Detection of Heterogeneous Variance in Herd Production Groups

THOMAS R. FAMULA
Department of Animal Science
University of California
Davis 95616

ABSTRACT

The detection of heterogeneous variance across herd groups is based largely on data stratified by mean herd production. Such stratification is analogous to selection on herd means. This introduces the possibility of biased estimates of genetic and residual variances. Results presented show that Henderson's Method 3 estimates of the residual variance are unbiased by selection on herd means. However, reductions in sums of squares for sires have different expected values in no selection and selection models. The magnitude of this bias is evaluated with a small-scale simulation. Extension of these results to other methods of variance estimation reveals a potential bias from selection on herd means.

INTRODUCTION

A topic of continuing interest to dairy geneticists is the identification and estimation of heterogeneous variances for milk production (1, 7, 8). Evidence for the existence of heteroscedasticity is abundant (6, 16) and generally reveals a positive relationship between mean herd production and variance, although heritability estimates have not shown a consistent rise with increasing mean herd production.

Homogeneity of variance is not a requirement of BLUP of breeding values, as recently discussed by Gianola (3). Provided that estimates of the heterogeneous covariance structure are available, BLUP has the capacity to weight appropriately individual and progeny information across herds with differing variability.

Received August 8, 1988.
Accepted October 19, 1988.

In attempting to establish the existence of changing variances from one herd to the next, the most common procedure is to stratify herds based on mean milk production. Thus, for example, De Veer and Van Vleck (1) divided their data into low, medium, and high producing herds. Using REML, their next step was to estimate the genetic and environmental variance within each production stratum (as well as genetic covariances across strata) using all data simultaneously. Other reports (7) have segmented the data into low, medium, and high groups and estimated variances from each set independently, usually with least squares based methods of variance estimation (4). The problem with either process of variance estimation is that the data are divided (selected) into strata based on the trait in which variances are to be estimated. Stratification on a correlated trait will also affect the estimation of variances. The effect of this stratification is to change the nature of the problem to one of estimating variance from selected data. The entire data set may, perhaps, be regarded as an independent sample. However, the production groups are obviously the result of selection.

Schaeffer (14) recently examined the estimation of variance components in a selection model in the more usual setting of multiple traits and sequential selection. A similar presentation will be considered here, for the specific case of selection that governs the dividing of data into herd production groups. The methods discussed will be those most typically used in problems of this sort, Henderson's (4) Method 3 and minimum variance quadratic unbiased estimation (MIVQUE) (10), which has an extension to REML. The objective of this report is to present and evaluate the expected bias in variance estimates when data are subjected to the selection of dividing data into herd production groups.

THE LINEAR MODEL

For the purpose of identifying and quantifying the bias of selection on herd means let us consider the data to conform to the following mixed linear model [after (1)] :

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_3 \end{bmatrix} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{Z}_3 \end{bmatrix} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix} \quad [1]$$

where $\mathbf{y}_i$ is a vector of n_i records made at herd production i (i = 1, 2, 3), $\mathbf{X}_i$ is a known incidence matrix relating records to herd-year-season effects for i, $\mathbf{h}_i$ is a vector of fixed herd-year-season effects for i, $\mathbf{Z}_i$ is a known incidence matrix relating records to sires for i (with null columns corresponding to sires without daughters in i), $\mathbf{u}_i$ is a random vector of sire transmitting abilities for i, and $\mathbf{e}_i$ is a vector of random residuals associated with records in i.

In a no selection (unconditional) model:

$$E[\mathbf{y}_i] = \mathbf{X}_i\mathbf{h}_i$$
$$E[\mathbf{u}_i] = \mathbf{0}$$
$$E[\mathbf{e}_i] = \mathbf{0}$$
$$\text{cov}(\mathbf{u}, \mathbf{e}') = \mathbf{0}$$

$$\mathbf{G} = \text{var}\begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \mathbf{u}_3 \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ u_{12} & u_{22} & u_{23} \\ u_{13} & u_{23} & u_{33} \end{bmatrix} * \mathbf{A}$$
$$= \mathbf{U} * \mathbf{A}$$

where $\mathbf{U}$ is a symmetric matrix of order three representing variances and covariances among sire transmitting abilities of the three production groups and $\mathbf{A}$ is the numerator relationship matrix among sires. The asterisk represents a direct (Kronecker) product. Moreover:

$$\text{var}\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{I}e_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}e_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}e_{33} \end{bmatrix}$$

where e_{ii} represents the residual variance common to all records in production i and the identity matrices are each of appropriate order (i.e., n_i).

