Journal of Animal and Veterinary Advances 3 (7): 434-441, 2004

Accuracy Comparison on Predicted Variance and Genetic Merit in Different Models with/out Consideration of Relationships in Animal Breeding

H. Mirtaghizadeh
Department of Animal Science, University of Yüzüncü Yil, Van, 65080, Turkey

Abstract: The objective of this article was to determine the effects of several characteristics of data structure on variance of prediction error using both sire and animal models. Data were collected according to mixed model in Ceylanpýnar State Farm of Dairy Cattle's Breeding Unit. Data were included milking records from 1987 to 1999. Data structures were replicated 300 times for each combination of variance and covariance assumption and proportion of occupied subclasses. Data were also evaluated in mixed models appropriate on sire and animal models. Results were comprised correlations between variables and variance of prediction error was obtained by evaluating the model and scheme. Simple curvilinear regression analysis was used to study several design variables in further details and to determine the effects on variance of prediction error. According to results, the sire models yielded a wider range of values for the design variables and in animal model analysis, the number of animals with a progeny test, individual test, and their combination, were 0.34, 0.41 and 0.25, respectively. Progeny tests yielded larger variances of prediction error than did individual performance. Genetic connections were not strongly associated with the variance of prediction error. No single piece of information was useful for predicting accuracy and several other contributions to accuracy are necessary.

Key words: Accuracy, animal model, sire model, BLUP, variance

Introduction

Estimation of co(variances) is the most important part of the breeding strategy. Mixed model methodology was developed by Henderson (1973) for genetic evaluation in large sets of unbalanced data. In animal breeding, data structure is very important for estimating of (co) variances and predicting genetic parameters. However, most of case data structure has infinite possibilities in animal breeding (Tosh and Wilton, 1994).

Basic knowledge on the effects of data structure is derived from selection index theory, which ignores non-genetic contributions and usually considers simple situations (Lush, 1931, 1935; Searle, 1964, Van Vleck et al., 1987). Actual variances of prediction error precisely reflect even the most complex data structure.

If breeders are using animal models for genetic evaluation, it can be assumed that animals have some individuals or sire-dam relationships. If this is the case, animal model is a correct model and animal's relationships must be evaluated for an appropriate assumption. Often, impractical to obtain directly from the inverse of the coefficient matrix, variances of prediction error can be approximated from various pieces of information (Robinson and Jones, 1987; Meyer, 1989). A better understanding of the effects of data structure would aid the improvement of approximation techniques. The objective of this study was to determine the effects of several characteristics of data structure on variance of prediction error using both sire and animal models.

Materials and Methods

Data subclasses of form (i, j), where I indicates the level of a fixed effect and j indicates an animal, were collected according to mixed model in Ceylanpýnar State Farm of Dairy Cattle's breeding records observed from 1987 to 1999. The general form of the model may be written as follows:

$$Y = Xb + Zu + e \quad (1)$$

where, y is an nx1 vector of observations, X is an nxp design matrix relating fixed effects to observations, b is a px1 vector of fixed effects, Z is an nxq design matrix relating to observations, u is a qx1 vector of random animal additive genetic effects, e is an nx1 vector of random residuals and expectations and variance-covariances are observed as:

$$E(y) = Xb,$$

$$E\begin{bmatrix} u \\ e \end{bmatrix} = \begin{vmatrix} 0 \\ 0 \end{vmatrix},$$

and

$$V\begin{vmatrix} u \\ e \end{vmatrix} = \begin{vmatrix} A\sigma_u^2 & 0 \\ 0 & I\sigma_e^2 \end{vmatrix} \quad (2)$$

where A is the additive genetic relationship matrix and I is an identity matrix. There is one fixed factor in b; the levels are referred to as contemporary groups. Design matrices were composed of zeros and ones.
Symmetric generalized inverse of the coefficient matrix of mixed-model equations is

$$\begin{vmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{vmatrix} = \begin{vmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{vmatrix}^{-} \quad (3)$$

where, $\alpha = \sigma_e^2 / \sigma_u^2$. Let the i diagonal element of A be a_{ij} and of C_{22} be e_{ij}, then

