Group Effects and Computing Strategies for Models for Estimating Breeding Values

G. K. ROBINSON
Department of Agriculture Victoria, Australia

ABSTRACT

No criterion to define group effects in models for estimating breeding values is universally accepted. In this paper it is argued that models that assign group effects to classes of unidentified animals allow group effects and relationships between animals to be combined in a coherent manner. Such models are discussed together with computing strategies to fit them to large amounts of data. In particular, a computing strategy is presented to estimate genetic merits of bulls and cows using a best linear unbiased prediction model, which uses all relationships and all lactations.

INTRODUCTION

This paper will only be concerned with models that are to be fitted using the best linear unbiased prediction (BLUP) approach of Henderson.

Henderson (2) discusses the model:

\[ Y_{ijkl} = h_i + g_j + s_{jk} + e_{ijkl} \]  \[1\]

where \( h_i \) is the \( i \)th herd-year-season effect, \( g_j \) is the group effect for the \( j \)th group of sires, \( s_{jk} \) is the departure from group mean for the \( k \)th sire in the \( j \)th group, and \( e_{ijkl} \) is the error associated with the first lactation yield, \( Y_{ijkl} \), for the \( i \)th daughter of the \( k \)th sire in the \( j \)th group. This model has been used since 1973 at Cornell for the Northeastern Sire Comparison. Groups of sires were defined as sets of sires from a single artificial insemination breeding association entering service during specified periods of time. The basic reason for including group effects in the model was to avoid the assumption that sires came from a single population. In theory, there may be anomalies with grouping schemes like this, such as sires that are full sibs being allocated to different groups. But for sires with large numbers of daughters, such anomalies are unimportant.

Henderson (3) discovered that relationships could readily be included explicitly in BLUP models because the inverse of a relationship matrix is easier to compute than the relationship matrix itself. Henderson (6) presents the most complete discussion of this and he (4) discusses a particular case commonly used in dairy sire evaluation.

Denoting the relationship matrix by \( A \), one way of understanding why \( A^{-1} \) might be expected to be simpler than \( A \) is that nonzero elements of \( A^{-1} \) correspond to nonzero partial correlations of genetic merits of two animals given all other animals in the model. For instance, the partial correlation of an animal's genetic merit with that of a grandparent given the genetic merits of its parents is zero but the simple correlation is not zero. See (9), equation [4g,2.8] and associated discussion of partial correlation. In general, it is reasonable to expect that the \((i, j)\) element of \( A^{-1} \) will be nonzero only if \( i \) and \( j \) are parent and offspring, not necessarily respectively.

Henderson’s (6) demonstration that \( A^{-1} \) may be easily computed expresses \( A \) in the form \( TDT' \) where \( D \) is diagonal, \( T \) is a lower triangular matrix with unit diagonal, and animals have been ordered so that parents precede offspring. Thompson (10) uses the matrix \( T \) in his approach to include group effects in a model that includes relationships. From Henderson’s recursive definition it may be seen that the \((i, j)\) element of \( T \) is the total of 1/2 if \( i \) is a parent of \( j \), 1/4 for each time \( i \) is a grandparent of \( j \), 1/8 for each time \( i \) is a great grandparent of \( j \), etc. If there is no inbreeding, there can be at most one contribution to any element of \( T \).

Received February 13, 1986.

Current address: CSIRO, Division of Mathematics and Statistics, Private Bag 10, Clayton, Victoria, 3168, Australia.

MODELS TO ESTIMATE BREEDING VALUES

In sire evaluation, relationships and group effects allow for genetic trends and genetic differences between subpopulations to some extent. It is not obvious how the two should be combined, if indeed they should.

One line of development has been to investigate the properties of known models that use both relationships between sires and group effects. This approach has not been used here. The interested reader is referred to Quaas and Pollak (8). Another line of development has been to develop models in which relationships and group effects are coherent in some sense. Thompson (10) suggested such a model and the models suggested in this paper are similar to his.

Thompson [(10), p. 346] shows that each sire effect can be written as an accumulation of independent terms from its ancestors and itself and suggests that group effects should be associated with these independent terms. This suggestion cannot conveniently be discussed in terms of Henderson's (2) notation.

