REML Variance-Component Estimation

In the numerous forms of analysis of variance (ANOVA) discussed in previous chapters, variance components were estimated by equating observed mean squares to expressions describing their expected values, these being functions of the variance components. ANOVA has the nice feature that the estimators for the variance components are unbiased regardless of whether the data are normally distributed, but it also has two significant limitations. First, field observations often yield records on a variety of relatives, such as offspring, parents, or sibs, that cannot be analyzed jointly with ANOVA. Second, ANOVA estimates of variance components require that sample sizes be well balanced, with the number of observations for each set of conditions being essentially equal. In field situations, individuals are often lost, and even the most carefully crafted balanced design can quickly collapse into an extremely unbalanced one. Although modifications to the ANOVA sums of squares have been proposed to account for unbalanced data (Henderson 1953, Searle et al. 1992), their sampling properties are poorly understood.

Unlike ANOVA estimators, maximum likelihood (ML) and restricted maximum likelihood (REML) estimators do not place any special demands on the design or balance of data. Such estimates are ideal for the unbalanced designs that arise in quantitative genetics, as they can be obtained readily for any arbitrary pedigree of individuals. Since many aspects of ML and REML estimation are quite difficult technically, the detailed mathematics can obscure the general power and flexibility of the methods. Therefore, our main concern is to make the theory more accessible to the nonspecialist, and as a consequence, we are not as thorough in our coverage of the literature as in previous chapters. Also, unlike elsewhere in this book, we occasionally rely upon mathematical machinery (such as matrix derivatives) that is not fully developed here (see Appendix 3 for an introduction). This chapter is mathematically difficult in places, and the reader will do well to review some of the advanced topics in Chapter 8 (such as the multivariate normal and expectations of quadratic products) and Appendix 4.

We start at a relatively elementary level, providing a simple example to show how ML and REML procedures can be used to estimate variance components and how these estimates differ. We then develop the ML and REML equations for variance-component estimation under the general mixed model (introduced in Chapter 26). Extension of these methods to multiple traits, wherein full covariance matrices, rather than single variance components, must be estimated, are then reviewed. We conclude our coverage of ML/REML by examining a number of computational methods for solving the ML/REML equations.

ML VERSUS REML ESTIMATES OF VARIANCE COMPONENTS

Although algebraically tedious, maximum likelihood (ML) is conceptually very simple. It was introduced to variance component-estimation by Hartley and Rao (1967). For a specified model, such as Equation 26.1, and a specified form for the joint distribution of the elements of $\mathbf{y}$, ML estimates the parameters of the distribution that maximize the likelihood of the observed data. This distribution is almost always assumed to be multivariate normal. An advantage of ML estimators is their efficiency — they simultaneously utilize all of the available data and account for any nonindependence.

One drawback with variance-component estimation via the usual maximum likelihood approach is that all fixed effects are assumed to be known without error. This is rarely true in practice, and as a consequence, ML estimators yield biased estimates of variance components. Most notably (as we show below), estimates of the residual variance tend to be downwardly biased. This bias occurs because the observed deviations of individual phenotypic values from an estimated population mean tend to be smaller than their deviations from the true (parametric) mean. Such bias can become quite large when a model contains numerous fixed effects, particularly when sample sizes are small.

Unlike ML estimators, restricted maximum likelihood (REML) estimators maximize only the portion of the likelihood that does not depend on the fixed effects. In this sense, REML is a *restricted* version of ML. The elimination of bias by REML is analogous to the removal of bias that arises in the estimate of a variance component when the mean squared deviation is divided by the degrees of freedom instead of by the sample size (Chapter 2, and below). REML does not always eliminate all of the bias in parameter estimation, since many methods for obtaining REML estimates cannot return negative estimates of a variance component. However, this source of bias also exists with ML, so REML is clearly the preferred method for analyzing large data sets with complex structure. In the ideal case of a completely balanced design, REML yields estimates of variance components that are identical to those obtained by classical analysis of variance. Since it was first introduced to breeders by Patterson and Thompson (1971), many thorough references to REML, its justification, and its various applications have been published (Harville 1977; Ott 1979; Henderson 1984b, 1986; Gianola and Fernando 1986; Little and Rubin 1987; Robinson 1987; Searle 1987; Shaw 1987; Searle et al. 1992).

A Simple Example of ML versus REML

In an attempt to make the distinction between ML and REML likelihood equations as simple and transparent as possible, we start with a useful pedagogical connection between ML and REML noticed by Foulley (1993), confining our attention to a very simple application — the estimation of the mean and variance

of a set of independent observations. In this case, the mixed model reduces to

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{e} \qquad (27.1)$$

where μ is the population mean (the fixed effect), $\mathbf{1}$ is a $n \times 1$ column vector of ones (equivalent to the design matrix $\mathbf{X}$ in Equation 26.1), and the covariance matrix of residuals about the mean is assumed to be $\mathbf{R} = \sigma^2\mathbf{I}$.

What are the ML estimates of μ and σ^2 based on the n sampled individuals? Assuming the phenotypes are independent of each other and normally distributed, the probability density of the data $\mathbf{y}$ conditional on the parametric mean and variance is the product of the n univariate normal densities,

$$p(\mathbf{y} \,|\, \mu, \sigma^2) = \prod_{i=1}^{n} p(y_i \,|\, \mu, \sigma^2)$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left[-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2}\right] \qquad (27.2)$$

where y_i is the phenotypic value of the ith individual. Taking the natural logarithm of the expression on the right, the log-likelihood (Appendix 4) for the observed data set is

$$L(\mathbf{y} \,|\, \mu, \sigma^2) = -\frac{n}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{1}{n\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \qquad (27.3\text{a})$$

Although this is the logarithm of the likelihood of the data given the moments of the normal distribution (μ and σ^2), it can also be viewed as the log-likelihood of the parameter estimates, $L(\mu, \sigma^2 \,|\, \mathbf{y})$, treating the y_i as constants and μ and σ^2 as variables. To obtain estimates of these two distributional parameters, we need at least two observable statistics. Letting

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{and} \qquad V = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

we have

$$\sum_{i=1}^{n}(y_i - \mu)^2 = \sum_{i=1}^{n}(y_i - \overline{y} + \overline{y} - \mu)^2$$

$$= \sum_{i=1}^{n}(y_i - \overline{y})^2 + \sum_{i=1}^{n}(\overline{y} - \mu)^2 + 2(\overline{y} - \mu)\sum_{i=1}^{n}(y_i - \overline{y})$$

$$= n\left[V + (\overline{y} - \mu)^2\right] \qquad (27.3\text{b})$$

Substituting this final expression into Equation 27.3a, the log-likelihood can be expressed as

$$L(\mu, \sigma^2 \mid \mathbf{y}) = -\frac{n}{2}\left[\ln(2\pi) + \ln(\sigma^2) + \frac{V + (\overline{y} - \mu)^2}{\sigma^2}\right] \qquad (27.3c)$$

