January 1979

An Approximate Procedure for Determining Prediction Error Variances of Sire Evaluations

G. R. Ufford
Cornell University

C. R. Henderson
Cornell University

L. Dale Van Vleck
University of Nebraska-Lincoln, dvan-vleck1@unl.edu

An Approximate Procedure for Determining Prediction Error Variances of Sire Evaluations

G. R. UFFORD[1], C. R. HENDERSON, and L. D. VAN VLECK
Department of Animal Science
Cornell University
Ithaca, NY 14853

ABSTRACT

Prediction errors of sire evaluations can be obtained directly from the inverse of the appropriate coefficient matrix. Considerably more effort is required to obtain the inverse in practical situations than can be justified for publication of a confidence figure. An approximate prediction error variance, k/(n + 20), is used currently in the Northeast Artificial Insemination Sire Comparison where n is the number of daughters and k is an appropriate breed constant corresponding to the residual variance. This procedure, however, does not account for distribution of sires across herds nor several lactations per daughter. Thus, the diagonal elements of the sire equations after absorption of cow, sire-by-herd, natural service sire, and herd-year-season equations were chosen as likely indicators of the prediction error variance for this more complicated model. Simple regression was used to relate prediction error variance obtained from the inverse to the diagonal after absorption. The coefficient of determination was .995 or greater in all cases. A single approximate prediction error variance of sire evaluation (group plus sire solution) could be used for Ayrshire, Guernsey, Jersey, and Brown Swiss bulls (and probably for Holsteins, which were not studied). The approximate prediction error variance is [−.0014 + 1.08/diagonal] times the appropriate residual variance. An approximation comparable to repeatability for herdmate comparisons also was derived as [1.01 − 9/diagonal].

Received September 25, 1978.
[1] 34516 Tennessee Road, Lebanon, OR 97355.

INTRODUCTION

Herdmate sire evaluations have been published with a figure called "repeatability." Repeatability is the square of an approximation derived from selection index procedures of the correlation of the index with the true value it predicts (sometimes given symbols r_{TI}, $r_{G\hat{G}}$, or r_{HI}). If the assumptions for selection index are true, the correlation has a one-to-one correspondence with prediction error variance (1).

In mixed model prediction, predictands may be sums of fixed and random effects, and the correlation is not as appropriate to indicate accuracy as for selection indexes where all effects are random. In this case, prediction error variance is a more informative indicator of accuracy or reliability of sire evaluation (4). Prediction error variances are usually difficult to calculate directly, requiring inversion of a large matrix. In special cases with small numbers of herds the inverse can be obtained relatively easily (6). Thus, when in practice the number of herd-year-seasons is large, it often becomes desirable to determine a procedure which will give a close approximation with much less computing cost.

An approximate empirical procedure derived by Henderson (2) is used in the Northeast AI Sire Comparison (NEAISC). This procedure approximates the prediction error variance by k/(n + 20), where n is the number of daughters of a sire, and k is a different constant for each breed related to σ_e^2. The constants (kg^2) for Ayrshires, Guernseys, Holsteins, Jerseys, and Brown Swiss are 1,183,171, 1,131,729, 1,831,343, 874,518, and 1,440,383. This procedure is probably as accurate as needed for the NEAISC since only AI (artificial insemination) first-lactation records are used, but it does not consider the distribution of sires across herds and is not appropriate when all lactations are used.

Once the prediction error variance is deter-

mined, it can be converted to a measure analogous to repeatability of herdmate comparisons for publication. The purpose of this study was to derive an approximate method to calculate prediction error variances and repeatabilities when all lactation records are used with a best linear unbiased prediction (BLUP) procedure.

METHODS

The model with sire-by-herd interaction was:

$$y_{fijklm} = g_i + s_{fij} + h_{kl} + sh_{fijk} + c_{fijkm} + e_{fijklm} \quad [1]$$

where:

- y is an age-season adjusted record,
- g is a fixed sire group effect,
- s is a random sire-within-group effect with $\sigma_e^2/\sigma_s^2 = 8.33$ (f indicates whether the sire is AI or natural service [NS]),
- h is a fixed herd-year-season (HYS) effect (i indicates the year-season within herd k),
- sh is a random sire-by-herd interaction effect (accounting for environmental correlation among paternal half-sisters) with $\sigma_e^2/\sigma_{sh}^2 = 3.57$,
- c is a random cow-within-sire effect with $\sigma_e^2/\sigma_c^2 = 1.67$, and
- e is a random residual effect with variance σ_e^2.