Although not broadly applicable, model [1] is sufficiently general to illustrate how the selection imposed by stratifying data on herd production can bias estimates of sire and residual variances. To develop these results we must consider a selection model developed by Henderson (5). As Schaeffer (14) mentions, there has been a reluctance on the part of many animal geneticists to completely accept Henderson's (5) formulation of selection. Yet, for the purpose at hand, this formulation of the effects of nonrandom sampling on the means and variance-covariance structure of $\mathbf{y}$ is adequate for gaining insight in variance estimation problems. The notation for the selection (conditional) model is identical to that presented by Henderson (5).

Consider an additional unselected random vector variable $\mathbf{w}$ with $E[\mathbf{w}] = \mathbf{d}$, $\text{var}(\mathbf{w}) = \mathbf{H}$ and:

$$\text{var}\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{V} & \mathbf{ZG} & \mathbf{B} \\ \mathbf{GZ}' & \mathbf{G} & \mathbf{B}_u \\ \mathbf{B}' & \mathbf{B}_u' & \mathbf{H} \end{bmatrix} \quad [2]$$

where $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$ and $\mathbf{Z}$ is the entire incidence matrix for sires in model [1]. At this point it is also necessary to assume that $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{w}$ are jointly normally distributed.

The essence of the selection model is to suppose that there has been some sort of selection on $\mathbf{w}$ such that $E[\mathbf{w}] = \mathbf{s} \neq \mathbf{d}$ and $\text{var}(\mathbf{w}) = \mathbf{H}_s \neq \mathbf{H}$. Applying the result of Pearson (9), as was done by Henderson (5), to the parameters of the mixed model [1] we obtain:

$$E\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{Xh} + \mathbf{Bt} \\ \mathbf{B}_u\mathbf{t} \\ \mathbf{s} \end{bmatrix} \quad [3]$$

where $t = H^{-1}(s - d)$, and:

$$\text{var}\begin{bmatrix} y \\ u \\ w \end{bmatrix} = \begin{bmatrix} V - BH_0B' & ZG - BH_0B_u' & BH^{-1}H_S \\ & G - B_uH_0B_u' & B_uH^{-1}H_S \\ \text{symmetric} & & H_S \end{bmatrix} \quad [4]$$

where $H_0 = H^{-1}(H - H_S)H^{-1}$.

Henderson (5) then derives BLUE and predictors with the "after selection" parameters. In particular, he discusses several forms for w and their possible application. The vector w is considered the conditional variable, the value upon which selection acts. For our specific problem of stratifying herds based on average milk production, w is a vector of herd means. Accordingly:

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} (P_1'X_1'X_1P_1)^{-1}P_1'X_1'y_1 \\ (P_2'X_2'X_2P_2)^{-1}P_2'X_2'y_2 \\ (P_3'X_3'X_3P_3)^{-1}P_3'X_3'y_3 \end{bmatrix} = \begin{bmatrix} Q_1' & 0 & 0 \\ 0 & Q_2' & 0 \\ 0 & 0 & Q_3' \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \quad [5]$$

where P_i is a matrix of appropriate order which identifies records across herd-year-seasons by the herd in which they are recorded. Having computed these herd means, it is upon these values that the records are stratified and upon which selection is made.

Under the definition for w in [5] $\text{cov}(y, w') = B = VQ$ and $H = Q'VQ$. With these parameters defined, we consider next the expected value of Method 3 estimators for this model of selection.

METHOD 3

One of the most widely used and discussed techniques of variance estimation is Henderson's (4) Method 3. Based on reductions in sums of squares (fitting constants), Method 3 estimators are unbiased and translation invariant (in unconditional models). When applied to data that have been subjected to selection, most investigations have demonstrated a significant selection bias on estimates of variance components (11, 12). However, selection practiced in these studies has been quite different from the herd mean selection examined herein.

To study the effect of herd mean selection used to stratify herds into production groups we consider the methods and models used by Mirande and Van Vleck (8) as an example. Other studies have employed a similar approach. Based on herd mean milk production, records are classified into one of three groups (low, medium, or high). Each group is then treated as an "independent" data set for the purpose of variance estimation. Thus, y_1, y_2, and y_3 of model [1] may be considered separately. For example, let us consider the model:

$$y_1 = X_1h_1 + Z_1u_1 + e_1 \quad [6]$$

with parameters as defined in [1], as our model for Method 3 estimation of u_{11} and e_{11}.