Differentiating with respect to μ and σ^2 yields

$$\frac{\partial\, L(\mu, \sigma^2|\mathbf{y})}{\partial\mu} = \frac{n(\overline{y} - \mu)}{\sigma^2} \qquad (27.4a)$$

$$\frac{\partial\, L(\mu, \sigma^2|\mathbf{y})}{\partial\sigma^2} = -\frac{n}{2\sigma^2}\left[1 - \frac{V + (\overline{y} - \mu)^2}{\sigma^2}\right] \qquad (27.4b)$$

By setting these equations equal to zero and solving, we obtain estimators for the population mean and variance that maximize the likelihood function given the observed data $\mathbf{y}$. From Equation 27.4a, we obtain an estimator for the mean that is completely independent of the variance,

$$\widehat{\mu} = \overline{y} \qquad (27.5a)$$

where $\widehat{}$ denotes an estimate. This shows that the standard definition of a sample mean is, in fact, the ML estimate of the parametric value. Unfortunately, the solution to Equation 27.4b,

$$\widehat{\sigma}^2 = V + (\overline{y} - \mu)^2 \qquad (27.5b)$$

is not independent of the estimated mean, $\overline{y}$, unless the estimated mean happens to coincide perfectly with the true mean μ. The maximum likelihood estimator of σ^2 is obtained by assuming that the mean is, in fact, estimated without error, yielding

$$\widehat{\sigma}^2 = V \qquad (27.5c)$$

Since the term ignored in Equation 27.5b is necessarily positive, Equation 27.5c gives a downwardly biased estimate of the true variance σ^2.

REML removes this bias by accounting for the error in the estimation of μ. From Equation 27.5b, the expected amount by which $\widehat{\sigma}^2$ underestimates σ^2 is the expected value of $(\overline{y} - \mu)^2$, which is simply the sampling variance of the mean, σ^2/n. Thus, an improved estimator is

$$\widehat{\sigma}^2 = V + E[(\overline{y} - \mu)^2] = V + \frac{\sigma^2}{n} \qquad (27.5d)$$

We cannot, of course, know exactly what this bias is because we do not know σ^2 with certainty (indeed, we are trying to estimate it). However, the bias is estimable

because we have a preliminary estimate of σ^2, the maximum likelihood estimate V. Thus, starting with the initial estimate of $\widehat{\sigma}^2(0) = V$, a second improved estimate of the variance is

$$\widehat{\sigma}^2(1) = V + \frac{\widehat{\sigma}^2(0)}{n} = V + \frac{V}{n}$$

However, just as this changes the estimate of the variance, it also changes the estimate of $(\overline{y} - \mu)^2$. Hence, a third estimate of σ^2 would be

$$\widehat{\sigma}^2(2) = V + \frac{\widehat{\sigma}^2(1)}{n} = V + \frac{V + (V/n)}{n}$$

This sequence suggests an iterative approach for estimating the variance,

$$\widehat{\sigma}^2(t+1) = V + \frac{\widehat{\sigma}^2(t)}{n} \tag{27.6a}$$

The final (stable) solution to this equation, $\widehat{\sigma}^2$, is obtained by setting $\widehat{\sigma}^2(t+1) = \widehat{\sigma}^2(t)$, yielding

$$\widehat{\sigma}^2 = \frac{n}{n-1}V = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1} \tag{27.6b}$$

which is the unbiased estimator of the variance that we normally use (Chapter 2).

To obtain a solution for this particular example, iteration of Equation 27.6a is not really necessary. However, with models containing multiple fixed effects in the form of the vector **u**, closed solutions such as Equation 27.6b are not usually possible, particularly in complex pedigree analyses involving unbalanced data. In those cases, as we will see below, iterative procedures can still yield solutions that are asymptotically unbiased.

Note that the REML estimators given by Equations 27.5a and 27.6b were derived under the assumption of normality. That these same solutions can be acquired without reference to any particular distribution (Chapter 2) provides some evidence that REML estimators may often be fairly robust to violations of the normality assumption.

ML ESTIMATES OF VARIANCE COMPONENTS IN THE GENERAL MIXED MODEL

In light of the fundamental role that the mixed model plays in quantitative genetics, we attempt in this section to give a clear step-by-step development of the maximum likelihood procedures, following the same steps that were used above for the simple model ($\mathbf{y} = \mathbf{1}\,\mu + \mathbf{e}$). Although REML is preferred over ML as a method of analysis, we start with ML, since REML estimation can be expressed as an ML problem by a simple linear transform.

We start with the general mixed model (Equation 26.1), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$, and we assume that $\mathbf{u} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{G})$ and $\mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{R})$. Under this model, $\mathbf{y}$ is also multivariate normal, with mean $\mathbf{X}\boldsymbol{\beta}$ and variance-covariance matrix $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$. Recalling the form of the multivariate normal distribution (Equation 8.24), the probability density of the data $\mathbf{y}$, analogous to that in Equation 27.2, is

$$p(\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}, \mathbf{V}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \quad (27.7a)$$

The next step, analogous to Equation 27.3a, is to take the natural logarithm of the expression on the right of Equation 27.7a. This yields the log-likelihood of $\boldsymbol{\beta}$ and $\mathbf{V}$ given the observed data $(\mathbf{X}, \mathbf{y})$ as

$$L(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{X}, \mathbf{y}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (27.7b)$$

The following discussion considers $\mathbf{u} = \mathbf{a}$ to be the vector of additive genetic (breeding) values. The variance components that we are trying to estimate are embedded within $\mathbf{G}$ and $\mathbf{R}$, and we assume that $\mathbf{G} = \sigma_A^2 \mathbf{A}$, where $\mathbf{A}$ is the additive genetic relationship matrix, and that $\mathbf{R} = \sigma_E^2 \mathbf{I}$, i.e., the residual deviations of different individuals are independent and homoscedastic.

