

Linear Mixed Models: Part II

2.1 Tests in Linear Mixed Models

The previous section dealt with point estimation and related problems in linear mixed models. In this section, we consider a different type of inference, namely, tests in linear mixed models. Section 2.1.1 discusses statistical tests in Gaussian mixed models. As shown, exact F -tests can often be derived under Gaussian ANOVA models. Furthermore, in some special cases, optimal tests such as uniformly most powerful unbiased (UMPU) tests exist and coincide with the exact F -tests. Section 2.1.2 considers tests in non-Gaussian linear mixed models. In such cases, exact/optimal tests typically do not exist. Therefore, statistical tests are usually developed based on asymptotic theory.

2.1.1 Tests in Gaussian Mixed Models

1. *Exact tests.* For ANOVA models, exact F -tests can often be derived using the following method. The original idea was due to Wald (1947). Consider the mixed ANOVA model (1.1) and (1.2). Suppose that one wishes to test the hypothesis $H_0: \sigma_1^2 = 0$. Note that the model can be written as

$$y = X\beta + Z_1\alpha_1 + Z_{-1}\alpha_{-1} + \epsilon, \quad (2.1)$$

where $\alpha_{-1} = (\alpha'_2, \dots, \alpha'_s)'$ and $Z_{-1} = (Z_2, \dots, Z_s)$. Consider the quadratic form $q_1 = \tau^{-2}y'P_{Z_1\ominus(X, Z_{-1})}y = y'\{P_{Z_1\ominus(X, Z_{-1})}/\tau^2\}y$, where \ominus is introduced in Example 1.9. Note that, under the null hypothesis, we have $y \sim N(X\beta, V_0)$, where $V_0 = \tau^2I + \sum_{i=2}^s \sigma_i^2 Z_i Z_i'$. Furthermore, we have

$$\begin{aligned} \left(\frac{P_{Z_1\ominus(X, Z_{-1})}}{\tau^2} \right) V_0 &= P_{Z_1\ominus(X, Z_{-1})} + \sum_{i=2}^s \left(\frac{\sigma_i^2}{\tau^2} \right) P_{Z_1\ominus(X, Z_{-1})} Z_i Z_i' \\ &= P_{Z_1\ominus(X, Z_{-1})}, \end{aligned}$$

which is idempotent. Therefore, by Theorem C.1 in Appendix C, we have $q_1 \sim \chi_{r_1}^2$, where $r_1 = \text{rank}\{P_{Z_1\ominus(X, Z_{-1})}\} = \text{rank}\{(X, Z)\} - \text{rank}\{(X, Z_{-1})\}$.

Note that $P_{Z_1 \ominus (X, Z_{-1})}X = 0$ and $P_{(X, Z)} = P_{(X, Z_{-1})} + P_{Z_1 \ominus (X, Z_{-1})}$, where the two projections on the right side are orthogonal to each other (see Example 1.9 and Exercise 1.17).

On the other hand, consider the quadratic form $q_2 = \tau^{-2}y'P_{(X, Z)^\perp}y = y'\{P_{(X, Z)^\perp}/\tau^2\}y$. Note that $y \sim N(X\beta, V)$, where $V = \tau^2I + \sum_{i=1}^s \sigma_i^2 Z_i Z_i'$. Thus, we have

$$\begin{aligned} \left(\frac{P_{(X, Z)^\perp}}{\tau^2} \right) V &= P_{(X, Z)^\perp} + \sum_{i=1}^2 \left(\frac{\sigma_i^2}{\tau^2} \right) P_{(X, Z)^\perp} Z_i Z_i' \\ &= P_{(X, Z)^\perp}, \end{aligned} \quad (2.2)$$

which is idempotent. Therefore, by the same theorem, we have $q_2 \sim \chi_{r_2}^2$, where $r_2 = \text{rank}\{P_{(X, Z)^\perp}\} = n - \text{rank}\{(X, Z)\}$. Note that $P_{(X, Z)^\perp}X = 0$. Also note that, unlike q_1 , the distribution of q_2 is unaffected by the null hypothesis.

Finally, because $P_{(X, Z)^\perp}VP_{Z_1 \ominus (X, Z_{-1})} = \tau^2 P_{(X, Z)^\perp}P_{Z_1 \ominus (X, Z_{-1})} = 0$ by (2.2), the two quadratic forms q_1 and q_2 are independent (again, this fact does not depend on the null hypothesis; see Appendix C). It follows that

$$\begin{aligned} F_1 &= \frac{y'P_{Z_1 \ominus (X, Z_{-1})}y/r_1}{y'P_{(X, Z)^\perp}y/r_2} \\ &= \frac{q_1/r_1}{q_2/r_2} \sim F_{r_1, r_2}. \end{aligned} \quad (2.3)$$

In words, F_1 has an exact (central) F -distribution with degrees of freedom r_1 and r_2 for testing the hypothesis $H_0: \sigma_1^2 = 0$.

It should be pointed out that, for the above test to be effective one must have $Z_1 \ominus (X, Z_{-1}) \neq \emptyset$. For example, if $\mathcal{L}(Z_1) \subset \mathcal{L}(Z_{-1})$, then the test will not work. We now consider an example.

Example 2.1 (Balanced two-way random effects model). First consider the case where there is no interaction between the random effect factors. The model can be expressed as

$$y_{ijk} = \mu + u_i + v_j + e_{ijk},$$

$i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, c$, where u_i s and v_j s are random effects and e_{ijk} s are errors such that u_i s are independent $N(0, \sigma_1^2)$, v_j s are independent $N(0, \sigma_2^2)$, e_{ijk} s are independent $N(0, \tau^2)$, and u , v , e are independent. Using matrix expressions, we have

$$y = X\mu + Z_1u + Z_2v + e,$$

where $X = 1_a \otimes 1_b \otimes 1_c$, $Z_1 = I_a \otimes 1_b \otimes 1_c$, and $Z_2 = 1_a \otimes I_b \otimes 1_c$. Clearly, $Z_1 \ominus (X, Z_2) \neq \emptyset$, thus (2.3) may be applied for testing $H_0: \sigma_1^2 = 0$. In this case, we have $r_1 = (a+b-1)-b = a-1$ and $r_2 = n - (a+b-1) = abc - a - b + 1$.

Next, we consider the case where there is interaction between u and v . In this case, the model can be expressed as

$$y_{ijk} = \mu + u_i + v_j + w_{ij} + e_{ijk},$$

$i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, c$, where, in addition, the interactions w_{ij} s are independent $N(0, \sigma_3^2)$, and u , v , w , e are independent. Similarly, the model may be written as

$$y = X\mu + Z_1u + Z_2v + Z_3w + e,$$

where $Z_3 = I_a \otimes I_b \otimes 1_c$. However, neither $\sigma_1^2 = 0$ nor $\sigma_2^2 = 0$ can be tested using the exact F -test derived above, because $\mathcal{L}(Z_j) \subset \mathcal{L}(Z_3)$, $j = 1, 2$. Nevertheless, the hypothesis $H_0: \sigma_3^2 = 0$ can be tested using (2.3). In this case, $r_1 = ab - (a + b - 1) = (a - 1)(b - 1)$ and $r_2 = n - ab = ab(c - 1)$ (Exercise 2.1).

Further results on exact tests in Gaussian mixed models can be found in Khuri et al. (1998).

2. Optimal tests. It is known that optimal tests, such as UMPU and uniformly most powerful invariant unbiased tests (UMPIU), exist in some special cases of the mixed ANOVA models, assuming that normality holds. For example, Mathew and Sinha (1988) considered a balanced mixed ANOVA model, which can be expressed as

$$y = X_1\beta_1 + \dots + X_t\beta_t + Z_1\alpha_1 + \dots + Z_s\alpha_s + \epsilon, \quad (2.4)$$

where the β s and α s are, respectively, vectors of fixed and random effects in the analysis of variance; that is, they are main effects, interactions, nested effects, and the like. (e.g., Scheffé 1959), and ϵ is a vector of errors. Furthermore, assume that the random effects and errors are independent such that the components of α_i are distributed as $N(0, \sigma_i^2)$, and the components of ϵ are distributed as $N(0, \tau^2)$. The design matrices X_1, \dots, X_t and Z_1, \dots, Z_s are assumed known with $X_1 = 1_n$. Let P_i , $i = 1, \dots, t$ and Q_i , $i = 1, \dots, s$ be projection matrices such that $P_1 = n^{-1}J_n$, where $J_n = 1_n 1_n'$, $y'P_i y$ the sum of squares due to β_i , $2 \leq i \leq t$, and $y'Q_i y$ the sum of squares due to α_i (as in a fixed effects model), $1 \leq i \leq s$ (Searle 1971, §9.6). Note that each P_i (Q_i) is a Kronecker product of matrices of the form I_a , $a^{-1}J_a$ or $I_a - a^{-1}J_a$, so that P_i , $i = 1, \dots, t$ and Q_i , $i = 1, \dots, s + 1$ are orthogonal to each other, where $Q_{s+1} = I_n - \sum_{i=1}^t P_i - \sum_{i=1}^s Q_i$. With these notations, the likelihood function can be expressed as

$$f(y) = c(\theta) \times \exp \left[-\frac{1}{2} \left\{ \sum_{i=1}^{s+1} \xi_i y' Q_i y + \sum_{i=1}^t \eta_i (S_i' y - \lambda_i)' (S_i' y - \lambda_i) \right\} \right], \quad (2.5)$$

where $c(\theta)$ depends only on the variance components, $\theta = (\sigma_1^2, \dots, \sigma_s^2, \tau^2)'$; ξ_i and η_i are linear functions of the variance components; $S_i S_i' = P_i$ and

$\lambda_i = S'_i X\beta$, $1 \leq i \leq t$. Here $X\beta$ is as in (1.1) when (2.4) is written in this way. By (2.5), it can be shown that $S'_i y$, $i = 1, \dots, t$ and $y'Q_i y$, $i = 1, \dots, s+1$ are complete sufficient statistics for the parameters ξ_i s, η_i s and λ_i s. Furthermore, standard theory for the multiparameter exponential family (e.g., Lehmann and Casella 1998, §1) may be applied to derive UMPU and other optimal tests. For example, Mathew and Sinha (1988) obtained the following results.

1. Suppose that the hypothesis of interest is $H_0: \lambda_i = 0$ versus $H_1: \lambda_i \neq 0$. If η_i equals some ξ_j , say, ξ_1 , an exact F -test is based on $y'P_i y/y'Q_1 y$; if λ_i is a scalar, then this test is UMPU; if λ_i is multidimensional, a UMPU test does not exist, however, the above F -test is UMPIU.

2. Suppose that the hypothesis of interest is $H_0: \xi_1 = \xi_2$ versus $H_1: \xi_2 > \xi_1$. The F -test based on $y'Q_2 y/y'Q_1 y$ is UMPU and UMPIU.

Note that, in some cases, a hypothesis such as $\sigma_i^2 = 0$ is equivalent to the equality of two ξ_i s. We consider some examples.

Example 2.2 (Balanced one-way random effects model). Consider a special case of the one-way random effects model of Example 1.1 with $k_i = k$, $1 \leq i \leq m$. In this case, $y'Q_1 y$ is equal to the treatment sum of squares and $y'Q_2 y$ error sum of squares, that is, $y'Q_1 y = \text{SSA} = k \sum_{i=1}^m (\bar{y}_{i.} - \bar{y}_{..})^2$, $y'Q_2 y = \text{SSE} = \sum_{i=1}^m \sum_{j=1}^k (y_{ij} - \bar{y}_{i.})^2$, and $S'_1 y = \sqrt{mk}\bar{y}_{..}$. Furthermore, we have $\xi_1^{-1} = \tau^2 + k\sigma^2$, $\xi_2^{-1} = \tau^2$, $\eta_1^{-1} = \tau^2 + k\sigma^2$, and $\lambda_1 = \sqrt{mk}\mu$.

Consider the hypothesis $\mu = 0$. Because $\eta_1 = \xi_1$ and λ_1 is a scalar, by the first result above, the F -test based on $\bar{y}_{..}/\text{SSA}$ is UMPU and UMPIU. As for the hypothesis $\sigma^2 = 0$, because it is equivalent to $\xi_1 = \xi_2$, the F -test based on SSA/SSE is UMPU and UMPIU.

Example 2.1 (Continued). Consider the case without interaction and that $k = 1$. In this case, the model can simply be expressed as

$$y_{ij} = \mu + u_i + v_j + e_{ij},$$

$i = 1, \dots, a$, $j = 1, \dots, b$. In this case, we have $y'Q_1 y = b \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2 \equiv \text{SSA}$, $y'Q_2 y = a \sum_{j=1}^b (\bar{y}_{.j} - \bar{y}_{..})^2 \equiv \text{SSB}$, and $y'Q_3 y = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \equiv \text{SSE}$, which correspond to $\xi_1^{-1} = \tau^2 + b\sigma_1^2$, $\xi_2^{-1} = \tau^2 + a\sigma_2^2$, and $\xi_3^{-1} = \tau^2$, respectively. Furthermore, we have $S'_1 y = \sqrt{ab}\bar{y}_{..}$ with $\eta_1^{-1} = \tau^2 + a\sigma_2^2 + b\sigma_1^2$ and $\lambda_1 = \sqrt{ab}\mu$.

The hypotheses $\sigma_1^2 = 0$ and $\sigma_2^2 = 0$ correspond to $\xi_1 = \xi_3$ and $\xi_2 = \xi_3$, respectively. Thus, the F -tests based on SSA/SSE and SSB/SSE are, respectively, optimal (i.e., UMPU and UMPIU) for testing these hypotheses. However, unlike the previous example, no exact optimal test (in the same sense) exists for testing $\mu = 0$, because η_1 is not equal to any of the ξ s.

These examples show that the results of Mathew and Sinha (1988) may be useful in some cases to obtain optimal tests, but there are cases where these

results do not yield optimal tests (see Exercise 2.2 for an additional example). For more discussion on optimal tests, see Khuri et al. (1998).

3. Likelihood-ratio tests. The likelihood-ratio is a well-known method of constructing statistical tests. The theory of likelihood-ratio tests is fully developed in the i.i.d. case (e.g., Lehmann 1999, §7.7). However, the literature on likelihood-ratio tests in the context of linear mixed models is much less extensive, from a theoretical point of view. Hartley and Rao (1967) was the first paper that addressed the issue. Let $\psi = (\beta', \theta')'$ be the vector of all the unknown parameters involved in a Gaussian mixed model, where θ represents the vector of variance components. Many of the hypotheses are concerned with testing whether a subvector of θ , say, $\theta^{(1)}$, is identical to a known vector, $\theta_0^{(1)}$. Let $\theta^{(2)}$ denote the subvector of θ complementary to $\theta^{(1)}$. Then, the likelihood function may be expressed as $L(\theta) = L(\theta^{(1)}, \theta^{(2)})$. [Note that $L(\theta)$ depends on y and therefore should be properly denoted by $L(\theta|y)$, but we suppress y for notational simplicity.] Let $\hat{\theta}$ be the (global) maximizer of $L(\theta|y)$ over $\theta \in \Theta$, where Θ is the parameter space, and $\hat{\theta}^{(2)}$ be the (global) maximizer of $L(\theta_0^{(1)}, \theta^{(2)})$ over $\theta^{(2)} \in \Theta^{(2)}$, where $\Theta^{(2)}$ is the parameter space for $\theta^{(2)}$. Then, the likelihood ratio is given by

$$\mathcal{R} = \frac{L(\theta_0^{(1)}, \hat{\theta}^{(2)})}{L(\hat{\theta})}. \quad (2.6)$$

Hartley and Rao (1967) stated without giving a proof that the asymptotic null distribution of $-2 \log \mathcal{R}$ is a central χ^2 with r degrees of freedom, where r is the dimension of $\theta^{(1)}$. See Jiang (2005c) for a rigorous proof of this result, which also applies to non-Gaussian linear mixed models (see Section 2.1.2.4). We consider a simple example.

Example 2.3 (One-way random effects model). Consider the one-way random effects model of Example 1.1 with normality assumption. It was shown in Section 1.3.1 (see Example 1.1 (Continued)) that the log-likelihood function is given by

$$\begin{aligned} l(\mu, \sigma^2, \tau^2) = & c - \frac{1}{2}(n - m) \log(\tau^2) - \frac{1}{2} \sum_{i=1}^m \log(\tau^2 + k_i \sigma^2) \\ & - \frac{1}{2\tau^2} \sum_{i=1}^m \sum_{j=1}^{k_i} (y_{ij} - \mu)^2 + \frac{\sigma^2}{2\tau^2} \sum_{i=1}^m \frac{k_i^2}{\tau^2 + k_i \sigma^2} (\bar{y}_{i\cdot} - \mu)^2, \end{aligned}$$

where c is a constant, $n = \sum_{i=1}^m k_i$, and $\bar{y}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} y_{ij}$. Let $\hat{\mu}$, $\hat{\sigma}^2$ and $\hat{\tau}^2$ be the MLE of μ , σ^2 , and τ^2 . Suppose that one is interested in testing the hypothesis $\sigma^2 = 0$. Under the null hypothesis, we have

$$l(\mu, 0, \tau^2) = c - \frac{n}{2} \log(\tau^2) - \frac{1}{2\tau^2} \sum_{i=1}^m \sum_{j=1}^{k_i} (y_{ij} - \mu)^2.$$

The MLE under the null are $\tilde{\mu} = \bar{y}_{..}$ and $\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^m \sum_{j=1}^{k_i} (y_{ij} - \bar{y}_{..})^2$, where $\bar{y}_{..} = n^{-1} \sum_{i=1}^m \sum_{j=1}^{k_i} y_{ij}$. Thus, an expression for $-2 \log \mathcal{R}$ can be easily derived (Exercise 2.3).

2.1.2 Tests in Non-Gaussian Linear Mixed Models

For non-Gaussian linear mixed models, exact or optimal tests typically do not exist. This is because under a non-Gaussian model, the distribution of y is not fully specified, therefore it is (usually) not possible either to derive the exact distribution of a test statistic or to study the power function of the test. In such cases, statistical tests are usually based on asymptotic theory. In this section, we consider asymptotic tests in non-Gaussian linear mixed models. Please note that the results of this section also apply to Gaussian mixed models, especially in cases where exact/optimal tests do not exist.