In Thompson's notation, model [1] becomes:

\[ Y_{ijkl} = \beta_i + u_j + G_{g(j)} + e_{ijk} \]

where \( \beta_i = \mu_i \), the \( j \text{th} \) herd-year-season effect; \( u_j = s_{jk} \), the sire effect for the \( j \text{th} \) sire; \( G_{g(j)} = g_j \), the group effect for the group \( g(j) \) to which the \( j \text{th} \) sire belongs; and \( k = 1 \), the number of the daughter.

Thompson shows that:

\[ u_j = \Sigma T_{lj} u^* \]

where \( u^* \) are independent effects with variance-covariance matrix \( D \) times the additive genetic variance for the trait, \( T_{lj} \) is the \((l, j)\) element of the matrix \( T \) such that \( A = TDT' \), as explained earlier, and the summation is over ancestors of \( j \) together with \( j \) itself (all other \( T_{lj} \) being zero).

Thompson's suggested model is:

\[ Y_{ijk} = \beta_i + \Sigma T_{lj} (u^* + G_{g(l)}^*) + e_{ijk} \]

or equivalently:

\[ Y_{ijk} = \beta_i + u_j + \Sigma T_{lj} G_{g(l)}^* + e_{ijk} \]

where the group effects \( G_{g(l)}^* \), which are associated with the independent \( u^* \), need bear no particular relationship to the group effects written without asterisks.

Both Henderson's and Thompson's models were intended to be used for sire evaluation in which daughter-dam relationships are neglected. Henderson (5) showed that all relationships could be used when simultaneously estimating genetic merits of bulls and cows at least provided that the total number of animals was small. A desire to extend the concept of group effects to models using all relationships was a primary motivation for developing a new way to define group effects.

The computing strategy in this paper shows that the model of Henderson (5) may be fitted to large data sets. This development is likely to be of interest to people who are not interested in group effects. The section of the paper on computing strategies for sire evaluation is only of interest when models with group effects are to be fitted.

PRINCIPLES FOR ASSIGNING GROUP EFFECTS

One way of defining group effects for a model to be used for estimating the genetic merits of animals is to take the following four steps:

1. List animals that are females with records in the data, sires of such females, or ancestors needed to indicate relationship between animals already listed.

2. List animals that are parents of animals listed in step 1 but were not themselves listed in step 1. Generally, these animals can only be listed as "sire of A" or "dam of B" where A and B are well identified.

3. Classify the animals listed in step 2 into groups of animals thought likely to be of similar genetic merit and assign group effects to those groups. No particular guidelines for this step are being suggested.

4. The group effects for animals listed in step 1 can be worked out recursively on the basis that each animal's group effect is the average of its parents' group effects.

A desirable feature of this method of defining group effects is that they are guaranteed to be consistent with relationship between animals. A situation where the result of fol-
lowing these steps is easy to envisage is where the genetic merits of various pure bred animals and animals of mixed breed are to be estimated simultaneously. The animals listed in step 2 could be classified on the basis of breed and the group effects for all animals would correspond to breed proportions as estimated by pedigree. In notation like that of Thompson (10), an explicit formula for the group effect of the \( j \)th animal is:

\[
\sum_{t \in P} T_{jt} G_{g(t)^*}
\]

where \( P \) denotes the set of parents listed in step 2, \( g(t) \) denotes the group to which animal \( t \) belongs, and \( G_{g(t)^*} \) denotes the group effect for that group. The recursive definition in the four steps is generally more convenient, both conceptually and computationally, than this formula. Consider an example involving three bulls and two cows related in that bull A is the sire of bull B, which is the sire of cow 1, which is the dam of cow 2 and bull C is the sire of cow 2. These relationships can be represented diagrammatically (Figure 1).

This prescription is easiest to follow when animals of both sexes are being evaluated using all relationships between animals in the manner of Henderson (5). Step 1 lists the five animals already mentioned. Step 2 lists the animals: sire of bull A, dam of bull A, dam of bull B, dam of cow 1, sire of bull C, dam of bull C.

The basis for grouping animals in step 3 will vary among applications of the prescription. Step 4 is straightforward to apply. For instance, the group effect for cow 2 will be the group effects for the animals listed above multiplied by the weights 1/16, 1/16, 1/8, 1/4, 1/4, and 1/4, respectively. Note that these weights add to unity. In the matrix notation of Henderson (5), group effects are part of the vector \( \beta \) of unknown fixed effects and these weights are part of the matrix \( X \).

