data containing selection on sires, for the estimation of residual variance, the IAM-M3 and the other methods also performed well and showed MSEs similar to those with unselected data.

For the estimation of sire variance with selected data, all methods underestimated the parameter in different degrees. The MIVQUE estimator was least biased for this intensity of selection, as would be expected. The estimator by IAM-M1 was biased downwards by 22%. The current IAM-M3 method reduced selection bias as much as 62%, which was comparable to the HSM-D2 method and was better than the HSM-D1 method. On the other hand, variance of the IAM-M3 estimator was .067 and was similar to that of the HSM-D2 estimator. Variance of the MIVQUE estimator was about two times larger than those with IAM-M3 and HSM-D2 and was largest among the five methods. The magnitude of variance of IAM-M1 estimator ranked in the middle of MIVQUE and IAM-M3. As a result, based on MSE, the approximate methods ranked from best to worst: IAM-M1, IAM-M3 and HSM-D2, and HSM-D1. MSEs with IAM-M3 were about 1.6 times larger than that with MIVQUE.

5. Discussion

In this paper, a method is proposed for the unbiased estimation of variance components, which is applicable to unbalanced, mixed model situations. The present method is a simpler approximation to MIVQUE than the IAM-M1 method. Although the IAM-M1 method uses $\bar{u}_i$ on both sides of the quadratics other than for error, the current IAM-M3 method uses $\bar{u}_i$ only on one side in the quadratic. $Z'_i My$ is used on another side. $Z'_i My$ is used on both sides of the quadratics in MIVQUE-0 (Seely, 1971) which is a good strategy for variances approaching zero. $Z'_i My$ is also contained on one side in the HSM-D1 quadratics. Use of $Z'_i My$ gives relatively more weight the levels with many observations. The results in the simulation study indicate that the IAM-M3 proposed, as well as the MIVQUE and the other approximate methods, performs well with cross-classified, unbalanced and unselected data.

MIVQUE has a considerable ability to eliminate bias due to selection. MIVQUE uses the exact solutions of, and inverse elements of the coefficient matrix of, mixed model equations, which contain the relevant information to adjust for selection. The HSM-D1 and HSM-D2 approaches utilize only information included in the diagonals of the coefficient matrix after absorbing fixed effects. In other words, they do not utilize any information contained in the off-diagonal elements. The IAM-M3 method utilizes information of the off-diagonal elements in addition to the diagonal elements, though use of the information is further simplified compared to IAM-M1. For the data structure containing selection, IAM-M3 is inferior to IAM-M1 in the power to estimate the parameter for the factor on which selection operated. An advantage of the IAM-M3 method, however, would be that this method is computationally less demanding than IAM-M1 and is likely to reduce selection bias at least as well as HSM-D2. With hierarchical and selected data, it might be possible that IAM-M3 is significantly better than HSM-D2, since the off-diagonal elements are at least partially accounted for by IAM-M3. The hierarchical nature of data structure is completely ignored by HSM-D1 and HSM-D2 methods.

MIVQUE, disregarding differences in constraints on the parameter space, is equal to one round of REML (REstricted Maximum Likelihood) of Patterson and Thompson (1971) for a Fisher's Method of Scoring algorithm. In addition to MIVQUE, the likelihood methods like REML and ML (Hartley and Rao, 1967) are also known to be free of some forms of selection bias (e.g., Thompson, 1973; Meyer and Thompson, 1984). Some researchers have recently developed REML-like methods

known as PE (Pseudo-Expectation) methods (Schaeffer, 1986; Van Raden and Yung, 1988; Iwaisaki, 1992b). PE methods mimic the computational steps in REML by the Expectation-Maximization algorithm (Dempster et al., 1977) in different degrees and provide less computational burden than EM-REML. These iterative methods, as well as the approximations to MIVQUE, do not require inversion of the coefficient matrix of mixed model equations. In this context, it would also be of interest to further examine the relative ability of IAM-M3 to account for selection compared to PE methods.

REFERENCES

- Dempfle, L., Hagger, C., and Schneeberger, M. (1983). On the estimation of genetic parameters via variance components. *Genetics Selection Evolution* 15, 425-444.
- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data with EM algorithm. *Journal of the Royal Statistical Society B* 39, 1-38.
- Hartley, H.O. and Rao, J.N.K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* 54, 93-108.
- Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* 9, 226-252.
- Henderson, C.R. (1973). Sire evaluation and genetic trends. In: *Proceedings of the Animal Breeding and Genetics Symposium in honor of Dr. J. L. Lush*, pp.10-41. Champaign, Illinois, ASAS and ADSA.
- Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* 31, 423-447.
- Henderson, C.R. (1980). A simple method for unbiased estimation of variance components in the mixed model. Mimeo., Department of Animal Science, Cornell University, Ithaca, New York.
- Henderson, C.R. (1984a). ANOVA, MIVQUE, REML, and ML algorithms for estimation of variances and covariances. In: *Statistics: An appraisal-Proceedings of the 50th Anniversary Conference in Iowa State Statistical Laboratory*, H.A. David and H.T. David (Eds.). pp.257-280. Ames, Iowa State University Press.
- Henderson, C.R. (1984b). *Applications of linear models in animal breeding*. Guelph, Ontario, Canada, University of Guelph Press.
- Hudson, G.F.S. and Van Vleck, L.D. (1982). Estimation of components of variance by Method 3 and Henderson's new method. *Journal of Dairy Science* 65, 435-441.
- Iwaisaki, H. (1992a). An approximate MIVQUE method for unbiased estimation of variance components in the mixed linear model. *Biometrical Journal* (in press).
- Iwaisaki, H. (1992b). A modified pseudoexpectation method for variance component estimation. *Journal of Animal Breeding and Genetics* (in press).
- LaMotte, L.R. (1970). A class of estimators of variance components. Technical Report No. 10, Department of Statistics, University of Kentucky, Lexington, Kentucky.
- LaMotte, L.R. (1971). Locally best quadratic estimators of variance components. Technical Report No. 22, Department of Statistics, University of Kentucky, Lexington, Kentucky.
- Meyer, K. and Thompson, R. (1984). Bias in variance and covariance component estimators due to selection on a correlated trait. *Journal of Animal Breeding and Genetics* 101, 33-50.
- Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554.

- Rao, C.R. (1971). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis* **1**, 445-456.
- Schaeffer, L.R. (1986). Pseudo expectation approach to variance component estimation. *Journal of Dairy Science* **69**, 2884-2889.
- Seely, J. (1971). Quadratic subspaces and completeness. *Annals of Mathematical Statistics* **42**, 710-721.
- Sorensen, D.A. and Kennedy, B.W. (1984). Estimation of genetic variances from unselected and selected populations. *Journal of Animal Science* **59**, 1213-1223.
- Thompson, R. (1973). The estimation of variance and covariance components with an application when records are subject to culling. *Biometrics* **29**, 527-550.
- Van Raden, P.M. and Jung, Y.C. (1988). A general purpose approximation to restricted maximum likelihood: The tilde-hat approach. *Journal of Dairy Science* **71**, 187-194.

Recieved January 1992; Accepted March 1992