A basic idea of deriving an asymptotic test is the following. Consider a non-Gaussian linear mixed model (1.1). Let $\psi = (\beta', \theta')'$, where θ represents the vector of variance components involved. Then ψ is the vector of all the unknown parameters involved in the model. Suppose that an estimator of ψ , say, $\hat{\psi}$, can be obtained, which is asymptotically normal, that is, there exists a sequence of positive definite matrices, $\Sigma = \Sigma_n$, such that

$$\Sigma^{-1/2}(\hat{\psi} - \psi) \longrightarrow N(0, I), \quad \text{in distribution,} \quad (2.7)$$

where I is the $(p + q)$ -dimensional identity matrix with $p = \dim(\beta)$ and $q = \dim(\theta)$. Σ is called the asymptotic covariance matrix of $\hat{\psi}$. Suppose that one wishes to test a linear hypothesis of the form

$$H_0 : K' \psi = c, \quad (2.8)$$

where K is a known matrix of full (column) rank, say, r , and c is a known vector. Under (2.8), (2.7) implies that

$$(K' \hat{\psi} - c)' (K' \Sigma K)^{-1} (K' \hat{\psi} - c) \longrightarrow \chi_r^2, \quad \text{in distribution.} \quad (2.9)$$

Thus, (2.9) can be used to test the hypothesis (2.8).

Typically, the asymptotic covariance matrix depends not only on θ but also on some additional parameters. For example, under the mixed ANOVA model Section 1.2.2.1, the asymptotic covariance matrix of the REML estimator of $\theta = (\tau^2, \sigma_1^2, \dots, \sigma_s^2)'$ depends not only on θ but also on the kurtoses of the random effects and errors; the asymptotic covariance matrix of the ML estimator of ψ depends not only on θ but also on the kurtoses as well as the third moments of the random effects and errors. (See Section 2.2.2 for more details; note that, under normality both the third moments and the kurtoses vanish, so there is no such problem for Gaussian mixed models.) Therefore, for the asymptotic test (2.9) to be applicable, one has to find some way to consistently estimate Σ , because standard procedures in mixed model analysis

such as ML and REML do not produce estimators of higher (i.e., third and fourth) moments of the random effects and errors. In the following we discuss several methods of estimating Σ . Typically, when Σ in (2.9) is replaced by a consistent estimator, say $\hat{\Sigma}$, the asymptotic distribution on the right side does not change. The test therefore rejects if

$$(K'\hat{\psi} - c)'(K'\hat{\Sigma}K)^{-1}(K'\hat{\psi} - c) > \chi_{r,\rho}^2, \quad (2.10)$$

where ρ is the significance level.

1. *Empirical method of moments.* Consider the case of the mixed ANOVA model (1.1) and (1.2). As mentioned, the asymptotic covariance matrix of the REML (ML) estimator involves higher moments, thus, a natural approach would be to find consistent estimators of those higher moments. Jiang (2003) proposed an empirical method of moments and gave a number of applications, including estimation of the kurtoses in mixed ANOVA models. The basic idea is the following. Let θ be a vector of parameters. Suppose that a consistent estimator of θ , $\hat{\theta}$, is available. Let ϕ be a vector of additional parameters about which knowledge is needed. Let $\vartheta = (\theta' \phi')'$, and $M(\vartheta, y) = M(\theta, \phi, y)$ be a vector-valued function of the same dimension as ϕ that depends on ϑ and y , a vector of observations. Suppose that $E\{M(\vartheta, y)\} = 0$ when ϑ is the true vector of parameters. Then, if θ were known, a method of moments estimator of ϕ would be obtained by solving

$$M(\theta, \phi, y) = 0 \quad (2.11)$$

for ϕ . Note that this is more general than the classical method of moments, in which the function M is a vector of sample moments minus their expected values. In econometric literature, this is referred to as the generalized method of moments (e.g., Hansen 1982, Newey 1985). Because θ is unknown, we replace it in (2.11) by $\hat{\theta}$. The result is called an empirical method of moments (EMM) estimator of ϕ , denoted by $\hat{\phi}$, which is obtained by solving

$$M(\hat{\theta}, \phi, y) = 0. \quad (2.12)$$

Note that here we use the words “an EMM estimator” instead of “the EMM estimator”, because sometimes there may be more than one consistent estimator of θ , and each may result in a different EMM estimator of ϕ . In general, ML estimators may be viewed as a special kind of EMM estimator (Exercises 2.4 and 2.5). To see this, let $l(\vartheta; y) = l(\theta, \phi; y)$ be the log-likelihood function. Then, the ML estimator, $\hat{\vartheta} = (\hat{\theta}' \hat{\phi}')$ satisfies $\partial l / \partial \vartheta = 0$, and hence $\hat{\phi}$, the ML estimator of ϕ , satisfies

$$\frac{\partial}{\partial \vartheta} l(\hat{\theta}, \phi; y) = 0. \quad (2.13)$$

On the other hand, (2.13) is a special case of (2.12), in which $M(\theta, \phi, y) = \partial l / \partial \vartheta$. Note that $E(\partial l / \partial \vartheta) = 0$ when ϑ is the true vector of parameters. Jiang (2003) showed that, under mild conditions, $\hat{\phi}$ is consistent.

To apply EMM to non-Gaussian mixed ANOVA models, let θ be the vector of variance components. It is clear that a consistent estimator of θ , $\hat{\theta}$, exists. For example, $\hat{\theta}$ can be the REML or ML estimator (e.g., Jiang 1996). Furthermore, assume that the third moments of the random effects and errors vanish; that is,

$$E(\epsilon_1^3) = 0 \quad \text{and} \quad E(\alpha_{i1}^3) = 0, \quad 1 \leq i \leq s, \quad (2.14)$$

where α_{i1} is the first component of α_i and ϵ_1 the first component of ϵ . Then, the asymptotic covariance matrix of the REML (ML) estimator involves only the kurtoses, in addition to the variance components [in fact, the asymptotic covariance matrix of REML estimator does not involve the third moments regardless of (2.14)]. For notational convenience, write $\sigma_0^2 = \tau^2$. Then, the (unscaled) kurtoses are defined by $\kappa_0 = E(\epsilon_1^4) - 3\sigma_0^4$, $\kappa_i = E(\alpha_{i1}^4) - 3\sigma_i^4$, $1 \leq i \leq s$. For any matrix $A = (a_{ij})$, we define $\|A\|_4 = (\sum_{i,j} a_{ij}^4)^{1/4}$. Similarly, if $a = (a_i)$ is a vector, then $\|a\|_4 = (\sum_i a_i^4)^{1/4}$. Let L be a linear space, then L^\perp represents the linear space $\{a : a'b = 0, \forall b \in L\}$. If L_1, L_2 are linear spaces such that $L_1 \subset L_2$, then $L_2 \ominus L_1$ represents the linear space $\{a : a \in L_2, a'b = 0, \forall b \in L_1\}$ (note that the notation is consistent with that in Example 1.9). If M_1, \dots, M_k are matrices with the same number of rows, then $\mathcal{L}(M_1, \dots, M_k)$ represents the linear space spanned by the columns of M_1, \dots, M_k . Let the matrices Z_1, \dots, Z_s be suitably ordered such that

$$L_i \neq \{0\}, \quad 0 \leq i \leq s, \quad (2.15)$$

where $L_0 = \mathcal{L}(Z_1, \dots, Z_s)^\perp$, $L_i = \mathcal{L}(Z_i, \dots, Z_s) \ominus \mathcal{L}(Z_{i+1}, \dots, Z_s)$, $1 \leq i \leq s-1$, and $L_s = \mathcal{L}(Z_s)$. Let C_i be a matrix whose columns constitute a base of L_i , $0 \leq i \leq s$. We define $a_{ij} = \|Z_j' C_i\|_4^4$, $0 \leq j \leq i \leq s$, where $Z_0 = I$, the identity matrix. It is easy to see that, under (2.15), $a_{ii} > 0$, $0 \leq i \leq s$. Let n_i be the number of columns of C_i , and c_{ik} the k th column of C_i , $1 \leq k \leq n_i$, $0 \leq i \leq s$. Define

$$b_i(\sigma^2) = 3 \sum_{k=1}^{n_i} \left(\sum_{j=0}^i |Z_j' c_{ik}|^2 \sigma_j^2 \right)^2, \quad 0 \leq i \leq s.$$

where $\sigma^2 = (\sigma_j^2)_{0 \leq j \leq s}$. Let $\kappa = (\kappa_j)_{0 \leq j \leq s}$, and $M(\beta, \sigma^2, \kappa, y)$ be the vector whose i th component is

$$M_i(\beta, \sigma^2, \kappa, y) = \|C_i'(y - X\beta)\|_4^4 - \sum_{j=0}^i a_{ij} \kappa_j - b_i(\sigma^2), \quad 0 \leq i \leq s.$$

Then, by the following lemma and the definition of the C_i s, it can be shown that $E\{M(\beta, \sigma^2, \kappa, y)\} = 0$ when β, σ^2, κ correspond to the true parameters (Exercise 2.6). Thus, a set of EMM estimators can be easily obtained by

solving $M(\hat{\beta}, \hat{\sigma}^2, \kappa, y) = 0$, where $\hat{\beta}$ and $\hat{\sigma}^2$ are the REML or ML estimators. Furthermore, the EMM estimators can be computed recursively as follows.

$$\begin{aligned}\hat{\kappa}_0 &= a_{00}^{-1} \hat{d}_0, \\ \hat{\kappa}_i &= a_{ii}^{-1} \hat{d}_i - \sum_{j=0}^{i-1} \left(\frac{a_{ij}}{a_{ii}} \right) \hat{\kappa}_j, \quad 1 \leq i \leq s,\end{aligned}\tag{2.16}$$

where $\hat{d}_i = \|C'_i(y - X\hat{\beta})\|_4^4 - b_i(\hat{\sigma}^2)$, $0 \leq i \leq s$.

Lemma 2.1. Let ξ_1, \dots, ξ_n be independent random variables such that $E\xi_i = 0$ and $E\xi_i^4 < \infty$, and $\lambda_1, \dots, \lambda_n$ be constants. Then,

$$E \left(\sum_{i=1}^n \lambda_i \xi_i \right)^4 = 3 \left[\sum_{i=1}^n \lambda_i^2 \text{var}(\xi_i) \right]^2 + \sum_{i=1}^n \lambda_i^4 \{E\xi_i^4 - 3[\text{var}(\xi_i)]^2\}.$$

Example 2.2 (Continued). Here we have $\kappa_0 = E(\epsilon_{11}^4) - 3\tau^4$ and $\kappa_1 = E(\alpha_1^4) - 3\sigma^4$. The model can be written as $y = X\mu + Z\alpha + \epsilon$, where $X = 1_m \otimes 1_k$, and $Z = I_m \otimes 1_k$. Let

$$D_k = \begin{pmatrix} 1 & \cdots & 1 \\ -1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & -1 \end{pmatrix}_{k \times (k-1)}.$$

Then, it is easy to show that $C_0 = I_m \otimes D_k$, $C_1 = Z = I_m \otimes 1_k$. It follows from (2.16) that, in closed form,

$$\begin{aligned}\hat{\kappa}_0 &= \frac{1}{2m(k-1)} \sum_{i=1}^m \sum_{j=2}^k (y_{i1} - y_{ij})^4 - 6\hat{\tau}^4, \\ \hat{\kappa}_1 &= \frac{1}{mk^4} \sum_{i=1}^m (y_{i\cdot} - k\hat{\mu})^4 - \frac{1}{2mk^3(k-1)} \sum_{i=1}^m \sum_{j=2}^k (y_{i1} - y_{ij})^4 \\ &\quad - \frac{3}{k^2} \left(1 - \frac{2}{k}\right) \hat{\tau}^4 - \frac{6}{k} \hat{\tau}^2 \hat{\sigma}^2 - 3\hat{\sigma}^4,\end{aligned}$$

where $y_{i\cdot} = \sum_{j=1}^k y_{ij}$, $\hat{\mu} = \bar{y}_{\cdot\cdot}$, and $\hat{\tau}^2, \hat{\sigma}^2$ are the REML or ML estimators. It can be shown (Exercise 2.7) that the EMM estimators are consistent provided that $m \rightarrow \infty$ and $k \geq 2$.

2. Partially observed information. One important assumption that we have made in the application of EMM is (2.14). This assumption holds, for example, if the random effects and errors are symmetrically distributed. However, from a practical point of view, such an assumption is not very pleasant because, like normality, symmetry may not hold in practice. On the other hand,

a method called partially observed information was proposed in Section 1.4.2 for estimating the asymptotic covariance matrices of REML or ML estimators. This method applies to a general non-Gaussian mixed ANOVA model regardless of (2.14). We consider an example.

Example 2.2 (Continued). Suppose that one wishes to test the hypothesis $H_0: \gamma_1 = 1$; that is, the variance contribution due to the random effects is the same as that due to the errors. Note that in this case $\theta = (\lambda, \gamma_1)'$, so the null hypothesis corresponds to (2.8) with $K = (0, 1)'$ and $c = 1$. Furthermore, we have $K' \Sigma_R K = \Sigma_{R,11}$, which is the asymptotic variance of $\hat{\gamma}_1$, the REML estimator of γ_1 . Thus, the test statistic is $\hat{\chi}^2 = (\hat{\gamma}_1 - 1)^2 / \hat{\Sigma}_{R,11}$, where $\hat{\Sigma}_{R,11}$ is the POQUIM estimator of $\Sigma_{R,11}$ (see Section 1.8.5) given by

$$\hat{\Sigma}_{R,11} = \frac{\hat{\mathcal{I}}_{1,11} \hat{\mathcal{I}}_{2,00}^2 - 2 \hat{\mathcal{I}}_{1,01} \hat{\mathcal{I}}_{2,00} \hat{\mathcal{I}}_{2,01} + \hat{\mathcal{I}}_{1,00} \hat{\mathcal{I}}_{2,01}^2}{(\hat{\mathcal{I}}_{2,00} \hat{\mathcal{I}}_{2,11} - \hat{\mathcal{I}}_{2,01}^2)^2},$$

where $\hat{\mathcal{I}}_{1,st} = \hat{\mathcal{I}}_{1,1,st} + \hat{\mathcal{I}}_{1,2,st}$, $s, t = 0, 1$, and $\hat{\mathcal{I}}_{1,r,st}$, $r = 1, 2$ are given in Example 1.1 (continued) in Section 1.8.5 but with $\hat{\gamma}_1$ replaced by 1, its value under H_0 ; furthermore, we have

$$\begin{aligned} \hat{\mathcal{I}}_{2,00} &= -\frac{(mk - 1)}{2\hat{\lambda}^2}, \\ \hat{\mathcal{I}}_{2,01} &= -\frac{(m - 1)k}{2\hat{\lambda}(1 + \hat{\gamma}_1 k)}, \\ \hat{\mathcal{I}}_{2,11} &= -\frac{(m - 1)k^2}{2(1 + \hat{\gamma}_1 k)^2}, \end{aligned}$$

again with $\hat{\gamma}_1$ replaced by 1, where $\hat{\lambda}$ is the REML estimator of λ (Exercise 2.8). The asymptotic null distribution of the test is χ_1^2 .

3. Jackknife method. For non-Gaussian longitudinal models, the asymptotic covariance matrix of the REML (ML) estimator may be estimated using the jackknife method discussed in Section 1.4.4. One advantage of the jackknife method is that it is one-formula-works-for-all. In fact, the same jackknife estimator not only applies to longitudinal linear mixed models, it also applies to longitudinal generalized linear mixed models, which we discuss in Chapters 4 and 5. Let ψ be the vector of all the parameters involved in a non-Gaussian longitudinal model, which includes fixed effects and variance components. Let $\hat{\psi}$ be the REML or ML estimator of ψ . Then, the jackknife estimator of the asymptotic covariance matrix of $\hat{\psi}$ is given by (1.43). Jiang and Lahiri (2004) showed that, under suitable conditions, the jackknife estimator is consistent in the sense that $\hat{\Sigma}_{\text{Jack}} = \Sigma + O_P(m^{-1-\delta})$ for some $\delta > 0$. Therefore, one may use $\hat{\Sigma} = \hat{\Sigma}_{\text{Jack}}$ on the left side of (2.10) for the asymptotic test. We consider a simple example.

Example 2.4 (The James–Stein estimator). Let y_i , $i = 1, \dots, m$ be independent such that $y_i \sim N(\theta_i, 1)$. In the context of simultaneous estimation of

$\theta = (\theta_1, \dots, \theta_m)'$, it is well known that for $m \geq 3$, the James–Stein estimator dominates the maximum likelihood estimator, given by $y = (y_1, \dots, y_m)'$ in terms of the frequentist risk under the sum of squared error loss function (e.g., Lehmann and Casella 1998, pp. 272–273). Efron and Morris (1973) provided an empirical Bayes justification of the James–Stein estimator. Their Bayesian model can be equivalently written as the following simple random effects model: $y_i = \alpha_i + \epsilon_i$, $i = 1, \dots, m$, where the sampling errors $\{\epsilon_i\}$ and the random effects $\{\alpha_i\}$ are independently distributed with $\alpha_i \sim N(0, \psi)$ and $\epsilon_i \sim N(0, 1)$, and ϵ and α are independent.

Now we drop the normality assumption. Instead, we assume that y_i , $1 \leq i \leq m$ ($m > 1$) are i.i.d. with $E(y_1) = 0$, $\text{var}(y_1) = \psi + 1$ and $E(|y_1|^d) < \infty$ ($d > 4$). Then, an M-estimator for ψ , which is the solution to the ML equation, is given by $\hat{\psi} = m^{-1} \sum_{i=1}^m y_i^2 - 1$. The delete- i M-estimator is $\hat{\psi}_{-i} = (m - 1)^{-1} \sum_{j \neq i} y_j^2 - 1$. The jackknife estimator of the asymptotic variance of $\hat{\psi}$ is given by

$$\hat{\sigma}_{\text{jack}}^2 = \frac{m-1}{m} \sum_{i=1}^m (\hat{\psi}_{-i} - \hat{\psi})^2.$$

4. Robust versions of classical tests. Robust testing procedures have been studied extensively in the literature. In particular, robust versions of the classical tests, that is, the Wald, score, and likelihood-ratio tests (e.g., Lehmann 1999, §7) have been considered. In the case of i.i.d. observations, see, Foutz and Srivastava (1977), Kent (1982), Hampel et. al. (1986), and Heritier and Ronchetti (1994), among others. In the case of independent but not identically distributed observations, see, for example, Schrader and Hettmansperger (1980), Chen (1985), Silvapulle (1992), and Kim and Cai (1993). In contrast to the independent cases, the literature on robust testing with dependent observations is not extensive. In particular, in the case of linear mixed models, such tests as the likelihood-ratio test were studied only under the normality assumption (e.g., Hartley and Rao 1967). Because the normality assumption is likely to be violated in practice, it would be interesting to know if the classical tests developed under normality are robust against departure from such a distributional assumption.

