Ordinary Least Squares Estimation of General and Specific Combining Abilities from Half-Diallel Mating Designs

By D. A. Huber, T. L. White, R. C. Littell, and G. R. Hodge

Department of Forestry, Department of Statistics, University of Florida, Gainesville, FL 32611, USA

(Received 16th December 1981)

Abstract

This paper presents a matrix algebra approach to the solution of ordinary least squares (OLS) equations for estimation of genetic parameters as fixed effects in a half-diallel mating system. The relative ease of implementation and omnipresence of OLS based analysis in forest genetics applications motivates the current discussion. Few statistical packages are available that allow direct specification of a half-diallel model and dependence on standard formulae for estimation can lead to erroneous results if the data are unbalanced (plots or crosses are missing). The methods offered herein can be implemented with any software capable of matrix manipulations and can be extended to any linear model.

The mechanics of the OLS analysis are described to foster understanding of the assumptions and mathematical properties of ordinary least squares analysis and sum-to-zero restrictions. The impacts of OLS assumptions and sum-to-zero restrictions on the interpretation of genetic parameter estimates are examined, and the relationship between OLS assumptions and other types of analyses is presented. The analysis is developed through formation of the design matrix by modeled parameters as an extension of the scalar linear model for half-diallel designs. The overparameterized design matrix is created for balanced data (plot-mean basis) and is then reparameterized to full-rank in sum-to-zero format. Adjustments to the design matrix prompted by various common causes of data imbalance are demonstrated. Numerical examples are provided along with the data set so that the reader may recreate the analysis.

Two problems which are basic in a genetic analysis are addressed: 1) the variance-covariance matrix for the observations, and 2) the desire to compare breeding value estimates from disconnected experiments. In general, OLS assumptions about the variance-covariance matrix for cell (plot) means are not appropriate and evaluation of these assumptions often suggests other types of analyses especially for large data bases. Comparison of breeding value estimates from disconnected experiments is problematic with any analysis method; however, a method is offered to improve OLS estimate comparability.

Key words: Unbalanced data, estimability.

Introduction

The diallel mating system is an altered factorial design in which the same individuals (or lines) are used as both male and female parents. A full diallel contains all crosses, including reciprocal crosses and selfs, resulting in a total of $p^2$ combinations, where $p$ is the number of parents. Assumptions that reciprocal effects, maternal effects, and paternal effects are negligible lead to the use of the half-diallel mating system (Griffing, 1956, method 4) which has $p(p-1)/2$ parental combinations and is the mating system addressed in this paper.

Half diallels have been widely used in crop and tree breeding (Sprague and Tatum, 1942; Gilbert, 1958; Matzinger et al., 1958; Burley et al., 1966; and Squillace, 1973) and the widespread use of this mating system continues today (Weir and Zobel, 1975; Welcox et al., 1975; Snyder and Namkoong, 1978; Hallauer and Miranda, 1981; Singh and Singh, 1984; Greenwood et al., 1986; and Weir and Goodard, 1986).

Most of the statistical packages available treat fixed effect estimation as the objective of the program with random variables representing nuisance variation. Within this context a common analysis of half-diallel experiments is conducted by first treating genetic parameters as fixed effects for estimation of general (GCA) and specific (SCA) combining abilities and subsequently as random variables for variance component estimation (used for estimating heritabilities, genetic correlations, and general to specific combining ability variance ratios for determining breeding strategies). This paper focuses on the estimation of GCA's and SCA's as fixed effects. The treatment of GCA and SCA as fixed effects in OLS (ordinary least squares) is an entirely appropriate analysis if the comparisons are among parents and crosses in a particular experiment. If, as forest geneticists often wish to do, GCA estimates from disconnected experiments are to be compared, then methods such as checklists must be used to place the estimates on a common basis.

Formulae (Griffing, 1956; Falconer, 1981; Hallauer and Miranda, 1981; and Becker, 1975) for hand calculation of general and specific combining abilities are based on a solution to the OLS equations for half-diallels created by sum-to-zero restrictions i.e. the sum of all effect estimates for an experimental factor equals zero). These formulae will yield correct OLS solutions for sum-to-zero genetic parameters provided the data have no missing cells. If cell (plot) means are used as the basis for the estimation of effects there must be at least one observation per cell (plot) where a cell is a subclassification of the data defined by one level of each factor (Sturle, 1987). An example of a cell is the group of observations denoted by $AB_1$ for a randomized complete block design with factor $A$ across blocks ($B$). If the above formulae are applied without accounting for missing cells, incorrect and possibly misleading solutions can result. The matrix algebra approach described in this paper for these reasons: 1) in forest tree breeding applications data sets with missing cells are extremely common; 2) many statistical packages do not allow direct specification of the half-
diallel model; 3) the use of a linear model and matrix algebra can yield relevant OLS solutions for any degree of data imbalance; and 4) viewing the mechanics of the OLS approach is an aid to understanding the properties of the estimates.

The objectives of this paper are to: (1) detail the construction of ordinary least squares (OLS) analysis of half-diallel data sets to estimate genetic parameters (GCA and SCA) as fixed effects, (2) recount the assumptions and mathematical features of this type of analysis, (3) facilitate the reader's implementation of OLS analyses for diallels of any degree of imbalance and suggest a method for combining estimates from disconnected experiments, and (4) aid the reader in ascertaining what method is an appropriate analysis for a given data set.

\[ y_{jk} = \mu + B_j + GCA_j + GCA_k + SCA_{jk} + e_{jk} \]

where: \( y_{jk} \) is the mean of the \( j \text{th} \) block for the \( k \text{th} \) cross
\( \mu \) = an overall mean
\( B_j \) = fixed effect of block \( j \) for \( j = 1 \) to \( b \);
\( GCA_j \) = fixed general combining ability effect of the \( j \text{th} \) female
parent or \( k \text{th} \) male parent, \( j \text{ or } k = 1, \ldots , p \ (j + k) \);
\( SCA_{jk} \) = fixed specific combining ability effect of parents \( j \text{ and } k \);
\( e_{jk} \) = the random error associated with the observation of the \( j \text{th} \) block in the \( k \text{th} \) block, \( e_{jk} \sim (0, \sigma^2_e) \).

Cross by block interaction as genotype by environment interaction is treated as confounded with between plot variation as for contiguous plots.