$$V(\hat{u}_j) = Cov(\hat{u}_j, u_j) = a_{jj}\sigma_u^2 - C_{jj}\sigma_e^2$$

and

$$V(\hat{u}_j - u_j) = C_{jj}\sigma_e^2 \quad (4)$$

Variance of prediction error given by Henderson (1973), depends on the model and data structure but not on the observations, unknown true effects, or their estimates. To obtain variances of prediction error, occupied subclasses were needed to be identified; the actual records were not required.
For the sire model, the u in equation 1 represented sire effects equal to one-half the additive genetic value. Populations consisted of q = 63 sires and p = 10, 20 or 30 contemporary groups by subsets to form k cross-classified blocks. Within a block, the number of contemporary groups or the number of sires, whichever was larger, was the minimum number of occupied subclasses possible. Subclasses contained 1 to 23 progeny records, a discrete uniform random variable with mean of 10. These methods ensured sparse non-random usage of sires across contemporary groups, characteristic of livestock populations with limited natural or artificial insemination (Tosh and Wilton, 1994).
Probabilities were continuous uniform random variables determined prior to each pedigree file. Although complete relationship matrices were formed, only the sub-matrix corresponding to the q sires was inverted and included in the coefficient matrix of (3).
Six assumptions regarding variances and covariances of random effects were considered. The model with or without relationships among sires were $'(u) = A\sigma_u^2$ *or* $I\sigma_u^2$ and used one of three levels of heritability, (h^2), 0.10, 0.25 or 0.40 (α = 97.5/2.5; 93.75/6.25 or 90/10).
Data structures were replicated 300 times for each combination of variance-covariance assumption and proportion of occupied subclasses (6x3 total combinations). Table 1 shows the replicates according to p and k. When the number of contemporary groups (p) was low, they were not divided to form large numbers of cross-classified blocks (k) because sires would became nested within contemporary groups. For the number of contemporary groups increased, or the proportion of happened subclasses diminished, the range widened as k decreased.
The coefficient matrix (CM) of equation 3 was set up and divided to obtain C_{22}. Variances of prediction error, as defined by expression equation 3, resulting from the data structures are summarized in Table 2. Variance of prediction error can have values from zero $a_{jj}\sigma_u^2$. For increasing heritability, the additive genetic variance of prediction error and standard deviation and mean also increased. For the jth sire, $a_{ij} = 1 + F_j$, where F_j is the inbreeding coefficient. If the jth sire is non-inbreed, the correlation between true and estimated breeding value and the accuracy of evaluation is Accuracies were calculated according to the equation five for all sires, although approximately 1% of individuals in models with relationships among sires were inbred and their accuracies would had been underestimated. Mean accuracies were 0.59, 0.72 and 0.77 for $h^2 = 0.10$, 0.25 and 0.40, respectively.

Table 1: Number of replicates of data structures by number of contemporary groups (p) and number of cross-classified blocks (k)

	k			
P	2	5	10	Total
10	300	-	-	300
25	150	150	-	300
50	100	100	100	300

Table 2: Sample statistics of variances of prediction error (squared units) obtained from data structures, by model and heritability (h^2)

h^2	Mean	df	Minimum	Maximum
		Sire model		
0.10	2.12	0.50	0.12	2.03
0.25	2.53	1.33	0.13	6.23
0.40	3.62	2.01	0.33	9.06
		Animal Model		
0.10	8.18	0.53	7.11	10.93
0.25	19.67	2.34	12.35	27.44
0.40	27.24	5.36	14.72	40.00

Table 3: Sample statistics of design variables for individuals in data structures simulated according to a sire model

Variable	Mean	SD	Minimum	Maximum
No. of Progeny	39.3	45.6	1	511
Effective no. of Progeny	32.5	40.1	0	352.9
No. of Contemporary groups	3.9	4.4	1	0.48
No of direct connections	19.8	13.4	0	48
No of genetic connections	11.4	10.1	0	47
Value of genetic connections	1.9	1.6	0	13.4

Table 4: Sample statistics of design variables for individuals in data structures designed according to an animal model

Variable	Mean	SD	Min.	Max.
Contemporary group size[a]	23.4	13.6	1	49
No. of Progeny[b]	2.0	2.0	1	34
Effective no of progeny[b]	1.7	1.4	0	12.4
No. of Contemporary Group	1.3	0.7	1	9
No of direct connections[b]	12.3	6.3	0	43
NO of genetic connections[b]	15.3	13.8	0	56
Value of genetic connections	3.4	3.0	0	24.1

a Variable applies only to animals with own record

[b]. Variable applies only to animals with progeny records

Table 5: Correlations of design variables with variance of prediction error, for data structures described according to a sire model, by heritability