Method 3 quadratic forms for estimating s_{11} and e_{11} are:

$$y_1'(I - W(W'W)^{-}W')y_1 = SSE_1 \quad [7]$$

$$y_1'(M_1Z_1(Z_1'M_1Z_1)^{-}Z_1'M_1)y_1 = R(u_1|h_1)$$

where $W = [X_1 \quad Z_1]$ and $M_1 = I - X_1(X_1'X_1)^{-}X_1'$. The SSE_1 is the error sum of squares and $R(u_1|h_1)$ is the reduction in sums of squares due to fitting sires "after" herds. In the unconditional model (usual, no selection):

$$E[SSE_1] = \text{trace}[I - W(W'W)^{-}W']e_{11} \quad [8]$$

$$E[R(u_1|h_1)] = \text{trace}[Z_1M_1Z_1A]u_{11} + \text{trace}[(Z_1'M_1Z_1)(Z_1'M_1Z_1)^{-}]e_{11}$$

In the herd mean selection model, where selection is defined by $w_1 = Q_1'y_1$ in line [5],

$\text{var}_u(y_1) = V_s = V - BH_0B'$ as in line [4] for $B = VQ_1$ and $E_s[y_1] = X_1h_1 + Bt$. Accordingly, in the selection model:

$$E_s[SSE_1] = \text{trace}[(I - W(W'W)^{-}W') \times (V - BH_0B')] + h_1'X_1'(I - W(W'W)^{-}W')X_1h_1 + t'B'(I - W(W'W)^{-}W')Bt \quad [9]$$

The fortunate result is that $W(W'W)^{-}W'B = B$ (see Appendix), and thus:

$$E_s[SSE_1] = \text{trace}(I - W(W'W)^{-}W')e_{11}$$

the identical result for the unconditional model.

The quadratic form for sires has the following expected value, in the selection model:

$$E_s[R(u_1|h_1)] = E_s[y_1'(M_1Z_1(Z_1'M_1Z_1)^{-}Z_1'M_1)y_1] = \text{trace}[M_1Z_1(Z_1'M_1Z_1)^{-}Z_1'M_1(V - BH_0B')] + h_1'X_1'(M_1Z_1(Z_1'M_1Z_1)^{-}Z_1'M_1)X_1h_1 + t'B'(M_1Z_1(Z_1'M_1Z_1)^{-}Z_1'M_1)Bt \quad [10]$$

Given that $M_1X_1 = 0$ and $M_1Q_1 = 0$ as well, [10] can be simplified to:

$$E_s[R(u_1|h_1)] = \text{trace}[Z_1'M_1Z_1A]s_{11} + \text{trace}[Z_1'M_1Z_1(Z_1'M_1Z_1)^{-}]e_{11} - \text{trace}[Q_1'Z_1A_1Z_1'M_1Z_1A_1Z_1'Q_1H_0]u_{11}^2 + t'Q_1'Z_1A_1Z_1'M_1Z_1A_1Z_1'Q_1tu_{11}^2 \quad [11]$$

The first two terms of model [11] are identical to those for the unconditional model (line [8]). However, the second two terms, which involve t and H_0, will not reduce to zero. Accordingly, this form of selection will bias estimates of the sire variance. The only means of removing this bias is to evaluate H_S and t and include these values in the computations. However, such a scheme is unworkable.

DISCUSSION

Although a great deal of evidence has been collected to support the notion of unequal variances across dairy herds, the present report indicates that much of this evidence should be treated cautiously. With the exception of Everett et al. (2), most investigations into heterogeneity of herd variances have relied upon stratifying the data into production groups based on herd milk averages. Estimates of variance components for milk production are then made within these herd production groups. Simply put, the data are selected on functions of the random variables from which we intend to estimate variances.

This type of selection problem is quite different from that usually considered by animal breeders. Typically, we examine selection problems where some part of the data is a random sample from an unselected population. For example, in the estimation of genetic variances for first and second lactations, we assume that the opportunity to have a second record is based on the performance of the first record. For this kind of selection, estimates of variances can be found unbiased by selection if we analyze the data jointly (13, 15). Considering the first records as an unselected sample, any selection practiced to obtain second records can be accounted for by the inclusion of the first, unselected, records.