The variance ratios 8.33, 3.57, and 1.67 were suggested by Norman (5) because they correspond to h^2 of .24, c^2 of .14, and cow repeatability of .50.

Data from the New York Dairy Record Processing Laboratory were used. Up to ten lactations were accepted on cows that had a first lactation record in the same herd. The data set included records normally used for AI sire evaluation plus records on daughters of NS sires. AI and NS sires were assigned to separate groups. Records of Ayrshires, Guernseys, Jerseys, and Brown Swiss were studied.

Solution of BLUP Equations

The BLUP solutions for fixed and random effects were obtained by solving the mixed model equations (4):

$$\begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{X} & \underset{\sim}{X}'\underset{\sim}{Z} \\ \underset{\sim}{Z}'\underset{\sim}{X} & \underset{\sim}{Z}'\underset{\sim}{Z} + \underset{\sim}{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\underset{\sim}{\beta}} \\ \hat{\underset{\sim}{u}} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{y} \\ \underset{\sim}{Z}'\underset{\sim}{y} \end{bmatrix} \quad [2]$$

where:

- $\underset{\sim}{y}$ is the vector of observed records,
- $\hat{\underset{\sim}{\beta}}$ is the vector of unknown fixed effects for groups and herd-year-seasons,
- $\hat{\underset{\sim}{u}}$ is a vector of unknown random effects for sires, sires-by-herds, and cows,
- $\underset{\sim}{X}$ and $\underset{\sim}{Z}$ are design matrices describing the contribution of fixed and random effects, respectively, to the records, and
- $\underset{\sim}{G}^{-1}$ is the inverse of the variance-covariance matrix of random effects multipled by σ_e^2 so that the diagonal elements are σ_e^2/σ_s^2, σ_e^2/σ_{sh}^2, and σ_e^2/σ_c^2 corresponding to the equations for s, sh, and c.

The group and sire values are the elements of interest. The sire evaluation is defined as $\hat{g}_i + \hat{s}_{ij}$.

Equations for cows, sire-by-herds, herd-year-seasons, and NS sires were absorbed with both cows and NS sires treated as though they were nested within a herd. After absorption, only group and AI sire equations remained. Because the herd-year-season equations sum to the sum of the group equations, the group equations sum to zero after the herd-year-season equations are absorbed. A Lagrange equation was added to obtain a solution and at the same time to define a base that made the solutions comparable to the NEAISC evaluations. The Lagrange equation in this study was used to tie the solutions to the same 1968 base used in the NEAISC by forcing the sum of all sires with NEAISC evaluations (one or more daughters) to be the same in this study as in the NEAISC evaluations. The diagonal of the Lagrange equation was zero. Sire coefficients were one if the sire had a NEAISC evaluation and zero otherwise. Group coefficients were the number of sires in that group with a previous evaluation. In the Lagrange equation, sire elements within a group sum to the corresponding group elements. The right-hand side for the Lagrange equation is the sum of the previous NEAISC evaluations of the sires.

After absorption, let the group, Lagrange, and AI sire equations be:

$$\begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B} \\ \underset{\sim}{B}' & \underset{\sim}{D} \end{bmatrix} \begin{bmatrix} \hat{\underset{\sim}{g}} \\ \hat{\underset{\sim}{s}} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{r}_1 \\ \underset{\sim}{r}_2 \end{bmatrix} \quad [3]$$

where $\hat{\underset{\sim}{g}}$ is a vector containing the group solutions and Lagrange multiplier and $\hat{\underset{\sim}{s}}$ is the vector of sire solutions.

Group and sire solutions were obtained by modified Gauss-Siedel iteration similar to the method used in the Northeast (4, 7). The notation $\hat{\underset{\sim}{g}}^{(k)}$ and $\hat{\underset{\sim}{s}}^{(k)}$ represents the group and sire solutions at the end of the kth round of iteration. To simplify the description of the iterative procedure, let p substitute for the subscripts ij referring to the jth sire in the ith group. Thus, $d_{pp'}$ is the element of $\underset{\sim}{D}$ corresponding to the pth and the p'th sires; b_{ip} is the element of $\underset{\sim}{B}$ corresponding to the ith group and pth sire; and $\underset{\sim}{r}_{2p}$ is the element of $\underset{\sim}{r}_2$ corresponding to the pth sire. Likewise, $\hat{s}_p^{(k)}$ is the pth element of $\hat{\underset{\sim}{s}}^{(k)}$ and $\hat{g}_i^{(k)}$ is the ith element of $\hat{\underset{\sim}{g}}^{(k)}$. Assume that there are n_i sires in each group for a total of n sires.