Jiang (2005c) considered robust versions of the Wald, score, and likelihood-ratio tests in the case of dependent observations, which he called W -, S - and L -tests, and applied the results to non-Gaussian linear mixed models. The approach is briefly described as follows with more details given in Section 2.7. Let $y = (y_k)_{1 \leq k \leq n}$ be a vector of observations not necessarily independent. Let ψ be a vector of unknown parameters that are associated with the joint distribution of y , but the entire distribution of y may not be known given ψ (and possibly other parameters). We are interested in testing the hypothesis:

$$H_0 : \psi \in \Psi_0 \tag{2.17}$$

versus $H_1: \psi \notin \Psi_0$, where $\Psi_0 \subset \Psi$, and Ψ is the parameter space. Suppose that there is a new parameterization ϕ such that, under the null hypothesis (2.17), $\psi = \psi(\phi)$ for some ϕ . Here $\psi(\cdot)$ is a map from Φ , the parameter space of ϕ , to Ψ . Note that such a reparameterization is almost always possible, but the key is to try to make ϕ unrestricted (unless completely specified, such as in Example 2.5 below). The following are some examples.

Example 2.5. Suppose that, under the null hypothesis, ψ is completely specified; that is, $H_0: \psi = \psi_0$. Then, under H_0 , $\psi = \phi = \psi_0$.

Example 2.6. Let $\psi = (\psi_1, \dots, \psi_p, \psi_{p+1}, \dots, \psi_q)'$, and suppose that one wishes to test the hypothesis $H_0: \psi_j = \psi_{0j}$, $p+1 \leq j \leq q$, where ψ_{0j} , $p+1 \leq j \leq q$ are known constants. Then, under the null hypothesis, $\psi_j = \phi_j$, $1 \leq j \leq p$, and $\psi_j = \psi_{0j}$, $p+1 \leq j \leq q$ for some (unrestricted) $\phi = (\phi_j)_{1 \leq j \leq p}$.

Example 2.7. Suppose that the null hypothesis includes inequality constraints: $H_0: \psi_j > \psi_{0j}$, $p_1+1 \leq j \leq p$, and $\psi_j = \psi_{0j}$, $p+1 \leq j \leq q$, where $p_1 < p < q$. Then, under the null hypothesis, $\psi_j = \phi_j$, $1 \leq j \leq p_1$, $\psi_j = \psi_{0j} + e^{\phi_j}$, $p_1+1 \leq j \leq p$, and $\psi_j = \psi_{0j}$, $p+1 \leq j \leq q$ for some (unrestricted) $\phi = (\phi_j)_{1 \leq j \leq p}$.

Let $L(\psi, y)$ be a function of ψ and y that takes positive values, and $l(\psi, y) = \log L(\psi, y)$. Let $L_0(\phi, y) = L(\psi(\phi), y)$, and $l_0(\phi, y) = \log L_0(\phi, y)$. Let q and p be the dimensions of θ and ϕ , respectively. Let $\hat{\psi}$ be an estimator of ψ , and $\hat{\phi}$ an estimator of ϕ . Note that here we do not require that $\hat{\psi}$ and $\hat{\phi}$ be the (global) maximizers of $l(\psi, y)$ and $l_0(\phi, y)$, respectively. But we require that $\hat{\psi}$ be a solution to $\partial l / \partial \psi = 0$, and $\hat{\phi}$ a solution to $\partial l_0 / \partial \phi = 0$.

We now loosely define matrices A , B , C , and Σ with the exact definitions given in section 2.7: A is the limit of the matrix of second derivatives of l with respect to θ ; B is the limit of the matrix of second derivatives of l_0 with respect to ϕ ; C is the limit of the matrix of first derivatives of θ with respect to ϕ ; and Σ is the asymptotic covariance matrix of $\partial l / \partial \theta$, all subject to suitable normalizations. As shown in Section 2.7, the normalizations are associated with sequences of nonsingular symmetric matrices G and H . The W -test is closely related to the following quantity.

$$\mathcal{W} = [\hat{\psi} - \psi(\hat{\phi})]' G Q_w^- G [\hat{\psi} - \psi(\hat{\phi})], \quad (2.18)$$

where Q_w^- is the unique Moore–Penrose inverse (see Appendix B) of

$$Q_w = [A^{-1} - C(C'AC)^{-1}C']\Sigma[A^{-1} - C(C'AC)^{-1}C'].$$

Let \hat{Q}_w^- be a consistent estimator of Q_w^- in the sense that $\|\hat{Q}_w^- - Q_w^-\| \rightarrow 0$ in probability. The W -test statistic, $\hat{\mathcal{W}}$, is defined by (2.18) with Q_w^- replaced by \hat{Q}_w^- . Similarly, we define the following:

$$\mathcal{S} = \left(\frac{\partial l}{\partial \psi} \bigg|_{\psi(\hat{\phi})} \right)' G^{-1} A^{-1/2} Q_s^- A^{-1/2} G^{-1} \left(\frac{\partial l}{\partial \psi} \bigg|_{\psi(\hat{\phi})} \right), \quad (2.19)$$

where Q_s^- is the unique Moore-Penrose inverse of

$$Q_s = (I - P)A^{-1/2}\Sigma A^{-1/2}(I - P),$$

and $P = A^{1/2}C(C'AC)^{-1}C'A^{1/2}$. Let \hat{A} and \hat{Q}_s^- be consistent estimators of A and Q_s^- , respectively, in the same sense as above. Note that, quite often, A only depends on ψ , of which a consistent estimator; that is, $\hat{\psi}$, is available. The S -test statistic, \hat{S} , is defined by (2.19) with A and Q_s^- replaced by \hat{A} and \hat{Q}_s^- , respectively. Finally, the L -ratio for testing (2.17) is defined as

$$\mathcal{R} = \frac{L_0(\hat{\phi}, y)}{L(\hat{\psi}, y)}.$$

Note that the L -ratio is the same as the likelihood ratio when $L(\psi, y)$ is a likelihood function. The L -test statistic is then $-2 \log \mathcal{R}$.

Jiang (2005c) showed that, under some regularity conditions, both the W - and S -tests have an asymptotic χ_r^2 distribution, where the degrees of freedom $r = \text{rank}\{\Sigma^{1/2}A^{-1/2}(I - P)\}$ with P given below (2.19). As for the L -test, the asymptotic distribution of $-2 \log \mathcal{R}$ is the same as $\lambda_1 \xi_1^2 + \dots + \lambda_r \xi_r^2$, where r is the same as before, $\lambda_1, \dots, \lambda_r$ are the positive eigenvalues of

$$Q_l = [A^{-1} - C(C'AC)^{-1}C']^{1/2}\Sigma[A^{-1} - C(C'AC)^{-1}C']^{1/2}, \quad (2.20)$$

and ξ_1, \dots, ξ_r are independent $N(0, 1)$ random variables. In particular, if Σ is nonsingular, then $r = q - p$. These general results apply, in particular, to non-Gaussian linear mixed models. See Section 2.7 for more details.

We now consider application of the robust versions of classical tests to non-Gaussian mixed ANOVA models. The models are defined in Section 1.2.2 and the estimation problems discussed in Section 1.4. Consider the Hartley-Rao variance components: $\lambda = \sigma_0^2$, $\gamma_i = \sigma_i^2/\sigma_0^2$, $1 \leq i \leq s$. Let $\gamma = (\gamma_i)_{1 \leq i \leq s}$, and $\psi = (\beta' \ \lambda \ \gamma')'$. Then, ψ is a vector of parameters, which alone may not completely determine the distribution of y . Nevertheless, in many cases, people are interested in testing hypotheses of the form (2.17), where $\Psi_0 \subset \Psi = \{\psi : \lambda > 0, \gamma_i \geq 0, 1 \leq i \leq s\}$, versus $H_1: \psi \notin \Psi_0$. We assume that there is a new parameterization ϕ such that, under the null hypothesis, $\psi = \psi(\phi)$ for some $\phi = (\phi_k)_{1 \leq k \leq d}$. Here $\psi(\cdot)$ is a map from Φ , the parameter space of ϕ , to Ψ . More specifically, let $q = p + s + 1$, which is the dimension of ψ . We assume that there is a subset of indices $1 \leq i_1 < \dots < i_d \leq q$ such that

$$\begin{cases} \psi_{i_k}(\phi) \text{ is a function of } \phi_k, & 1 \leq k \leq d, \quad \text{and} \\ \psi_i(\phi) \text{ is a constant,} & i \in \{1, \dots, q\} \setminus \{i_1, \dots, i_d\}. \end{cases} \quad (2.21)$$

Intuitively, the null hypothesis imposes constraints on ψ , therefore there are less free parameters under H_0 , and ϕ represents the vector of free parameters after some changes of variables. Note that such a reparameterization almost always exists, but the key is to try to make ϕ unrestricted unless completely specified.

When normality is assumed, the use of the likelihood-ratio test for complex hypotheses and unbalanced data was first proposed by Hartley and Rao (1967), although rigorous justification was not given. Welham and Thompson (1997) showed the equivalence of the likelihood ratio, score, and Wald tests under normality. On the other hand, Richardson and Welsh (1996) considered the likelihood-ratio test without assuming normality, whose approach is similar to our L -test, but their goal was to select the (fixed) covariates. Under the normality assumption, the log-likelihood function for estimating θ is given by

$$l(\psi, y) = \text{constant} - \frac{1}{2} \left\{ n \log \lambda + \log(|V|) + \frac{1}{\lambda} (y - X\beta)' V^{-1} (y - X\beta) \right\},$$

where $V = V_\gamma = I + \sum_{i=1}^s \gamma_i V_i$ with I being the n -dimensional identity matrix, $V_i = Z_i Z_i'$, $1 \leq i \leq s$, and $|V|$ the determinant of V . The restricted log-likelihood for estimating λ, γ is given by

$$l_R(\lambda, \gamma, y) = \text{constant} - \frac{1}{2} \left\{ (n-p) \log \lambda + \log(|K'VK|) + \frac{y'Py}{\lambda} \right\},$$

where K is any $n \times (n-p)$ matrix such that $\text{rank}(K) = n-p$ and $K'X = 0$, and $P = P_\gamma = K(K'VK)^{-1}K' = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}$ (see Appendix B). The restricted log-likelihood is only for estimating the variance components. It is then customary to estimate β by the empirical best linear unbiased estimator:

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y,$$

where $\hat{V} = V_{\hat{\gamma}}$, and $\hat{\gamma} = (\hat{\gamma}_i)_{1 \leq i \leq s}$ is the REML estimator of γ . Alternatively, one may define the following “restricted log-likelihood” for ψ .

$$l_R(\psi, y) = \text{constant} - \frac{1}{2} \left\{ (n-p) \log \lambda + \log |K'VK| + \frac{1}{\lambda} (y - X\beta)' V^{-1} (y - X\beta) \right\}.$$

It is easy to show that the maximizer of $l_R(\psi, y)$ is $\hat{\psi} = (\hat{\beta}' \hat{\lambda} \hat{\gamma}')'$, where $\hat{\lambda}$ and $\hat{\gamma}$ are the REML estimators, and $\hat{\beta}$ is given above with $\hat{V} = V_{\hat{\gamma}}$. The difference is that, unlike $l(\psi, y)$, $l_R(\psi, y)$ is not a log-likelihood even if normality holds. Nevertheless, it can be shown that both $l(\psi, y)$ and $l_R(\psi, y)$ can be used as the objective function to test (2.17) under a non-Gaussian mixed linear model. The details are given in Section 2.7.1. We now consider an example.

Example 2.2 (Continued). In this case, we have $q = 3$, $\psi_1 = \mu$, $\psi_2 = \lambda = \tau^2$, and $\psi_3 = \gamma = \sigma^2/\tau^2$. Consider the hypothesis $H_0: \lambda = 1, \gamma > 1$. Note that under H_0 we have $\mu = \phi_1$, $\lambda = 1$, and $\gamma = 1 + e^{\phi_2}$ for unrestricted ϕ_1, ϕ_2 . Thus, (2.21) is satisfied with $d = 2$, $i_1 = 1$, and $i_2 = 3$. The Gaussian log-likelihood is given by (Exercise 2.9)

$$l(\psi, y) = c - \frac{1}{2} \left\{ mk \log(\lambda) + m \log(1 + k\gamma) + \frac{\text{SSE}}{\lambda} + \frac{\text{SSA}}{\lambda(1 + k\gamma)} + \frac{mk(\bar{y}_{..} - \mu)^2}{\lambda(1 + k\gamma)} \right\},$$

where c is a constant, $\text{SSA} = k \sum_{i=1}^m (\bar{y}_{i.} - \bar{y}_{..})^2$, $\text{SSE} = \sum_{i=1}^m \sum_{j=1}^k (y_{ij} - \bar{y}_{i.})^2$, $\bar{y}_{..} = (mk)^{-1} \sum_{i=1}^m \sum_{j=1}^k y_{ij}$, and $\bar{y}_{i.} = k^{-1} \sum_{j=1}^k y_{ij}$. Here we have $\psi = (\mu, \lambda, \gamma)'$, $\phi = (\phi_1, \phi_2)'$, and $\psi(\phi) = (\phi_1, 1, 1 + e^{\phi_2})'$, where ϕ_1 and ϕ_2 are unrestricted. The solution to the (Gaussian) ML equation is given by $\hat{\psi}_1 = \hat{\mu} = \bar{y}_{..}$, $\hat{\psi}_2 = \hat{\lambda} = \text{MSE}$, and $\hat{\psi}_3 = \hat{\gamma} = (1/k) \{ (1 - 1/m)(\text{MSA}/\text{MSE}) - 1 \}$, where $\text{MSA} = \text{SSA}/(m - 1)$ and $\text{MSE} = \text{SSE}/m(k - 1)$. On the other hand, it is easy to show that the solution to the ML equation under the null hypothesis is given by $\hat{\phi}_1 = \bar{y}_{..}$, $\hat{\phi}_2 = \log \{ (1/k)(1 - 1/m)\text{MSA} - (1 + 1/k) \}$, provided that the term inside the logarithm is positive. Because $E(\text{MSA}) = 1 + k\gamma > k + 1$ under H_0 (Exercise 2.9), it is easy to show that, as $m \rightarrow \infty$, the logarithm is well defined with probability tending to one.

We now specify the matrices A , C , G , and Σ under the additional assumption that $E(\alpha_1^3) = E(\epsilon_{11}^3) = 0$. According to Theorem 2.4, A is given by (2.62), and it can be shown that $X'V^{-1}X/\lambda n = 1/\lambda^2(1 + k\gamma)$, $A_1 = \sqrt{k}/2\lambda^2(1 + k\gamma)$, and $A_2 = k^2/2\lambda^2(1 + k\gamma)^2$. Again, by Theorem 2.4, $G = \text{diag}(\sqrt{mk}, \sqrt{mk}, \sqrt{m})$; C is the 3×2 matrix whose first column is $(1, 0, 0)'$ and second column is $(0, 0, e^{\phi_2})'$. Finally, $\Sigma = A + \Delta$ with Δ given by (2.63), and it can be shown that

$$\begin{aligned} \frac{\Delta_{00}}{n} &= \frac{1}{4\lambda^4(1 + k\gamma)^2} [\kappa_0 \{1 + (k - 1)\gamma\}^2 + \kappa_1 k\gamma^2], \\ \Delta_1 &= \frac{\sqrt{k}}{4\lambda^4(1 + k\gamma)^3} [\kappa_0 \{1 + (k - 1)\gamma\} + \kappa_1 k^2\gamma^2], \\ \Delta_2 &= \frac{k}{4\lambda^4(1 + k\gamma)^4} (\kappa_0 + \kappa_1 k^3\gamma^2), \end{aligned}$$

where $\kappa_0 = \{E(\epsilon_{11}^4)/\tau^4\} - 3$ and $\kappa_1 = \{E(\alpha_1^4)/\sigma^4\} - 3$.

It can be shown that, in this case, the W -test statistic reduces to

$$\hat{\chi}_w^2 = \left(\frac{2k}{k - 1} + \hat{\kappa}_0 \right)^{-1} mk(\text{MSE} - 1)^2,$$

where $\hat{\kappa}_0$ is the EMM estimator of κ_0 given in Example 2.2 (Continued) below Lemma 2.1. Note that, by the consistency of $\hat{\kappa}_0$ (Exercise 2.7), we have, as $m \rightarrow \infty$,

$$\begin{aligned} \frac{2k}{k - 1} + \hat{\kappa}_0 &\xrightarrow{P} \frac{2k}{k - 1} + \kappa_0 \\ &\geq E(\epsilon_{11}^4) - 1 > 0, \end{aligned}$$

under H_0 , unless ϵ_{11}^2 is degenerate. Thus, with the exception of this extreme case, the denominator in $\hat{\chi}_w^2$ is positive with probability tending to one under the null hypothesis. By Theorem 2.4, as $m \rightarrow \infty$, the asymptotic null distribution of the W -test is χ_1^2 (Exercise 2.10).

As it turns out, the S -test statistic is identical to the W -test statistic in this case, and it has the same asymptotic null distribution (Exercise 2.11).

Finally, the L -test statistic is equal to

$$-2 \log R = m(k-1)\{\text{MSE} - 1 - \log(\text{MSE})\}$$

in this case. Suppose that $m \rightarrow \infty$ and k is fixed ($k \geq 2$). Then, it can be shown that $r = 1$ in this case, therefore, by Theorem 2.5, the asymptotic null distribution of $-2 \log R$ is the same as $\lambda_1 \chi_1^2$, where λ_1 is the positive eigenvalue of Q_l given by (2.20) evaluated under H_0 . It can be shown that $\lambda_1 = 1 + \{(k-1)/2k\}\kappa_0$, which is estimated by $1 + \{(k-1)/2k\}\hat{\kappa}_0$. Note that if $\kappa_0 = 0$, as will be the case if the errors are normal, the asymptotic null distribution of the L -test is χ_1^2 , which is the same as that for the W - and S -tests. Interestingly, the latter result does not require that the random effects are normal (Exercise 2.12).

2.2 Confidence Intervals in Linear Mixed Models

2.2.1 Confidence Intervals in Gaussian Mixed Models

Confidence intervals in linear mixed models include confidence intervals for fixed effects, confidence intervals for variance components, and confidence intervals for functions of variance components. Among the latter, difference and ratio are two simple functions that are frequently used. Other functions such as the heritability, an important quantity in genetics, may be expressed as functions of these two simple functions. For simplicity, the term confidence intervals for variance components is here understood as including functions of variance components. We first consider confidence intervals under Gaussian linear mixed models.

