9-5-2007

Technical Note: Calculation of standard errors of estimates of genetic parameters with the multiple-trait derivative-free restricted maximal likelihood programs

Stephen D. Kachman
University of Nebraska-Lincoln, steve.kachman@unl.edu

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

Technical Note: Calculation of standard errors of estimates of genetic parameters with the multiple-trait derivative-free restricted maximal likelihood programs

S. D. Kachman*[1] and L. D. Van Vleck†[2]

*Department of Statistics, University of Nebraska, Lincoln 68583-0963; and †USDA, ARS, Roman L. Hruska US Meat Animal Research Center, Lincoln, NE 68583-0908

ABSTRACT: The multiple-trait derivative-free REML set of programs was written to handle partially missing data for multiple-trait analyses as well as single-trait models. Standard errors of genetic parameters were reported for univariate models and for multiple-trait analyses only when all traits were measured on animals with records. In addition to estimating (co)variance components for multiple-trait models with partially missing data, this paper shows how the multiple-trait derivative-free REML set of programs can also estimate SE by augmenting the data file when not all animals have all traits measured. Although the standard practice has been to eliminate records with partially missing data, that practice uses only a subset of the available data. In some situations, the elimination of partial records can result in elimination of all the records, such as one trait measured in one environment and a second trait measured in a different environment. An alternative approach requiring minor modifications of the original data and model was developed that provides estimates of the SE using an augmented data set that gives the same residual log likelihood as the original data for multiple-trait analyses when not all traits are measured. Because the same residual vector is used for the original data and the augmented data, the resulting REML estimators along with their sampling properties are identical for the original and augmented data, so that SE for estimates of genetic parameters can be calculated.

Key words: average information matrix, genetic parameter, restricted maximal likelihood, standard error

J. Anim. Sci. 2007. 85:2375–2381
doi:10.2527/jas.2007-0202

INTRODUCTION

The multiple-trait derivative-free REML (**MTDFREML**; Boldman et al., 1995) set of programs was written to handle single-trait models and multiple-trait models with partially missing data in an expedient manner. When estimating (co)variance components and genetic parameters for multiple-trait models, the programs have not been able to estimate SE of those estimates for multiple-trait models when animals with observations do not have all traits measured.

When some traits were not recorded on some units (e.g., animals), the standard approach used has been to discard incompletely recorded units. In the worst case, when males have one trait measured and females have another trait measured, there are no animals that have both traits measured. A similar case is observed for genotype by environment interaction, where some animals have records in one environment and other (some related) animals have records in another environment.

Although the program uses a derivative-free algorithm (simplex) to minimize -2logL|y, where L|y is the likelihood (**L**) given the data (**y**), the asymptotic SE for single-trait analyses and multiple-trait analyses with all traits measured are based on the average information matrix (**AIM**; Johnson and Thompson, 1995) as implemented by Dodenhoff et al. (1998).

MATERIALS AND METHODS

Animal Care and Use Committee approval was not obtained for this study because no animals were used.

The limitation that all animals with observations have all traits measured can be overcome without any changes in the set of MTDFREML programs. A model-based procedure will be described to accomplish that goal, which makes use of properties of the mixed model

[1]Current address: Dep. Statistics, 340 Hardin Hall, University of Nebraska, Lincoln 68583-0963.
[2]Corresponding author: lvanvleck@unlnotes.unl.edu
Received April 5, 2007.
Accepted July 2, 2007.

equations. The relatively trivial changes in the data file and model will be described.

Each missing observation for a trait is assigned a unique level of a dummy factor associated with that trait. Each missing observation can be assigned the same arbitrary pseudo-value. The program accepts that value as a real observation. The result is that the program handles the analysis as if all traits were observed for each animal with records. In essence, the estimated residual element associated with the pseudo-observation is a structural zero because the pseudo-observation is alone in a unique level of the newly created fixed factor. Because structural zeros in the estimated residual vector contribute nothing to the AIM, the resulting AIM is the same as if it was computed with a more complex algorithm. Similarly, L|y also is the same as if it was computed with missing observations ignored. (See Appendix for proof of equivalence.)