The first-round starting points were:

$$\hat{\underset{\sim}{g}}^{(0)} = \underset{\sim}{A}^{-1}\underset{\sim}{r}_1$$

$$\hat{\underset{\sim}{s}}^{(0)} = \underset{\sim}{0}$$

Then for the kth round the revised solution for the pth sire is

$$\hat{s}_p^{(k)} = \frac{1}{d_{pp}}\left(r_{2p} - \sum_i b_{ip}\hat{g}_i^{(k-1)} - \sum_{p'=1}^{p-1} d_{pp'}\hat{s}_{p'}^{(k)} - \sum_{p'=p+1}^{n} d_{pp'}\hat{s}_{p'}^{(k-1)}\right) \quad [4]$$

This just says that the solution for the pth sire is obtained from the pth equation by substituting in the current values for the sire and group solutions except for the pth sire, then solving for that sire as the only remaining unknown in the equation. As the sire solutions are calculated, it is convenient to retain the row of $\underset{\sim}{B}'$ in storage, then when the new $\hat{s}_p^{(k)}$ is calculated, $b_{ip}\hat{s}_p^{(k)}$ can be substracted from the corresponding groups, as will be described in equation [5].

After all the sire solutions are updated, they can be adjusted to sum to zero within a group. When $\underset{\sim}{G}$ is of diagonal form, i.e. sires are assumed uncorrelated, the sire solutions will sum to zero within a group (4). To invoke this property, AI and NS sires must be in separate groups. More than 200 rounds of iteration were required to achieve reasonable convergence with no adjustment, and less than 15 rounds were required with the adjustment to force sires to sum to zero.

Sire solutions are adjusted to sum to zero as follows: calculate

$$a_i = \frac{1}{n_i}\sum_j \hat{s}_{ij}^{(k)}$$

Then,

$$\text{adjusted } \hat{s}_{ij}^{(k)} = \hat{s}_{ij}^{(k)} - a_i$$

Since the group right-hand sides have had the unadjusted sire contribution subtracted out as

$$\underset{\sim}{r}_1^* = \underset{\sim}{r}_1 - \hat{\underset{\sim}{B}}\hat{\underset{\sim}{s}}^{(k)} \quad [5]$$

where $\hat{\underset{\sim}{s}}^{(k)}$ was unadjusted, the next group solution must be calculated and adjusted upward correspondingly.

$$\hat{\underset{\sim}{g}}^{(k)} = \underset{\sim}{A}^{-1}\underset{\sim}{r}_1^*$$

$$\text{adjusted } \hat{\underset{\sim}{g}}^{(k)} = \hat{\underset{\sim}{g}}^{(k)} + \underset{\sim}{a} \quad [6]$$

where $\underset{\sim}{a}$ is a vector with elements a_i corresponding to the AI groups and zeroes corresponding to the NS sire groups and Lagrange multiplier. While NS sires also sum to zero within a group, no adjustment is made to NS groups because NS sire equations have been absorbed and are not involved in the iterative solutions.

Variance of Prediction Errors

Prediction error variance is obtained from the inverse of the coefficient matrix of [3] after absorption (4).

$$\text{Var}\begin{bmatrix} \hat{\underset{\sim}{g}} \\ \hat{\underset{\sim}{s}}-\underset{\sim}{s} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B} \\ \underset{\sim}{B}' & \underset{\sim}{D} \end{bmatrix}^{-1} \sigma_e^2 = \begin{bmatrix} \underset{\sim}{C}_{11} & \underset{\sim}{C}_{12} \\ \underset{\sim}{C}_{12}' & \underset{\sim}{C}_{22} \end{bmatrix} \sigma_e^2 \quad [7]$$

The inverse in this case may be difficult or impossible to calculate by direct inversion techniques because of limitations on core storage. Therefore an iterative procedure similar to that for obtaining solutions for milk evaluations was used to obtain the inverse. By using columns of the identity matrix for right-hand sides in [3], corresponding columns of the inverse can be found by iteration. The procedure was to solve simultaneously for all the group columns of the inverse by solving the equations

$$\begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B} \\ \underset{\sim}{B}' & \underset{\sim}{D} \end{bmatrix} \quad \begin{bmatrix} \underset{\sim}{C}_{11} \\ \underset{\sim}{C}'_{12} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{I} \\ \underset{\sim}{O} \end{bmatrix} \quad [8]$$

To be feasible, the convergence of these solutions was speeded up as were solutions for groups and sires by imposing knowledge of properties of the final solutions at the end of each round of iteration (3). In this model with no covariances among the sires, the sire-by-group elements will sum to zero in all groups except the group represented by the columns, i.e. the group with the nonzero right-hand side. Sires in this group sum to $-\sigma_s^2/\sigma_e^2$. Sire and group solutions for inverse elements are adjusted in the same manner as described for milk and fat solutions.

