MIXED MODELS

This chapter introduces best linear unbiased prediction (BLUP), a general method for predicting random effects, while Chapter 27 is concerned with the estimation of variances by restricted maximum likelihood (REML). These two methods are related in that BLUP assumes that the appropriate variance components are known, while REML procedures estimate variance components in an iterative fashion from BLUP estimates of random effects. Although the basic properties of these techniques have been known for decades, because of their computational demands, their practical application is a fairly recent phenomenon. BLUP is now by far the dominant methodology for estimating breeding values.

After a brief introduction to the general mixed model, we will develop expressions for BLUEs (best linear unbiased estimators) of fixed effects and for BLUPs of random effects under the assumption that variances are known in the base population.

THE GENERAL MIXED MODEL

Consider a column vector $\mathbf{y}$ containing the phenotypic values for a trait measured in n individuals. We assume that these observations are described adequately by a linear model with a $p \times 1$ vector of fixed effects ($\boldsymbol{\beta}$) and a $q \times 1$ vector of random effects ($\mathbf{u}$). The first element of the vector $\boldsymbol{\beta}$ is typically the population mean, and other factors included may be gender, location, year of birth, experimental treatment, and so on. The elements of the vector $\mathbf{u}$ of random effects are usually genetic effects such as additive genetic values. In matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{26.1}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are respectively $n \times p$ and $n \times q$ **incidence matrices** ($\mathbf{X}$ is also called the **design matrix**), and $\mathbf{e}$ is the $n \times 1$ column vector of residual deviations assumed to be distributed independently of the random genetic effects. Usually, all of the elements of the incidence matrices are equal to 0 or 1, depending upon whether the relevant effect contributes to the individual's phenotype. Because this model jointly accounts for fixed and random effects, it is generally referred to as a **mixed model** (Eisenhart 1947). Analysis of Equation 26.1 forms the basis for the remainder of this chapter and the next.

Example 1. Suppose that three sires are chosen at random from a population,

and each mated to a randomly chosen dam. Two offspring from each mating are evaluated, some in environment 1 and some in environment 2. Let y_{ijk} denote the phenotypic value of the kth offspring of sire i in environment j. The model is then

$$y_{ijk} = \beta_j + u_i + e_{ijk}$$

This model has three random effects (u_1, u_2, u_3), which measure the contribution from each sire, and two fixed effects (β_1, β_2), which describe the influence of the two environments. The model assumes an absence of sire $\times$ environment interaction.

As noted above, a total of six offspring were measured. One offspring of sire 1 was assigned to environment 1 and had phenotypic value $y_{1,1,1} = 9$, while the second offspring was assigned to environment 2 and had phenotypic value $y_{1,2,1} = 12$. The two offspring of sire 2 were both assigned to environment 1 and had values of $y_{2,1,1} = 11$ and $y_{2,1,2} = 6$. One offspring of sire 3 was assigned to environment 1 and had phenotypic value $y_{3,1,1} = 7$, while the second offspring was assigned to environment 2 and had phenotypic value $y_{3,2,1} = 14$. The resulting vector of observations can be written as

$$\mathbf{y} = \begin{pmatrix} y_{1,1,1} \\ y_{1,2,1} \\ y_{2,1,1} \\ y_{2,1,2} \\ y_{3,1,1} \\ y_{3,2,1} \end{pmatrix} = \begin{pmatrix} 9 \\ 12 \\ 11 \\ 6 \\ 7 \\ 14 \end{pmatrix}$$

giving the mixed model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

where the incidence matrices for fixed and random effects and the vectors of these effects are respectively

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Now consider the means and variances of the component vectors of the mixed model. Since $E(\mathbf{u}) = E(\mathbf{e}) = \mathbf{0}$ by definition, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Denote the $(n \times n)$ covariance matrix for the vector $\mathbf{e}$ of residual errors by $\mathbf{R}$ and the $(q \times q)$

covariance matrix for the vector $\mathbf{u}$ of random genetic effects by $\mathbf{G}$. Excluding the difference among individuals due to fixed effects, from Equation 8.21b and the assumption that $\mathbf{u}$ and $\mathbf{e}$ are uncorrelated, the covariance matrix for the vector of observations $\mathbf{y}$ is

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R} \tag{26.2}$$

The first term accounts for the contribution from random genetic effects, while the second accounts for the variance due to residual effects. We will generally assume that residual errors have constant variance and are uncorrelated, so that $\mathbf{R}$ is a diagonal matrix, with $\mathbf{R} = \sigma_E^2\,\mathbf{I}$.