1. *Exact confidence intervals for variance components.* It is known that in some special cases, mostly balanced cases, exact confidence intervals for variance components can be derived. Here we do not attempt to list all such cases where exact confidence intervals are available. For more details, see Burdick and Graybill (1992). Instead, our approach is to introduce a basic method used to derive exact confidence intervals, so that it may be applied to different cases whenever applicable. The basic idea is to find a *pivotal quantity*, that is, a random variable that depends on both the observations and the variance component, for which an exact confidence interval is to be constructed. Quite often, such a pivotal quantity is in the form of either an “ F -statistic” or a “ χ^2 -statistic”. Here the quotes indicate that the quantity is not

really a statistic because it involves the variance component. We illustrate the method by examples.

Example 2.2 (Continued). Consider the Hartley–Rao form of variance components $\lambda = \tau^2$ and $\gamma = \sigma^2/\tau^2$. Suppose that one is interested in constructing an exact confidence interval for γ . Consider the following quantity

$$F = \frac{\text{MSA}}{(1 + k\gamma)\text{MSE}},$$

where $\text{MSA} = \text{SSA}/(m-1)$ and $\text{MSE} = \text{SSE}/m(k-1)$. It can be shown that, under normality, F has an F -distribution with $m-1$ and $m(k-1)$ degrees of freedom (Exercise 2.13). It follows that, given ρ ($0 < \rho < 1$), an exact $(1-\rho)\%$ confidence interval for γ is

$$\left[\frac{1}{k} \left(\frac{R}{F_U} - 1 \right), \frac{1}{k} \left(\frac{R}{F_L} - 1 \right) \right],$$

where $R = \text{MSA}/\text{MSE}$, $F_L = F_{m-1, m(k-1), 1-\rho/2}$, and $F_U = F_{m-1, m(k-1), \rho/2}$ (Exercise 2.13).

Example 2.3 (Continued). Suppose that the problem of interest is to construct an exact confidence interval for the variance of any single observation y_{ij} ; that is, $\text{var}(y_{ij}) = \sigma^2 + \tau^2$. Let c_{ij} , $1 \leq j \leq k_i$ be constants such that $\sum_{j=1}^{k_i} c_{ij} = 0$ and $\sum_{j=1}^{k_i} c_{ij}^2 = 1 - 1/k_i$. Define $u_i = \bar{y}_i + \sum_{j=1}^{k_i} c_{ij}y_{ij}$, $1 \leq i \leq m$. It can be shown that u_1, \dots, u_m are independent and normally distributed with mean μ and variance $\sigma^2 + \tau^2$ (Exercise 2.14). Thus, the quantity

$$\chi^2 = \frac{\sum_{i=1}^m (u_i - \bar{u})^2}{\sigma^2 + \tau^2}$$

is distributed as χ_{m-1}^2 . It follows that an exact $(1-\rho)\%$ confidence interval for $\sigma^2 + \tau^2$ is

$$\left[\frac{\sum_{i=1}^m (u_i - \bar{u})^2}{\chi_{m-1, \rho/2}^2}, \frac{\sum_{i=1}^m (u_i - \bar{u})^2}{\chi_{m-1, 1-\rho/2}^2} \right].$$

The method used in the above example for constructing an exact confidence interval for $\sigma^2 + \tau^2$ is due to Burdick and Sielken (1978). In fact, the authors developed a method that can be used to obtain an exact confidence interval for $a\sigma^2 + b\tau^2$, where a, b are positive constants subject to some additional constraints. One such constraint is that $b \neq 0$. Thus, for example, the method cannot give an exact confidence interval for σ^2 (see Exercise 2.15). This example shows the limitation of the method used to construct exact confidence intervals. In fact, no existing method is known to be able to obtain an

exact confidence interval for σ^2 in an analytic form. On the other hand, approximate confidence intervals do exist for σ^2 and other variance components. We discuss such methods next.

2. *Approximate confidence intervals for variance components.* Satterthwaite (1946) proposed a method, which extended an earlier approach of Smith (1936), for balanced ANOVA models. The goal was to construct a confidence interval for a quantity in the form $\zeta = \sum_{i=1}^h c_i \lambda_i$, where $\lambda_i = E(S_i^2)$ and S_i^2 is the mean sum of squares corresponding to the i th factor (fixed or random) in the model (e.g., Scheffé 1959). Note that many variance components can be expressed in this form; for example, the variance of y_{ij} , $\sigma^2 + \tau^2$, in Example 2.3 can be expressed as $(1/k)E(S_1^2) + (1 - 1/k)E(S_2^2)$, where S_1^2 is the mean sum of squares corresponding to α and S_2^2 that corresponding to ϵ . The idea was to find an appropriate “degrees of freedom,” say, d , such that the first two moments of the random variable $d \sum_{i=1}^h c_i S_i^2 / \zeta$ match those of a χ_d^2 random variable. This approach is known as Satterthwaite’s procedure. Graybill and Wang (1980) proposed a method that improved Satterthwaite’s procedure. The authors called their method the modified large sample (MLS) method. The method provides an approximate confidence interval for a nonnegative linear combination of the λ_i s, which is exact when all but one of the coefficients in the linear combination are zero. We describe the Graybill–Wang for the special case of balanced one-way random effects model (Example 2.2).

Suppose that one is interested in constructing a confidence interval for $\zeta = c_1 \lambda_1 + c_2 \lambda_2$, where $c_1 \geq 0$ and $c_2 > 0$. This problem is equivalent to constructing a confidence interval for $\zeta = c \lambda_1 + \lambda_2$, where $c \geq 0$. A uniformly minimum variance unbiased estimator (UMVUE, e.g., Lehmann and Casella 1998) of the quantity is given by $\hat{\zeta} = c S_1^2 + S_2^2$. Furthermore, it can be shown that $\hat{\zeta}$ is asymptotically normal such that $(\hat{\zeta} - \zeta) / \sqrt{\text{var}(\hat{\zeta})}$ has a limiting $N(0, 1)$ distribution (Exercise 2.16). Furthermore, the variance of $\hat{\zeta}$ is given by $c^2 \{2\lambda_1^2 / (m - 1)\} + 2\lambda_2^2 / m(k - 1)$. Again, recall that S_j^2 is an unbiased (and consistent) estimator of λ_j $j = 1, 2$ (Exercise 2.16). This allows one to obtain a large sample confidence interval for ζ as follows.

$$\left[\hat{\zeta} - z_{\rho/2} \sqrt{c^2 \left(\frac{2S_1^4}{m-1} \right) + \frac{2S_2^4}{m(k-1)}}, \right. \\ \left. \hat{\zeta} + z_{\rho/2} \sqrt{c^2 \left(\frac{2S_1^4}{m-1} \right) + \frac{2S_2^4}{m(k-1)}} \right], \quad (2.22)$$

where $1 - \rho$ is the confidence coefficient. The confidence interval (2.22) is expected to be accurate when the sample size is large, that is, when $m \rightarrow \infty$. However, small sample performance is not guaranteed. Graybill and Wang proposed to modify the constants $z_{\rho/2}$, $2/(m - 1)$ and $2/m(k - 1)$, so that the confidence interval will be exact when either $\lambda_1 = 0$ or $\lambda_2 = 0$. Their confidence interval is given by

$$\left[\hat{\zeta} - \sqrt{G_1^2 c^2 S_1^4 + G_2^2 S_2^4}, \hat{\zeta} + \sqrt{H_1^2 c^2 S_1^4 + H_2^2 S_2^4} \right],$$

where $G_1 = 1 - (m-1)/\chi_{m-1, \rho/2}^2$, $G_2 = 1 - m(k-1)/\chi_{m(k-1), \rho/2}^2$, $H_1 = (m-1)/\chi_{m-1, 1-\rho/2}^2 - 1$, and $H_2 = m(k-1)/\chi_{m(k-1), \rho/2}^2 - 1$. Using numerical integration, Graybill and Wang compared confidence coefficients of the MLS confidence intervals with those of Satterthwaite and Welch (Welch 1956). They concluded that the confidence coefficients of the MLS are closer to the nominal levels than those of Satterthwaite and Welch. As for the length of the confidence intervals, Graybill and Wang conducted a simulation study. The results showed that average widths of two types of MLS confidence intervals, namely, the shortest unbiased confidence interval and shortest confidence interval, are generally smaller than those of Welch's.

Sometimes, the variance components of interest cannot be expressed as a nonnegative linear combination of the λ_i s. For example, in Example 2.2, the variance $\sigma^2 = (\lambda_1 - \lambda_2)/k$, so the coefficients in the linear combination have different signs. It is therefore of interest to obtain confidence intervals for $\zeta = \sum_{i=1}^h c_i \lambda_i$, where the c_i s may have different signs. Healy (1961) proposed a procedure that may be used to obtain an exact confidence interval for $c_1 \lambda_1 - c_2 \lambda_2$, where c_1 and c_2 are nonnegative. However, the procedure requires a randomization device. In other words, the confidence interval is not solely determined by the data. Several authors have proposed (solely data-based) approximate confidence intervals for ζ . For example, Ting et al. (1990) proposed a procedure similar to Graybill and Wang (1980) discussed above. Note that a large sample confidence interval such as (2.22) based on asymptotic normality of $\hat{\zeta}$ does not require that the signs of the c_i s be the same. All one has to do is to modify the coefficients of the large sample confidence interval so that it performs better in small sample situations. See Ting et al. (1990) for details. Burdick and Graybill (1992) reviewed several approximate procedures for constructing confidence intervals for ζ . They conclude that there is little difference in terms of performance of the proposed procedures.

Finally, one should bear in mind that, in cases of large samples, a confidence interval as simple as (2.22) can be used without modification. Such a procedure is much easier to derive and calculate. We return to this method in the next section.

3. Simultaneous confidence intervals. Hartley and Rao (1967) derived a simultaneous confidence region for the variance ratios $\gamma_i = \sigma_i^2/\tau^2$, $i = 1, \dots, s$ (i.e., the Hartley-Rao form of variance components; see Section 1.2.1.1) in a Gaussian mixed ANOVA model. Their derivation is based on maximum likelihood estimation, a method that we visit again in the next section. The Hartley-Rao confidence region is quite general, that is, it applies to a general ANOVA model, balanced or unbalanced. On the other hand, in some special cases different methods may result in confidence intervals that are easier to interpret. For example, Khuri (1981) developed a method of constructing simultaneous confidence intervals for all continuous functions of variance

components in the balanced random effects model, a special case of the mixed ANOVA model.

It should be pointed out that, provided one knows how to construct confidence intervals for the individual variance components, by then Bonferroni inequality, a conservative simultaneous confidence interval for the variance components can always be constructed. Suppose that $[L_i, U_i]$ is a $(1 - \rho_i)\%$ confidence interval for the variance component θ_i , $i = 1, \dots, q$. Then, by the Bonferroni inequality, the set of intervals $[L_i, U_i]$, $i = 1, \dots, q$ is a (conservative) simultaneous confidence interval for θ_i , $i = 1, \dots, q$ with confidence coefficient greater than or equal to $1 - \sum_{i=1}^q \rho_i$. Sometimes, the confidence coefficient may be improved if there is independence among the individual confidence intervals. For example, in the balanced normal random effects model, let n_i be the degrees of freedom associated with S_i^2 , the mean sum of squares corresponding to the i th factor (fixed or random). Then, it is known that $n_i S_i^2 / \lambda_i$ has a χ^2 distribution with n_i degrees of freedom, where $\lambda_i = E(S_i^2)$. Furthermore, the random variables $n_i S_i^2 / \lambda_i$, $i = 1, \dots, h$ are independent (e.g., Scheffé 1959). It follows that a $(1 - \rho)\%$ confidence interval for λ_i is

$$\left[\frac{n_i S_i^2}{\chi_{n_i, \rho/2}^2}, \frac{n_i S_i^2}{\chi_{n_i, 1-\rho/2}^2} \right], \quad (2.23)$$

and, furthermore, the set of intervals (2.23) with $i = 1, \dots, h$ is a simultaneous confidence interval for λ_i , $i = 1, \dots, h$ with confidence coefficient $(1 - \rho)^h$. Note that $(1 - \rho)^h \geq 1 - h\rho$ for any integer $h \geq 1$.

4. *Confidence intervals for fixed effects.* For the vector of fixed effects β in (1.1), the best linear unbiased estimator, or BLUE, is given by (1.36), provided that the expression does not involve unknown variance components. Furthermore, we have

$$\text{Var}(\hat{\beta}_{\text{BLUE}}) = (X'V^{-1}X)^{-1}. \quad (2.24)$$

In fact, under mild conditions, $\hat{\beta}_{\text{BLUE}}$ is asymptotically normal with mean vector β and asymptotic covariance matrix given by the right side of (2.24). It is known that in some special cases, mostly in the balanced situations, the right side of (1.36) does not depend on the variance components, therefore $\hat{\beta}_{\text{BLUE}}$ can be used as an estimator. However, even in those cases the right side of (2.24) typically depends on the variance components. Of course, in general, both $\hat{\beta}_{\text{BLUE}}$ and its covariance matrix depend on the variance components. Therefore, to construct a confidence interval for a fixed effect, or more generally, any linear function of β , one needs to replace the unknown variance components by consistent estimators, for example, REML estimators. Except for some special cases (see Example 2.8 below), the resulting confidence interval will be approximate in the sense that its confidence coefficient approaches the nominal level as sample size increases. We consider an example.

Example 2.8. Consider the following model, which is a special case of the so-called *nested error regression model*:

$$y_{ij} = \beta_0 + \beta_1 x_i + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, k_i,$$

where β_0, β_1 are unknown regression coefficients, x_i s are known covariates, α_i s are random effects, and ϵ_{ij} s are errors. Suppose that the random effects and errors are independent and normally distributed such that $E(\alpha_i) = 0$, $\text{var}(\alpha_i) = \sigma^2$, $E(\epsilon_{ij}) = 0$, and $\text{var}(\epsilon_{ij}) = \tau^2$.

It can be shown (Exercise 2.17) that, in this case, (1.36) gives the following expressions for the BLUE,

$$\hat{\beta}_{\text{BLUE},0} = \frac{(\sum_{i=1}^m d_i x_i^2)(\sum_{i=1}^m d_i \bar{y}_{i\cdot}) - (\sum_{i=1}^m d_i x_i)(\sum_{i=1}^m d_i x_i \bar{y}_{i\cdot})}{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2}, \quad (2.25)$$

$$\hat{\beta}_{\text{BLUE},1} = \frac{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i \bar{y}_{i\cdot}) - (\sum_{i=1}^m d_i x_i)(\sum_{i=1}^m d_i \bar{y}_{i\cdot})}{(\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2}, \quad (2.26)$$

where $d_i = k_i/(\tau^2 + k_i\sigma^2)$. It follows that, when $k_i = k$, $1 \leq i \leq m$ (i.e., in the balanced case), we have

$$\begin{aligned} \hat{\beta}_{\text{BLUE},0} &= \frac{(\sum_{i=1}^m x_i^2)(\sum_{i=1}^m \bar{y}_{i\cdot}) - (\sum_{i=1}^m x_i)(\sum_{i=1}^m x_i \bar{y}_{i\cdot})}{m \sum_{i=1}^m x_i^2 - (\sum_{i=1}^m x_i)^2}, \\ \hat{\beta}_{\text{BLUE},1} &= \frac{\sum_{i=1}^m (x_i - \bar{x})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}{\sum_{i=1}^m (x_i - \bar{x})^2}. \end{aligned}$$

It is seen that in the balanced case, the BLUE does not depend on the variance components but in the unbalanced case it does. Furthermore, $\hat{\beta}_{\text{BLUE}} = (\hat{\beta}_{\text{BLUE},0}, \hat{\beta}_{\text{BLUE},1})'$. It can be shown by (2.24) that

$$\text{Var}(\hat{\beta}_{\text{BLUE}}) = \frac{1}{\tau^2 D} \begin{pmatrix} \sum_{i=1}^m d_i x_i^2 & -\sum_{i=1}^m d_i x_i \\ -\sum_{i=1}^m d_i x_i & \sum_{i=1}^m d_i \end{pmatrix}, \quad (2.27)$$

where $D = (\sum_{i=1}^m d_i)(\sum_{i=1}^m d_i x_i^2) - (\sum_{i=1}^m d_i x_i)^2$ (Exercise 2.17). So even in the balanced case the covariance matrix of BLUE depends on the variance components.

When the variance components involved in BLUE are replaced by their estimators, the resulting estimator is often called empirical BLUE, or EBLUE. It is easy to see that, under normality, EBLUE is the same as the MLE of β , if the variance components are replaced by their MLE. It should be pointed out that EBLUE is more complicated and, in particular, not linear in y . Furthermore, if one replaces the variance components involved on the right side of (2.24) by their estimators, the result would underestimate the true variation of EBLUE. In fact, Kackar and Harville (1981) showed that EBLUE, denoted by $\hat{\beta}$, is still an unbiased estimator of β , that is $E(\hat{\beta}) = \beta$, provided that the data are normal and estimators of the variance components are even

and translation invariant (see Section 2.8 for more detail). In addition, the authors showed that, under normality

$$\text{var}(a'\hat{\beta}) = \text{var}(a'\hat{\beta}_{\text{BLUE}}) + E\{a'(\hat{\beta} - \hat{\beta}_{\text{BLUE}})\}^2 \quad (2.28)$$

for any given vector a . Because $\text{var}(a'\hat{\beta}_{\text{BLUE}}) = a'\text{Var}(\hat{\beta}_{\text{BLUE}})a$, the first term on the right side of (2.28) can be estimated by the right side of (2.24) with the variance components replaced by their estimators. However, there is a second term on the right side of (2.28) that cannot be estimated this way. Fortunately, for constructing confidence intervals for the fixed effects, this complication does not necessarily cause any problem, at least in the large-sample situation. In fact, for ANOVA models, Jiang (1998b) showed that, when the variance components are estimated by the REML estimators, the asymptotic covariance matrix of $\hat{\beta}$ is still given by the right side of (2.24) (in spite of estimation of the variance components). It is known (e.g., Miller 1977) that, when the variance components are estimated by the MLE, the asymptotic covariance matrix of $\hat{\beta}$ is also given by the right side of (2.24). Thus, in such cases, a (large-sample) confidence interval for $a'\beta$ is given by

$$\left[a'\hat{\beta} - z_{\rho/2}\{a'(X'\hat{V}^{-1}X)^{-1}a\}^{1/2}, \right. \\ \left. a'\hat{\beta} + z_{\rho/2}\{a'(X'\hat{V}^{-1}X)^{-1}a\}^{1/2} \right], \quad (2.29)$$

where \hat{V} is V with the variance components replaced by their REML or ML estimators. It is shown in section 2.3 that the complication in EBLUE becomes important in the prediction of a mixed effect, that is, a linear combination of fixed and random effects.