Next the sire-by-sire elements could be determined by solving

$$\begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B} \\ \underset{\sim}{B}' & \underset{\sim}{D} \end{bmatrix} \quad \begin{bmatrix} \underset{\sim}{C}_{12} \\ \underset{\sim}{C}_{22} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{O} \\ \underset{\sim}{I} \end{bmatrix}$$

Since $\underset{\sim}{C}_{12}$ has been determined by solving [8], $\underset{\sim}{C}_{22}$ can be determined more readily by solving

$$\underset{\sim}{D}\underset{\sim}{C}_{22} = \underset{\sim}{I} - \underset{\sim}{B}'\underset{\sim}{C}_{12} \quad [9]$$

Convergence is speeded up by imposing the property that the sire-by-sire elements in a column sum to zero within all groups with one exception; that is, the elements for sires in the group of the sire whose inverse column is being sought sum to σ_s^2/σ_e^2. The adjustment is made by simply adjusting elements of $\underset{\sim}{C}_{22}$ up or down so that averages within each group are zero or σ_s^2/σ_e^2.

RESULTS AND DISCUSSION

Standard errors of sire evaluations can be obtained directly from the inverse of the coefficient matrix (4). Whether the inverse is calculated directly or by iteration, considerably more effort is required than can be justified for publication of a confidence figure. Approximately the same effort is required to obtain one row of the inverse as is required to obtain evaluations for milk or fat once the coefficient matrix has been formed.

The diagonal element of each sire equation in the coefficient matrix after absorption of cow, sire-by-herd, NS sire, and herd-year-season equations was chosen as a likely candidate for a good predictor of error variance. The elements referred to are the diagonals of $\underset{\sim}{D}$ in equation [3]. This is the mixed model equation with variance ratio, σ_e^2/σ_s^2, added. A simple regression was fit to the inverses of the diagonal elements. Each sire was considered a separate observation. Dependent variables were the sire prediction error variance $V(\hat{s} - s)$ for the regressions shown in Table 1 and evaluation error variance $V(\hat{g} + \hat{s} - s)$ for the regressions in Table 2. The coefficient of determination, R^2, was greater than .995 in all cases. Both the sire effect and evaluation error variances appear to be predicted with equal accuracy. Regressions for prediction of the evaluation error variances are more consistent from breed to breed than regressions for prediction of sire error variances. The following single equation could be used for all breeds (probably Holsteins included) to approximate the evaluation error variance.

$$V(\hat{g} + \hat{s} - s) \approx [-.0014 + 1.08(1/\text{diagonal})]\,\sigma_e^2 \quad [10]$$

While "repeatability" of sire evaluation is not appropriate or even discussed in connection with mixed model sire evaluation, it may be useful to provide a measure of accuracy comparable to the repeatability obtained from evaluations using some form of herdmate comparison. If selection within a sire group is of interest or if the mixed model procedure is applied to a model without genetic groups, the

TABLE 1. Regression of sire prediction error variance, $V(\hat{s} - s)$, on inverse of sire equation diagonal element ($\sigma_e = 1$).

Breed	R^2	Intercept		Slope	
		—(β_0 ± SE) —		—(β_1 ± SE)—	
Ayrshire	.9963	.00945	.00031	.909	.005
Guernsey	.9977	.00552	.00016	.947	.002
Jersey	.9971	.00524	.00019	.950	.003
Brown Swiss	.9975	.00797	.00031	.923	.004

following definition of repeatability may be appropriate:

$$R_1^2 = \left[\frac{\text{cov}(\hat{s},s)}{\hat{\sigma}_s \sigma_s}\right]^2 \qquad [11]$$