We are now in a position to contrast the mixed model and the general linear model. Under the general linear model (Chapter 8),

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^* \qquad \text{where} \quad \mathbf{e}^* \sim (\mathbf{0}, \mathbf{V}) \quad \text{implying} \;\; \mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

where the notation $\sim\ (a, b)$ means that the random variable has mean a and variance b. On the other hand, the mixed model partitions the vector of residual effects into two components, with $\mathbf{e}^* = \mathbf{Z}\mathbf{u} + \mathbf{e}$, giving

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \qquad \text{where} \quad \mathbf{u} \sim (\mathbf{0}, \mathbf{G}) \;\text{ and }\; \mathbf{e} \sim (\mathbf{0}, \mathbf{R})$$

$$\text{implying} \;\; \mathbf{y} \sim (\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) = (\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R})$$

When analyzed in the appropriate way, both formulations yield the same estimate of the vector of fixed effects $\boldsymbol{\beta}$, while the mixed-model formulation further allows estimates of the vector of random effects $\mathbf{u}$.

For the mixed model, we observe $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$, while $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{R}$, and $\mathbf{G}$ are generally unknown. Thus, mixed-model analysis involves two complementary estimation issues: (1) estimation of the vectors of fixed and random effects, $\boldsymbol{\beta}$ and $\mathbf{u}$, and (2) estimation of the covariance matrices $\mathbf{G}$ and $\mathbf{R}$. These covariance matrices are generally assumed to be functions of a few unknown variance components. For the remainder of this chapter, we consider estimators of $\boldsymbol{\beta}$ and $\mathbf{u}$ under the assumption that $\mathbf{y}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{R}$ are all known. Estimation of the variance components (and hence $\mathbf{R}$ and $\mathbf{G}$) from $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$ is the subject of the next chapter.

Estimating Fixed Effects and Predicting Random Effects

As outlined in the preceding chapters, the primary goal of a quantitative-genetic analysis is often solely to estimate variance components. However, there are also numerous situations in which inferences about fixed effects (such as the effect of a particular environment or year) and / or random effects (such as the breeding value of a particular individual) are the central motivation. Inferences about fixed effects have come to be called **estimates,** whereas those that concern random effects are known as **predictions.** Procedures for obtaining such estimators and predictors

have been developed using a variety of approaches, such as likelihood theory (Appendix 4). The most widely used procedures are BLUE and BLUP, referring respectively to **best linear unbiased estimator** and **best linear unbiased predictor.** They are *best* in the sense that they minimize the sampling variance, *linear* in the sense that they are linear functions of the observed phenotypes $\mathbf{y}$, and *unbiased* in the sense that $E[\,\text{BLUE}(\boldsymbol{\beta})\,] = \boldsymbol{\beta}$ and $E[\,\text{BLUP}(\mathbf{u})\,] = \mathbf{u}$.

For the mixed model given by Equation 26.1, the BLUE of $\boldsymbol{\beta}$ is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{26.3}$$

with $\mathbf{V}$ as given by Equation 26.2. Notice that this is just the generalized least-squares (GLS) estimator discussed in Chapter 8. Henderson (1963) showed that the BLUP of $\mathbf{u}$ is

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{26.4}$$

which is equivalent to the conditional expectation of $\mathbf{u}$ given $\mathbf{y}$ under the assumption of multivariate normality (cf. Equation 8.27). As noted above, the practical application of both of these expressions requires that the variance components be known. Thus, prior to a BLUP analysis, the variance components need to be estimated by ANOVA or REML.