Example 2.8 (Continued). Suppose that one is interested in constructing a confidence interval for $\hat{\beta}_1$. By (2.29) and (2.27), taking $a = (0, 1)'$, a large sample confidence interval is

$$\left[\hat{\beta}_1 - z_{\rho/2} \left(\frac{\sum_{i=1}^m \hat{d}_i}{\hat{\tau}^2 \hat{D}} \right)^{1/2}, \hat{\beta}_1 + z_{\rho/2} \left(\frac{\sum_{i=1}^m \hat{d}_i}{\hat{\tau}^2 \hat{D}} \right)^{1/2} \right],$$

where $\hat{d}_i = k_i/(\hat{\tau}^2 + k_i\hat{\sigma}^2)$, $\hat{\beta}_1$ is given by (2.26) with d_i replaced by \hat{d}_i , $1 \leq i \leq m$, and \hat{D} is D with d_i replaced by \hat{d}_i , $1 \leq i \leq m$. Here $\hat{\sigma}^2$ and $\hat{\tau}^2$ are understood as the REML (or ML) estimators.

2.2.2 Confidence Intervals in Non-Gaussian Linear Mixed Models

For non-Gaussian linear mixed models, exact confidence intervals for parameters of interest usually do not exist. Therefore, methods of constructing confidence intervals will be based on large sample theory. Suppose that one is

interested in obtaining a confidence interval for a linear function of the parameters, which may include fixed effects and variance components. Let ψ be the vector of all fixed parameters involved in a non-Gaussian linear mixed model. Suppose that an estimator of ψ , say $\hat{\psi}$, is available which is consistent and asymptotically normal; that is, (2.7) holds. If a consistent estimator of Σ , the asymptotic covariance matrix of $\hat{\psi}$, is available, say $\hat{\Sigma}$, then, for any linear function $a'\psi$, where a is a known vector, one may be able to show that $a'(\hat{\psi} - \psi)/\sqrt{a'\hat{\Sigma}a}$ is asymptotically standard normal. Therefore, a large-sample $(1 - \rho)\%$ confidence interval ($0 < \rho < 1$) for $a'\psi$ is

$$\left[a'\hat{\psi} - z_{\rho/2}\sqrt{a'\hat{\Sigma}a}, a'\hat{\psi} + z_{\rho/2}\sqrt{a'\hat{\Sigma}a} \right].$$

We now consider two special cases of non-Gaussian linear mixed models and discuss how to estimate Σ in those cases.

1. *ANOVA models.* For ANOVA models, Jiang (1996) derived asymptotic distributions of both REML and ML estimators without the normality assumption. Jiang (1998b) extended these results to include estimators of fixed effects. The main result of the latter is summarized as follows. Consider the Hartley–Rao form of variance components (see Section 1.2.1.1). Let the normalizing constants p_i , $0 \leq i \leq s$ and matrices \mathcal{M} , \mathcal{J} be defined as in Theorem 1.1 of Chapter 1. Define $\mathcal{P} = \mathcal{M}\text{diag}(p_0, p_1, \dots, p_s)$, $\mathcal{Q} = (X'V^{-1}X)^{1/2}$, $\mathcal{R} = \mathcal{J}^{1/2}\mathcal{T}\mathcal{C}$, where

$$\mathcal{T} = \left(\frac{V_{i,l}(\gamma)E(\omega_l^3)}{p_i\lambda^{1(i=0)}} \right)_{0 \leq i \leq s, 1 \leq l \leq n+m},$$

$$\mathcal{C} = \lambda^{1/2}b(\gamma)V^{-1}X\mathcal{Q}^{-1}$$

with $V_{i,l}$, ω_l and $b(\gamma)$ defined above Theorem 1.1. Then, under suitable conditions, we have

$$\begin{pmatrix} \mathcal{P} & \mathcal{R}\mathcal{Q} \\ \mathcal{R}'\mathcal{P} & \mathcal{Q} \end{pmatrix} \begin{pmatrix} \hat{\theta} - \theta \\ \hat{\beta} - \beta \end{pmatrix} \xrightarrow{\mathcal{D}} N(0, I_{p+s+1}), \quad (2.30)$$

where $\hat{\beta}$ is the EBLUE with REML estimators of variance components (in other words, $\hat{\beta}$ is the REML estimator of β ; see Section 1.3.2). Because, under normality, $\mathcal{T} = 0$ hence $\mathcal{R} = 0$, the normalizing matrix on the left side of (2.30) reduces to $\text{diag}(\mathcal{P}, \mathcal{Q})$ in this case. However, for non-Gaussian linear mixed models, the normalizing matrix in (2.30) may involve additional parameters such as the third and fourth moments of the random effects and errors. A method of estimating the higher moments, known as EMM, has been introduced earlier (see Section 2.1.2.1), under the assumption (2.14) (which implies $E(\omega_l) = 0$, $1 \leq l \leq n + m$). To see how much difference there may be if one ignores the higher moments, consider the following example.

Example 2.2 (Continued). If normality is not assumed, it can be shown, by (2.30), that the asymptotic variance of $\sqrt{mk}(\hat{\lambda} - \lambda)$ is $\lambda^2/2 + \kappa_0$, that

is, $\sqrt{mk}(\hat{\lambda} - \lambda) \rightarrow N(0, \lambda^2/2 + \kappa_0)$ in distribution, where κ_0, κ_1 are defined below (2.14). Similarly, the asymptotic variance of $\sqrt{m}(\hat{\gamma} - \gamma)$ is $\gamma^2/2 + \kappa_1/\lambda^2$. Therefore, the difference in asymptotic variance from that under normality is κ_0 for the estimation of λ , and κ_1/λ^2 for the estimation of γ .

If (2.14) is not known to hold, the EMM may not apply. In this case, an alternative method would be that of partially observed information introduced in Section 1.4.2. Note that the latter method applies more generally not only to mixed ANOVA models but also to other types of non-Gaussian linear mixed models for estimating the asymptotic covariance matrix of the REML or ML estimator.

2. Longitudinal models. For longitudinal models, the asymptotic covariance matrix of the vector of parameters of interest, which may include fixed effects and variance components, may be estimated using the jackknife method introduced in Section 1.4.4 [see (1.43)]. Alternatively, the asymptotic covariance matrix may also be estimated by partially observed information. See the remark at the end of Section 1.4.2.

2.3 Prediction

There are two types of prediction problems in the context of linear mixed models. The first is the prediction of a random effect, or, more generally, a mixed effect. Here we focus on a linear mixed effect, which can be expressed as $\eta = a'\alpha + b'\beta$, where a, b are known vectors, and α and β are the vectors of random and fixed effects, respectively, in (1.1). This type of prediction problem has a long history, starting with C. R. Henderson in his early work in the field of animal breeding (e.g., Henderson 1948). The best-known method for this kind of prediction is best linear unbiased prediction, or BLUP. Robinson (1991) gives a wide-ranging account of BLUP with examples and applications. The second type of prediction is that of a future observation. In contrast to the first type, prediction of the second type has received much less attention, although there are plenty of such prediction problems with practical interest (e.g., Jiang and Zhang 2002). In the next two sections we discuss these two types of predictions.

2.3.1 Prediction of Mixed Effect

1. Best prediction when all the parameters are known. When the fixed effects and variance components are known, the best predictor for $\xi = a'\alpha$, in the sense of minimum mean squared error (MSE), is its conditional expectation given the data; that is,

$$\tilde{\xi} = E(\xi|y) = a'E(\alpha|y) . \quad (2.31)$$

Assuming normality of the data, we have, by (1.1), that

$$\begin{pmatrix} \alpha \\ y \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ X\beta \end{pmatrix}, \begin{pmatrix} G & GZ' \\ ZG & V \end{pmatrix} \right),$$

where $G = \text{Var}(\alpha)$, $R = \text{Var}(\epsilon)$, and $V = \text{Var}(y) = ZGZ' + R$. It follows that

$$E(\alpha|y) = GZ'V^{-1}(y - X\beta)$$

(see Appendix C). Therefore, by (2.31), the best predictor of ξ is

$$\tilde{\xi} = a'GZ'V^{-1}(y - X\beta).$$

Once the best predictor of $\xi = a'\alpha$ is obtained, the best predictor of $\eta = a'\alpha + b'\beta$ is

$$\hat{\eta}_B = b'\beta + a'GZ'V^{-1}(y - X\beta). \quad (2.32)$$

Here the subscript B refers to the best predictor.

It can be shown that, without assuming normality, (2.32) gives the best linear predictor of η in the sense that it minimizes the MSE of a predictor that is linear in y . See Searle et al. (1992, §7.3). The following example was given by Mood et al. (1974, pp. 370).

Example 2.9 (IQ tests). Suppose intelligence quotients for students in a particular age group are normally distributed with mean 100 and standard deviation 15. The IQ, say x_1 , of a particular student is to be estimated by a test on which he scores 130. It is further given that test scores are normally distributed about the true IQ as a mean with standard deviation 5. What is the best prediction on the student's IQ? (The answer is not 130.)

The solution may be found by applying the method of best prediction. Here we have $y = \mu + \alpha + \epsilon$, where y is the student's test score, which is 130; α is the realization of a random effect corresponding to the student, so that $\mu + \alpha$ is the student's true IQ, which is unobservable. The question is to predict $\mu + \alpha$, a mixed effect. It is known that $\text{IQ} \sim N(100, 15^2)$ and $\text{score}|\text{IQ} \sim N(\text{IQ}, 5^2)$. Also, $\mu = 100$ is given. It follows that $Z = 1$, $G = \text{var}(\text{IQ}) = 15^2$, $V = \text{var}(\text{score}) = \text{var}(E(\text{score}|\text{IQ})) + E(\text{var}(\text{score}|\text{IQ})) = \text{var}(\text{IQ}) + E(5^2) = 15^2 + 5^2$. Therefore, by (2.32), the best prediction of the student's IQ is

$$\widetilde{\text{IQ}} = \mu + \frac{15^2}{15^2 + 5^2}(\text{score} - \mu) = 127.$$

2. Best linear unbiased prediction. If the fixed effects are unknown but the variance components are known, Equation (2.32) is not a predictor. In such a case, it is customary to replace β by $\hat{\beta}$, its maximum likelihood estimator under normality, which is

$$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y. \quad (2.33)$$

Here, for simplicity, we assume that X is of full rank p . (2.33) is also known as the best linear unbiased estimator, or BLUE, whose derivation does not

require normality. Henderson (1973) showed that, after β in (2.32) is replaced by the BLUE (2.33), the resulting predictor is the best linear unbiased predictor of η in the sense that (i) it is linear in y , (ii) its expected value is equal to that of η , and (iii) it minimizes the MSE among all linear unbiased predictors, where the MSE of a predictor $\tilde{\eta}$ is defined as $\text{MSE}(\tilde{\eta}) = E\{(\tilde{\eta} - \eta)(\tilde{\eta} - \eta)'\}$. Again, the result does not require normality. Thus, the BLUP is given by

$$\hat{\eta}_{\text{BLUP}} = b'\tilde{\beta} + a'GZ'V^{-1}(y - X\tilde{\beta}), \quad (2.34)$$

where $\tilde{\beta}$ is the BLUE given by (2.33). The vector

$$\tilde{\alpha} = GZ'V^{-1}(y - X\tilde{\beta}) \quad (2.35)$$

is also called the BLUP of α .

Henderson's original derivation of BLUP was based on what he called "joint maximum likelihood estimates" of fixed and random effects. Consider a Gaussian mixed model (1.1), where $\alpha \sim N(0, G)$, $\epsilon \sim N(0, R)$, and α and ϵ are independent. Suppose that both G and R are nonsingular. Then, it can be shown that the logarithm of the joint pdf of α and y can be expressed as (Exercise 2.18)

$$c - \frac{1}{2} \{ (y - X\beta - Z\alpha)'R^{-1}(y - X\beta - Z\alpha) + \alpha'G^{-1}\alpha \}, \quad (2.36)$$

where c is a constant. Henderson (1950) proposed to find the "maximum likelihood estimates" of β and α , treating the latter as (fixed) parameters, by differentiating (2.36) with respect to β and α and setting the partial derivatives equal to zero. This leads to Henderson's mixed model equations:

$$\begin{pmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & G^{-1} + Z'R^{-1}Z \end{pmatrix} \begin{pmatrix} \tilde{\beta} \\ \tilde{\alpha} \end{pmatrix} = \begin{pmatrix} X'R^{-1} \\ Z'R^{-1} \end{pmatrix} y, \quad (2.37)$$

the solution to which leads to (2.33) and (2.35) (Exercise 2.18). Later, Henderson (1963) showed that the "maximum likelihood estimates" he derived earlier are indeed the BLUP. A more intuitive approach to show that the BLUP has minimum mean squared error within the class of linear unbiased estimators was given by Harville (1990). Also see Robinson (1991). In particular, this derivation does not require normality assumptions. In other words, the BLUP is well defined for non-Gaussian linear mixed models. The BLUP may also be regarded as the maximum likelihood estimator of the best predictor, because, assuming that the variance components are known, the BLUP may be obtained by replacing β in the expression of the best predictor (2.32) by its maximum likelihood estimator under normality, that is, (2.33). Finally, Jiang (1997b) showed that BLUP is the best predictor based on error contrasts; that is, (2.35) is identical to $E(\alpha|A'y)$, where A is any $n \times (n - p)$ matrix of full rank such that $A'X = 0$.

Robinson (1991) used the following example to illustrate the calculation of BLUE and BLUP.

Example 2.10. Consider a linear mixed model for the first lactation yields of dairy cows with sire additive genetic merits being treated as random effects and herd effects being treated as fixed effects. The herd effects are represented by β_j , $j = 1, 2, 3$ and sire effects by α_i , $i = 1, 2, 3, 4$, which correspond to sires A, B, C, D. The matrix R is taken to be the identity matrix, while the matrix G is assumed to be 0.1 times the identity matrix. This would be a reasonable assumption, provided that the sires were unrelated and that the variance ratio σ^2/τ^2 had been estimated previously, where $\sigma^2 = \text{var}(\alpha_i)$ and $\tau^2 = \text{var}(\epsilon_{ij})$. Suppose that the data are given below. It can be shown (Exercise 2.20) that

Herd	1	1	2	2	2	3	3	3	3
Sire	A	D	B	D	D	C	C	D	D
Yield	110	100	110	100	100	110	110	100	100

the mixed model equations (2.37) are

$$\begin{pmatrix} 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 0 & 1 & 0 & 2 \\ 0 & 0 & 4 & 0 & 0 & 2 & 2 \\ 1 & 0 & 0 & 11 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 11 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 12 & 0 \\ 1 & 2 & 2 & 0 & 0 & 0 & 15 \end{pmatrix} \begin{pmatrix} \tilde{\beta}_1 \\ \tilde{\beta}_2 \\ \tilde{\beta}_3 \\ \tilde{\alpha}_1 \\ \tilde{\alpha}_2 \\ \tilde{\alpha}_3 \\ \tilde{\alpha}_4 \end{pmatrix} = \begin{pmatrix} 210 \\ 310 \\ 420 \\ 110 \\ 110 \\ 220 \\ 500 \end{pmatrix},$$

which have the solution

$$\begin{aligned} \tilde{\beta} &= (105.64, 104.28, 105.46)', \\ \tilde{\alpha} &= (0.40, 0.52, 0.76, -1.67)'. \end{aligned}$$

3. Empirical BLUP. In practice, the fixed effects and variance components are typically unknown. Therefore, in most cases neither the best predictor nor the BLUP is computable, even though they are known to be best in their respective senses. In such cases, it is customary to replace the vector of variance components, θ , which is involved in the expression of BLUP by a consistent estimator, $\hat{\theta}$. The resulting predictor is often called empirical BLUP, or EBLUP.

Kackar and Harville (1981) showed that, if $\hat{\theta}$ is an even and translation-invariant estimator and the data are normal, the EBLUP remains unbiased. An estimator $\hat{\theta} = \hat{\theta}(y)$ is even if $\hat{\theta}(-y) = \hat{\theta}(y)$, and it is translation invariant if $\hat{\theta}(y - X\beta) = \hat{\theta}(y)$. Some of the well-known estimators of θ , including ANOVA, ML, and REML estimators (see Sections 1.3–1.5), are even and translation invariant. In their arguments, however, Kackar and Harville had assumed the existence of the expected value of EBLUP, which is not obvious because,

unlike BLUP, EBLUP is not linear in y . The existence of the expected value of EBLUP was proved by Jiang (1999b, 2000a). See Section 2.7 for details.

Harville (1991) considered the one-way random effects model of Example 1.1, and showed that, in this case, the EBLUP of the mixed effect, $\mu + \alpha_i$, is identical to a parametric empirical Bayes (PEB) estimator. In the meantime, Harville noted some differences between these two approaches, PEB and EBLUP. One of the differences is that much of the work on PEB has been carried out by professional statisticians and has been theoretical in nature. The work has tended to focus on relatively simple models, such as the one-way random effects model, because it is only these models that are tractable from a theoretical standpoint. On the other hand, much of the work on EBLUP has been carried out by practitioners such as researchers in the animal breeding area, and has been applied to relatively complex models.

One problem of practical interest is estimation of the MSE of EBLUP. Such a problem arises, for example, in small area estimation (e.g., Ghosh and Rao 1994). The EBLUP method has been used in small area estimation for estimating small area means, which are in the form of mixed effects. However, the MSE of EBLUP is complicated. A naive estimator of MSE of EBLUP may be obtained by replacing θ by $\hat{\theta}$ in the expression of the MSE of BLUP. However, this is an underestimation. To see this, let $\hat{\eta} = a'\hat{\alpha} + b'\hat{\beta}$ denote the EBLUP of a mixed effect $\eta = a'\alpha + b'\beta$, where $\hat{\alpha}$ and $\hat{\beta}$ are the BLUP of α , given by (2.35), and BLUE of β , given by (2.33), with the variance components θ replaced by $\hat{\theta}$. Kackar and Harville (1981) showed that, under normality assumptions, one has

$$\text{MSE}(\hat{\eta}) = \text{MSE}(\tilde{\eta}) + E(\hat{\eta} - \tilde{\eta})^2, \quad (2.38)$$

where $\tilde{\eta}$ is the BLUP of η given by (2.34). It is seen that the MSE of BLUP is only the first term on the right side of (2.38). In fact, it can be shown that $\text{MSE}(\tilde{\eta}) = g_1(\theta) + g_2(\theta)$, where

$$\begin{aligned} g_1(\theta) &= a'(G - GZ'V^{-1}ZG)a, \\ g_2(\theta) &= (b - X'V^{-1}ZGa)'(X'V^{-1}X)^{-1}(b - X'V^{-1}ZGa) \end{aligned}$$

(e.g., Henderson 1975). It is clear that, using $g_1(\hat{\theta}) + g_2(\hat{\theta})$ as an estimator would underestimate the MSE of $\hat{\eta}$, because it does not take into account the additional variation associated with the estimation of θ , represented by the second term on the right side of (2.38). Such a problem may become particularly important when, for example, large amounts of funds are involved. For example, over \$7 billion of funds are allocated annually based on EBLUP estimators of school-age children in poverty at the county and school district levels (National Research Council 2000).