For the problem considered here, there is no such section of the data that has been unselected. If the data have been stratified by herd production, all the data have been subject to selection and, thus, the opportunity for some part of the data to "correct" for selection is not possible.

The most interesting aspect of the results established herein is the ability of Method 3 to provide an estimate of the residual variance unbiased by selection. In other more typical selection models, the failure of the Henderson methods to provide estimates unbiased by selection has been well established. Rothschild et al. (13) were among the first to present maximum likelihood-based methods as the solution to estimating variances from data subjected to selection. However, for the problem discussed herein of $Q'y$ selection, REML is not capable of providing estimates free of selection bias. This can be evaluated by using the quadratics and expected values under the selection model as presented by Schaeffer (14). Method 3, being based on the least squares equations, provides for some very simple matrix identities under $Q'y$ selection which REML and the mixed

TABLE 1. Mean estimates of sire and residual variances from simulated data divided into four production groups.

	All data	Herd average group			
		Low	Medium low	Medium high	High
Sire variance					
Mean	119,710	127,392	125,257	115,453	102,917
Minimum	92,616	65,054	82,310	58,309	56,894
Maximum	154,410	178,024	143,367	167,228	145,189
Residual variance					
Mean	1,835,129	1,827,524	1,847,240	1,849,835	1,827,989
Minimum	1,806,984	1,762,436	1,800,802	1,760,732	1,740,202
Maximum	1,860,407	1,958,919	1,903,726	1,923,528	1,953,379
Mean h^2	.24	.26	.25	.23	.21
Percent bias of sire variance[1]	−2.3	4.0	2.3	−5.8	−16.0
Percent bias of residual variance[2]	−.1	−.5	.5	.7	.5

[1] Percent bias of sire variance = [(estimated sire variance − 122,500)/122,500] × 100.

[2] Percent bias of residual variance = [(estimated residual variance − 1,873,499.54)/1,837,499.54] × 100.

model equations cannot. For example, one can show that:

$$\text{trace}(\mathbf{VQH_0Q'V}) = \text{trace}(\mathbf{W(W'W)^{-}W'VQH_0Q'V}) \neq \text{trace}(\mathbf{WCW'VQH_0Q'V})$$

for $\mathbf{C}$ defined as the generalized inverse of the coefficient matrix of the mixed model equations. The $\mathbf{WCW'}$ is part of the expected value of the quadratic form for the estimation of the residual value of the quadratic form for the estimation of the residual variance with REML. Such equalities permit the residual variance to be estimated without any bias under Method 3 but lead to bias in estimates from mixed model based methods.

Evaluating the magnitude of the bias in estimating the sire variance remains problematic. This bias, presented in equation [11], contains the matrix $\mathbf{H_0}$ and the vector $\mathbf{t}$, both of which are difficult to represent in any general fashion.

The results of a small, yet representative, simulation are presented in Table 1. The design of the simulation is to create data sets similar in structure to that of Mirande and Van Vleck (8). The objective is to simulate data with a homogeneous variance across herds. Accordingly, the data are divided into strata by herd means. Variances are subsequently estimated from the herd production groups to examine how the segmentatio of the data into groups may bias estimates of the variances.

There are 1800 herd-year-season effects from 150 herds and 150 unrelated sires represented in each of 10 simulated data sets. Assigning simulated daughter records to sires and herd-year-seasons is done at random with probabilities set so as to create an average of 100 progeny per sire (i.e., the probability that a record would appear in a given herd-year-season by sire subclass is .056). Records are generated with pseudorandom normal deviates, including a fixed herd-year-season effect, a random sire component (with null mean and variance of 122500 kg^2), and random residual (with null mean and variance 1837499.54 kg^2). Thus, all records are generated with homogeneous variances, corresponding to a heritability of .25.

With 10 data sets generated, each is divided into four groups based on mean herd production. Cutpoints for the four groups are 5776, 7000, and 8224 kg. The model used for variance estimation is a two-way mixed model without interaction, with fixed herd-year-season effects and random, unrelated sires. The method of estimation presented is Method 3.

Table 1 provides a brief summary of the results of this simulation. As expected, estimates of the residual variance are unaffected (a maximum of .7% relative bias) by the division of data into herd production groups. However, the sire variance can be quite substantially influenced by the stratification of production data. Estimates of heritability are affected only slightly, on average. Nevertheless, estimates of genetic variance across herd production groups should be evaluated with caution.