Example 2. What are the BLUP values for the sire effects (u_1, u_2, u_3) in Example 1? In order to proceed, we require the covariance matrices for sire effects and errors. We assume that the residual variances within both environments are the same (σ^2_E), so $\mathbf{R} = \sigma^2_E\,\mathbf{I}$, where $\mathbf{I}$ is the 6×6 identity matrix. Assuming that all three sires are unrelated and drawn from the same population, $\mathbf{G} = \sigma^2_S\,\mathbf{I}$, where $\mathbf{I}$ is the 3×3 identity matrix and σ^2_S is the variance of sire effects. Assuming only additive genetic variance, the sire effects (breeding values) are half the sires' additive genetic values. Thus, since the sires are sampled randomly from an outbred base population, $\sigma^2_S = \sigma^2_A/4$, where σ^2_A is the additive genetic variance. Assuming that $\sigma^2_A = 8$ and $\sigma^2_E = 6$, the covariance matrix $\mathbf{V}$ for the vector of observations $\mathbf{y}$ is given by $\mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$, or

$$\mathbf{V} = \frac{8}{4}\begin{pmatrix}1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}1&1&0&0&0&0\\0&0&1&1&0&0\\0&0&0&0&1&1\end{pmatrix} + 6\begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}$$

$$= \begin{pmatrix} 8 & 2 & 0 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & 0 & 2 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{pmatrix} \quad \text{giving} \quad \mathbf{V}^{-1} = \frac{1}{30} \cdot \begin{pmatrix} 4 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{pmatrix}$$

Using this result, a few simple matrix calculations give

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \end{pmatrix}$$

and

$$\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \frac{1}{18}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Example 3. As mentioned in Chapter 13, the effects of different genotypes at a single QTL are often estimated by ordinary least squares (OLS), using the model

$$y_{ij} = g_i + e_{ij}$$

where y_{ij} is the observed phenotype of the jth individual of genotype i, g_i is the mean genotypic value for the ith genotype at the locus of interest, and e_{ij} is a residual deviation assumed to be independently distributed among individuals. While this model may be reasonable for a random collection of individuals from a large population, when some sampled individuals are relatives, the sharing of alleles at other loci influencing the trait will induce correlations between residuals. If this is the case, OLS analysis can produce biased estimates of the QTL effects. When one of the QTL genotypes is very rare, as is often the case, the sampled individuals may be intentionally selected from the same pedigree, so the problem of bias is not trivial.

Use of a mixed model provides a means for accounting for associations among background QTLs in a way that eliminates bias in estimates of QTL effects. If the relatives in question share only additive effects (as in a pedigree with no full sibs or double first cousins, or when there is no nonadditive gene action), the correlations among residuals are accounted for by the additive genetic relationship matrix $\mathbf{A}$, where A_{ij} is twice the coefficient of coancestry, $2\Theta_{ij}$. When sibs are included and dominance is present at background QTLs, both $\mathbf{A}$ and a dominance relationship matrix (see below) are required.

Here we assume that all of the background genetic effects are additive, in which case the simplest mixed model can be applied,

$$y_{ij} = g_i + a_{ij} + e_{ij}$$

with the contribution from the different single-locus genotypes (g_i) being treated as fixed effects. The additive genetic background effects (a_{ij}) and the residual environmental deviations (e_{ij}) are treated as random effects, both with expected values equal to zero, and with respective variances σ_A^2 and σ_E^2. Note that σ_A^2 is the background additive genetic variance for the trait in excess of that caused by the QTL.

In matrix form,

$$\mathbf{y} = \mathbf{Xg} + \mathbf{Za} + \mathbf{e}$$

If there is a single observation for each individual, as we assume below, then $\mathbf{Z} = \mathbf{I}$ and the covariance matrix for the vector of observations $(\mathbf{y})$ is

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

Thus, the covariance between the residual errors of two individuals (i and j) is just $2\Theta_{ij}\sigma_A^2$, while the variance of individual errors is $\sigma_A^2 + \sigma_E^2$. The error in using OLS to estimate single gene effects is that $\mathbf{A}$ is assumed to equal an identity matrix, so that $\mathbf{V}$ is incorrectly assumed to be a diagonal matrix.