Kackar and Harville (1984) gave an approximation to the MSE of EBLUP under the linear mixed model (1.1), taking account of the variability in $\hat{\theta}$, and proposed an estimator of $\text{MSE}(\hat{\eta})$ based on this approximation. But the

approximation is somewhat heuristic, and the accuracy of the approximation and the associated MSE estimator was not studied. Prasad and Rao (1990) studied the accuracy of a second-order approximation to $\text{MSE}(\hat{\eta})$ for two important special cases of longitudinal linear mixed models (see Section 1.2): (i) the Fay–Herriot model (Fay and Herriot 1979), and (ii) the nested error regression model (e.g., Battese et al. 1988). Both models are very popular in the context of small area estimation. Recently, Das et al. (2004) extended the result of Prasad and Rao to general linear mixed models (1.1). For example, for Gaussian mixed ANOVA models with REML estimation of θ , Das *et al.* (2004) showed that $\text{MSE}(\hat{\eta}) = g_1(\theta) + g_2(\theta) + g_3(\theta) + o(d_*^{-2})$, where

$$g_3(\theta) = \text{tr} \left[\{(\partial/\partial\theta')V^{-1}ZGa\}'V\{(\partial/\partial\theta')V^{-1}ZGa\}H^{-1} \right], \quad (2.39)$$

where $H = E(\partial^2 l_R / \partial\theta\partial\theta')$ and l_R is the restricted log-likelihood given by (1.17), and $d_* = \min_{1 \leq i \leq s} d_i$ with $d_i = \|Z_i' P Z_i\|_2$ and P given by (1.11). The same result also holds for ML estimation. Based on the approximation, the authors obtained an estimator of $\text{MSE}(\hat{\eta})$ whose bias is corrected to the second order. More specifically, an estimator $\widehat{\text{MSE}}(\hat{\eta})$ was obtained such that $E\{\widehat{\text{MSE}}(\hat{\eta})\} = \text{MSE}(\hat{\eta}) + o(d_*^{-2})$. See Das et al. (2004) for details.

Alternatively, Jiang et al. (2002) proposed a jackknife method that led to second-order approximation and estimation of the MSE of EBLUP in the case of longitudinal linear mixed models. Denote $\text{MSE}(\tilde{\eta})$ by $b(\theta)$, where $\tilde{\eta}$ is the BLUP given by (2.34). The jackknife estimator of the MSE of $\hat{\eta}$ is given by $\widehat{\text{MSE}}(\hat{\eta}) = \widehat{\text{MSAE}}(\hat{\eta}) + \widehat{\text{MSE}}(\tilde{\eta})$, where

$$\begin{aligned} \widehat{\text{MSAE}}(\hat{\theta}) &= \frac{m-1}{m} \sum_{i=1}^m (\hat{\eta}_{-i} - \hat{\eta})^2, \\ \widehat{\text{MSE}}(\tilde{\eta}) &= b(\hat{\theta}) - \frac{m-1}{m} \sum_{i=1}^m \left\{ b(\hat{\theta}_{-i}) - b(\hat{\theta}) \right\}. \end{aligned} \quad (2.40)$$

Here m represents the number of clusters (e.g., number of small areas), $\hat{\theta}_{-i}$ denotes an M-estimator of θ using the data without the i th cluster (e.g., the i th small area), and $\hat{\eta}_{-i}$ the EBLUP of η in which the fixed parameters are estimated using the data without the i th cluster. Jiang et al. (2002) showed that $E\{\widehat{\text{MSE}}(\hat{\eta})\} = \text{MSE}(\hat{\eta}) + o(m^{-1})$. The result holds, in particular, when $\hat{\theta}$ is either the REML or the ML estimator. Furthermore, the result holds for non-Gaussian (longitudinal) linear mixed models. In fact, the jackknife method also applies to longitudinal generalized linear mixed models, in which EBLUP is replaced by the empirical best predictor (EBP). See Jiang et al. (2002) for details.

Example 2.4 (Continued). Consider, once again, the James–Stein estimator of Example 2.4. Consider the prediction of the random effect $\eta = \alpha_1$. The BLUP is given by $\tilde{\eta} = (1 - \omega)y_1$, where $\omega = (1 + \psi)^{-1}$. The EBLUP is given by $\hat{\eta} = (1 - \hat{\omega})y_1$. Efron and Morris (1973) used the following unbiased

estimator, $\hat{\omega} = (m-2)/\sum_{i=1}^m y_i^2$. Note that the MSE of $\tilde{\eta}$ is given by $1 - \omega$. The jackknife estimator of the MSE of $\hat{\eta}$ is given by

$$\begin{aligned}\widehat{\text{MSE}} &= 1 - \hat{\omega} + \frac{m-1}{m} \sum_{i=1}^m (\hat{\eta}_{-i} - \hat{\eta})^2 \\ &= 1 - \hat{\omega} + y_1^2 \left(\frac{m-1}{m} \right) \sum_{i=1}^m (\hat{\omega}_{-i} - \hat{\omega})^2.\end{aligned}$$

Note that, because in this case $1 - \hat{\omega}$ is an unbiased estimator of $1 - \omega$, no bias correction is needed; that is, the second term on the right side of (2.40) is not needed.

Example 2.11 (The baseball example). Efron and Morris (1975) considered a Bayesian model to predict the true 1970 season batting average of each of 18 major league baseball players using the data on batting averages based on the first 45 official at-bats. Their model can be obtained as a simple linear mixed model by adding an unknown μ term to the previous example. The prediction of the true season batting average of player 1 is the same as that of the mixed effect: $\eta = \mu + \alpha_1$. The best predictor of η (see Section 2.3.1.1) is given by $\tilde{\eta} = \mu + (1 - \omega)(y_1 - \mu)$. The EBLUP is given by $\hat{\eta} = \bar{y} + (1 - \hat{\omega})(y_1 - \bar{y})$, where \bar{y} is the sample mean. As for $\hat{\omega}$, Morris (1983) suggested a different estimator:

$$\hat{\omega} = \min \left\{ \frac{m-3}{m-1}, \frac{m-3}{\sum_{i=1}^m (y_i - \bar{y})^2} \right\}.$$

It can be shown that the bias of $1 - \hat{\omega}$ for estimating $1 - \omega$, the MSE of $\tilde{\eta}$, is $o(m^{-1})$, thus, again, bias correction is not needed. It follows that the jackknife estimator of the MSE of $\hat{\eta}$ is

$$\widehat{\text{MSE}} = 1 - \hat{\omega} + \frac{m-1}{m} \sum_{i=1}^m (\hat{\eta}_{-i} - \hat{\eta})^2,$$

where $\hat{\eta}_{-i} = \bar{y}_{-i} + (1 - \hat{\omega}_{-i})(y_1 - \bar{y}_{-i})$, $\bar{y}_{-i} = (m-1)^{-1} \sum_{j \neq i} y_j$ and

$$\hat{\omega}_{-i} = \min \left\{ \frac{m-4}{m-2}, \frac{m-4}{\sum_{j \neq i} (y_j - \bar{y}_{-i})^2} \right\}.$$

We return to this example later in this chapter.

2.3.2 Prediction of Future Observation

We now consider the problem of predicting a future observation under a non-Gaussian linear mixed model. Because normality is not assumed, the approach is distribution-free; that is, it does not require any specific assumption about the distribution of the random effects and errors. First note that for this

type of prediction, it is reasonable to assume that a future observation is independent of the current ones. We offer some examples.

Example 2.12. In longitudinal studies, one may be interested in prediction, based on repeated measurements from the observed individuals, of a future observation from an individual not previously observed. It is of less interest to predict another observation from an observed individual, because longitudinal studies often aim at applications to a larger population (e.g., drugs going to the market after clinical trials).

Example 2.13. In surveys, responses may be collected in two steps: in the first step, a number of families are randomly selected; in the second step, some family members (e.g., all family members) are interviewed for each of the selected families. Again, one may be more interested in predicting what happens to a family not selected, because one already knows enough about selected families (especially when all family members in the selected families are interviewed).

Therefore, we assume that a future observation, y_* , is independent of the current ones. Then, we have $E(y_*|y) = E(y_*) = x_*^t\beta$, so the best predictor is $x_*^t\beta$, if β is known; otherwise, an empirical best predictor (EBP) is obtained by replacing β by an estimator. So the point prediction is fairly straightforward. A question that is often of practical interest but has been so far neglected, for the most part, is that of prediction intervals.

1. *Distribution-free prediction intervals.* A prediction interval for a single future observation is an interval that will, with a specified coverage probability, contain a future observation from a population. In model-based statistical inference, it is assumed that the future observation has a certain distribution. Sometimes, the distribution is specified up to a finite number of unknown parameters, for example, those of the normal distribution. Then, a prediction interval may be obtained, if the parameters are adequately estimated, and the uncertainty in the parameter estimations is suitably assessed. Clearly, such a procedure is dependent on the underlying distribution in that, if the distributional assumption fails, the prediction interval may be seriously off: it either is wider than necessary, or does not have the claimed coverage probability. An alternative to the parametric method is a distribution-free one, in which one does not assume that the form of the distribution is known.

The problem of prediction intervals is, of course, an old one. One of the earliest works in this field is that of Baker (1935). Patel (1989) provided a review of the literature on prediction intervals when the future observation is independent of the observed sample, including results based on parametric distributions and on distribution-free methods. Hahn and Meeker (1991) reviewed three types of statistical intervals that are used most frequently in practice: the confidence interval, the prediction interval, and the tolerance interval. For a more recent overview, and developments on nonparametric prediction intervals, see Zhou (1997). Although many results on prediction

intervals are for the i.i.d. case, the problem is also well studied in some non-i.i.d. cases, such as linear regression (e.g., Sen and Srivastava 1990, §3.8.2). In the context of linear mixed models, Jeske and Harville (1988) considered prediction intervals for mixed effects, assuming that the joint distribution of α and $y - E(y)$ is known up to a vector of unknown parameters. Thus, their approach is not distribution-free.

Note that, even if β is unknown, it is still fairly easy to obtain a prediction interval for y_* if one is willing to make the assumption that the distributions of the random effects and errors are known up to a vector of parameters (e.g., variance components). To see this, consider a simple case: $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, where the random effect α_i and error ϵ_{ij} are independent such that $\alpha_i \sim N(0, \sigma^2)$ and $\epsilon_{ij} \sim N(0, \tau^2)$. It follows that the distribution of y_{ij} is $N(x'_{ij}\beta, \sigma^2 + \tau^2)$. Because methods are well developed for estimating fixed parameters such as β , σ^2 , and τ^2 (see Section 1.3), a prediction interval with asymptotic coverage probability $1 - \rho$ is easy to obtain. However, it is much more difficult if one does not know the forms of the distributions of the random effects and errors, and this is the case that we consider. In the following, we propose a distribution-free approach to prediction intervals. Our results do not require normality or any specific distributional assumptions about the random effects and errors, and therefore are applicable to non-Gaussian linear mixed models.

First note that to consistently estimate the fixed effects and variance components in a linear mixed model, one does not need to assume that the random effects and errors are normally distributed (see Section 1.4). We categorize (non-Gaussian) linear mixed models into two classes: the standard and the nonstandard ones. A linear mixed model (1.1), (1.2) is standard if each Z_i consists only of 0s and 1s, there is exactly one 1 in each row and at least one 1 in each column. Our approaches are quite different for standard and nonstandard linear mixed models.

2. Standard linear mixed models. For standard linear mixed models, the method is surprisingly simple, and can be described as follows. First, one throws away the middle terms in (1.1) that involve the random effects, that is, (1.2), and pretends that it is a linear regression model with i.i.d. errors: $y = X\beta + \epsilon$. Next, one computes the least squares (LS) estimator $\hat{\beta} = (X'X)^{-1}X'y$ and the residuals $\hat{\epsilon} = y - X\hat{\beta}$. Let \hat{a} and \hat{b} be the $\rho/2$ and $1 - \rho/2$ quantiles of the residuals. Then, a prediction interval for y_* with asymptotic coverage probability $1 - \rho$ is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x'_*\hat{\beta}$. Note that, although the method sounds almost the same as the residual method in linear regression, its justification is not so obvious because, unlike linear regression, the observations in a (standard) linear mixed model are not independent. The method may be improved if one uses more efficient estimators such as the empirical BLUE (EBLUE; see Section 2.3) instead of the LS estimator. We study this in a simulated example in the sequel.

Let y_* be a future observation that we wish to predict. Suppose that y_* satisfies a standard linear mixed model. Then, y_* can be expressed as

$$y_* = x_*' \beta + \alpha_{*1} + \cdots + \alpha_{*s} + \epsilon_* ,$$

where x_* is a known vector of covariates (not necessarily present with the data), α_{*r} s are random effects, and ϵ_* is an error, such that $\alpha_{*i} \sim F_{ir}$, $\leq i \leq s$, $\epsilon_* \sim F_0$, where the F s are unknown distributions (not necessarily normal), and $\alpha_{*1}, \dots, \alpha_{*s}, \epsilon_*$ are independent. According to earlier discussion, we assume that y_* is independent of $y = (y_i)_{1 \leq i \leq n}$. It follows that the best (point) predictor of y_* , when β is known, is $E(y_*|y) = E(y_*) = x_*' \beta$. Because β is unknown, it is replaced by a consistent estimator, $\hat{\beta}$, which may be the OLS estimator or EBLUE (e.g., Jiang and Zhang 2002, Theorem 1; Jiang 1998b). This results in an empirical best predictor:

$$\hat{y}_* = x_*' \hat{\beta} . \quad (2.41)$$

Let $\hat{\delta}_i = y_i - x_i' \hat{\beta}$. Define

$$\hat{F}(x) = \frac{\#\{1 \leq i \leq n : \hat{\delta}_i \leq x\}}{n} = \frac{1}{n} \sum_{i=1}^n 1_{(\hat{\delta}_i \leq x)} . \quad (2.42)$$

Note that, although (2.42) resembles the empirical distribution, it is not one in the classic sense, because the $\hat{\delta}_i$ s are not independent (the y_i s are dependent, and $\hat{\beta}$ depends on all the data). Let $\hat{a} < \hat{b}$ be any numbers satisfying $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1 - \rho$ ($0 < \rho < 1$). Then, a prediction interval for y_* with asymptotic coverage probability $1 - \rho$ is given by

$$[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}] . \quad (2.43)$$

See Jiang and Zhang (2002). Note that a typical choice of \hat{a} , \hat{b} has $\hat{F}(\hat{a}) = \rho/2$ and $\hat{F}(\hat{b}) = 1 - \rho/2$. Another choice would be to select \hat{a} and \hat{b} to minimize $\hat{b} - \hat{a}$, the length of the prediction interval. Usually, \hat{a} , \hat{b} are selected such that the former is negative and the latter positive, so that \hat{y}_* is contained in the interval. Also note that, if one considers linear regression as a special case of the linear mixed model, in which the random effects do not appear, $\hat{\delta}_i$ is the same as $\hat{\epsilon}_i$, the residual, if $\hat{\beta}$ is the least squares estimator. In this case, \hat{F} is the empirical distribution of the residuals, and the prediction interval (2.43) corresponds to that obtained by the bootstrap method (Efron 1979). The difference is that our prediction interval (2.43) is obtained in closed form rather than by a Monte Carlo method. For more discussion on bootstrap prediction intervals, see Shao and Tu (1995, §7.3).

3. Nonstandard linear mixed models. Although most linear mixed models used in practice are standard, nonstandard linear mixed models are also used. First, the method developed for standard models may be applied to some of the nonstandard cases. To illustrate this, consider the following example.

Example 2.14. Suppose that the data are divided into two parts. For the first part, we have $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, m$, $j = 1, \dots, n_i$, where $\alpha_1, \dots, \alpha_m$ are i.i.d. random effects with mean 0 and distribution F_1 ; ϵ_{ij} s are i.i.d. errors with mean 0 and distribution F_0 , and the α s and ϵ s are independent. For the second part of the data, we have $y_k = x'_k\beta + \epsilon_k$, $k = N + 1, \dots, N + K$, where $N = \sum_{i=1}^m n_i$, and the ϵ_k s are i.i.d. errors with mean 0 and distribution F_0 . Note that the random effects only appear in the first part of the data (and hence there is no need to use a double index for the second part).

For the first part, let the distribution of $\delta_{ij} = y_{ij} - x'_{ij}\beta$ be $F (= F_0 * F_1)$. For the second part, let $\delta_k = y_k - x'_k\beta$. If β were known, the δ_{ij} s (δ_k s) would be sufficient statistics for F (F_0). Therefore it suffices to consider an estimator of F (F_0) based on the δ_{ij} s (δ_k s). Note that the prediction interval for any future observation is determined either by F or by F_0 , depending on to which part the observation corresponds. Now, because β is unknown, it is customary to replace it by $\hat{\beta}$. Thus, a prediction interval for y_* , a future observation corresponding to the first part, is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x'_*\hat{\beta}$, \hat{a} , \hat{b} are determined by $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1 - \rho$ with

$$\hat{F}(x) = \frac{1}{N} \#\{(i, j) : 1 \leq i \leq m, 1 \leq j \leq n_i, \hat{\delta}_{ij} \leq x\}$$

and $\hat{\delta}_{ij} = y_{ij} - x'_{ij}\hat{\beta}$. Similarly, a prediction interval for y_* , a future observation corresponding to the second part, is $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$, where $\hat{y}_* = x'_*\hat{\beta}$, \hat{a} , \hat{b} are determined similarly with \hat{F} replaced by

$$\hat{F}_0(x) = \frac{1}{K} \#\{k : N + 1 \leq k \leq N + K, \hat{\delta}_k \leq x\}$$

and $\hat{\delta}_k = y_k - x'_k\hat{\beta}$. The prediction interval has asymptotic coverage probability $1 - \rho$ (see Jiang and Zhang 2002).

If one looks more carefully, it is seen that the model in Example 2.14 can be divided into two standard submodels, so that the previous method is applied to each submodel. Of course, not every nonstandard linear mixed model can be divided into standard submodels. For such nonstandard models we consider that a different approach may need to be used.