Using the BLUP properties $V(\hat{s}) = \text{cov}(\hat{s},s) = V(s) - V(\hat{s} - s)$ given by Henderson (4),

$$R_1^2 = 1 - \frac{V(\hat{s}-s)}{V(s)} \qquad [12]$$

Letting the diagonal element of C_{22} be $c_{ij,ij}$ for the ijth sire,

$$R_1^2 = 1 - c_{ij,ij}\sigma_e^2/\sigma_s^2 \qquad [13]$$

and for the model used in this study where $\sigma_e^2/\sigma_s^2 = 8.33$,

$$R_1^2 = 1 - 8.33c_{ij,ij} \qquad [14]$$

When genetic groups are included in the model and selection is among sires in different groups, it may be appropriate to depart from the strict definition of a correlation to define the "*repeatability*" *in terms of the error variance* of the evaluation $V(\hat{g} + \hat{s} - s)$ rather than in terms of the error variance of the sire prediction $V(\hat{s} - s)$, i.e.

$$R_2^2 = 1 - \frac{V(\hat{g} + \hat{s} - s)}{V(s)} \qquad [15]$$

Letting $c_{i,i}$ be the diagonal element of C_{11} and $c_{i,ij}$ be the element of C_{12} appropriate for $\text{cov}(\hat{g}, \hat{s} - s)$,

$$R_2^2 = 1 - (c_{i,i} + c_{ij,ij} + 2c_{i,ij})\sigma_e^2/\sigma_s^2 \qquad [16]$$

$$= 1 - 8.33(c_{i,i} + c_{ij,ij} + 2c_{i,ij}) \qquad [17]$$

with the variance ratio 8.33 used in this study. The latter definition, [15], would seem to be most appropriate for general publication of sire evaluations and using the regression derived in this study, it could be approximated by

$$R_2^2 \approx 1 - 8.33\ [-.0014 + 1.08(1/\text{diagonal})] \qquad [18]$$
$$\approx 1.01 - 9(1/\text{diagonal})$$

The lower and upper limits of this approximation are −.0704 and 1.01.

Why should the diagonal serve as a good predictor of evaluation error variances? The diagonal of the sire equation prior to absorption of any effects is just the number of total lactation records plus the variance ratio. During absorption, this element is modified to reflect the number of records per cow, distribution across herds and herd-year-seasons, and association with NS sires. Absorption of NS sires probably has a detrimental effect on accuracy of predicting the error variance and may be the reason why the intercept is not zero and the slope is not one. Although this is just speculation, it should be investigated if NS sires are not absorbed.

Simple regressions as in equations [10] and [18] can be used to approximate the evaluation error variances or repeatability with more accuracy than is needed for published confidence ranges. However, they should not be used with evaluations under different models without further investigation since the regression equations were derived from the all-lactation procedure. This caution includes evaluations with only first lactation records. If this pro-

TABLE 2. Regression of sire evaluation error variance, $V(\hat{g} + \hat{s} - s)$, on inverse of sire equation diagonal element ($\sigma_e = 1$).

Breed	R^2	Intercept		Slope	
		— (β_0 ± SE)——		—(β_1 ± SE)—	
Ayrshire	.9982	−.00142	.00025	1.088	.004
Guernsey	.9957	−.00139	.00025	1.081	.004
Jersey	.9970	−.00138	.00022	1.078	.003
Brown Swiss	.9958	−.00173	.00047	1.090	.007

cedure is to be used in research as an indication of the accuracy of evaluation, extreme caution should be observed, including tests that the regression is providing good estimates for the situation under consideration.

ACKNOWLEDGMENTS

This research was supported in part by grants from Eastern Artificial Insemination Cooperative, Inc., Ithaca, NY, and the USDA-ARS Animal Improvement Laboratory, Beltsville, MD.

REFERENCES

1 Dickinson, F. N., B. T. McDaniel, H. D. Norman, G. J. King, J. F. Keown, and R. L. Powell. 1974. USDA-DHIA sire summary list (May). ARS-NE-24.

2 Henderson, C. R. 1971. Personal communication. Cornell University, Ithaca, NY.

3 Henderson, C. R. 1972. Some properties of the sire comparison method. Mimeograph, Cornell University, Ithaca, NY.

4 Henderson, C. R. 1973. Sire evaluation and genetic trends. Page 10 *in* Proc. Anim. Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush. ASAS, Champaign, IL.

5 Norman, H. D. 1973. Personal communication. USDA-ARS, Beltsville, MD.

6 Szkotnicki, W. J., A. K. W. Tong, M. A. Sharaby, K. M. Krotch, L. P. Johnson, and L. R. Schaeffer. 1978. Sire and cow evaluation in Brown Swiss, Canadienne, and Milking Shorthorn. J. Dairy Sci. 61:497.

7 Van Norton, R. 1959. The solution of linear equations by the Gauss-Siedel method. Mathematical methods of digital computers. A. Ralston and H. S. Wolf, ed. John Wiley and Sons, Inc., New York.