From Equation 26.3, the estimates of the QTL means are given by

$$\widehat{\mathbf{g}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

Kennedy et al. (1992) showed that mixed-model estimates of QTL effects are much more reliable than OLS estimates, especially in small selected populations. Building on this approach, several authors (Hoeschele 1988, Hofer and Kennedy 1993, Kinghorn et al. 1993) have proposed BLUP-based segregation analysis for estimating the effects of an unknown major gene. Here the elements in the design matrix $\mathbf{X}$ associated with g_i are probabilistic estimates for the major-locus genotypes of each individual.

Note that the solution of Equations 26.3 and 26.4 requires the inverse of the covariance matrix $\mathbf{V}$. In the preceding example, $\mathbf{V}^{-1}$ was not particularly difficult to obtain. However, when $\mathbf{y}$ contains many thousands of observations, as is commonly the case in cattle breeding, the computation of $\mathbf{V}^{-1}$ can be quite difficult. As a way around this problem, Henderson (1950, 1963, 1973, 1984a)

offered a more compact method for jointly obtaining $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ in the form of his **mixed-model equations** (MME),

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{26.5}$$

While these expressions may look considerably more complicated than Equations 26.3 and 26.4, $\mathbf{R}^{-1}$ and $\mathbf{G}^{-1}$ are trivial to obtain if $\mathbf{R}$ and $\mathbf{G}$ are diagonal, and hence the submatrices in Equation 26.5 are much easier to compute than $\mathbf{V}^{-1}$. A second advantage of Equation 26.5 can be seen by considering the dimensionality of the matrix on the left. Recalling that $\mathbf{X}$ and $\mathbf{Z}$ are $n \times p$ and $n \times q$ respectively, $\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}$ is $p \times p$, $\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}$ is $p \times q$, and $\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}$ is $q \times q$. Thus, the matrix that needs to be inverted to obtain the solution for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ is of order $(p+q) \times (p+q)$, which is usually considerably less than the dimensionality of $\mathbf{V}$ (an $n \times n$ matrix).

Although there are several ways to derive the mixed-model equations (Robinson 1991), Henderson (1950) originally obtained them by assuming that the covariance matrices $\mathbf{G}$ and $\mathbf{R}$ are known and that the densities of the vectors $\mathbf{u}$ and $\mathbf{e}$ are each multivariate normal with no correlations between them. Equation 26.5 then yields the maximum likelihood estimates of the fixed and random effects. Henderson (1963) later showed that the mixed-model equations do not actually depend on normality, and that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$ are BLUE and BLUP, respectively, under general conditions provided the variances are known.

Example 4. Using the values from Examples 1 and 2, we find that

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} = \frac{1}{6}\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} = \left(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\right)^T = \frac{1}{6}\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{G}^{-1}+\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} = \frac{5}{6}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 33 \\ 26 \end{pmatrix}, \quad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 21 \\ 17 \\ 21 \end{pmatrix}$$

Thus, after factoring out $1/6$ from both sides, the mixed-model equations for these data become

$$\begin{pmatrix} 4 & 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 1 & 5 & 0 & 0 \\ 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix}$$

Taking the inverse gives the solution

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \frac{1}{270} \begin{pmatrix} 100 & 25 & -25 & -40 & -25 \\ 25 & 175 & -40 & -10 & -40 \\ -25 & -40 & 67 & 10 & 13 \\ -40 & -10 & 10 & 70 & 10 \\ -25 & -40 & 13 & 10 & 67 \end{pmatrix} \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix} = \frac{1}{18} \begin{pmatrix} 148 \\ 235 \\ -1 \\ 2 \\ -1 \end{pmatrix}$$

which is identical to the results obtained in Example 2.