Variable	Heritability		
	0.10	0.25	0.40
NP	-0.81	-0.70	-0.63
EP	-0.86	-0.77	-0.71
NCG	-0.78	0-0.67	-0.59
NDC	-0.58	-0.62	-0.61
NGC	-0.06	-0.05	-0.02
VGC	-0.08	-0.07	-0.04

$$r_{u,\hat{u}} = \frac{Cov(\hat{u}_j, u_j)}{\left[V(u_j)\right]^{1/2}\left[V(\hat{u}_j)\right]^{1/2}} = (1 - c_{jj}\sigma)^{1/2} \qquad (5)$$

Table 6: Fitted regression equations for estimating variance of prediction error (σ_e) from 1) number of progeny (N) and 2) both N and Number of direct connections (DC)

Model[a]	Equation	SwR²
$h^2 = 0.40$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 6.67 - 0.0798\,N + 0.00256\,N^2$	0.56
	2 $\hat{y} = 8.54 - 0.06454\,N + 0.000201\,N^2 - 0.215DC + 0.0034\,DC^2$	0.76
$h^2 = 0.40$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 6.85 - 0.0872N + 0.00028\,N^2$	0.59
	2 $\hat{y} = 8.74 - 0.0709\,N + 0.000225N^2 - 0.213\,DC + 0.0033\,DC^2$	0.76
$h^2 = 0.25$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 4.65 - 0.0508\,N + 0.000155\,N^2$	0.62
	2 $\hat{y} = 5.64 - 0.0427\,N + 0.000128N^2 - 0.111\,DC + 0.001\,7DC^2$	0.82
$h^2 = 0.25$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 4.83 - 0.0561\,N + 0.000173\,N^2$	0.69
	2 $\hat{y} = 5.90 - 0.0466\,N + 0.000142\,N^2 - 0.120\,DC + 0.0019\,DC^2$	0.83
$h^2 = 0.10$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 2.18 - 0.0185N + 0.000052\,N^2$	0.78
	2 $\hat{y} = 2.46 - 0.0185\,N + 0.000043\,N^2 - 0.03\,DC + 0.0004DC^2$	0.88
$h^2 = 0.10$; $v(u) = A\,\delta^2_u$	1 $\hat{y} = 2.27 - 0.0197\,N + 0.000053\,N^2$	0.82
	2 $\hat{y} = 2.25 - 0.0170\,N + 0.000044N^2 - 0.030\,DC + 0.0004\,DC^2$	0.90

[a]sire model assuming one of three levels of heritability (h^2.9, with and without relationships among sires (equations 1 and 2, respectively) and δ^2_u is the residual variance

Table 7: Mean variances of prediction error (squared units) for animal swith different sources of information, by heritability

Records	Heritability		
	0.10	0.25	0.40
Progeny	9.5	22.2	33.4
Individual	9.0	18.7	24.5
Progeny, individual	8.8	17.9	23.4

aMean within a column differ ($p<0.01$) from each other

Several variables that describe the data structure with respect to sire were defined with respect to number of progeny (NP), effective number of progeny (NEP), number of contemporary groups in which the sires have progeny (M), and direct connections (DC = number of other sires with progeny in the same contemporary groups). Let n_{ij} be the number of progeny of the jth sire in the ith contemporary of the jth sire is

$$n_{e_j} = \sum_i n_{ij}(n_{i.} - n_{ij})/n_{i.} = n_{.j} - \sum_i (n_{ij}^2/n_{i.})$$

This is traditionally the weighting factor for comparing records of progeny with those of their contemporaries (Searle, 1964).

The relationship matrix also contributes to data structure. Based on A, the following design variables were defined with respect to sire, genetic connections (GC) and value of genetic connections (VGC).

For the animal model, the u of the equation 1 represented the animal additive genetic effects. Design of data under this model differs from that of the sire model, being centered on individuals with records themselves, and all related animals. Populations included n = 45 animals, each with one observation within p = 2, 5 or 10 contemporary groups, as well as their relatives. All animals had either no records or a single record. The vector of q animal effects, q_{od} base population dams with no records, and $q_1 = n$ animals with records.