This approach extends readily to the estimation of additional variance components by using the generalized model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{m} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e} \quad (27.8a)$$

where the m vectors of random effects ($\mathbf{u}_i$) are assumed to be uncorrelated, with $\mathbf{u}_i \sim \mathrm{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{B}_i)$ and $\mathbf{B}_i$ being a matrix of known constants. This more general model can incorporate estimates of dominance and other nonadditive variances, and maternal effects variances, to name a few (see Chapter 26). The log-likelihood is still given by Equation 27.7b, but now the covariance matrix $\mathbf{V}$ consists of $m+1$ (unknown) variances,

$$\mathbf{V} = \sum_{i=1}^{m} \sigma_i^2 \mathbf{Z}_i \mathbf{B}_i \mathbf{Z}_i^T + \sigma_E^2 \mathbf{I} \quad (27.8b)$$

We now move on to the partial derivatives of the log-likelihood required for the derivation of the ML estimators. Consider first the derivative with respect to the vector of fixed effects, $\boldsymbol{\beta}$. This derivative involves only the final term of Equation 27.7b, and its procurement is facilitated by using a general result for matrix derivatives. Applying Equation A3.25d,

$$\frac{\partial\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (27.9)$$

which yields

$$\frac{\partial L(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{X}, \mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{27.10}$$

Obtaining the partial derivatives with respect to the variances σ_A^2 and σ_E^2 involves two other general results from matrix theory (Searle 1982, pp. 335–336). If $\mathbf{M}$ is a square matrix whose elements are functions of a scalar variable x, then

$$\frac{\partial \ln |\mathbf{M}|}{\partial x} = \text{tr} \left(\mathbf{M}^{-1} \frac{\partial \mathbf{M}}{\partial x} \right) \tag{27.11a}$$

$$\frac{\partial \mathbf{M}^{-1}}{\partial x} = -\mathbf{M}^{-1} \frac{\partial \mathbf{M}}{\partial x} \mathbf{M}^{-1} \tag{27.11b}$$

where tr, the **trace**, denotes the sum of the diagonal elements of a square matrix (Chapter 8). The trace operator appears frequently in this chapter, and the following properties will prove useful

$$\text{tr}(a \, \mathbf{A}) = a \, \text{tr}(\mathbf{A}) \tag{27.12a}$$

$$\text{tr}(\mathbf{I}_n) = n \tag{27.12b}$$

$$\text{tr}(\mathbf{B}_{n \times m} \mathbf{A}_{m \times n}) = \text{tr}(\mathbf{A}_{m \times n} \mathbf{B}_{n \times m}) \tag{27.12c}$$

$$\text{tr}(\mathbf{A} + \mathbf{C}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{C}) \tag{27.12d}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Recall that prior to the differentiation of Equation 27.3a, we rewrote the sum of squared deviations of observed mean phenotypes from the population mean in terms of $(y_i - \mu)$ and $\overline{y} - \mu$. Performing the analogous changes in matrix form, we find that

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$
$$+ (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \tag{27.13}$$

where $\widehat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}$. (This step is not really necessary here, but its incorporation will allow us to see the bias in ML estimates of the variance components, as it did in the previous section.)

Moving now to the derivatives with respect to the variance components, we first assume the simple case of only two unknown variances, typically σ_E^2 and σ_A^2. Writing $\mathbf{V}$ in terms of these two components, we have $\mathbf{V} = \sigma_A^2 \, \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \sigma_E^2 \, \mathbf{I}$. Using the notation of σ_i^2 to denote the variance component being estimated, we have

$$\frac{\partial \mathbf{V}}{\partial \sigma_i^2} = \mathbf{V}_i = \begin{cases} \mathbf{I} & \text{when } \sigma_i^2 = \sigma_E^2 \\ \mathbf{Z}\mathbf{A}\mathbf{Z}^T & \text{when } \sigma_i^2 = \sigma_A^2 \end{cases} \tag{27.14a}$$

Substituting Equation 27.13 into Equation 27.7b, using Equations 27.11a,b, and letting σ_i^2 denote either σ_A^2 or σ_E^2, we obtain the general equation

$$\frac{\partial L(\boldsymbol{\beta}, \mathbf{V} \mid \mathbf{X}, \mathbf{y})}{\partial \sigma_i^2} = -\frac{1}{2}\text{tr}(\mathbf{V}^{-1}\mathbf{V}_i) + \frac{1}{2}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$$

$$+ \frac{1}{2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad (27.14\text{b})$$

where $\mathbf{V}_i$ is given by Equation 27.14a. Equations 27.10 and 27.14b are directly analogous to Equations 27.4a,b derived above. Note that $\mathbf{V}_i$ is a fixed matrix of known constants, whereas $\mathbf{V} = \sigma_A^2\,\mathbf{Z}\mathbf{A}\mathbf{Z}^T + \sigma_E^2\,\mathbf{I}$ is a function of the variance-component estimates. More generally, with m random effects plus a residual error (Equation 27.8a), Equation 27.14b holds for each of the $m+1$ variance components with

$$\frac{\partial\,\mathbf{V}}{\partial\,\sigma_i^2} = \mathbf{V}_i = \begin{cases} \mathbf{I} & \text{when } \sigma_i^2 = \sigma_E^2 \\ \mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T & \text{otherwise} \end{cases} \qquad (27.15)$$

The maximum likelihood (ML) estimators are obtained by setting Equations 27.10 and 27.14b equal to zero and solving. Using Equation 27.10 alone, a little rearranging gives the ML estimate of the vector of fixed effects as

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{y} \qquad (27.16)$$

Note that this is the BLUE (best linear unbiased estimator) of $\boldsymbol{\beta}$ obtained in the previous chapter (Equation 26.3). The ML estimators for the variance components are obtained by setting $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ in Equation 27.14b, rendering the last term equal to zero. Rearranging, we obtain

$$\text{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{V}_i) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T\widehat{\mathbf{V}}^{-1}\mathbf{V}_i\widehat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \qquad (27.17\text{a})$$

This equation can be simplified by using the matrix

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} \qquad (27.17\text{b})$$

which will appear frequently throughout the rest of the chapter. In particular, we have the very useful result that

$$\mathbf{P}\mathbf{y} = \mathbf{V}^{-1}\mathbf{y} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \qquad (27.17\text{c})$$

Using this identity, Equation 27.17a can be more compactly written as

$$\text{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{V}_i) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{V}_i\widehat{\mathbf{P}}\mathbf{y} \qquad (27.17\text{d})$$

where we use the notation $\widehat{\mathbf{P}}$ to remind the reader that $\mathbf{P}$, being a function of $\mathbf{V}$, depends on the variance components that we are trying to estimate. Although

it may not be immediately apparent, Equation 27.17d is directly analogous to Equation 27.5c. The variance estimates that we wish to obtain, $\widehat{\sigma}_A^2$ and $\widehat{\sigma}_E^2$, are contained on both sides of Equation 27.17d, embedded in the inverted variance-covariance matrix $\widehat{\mathbf{V}}^{-1}$ that appears in $\mathbf{P}$.