To summarize, stratifying milk records by herd production may be a simple strategy for the detection of heterogeneous variance. However, interpretation of such results is not straightforward. Fortunately, Method 3 provides an unbiased estimate of the residual variance in stratified data. Yet estimates of the sire variance can be substantially influenced by stratification on herd mean. Other methods of variance estimation are also biased by this form of selection. Detection of heterogeneous variance will be best done on a herd by herd basis or some other strategy free of selection.

REFERENCES

1 De Veer, J. C., and L. D. Van Vleck. 1987. Genetic parameters for first lactation milk yields at three levels of herd production. J. Dairy Sci. 70:1434.

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

3 Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogenous. Theor. Appl. Genet. 72:671.

4 Henderson, C. R. 1953. Estimation of variance and covariance components. Biometrics 9:226.

5 Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423.

6 Hill, W. G., M. R. Edwards, M.K.A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition of different levels and variability of production. Anim. Prod. 36:59.

7 Lofgren, D. L. W. E. Vinson, R. E. Pearson, and R. L. Powell. 1985. Heritability of milk yield at different levels of herd mean and standard deviation for production. J. Dairy Sci. 68:2737.

8 Mirande, S. L., and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.
9 Pearson, K. 1903. Mathematical contributions to the theory of evolution.–XI. On the influence of natural selection on the variability and correlation of organs. Phil. Trans. Royal. Soc. Lond. A 200:1.
10 Rao, C. R. 1971. Minimum variance quadratic unbiased estimation of variance components. J. Multivar. Anal. 1:445.
11 Reeve, E.C.R. 1953. Studies in quantitative inheritance. III. Heritability and genetic correlation in progeny tests using different mating systems. J. Genet. 51:520.
12 Robertson, A. 1977. The effect of selection on the estimation of genetic parameters. Z. Tierz. Züchtungsbiol. 94:131.
13 Rothschild, M. F., C. R. Henderson, and R. L. Quaas. 1979. Effects of selection on variances and covariances of simulated first and second lactations. J. Dairy Sci. 62:996.
14 Schaeffer, L. R. 1987. Estimation of variance components under a selection model. J. Dairy Sci. 70: 661.
15 Sorenson, D. A., and B. W. Kennedy. 1984. Estimation of genetic variances from unselected and selected populations. J. Anim. Sci. 59:1213.
16 Van Vleck, L. D., and G. E. Bradford. 1964. Heritability of milk yield at different environmental levels. Anim. Prod. 6:285.

APPENDIX

Prove $\mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{B} = \mathbf{B}$.

As outlined in the text $\mathbf{W} = (\mathbf{X}_1 \quad \mathbf{Z}_1)$ and $\mathbf{B} = \mathbf{V}\mathbf{Q}_1$, where $\mathbf{V} = \mathbf{Z}_1\mathbf{A}\mathbf{Z}_1 u_{11} + \mathbf{I}e_{11}$ and $\mathbf{Q}_1' = (\mathbf{P}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{P}_1)^{-1}\mathbf{P}_1'\mathbf{X}_1'$. Recall that for generalized inverses $\mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{W} = \mathbf{W}$ and thus $\mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{X}_1 = \mathbf{X}_1$ and $\mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{Z}_1 = \mathbf{Z}_1$. Accordingly,

$$\begin{aligned}
\mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'\mathbf{B} &= \mathbf{W}(\mathbf{W}'\mathbf{W})^{-}\mathbf{W}'(\mathbf{Z}\mathbf{A}\mathbf{Z}'u_{11} + \mathbf{I}e_{11})\mathbf{Q} \\
&= \mathbf{Z}\mathbf{A}\mathbf{Z}'u_{11} + \mathbf{W}(\mathbf{W}'\mathbf{W})\mathbf{W}'\mathbf{X}_1\mathbf{P}_1(\mathbf{P}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{P}_1)^{-1}e_{11} \\
&= \mathbf{Z}\mathbf{A}\mathbf{Z}'u_{11} + \mathbf{X}_1\mathbf{P}_1(\mathbf{P}_1'\mathbf{X}_1'\mathbf{X}_1\mathbf{P}_1)^{-1}e_{11} \\
&= (\mathbf{Z}\mathbf{A}\mathbf{Z}'u_{11} + \mathbf{I}e_{11})\mathbf{Q} \\
&= \mathbf{V}\mathbf{Q} = \mathbf{B}
\end{aligned}$$

the desired result.