Statistical Analysis: Correlations between design variables and variance of prediction error were obtained by the evaluated model and its scheme. Simple (one independent variable) curvilinear regression analysis was used to study several design variables in further details and to determine effects on variance of prediction error. According to the results, sire model yielded a wider range of values for the design variables, so these results were used to assess the effects of design variables on the prediction error variance of progeny tests. Two design variables, N and DC, were analyzed jointly using a multiple regression model that included linear and quadratic terms and an interaction.

Results

Sire Model: Many of the design variables were highly inter correlated, which had implications for interpreting results. In simple regression analyses, it is difficult to ascribe a particular effect to the variable in the model or to another highly correlated one (M vs N; r=0.97). It is difficult to dealt with the variables measured as 0 . There fore 0's can be replaced with 1's that are highly correlated with each them (N instead of NE; r=0.95). Moreover, in multiple regression analyses, a regression coefficient describes only a marginal effect given what other correlated independent variables are in the model.

Correlations of design variables with variance of prediction error indicate the strength of linear association, (Table 5). They tended to be higher when variability in variance of prediction error was low, that is, when heritability was low. The largest correlation was always for effective number of progeny, followed by number of progeny. Because sample sizes were very large, even the smallest correlations, for number and value of genetic connections, were different from zero (P<0.01).

Simple linear regression of variance of prediction error on each of the design variables gave the effect of a unit change in that variable. Results were divided by the sire variance to remove differences between heritabilities. For M, the regression coefficient was 10 times that of N but sires had, on average, 10 offspring per contemporary group so the effect of number of contemporary groups was attributed to number of progeny. Number and effective number of progenies were highly intercorrelated and had similar effects on the variance of prediction error.

Number of Progeny: Fitted regression curves in Fig. 1 illustrate the observed association between the accuracy of evaluation and the number of progeny. With the regression models, the number of progeny accounted for 56 to 83% of the variation in variance of prediction error (Table 6) larger R^2 values coincided with lower heritability. Quadratic terms (P<0.01) suggested an on-linear response. As number of progeny increased from a minimum of one, accuracy increased but gradually began to reach a plateau. Reliability of progeny tests can be improved, but with a diminishing rate, by increasing the number of offspring included in the evaluation (Lush , 1931).

Selection index theory defines accuracy of a progeny test based on n_{phs} paternal half-sibs with one record each as follows:

$$r_{u,\hat{u}} = \sqrt{\frac{n_{phs}h^2}{4+(n_{phs}-1)h^2}} \qquad (6)$$

a function of only the number of progeny and heritability (Van Vleck et al., 1987, p. 264). The design variable N is equivalent to n_{phs}. Derived curves (6), start at lower level and show sharper increases to plateaus compared with those of Figure 1. Both selection index and simulation that gave observed accuracies used several simplifying assumptions: no dominance genetic effects, random mating, and no environmental covariances among progeny. The mixed model evaluation, however, took into account all information in the coefficient matrix of (3). Other sires with progeny in the same contemporary groups were important when the number of progeny was low (N<10), where the observed accuracies were greater than those from the selection index,. although selection index values are unknown and must be estimated. The mixed model evaluation simultaneously estimated contemporary group and sire effects.

Therefore, when progeny became very numerous (N>100), observed accuracies approached selection index values. The fitted regression equations in Table 6 can be used to estimate or predict variance of prediction error from number of progeny, or to quantify the effect of a specific change in N, and can be solved for the number of progeny needed to achieve a particular level of accuracy. Mean values of the dependent variable given N are represented by these regression equations. Alternatively, confidence or prediction interval limits of regression equations can be underestimate the necessary number of progeny numbers based on actual accuracies should be used as criteria for improving reliability of progeny tests.

Direct Connections: A direct connection was defined as another sire with progeny in the same contemporary group. Fitted regression curves demonstrate the observed association between accuracy and the number of direct connections. For no genetic relationships in the evaluation model, curves were fitted for DC≥1 and extended through the origin because $r_{u,\hat{u}} = 0$ (Tosh, 1992) with genetic relationships, accuracy could be zero if the related wires also had no direct connections. Zeros pulled the curves downward for low numbers of direct connections. Quadratic terms were highly significant (P<0.01). Curves rose gradually to plateaus. Increasing the number of other sires within contemporary groups improved accuracy, but one other sire was critical.