In summary, the ML estimates satisfy the solutions to Equation 27.16 (for the fixed effects) and the set of equations for the variance components (Equation 27.17d). For the additive model assumed above, the two variance equations are

$$\text{tr}(\widehat{\mathbf{V}}^{-1}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2 \qquad (27.18a)$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{Z}\mathbf{A}\mathbf{Z}^T) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{Z}\mathbf{A}\mathbf{Z}^T\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_A^2 \qquad (27.18b)$$

More generally, with m random effects plus a residual (Equation 27.8a), the set of $m + 1$ ML equations for the variances of random effects is

$$\text{tr}(\widehat{\mathbf{V}}^{-1}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2 \qquad (27.19a)$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{Z}_i\mathbf{B}_i\mathbf{Z}_i^T\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_i^2,\, 1 \le i \le m \qquad (27.19b)$$

where $\widehat{\mathbf{P}}$ now uses

$$\widehat{\mathbf{V}} = \sum_{i=1}^{m} \widehat{\sigma}_i^2\, \mathbf{Z}_i\, \mathbf{B}_i\, \mathbf{Z}_i^T + \widehat{\sigma}_E^2\, \mathbf{I} \qquad (27.19c)$$

These solutions have two troublesome properties. First, unlike our simple example at the start of this chapter where there was a closed form estimator for the fixed effect μ, the ML vector of fixed effects $\widehat{\boldsymbol{\beta}}$ is a function of the variance-covariance matrix $\widehat{\mathbf{V}}$, which in turn contains the variance components that we wish to estimate. Second, because these solutions involve the inverse of $\widehat{\mathbf{V}}$, they are nonlinear functions of the variance components. As a consequence, there is no simple one-step solution. ML estimation of $\boldsymbol{\beta}$, σ_A^2, and σ_E^2 requires an iterative procedure, several steps of which are described below.

Example 1. Consider the simple animal model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\mathbf{a}+\mathbf{e}$, where there is only one observation per individual ($\mathbf{Z} = \mathbf{I}$), and we assume $\mathbf{a} \sim \text{MVN}(\mathbf{0}, \sigma_A^2\, \mathbf{A})$ and $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma_E^2\, \mathbf{I})$. In this case, the ML equations become

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\widehat{\mathbf{V}}^{-1}\mathbf{y}$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1}) = \mathbf{y}^T\widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y}$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1}\mathbf{A}) = \mathbf{y}^T\widehat{\mathbf{P}}\mathbf{A}\widehat{\mathbf{P}}\mathbf{y}$$

where

$$\widehat{\mathbf{V}} = \widehat{\sigma}_A^2\, \mathbf{A} + \widehat{\sigma}_E^2\, \mathbf{I}$$

and $\widehat{\mathbf{P}}$ is obtained by substituting $\widehat{\mathbf{V}}$ into Equation 27.17b.

If we further allow for dominance, the model becomes modified to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{d} + \mathbf{e}$. Assuming $\mathbf{a} \sim \text{MVN}(\mathbf{0}, \sigma_A^2 \mathbf{A})$, $\mathbf{d} \sim \text{MVN}(\mathbf{0}, \sigma_D^2 \mathbf{D})$, and $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \sigma_E^2 \mathbf{I})$, the ML equations now become

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \widehat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{V}}^{-1} \mathbf{y}$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1}) = \mathbf{y}^T \widehat{\mathbf{P}} \widehat{\mathbf{P}} \mathbf{y}$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1} \mathbf{A}) = \mathbf{y}^T \widehat{\mathbf{P}} \mathbf{A} \widehat{\mathbf{P}} \mathbf{y}$$

$$\text{tr}(\widehat{\mathbf{V}}^{-1} \mathbf{D}) = \mathbf{y}^T \widehat{\mathbf{P}} \mathbf{D} \widehat{\mathbf{P}} \mathbf{y}$$

where $\widehat{\mathbf{P}}$ is a function of

$$\widehat{\mathbf{V}} = \widehat{\sigma}_A^2 \mathbf{A} + \widehat{\sigma}_D^2 \mathbf{D} + \widehat{\sigma}_E^2 \mathbf{I}$$

Standard Errors of ML Estimates

Recall from the theory of maximum likelihood (Appendix 4) that standard errors of ML estimates can be obtained from the appropriate elements of the inverse of the Fisher information matrix ($\mathbf{F}$) involving the vector of parameters being estimated ($\boldsymbol{\Theta}$). The elements of $\mathbf{F}$ are functions of the second derivatives of the log-likelihood function, evaluated by substituting ML estimates of the parameters,

$$\mathbf{F}_{ij} = -E\left(\frac{\partial^2 L}{\partial\,\theta_i\,\partial\,\theta_j} \right) \simeq -\left. \frac{\partial^2 L}{\partial\,\theta_i\,\partial\,\theta_j} \right|_{\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}} \tag{27.20}$$

The sampling variance of the ML estimate of the parameter θ_i is approximated by F_{ii}^{-1} (the ith diagonal element of $\mathbf{F}^{-1}$), while the sampling covariance between the ML estimates of θ_i and θ_j is approximated by F_{ij}^{-1}.

Computing the partials for the mixed model gives the information matrix for the ML estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ (the vector of variance-component estimates) as

$$\mathbf{F} = \begin{pmatrix} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix} \tag{27.21}$$

where

$$S_{ij} = \frac{1}{2}\, \text{tr}\left(\mathbf{V}^{-1} \mathbf{V}_i \mathbf{V}^{-1} \mathbf{V}_j \right) \tag{27.22}$$

with $\mathbf{V}_i$ given by Equation 27.15 (Searle et al. 1992). Inverting gives

$$\mathbf{F}^{-1} = \begin{pmatrix} \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{-1} \end{pmatrix} \tag{27.23}$$

Hence,

$$\sigma(\beta_i, \beta_j) = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X} \right)^{-1}_{ij}, \qquad \sigma(\sigma_i^2, \sigma_j^2) = \left(\mathbf{S}^{-1} \right)_{ij} \qquad (27.24)$$

The ML estimates for fixed effects are uncorrelated with those for variance components, i.e., $\sigma(\beta_i, \sigma_j^2) = 0$.

Example 2. For the simple model with dominance (Example 1), the Fisher information submatrix $\mathbf{S}$ dealing with the ML variance estimates $(\sigma_A^2, \; \sigma_D^2, \; \sigma_E^2)$ is

$$\mathbf{S} = \frac{1}{2} \begin{pmatrix} \mathrm{tr}(\mathbf{V}^{-1}\mathbf{A}\mathbf{V}^{-1}\mathbf{A}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{A}\mathbf{V}^{-1}\mathbf{D}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{A}\mathbf{V}^{-1}) \\ \mathrm{tr}(\mathbf{V}^{-1}\mathbf{A}\mathbf{V}^{-1}\mathbf{D}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{D}\mathbf{V}^{-1}\mathbf{D}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{D}\mathbf{V}^{-1}) \\ \mathrm{tr}(\mathbf{V}^{-1}\mathbf{A}\mathbf{V}^{-1}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{D}\mathbf{V}^{-1}) & \mathrm{tr}(\mathbf{V}^{-1}\mathbf{V}^{-1}) \end{pmatrix}$$

where $\mathbf{V}$ is as given in Example 1.