For a sire evaluation as discussed in Henderson (4), the only relationships used are sire-daughter (where the daughter's first lactation is recorded), sire-son, and maternal grand sire-grandson. In the preceding sample, the dam-daughter relationship between cow 1 and cow 2 would be neglected and the list of animals formed in step 2 would be extended to include dam of cow 2. The group effect of cow 2 would now be 1/4 of the group effect for the sire of bull C plus 1/4 of the group effect for the dam of bull C plus 1/2 of the group effect for the dam of cow 2.

The prescription requires careful application when maternal grandsire-grandson relationships are processed. One method is to insert the missing female relatives for the purpose of finding group effects. Equivalently, maternal granddams rather than dams of bulls whose maternal grandsires are known could be listed in step 2 and for these bulls step 4 would calculate their group effects as 1/4 of maternal granddam group effect plus 1/4 of maternal grandsire group effect plus 1/2 of sire group effect.

In principle, this method of defining group effects would use cow group effects rather than sire group effects in sire evaluation models. This distinction is of no practical consequence provided that dams of cows having their first lactations in the same herd-year-season are always allocated to the same group, that group effects used for cow dams are not used for other animals, and that cows of unknown sires are excluded. In such circumstances cow dam group effects cannot be separated from herd-year-season effects and cow group effects are always the average of a cow dam group effect and the group effect of a known sire, so only the group effects of known sires need to be included in the estimating equations. In other cases the distinction is important.

Care must again be exercised when applying the prescription to sire evaluation using cows' maternal grandsires as discussed in Everett et al. (1) and Quaas et al. (7). For the example, cow

```
Bull A

Bull B

Cow 1

Cow 2

Bull C
```

Figure 1. Relationships between animals in example.
1 can be considered to be replaced by two cows both of which are daughters of bull B. One of these, say cow 1.1, is considered to have produced lactations of cow 1 and the other, cow 1.2, is considered to be the dam of cow 2 but to have no known lactations. The list of animals produced in step 2 would now omit “dam of cow 1” and add “dam of cow 1.1” and “dam of cow 1.2”. The group effect for cow 2 would be 1/16 of group effect for sire of bull A plus 1/16 of group effect for dam of bull A plus 1/8 of group effect for dam of bull B plus 1/4 of group effect for dam of cow 1.2 plus 1/4 of group effect for sire of bull C plus 1/4 of group effect for dam of bull C.

COMPUTING STRATEGIES FOR SIRE EVALUATION

As Thompson (10) indicated, models with complicated group effects require more computational effort than models with simple group effects. The most commonly used computing strategy was developed at Cornell and is little affected by whether cows' maternal grandsires are included in the model or not. See Henderson [(2), p. 22] and Everett et al. [(1), p. 1306] for summaries of two versions of this basic computing strategy.

To fit a model with complicated group effects, either as suggested in this paper or according to Thompson's method, the first step is to express each sire's group effect as a weighted sum of simple group effects. The method of doing this will depend on the basis used for assigning group effects. Computationally, this is not an expensive part of the total process of fitting a model so it does not need to be performed particularly efficiently.

For Thompson's method of modeling group effects, the only further change required to the computing strategy is that the method of deriving the group effect equations from the sire by sire portion of the BLUP equations is more complicated than for a model with simple group effects. Using the notation of Everett et al. (1) for the BLUP equations with herd-year-season equations absorbed:

\[
\begin{bmatrix}
P & Q \\
Q' & C + \theta A^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{g} \\
\hat{s}
\end{bmatrix} =
\begin{bmatrix}
t \\
r
\end{bmatrix}
\]

P and Q and t can be derived from C and r using the known group effect expressions for the sires in the model.

When cow group effects, as suggested by this paper, are used in sire evaluation it is generally not true that P, Q, and t can be derived from C and r. This means that group equations cannot be ignored during the absorption of herd-year-season equations. One feasible computing strategy would be to find the coefficients for each cow's group effect as a weighted sum of pure group effects as that cow's lactation was processed and to include group effect equations in the process of absorbing herd-year-season effects. For large data sets, contributions to P could normally be accumulated in the computer memory, but contributions to Q would need to be written to disk and later sorted and contributions to the same elements added together.

This procedure requires that numbers expressing sire group effects as sums of group effects either be in computer memory during the absorption of herd-year-season effects or be included on the lactation file.