The model in matrix notation is:

\[ y = X\beta + e \]

where: \( y \) is the vector of observation vectors (nx1 = n rows and 1 column) where \( n \) equals the number of observations;
\( X \) is the design matrix (nxm) whose function is to select the appropriate parameters for each observation where \( m \) equals the number of fixed effect parameters in the model;
\( \beta \) is the vector (mx1) of fixed effect parameters ordered in a column; and
\( e \) is the vector (nx1) of deviations (errors) from the expectation associated with each observation.

2.1 Linear Model

Plot means are used as the unit of observation for this analysis with unequal numbers of observations per plot. Plot (cell) means are always estimable as long as there is one observation per plot, and linear combinations of these means (least squares means) provide the most efficient way of estimating OLS fixed effects (Yates, 1934). Throughout this paper, estimates are denoted by lower case letters while the parameters are designated by upper case letters and matrices are in bold print.

Using plot means as observations, a common scalar linear model for an analysis of a half-diallel mating design with \( p(p-1)/2 \) crosses planted at a single location in a randomized complete block design with one plot per block is:

\[ (I) \]

Observations \( (y) \) is equal to \( I_{nxn} \), where \( I \) is an nxn identity matrix. The elements of an identity matrix are 1's on the main diagonal and all other elements are 0. Multiplying \( I \) by \( \sigma^2_e \) places \( \sigma^2_e \) on the main diagonal. In the variance-covariance matrix for the observations, the variance of the observations appears on the main diagonal and the covariance between observations appears in the off-diagonal elements. Thus, \( V = I_{nxn} \) states that the variance of the observations is equal to \( \sigma^2_e \) for each observation and there are no covariances between the observations (which is one direct result of considering genetic parameters as fixed effects).

2.1.2 Sum-to-Zero Restrictions

The design matrix presented in this paper is reparameterized by sum-to-zero restrictions to: (1) reduce the dimension of the matrices to a minimal size, and (2) yield estimates of fixed effects with the same solution as common formulae in the balanced case. Other restrictions such as set-to-zero could also be applied so the discussion that follows treats sum-to-zero restrictions as a specific solution to the more general problem which is finding an inverse for \( X'X \). The subscripts '0' and 's' refer to the overparameterized model and the reparameterized model with sum-to-zero restrictions, respectively.

The matrix \( X_o \) of figure 1 is the design matrix for an overparameterized linear model (Mikkelson and Johnson, 1984, page 96). Overparameterization means that the equations are written in more unknowns (parameters, in this case 13) than there are equations (number of observations minus degrees of freedom for error, in this case 12 - 5 = 7) with which to estimate the parameters. Re-parameterization as a sum-to-zero matrix overcomes this dilemma by reducing the number of parameters through making some of the parameters linear combinations of others. Sum-to-zero restrictions make the resulting parameters and estimates sum to zero even though the unrestricted parameters (for example the true GCA values as applied to a broader population) do not necessarily sum-to-zero within a diallel. This is the problem of comparability of GCA estimates from disconnected experiments.

To illustrate the concept of sum-to-zero estimates versus population parameters, we use the expectation of a common formula. Becker (1976) gives equation 4 (which for balanced cases is equivalent to \( \sigma^2 = ((p-1)/(p-2))(\zeta - \zeta) \)) as the estimate for general combining ability for the \( j \text{th} \)
line with \( p \) equalling the number of parents and \( Z \) equalling the site mean of the \( j \times k \) cross. This equation yields the same solution as the matrix equations with no missing plots or crosses and with a design matrix which contains the sum-to-zero restrictions. An evaluation of this formula in a four-parent half-diallel planted in \( b \) blocks for the GCA of parent 1 is obtained by substituting the expectation of the linear model (Equation 1) for each observation:

\[
g_j = \frac{1}{(p(p-2))}(pZ_j - 2Z_\cdot) \]

\[
E\{g_1\} = E\{\frac{1}{(p(p-2))}(pZ_1 - 2Z_\cdot)\}
\]

\[
E\{g_1\} = \frac{3}{4}(GCA_1) - \frac{1}{4}(GCA_2 + GCA_3 + GCA_4) + \frac{1}{4}(SCA_{12} + SCA_{13} + SCA_{14}) - \frac{1}{4}(SCA_{23} + SCA_{24} + SCA_{34})
\]

The result of equation 4 is obviously not GCA from the unrestricted model (Equation 1). Thus, \( g_1 \) is an estimable function and an estimate of parameter GCA\(_{1\cdot}\) (the estimate of the GCA of parent 1 given the sum-to-zero restrictions), does not have the same meaning as GCA\(_1\) in the unrestricted model. An estimable function is a linear combination of the observations; but in order for an individual parameter in a model to be estimable, one must devise a linear combination of the observations such that the expectation has a weight of one on the parameter one wishes to estimate while having a weight of zero on all other parameters. A solution such as this does not exist for the individual parameters in the overparameterized model (Equation 1). So although the sum-to-zero restricted GCA parameters and estimates are forced to sum-to-zero for the sample of parents in a given diallel, the unrestricted GCA parameters only sum-to-zero across the entire population (Falconer, 1981) and an evaluation of GCA\(_{1\cdot}\) demonstrates that the estimate contains other model parameters.

The result of sum-to-zero restrictions is that the degrees of freedom for a factor equals the number of columns (parameters) for that factor in \( X_0 \) (Figure 2). Thus, a generalized inverse for \( X_0'X_0 \) is not required since the number of columns in the sum-to-zero \( X_0 \) matrix for each factor equals the degrees of freedom for that factor in the model \( X_0 \) is full column rank and provides a solution to Equation 3.

**Figure 1.** The overparameterized linear model for a four-parent half-diallel planted on a single site in two blocks displayed as matrices. The design matrix \( X_0 \) and parameter vector \( \beta_0 \) are shown in overparameterized form. 1's and 0's denote the presence or absence of a parameter in the model for the observed means (data vector, \( y \)). The parameters displayed above the design matrix label the appropriate columns for each parameter. Error vector not exhibited.

\[
y = X_0 \beta_0 + \epsilon
\]

**Figure 2.** The linear model for a four-parent half-diallel planted on a single site in 2 blocks displayed as matrices. The design matrix \( X_0 \) and the parameter vector \( \beta_0 \) are presented in sum-to-zero format. The parameters displayed above the design matrix label the appropriate column for each parameter.
2.2 Components of the Matrix Equation

The equational components of 2 are now considered in greater detail.