RESTRICTED MAXIMUM LIKELIHOOD

REML is based on a linear transformation of the observation vector $\mathbf{y}$ that removes the fixed effects from the model. The simplest way to see how this is done is to imagine a transformation matrix $\mathbf{K}$ associated with the design matrix $\mathbf{X}$ for the model under consideration such that

$$\mathbf{K}\mathbf{X} = \mathbf{0} \qquad (27.25)$$

Applying this transformation matrix to the mixed model yields

$$\begin{aligned} \mathbf{y}^* = \mathbf{K}\mathbf{y} &= \mathbf{K}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e}) \\ &= \mathbf{K}\mathbf{Z}\mathbf{a} + \mathbf{K}\mathbf{e} \end{aligned} \qquad (27.26a)$$

The linear contrasts $\mathbf{y}^*$ are equivalent to residual deviations from the estimated fixed effects, akin to using $y_i^* = y_i - \overline{y}$ in the introductory example used at the start of this chapter. REML estimates of variance components are equivalent to ML estimates of the transformed variables. Thus, we can use the ML solutions outlined above by making the following substitutions:

$$\mathbf{K}\mathbf{y} \text{ for } \mathbf{y}, \qquad \mathbf{K}\mathbf{X} = \mathbf{0} \text{ for } \mathbf{X}, \qquad \mathbf{K}\mathbf{Z} \text{ for } \mathbf{Z}, \qquad \mathbf{K}\mathbf{V}\mathbf{K}^T \text{ for } \mathbf{V} \quad (27.26b)$$

While REML appears to require the additional task of finding a matrix $\mathbf{K}$ that satisfies Equation 27.25, the REML equations can actually be expressed directly in terms of $\mathbf{V}$, $\mathbf{y}$, and $\mathbf{P}$. This result follows from the very useful identity, proven in Searle et al. (1992), that $\mathbf{K}$ satisfies

$$\mathbf{P} = \mathbf{K}^T (\mathbf{K}\mathbf{V}\mathbf{K}^T)^{-1} \mathbf{K} \qquad (27.27a)$$

Noting that

$$(\mathbf{y}^*)^T (\mathbf{V}^*)^{-1} \mathbf{y}^* = (\mathbf{y}^T \mathbf{K}^T)(\mathbf{K}\mathbf{V}\mathbf{K}^T)^{-1}(\mathbf{K}\mathbf{y}) = \mathbf{y}^T \mathbf{P} \mathbf{y} \qquad (27.27b)$$

and substituting the expressions given as 27.26b into Equation 27.17a, after some rearrangement, the ML equations yield the REML estimators,

$$\text{tr}(\widehat{\mathbf{P}}) = \mathbf{y}^T \widehat{\mathbf{P}}\widehat{\mathbf{P}} \mathbf{y} \qquad \text{for } \sigma_E^2 \qquad (27.28a)$$

$$\text{tr}(\widehat{\mathbf{P}}\mathbf{Z}\mathbf{A}\mathbf{Z}^T) = \mathbf{y}^T \widehat{\mathbf{P}}\mathbf{Z}\mathbf{A}\mathbf{Z}^T \widehat{\mathbf{P}} \mathbf{y} \qquad \text{for } \sigma_A^2 \qquad (27.28b)$$

Note that REML does not return estimates of $\boldsymbol{\beta}$, since the fixed effects are removed by setting $\boldsymbol{\beta}^* = \mathbf{0}$.

Since the transformation $\mathbf{y}^* = \mathbf{K}\mathbf{y}$ satisfying Equation 27.25 solely depends on the design matrix, this general approach still holds with m uncorrelated random vectors. In this case, Equation 27.8a expands to

$$\mathbf{y}^* = \sum_{i=1}^{m} \mathbf{K}\, \mathbf{Z}_i\, \mathbf{u}_i + \mathbf{K}\mathbf{e} \qquad (27.29)$$

and the REML equations for the $m + 1$ variance components become

$$\text{tr}(\widehat{\mathbf{P}}) = \mathbf{y}^T \widehat{\mathbf{P}}\widehat{\mathbf{P}} \mathbf{y} \qquad \text{for } \sigma_E^2 \qquad (27.30a)$$

$$\text{tr}(\widehat{\mathbf{P}}\, \mathbf{Z}_i\, \mathbf{B}_i\, \mathbf{Z}_i^T) = \mathbf{y}^T\, \widehat{\mathbf{P}}\, \mathbf{Z}_i\, \mathbf{B}_i\, \mathbf{Z}_i^T\, \widehat{\mathbf{P}} \mathbf{y} \qquad \text{for } \sigma_i^2,\ 1 \leq i \leq m \qquad (27.30b)$$

where $\widehat{\mathbf{P}}$ is now a function of $\widehat{\mathbf{V}} = \sum \widehat{\sigma}_i^2\, \mathbf{Z}_i\, \mathbf{B}_i\, \mathbf{Z}_i^T + \widehat{\sigma}_E^2\, \mathbf{I}$.

With REML, the information matrix contains only items corresponding to variance-component estimates, so $\mathbf{F} = \mathbf{S}$, where

$$S_{ij} = \frac{1}{2} \text{tr}(\, \mathbf{P}\mathbf{V}_i \mathbf{P}\mathbf{V}_j\,) \qquad (27.31a)$$

with $\mathbf{V}_i$ given by Equation 27.15. Estimates of the sampling variances and covariances of the variance-component estimates are obtained from the inverse of the matrix $\mathbf{S}$, as described above.