RESULTS

Revised Data File

The simplest case is when the data line for an animal that includes a missing measurement for a trait includes the levels of fixed factors that would have been associated with an actual measurement. (Often these would be the same as for another trait that has an observation.) As an example, consider the 2-trait analysis for the data file partially represented in Table 1, which is the data file used by K. Meyer (University of New England, Armidale, New South Wales, Australia, unpublished data) as an example with early versions of her DFREML program but now with some observations missing. The table presents data for the first 20 of 284 mice. Observations on the 2 traits are in the last 2 columns. A –99.00 represents a pseudo-value for a missing observation. Actual observations will be assigned level 1 for the dummy factor (MTDFREML does not accept 0 as a valid level). Each missing observation for a trait will be assigned a unique level for the dummy factor. An easy-to-program method is to assign level 2 for the first missing observation, level 3 for the next, etc. This assignment of levels would be done for each trait. Table 1 shows the result for the first 20 animals that includes all animals with missing observations. The first animal has trait 1 measured so field m1 has a 1 for the level, but the first animal does not have trait 2 measured so field m2 has a 2 as the first unique level. Animal 2 has both traits measured, so fields m1 and m2 are both 1. The third animal is missing trait 1 but has trait 2, so m1 is 2 and m2 is 1. Animal 8 is missing both traits and both would represent the second missing observation, so m1 and m2 will both be 3. The pattern for the remaining missing observations is obvious. For the first 20 animals shown for the data file, trait 1 has 3 missing observations with corresponding unique levels for m1 of 2, 3, and 4. Trait 2 has 2 missing observations with corresponding unique levels for m2 of 2 and 3.

For the analysis with MTDFREML, an additional fixed factor is used for each trait (fields m1 and m2). In this example, levels for gn (generation), sx (sex), and nℓ (number in litter) are the same for both traits and are present even for the eighth animal, which is missing both traits. A suitable strategy for handling the case when some of the levels for those 3 factors are missing will be described later. Note that the pseudo-values –99.00 will now be treated as actual observations in the analysis with MTDFREML.

For trait 1 alone with the full model including maternal effects, –2logL|y is 744.53 for the analysis with missing observations treated as missing and also is 744.53 for the analysis with –99.00 treated as observed measurements for trait 1 but with the fixed factor m1 added to the model. Estimates of variance components and genetic parameters as well as the average information matrix and asymptotic SE of the genetic parameters are also the same for both analyses. Similarly, estimates of estimable functions of the fixed effects are the same although the analysis with –99.00 as pseudo measurements also has estimates for levels 1,..., 6 of m1.

Choice of a Pseudo-Value for a Missing Observation

When 0.00 is substituted for –99.00 to denote a missing measurement, the –2logL and estimates of (co)variance components and genetic parameters with SE are the same as when –99.00 is used. Although estimates of levels for some fixed factors can be different, estimates of estimable functions will be the same. Estimates for levels of m1 are greater by 99.00 when 0.00 rather than –99.00 is substituted for missing values. That difference forces the residual effects for missing observations to be zero in both cases. The solution for level 1 of m1 was constrained to zero for both analyses (when –99.00 and 0.00 were substituted for missing values). As had to happen, similar analyses of trait 2 resulted in the same results.

Different pseudo-values could be assigned to different missing observations because the key step is to have only 1 missing observation within a level of m1 or m2. Assignment of the same pseudo-value for each missing observation is usually the easiest way to modify the data file.

Assignment of Levels of Other Factors for Missing Observations

Often a missing observation will not have levels associated with fixed factors as will actual observations. The question then is what levels of these fixed factors should be assigned when pseudo-observations are analyzed within unique levels of factor m1 (and/or m2). Three options for an analysis of a single trait were tried: 1) a unique level different from those associated with actual observations was used (in the example, level 10 was used for generation, sex, and number in the litter when actual levels were 1 to 3, 1 to 2, and 1 to 7); 2)

Table 1. Data file for the first 20 of 284 mice, modified to create levels for a dummy factor (m1) for trait 1 and a dummy factor (m2) for trait 2 which will allow analysis as if the missing value indicators (–99.00) are actual observations for m1 and m2[1,2]