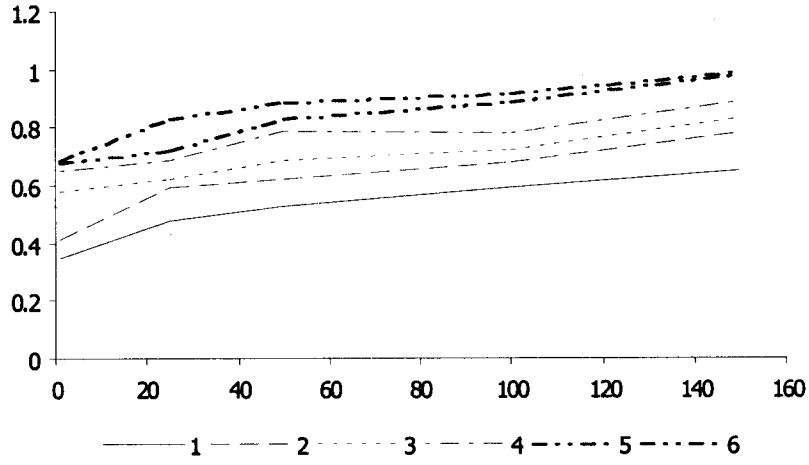

Fig. 1: The observed associatio between accuracy of a progeny test and number of progeny by heritability (h^2) with and without relationships among sires (details in Tabel 6, For example 1 = h2 = 0.40; V(u) = A δ^2_u and 2 = h2 = 0.40' V(u) = I δ^2_u respectively)

Animal Model: The proportion of animals with a progeny test, individual performance test, and their combination, were 0.34, 0.41, and 0.25, respectively. Mean variances of prediction error are given in Table 7. Progeny tests yielded larger variances of prediction error than did individual performance, particularly when heritability was high. The average number of progeny was only two. Lush (1935) showed that even when heritability approaches zero, a progeny test requires more than four progenies to be more accurate in estimating breeding value than the individuals' own record. The small amount of information provided by progeny did not greatly reduce variance of prediction error for the individuals that already had a record themselves. However, differences were significant because sample sizes were large.

Performance Test: Animals with their own record but no progeny had data structure representing performance tests. Contemporary group size and number and value of genetic connections were the design variables for these individuals. Correlations between contemporary group size and variance of prediction error were -0.29, -0.50 and -0.55 for these three groups, respectively. Heritability estimates for these three groups were 0.10, 0.25 and 0.40, respectively. There was no obvious reason for weaker correlations when heritability was lower. Fitted regression curves illustrating the observed association between accuracy of evaluation and contemporary groups were small and seemed to be linear, although quadratic terms were highly significant ($P<0.01$).
With the regression models, contemporary group size explained only 12 to 52% of the variation in variance of prediction error. Therefore, contemporary group size was not very promising for predicting accuracy.
Ignoring genetic relationships, accuracy of a performance test of an individual that is one of n_c contemporaries is:

$$r_{u,\hat{u}} = \sqrt{\frac{h(n_c - 1)}{n_c}} \qquad (7)$$

The equation 7 gives the function of only heritability and contemporary group size (Tosh, 1992).

Discussion

Number of progeny is recognized as useful for predicting accuracy of progeny test evaluations, but the number of direct connections together with progeny numbers has greater predictive ability than either variable alone such that direct connections cannot be ignored. More than one sire should be used per contemporary group to ensure that every individual has at least one direct connection and non-zero accuracy. Relatives other than progeny are not always advantageous because there is no guarantee they provide quality information. With genetic evaluation by an animal model, a portion of the sires and dams are progeny tested: for them, accuracy can also be predicted from

the number of genetic connections, which are mostly progeny. For animals with performance test (and no progeny), accuracy is affected by contemporary group size, particularly for animals without relatives. Contemporary groups should contain at least two individuals to prevent zero accuracy. Larger contemporary groups increase accuracy but there will be little advantage of >5 animals per group, even for animals without relatives.