A procedure that will be more efficient, but probably require a more complicated computer program, is to neglect the group effects of known bulls during the absorption of herd-year-season equations but include the group effects of unknown bulls and all group effects of cows. Group effects of cows and unknown bulls should be deducible from information (like dates of birth) easy to include on the lactation file. The contributions to Equation [4] in the P, Q, Q', and t portions, which come from the group effects of known bulls, can be computed after contributions to elements of C and r have been accumulated in the usual way.

COMPUTING STRATEGY FOR ALL RELATIONSHIPS

Henderson (5) showed how BLUP could use all relationships when estimating genetic merits of animals. This method has been used at Cornell to estimate genetic merits of cows one herd at a time using estimated genetic merits of bulls found by other methods as given.

The Australian Dairy Herd Improvement Scheme has used this method for simultaneous evaluation of cows and bulls on a nationwide basis. The developments from Henderson (5) are all computational.
Features of the computing strategy that do not concern group effects are as follows:

1. Daughter-dam relationships where the cows are in different herds are neglected.
2. Where a cow has lactations in more than one herd she is treated as being a different cow in each of the herds. These two features are essential so that some computations can be done on a within-herd basis.
3. All lactations are used but they are weighted according to age at calving. Weights are unity for age 2 and $.8 \times (.85)^{n-3}$ for age $n(>3)$. These weights were chosen as a compromise between using first lactations only and using all lactations.
4. The influences of age at calving, season of calving, and region are reduced as much as possible before BLUP analysis. The procedures outlined in Beard and Robinson (1985, unpublished data) are now used whenever test day data are available.
5. Equation numbers are allocated to bulls, within-herd sequence numbers are allocated to cows, and these identifiers are used on all data input to the analysis programs.
6. No equations are absorbed. The computing cost of the estimation would probably be slightly less if equations for nonadditive genetic merit plus permanent environmental effects ($\varpi$) were absorbed, but the computer program would have to be more complex. For most minor modifications of the estimation (e.g., to use different heritabilities for milk and fat yields) to not absorb $\varpi$ is definitely better.
7. For each iteration of the equation solving technique, data concerning cows are processed one herd at a time. The most recent approximations for within-herd effects are read in, data concerning the herd are read in, the next approximations to within-herd effects are computed and written out, and computations required toward the next approximations to bull genetic merit estimates are performed.
8. There is a computer program that preprocesses the lactation and pedigree data into a form that enables the equation-solving program to proceed as rapidly as possible. These are separate programs so that equation solving can be restarted from any completed iteration.

Addition of group effects of this model requires three main extra computation jobs. 1) Group effects of bulls need to be expressed as weighted combinations of pure group effects. Computer programming has not yet been done for this step. 2) The program that preprocesses the lactation and pedigree data must compute the portion of the $X'X$ matrix [in Henderson's (5) notation] for group effects by group effects. This is a straightforward but computationally expensive accumulation of sums of squares and crossproducts. 3) To process each herd within each iteration of the equation-solving technique, the latest approximation to the group effect for each cow can be evaluated using the recursive definition of group effects. Contributions from within-herd effects to the equations for group effects can be evaluated using a procedure, which is essentially the same but in reverse. This requires only a small number of operations per cow per iteration.

There are a number of other computational steps necessitated by addition of group effects to the model. The group by group submatrix must be inverted by the preprocessing program and the inverse used during each iteration; group by herd-year-season effects must be handled; and several other minor jobs. The three jobs listed are the ones that need to be most carefully noted to estimate programming and computing costs.

**DISCUSSION**

The computer programs for using all relationships in evaluating bulls and cows have all been written except for a few required to express bull group effects as weighted combinations of pure group effects as noted. Computer programs for sire evaluation using cow group effects rather than the more common bull group effects have not been written.

The situation where the cow group effects approach seems most desirable in sire evaluation is where cows' maternal grandsires are used, but not all cows in the data set have known maternal grandsire (which would be in essentially every case). The usual approach makes an implicit assumption that all unknown maternal grandsires are of equal genetic merit. This is unlikely to be a reasonable assumption, particularly if the extent of pedigree information is not uniform.

Group effects could be regarded as including any fixed effects that apply to animals rather than lactations. Conversion of estimates of
breeding values obtained in another evaluation system could be performed by constructing two variables, a zero-one variable, which indicates whether the animal has an evaluation in that system and the departure from average merit as estimated by that system. These variables could be included as group effects and the estimates of their coefficients could be used to convert evaluations of animals from that system to one's own for animals without a local estimate of breeding value.

REFERENCES