An ID	Sire	Dam	gn	sx	nl	lit	m1	m2	n1	t1	t2
20101	11012	10101	1	1	4	1	1	2	4.	22.50	–99.00
20102	11012	10101	1	1	4	1	1	1	4.	22.60	52.40
20103	11012	10101	1	1	4	1	2	1	4.	–99.00	61.10
20104	11012	10101	1	1	4	1	1	1	4.	23.00	57.90
20112	11012	10101	1	2	4	1	1	1	4.	24.60	69.30
20113	11012	10101	1	2	4	1	1	1	4.	26.40	66.40
20114	11012	10101	1	2	4	1	1	1	4.	24.10	61.60
20115	11012	10101	1	2	4	1	3	3	4.	–99.00	–99.00
20302	10614	10506	1	1	7	2	1	1	7.	23.60	60.10
20303	10614	10506	1	1	7	2	1	1	7.	24.20	60.80
20304	10614	10506	1	1	7	2	1	1	7.	22.50	62.80
20306	10614	10506	1	1	7	2	1	1	7.	22.30	59.40
20312	10614	10506	1	2	7	2	1	1	7.	23.70	66.90
20313	10614	10506	1	2	7	2	1	1	7.	27.50	72.90
20314	10614	10506	1	2	7	2	1	1	7.	17.30	58.40
20401	10813	10701	1	1	2	3	1	1	2.	20.30	58.10
20402	10813	10701	1	1	2	3	4	1	2.	–99.00	58.70
20403	10813	10701	1	1	2	3	1	1	2.	21.90	53.20
20404	10813	10701	1	1	2	3	1	1	2.	22.10	57.10
20413	10813	10701	1	2	2	3	1	1	2.	28.80	62.50

[1]Level 1 is for the actual records. Other levels for m1 and m2 are unique for each missing observation within trait (2,..., 6 for trait 1 and 2,..., 9 for trait 2 for the augmented data set).

[2]gn = generation; sx = sex; nl = number in litter; m1 = dummy factor for trait 1; m2 = dummy factor for trait 2; n1 = number in litter if used as a covariate; t1 = trait 1; t2 = trait 2.

level 1 was assigned to gn, sx, and nl when a pseudo-observation was analyzed as a real observation; and 3) the last level of gn (3), sx (2), and nl (7) was assigned. As expected, results were the same as those described before for –2logL|y, estimates of variance components and genetic parameters, and estimable functions of estimates of fixed effects.

From a practical standpoint, assigning a different unique level from those for actual observations for each fixed factor would seem preferable. The output file from MTDFPREP, MTDF66, will then provide means for each level of each factor for checking. This combines both the pseudo-observations and actual observations (see Table 2). The means for levels with actual observations would be actual means and means for the different level would be the missing value (pseudo-values of 0.00 or –99.00, etc.). The overall mean and unadjusted SD in MTDF66, however, would be calculated from a mixture of actual and pseudo-observations. Note that the mean for level 1 of m1 (missing 1) is the unadjusted mean for actual measurements.

The proceeding analyses were exploratory as SE have been available for single-trait analyses even without the trick of assigning pseudo-values for missing observations nested within unique levels of another fixed factor. The real strength is in the enhanced capability to estimate SE of genetic correlations with partially missing data.

Two-Trait Analyses

Two-trait analyses were also done with the preceding ways of handling missing observations. As before, –2logL, estimates of variance components and genetic parameters, and estimates of estimable functions of fixed effects were the same when missing observations were excluded or included as pseudo-observations assigned unique levels of fixed factor m1 for trait 1 and fixed factor m2 for trait 2 for the case when levels of the other fixed factors were already available.

When levels of the other fixed factors are not available for missing observations, then modifications of the fields for fixed factors are needed. The options could be as described earlier, taking care not to modify the levels for a field shared by a trait with real measures when modifying levels of the same field for a pseudo measurement of another trait. An easy solution is to have separate fields for each trait. In contrast to the usual analysis that would include fixed effects for gn, sx, and nl for trait 1 and trait 2, the augmented analysis has a model equation for trait 1 that has the fixed effects gn1, sx1, and nl1, whereas the model equation for trait 2 has gn2, sx2, and nl2. Table 3 shows how to accomplish this in the example. Two extra sets of 3 fields were added in addition to m1 and m2. The first line of Table 3 for missing trait 2 shows that the 3 extra fields for trait 1 are the same as the original 3 fields, but the 3 extra fields for trait 2 now contain the level 10. For data line 3, the 3 extra fields for trait 1 contain level 10, and the 3 extra fields for trait 2 are the original levels. For data line 8, both sets of 3 extra fields contain level 10. (Assigning first or last levels for each factor would also work). In all 3 cases, the –2logL|y, estimates of (co)variance components, genetic parameters, and estimable functions of fixed effects were the same as with excluding the missing observations from the anal-