Wilton *et al.* (1975) presented a similar equation for performance tested bulls that allowed for common sires. Derived curves from equation 7, differ markedly from those in precisely when contemporary groups contain few animals. An individual was alone and had no relatives, accuracy could be zero. At this time, there is no difference between animal and sire model. However, increasing the size of the contemporary group improved information for groups and animals, thereby, increasing the accuracy on the other hand, there was little advantage of >5 animals within a group. At this point, animal model is more efficient than sire model. There are high individual relationships between animals. There, when contemporary groups became large, accuracy approached the square root of heritability. As regard of selection index theory determines accuracy only from the square root of heritability (Van Vleck *et al*, 1987) while assuming known fixed effects.

Genetic connections were not strongly associated with variance of prediction error. Correlations for GC and VGC with variance of prediction error were low ($r=-0.10$ and -0.25, respectively, across h^2). However, relatives were demonstrated to prevent zero accuracy when contemporary groups contained solely one animal. Wood *et al* (1991) found that increasing half sib connections across contemporary groups decreased the mean variance of prediction error for unrelated contemporaries but not necessarily for the half-sibs (Misztal and Wiggans, 1988; Meyer, 1989). Animals that had records were tested according to both their progeny and performance. Corresponding regression equations including quadratic terms ($P<0.01$) had R^2 values of only 6 to 35%. All design variables were less closely associated with accuracy when animals had both progeny and performance tests rather than a single test. No single piece of information was useful for predicting accuracy. An approximation procedure such as that described by Boichard and Lee (1991), which takes into account progeny, parents, contemporaries, and several other contributions to accuracy, is necessary.

References

Boichard, D. and A. J. Lee, 1991. Approximate prediction error variance of breeding values under an animal model. J. Anim. Sci., 69: 188

Henderson, C. R., 1973. Sire evaluation and genetic trends in: Proc. of The Anim. Breed. And Genet. Symp. in Honor of Dr. J. L. Lush. P10. ASAS and ADSA, Champaign, 1L.

Henderson, C. R., 1988. Theoretical basis and computational methods for a number of different animal models. J. Dairy Sci., 71:1.

Lush, J. L., 1931. The number of daughters necessary to prove a sire. J. Dairy Sci., 14: 209.

Lush, 1935. Progeny test and individual performance as indicators of an animals breeding value. J. Dairy Sci., 18: 1.

Meyer, K., 1989. Approximate accuracy of genetic evaluation under an animal model. Livest. Prod. Sci., 21: 87.

Misztal, I. and G. R. Wiggans, 1988. Approximation of prediction error variance: in large scale animal models. J. Dairy Sci., 71: 27.

Petersen, P. H., 1978. A test for connectedness fitted for the two-way BLUP sire evaluation. Acta. Agric. Scand. 28: 360.

Robinson, G. K. and L. P. Jones, 1987. Approximations for prediction error variances. J. Dairy Sci., 70: 1623.

Searle, S. R., 1964. Review of sire proving methods in New Zealand, Great Britain, and New York State. J. Dairy Sci., 47: 402.

Searle, S. R., 1964. Progeny tests of sire and son. J. Dairy Sci., 47: 414.

Tosh, J. J., 1992. Effects of data structure on genetic evaluation of livestock. Ph. D., Thesis. Univ. of Guelph, On Tario, Canada.

Ufford, G. R., C. R., Henderson and L. D. Van Vleck, 1979. An approximate procedure for determining prediction error variances of sire evaluations. J. Dairy Sci., 62: 621.

Van Vleck, L. D., E. J. Pollok and E. A. B. Oltenaca, 1987. Genetics for the Animal Sci. W. H. Freeman and Co., New York.

Wood, C. M., L. L., Christian and M. F. Rothschild, 1991. Use of an animal model in situations of limited subclass numbers and high degress of relationships. J. ANim. Sci., 69: 1420.

Wilmink, J. B. M. and J. Dommerholt, 1985. Approximate reliability of best linear unbiased prediction in models with and without relationships. J. Dairy Sci., 68: 946.

Henderson, C. R., 1984. Applications of linear models in animal breeding. Univ. of Guelph, ON, Canada.

Sorensen, D. A. and B. W., Kennedy, 1984. Estimation of genetic variances from unselected and selected populations. J. Anim. Sci., 59: 1213.

Van der Werf, J. H. J. and R. Thompson, 1992. Variance decomposition in the estimation of genetic variance with selected data. J. Anim. Sci., 70: 2975-2985.

Tosh, J. J. and J. W. Wilton, 1994. Effects of data structure on variance of prediction error and accuracy of genetic evaluation. J. Anim. Sci., 72: 2568-2577.