Although the method of predicting random effects using BLUP methodology was first discussed by Henderson (1949, 1950), the expression "best linear unbiased predictor" was apparently first used by Goldberger (1962), with the acronym BLUP due to Henderson (1973). In a relatively short time, BLUP has become the method of choice for estimating the breeding values of individuals from field records of large and complex pedigrees. For BLUPs to be the best unbiased estimates, the appropriate genetic variances must be known without error. Kackar and Harville (1981) show that BLUP estimates remain unbiased when estimates of genetic variances are used in place of actual values (as is usually the case), although they are not guaranteed to be the best of all unbiased linear estimators.

Estimability of Fixed Effects

It is sometimes impossible to obtain unique BLUE estimates for all of the fixed factors in a model. Suppose, for example, that

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \text{with} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Here, factors 1 and 2 are completely confounded, as they contribute equally to all individuals, so unique estimates of β_1 and β_2 cannot be acquired. Generally, when two or more columns of $\mathbf{X}$ are not independent, it is still possible to obtain unique BLUEs for certain linear combinations of $\boldsymbol{\beta}$ through the use of **generalized inverses** (Appendix 3). With the preceding design matrix $\mathbf{X}$, the solution is simple — by combining the two factors into a single new factor, $\beta_1 + \beta_2$, the new model becomes

$$\boldsymbol{\beta}_* = \begin{pmatrix} \beta_1 + \beta_2 \\ \beta_3 \end{pmatrix} \quad \text{with} \quad \mathbf{X}_* = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Since the columns of $\boldsymbol{\beta}_*$ are now independent, a unique solution exists for $\mathbf{X}_*^T \mathbf{V}^{-1} \mathbf{X}_*$, and from Equation 26.3, the two BLUEs of the fixed effects are given by

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

Situations in which linear combinations of fixed effects are required commonly arise when a very large number of fixed factors are included in the model, as occurs in large breeding programs involving multiple environments (such as different herds and different years.) Henderson (1984a) provides an extended discussion of the issues. Throughout the remainder of the book, we assume that $\boldsymbol{\beta}$ is estimable, either immediately or after an appropriate transformation. Appendix 3 discusses how to determine which combinations of effects are estimable when singular matrices exist.

Standard Errors

A relatively straightforward extension of Henderson's mixed-model equations provides estimates of the standard errors of the fixed and random effects. Let the inverse of the leftmost matrix in Equation 26.5 be

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{26.6}$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $p \times p$, $p \times q$, and $q \times q$ submatrices. Using this notation, Henderson (1975) showed that the sampling covariance matrix for the BLUE of $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}) = \mathbf{C}_{11} \tag{26.7a}$$

that the sampling covariance matrix of the prediction errors $(\widehat{\mathbf{u}} - \mathbf{u})$ is given by

$$\boldsymbol{\sigma}(\,\widehat{\mathbf{u}} - \mathbf{u}\,) = \mathbf{C}_{22} \tag{26.7b}$$

and that the sampling covariance of estimated effects and prediction errors is given by

$$\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}} - \mathbf{u}\,) = \mathbf{C}_{12} \tag{26.7c}$$

(We consider $\widehat{\mathbf{u}} - \mathbf{u}$ rather than $\widehat{\mathbf{u}}$ as the latter includes variance from both the prediction error and the random effects $\mathbf{u}$ themselves.) The standard errors of the fixed and random effects are obtained, respectively, as the square roots of the diagonal elements of $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$. For very large animal breeding designs where the inverse of the MME matrix may be difficult to compute, Meyer (1989a) presents methods for approximating the diagonal elements of the inverse of this matrix (and hence the standard errors).