Table 2. A condensed copy of the summary file (MTDF66) from running MTDFPREP for trait 1, with missing values (0.00) as actual observations using unique levels for each missing value for factor (m1) and one unique level (10) for each other factor (generation, sex, litter size) for missing values

No. of data lines in Unit 33	=	284
No. of integer variables per record	=	15
No. of real variables per record	=	3
No. of traits	=	1
No. of animals with valid records	=	284
No. of animals in A-1	=	329
Order of MME (before constraints)	=	721

Results for trait 1 – body weight (position 2)
No. of records = 284 (missing value: –999.0000; No. Missing = 0)

Trait	Mean	SD	CV	Min	Max
1	23.6989	4.55254	19.21	0.000	34.500

No. of fixed effects = 4

1, 4 levels for generation

Levels	Value	No.	%	Mean
1	1	88	30.99	23.878
2	2	84	29.58	23.063
3	3	107	37.68	25.158
4	10	5	1.76	0.000

2, 3 levels for sex

Levels	Value	No.	%	Mean
1	1	147	51.76	22.700
2	2	132	46.48	25.709
3	10	5	1.76	0.000

3, 8 levels for litter size

Levels	Value	No.	%	Mean
1	1	11	3.87	26.609
2	2	40	14.08	23.783
3	3	25	8.80	24.864
4	4	34	11.97	24.053
5	5	94	33.10	24.393
6	6	45	15.85	24.333
7	7	30	10.56	21.973
8	10	5	1.76	0.000

4, 6 levels for ml

Level	Value	No.	%	Mean
1	1	279	98.24	24.124
2	2	1	0.35	0.000
3	3	1	0.35	0.000
4	4	1	0.35	0.000
5	5	1	0.35	0.000
6	6	1	0.35	0.000

Fixed effects = 4

Trait	No.	Name	Position	Levels	Rows
1	1	generation	10	4	1 to 4
1	2	sex	11	3	5 to 7
1	3	litter size	12	8	8 to 15
1	4	m1	8	6	16 to 21
1	1	animal w/ full A-1	1	329	22 to 350
1	1	maternal genetic effect	3	329	351 to 679
1	1	maternal permanent env	3	42	680 to 721

ysis. With each of these 3 assignments of fixed levels along with dummy fixed factors m1 and m2, estimates of SE of estimates of genetic parameters were obtained as well as the same –2logL, estimates of genetic parameters, and estimates of estimable functions of levels of fixed factors.

Table 3. Data file for the first 20 of 284 mice for a 2-trait analysis modified to create levels for dummy fixed factors (m1 for trait 1 and m2 for trait 2) and also modified to create separate sets of fixed factors for traits 1 and 2, so that the actual levels are used for actual observations but with a unique level used for a missing value (0.00) analyzed as an actual observation[1]

								Trait 1				Trait 2					
An ID	Sire	Dam	gn	sx	nl	lit	m1	m2	Gn1	Sx1	Nl1	Gn2	Sx2	Lit2	Nl2	t1	t2
20101	11012	10101	1	1	4	1	1	2	1	1	4	10	10	10	4.	22.50	0.00
20102	11012	10101	1	1	4	1	1	1	1	1	4	1	1	4	4.	22.60	52.40
20103	11012	10101	1	1	4	1	2	1	10	10	10	1	1	4	4.	0.00	61.10
20104	11012	10101	1	1	4	1	1	1	1	1	4	1	1	4	4.	23.00	57.90
20112	11012	10101	1	2	4	1	1	1	1	2	4	1	2	4	4.	24.60	69.30
20113	11012	10101	1	2	4	1	1	1	1	2	4	1	2	4	4.	26.40	66.40
20114	11012	10101	1	2	4	1	1	1	1	2	4	1	2	4	4.	24.10	61.60
20115	11012	10101	1	2	4	1	3	3	10	10	10	10	10	10	4.	0.00	0.00
20302	10614	10506	1	1	7	2	1	1	1	1	7	1	1	7	7.	23.60	60.10
20303	10614	10506	1	1	7	2	1	1	1	1	7	1	1	7	7.	24.20	60.80
20304	10614	10506	1	1	7	2	1	1	1	1	7	1	1	7	7.	22.50	62.80
20306	10614	10506	1	1	7	2	1	1	1	1	7	1	1	7	7.	22.30	59.40
20312	10614	10506	1	2	7	2	1	1	1	2	7	1	2	7	7.	23.70	66.90
20313	10614	10506	1	2	7	2	1	1	1	2	7	1	2	7	7.	27.50	72.90
20314	10614	10506	1	2	7	2	1	1	1	2	7	1	2	7	7.	17.30	58.40
20401	10813	10701	1	1	2	3	1	1	1	1	2	1	1	2	2.	20.30	58.10
20402	10813	10701	1	1	2	3	4	1	10	10	10	1	1	2	2.	0.00	58.70
20403	10813	10701	1	1	2	3	1	1	1	1	2	1	1	2	2.	21.90	53.20
20404	10813	10701	1	1	2	3	1	1	1	1	2	1	1	2	2.	22.10	57.10
20413	10813	10701	1	2	2	3	1	1	1	2	2	1	2	2	2.	28.80	62.50