2.2.1 Data Vector y

Observations (plot means) in the data vector are ordered in the manner demonstrated in Figure 1. For our example, Figure 1 is the matrix equation of a four parent half-diallel mating design planted in two randomized complete blocks on a single site. There are six crosses present in the two blocks for a total of 12 observations in the data vector, y. The observations are first sorted by block. Second, within each block the observations should be in the same sequence (for simplicity of presentation only). This sequence is obtained by assigning numbers 1 through p to each of the p parents and then sorting all crosses containing parent 1 (whether as male or female) as the primary index in descending numerical order by the other parent of the cross as the secondary index. Next all crosses containing parent 2 (primary index, as male or female) in which the other parent in the cross (secondary index) has a number greater than 2 are then also sorted in descending order by the secondary index. This procedure is followed through using parent p—1 as the primary index.

2.2.2 Design Matrix and Parameter Vector, X and β

The design matrix for a model is conceptually a listing of the parameters present in the model for each observation (Seahle, 1987, page 243). In Figure 1, y and β are exhibited and the parameters in βp are displayed at the tops of the columns of Xp (a visually correct interpretation of the multiplication of a matrix by a vector). For each observation in y, the scalar model (Equation 1) may be employed to obtain the listing of parameters for that observation (the row of the design matrix corresponding to the particular observation). The convention for design matrices is that the columns for the factors occur in the same order as the factors in the linear model (Equation 1 and Figure 1). Since design matrices can be devised by first creating th columns pertinent to each factor in the model (submatrices) and then horizontally and/or vertically stacking the submatrices, the discussion of the parameterized design matrix formulation will proceed by factor.

2.2.3 Mean

The first column of Xp is for μ and is a vector of I's with the number of rows equaling the number of observations (Figure 2). The linear model (Equation 1) indicates that all observations contain μ and the deviation of the observations from μ is explained in terms of the factors and interactions in the model plus error.

2.2.4 Block

The number of columns for block is equal to the number of blocks minus one (column 2, X2). Each row of a block submatrix consists of I's and 0's or —1's according to the identity of the observation for which the row is being formed. The normal convention is that the first column represents block 1 and the second column block 2, etc. through block b—1. Since we have used a sum-to-zero solution (Σβp = 0), the effect due to block b is a linear combination of the other b—1 effects, i.e., βb = —Σb=1b+βp. This, in our example is 0 = b1 + b2 and β3 = b3. Thus, the row of the block submatrix for an observation in block (b the last block) has a —1 in each of the b—1 columns signifying that the block b effect is indeed a linear combination of the other b—1 block effects. Columns 2 and 3 of Xp (Figure 1) have become column 2 of Xp (Figure 2).

2.2.5 General Combining Ability

This submatrix of Xp is slightly more complex than previous factors as a result of having two levels of a main effect present per observation, i.e., the deviation of an observation from μ is modeled as the result of the GCA's of both the male and female parents (Equation 1). Again we have imposed a restriction, Σgacap = 0. Since GCA has b—1 degrees of freedom, the submatrix for Xp should have p—1 columns, i.e., gcap = —Σp=1b+gcap. The GCA submatrix for Xp (columns 3 through 5 in Figure 2) is formed from Xp (columns 4 through 7 in Figure 1) in the same manner as the block matrix: (1) add minus one to the elements in the other columns along each row containing a one for gcap (p = 4 in our example); and (3) delete the column from Xp corresponding to gcap. The GCA submatrix has p(p—1)/2 rows (the number of crosses). This, with no missing cells (plots), equals the number of observations per block. To form the GCA factor submatrix for a site, the GCA submatrix is vertically concatenated (stacked on itself) b times. This completes the portion of the Xp matrix for GCA.

2.2.6 Specific Combining Ability

In order to facilitate construction of the SCA submatrix, a horizontal direct product should be defined. A horizontal direct product, as applied to two column vectors, is the element by element product between the two vectors (SAS/IML User's Guide, 1989) such that the element in the ith row of the resulting product vector is the product of the elements in the ith rows of the two initial vectors. The resultant product vector has dimension n x 1. A horizontal direct product is useful for the formation of interaction or nested factor submatrices where initial matrices represent the main factors and the resulting matrix represents an interaction or a nested factor (product rule, Seahle, 1987).

The SCA submatrix can be formulated from the horizontal direct products of the columns of the GCA submatrix in Xp (Figure 2). The results from the GCA columns require manipulation to become the SCA submatrix (since degrees of freedom for SCA do not equal those of an interaction for a half-diallel analysis), but the GCA column products provide a convenient starting point. The column of the SCA submatrix representing the cross between the i-th and the k-th parents (SCAIj) is formed as the product between the GCAj and GCAk columns (Figure 3). The SCA columns in Figure 2 are multiplied in this order: column 1 times column 2 forming the first SCA column, column 1 times column 3 forming the second SCA, and column 2 times column 3 forming the third SCA column (Figure 3). With four parents (six crosses) there are three degrees of freedom for GCA (p—1) and two degrees of freedom for SCA (6 crosses — 3 for GCA — 1 for the mean). Since SCA has only two degrees of freedom, a sum-to-zero design matrix can have only two columns for SCA. Imposing the restriction that the sum of the SCA's across all parents equals zero is equivalent to making the last column for the SCA submatrix (Figure 3).

9) SAS/IML is the registered trademark of the SAS Institute Inc. Cary, North Carolina.
3) A linear combination of the others (Figure 2). The procedure for deleting the third column product is identical to that for the GCA submatrix: add minus one to every element in the rows of the remaining SCA columns in which a one appears in the column which is to be deleted (Figure 2, columns 6 and 7). The number of rows in the SCA submatrix equals the number of observations in a block and must be vertically concatenated b times to create the SCA submatrix for a site.