Example 5. Consider the mixed-model equation from Example 4. Here for the fixed factors β_1, β_2 and the random effects u_1, u_2, u_3, the inverse of the coefficient matrix is

$$\begin{pmatrix} 4 & 0 & \vdots & 1 & 2 & 1 \\ 0 & 2 & & 1 & 0 & 1 \\ \cdots & \cdots & & \cdots & & \\ 1 & 1 & & 5 & 0 & 0 \\ 2 & 0 & \vdots & 0 & 5 & 0 \\ 1 & 1 & & 0 & 0 & 5 \end{pmatrix}^{-1} = \frac{1}{270} \begin{pmatrix} 100 & 25 & \vdots & -25 & -40 & -25 \\ 25 & 175 & & -40 & -10 & -40 \\ \cdots & & \cdots & & \cdots & \\ -25 & -40 & & 67 & 10 & 13 \\ -40 & -10 & \vdots & 10 & 70 & 10 \\ -25 & -40 & & 13 & 10 & 67 \end{pmatrix}$$

Hence,

$$\mathbf{C}_{11} = \frac{1}{270} \begin{pmatrix} 100 & 25 \\ 25 & 175 \end{pmatrix} \quad \text{and} \quad \mathbf{C}_{22} = \frac{1}{270} \begin{pmatrix} 67 & 10 & 13 \\ 10 & 70 & 10 \\ 13 & 10 & 67 \end{pmatrix}$$

so that, for example,

$$\sigma^2(\widehat{\beta_1}) = \frac{100}{270}, \quad \sigma^2(\widehat{\beta_2}) = \frac{175}{270}, \quad \sigma(\widehat{\beta_1}, \widehat{\beta_2}) = \frac{25}{270}$$

and, likewise,

$$\sigma^2(\widehat{u_2} - u_2) = \frac{70}{270}, \quad \sigma(\widehat{u_1} - u_1, \widehat{u_3} - u_3) = \frac{13}{270}, \quad \text{and so on.}$$

JOINT ESTIMATION OF SEVERAL VECTORS OF RANDOM EFFECTS

The mixed-model equations can be easily extended to situations where two (or more) vectors of random effects are of interest, as for example, in the estimation of both additive and dominance values or in the estimation of breeding values and maternal effects. With two vectors of random effects ($\mathbf{u}_1$ and $\mathbf{u}_2$) uncorrelated with each other, the mixed model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e} \tag{26.19a}$$

The vectors of random effects can have different dimensions (q_1 for $\mathbf{u}_1$, q_2 for $\mathbf{u}_2$), so with n individuals in the vector $\mathbf{y}$, the incidence matrix $\mathbf{Z}_i$ is $n \times q_i$ (for $i = 1, 2$). Letting $\mathbf{R}$ be the $n \times n$ covariance matrix for the vector of residual errors $\mathbf{e}$, and $\mathbf{G}_i$ be the $q_i \times q_i$ covariance matrix for $\mathbf{u}_i$, the MMEs become

$$\begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{G}_1^{-1} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{G}_2^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}}_1 \\ \widehat{\mathbf{u}}_2 \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \tag{26.19b}$$

Equation 26.19b can be extended in an obvious fashion to incorporate additional uncorrelated vectors of random effects. The following sections outline a few common applications of this extension of the mixed model.

Repeated Records

Another situation in which correlations are expected among residual errors arises when multiple observations are made on individuals, a common procedure used to reduce measurement error. Here, assuming dominance is of negligible importance, the residual error can be described as $p + e$, where p is the "permanent" environmental effect common to all observations on the same individual, and e is the residual error between observations of the same individual due, for example, to measurement error and changes in some environmental factors. Recall from Chapter 6 that the repeatability of a character (r) is defined to be the correlation between different measurements in the same individual. If j and k denote different observations on the same individual i, the covariance between repeated measurements is

$$\begin{aligned} \sigma(y_{ij}, y_{ik}) = r\sigma_y^2 &= \sigma(a_i + p_i + e_{ij}, a_i + p_i + e_{ik}) \\ &= \sigma_A^2 + \sigma_P^2 \end{aligned}$$

which follows from the assumption that the residual errors for the same individual are uncorrelated. Assuming purely additive gene action, $\sigma_A^2 = h^2\sigma_y^2$ of the covariance is due to genetic effects, leaving $(r - h^2)\sigma_y^2 = \sigma_P^2$ as the covariance due to permanent environmental effects.