[1]In this example, level 10 for gn, sx, and nl, which are not actual levels, where gn = generation, sx = sex, nl = number in litter; m1 = dummy factor for trait 1; m2 = dummy factor for trait 2. Gn1 = generation, Sx1 = sex, and nl1 = number in litter for trait 1. Gn2 = generation, Sx2 = sex, and Nl2 = number in litter for trait 2. t1 = trait 1; t2 = trait 2.

A rigorous test of the method was to develop a file (Table 4) with trait 1 measured only for sex 1 and trait 2 only for sex 2. Thus, trait 2 would always be missing for animals of sex 1 and trait 1 would always be missing for animals of sex 2. As before, each missing observation (0.00) for trait 1 was assigned a unique level for factor m1, and each missing observation for trait 2 was assigned a unique level for factor m2. In this case, the number of unique levels for m1 is number of the animals of sex 2 plus 1, and the number of unique levels for m2 is the number of animals of sex 1 plus 1. For this data file, the number of levels for m1 was 1 + 134 = 135 and for m2 was 1 + 150 = 151. With large data files, many extra equations will result: basically one extra equation for each animal with an observation on one trait but no observation for the other trait for a 2-trait analysis. In this case the original levels for gn, sx, and nℓ could be used, but with other genotype × environmental interaction models that would not always be true.

Analysis of the 2 sex-limited traits with missing observations excluded gave the same estimates of fixed and random effects, (co)variance components, heritabilities, and genetic correlations as for the analysis with missing observations for 1 trait or the other trait treated as actual observations (0.00) with corresponding unique levels of factors, m1 and m2. With factors m1 and m2 and pseudo-values as "real" observations, SE of the estimates of the genetic parameters were obtained for the 2-trait analysis even when no animals had actual measurements for both traits.

A similar multiple-trait model is often used to estimate covariance components due to genotype × environmental interaction based on sires with progeny in more than one environment. Environment is often region or country. Often thousands of animals provide observations. For a 2-trait G × E analysis, each animal would have a record only in its own environment, but to obtain a SE for the estimate of the genetic correlation between environments, each animal would also have a pseudo-value for the other environment. The numbers of unique levels for the dummy factors (m1 and m2) would effectively be the total number of animals with observations in the 2 environments, which could be many thousands, whereas the number of genetic effects would usually be twice the total number of animals, with a much smaller number with a sire model. Most of the mixed model equations would be associated with levels of the dummy factors for pseudo-observations. The 2-trait file would usually be created by joining files from the 2 regions with separate sets of fields for fixed factors and with 2 new fields for the pseudo missing observations for environment 1 or 2 very much the same as for Table 4. The ASREML program (Gilmour et al., 2002) appears to use a similar procedure but without external changes in the data file.