Example 3.   The REML variance-component estimates for the single-records dominance model of Example 2, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{d} + \mathbf{e}$, satisfy

$$\text{tr}(\widehat{\mathbf{P}}) = \mathbf{y}^T \widehat{\mathbf{P}}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_E^2$$

$$\text{tr}(\widehat{\mathbf{P}}\mathbf{A}) = \mathbf{y}^T \widehat{\mathbf{P}}\mathbf{A}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_A^2$$

$$\text{tr}(\widehat{\mathbf{P}}\mathbf{D}) = \mathbf{y}^T \widehat{\mathbf{P}}\mathbf{D}\widehat{\mathbf{P}}\mathbf{y} \qquad \text{for } \sigma_D^2$$

where $\widehat{\mathbf{P}}$ is defined as in Equation 27.17b with $\mathbf{V} = \widehat{\sigma}_A^2\,\mathbf{A} + \widehat{\sigma}_D^2\,\mathbf{D} + \widehat{\sigma}_E^2\,\mathbf{I}$. For purposes of estimating sampling variances and covariances of these estimates, the information matrix is given by

$$\mathbf{S} = \frac{1}{2} \begin{pmatrix} \text{tr}(\mathbf{PAPA}) & \text{tr}(\mathbf{PAPD}) & \text{tr}(\mathbf{PAP}) \\ \text{tr}(\mathbf{PAPD}) & \text{tr}(\mathbf{PDPD}) & \text{tr}(\mathbf{PDP}) \\ \text{tr}(\mathbf{PAP}) & \text{tr}(\mathbf{PDP}) & \text{tr}(\mathbf{PP}) \end{pmatrix}$$

When the estimate of $\mathbf{P}$ is inserted into this matrix, the standard errors of the variance-component estimates are obtained as the square roots of the diagonal elements of $\mathbf{S}^{-1}$, and the covariance between $\widehat{\sigma}_i^2$ and $\widehat{\sigma}_j^2$ is given by S_{ij}^{-1}.

SOLVING THE ML/REML EQUATIONS

Because the equations for the ML/REML solutions are highly nonlinear, closed analytical solutions are only available in very special cases (e.g., certain completely balanced designs). In principle, the solutions can be obtained by performing an exhaustive grid search — computing the log-likelihood of the data at each point on a grid covering the entire range of parameter space, and letting the solution be defined by the point on the grid giving the largest log-likelihood. However, this procedure is impractical under ML if $\boldsymbol{\beta}$ contains more than a few elements, since each element of $\boldsymbol{\beta}$ adds to the dimensionality of the search. Under REML, the dimensionality of parameter space can be greatly reduced, but the likelihood function is considerably more complicated to compute. Thus, simple grid searches are rarely used by themselves, although they are sometimes used in conjunction with other methods that restrict the search to one or a few dimensions.

   A wide variety of iterative techniques for solving ML/REML equations have been proposed based on various modifications of two basic approaches: the Newton-Raphson algorithm and the EM algorithm. Both procedures start with preliminary estimates of the parameters (obtained, for example, by ordinary least-squares analysis), and using information on the slope of the likelihood surface, these estimates are then moved in a direction that increases the log-likelihood of

the data. The revised estimates are subsequently modified in an iterative fashion, until a satisfactory degree of convergence on a final set of estimates has been achieved. With these types of approaches, the search for ML/REML solutions avoids spending huge amounts of computational time in regions of low likelihood. Such hill-climbing methods are not guaranteed to converge on the global maximum of the likelihood function, but potential problems with secondary peaks in the likelihood surface can be investigated through the use of different starting values.

Our review of numerical methods for obtaining solutions to the ML/REML equations is intentionally brief, focusing only on the general principles. All of the methods are very intensive computationally when large pedigrees are involved, as they usually require the inversion of large matrices at each step. Detailed reviews of this highly technical area appear in Meyer (1989b), Harville and Callanan (1990), and Searle et al. (1992).

Derivative-based Methods

The Newton-Raphson (NR) algorithm, a standard method for numerically solving coupled sets of nonlinear equations, has been used extensively to solve ML/REML equations (Harville 1977, Jennrich and Sampson 1976, Searle et al. 1992). Specific applications to genetic variance-component estimation include Lange et al. (1977) for ML estimates of additive and dominance variances for single characters and Meyer (1983, 1985) for REML estimates of the additive genetic covariance matrix for multiple characters. We confine our discussion of Newton-Raphson iteration to REML estimates, as applications to ML follow in a similar fashion.

The Newton-Raphson method obtains the REML estimate of the vector of parameters $\boldsymbol{\Theta}$ by starting with some initial value $\boldsymbol{\Theta}^{(0)}$ and then iterating to convergence to a final solution by using

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} - \left(\mathbf{H}^{(k)}\right)^{-1} \left.\frac{\partial L}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}^{(k)}} \tag{27.32}$$

where $\partial L / \partial \boldsymbol{\Theta}$ is a column vector of the partials of the log-likelihood function with respect to each parameter evaluated at the estimate $\boldsymbol{\Theta}^{(k)}$, and $\mathbf{H}$ is the **Hessian matrix** of all second-order partial derivatives of the log-likelihood L with respect to the variance components. $\mathbf{H}^{-1}$ and $\partial L / \partial \boldsymbol{\Theta}$ respectively provide measures of the curvature and the slope (and directionality) of the likelihood surface, given the current estimates. Their product gives a projected degree of movement of the vector $\boldsymbol{\Theta}$ towards an improved set of values to be used in the next iteration.

Consider again the mixed model with m random factors plus a residual,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{m} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e}$$

where $\mathbf{u}_i \sim \text{MVN}(\mathbf{0}, \sigma_i^2 \mathbf{B}_i)$ for $1 \leq i \leq m$, and $\mathbf{B}_i$ is a square symmetric $n_i \times n_i$ matrix of known constants. The residuals are also assumed to be multivariate

normal with $\mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \sigma_E^2\,\mathbf{I})$. Since $\mathbf{y}$ is the sum of multivariate normals, it is also multivariate normal with $\mathbf{y} \sim \mathrm{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, where

$$\mathbf{V} = \sum_{i=1}^{m} \sigma_i^2\,\mathbf{Z}_i\,\mathbf{B}_i\,\mathbf{Z}_i^T + \sigma_E^2\,\mathbf{I}$$

Under REML, $\boldsymbol{\Theta} = (\sigma_1^2,\,\sigma_2^2,\,\cdots,\,\sigma_E^2)^T$ and Equations 27.14b and 27.27b give the elements of $\partial L / \partial \boldsymbol{\Theta}$ as

$$\left.\frac{\partial L}{\partial \sigma_i^2}\right|_{\boldsymbol{\Theta}^{(k)}} = -\frac{1}{2}\operatorname{tr}\left(\mathbf{P}^{(k)}\mathbf{V}_i\right) + \frac{1}{2}\mathbf{y}^T\mathbf{P}^{(k)}\mathbf{V}_i\mathbf{P}^{(k)}\mathbf{y} \qquad (27.33)$$

where $\mathbf{V}_i$ is given by Equation 27.15 and $\mathbf{P}^{(k)}$ is calculated from Equation 27.17b using the current variance-component estimates in $\boldsymbol{\Theta}^{(k)}$. Searle et al. (1992) give the elements of $\mathbf{H}$ for REML as

$$\mathbf{H}_{ij}^{(k)} = \frac{\partial^2 L}{\partial \sigma_i^2\,\partial \sigma_j^2} = \frac{1}{2}\operatorname{tr}\left(\mathbf{P}^{(k)}\,\mathbf{V}_i\,\mathbf{P}^{(k)}\,\mathbf{V}_j\right) - \mathbf{y}^T\mathbf{P}^{(k)}\,\mathbf{V}_i\,\mathbf{P}^{(k)}\,\mathbf{V}_j\,\mathbf{P}^{(k)}\,\mathbf{y} \quad (27.34)$$

where again the partials are evaluated using $\boldsymbol{\Theta}^{(k)}$.