An algebraic evaluation of SCA sum-to-zero restrictions requires that \( \sum \Sigma_{i} \Sigma_{j} \text{sc}_{ji} = 0 \) for each k and that \( \sum \Sigma_{i} \text{sc}_{1i} = 0 \); thus, for observations in the ith block with i serving to denote the row of the SCA submatrix in block i, \( \text{sc}_{1i} \) = \( \text{sc}_{11} - \text{sc}_{12} \), and entries in the submatrix row for \( y_{1i} \) are \( y_{1i} \) = \( y_{11} \) - 1. \( \text{sc}_{1i} \) equals \( \text{sc}_{1i} \) because \( \text{sc}_{1i} \) is the negative of the sum of the independently estimated SCA's \( \text{sc}_{1i} - \text{sc}_{1j} \) from the restriction that the sum of the SCA's across all parents equals zero. Similarly, by sum-to-zero definition \( \text{sc}_{12} = - \text{sc}_{22} + \text{sc}_{12} \) and by substitution \( \text{sc}_{12} = \text{sc}_{12} - \text{sc}_{12} = \text{sc}_{12} \). By the same protocol, it can be shown that \( \text{sc}_{22} = \text{sc}_{22} \). The elements in the rows of the SCA submatrix are \( 1, -1 \) and 0's in accordance with the algebraic evaluation. Thus, while it may seem that there should be 6 SCA values (one for each cross), only 2 can be independently estimated and the remaining 4 are linear combinations of the independently estimated SCA's. Again, the SCA sum-to-zero estimates are not equal to the parametric population SCA's. An analogous illustration for SCA to that for GCA would show that the estimable function (linear combination of observations) for a given SCA contains a variety of other parameters.

2.3 Estimation of Fixed Effects

2.3.1 GCA Parameters

The GCA parameters can be estimated (without mean, block, and SCA in the design matrix) through the use of equation 3, if there are no missing cell means (plots) for any cross and no missing crosses. The design matrix consists only of the GCA submatrix. This design matrix has \( n \) (i.e. the vector \( \mathbf{gca}_n \)) as in Figure 2 and the linear and quadratic interactions for the estimation of gca = \( \mathbf{gca}_n = \sum_{i=1}^{n} \mathbf{gca} \). Parameters for any of the factors can be estimated independently using the pertinent submatrix as long as there are no missing cell means (plots) and no missing crosses; this uses a property known as orthogonality.

Orthogonality requires that the dot product between two vectors equals zero (Schnyder, 1987, page 168). The dot product (a scalar) is the sum of the values in a vector obtained from the horizontal direct product of two vectors. For two factors to be orthogonal, the dot products of all the column vectors making up the section of the design matrix for one factor with the column vectors making up the portion of the design matrix for the second must be zero. If all factors in the model are orthogonal, then the \( X_i \) = \( X_i \) block diagonal. A block-diagonal \( X_i \) matrix is composed of square factor submatrices (degrees of freedom x degrees of freedom) along the diagonal with all off-diagonal elements not in one of the square factor submatrices equaling zero. A property of block-diagonal matrices is that the inverse can be calculated by inverting each block separately and replacing the original blocks in the full \( X_i \) matrix by the inverted block. Because the blocks can be inverted separately and all other off-diagonal elements of the inverse are zero, the effects for factors which are orthogonal to all other factors may be estimated separately. i.e. there are no functions of other sum-to-zero factors in the sum-to-zero estimates.

2.3.2 Mean, Block, GCA, and SCA Parameters

All parameters are estimated simultaneously by horizontally concatenating the mean, block, GCA, and SCA matrices to create \( X_i \). Equation 3 is again utilized to solve the system of equations. The b vector for the four parent example is an estimate of \( \mathbf{b} \), Figure 2. Again, one parameter is estimated for each column in the \( X_i \) matrix and all parameter estimates are not present. line combinations of the parameter estimates in the b vector. So \( \mathbf{b} \) is equal to \( \sum_{i=1}^{n} \mathbf{b}_i \) and gca = equal to \( \sum_{i=1}^{n} \mathbf{gca} \). The linear combinations for SCA effects can be obtained by reading along the row of the SCA submatrix associated with the observation containing the parameter, i.e. in Figure 2 the observation \( y_{12i} \) contains the effect \( \text{sc}_{12} \) which is estimated as the linear combination \( \mathbf{sc}_{12} - \mathbf{sc}_{13} \).

This completes the estimation of fixed effect parameters from a data set which is balanced on a plot-mean basis. Since field data sets with such completeness are a rarity in forestry applications, the next step is OLS analysis for various types of data imbalance. Calculations of solutions based on a complete data set and simulated data sets with common types of imbalance are demonstrated in numerical examples.

Numerical Examples

The data set analyzed in the numerical examples is from a five-year-old, six-parent half-dialed slash pine (Pinus elliottii var. elliottii Englem.) progeny test planted on a single site in four complete blocks. Each cross is represented by a five-tree row plot within each block. Total height in meters and diameter at breast height (dbh in centimeters) are the traits selected for analysis. The data set is presented in Table 1 so that the reader may reconstruct the analysis and compare answers with the examples. The numbers 1 through 6 were arbitrarily assigned to the parents for analysis. Because of unequal survival within plots, plot means are used as the unit of observation.
3.1 Balanced Data (Plot-mean Basis)

The sum-to-zero design matrix for the balanced data set has (4 blocks) x (15 crosses) = 60 rows (which equals the number of observations in y) and has the following columns: one column for $\mu$, three columns for blocks (b-1), five columns for GCA (p-1), and nine columns for SCA (15 crosses - 5 - 1) for a total of 18 columns. With 60 plot means (degrees of freedom) and 18 degrees of freedom in the model, subtracting 18 from 60 yields 42 degrees of freedom for error which matches the degrees of freedom for cross by block interaction, thus verifying that degrees of freedom concur with the number of columns in the sum-to-zero design matrix.

To illustrate the principle of orthogonality in the balanced case, the $X'X_b$ and $(X_b'X_b)^{-1}$ matrices may be printed to show that they are block diagonal. In further illustration, the effects within a factor may also be estimated without any other factors in the design matrix and compared to the estimates from the full design matrix.

The vectors of parameter estimates for height and dbh (Table 2) were calculated from the same $X_b$ matrix because height and dbh measurements were taken on the same trees. In other words, if a height measurement was taken on a tree, a dbh measurement was also taken, so the design matrices are equivalent.
### Table 2. — Numerical results for examples of data imbalance using the OLS techniques presented in the text.