Jiang (1998b) considered estimation of the distributions of the random effects and errors. His approach is the following. Consider the EBLUP of the random effects: $\hat{\alpha}_i = \hat{\sigma}_i^2 Z'_i \hat{V}^{-1}(y - X\hat{\beta})$, $1 \leq i \leq s$, where $\hat{\beta}$ is the EBLUE (see Section 2.2.1.4). The “EBLUP” for the errors can be defined as $\hat{\epsilon} = y - X\hat{\beta} - \sum_{i=1}^s Z_i \hat{\alpha}_i$. It was shown that, if the REML or ML estimators of the variance components are used, then, under suitable conditions,

$$\hat{F}_i(x) = \frac{1}{m_i} \sum_{u=1}^{m_i} 1_{(\hat{\alpha}_{i,u} \leq x)} \xrightarrow{P} F_i(x), \quad x \in C(F_i),$$

where $\hat{\alpha}_{i,u}$ is the u th component of $\hat{\alpha}_i$, $1 \leq i \leq s$, and

$$\hat{F}_0(x) = \frac{1}{n} \sum_{u=1}^n 1_{(\hat{\epsilon}_u \leq x)} \xrightarrow{P} F_0(x), \quad x \in C(F_0),$$

where $\hat{\epsilon}_u$ is the u th component of $\hat{\epsilon}$. Here $C(F_i)$ represents the set of all continuity points of F_i , $0 \leq i \leq s$ (see Jiang 1998b).

For simplicity, we assume that all the distributions F_0, \dots, F_s are continuous. Let y_* be a future observation we would like to predict. As before, we assume that y_* is independent of y and satisfies a mixed linear model, which can be expressed componentwise as

$$y_i = x'_i \beta + z'_{i1} \alpha_1 + \dots + z'_{is} \alpha_s + \epsilon_i, \quad i = 1, \dots, n.$$

This means that y_* can be expressed as

$$y_* = x'_* \beta + \sum_{j=1}^l w_j \gamma_j + \epsilon_*,$$

where x_* is a known vector of covariates (not necessarily present with the data), w_j s are known nonzero constants, γ_j s are unobservable random effects, and ϵ_* is an error. In addition, there is a partition of the indices $\{1, \dots, l\} = \cup_{k=1}^q I_k$, such that $\gamma_j \sim F_{r(k)}$ if $j \in I_k$, where $r(1), \dots, r(q)$ are distinct integers between 1 and s (so $q \leq s$); $\epsilon_* \sim F_0$; $\gamma_1, \dots, \gamma_l, \epsilon_*$ are independent. Define

$$\hat{F}^{(j)}(x) = m_{r(k)}^{-1} \sum_{u=1}^{m_{r(k)}} 1_{(w_j \hat{\alpha}_{r(k),u} \leq x)}, \quad \text{if } j \in I_k$$

for $1 \leq k \leq q$. Let

$$\begin{aligned} \hat{F}(x) &= (\hat{F}^{(1)} * \dots * \hat{F}^{(l)} * \hat{F}_0)(x) \\ &= \frac{\#\{(u_1, \dots, u_l, u) : \sum_{k=1}^q \sum_{j \in I_k} w_j \hat{\alpha}_{r(k),u_j} + \hat{\epsilon}_u \leq x\}}{\left(\prod_{k=1}^q m_{r(k)}^{|I_k|}\right) n}, \end{aligned} \quad (2.44)$$

where $*$ represents convolution (see Appendix C), and $1 \leq u_j \leq m_{r(k)}$ if $j \in I_k$, $1 \leq k \leq q$; $1 \leq u \leq n$. It can be shown that

$$\sup_x |\hat{F}(x) - F(x)| \xrightarrow{P} 0,$$

where $F = F^{(1)} * \dots * F^{(l)} * F_0$, and $F^{(j)}$ is the distribution of $w_j \gamma_j$, $1 \leq j \leq l$. Note that F is the distribution of $y_* - x'_* \beta$. Let \hat{y}_* be defined by (2.41) with $\hat{\beta}$ a consistent estimator, and \hat{a}, \hat{b} defined by $\hat{F}(\hat{b}) - \hat{F}(\hat{a}) = 1 - \rho$, where \hat{F} is given by (2.44). Then, the prediction interval $[\hat{y}_* + \hat{a}, \hat{y}_* + \hat{b}]$ has asymptotic coverage probability $1 - \rho$ (see Jiang and Zhang 2002).

We conclude this section with a simulated example.

4. *A simulated example.* Consider the linear mixed model

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij}, \quad (2.45)$$

$i = 1, \dots, m, j = 1, \dots, n_i$, where the α_i s are i.i.d. random effects with mean 0 and distribution F_1 , and ϵ_{ij} s are i.i.d. errors with mean 0 and distribution F_0 . The model might be associated with a sample survey, where α_i is a random effect related to the i th family in the sample, and n_i is the sample size for the family (e.g., the family size, if all family members are to be surveyed). The x_{ij} s are covariates associated with the individuals sampled from the family and, in this case, correspond to people's ages. The ages are categorized by the following groups: 0-4, 5-9, ..., 55-59, so that $x_{ij} = k$ if the person's age falls into the k th category (people whose ages are 60 or over are not included in the survey). The true parameters for β_0 and β_1 are 2.0 and 0.2, respectively.

In the following simulations, four combinations of the distributions F_0, F_1 are considered. These are Case I: $F_0 = F_1 = N(0, 1)$; Case II: $F_0 = F_1 = t_3$; Case III: $F_0 = \text{logistic}$ [the distribution of $\log\{U/(1-U)\}$, where $U \sim \text{Uniform}(0, 1)$], $F_1 = \text{centralized lognormal}$ [the distribution of $e^X - \sqrt{e}$, where $X \sim N(0, 1)$]; Case IV: $F_0 = \text{double exponential}$ [the distribution of $X_1 - X_2$, where X_1, X_2 are independent $\sim \text{exponential}(1)$], $F_1 = \text{a mixture of } N(-4, 1) \text{ and } N(4, 1) \text{ with equal probability}$. Note that Cases II–IV are related to the following types of departure from normality: heavy-tail, asymmetry, and bimodal. In each case, the following sample size configuration is considered: $m = 100, k_1 = \dots = k_{m/2} = 2$, and $k_{m/2+1} = \dots = k_m = 6$. Finally, for each of the above cases, three prediction intervals are considered. The first is the prediction interval based on the OLS estimator of β ; the second is that based on the EBLUE of β , where the variance components are estimated by REML (see Section 1.4.1); and the third is the linear regression (LR) prediction interval (e.g., Casella and Berger 2002, pp. 558), which assumes that the observations are independent and normally distributed. The third one is considered here for comparison.

For each of the four cases, 1000 datasets are generated. First, the following are independently generated, (i) $x_{ij}, 1 \leq i \leq m, 1 \leq j \leq k_i$, uniformly from the integers $1, \dots, 12$ (twelve age categories); (ii) $\alpha_i, 1 \leq i \leq m$, from F_1 ; (iii) $\epsilon_{ij}, 1 \leq i \leq m, 1 \leq j \leq k_i$, from F_0 . Then y_{ij} is obtained by (2.45) with β_0, β_1 being the true parameters. Because of the way that the data are generated, condition on the x_{ij} s, the y_{ij} s satisfy (2.45) with its distributional assumptions. For each dataset generated, and for each of the 12 age categories, three prediction intervals are obtained, where $\rho = .10$ (nominal level 90%): OLS, EBLUE, and LR; then one additional observation is generated, which corresponds to a future observation in that category. The percentages of coverage and average lengths of the intervals over the 1000 data sets are reported.

The results are given in Table 2.1, in which the letters O, E, and L stand for OLS, EBLUE, and LR, respectively. The numbers shown in the table are coverage probabilities based on the simulations, in terms of percentages, and average lengths of the prediction intervals. Note that for OLS and EBLUE the lengths of the prediction intervals do not depend on the covariates, whereas

for LR the length of the prediction interval depends on the covariate, but will be almost constant if the sample size is large. This, of course, follows from the definition of the prediction intervals, but there is also an intuitive interpretation. Consider, for example, the normal case. The distribution of a future observation y_* corresponding to a covariate x_* is $N(\beta_0 + \beta_1 x_*, \sigma^2)$, where $\sigma^2 = \text{var}(\alpha_i) + \text{var}(\epsilon_{ij})$ is a constant. So, if the β s were known the length of any prediction interval for y_* would not depend on x_* . If the β s are unknown but replaced by consistent estimators, then if the sample size were large, one would also expect the length of the prediction interval to be almost constant (not dependent on x_*). For such a reason, there is no need to exhibit the lengths of the prediction intervals for different categories, and we only give the averages over all categories.

It is seen that in the normal case there is not much difference among all three methods. This is not surprising. The difference appears in the nonnormal cases. First, the LR prediction intervals are wider than the OLS and EBLUE ones. Second, as a consequence, the coverage probabilities for the LR prediction intervals seem to be higher than 90%. Overall, the OLS and EBLUE perform better than LR in the nonnormal cases. This is not surprising, because the OLS and EBLUE prediction intervals are distribution-free. The EBLUE does not seem to do better than the OLS. This was a bit unexpected. On the other hand, it shows that at least in this special case the OLS, although much simpler than the EBLUE in that one does not need to estimate the variance components, can do just as well as more sophisticated methods such as the EBLUE.

Table 2.1. Coverage probability and average length

Coverage Probability (%)												
x	Case I			Case II			Case III			Case IV		
	O	E	L	O	E	L	O	E	L	O	E	L
1	90	90	90	89	89	92	90	91	93	90	90	94
2	90	90	90	89	89	91	91	91	93	89	90	96
3	88	88	88	91	91	93	90	89	92	88	89	96
4	90	90	89	91	91	93	89	89	91	89	89	97
5	89	89	89	89	89	92	90	90	92	90	90	96
6	89	89	90	89	89	92	91	91	93	90	90	97
7	89	88	89	90	90	92	90	90	93	88	89	96
8	90	90	90	90	90	92	89	89	91	90	90	97
9	90	90	91	89	89	92	89	89	91	89	89	96
10	89	89	90	91	90	93	89	89	93	88	88	95
11	90	90	90	89	89	93	89	89	92	89	89	97
12	89	89	89	89	89	92	91	91	93	89	89	96
Average Length												
	4.6	4.6	4.7	7.0	7.0	7.9	8.1	8.1	9.0	12.1	12.1	14.3

2.4 Model Checking and Selection

The previous sections have been dealing with inference about linear mixed models. For the most part, we have assumed that the basic assumptions about the model, for example, those about the presence of the random effects and their distributions, are correct. In practice, however, these assumptions may also be subject to checking. Methods of model checking are also known as model diagnostics. Sometimes, it is not clear which is the best model to use when there are a number of potential, or candidate, models. Here being best is in the sense that the model is not only correct but also most economical, meaning that it is simplest among all correct models. In this section we deal with the problems of model diagnostics and selection.

2.4.1 Model Diagnostics

Unlike standard regression diagnostics, the literature on diagnostics of linear mixed models involving random effects is not extensive (e.g., Ghosh and Rao 1994, pp. 70–71, Verbeke and Molenberghs 2000, pp. 151–152). Limited methodology is available, mostly regarding assessing the distribution of the random effects and errors. For the most part, the methods may be classified as diagnostic plots and goodness-of-fit tests.

1. Diagnostic plots. Several authors have used the idea of EBLUP or empirical Bayes estimators (EB), discussed in the previous section, for diagnosing distributional assumptions regarding the random effects (e.g., Dempster and Ryan 1985; Calvin and Sedransk 1991). The approach is reasonable because the EBLUP or EB are natural estimators of the random effects. In the following we describe a method proposed by Lange and Ryan (1989) based on a similar idea.

One commonly used assumption regarding the random effects and errors is that they are normally distributed. If such an assumption holds, one has a case of Gaussian mixed models. Otherwise, one is dealing with non-Gaussian linear mixed models. Lange and Ryan considered the longitudinal model (see Section 1.2.1.2), assuming that $G_i = G$, $R_i = \tau^2 I_{k_i}$, $i = 1, \dots, m$, and developed a weighted normal plot for assessing normality of the random effects in a longitudinal model. First, under the model (1.3) and normality, one can derive the best predictors, or Bayes estimators, of the random effects α_i $i = 1, \dots, m$ (see Section 2.3.1.1 and Section 2.5.1.1), assuming that β and θ , the vector of variance components, are known. This is given by

$$\begin{aligned}\tilde{\alpha}_i &= E(\alpha_i | y_i) \\ &= GZ_i'V_i^{-1}(y_i - X_i\beta),\end{aligned}$$

where $V_i = \text{Var}(y_i) = \tau^2 I_{k_i} + Z_i G Z_i'$. Furthermore, the covariance matrix of $\tilde{\alpha}_i$ is given by

$$\text{Var}(\tilde{\alpha}_i) = GZ_i'V_i^{-1}Z_iG.$$

Lange and Ryan proposed to examine a Q-Q plot of some standardized linear combinations

$$z_i = \frac{c'\tilde{\alpha}_i}{\{c'\text{Var}(\tilde{\alpha}_i)c\}^{1/2}}, \quad i = 1, \dots, m, \quad (2.46)$$

where c is a known vector. They argued that, through appropriate choices of c , the plot can be made sensitive to different types of model departures. For example, for a model with two random effects factors, a random intercept and a random slope, one may choose $c_1 = (1, 0)'$ and $c_2 = (0, 1)'$ and produce two Q-Q plots. On the other hand, such plots may not reveal possible nonzero correlations between the (random) slope and intercept. Thus, Lange and Ryan suggested producing a set of plots ranging from one marginal to the other by letting $c = (1 - u, u)'$ for some moderate number of values $0 \leq u \leq 1$.

Dempster and Ryan (1985) suggested that the normal plot should be weighted to reflect the differing sampling variances of $\tilde{\alpha}_i$. Following the same idea, Lange and Ryan proposed a generalized weighted normal plot. They suggested plotting z_i against $\Phi^{-1}\{F^*(z_i)\}$, where F^* is the weighted empirical cdf defined by

$$F^*(x) = \frac{\sum_{i=1}^m w_i 1_{(z_i \leq x)}}{\sum_{i=1}^m w_i},$$

and $w_i = c'\text{Var}(\tilde{\alpha}_i)c = c'GZ_i'V_i^{-1}Z_iGc$.

In practice, however, the fixed effects β and variance components θ are unknown. In such cases, Lange and Ryan suggested using the ML or REML estimators in place of these parameters. They argued that, under suitable conditions, the limiting distribution of $\sqrt{n}\{\hat{F}^*(x) - \Phi(x)\}$ is normal with mean zero and variance equal to the variance of $\sqrt{n}\{F^*(x) - \Phi(x)\}$ minus an adjustment, where $\hat{F}^*(x)$ is $F^*(x)$ with the unknown parameters replaced by their ML (REML) estimators. See Lange and Ryan (1989) for details. This suggests that, in the case of unknown parameters, the Q-Q plot will be \hat{z}_i against $\Phi^{-1}\{\hat{F}^*(\hat{z}_i)\}$, where \hat{z}_i is z_i with the unknown parameters replaced by their ML (REML) estimates. However, the (asymptotic) variance of $\hat{F}^*(x)$ is different from that of $F^*(x)$, as indicated above. Therefore, if one wishes to include, for example, a ± 1 SD bound in the plot, the adjustment for estimation of parameters must be taken into account. See Lange and Ryan (1989). We consider an example.

Example 2.3 (Continued). Consider, again, the one-way random effects model of Example 1.1 with normality assumption. Because α_i is real-valued, $c = 1$ in (2.31). If μ , σ^2 , τ^2 are known, the EB estimator of α_i is given by

$$\hat{\alpha}_i = \frac{k_i \sigma^2}{\tau^2 + k_i \sigma^2}(\bar{y}_{i\cdot} - \mu),$$

where $\bar{y}_{i\cdot} = k_i^{-1} \sum_{j=1}^{k_i} y_{ij}$, with

$$w_i = \text{var}(\hat{\alpha}_i) = \frac{k_i \sigma^4}{\tau^2 + k_i \sigma^2}.$$

Therefore, in this case,

$$z_i = \frac{\hat{\alpha}_i}{\text{sd}(\hat{\alpha}_i)} = \frac{\bar{y}_{i\cdot} - \mu}{\sqrt{\sigma^2 + \tau^2/k_i}},$$

$i = 1, \dots, m$ and

$$F^*(x) = \left(\sum_{i=1}^m \frac{k_i \sigma^4}{\tau^2 + k_i \sigma^2} \right)^{-1} \sum_{i=1}^m \frac{k_i \sigma^4}{\tau^2 + k_i \sigma^2} 1_{(z_i \leq x)}.$$

In practice, μ , σ^2 , and τ^2 are unknown and therefore replaced by their REML (ML) estimators when making a Q-Q plot (Exercise 2.20).

2. Goodness-of-fit tests. Recently, several authors have developed tests for checking distributional assumptions involved in linear mixed models. Consider a mixed ANOVA model (1.1), where for $1 \leq i \leq s$, $\alpha_i = (\alpha_{ij})_{1 \leq j \leq m_i}$, where the α_{ij} s are i.i.d. with mean 0, variance σ_i^2 , which is unknown, and continuous distribution $F_i = F_i(\cdot | \sigma_i)$; and $\epsilon = (\epsilon_j)_{1 \leq j \leq N}$, where the ϵ_j s are i.i.d. with mean 0, variance τ^2 , which is unknown, and continuous distribution $G = G(\cdot | \tau)$; and $\alpha_1, \dots, \alpha_s, \epsilon$ are independent. We are interested in testing the following hypothesis,

$$\begin{aligned} H_0 : F_i(\cdot | \sigma_i) &= F_{0i}(\cdot | \sigma_i), \quad 1 \leq i \leq s, \\ \text{and } G(\cdot | \tau) &= G_0(\cdot | \tau); \end{aligned} \quad (2.47)$$

that is, the distributions of the random effects and errors, up to a set of unknown variance components $\sigma_1^2, \dots, \sigma_s^2, \tau^2$, are as assumed.