A common variant of the Newton-Raphson algorithm is **Fisher's scoring method**, which replaces the inverse of the Hessian matrix in Equation 27.32 by its expected value, which after allowing for a change in sign, turns out to be defined by the inverse of Fisher's information matrix, $-\mathbf{F}^{-1}$ (Equation 27.20). This reduces the iterative equation to

$$\boldsymbol{\Theta}^{(k+1)} = \boldsymbol{\Theta}^{(k)} + \left(\mathbf{F}^{(k)}\right)^{-1}\left.\frac{\partial L}{\partial \boldsymbol{\Theta}}\right|_{\boldsymbol{\Theta}^{(k)}} \qquad (27.35\text{a})$$

with

$$F_{ij}^{(k)} = \frac{1}{2}\operatorname{tr}(\mathbf{P}^{(k)}\,\mathbf{V}_i\,\mathbf{P}^{(k)}\,\mathbf{V}_j) \qquad (27.35\text{b})$$

There are several motivations for employing this modification. First, as noted above, the inverse of the information matrix, when evaluated at the REML values, estimates the standard errors for these estimates. Second, $\mathbf{F}$ is easier to compute than $\mathbf{H}^{-1}$ (compare Equations 27.34 and 27.35b). Finally, Fisher's scoring method appears to be slightly more robust to initial values than strict Newton-Raphson iteration (Jennrich and Sampson 1976).

Example 4. Again consider the simple animal model with a single observation per individual, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{e}$. For REML estimates, letting

$$\boldsymbol{\Theta}^{(k)} = \begin{pmatrix} (\sigma_A^2)^{(k)} \\ (\sigma_E^2)^{(k)} \end{pmatrix} \qquad \text{gives} \qquad \frac{\partial L}{\partial \boldsymbol{\Theta}} = \frac{1}{2}\begin{pmatrix} -\operatorname{tr}(\mathbf{P}) + \mathbf{y}^T\mathbf{P}\mathbf{P}\mathbf{y} \\ -\operatorname{tr}(\mathbf{P}\mathbf{A}) + \mathbf{y}^T\mathbf{P}\mathbf{A}\mathbf{P}\mathbf{y} \end{pmatrix}$$

Note that $\mathbf{P}$ is a function of the current variance-component estimates, with

$$\mathbf{P} = (\mathbf{V}^{-1})^{(k)} - (\mathbf{V}^{-1})^{(k)}\mathbf{X}(\mathbf{X}^T(\mathbf{V}^{-1})^{(k)}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{V}^{-1})^{(k)}$$

where

$$\mathbf{V}^{(k)} = (\sigma_A^2)^{(k)}\,\mathbf{A} + (\sigma_E^2)^{(k)}\,\mathbf{I}$$

with $\mathbf{A}$ being the relationship matrix for the inviduals being measured. Likewise, from Equation 27.34 the Hessian matrix $\mathbf{H}$ is given by

$$\left.\frac{\partial^2 L}{\partial\boldsymbol{\Theta}^2}\right|_{\boldsymbol{\Theta}^{(k)}} = \frac{1}{2}\begin{pmatrix} \text{tr}(\mathbf{PP}) - 2\mathbf{y}^T\mathbf{PPPy} & \text{tr}(\mathbf{PAP}) - 2\mathbf{y}^T\mathbf{PAPPy} \\ \text{tr}(\mathbf{PAP}) - 2\mathbf{y}^T\mathbf{PAPPy} & \text{tr}(\mathbf{PAPA}) - 2\mathbf{y}^T\mathbf{PAPAPy} \end{pmatrix}$$

and the Fisher information matrix by

$$\mathbf{F} = -E\left(\frac{\partial^2 L}{\partial\boldsymbol{\Theta}^2}\right) = \frac{1}{2}\begin{pmatrix} \text{tr}(\mathbf{PP}) & \text{tr}(\mathbf{PAP}) \\ \text{tr}(\mathbf{PPA}) & \text{tr}(\mathbf{PAPA}) \end{pmatrix}$$

Note that $\mathbf{P}$ is really indexed by k since it depends on the current estimates of the unknown variance components, $\widehat{\sigma}_A^2$ and $\widehat{\sigma}_E^2$.

EM Methods

The idea behind the EM (**expectation/maximization**) algorithm for variance-component analysis is that if we knew the values of the random effects, we could estimate the variances in a simple fashion directly from them. Focusing on the general mixed model defined by Equation 27.8a, the variances of the random and residual effects are defined respectively to be

$$\sigma_i^2 = \frac{E[\mathbf{u}_i^T\mathbf{B}_i^{-1}\mathbf{u}_i]}{n_i} \tag{27.36a}$$

$$\sigma_E^2 = \frac{E[\mathbf{e}_i^T\mathbf{e}_i]}{n} \tag{27.36b}$$

where n and n_i are, respectively, the number of elements in $\mathbf{e}$ and $\mathbf{u}_i$. Equation 27.36a follows from Equation 8.22, which, since $E[\mathbf{u}_i] = \mathbf{0}$, reduces to

$$E[\mathbf{u}_i^T\mathbf{B}_i^{-1}\mathbf{u}_i] = \text{tr}(\mathbf{B}_i^{-1}\sigma_i^2\mathbf{B}_i) = \sigma_i^2\text{tr}(\mathbf{I}_{n_i}) = n_i\,\sigma_i^2$$

The last two steps follow from Equations 27.12a and 27.12b, respectively. Equation 27.36b follows in a similar fashion. In actuality, of course, we only know $\mathbf{y}$, not the underlying vectors of random effects ($\mathbf{u}_i$) or residual deviations ($\mathbf{e}$).