One approach to estimating breeding values when repeated measures are contained in the data set is to continue to apply the simple animal model (Equation 26.8), with suitable $\mathbf{Z}$ to accommodate multiple records, modifying the residual covariance matrix $\mathbf{R}$ such that

$$R_{jk} = \begin{cases} (1 - h^2)\sigma_y^2 & j = k \text{ (the same measurement in an individual)} \\ (r - h^2)\sigma_y^2 & j \text{ and } k \text{ are repeated measures} \\ 0 & j \text{ and } k \text{ are measures on different individuals.} \end{cases} \tag{26.25}$$

Since the resulting covariance matrix is not diagonal, it is not always easily inverted, a potentially serious complication for extremely large data sets.

An alternative approach follows the same rationale as the model incorporating dominance, i.e., explicitly accounting for shared environmental effects (rather

than incorporating them into the residual error structure) by introducing a new random factor into the model, such that

$$y_{ij} = \mu + a_i + p_i + e_{ij} \tag{26.26}$$

(Henderson 1977a). With this structure, all residual effects are again uncorrelated with common variance $\sigma_E^2 = (1 - r)\sigma_y^2$.

Suppose a total of k measurements are made on n individuals (such a balanced design is not essential). We can write this in the general mixed-model framework as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathbf{a} + \mathbf{p}) + \mathbf{e} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{p} + \mathbf{e} \tag{26.27}$$

where $\mathbf{a}^T = (a_1, \cdots, a_n)^T$ and $\mathbf{p}^T = (p_1, \cdots, p_n)^T$, and each row of the $nk \times n$ incidence matrix $\mathbf{Z}$ has all zeros except a one at the position corresponding to the individual measured. This model has the resulting covariance matrices

$$\sigma(\mathbf{a}, \mathbf{a}) = h^2\sigma_y^2\,\mathbf{A}, \qquad \sigma(\mathbf{p}, \mathbf{p}) = (r - h^2)\,\sigma_y^2\,\mathbf{I}, \qquad \sigma(\mathbf{e}, \mathbf{e}) = (1 - r)\,\sigma_y^2\,\mathbf{I}$$

with $\mathbf{a}$, $\mathbf{p}$, and $\mathbf{e}$ being assumed to be uncorrelated. Applying Equation 26.19b and removing the common factor σ_y^2 from all expressions, the mixed-model equations become

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda_A\,\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} & \mathbf{Z}^T\mathbf{Z} + \lambda_P\,\mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix} \tag{26.28a}$$

where

$$\lambda_A = \frac{\sigma_E^2}{\sigma_A^2} = \frac{1 - r}{h^2}, \qquad \lambda_P = \frac{\sigma_E^2}{\sigma_P^2} = \frac{1 - r}{r - h^2} \tag{26.28b}$$

Example 10. To compare the two different methods for dealing with repeated records, suppose three unrelated and noninbred individuals are measured, with two observations on individual one ($y_1 = 7$, $y_2 = 8$), three observations on individual two ($y_3 = 6$, $y_4 = 6$, $y_5 = 5$), and one observation on individual three ($y_6 = 9$). Assume that the only fixed factor is the mean and that the character has heritability $h^2 = 0.4$ and repeatability $r = 0.5$, giving $1 - h^2 = 0.6$ and $r - h^2 = 0.1$. For either formulation, we have

$$\mathbf{y} = \begin{pmatrix} 7 \\ 8 \\ 6 \\ 6 \\ 5 \\ 9 \end{pmatrix}, \quad \boldsymbol{\beta} = (\,\mu\,), \quad \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Since all three individuals are assumed to be unrelated, $\mathbf{A} = \mathbf{I}$.