A method has been developed to obtain SE of genetic parameters with the MTDFREML program for multiple-trait analyses when some observations are missing for some traits. The method involves relatively small changes in the data file. Estimates of genetic parame-

Table 4. Data file for first 20 of 284 mice for a 2-trait analysis with trait 1 measured only on sex 1 and trait 2 on sex 2 modified to create unique levels for factors m1 and m2 for missing observations (0.00) for trait 1 (m1) or trait 2 (m2)[1]

An ID	Sire	Dam	gn	sx	nl	lit	m1	m2	n1	t1	t2
20101	11012	10101	1	1	4	1	1	2	4.	22.50	0.00
20102	11012	10101	1	1	4	1	1	3	4.	22.60	0.00
20103	11012	10101	1	1	4	1	1	4	4.	22.90	0.00
20104	11012	10101	1	1	4	1	1	5	4.	23.00	0.00
20112	11012	10101	1	2	4	1	2	1	4.	0.00	69.30
20113	11012	10101	1	2	4	1	3	1	4.	0.00	66.40
20114	11012	10101	1	2	4	1	4	1	4.	0.00	61.60
20115	11012	10101	1	2	4	1	5	1	4.	0.00	68.30
20302	10614	10506	1	1	7	2	1	6	7.	23.60	0.00
20303	10614	10506	1	1	7	2	1	7	7.	24.20	0.00
20304	10614	10506	1	1	7	2	1	8	7.	22.50	0.00
20306	10614	10506	1	1	7	2	1	9	7.	22.30	0.00
20312	10614	10506	1	2	7	2	6	1	7.	0.00	66.90
20313	10614	10506	1	2	7	2	7	1	7.	0.00	72.90
20314	10614	10506	1	2	7	2	8	1	7.	0.00	58.40
20401	10813	10701	1	1	2	3	1	10	2.	20.30	0.00
20402	10813	10701	1	1	2	3	1	11	2.	21.30	0.00
20403	10813	10701	1	1	2	3	1	12	2.	21.90	0.00
20404	10813	10701	1	1	2	3	1	13	2.	22.10	0.00
20413	10813	10701	1	2	2	3	9	1	2.	0.00	62.50

[1]Note that no animal has both traits measured. gn = generation; sx = sex; nl = number in litter; m1 = dummy factor for trait 1; m2 = dummy factor for trait 2; n1 = number in litter if used as a covariate; t1 = trait 1; t2 = trait 2.

ters, estimable functions of fixed effects, and L|y are the same as for the original data file.

The steps in changing the data file and analysis of the augmented data file may vary, but the following are suggested:

1) Do the analysis with missing observations treated as missing because there will be fewer equations and nonzero coefficients in the mixed model equations. Thus, the analysis will run faster and result in best possible starting values to run with missing observations treated as actual values. At convergence, copy updated starting answer file, MTDF4, to MTDF4.1.

2a) When levels of other factors are known, the data file is modified by using the missing value indicator as a real measurement and including a new fixed factor for each trait which would have a level of 1 for an actual measurement and a unique level (easiest as, 2, 3, ..., etc.) for each missing measurement.

2b) When levels of other factors are not known, the data file can also be modified by creating new fields for other fixed factors for each trait and using the original levels when the trait is actually measured and, for data checking, using a unique level different from any of the original levels when the missing value is used as the measurement. Other options will also work.

3) When the data file with extra fixed factors is created, change levels for other factors for missing observations to a level unique for that factor. Means from MTDF66 will be correct for other levels, and the unique level will have as a mean the original "missing value indicator".

4) Next run MTDFPREP with the modified data file with a different "missing value indicator". The MTDF66 file from MTDFPREP must be interpreted carefully as the original missing value indicator will be used to calculate overall mean, unadjusted SD, and high and low observations. These items of data description can be obtained from MTDF66 from running the analysis with missing observations treated as missing (suggestion 1).

5) Then restart MTDFRUN with command MTDFRUN<MTDF4.1. MTDFRUN should need only 1 round (i.e., modify MTDF4.1 to do 1 round). Disregard "not converged" message. To be sure, compare –2logL from 1) with –2logL from 5).

6) For the extra factor, do not always ask for a summary because some analyses may have 1,000s of levels (e.g., G × E analyses). For some analyses, the user may need to change sizes of vectors included in the PARAM.DAT file and recompile MTDFPREP and MTDFRUN.

APPENDIX

The REML estimators differ from maximum likelihood estimators because the maximum likelihood estimators are based on the distribution of the data vector, whereas REML estimators are based on the distribution of a lower dimensional linear function of the data

vector or residual vector. If the same residual vector can be used for both the observed data and the augmented data, then the REML estimators along with their sampling properties will also be the same as for the maximum likelihood estimators. To show that the same residual vector can be used for both the observed and augmented data, we will first define a residual vector for the observed data and then show that this residual vector can be used for the augmented data.