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Balanced</th>
<th>Missing Plot</th>
<th>Missing Cross</th>
<th>Five Missing Crosses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DBH</td>
<td>DBH</td>
<td>DBH</td>
<td>DBH</td>
</tr>
<tr>
<td>( \mu  )</td>
<td>2.5830</td>
<td>3.362</td>
<td>2.5787</td>
<td>3.346</td>
</tr>
<tr>
<td>B1</td>
<td>0.1203</td>
<td>0.292</td>
<td>0.1074</td>
<td>0.245</td>
</tr>
<tr>
<td>B2</td>
<td>0.5230</td>
<td>0.976</td>
<td>0.5274</td>
<td>0.992</td>
</tr>
<tr>
<td>B3</td>
<td>0.1264</td>
<td>0.205</td>
<td>0.1308</td>
<td>0.220</td>
</tr>
<tr>
<td>GCA1</td>
<td>0.0706</td>
<td>0.144</td>
<td>0.0760</td>
<td>0.163</td>
</tr>
<tr>
<td>GCA2</td>
<td>-1.0777</td>
<td>-0.180</td>
<td>-1.1186</td>
<td>-0.220</td>
</tr>
<tr>
<td>GCA3</td>
<td>-1.1316</td>
<td>-0.947</td>
<td>-1.1426</td>
<td>-0.366</td>
</tr>
<tr>
<td>GCA4</td>
<td>0.2489</td>
<td>0.398</td>
<td>0.2544</td>
<td>0.417</td>
</tr>
<tr>
<td>GCA5</td>
<td>0.1265</td>
<td>0.489</td>
<td>0.1320</td>
<td>0.509</td>
</tr>
<tr>
<td>SCA12</td>
<td>0.0665</td>
<td>0.172</td>
<td>0.0763</td>
<td>0.208</td>
</tr>
<tr>
<td>SCA13</td>
<td>-0.3374</td>
<td>-0.628</td>
<td>-0.3277</td>
<td>-0.592</td>
</tr>
<tr>
<td>SCA14</td>
<td>-0.0484</td>
<td>-0.128</td>
<td>-0.0550</td>
<td>-0.152</td>
</tr>
<tr>
<td>SCA15</td>
<td>0.0766</td>
<td>0.126</td>
<td>0.0700</td>
<td>0.102</td>
</tr>
<tr>
<td>SCA22</td>
<td>0.3995</td>
<td>0.912</td>
<td>0.3600</td>
<td>0.771</td>
</tr>
<tr>
<td>SCA24</td>
<td>0.1528</td>
<td>0.289</td>
<td>0.1627</td>
<td>0.324</td>
</tr>
<tr>
<td>SCA25</td>
<td>-0.3158</td>
<td>-0.706</td>
<td>-0.3084</td>
<td>-0.670</td>
</tr>
<tr>
<td>SCA26</td>
<td>-0.0592</td>
<td>0.164</td>
<td>-0.0493</td>
<td>0.129</td>
</tr>
<tr>
<td>SCA35</td>
<td>0.3580</td>
<td>0.677</td>
<td>0.3679</td>
<td>0.712</td>
</tr>
</tbody>
</table>

**a)** (numerical examples are for height)

\[
\begin{align*}
D_1 &= -2 \beta_1 b_1 = 0.7087; \\
GCA_1 &= -2 \beta_2 b_2 = -0.2087; \\
GCA_{jk} &= -2 \beta_{jk} b_{jk} \text{ for } j \neq k = p \text{ and } p = 1, 2, 3 \text{ then } GCA_{jk} = 0.2425; \\
SCA_{ij} &= -0.3052, \text{ and } SCA_{ij} = -0.3052; \\
SCA_{i} &= -2 \beta_i b_i \text{ and } SCA_{i} = -0.3052; \\
e &= \text{ independently estimated } SCA_1, \ldots, 9; \\
SCA_{10} &= \sum_{i=2}^{9} SCA_i + SCA_{15} + SCA_{16} + SCA_{17} + SCA_{18} + SCA_{19} = 0.2446; \text{ and } \\
SCA_{20} &= \sum_{i=2}^{9} SCA_i + SCA_{14} + SCA_{15} + SCA_{16} + SCA_{17} + SCA_{18} + SCA_{19} = 0.1737.
\end{align*}
\]

**b)** (where the linear combinations for parameter estimates are identical to the balanced example)

\[
\begin{align*}
\text{where } SCA_{i} &= -2 \beta_i b_i \text{ for } j \neq k = p \text{ and } p = 1 \text{ to } 3; \\
SCA_i &= -2 \beta_i b_i \text{ and } SCA_i = -0.3052; \\
SCA_{ij} &= SCA_{i2} + SCA_{i3} + SCA_{i1} + SCA_{i4} + SCA_{i5} + SCA_{i6}; \text{ and } \\
SCA_{ij} &= SCA_{i2} + SCA_{i3} + SCA_{i1} + SCA_{i4} + SCA_{i5} + SCA_{i6}.
\end{align*}
\]

**c)** (where \( SCA_{i} = -2 \beta_i b_i \text{ for } j \neq k = p \text{ and } p = 1 \text{ to } 3; \text{ SCAs are independently estimated SCAs' } 1, \ldots, 9; \))

\[
\begin{align*}
\text{SCA} &= SCA_{i} + SCA_{i} + SCA_{i} + SCA_{i} + SCA_{i}; \text{ and } \\
SCA &= SCA_{i} + SCA_{i} + SCA_{i} + SCA_{i} + SCA_{i}.
\end{align*}
\]

**d)** (where \( SCA_{i} = -2 \beta_i b_i \text{ for } j \neq k = p \text{ and } p = 1 \text{ to } 3; \text{ SCAs are independently estimated SCAs' } 1, \ldots, 9; \))

\[
\begin{align*}
\text{SCAs are the negative of the sum of the four independently estimated SCAs' } 1, \ldots, 9; \\
\text{SCAs are the negative of the sum of the four independently estimated SCAs' } 1, \ldots, 9; \\
\text{SCAs are the negative of the sum of the four independently estimated SCAs' } 1, \ldots, 9; \\
\text{SCAs are the negative of the sum of the four independently estimated SCAs' } 1, \ldots, 9.
\end{align*}
\]

### 3.2 Missing Plot

To illustrate the problem of a missing plot, the cross, parent two by parent three, was arbitrarily deleted in block one (as if observation \( y_{123} \) were missing). This deletion prompts adjustments to factor matrices in order to analyze the new data set. The new vector of observations \( y \) now has 59 rows. This necessitates deletion of the row of the design matrix \( (X_y) \) in block 1 which would have been associated with cross 2 x 3. This is the only matrix alteration required for the analysis. Thus, the resultant \( X_y \) matrix has 60 — 1 = 59 rows and 18 columns. With 59 means in \( y \) and 18 columns in \( X_y \), the degrees of freedom for error is 41.