To apply the permanent-effects model $y_i = \mu + a_i + p_i + e_i$, note that

$$\lambda_A = \frac{1-r}{h^2} = \frac{1-0.5}{0.4} = 1.25, \qquad \lambda_P = \frac{1-r}{r-h^2} = \frac{1-0.5}{0.5-0.4} = 5,$$

$$\mathbf{X}^T\mathbf{X} = 6, \qquad \mathbf{X}^T\mathbf{y} = 41,$$

$$\mathbf{Z}^T\mathbf{X} = (\mathbf{X}^T\mathbf{Z})^T = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}, \qquad \mathbf{Z}^T\mathbf{y} = \begin{pmatrix} 15 \\ 17 \\ 9 \end{pmatrix}, \qquad \mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{Z}^T\mathbf{Z} + \lambda_A\,\mathbf{A}^{-1} = \begin{pmatrix} 3.25 & 0 & 0 \\ 0 & 4.25 & 0 \\ 0 & 0 & 2.25 \end{pmatrix}, \qquad \mathbf{Z}^T\mathbf{Z} + \lambda_P\,\mathbf{I} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 6 \end{pmatrix}$$

giving the MMEs (Equation 26.28a) as

$$\begin{pmatrix} 6 & 2 & 3 & 1 & 2 & 3 & 1 \\ 2 & 3.25 & 0 & 0 & 2 & 0 & 0 \\ 3 & 0 & 4.25 & 0 & 0 & 3 & 0 \\ 1 & 0 & 0 & 2.25 & 0 & 0 & 1 \\ 2 & 2 & 0 & 0 & 7 & 0 & 0 \\ 3 & 0 & 3 & 0 & 0 & 8 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 6 \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{p}_1 \\ \widehat{p}_2 \\ \widehat{p}_3 \end{pmatrix} = \begin{pmatrix} 41 \\ 15 \\ 17 \\ 9 \\ 15 \\ 17 \\ 9 \end{pmatrix}$$

which has solutions

$$\widehat{\mu} \simeq 7.174, \qquad \begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \end{pmatrix} \simeq \begin{pmatrix} 0.174 \\ -0.904 \\ 0.730 \end{pmatrix}, \qquad \begin{pmatrix} \widehat{p}_1 \\ \widehat{p}_2 \\ \widehat{p}_3 \end{pmatrix} \simeq \begin{pmatrix} 0.043 \\ -0.226 \\ 0.183 \end{pmatrix}$$

Conversely, applying the simple animal model $y_i = \mu + a_i + e_i$, from Equation 26.25 the covariance matrix for the residual errors becomes

$$\mathbf{R} = \sigma_y^2 \begin{pmatrix} 0.5 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0.1 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0.5 & 0.1 & 0 \\ 0 & 0 & 0.1 & 0.1 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}$$

Likewise, $\mathbf{G} = h^2\sigma_y^2\,\mathbf{I}$, and hence $\mathbf{G}^{-1} = (h^2\sigma_y^2)^{-1}\,\mathbf{I}$. Removing the factor σ_y^2 common to all expressions gives

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} \simeq 8.27, \qquad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} = (\mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z})^T \simeq \begin{pmatrix} 2.86 \\ 3.75 \\ 1.67 \end{pmatrix}$$

$$\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} = \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \frac{1}{h^2}\mathbf{I} \simeq \begin{pmatrix} 5.36 & 0 & 0 \\ 0 & 6.25 & 0 \\ 0 & 0 & 4.17 \end{pmatrix}$$

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = 57.68, \qquad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \simeq \begin{pmatrix} 21.43 \\ 21.25 \\ 15 \end{pmatrix}$$

Substituting into Equation 26.5 gives the MMEs

$$\begin{pmatrix} 8.27 & 2.86 & 3.75 & 1.67 \\ 2.86 & 5.36 & 0 & 0 \\ 3.75 & 0 & 6.25 & 0 \\ 1.67 & 0 & 0 & 4.17 \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \end{pmatrix} = \begin{pmatrix} 57.68 \\ 21.43 \\ 21.25 \\ 15 \end{pmatrix}$$

which gives the same estimates as obtained with the permanent-effects model.