Let y_o be the $n_o \times 1$ vector of observed data, where

$$y_o = X_o\beta + Z_o u + e_o \text{ with } u \sim N(0,G) \text{ and } e_o \sim N(0,R_o).$$

The REML estimators of the (co)variance components of G and R_o are the maximum likelihood estimators of those (co)variance components based on the distribution of the residual vector $K_o'y_o$. The $(n_o - p_o) \times n_o$ matrix K_o' is selected such that 1) p_o is the rank of X_o, 2) $K_o'X_o = 0$, and 3) the rank of K_o' is $n_o - p_o$. These requirements define REML (e.g., Harville, 1977; Searle et al., 1992). The distribution of the residual vector is

$$K_o'y_o \sim N(0, K_o'Z_0GZ_o'K_0 + K_o'R_0K_0).$$

Because the selection of K_o' is based on the column space of X_o, augmenting the original design matrix X_o with additional columns, W_o, will not impact the selection of K_o' provided that the columns of W_o are a linear function of the columns of X_o. So without loss of generality, the vector of missing value effects, m_o, and design matrix, W_o, can be added to the observed model equations for the data,

$$y_o = X_o\beta + W_om_o + Z_ou + e_o.$$

The same residual vector, $K_o'y_o$, can be used provided that the columns of W_o are a linear function of the columns of X_o.

Next let y_u be the $n_u \times 1$ vector of unobserved data, where

$$y_u = X_u\beta + W_um_u + Z_uu + e_u,$$

with W_u is a matrix with rank n_u and with $u \sim N(0,G)$ and

$$\begin{pmatrix} e_o \\ e_u \end{pmatrix} \sim N(0,R).$$

The $n \times 1$ augmented data vector, y, is then formed by concatenating the observed and unobserved data vectors. The design matrix for the fixed effects for the augmented data is

$$X = \begin{pmatrix} X_o & W_o & 0 \\ X_u & 0 & W_u \end{pmatrix}.$$

The residual vector for the observed data, $K_o'y_o$, can be written in terms of the augmented data as

$$K'y = (K_0' \;\; 0)\begin{pmatrix} y_o \\ y_u \end{pmatrix}.$$

It remains to be shown that $K'y$ is a $(n - p) \times 1$ residual vector for the augmented data that satisfies: 1) p = rank of X, the design matrix for the augmented data, 2) $K'X = 0$, and 3) rank of K' is $n - p$. Because none of the last n_u rows of X are a linear function of the first n_o rows, the rank of X is equal to the rank of the first n_o rows, p_o, plus the rank of the last n_u rows, n_u. Therefore, the rank of X is $p = n_u + p_o$ and $K'y$ is a $(n - p) \times 1$ residual vector because $n - p = (n_o + n_u) - (n_u + p_o) = n_o - p_o$. With $K'X = K_o'X_o + K_o'W_o$ and K_o' selected such that $K_o'X_o + K_o'W_o = 0$, $K'X$ must also be equal to 0. Because $K' = (K_o' \; 0)$, the rank of K' is equal to the rank of K_o'. With K_o' selected to have rank $n_o - p_o$, which is equal to $n - p$, the rank of K' is also equal to $n - p$. Therefore, $K'y$ satisfies the 3 conditions for a REML residual vector.

LITERATURE CITED

Boldman, K. G., L. A. Kriese, L. D. Van Vleck, C. P. Van Tassell, and S. D. Kachman. 1995. A manual for USE of MTDFREML. A set of programs to obtain estimates of variances and covariances [DRAFT]. USDA-ARS, Washington, DC.

Dodenhoff, J., L. D. Van Vleck, S. D. Kachman, and R. M. Koch. 1998. Parameter estimates for direct, maternal and grandmaternal genetic effects for birth weight and weaning weight in Hereford cattle. J. Anim. Sci. 76:2521–2527.

Gilmour, A. R., B. J. Gogel, B. R. Cullis, J. J. Welham, and R. Thompson. 2002. ASReml User Guide. VSN International Ltd., Hemel Hempstead, UK.

Harville, D. A. 1977. Maximum-likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 72:320–340.

Johnson, D. L., and R. Thompson. 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci. 78:449–456.

Searle, S. R., G. Casella, and C. E. McCulloch. 1992. Variance Components. John Wiley and Sons Inc., New York, NY.