Comparisons between results of the analysis (Table 2) of the full data set and the data set missing observation \( y_{123} \) reveal that for this case the estimates of parameters have been relatively unaffected by the imbalance (magnitudes of GCA's changed only slightly and rankings by GCA were unaffected).

### 3.3 Missing Cross

Another common form of imbalance in diallel data sets, the missing cross, is examined through arbitrary deletion of the 2 x 3 cross from all blocks, i. e. \( y_{123}, y_{136}, y_{235}, y_{236} \) are missing in the data vector. This type of imbalance is representative of a particular cross that could not be made and is therefore missing from all blocks. The matrix manipulations required for this analysis are again presented by factor. For appropriate SCA restrictions, the data vector and design matrix should be ordered so that the parent with no missing crosses. Since the labeling of a parent as parent \( p \) is entirely subjective, any parent with all crosses may be designated as parent \( p \). The previous labeling directions are necessary since we generate the SCA submatrix as horizontal direct products of the columns of the GCA submatrix; and to account for missing crosses, the horizontal direct product for each particular missing parental combinations are not calculated which sets the missing SCA's to zero. If there is a cross missing from those of the \( p \)th parent, we cannot account for the missing cross with this technique (Searle, 1987, page 479).

For the mean, block, and GCA submatrices, the adjustment for the missing cross dictates deleting the rows in the submatrices which would have corresponded to the \( y_{123} \)
observations. The SCA submatrix must be reformed since a degree of freedom for SCA and hence a column of the submatrix has been lost. The SCA submatrix is reconstituted from the GCA horizontal direct products (remembering that one cross, 2x3, no longer exists and therefore that product GCA x GCA is inappropriate). Dropping the column for SCA to zero (Searle, 1987) so that the remaining SCA's will sum-to-zero. After that, the reformation is according to the established pattern. With one missing cross there are now 56 observations and hence 56 degrees of freedom available. The columns of the $X_i$ matrix are now: one for the mean, three for block, five for GCA, and eight for SCA for a total of 17 columns. The remaining degrees of freedom for error is 39, matching the correct degrees of freedom ((14−1) x (4−1) = 39).

For the missing cross example $\hat{x}$ is no longer equivalent to the mean of the plot means since $\mu = 2.5386$ and $\sum_{i=1}^{10} y_{ij}/N = 2.5715$ where N = 56 (number of plot means).

This is the result of GCA effects which are no longer orthogonal to the mean. Check the $X_i X_i$ matrix or try estimating factors separately and compare to the estimates when all factors are included in $X_i$.

If formulae for balanced data (Becker, 1975; Falconer, 1981; and Hallauer and Miranda, 1981) are applied to unbalanced data (plot-mean basis) estimates of parameters are no longer appropriate because factors in the model are no longer independent (orthogonal). Applying Becker's formula which uses totals of cross means for a site ($\gamma_{ij}$) to the missing cross example yields: $g_{ca} = 0.0299$, $g_{ca} = -0.5649$, $g_{ca} = -0.5888$, $g_{ca} = 0.4665$, $g_{ca} = 0.3552$, and $g_{ca} = 0.0219$. These answers are very different in magnitude from those in table 2 for this example and $g_{ca}$ also has a different sign. Employing these formulae in the analysis of unbalanced data is analogous to matrix estimation of GCA's without the other factors in the model which is inappropriate.

3.4 Several Missing Crosses

The concluding example (Table 2) is a drastically unbalanced data set resulting from the arbitrary deletion of five crosses (1 x 2, 1 x 3, 2 x 3, 3 x 3, and 4 x 3). The matrix manipulation for this example is an extension of the previous one cross deletion example. Raws corresponding to $y_{123}$, $y_{132}$, $y_{223}$, $y_{232}$, and $y_{332}$ are deleted from the mean, block and GCA submatrices for all blocks. The SCA matrix (now 4 columns = 10 crosses - 5 - 1 - 4 degrees of freedom) is again reformed with only the relevant products of the GCA columns. Counting degrees of freedom (columns of the sum-to-zero design matrix), the means has one, block has three, GCA has five, and SCA has four degrees of freedom for a total of 13. Error has $(4-1)(10-7) = 27$ degrees of freedom. Totalizing degrees of freedom for modeled effects and error yields 40 which equals the number of plot means.

In increasingly unbalanced cases (Table 2), the spread among the GCA estimates tends to increase with increasing imbalance (loss of information). This is a general feature of OLS analyses and the basis for the feature is that the spread among the GCA estimates is due to both the innate spread due to additive genetics effects as well as the error in estimation of the GCA's. When there is less information, GCA estimates tend to be more widely spread due to the increase in the error variance associated with their estimation. This feature has been noted (White and Hodge, 1989, page 54) as the tendency to pick as parental winners individuals in a breeding program which are the most poorly tested.

Discussion

After developing the OLS analysis and describing the inherent assumptions of the analysis, there are four important factors to consider in the interpretation of sum-to-zero OLS solutions: (1) the lack of uniqueness of the parameter estimates; (2) the weights given to plot means ($\gamma_{ij}$) and in turn site means ($\gamma_{ij}$) for crosses in data sets with missing crosses in parameter estimation; (3) the arbitrary nature of using a dialled mean (perforce a narrow genetic base) as the mean about which the GCA's sum-to-zero; and (4) the assumption that the variance-covariance matrix for the observations (V) is $10^2$.

4.1 Uniqueness of Estimates

Sum-to-zero restrictions furnishes what would appear to be unique estimates of the individual parameters, e.g. GCA, when, in fact, these individual parameters are not estimable (Graybill, 1976; Freund and Little, 1981; and Milliken and Johnson, 1984). The lack of estimability is again analogous to attempting to solve a set of equations in n unknowns with t equations where n is greater than t. Therefore, an infinite number of solutions exist for $\beta$.