Underlying the EM algorithm is the idea, discussed in Chapter 26, that the information in $\mathbf{y}$ provides a basis for making predictions about the elements of $\mathbf{u}_i$ and $\mathbf{e}$. In the context of variance-component analysis, we need to go a step beyond BLUP estimation of $\mathbf{u}_i$ and $\mathbf{e}$, as it is actually the quadratic products of $\mathbf{u}_i$ and $\mathbf{e}$ in the numerators of Equations 27.36a,b that we need to predict. Here, in the interest of clarity and space, we skip over a number of steps to the final solution (see Searle et al. 1992, pp. 297–304 for a complete derivation). Searle et al. show that the conditional distribution of $\mathbf{u}$ given the observed $\mathbf{y}$ is MVN, with

$$E[\mathbf{u}_i \,|\, \mathbf{y}] = \sigma_i^2 \, \mathbf{Z}_i^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sigma_i^2 \, \mathbf{Z}_i^T \mathbf{P}\mathbf{y}$$

and

$$\sigma^2(\mathbf{u}_i \,|\, \mathbf{y}) = \sigma_i^2 \, \mathbf{I}_{n_i} - \sigma_i^4 \, \mathbf{Z}_i^T \mathbf{V}^{-1} \mathbf{Z}_i$$

Substituting into Equation 8.22, after some simplication, the expectation of the quadratic product in Equation 27.36a, conditional on the observed $\mathbf{y}$, becomes

$$E[\, \mathbf{u}_i^T \mathbf{B}_i^{-1} \mathbf{u}_i \,|\, \mathbf{y}\,] = n_i \sigma_i^2 + \sigma_i^4 \, [\, \mathbf{y}^T \, \mathbf{P}\mathbf{V}_i \mathbf{P}\mathbf{y} - \mathrm{tr}(\, \mathbf{P}\mathbf{V}_i\,)\,] \qquad (27.37\mathrm{a})$$

where $\mathbf{V}_i = \mathbf{Z}_i \, \mathbf{B}_i \, \mathbf{Z}_i^T$ as given by Equation 27.15. Similar logic gives

$$E[\, \mathbf{e}^T \mathbf{e} \,|\, \mathbf{y}\,] = n\sigma_E^2 + \sigma_E^4 \, [\, \mathbf{y}^T \, \mathbf{P}\mathbf{P}\mathbf{y} - \mathrm{tr}(\, \mathbf{P})\,] \qquad (27.37\mathrm{b})$$

These expressions define expected quadratic values, conditional on the particular set of observations $\mathbf{y}$, under the assumption that the true variance components are known. The astute reader will immediately notice that our problem has hardly been solved, since we are trying to estimate the variance components.

The EM algorithm (Dempster et al. 1977) attempts to circumvent this problem by starting with some initial estimates of the variance components, and then substituting these as well as $\mathbf{y}$ into Equations 27.37a,b to obtain estimates of the quadratic products. These latter estimates are then substituted into Equations 27.36a,b to obtain improved estimates of the variance components, and then the entire process is repeated again and again until satisfactory convergence has been achieved. Defining the quantities estimated by Equations 27.37a,b in the kth iteration as $\widehat{q}_i^{(k)}$, and $\widehat{q}_E^{(k)}$, the EM algorithm can be summarized as follows: (1) the E step computes the *expected* quadratic products conditional upon $\mathbf{y}$, $\widehat{q}_i^{(k)}$, and $\widehat{q}_E^{(k)}$, and (2) the M step substitutes these conditional expectations into the *maximum* likelihood estimators (Equations 27.36a,b) to generate the next round of REML variance-component estimates, $(\widehat{\sigma}_E^2)^{(k+1)}$ and $(\widehat{\sigma}_i^2)^{(k+1)}$, which are then applied to the next E step. The final REML estimates are achieved when $(\widehat{\sigma}_E^2)^{(k)} \simeq (\widehat{\sigma}_E^2)^{(k+1)}$ and $(\widehat{\sigma}_i^2)^{(k)} \simeq (\widehat{\sigma}_i^2)^{(k+1)}$.

That estimates obtained via the EM algorithm do indeed correspond to the REML solutions can be seen by recalling the REML Equations 27.30a,b. Upon convergence of the EM algorithm, the terms in brackets on the right sides of Equations

27.37a,b must be equal to zero, which is equivalent to the REML solutions. When convergence is reached, the estimates of the variance components are used to obtain the final estimate of $\widehat{\mathbf{V}}$, and the vector of fixed effects is then estimated by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \widehat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{V}}^{-1} \mathbf{y}$$

In general, solutions via the EM algorithm can take considerably more iterations to converge than those via Newton-Raphson iteration, especially when heritabilities are low. Moreover, as with the Newton-Raphson algorithm, the EM algorithm is by no means guaranteed to converge on the REML solution; it sometimes generates multiple solutions for different starting conditions (Groeneveld and Kovac 1990b). Such problems can result from multiple peaks in the likelihood surface. Since the EM method in essence uses the first derivatives of the likelihood function to adjust the variance-component estimates (compare Equation 27.33 and the terms in brackets in Equations 27.37a,b), it can get stuck on inflection points in the likelihood surface as well. Rounding errors can also compromise the iterative solutions (Boichard et al. 1992). As in the case of derivative-based methods, many of these problems can be minimized by performing multiple analyses from different starting points.

Example 5.   Consider again the animal model with dominance and a single record per individual, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{a} + \mathbf{d} + \mathbf{e}$. The EM equations for the REML estimates of σ_A^2, σ_D^2, and σ_E^2 are

$$(\widehat{\sigma}_A^2)^{(k+1)} = (\widehat{\sigma}_A^2)^{(k)} + \frac{(\widehat{\sigma}_A^4)^{(k)}}{n} \cdot \left\{ \mathbf{y}^T \mathbf{P}^{(k)} \mathbf{A} \mathbf{P}^{(k)} \mathbf{y} - \operatorname{tr}[\mathbf{P}^{(k)} \mathbf{A}] \right\}$$

$$(\widehat{\sigma}_D^2)^{(k+1)} = (\widehat{\sigma}_D^2)^{(k)} + \frac{(\widehat{\sigma}_D^4)^{(k)}}{n} \left\{ \mathbf{y}^T \mathbf{P}^{(k)} \mathbf{D} \mathbf{P}^{(k)} \mathbf{y} - \operatorname{tr}[\mathbf{P}^{(k)} \mathbf{D}] \right\}$$

$$(\widehat{\sigma}_E^2)^{(k+1)} = (\widehat{\sigma}_E^2)^{(k)} + \frac{(\widehat{\sigma}_E^4)^{(k)}}{n} \left\{ \mathbf{y}^T \mathbf{P}^{(k)} \mathbf{P}^{(k)} \mathbf{y} - \operatorname{tr}[\mathbf{P}^{(k)}] \right\}$$

where $\mathbf{P}^{(k)}$ is defined by Equation 27.17b using $\mathbf{V}^{(k)}$ for $\mathbf{V}$ where

$$\mathbf{V}^{(k)} = (\widehat{\sigma}_A^2)^{(k)} \mathbf{A} + (\widehat{\sigma}_D^2)^{(k)} \mathbf{D} + (\widehat{\sigma}_E^2)^{(k)} \mathbf{I}$$