There are quantities in this system of equations that are unique (estimable), i.e. the estimate is invariant regardless of the restriction (sum-to-zero or set-to-zero) or generalized inverse (no restrictions) used (Milliken and Johnson, 1984) and the estimable functions include sum-to-zero GCA and SCA estimates since they are linear combinations of the observations; but, these estimable quantities do not estimate the individual parametric GCA's and SCA’s of the overparameterized model (Equation 4) since there is no unique solution for those parameters.

4.2 Weighting of Plot Means and Cross Means in Estimating Parameters

With at least one measurement tree in each plot and with plot means as the unit of observation, use of the matrix approach produces the same results as the basic formulae. The weight placed on each plot mean in the estimation of a parameter can be determined by calculating $X_i'X_i^{-1}X_i'$ which can be viewed as a matrix of weights W so that $equation 3$ can be written as $b = Wy$. The matrix W has these dimensions: the number of rows equals the number of parameters in $\beta_c$ and the number of columns equals the number of plot means in $y$. The ith row of the W contains the weights applied to y to estimate the ith parameter in b (6). In the discussion which follows $g_{ca}$ is utilized as $\beta_c$.

If there are no missing plots, the cross mean in every block ($\gamma_{ij}$) has the same weighting and weights can be combined across blocks to yield the weight on the overall cross mean ($\gamma_{ik}$). It can be shown that for the balanced numerical example $g_{ca}$ is calculated by weighting the overall cross means containing parent 1 by 1/6 and weighting all overall cross means not containing parent 1 by $-1/12$. Figure 4 (above the diagonal) demonstrates the weightings on the overall cross means for the balanced numerical example as well as the marginal weighting on the GCA parameters. These marginal weightings are obtained by summing along a row and/or column as one would to obtain the marginal totals for a parent (Becker,
![Figure 4](image)

Figure 4. — Weights on overall cross means ($\bar{y}_{24}$) for the three numerical examples for estimation of GCA$_1$. The weights for the balanced example (above the diagonal) are presented in both fractional and decimal form. The weights for the one-cross missing and the five-crosses missing are presented as the upper number and lower number, respectively, in cells below the diagonal. The marginal weights on GCA parameters (right margin) do not change although cells are missing.

1975). One feature of sum-to-zero solutions is that these marginal weightings will be maintained no matter the imbalance due to missing crosses, as will be seen by considering the numerical examples for a missing cross (Figure 4 below the diagonal, upper number) and five missing crosses (Figure 4 below the diagonal, lower number). The marginal weights have remained the same as in the balanced case while the weights on the cross means differ among the crosses containing parent 1 and also among the crosses not containing parent 1. In the five missing crosses example, crosses $\bar{y}_{24}$ and $\bar{y}_{25}$ even receive a positive weighting where in the prior examples they had negative weighting.

The expected value in all three examples is GCA$_{1s}$ (for sum-to-zero) despite the apparently nonsensical weightings to cross means with missing crosses; however, the evaluation of the estimates in terms of the original model changes with each new combination of missing cells, i.e.

$$
GCA_{1s} = \frac{(p-1)}{p}GCA_1 - \frac{1}{p}\Sigma_{j=2}^p GCA_j + (1/p)\Sigma_{k=2}^p SCA_{1k} - \frac{2}{(p(p-2)))\Sigma_{j=2}^p \Sigma_{k=2}^p SCA_{jk};
$$

$$
E\{\text{diadel mean}\} = \mu + (\Sigma_{i=1}^p B_i)/b + (2/p)\Sigma_{i=1}^p GCA_i + (2/(p(p-1)))\Sigma_{j=2}^p \Sigma_{k=2}^p SCA_{jk}; \quad \text{and}
$$

$$
E\{\text{checklet mean}\} = \mu + (\Sigma_{i=1}^p B_i)/b + r;
$$

The function used to properly locate GCA$_{rel}$ (the subscript rel denotes the relocated GCA$_{1s}$) is $gca_{rel} = gca_{1s} + (1/2)$ (diadel mean — checklet mean). The expectation of $gca_{rel}$ with negligible SCA is $GCA_{rel} = GCA_1 - r/2$; and since breeding value equals twice GCA, $BV_{rel} = BV_i - r$. If SCA is non-negligible then the expectation is:

$$
GCA_{rel} = GCA_1 + (1/(p-1))\Sigma_{p=2}^p SCA_{1k} - (1/((p-1)(p-2)))\Sigma_{j=2}^p \Sigma_{k=2}^p SCA_{jk} - r/2.
$$

271
In either case the function provides a reasonable manner by which GCA estimates from disconnected diallels are centered at the same location on a number scale and are then comparable.

4.4 Variance and Covariance of Plot Means

The variances of plot means with unequal numbers of trees per plot are by definition unequal, i.e., \( \text{Var}(Y_{ij}) = \sigma^2_g + \sigma^2_p/n_{ij} \) where \( \sigma^2_g \) is plot variance, \( \sigma^2_p \) is the within plot variance and \( n_{ij} \) is the number of observations per plot. Also, if blocks were considered random, there would be additional source of variance for plot means due to blocks (as well as a covariance between plot means in the same block) and this could be incorporated into the V matrix with \( \text{Var}(Y_{ij}) = \sigma^2_g + \sigma^2_p + \sigma^2_b/n_{ij} \). Since the variances of the means in the observation vector are not equal and there is a covariance between the means if blocks are being considered random, best linear unbiased estimates (BLUE) would be secured by weighting each mean by its true associated variance (Searle, 1987, page 316). This is the generalized least squares (GLS) approach as:

\[
b = (X^*V^{-1}X)^{-1}X^*V^{-1}y
\]

The GLS approach relaxes the OLS assumptions of equal variance of and no covariance between the observations (plot means) while still treating genetic parameters as fixed effects. The entries along the diagonal of the V matrix are the variances of the plot means (Var(Y_{ij})) in the same order as means in data vector. The off-diagonal elements of V would be either 0 or \( \sigma^2_b \) (the variance due to the random variable block) for elements corresponding to observations in the same block. BLUE requires exact knowledge of V; if estimates of \( \sigma^2_g \), \( \sigma^2_p \), and \( \sigma^2_b \) are utilized in the V matrix, estimable functions of \( \beta \) approximate BLUE.

The OLS assumption that SCA and GCA are fixed effects can also be relaxed to allow for covariances due to genetic relatedness. In particular, the information that means are from the same half- or full-sib family could be included in the V matrix. Relaxation of the zero covariance assumption implies that GCA and SCA are random variables. If GCA and SCA are treated as random variables, then the application of best linear prediction (BLP) or best linear unbiased prediction (BLUP) to the problem would be more appropriate (Wright and Hodge, 1989, page 64). The treatment of the genetic parameters as random variables is consistent with that used in estimating genetic correlations and heritabilities. The V matrix of such an application would include, in addition to the features of the GLS V matrix, the covariance between full-sib or half-sib families added to the off-diagonal elements in V, i. e. if the first and second plot means in the data vector had a covariance due to relationship, then that covariance is inserted twice in the V matrix. The covariance would appear as the second element in the first row and the first element in the second row of V (V is a symmetric matrix). Also the diagonal elements of V would increase by \( 2\sigma^2_{ca} \) (the variance due to treating GCA as a random variable) + \( \sigma^2_{ca} \) (the variance due to treating SCA as a random variable).

4.5 Comparison of Prediction and Estimation Methodologies

Which methodology (OLS, GLS, BLP, or BLUP) to apply to individual data bases is somewhat a subjective decision. The decision can be based both on the computational or conceptual complexity of the method and the magnitude of the data base with which the analyst is working. To aid in this decision, this discussion highlights the differences in the inherent properties and assumptions of the techniques.

For all practical purposes the answers from the four techniques will never be equal; however, there are two caveats. First, OLS estimates equal GLS estimates if all the cell means are known with the same precision (variance), Searle, 1987, page 490). Otherwise, GLS discounts the means that are known with less precision in the calculations and different estimates result. The second caveat is if the amount of data is infinite, i. e. all cross means are known without error, then all four techniques are equivalent (White and Hodge, 1989, pages 104 to 106). In all other cases BLP and BLUP shrink predictions toward the location parameter(s) and produce predictions which are different from OLS or GLS estimates even with balanced data. During calculations GLS, BLP, and BLUP place less weight on observations known with less precision, which is intuitively pleasing.

With OLS and GLS forest geneticists treat GCA’s and SCA’s as fixed effects for estimation and then as random variables for genetic correlations and heritabilities. BLP and BLUP provide a consistent treatment of GCA’s and SCA’s as random variables while differing in their assumptions about location parameters (fixed effects). In BLP fixed effects are assumed known without error (although they are usually estimated from the data) while with BLUP fixed effects are estimated using GLS. BLP and BLUP techniques also contain the assumption that the variance-covariance matrix of the observations is known without error (most often variances must be estimated). In many BLUP applications (Henderson, 1974), mixed model equations are utilized interactively to estimate fixed effects and to predict random variables from a data set. A BLUP treatment of fixed effects allows any connectedness between experiments to be utilized in the estimation of the fixed effects. This provides an intuitive advantage of BLUP over BLP in experimentation where connectedness among genetic experiments is available or where the data are so unbalanced that treating the fixed effects as known is less desirable than a GLS estimate of the fixed effects.

An ordering of computational complexity and conceptual complexity from least to most complex of the four methods is OLS, GLS, BLP and BLUP. The latter three methods require the estimation of the variance-covariance matrix of the observations either separately (a priori) or iteratively with the fixed effects. Precise estimation of the variance-covariance matrix for observations requires a great number of observations and the precision of GLS, BLP and BLUP estimations or predictions is affected by the error of estimation of the components of V.

Selection of a method can then be based on weighing the computational complexity and size of the available data base against the advantages offered by each method. Thus, if complexity of the computational problem is of paramount concern, the analyst necessarily would choose OLS. With a small data base (one that does not allow reasonable estimates of variances), the analyst would again choose OLS. With a large data base and no qualms with computational complexity, the analyst can choose between BLP and BLUP based on whether there is sufficient con-
nectedness or imbalance among the experiments to make BLUP advantageous.

Conclusions

Methods of solving for GCA and SCA estimates for balanced (plot-mean basis) and unbalanced data have been presented along with the inherent assumptions of the analysis. The use of plot means and the matrix equations will produce sum-to-zero OLS estimates for GCA and SCA for all types of imbalance. Formulæ in the literature which yield OLS solutions for balanced data can yield misleading solutions for unbalanced data because of the loss of orthogonality and also weightings on site means for crosses (or totals) are constants.

GCA's and SCA's obtained through sum-to-zero restrictions are not truly estimates of parametric population GCA's and SCA's. There are an infinite number of solutions for GCA's and SCA's from the system of equations as a result of the overparameterized linear model. Yet, if the only comparisons of interest are among the specific parents on a particular site, then the estimates calculated by sum-to-zero restrictions are appropriate. Checklists may be used to provide comparability among estimates derived from disconnected sets.

Having discussed the innate mathematical features of OLS analysis, knowledge of these features should help the data analyst decide if OLS is the most desirable technique for the data at hand. It may be desirable to relax OLS assumptions, which are in all likelihood invalid for the variance-covariance matrix of the observations. This could lead to GLS, BLP or BLUP as better alternatives.

References


Segregation and Linkage of Allozymes in Seed Tissues of the Hybrid Greek Fir Abies borisii regis Mattfeld

By B. FADY1) and M. T. CONKLIN2)

(Received 30th December 1981)

Summary

Seed tissues (haploid megagametophyte and diploid embryo tissues) of Abies borisii regis were used in starch gel electrophoresis to study inheritance and linkage of allozyme variants. The 10 enzyme systems studied are coded by a minimum of 15 isozyme loci. All loci code allozymes in both megagametophyte and embryo tissues. Mendelian segregation ratios were found for all enzyme systems except ACO and 6-PGD where distortion was observed. Segregation distortion could also exist in other enzyme systems (LAP, PG12, MNR1, MNR2). Evidence of total linkage is provided for one marker gene of Abies borisii regis (A1) that has never been tested before in conifers and tight linkage for another pair of loci (MNR1/PG11).

Key words: Abies borisii regis, allozymes, inheritance, linkage, electrophoresis, seed tissues.

Silvae Genetics 41, 4-5 (1992)