Such distributional assumptions are vital in many applications of linear mixed models, and this is true even in large sample situations. For example, in many cases the prediction of a mixed effect is of main interest. Consider, for example, a nested error regression model, a special case of linear mixed models, which is useful in small area estimation (e.g., Battese et al. 1988; Prasad and Rao 1990; Ghosh and Rao 1994; Arora, Lahiri, and Mukherjee 1997):

$$y_{ij} = x'_{ij} \beta + \alpha_i + \epsilon_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, k_i, \quad (2.48)$$

where x_{ij} is a known vector of covariates, β is an unknown vector of regression coefficients, α_i is a random effect associated with the i th small area, and ϵ_{ij} is an error. A mixed effect may be in the form $\eta = x' \beta + \alpha_i$, where x is known. If the sample size is large (i.e., m is large), one may consistently estimate β and even obtain an asymptotic confidence interval for it, and this

does not rely on distributional assumptions such as normality. However, large sample results may not help, for example, in obtaining a prediction interval for η , because the effective sample size for estimating α_i is k_i , the sample size for the i th small area, which is often very small. Therefore, unless one knows the form of the distribution of α_i (e.g., normal), an accurate prediction interval for η cannot be obtained no matter how large m is (provided that k_i is small). To see another example, consider the estimation of the MSE of the EBLUP. Prasad and Rao (1990) give approximation formulas for MSE of EBLUP in the context of small area estimation, which are correct to the order $o(m^{-1})$. Although their results are asymptotic, assuming that m is large, normality distributional assumption remains critical for the validity of their approximations.

Jiang, Lahiri, and Wu (2001) developed an asymptotic theory of Pearson's χ^2 -test with estimated cell frequencies, and applied the method to the case of nested error regression model (2.48) for checking the distributions of α and ϵ . The procedure requires splitting the data into two parts, one used for estimation and the other for testing, and thus raised some concerns about the power of the test. Jiang (2001) developed a method that applies to a general mixed ANOVA model as described above (2.47), which does not require data splitting. The method is described below.

The procedure is similar to Pearson's χ^2 -test with estimated cell probabilities (e.g., Moore 1978). Let E_1, \dots, E_M be a partition of R , the real line. Let a_n be a sequence of normalizing constants that is determined later on. Define

$$\hat{\chi}^2 = \frac{1}{a_n} \sum_{j=1}^M \{N_j - E_{\hat{\theta}}(N_j)\}^2, \quad (2.49)$$

where $N_j = \sum_{i=1}^n 1_{(y_i \in E_j)} = \#\{1 \leq i \leq n : y_i \in E_j\}$, and $\hat{\theta}$ is the REML estimator of the vector of parameters involved in the linear mixed model. Despite the similarity of (2.49) to Pearson's χ^2 -statistic, there are several major differences. First and most important, the observed count N_k is not a sum of independent random variables. In Pearson's χ^2 -test, one deals with i.i.d. observations, so that N_k is a sum of i.i.d. random variables, and hence the asymptotic result follows from the classic central limit theorem (CLT). In a mixed linear model, however, the observations are correlated. Therefore, the classic CLT cannot handle the asymptotics. Second, unlike Pearson's χ^2 -statistic, the normalizing constant in (2.49) is the same for all the squares in the sum. The choice of the normalizing constants in Pearson's χ^2 -test is such that the asymptotic distribution is χ^2 . However, even in the i.i.d. case, the asymptotic distribution of Pearson's χ^2 -statistic is not necessarily χ^2 , if the parameters are to be estimated (see Moore 1978). In fact, it may never be χ^2 no matter what normalizing constants are used. Thus, for simplicity, we choose a unified normalizing constant a_n . Note that, because of the dependence among the observations, a_n may not increase at the same rate as n , the sample size. Third, in a linear mixed model the number of fixed effects

may be allowed to increase with n (e.g., Jiang 1996). As a consequence, the dimension of θ may increase with n . This shows, from another angle, that one can no longer expect an asymptotic distribution such as χ^2_{M-q-1} , where q is the number of (independent) parameters being estimated.

Jiang (2001) showed that, under suitable conditions, the asymptotic distribution of $\hat{\chi}^2$ is a weighted χ^2 , that is, the distribution of $\sum_{j=1}^M \lambda_j Z_j^2$, where Z_1, \dots, Z_M are independent $N(0, 1)$ random variables, and $\lambda_1 \geq \dots \geq \lambda_M$ are eigenvalues of some nonnegative definite matrix, which depends on θ . Because the latter is unknown in practice, Jiang (2001) developed a method of estimating the critical value of the asymptotic distribution, and showed that $P(\hat{\chi}^2 > \hat{c}_\rho) \rightarrow \rho$ as $n \rightarrow \infty$, where $\rho \in (0, 1)$ is the level of the test. The estimated critical value, \hat{c}_ρ is determined as $c_\rho(\hat{\lambda}_1, \dots, \hat{\lambda}_M)$, where for any given $\lambda_1 \geq \dots \geq \lambda_M$ and $0 < \rho < 1$, $c_\rho(\lambda_1, \dots, \lambda_M)$ is the ρ -critical value of the random variable $\xi = \sum_{j=1}^M \lambda_j Z_j^2$, and $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_M$ are the eigenvalues of a matrix $\hat{\Sigma}_n = \Sigma_n(\hat{\theta})$. The definition of $\Sigma_n(\theta)$, which depends on θ , is given in Section 2.7.

It remains to specify the normalizing constant a_n . Jiang (2001) noted that the choice of a_n is not unique. However, in some special cases there are natural choices. For example, in the case of linear regression, which may be regarded as a special case of the linear mixed model [with $s = 0$ in (1.1)], one has $a_n = n$. In the case of the one-way random effects model of Example 1.1, if the k_i s are bounded, one has $a_n = m$. The choice is less obvious in the case of multiple random effects factors [i.e., $s > 1$ in (1.1)]. Jiang (2001) proposed the following principle that in many cases either uniquely determines a_n or at least narrows the choices. Note that there are a number of integers that contribute to the total sample size n , for example, m , k in Example 2.2; a , b , c in Example 2.1. Usually, a_n is a function of these integers. It is required that a_n depend on these integers in a way as simple as possible. In particular, no unnecessary constant is allowed in the expression of a_n . This is called a *natural choice* of a_n . A natural choice of a_n can be found by examining the leading term in the expression of the matrix $H_n + \Delta_n$ defined in Section 2.7. The following are some special cases.

Example 2.2 (Continued). In the case of the balanced one-way random effects model, it can be shown that $H_n + \Delta_n = mk^2\{\text{Var}(h_1) + o(1)\}$, where h_1 is some nondegenerate random vector (see Jiang 2001, Section 3). Thus, in this case, a natural choice is $a_n = mk^2$. If, in fact, k is bounded, a natural choice would be $a_n = m$.

Example 2.1 (Continued). Suppose, for simplicity, that $c = 1$; that is, there is a single observation per cell. Similarly, it can be shown that, in this case, a natural choice is $a_n = (ab)^{3/2}$ (see Jiang 2001, Example 4.1).

2.4.2 Model Selection

In a way, model selection and estimation are viewed as two components of a process called model identification. The former determines the form of the model, leaving only some undetermined coefficients or parameters. The latter finds estimators of the unknown parameters. A pioneering work on model selection criteria was Akaike's information criterion (AIC, Akaike 1972). One of the earlier applications of AIC and other procedures such as the Bayesian information criterion (BIC, Schwartz 1978) was determination of the orders of an autoregressive moving-average time series model (e.g., Choi 1992). Similar methods have also been applied to regression model selection (e.g., Rao and Wu 1989; Bickel and Zhang 1992; Shao 1993; and Zheng and Loh 1995). It was shown that most of these model selection procedures are asymptotically equivalent to what is called the generalized information criterion (GIC, e.g., Nishii 1984). Although there is extensive literature on parameter estimation in linear mixed models, so that one component of the model identification has been well studied, the other component, that is, mixed model selection, has received little attention. Only recently have some results emerged in a paper by Jiang and Rao (2003).

Consider a general linear mixed model (1.1), where it is assumed that $E(\alpha) = 0$, $\text{Var}(\alpha) = G$; $E(\epsilon) = 0$, $\text{Var}(\epsilon) = R$, where G and R may involve some unknown parameters such as variance components; and α and ϵ are uncorrelated. In the following we first consider the problem of mixed model selection when the random effect factors are not subject to selection.

1. Selection with fixed random factors. Consider the model selection problem when the random part of the model (i.e., $Z\alpha$) is not subject to selection. Let $\zeta = Z\alpha + \epsilon$. Then, the problem is closely related to a regression model selection problem with correlated errors. Consider a general linear model $y = X\beta + \zeta$, where ζ is a vector of correlated errors, and everything else is as above. We assume that there are a number of candidate vectors of covariates, X_1, \dots, X_l , from which the columns of X are to be selected. Let $L = \{1, \dots, l\}$. Then, the set of all possible models can be expressed as $\mathcal{B} = \{a : a \subseteq L\}$, and there are 2^l possible models. Let \mathcal{A} be a subset of \mathcal{B} that is known to contain the true model, so the selection will be within \mathcal{A} . In an extreme case, \mathcal{A} may be \mathcal{B} itself. For any matrix M , let $\mathcal{L}(M)$ be the linear space spanned by the columns of M ; P_M the projection onto $\mathcal{L}(M)$: $P_M = M(M'M)^{-1}M'$; and P_M^\perp the orthogonal projection: $P_M^\perp = I - P_M$ (see Appendix B). For any $a \in \mathcal{B}$, let $X(a)$ be the matrix whose columns are X_j , $j \in a$, if $a \neq \emptyset$, and $X(a) = 0$ if $a = \emptyset$. Consider the following criterion for model selection,

$$\begin{aligned} C_n(a) &= |y - X(a)\hat{\beta}(a)|^2 + \lambda_n |a| \\ &= |P_{X(a)}^\perp y|^2 + \lambda_n |a|, \end{aligned} \tag{2.50}$$

$a \in \mathcal{A}$, where $|a|$ represents the cardinality of a ; $\hat{\beta}(a)$ is the ordinary least squares (OLS) estimator of $\beta(a)$ for the model $y = X(a)\beta(a) + \zeta$; that is,

$\hat{\beta}(a) = \{X(a)'X(a)\}^{-1}X(a)'y$, and λ_n is a positive number satisfying certain conditions specified below. Note that $P_{X(a)}$ is understood as 0 if $a = \emptyset$. Denote the true model by a_0 . If $a_0 \neq \emptyset$, we denote the corresponding X and β by X and $\beta = (\beta_j)_{1 \leq j \leq p}$ ($p = |a_0|$), and assume that $\beta_j \neq 0$, $1 \leq j \leq p$. This is, of course, reasonable because, otherwise, the model can be further simplified. If $a_0 = \emptyset$, X , β , and p are understood as 0. Let $\nu_n = \max_{1 \leq j \leq q} |X_j|^2$ and $\rho_n = \lambda_{\max}(ZGZ') + \lambda_{\max}(R)$, where λ_{\max} means the largest eigenvalue. Let \hat{a} be the minimizer of (2.50) over $a \in \mathcal{A}$, which is our selection of the model. Jiang and Rao (2003) showed that, under suitable conditions, \hat{a} is consistent in the sense that $P(\hat{a} \neq a_0) \rightarrow 0$ as $n \rightarrow \infty$, provided that

$$\lambda_n/\nu_n \rightarrow 0 \quad \text{and} \quad \rho_n/\lambda_n \rightarrow 0. \quad (2.51)$$

Note 1. If $\rho_n/\nu_n \rightarrow 0$, there always exists λ_n that satisfies (2.51). For example, take $\lambda_n = \sqrt{\rho_n \nu_n}$. However, this may not be the best choice of λ_n , as a simulated example in the following shows.

Note 2. Typically, we have $\nu_n \sim n$. To see what the order of ρ_n may turn out to be, consider a special but important case of linear mixed models: the mixed ANOVA model of (1.1) and (1.2). Furthermore, assume that each Z_i ($1 \leq i \leq s$) is a standard design matrix in the sense that it consists only of 0s and 1s, there is exactly one 1 in each row, and at least one 1 in each column. Let n_{ij} be the number of 1s in the j th column of Z_i . Note that n_{ij} is the number of appearance of the j th component of α_i . Also note that $Z_i'Z_i = \text{diag}(n_{ij}, 1 \leq j \leq m_i)$. Thus, we have $\lambda_{\max}(ZGZ') \leq \sum_{i=1}^s \sigma_i^2 \lambda_{\max}(Z_i Z_i') = \sum_{i=1}^s \sigma_i^2 \max_{1 \leq j \leq m_i} n_{ij}$. Also, we have $\lambda_{\max}(R) = \sigma_0^2$. It follows that $\rho_n = O(\max_{1 \leq i \leq s} \max_{1 \leq j \leq m_i} n_{ij})$. Therefore, (2.51) is satisfied provided that $\lambda_n/n \rightarrow 0$ and $\max_{1 \leq i \leq s} \max_{1 \leq j \leq m_i} n_{ij}/\lambda_n \rightarrow 0$. The following is an example not covered by the above case, because the errors are correlated.

Example 2.15. Consider the following linear mixed model which is a special case of the nested error regression model of (2.48); $y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, m$, $j = 1, \dots, k$, where β_0, β_1 are unknown coefficients (the fixed effects). It is assumed that the random effects $\alpha_1, \dots, \alpha_m$ are uncorrelated with mean 0 and variance σ^2 . Furthermore, assume that the errors ϵ_{ij} s have the following exchangeable correlation structure: Let $\epsilon_i = (\epsilon_{ij})_{1 \leq j \leq k}$. Then, $\text{Cov}(\epsilon_i, \epsilon_{i'}) = 0$ if $i \neq i'$, and $\text{Var}(\epsilon_i) = \tau^2\{(1-\rho)I + \rho J\}$, where I is the identity matrix and J the matrix of 1s, and $0 < \rho < 1$ is an unknown correlation coefficient. Finally, assume that the random effects are uncorrelated with the errors. Suppose that $m \rightarrow \infty$, and

$$0 < \liminf \left[\frac{1}{mk} \sum_{i=1}^m \sum_{j=1}^k (x_{ij} - \bar{x}_{..})^2 \right]$$

$$\leq \limsup \left[\frac{1}{mk} \sum_{i=1}^m \sum_{j=1}^k x_{ij}^2 \right] < \infty,$$

where $\bar{x}_{..} = (mk)^{-1} \sum_{i=1}^m \sum_{j=1}^k x_{ij}$. It is easy to see that, in this case, $\rho_n \sim k$ and $\nu_n \sim mk$ (Exercise 2.21).

The above procedure requires selecting \hat{a} from all subsets of \mathcal{A} . Note that \mathcal{A} may contain as many as 2^l subsets. When l is relatively large, alternative procedures have been proposed in the (fixed effect) linear model context, which require less computation (e.g., Zheng and Loh 1995). In the following, we consider an approach similar to Rao and Wu (1989). First, note that one can always express $X\beta$ as $X\beta = \sum_{j=1}^l \beta_j X_j$ with the understanding that some of the coefficients β_j may be zero. It follows that $a_0 = \{1 \leq j \leq l : \beta_j \neq 0\}$. Let $X_{-j} = (X_u)_{1 \leq u \leq l, u \neq j}$, $1 \leq j \leq l$, $\eta_n = \min_{1 \leq j \leq l} |P_{X_{-j}}^\perp X_j|^2$, and δ_n be a sequence of positive numbers satisfying conditions specified below. Let \hat{a} be the subset of $L = \{1, \dots, l\}$ such that

$$(|P_{X_{-j}}^\perp y|^2 - |P_X^\perp y|^2) / (|P_{X_{-j}}^\perp X_j|^2 \delta_n) > 1 \quad (2.52)$$

for $j \in \hat{a}$. Jiang and Rao (2003) showed that, if $\rho_n/\eta_n \rightarrow 0$, where ρ_n is defined earlier, then \hat{a} is consistent, provided that

$$\delta_n \rightarrow 0 \quad \text{and} \quad \rho_n/(\eta_n \delta_n) \rightarrow 0.$$

Example 2.15 (Continued). It is easy to show that, in this case, $\eta_n \sim mk$. Recall that $\rho_n \sim k$ in this case. Thus, $\rho_n/\eta_n \rightarrow 0$ as $m \rightarrow \infty$.

To study the finite sample behavior of the proposed model selection procedures, we consider a simulated example.

Example 2.16 (A simulated example). The model here is similar to Example 2.15 except that it may involve more than one fixed covariate; that is, $\beta_0 + \beta_1 x_{ij}$ is replaced by $x'_{ij} \beta$, where x_{ij} is a vector of covariates and β a vector of unknown regression coefficients. Here we focus on the first model selection procedure, the one defined by (2.50), which we also call GIC (e.g., Nishii 1984). We examine it by simulating the probability of correct selection and also the overfitting (a1) and underfitting (a2) probabilities, respectively, of various GICs for some given model parameters and sample sizes. Five GICs with different choices of λ are considered: (1) $\lambda = 2$, which corresponds to the C_p method; (2) $\lambda = \log n$. The latter choice satisfies the conditions required for consistency of the model selection. A total of 500 realizations of each simulation were run.

In the simulation the number of fixed factors was $l = 5$ with \mathcal{A} being all subsets of $\{1, \dots, 5\}$. The first column of X is all ones, corresponding to the intercept, and the other four columns of X are generated randomly from $N(0, 1)$ distributions, then fixed throughout the simulation. Three β s are considered: $(2, 0, 0, 4, 0)'$, $(2, 0, 0, 4, 8)'$, and $(2, 9, 0, 4, 8)'$, which correspond to $a_0 = \{1, 4\}$, $\{1, 4, 5\}$, and $\{1, 2, 4, 5\}$, respectively.

Furthermore, we consider the case where the correlated errors have varying degrees of exchangeable structure as described in Example 2.15, where four values of ρ were considered: 0, 0.2, 0.5, 0.8. Variance components σ and τ were both taken to be equal to 1. We take the number of clusters (m) to be 50 and 100 and the number of repeats on a cluster to be fixed at $k = 5$. Table 2.2 presents the results.

Table 2.2. Selection probabilities under Example 1.10

Model	ρ	% correct		$a1$		$a2$	
		$\lambda_n = 2 \log(n)$	$2 \log(N)$	$2 \log(N)$	$2 \log(N)$	$2 \log(N)$	$2 \log(N)$
$M1(m = 50)$	0	59	94	41	6	0	0
	.2	64	95	36	5	0	0
	.5	59	90	40	9	1	1
	.8	52	93	47	5	1	2
$M1(m = 100)$	0	64	97	36	3	0	0
	.2	57	94	43	6	0	0
	.5	58	96	42	3	0	1
	.8	61	96	39	4	0	0
$M2(m = 50)$	0	76	97	24	3	0	0
	.2	76	97	24	3	0	0
	.5	73	96	27	4	0	0
	.8	68	94	31	4	1	2
$M2(m = 100)$	0	76	99	24	1	0	0
	.2	70	97	30	3	0	0
	.5	70	98	30	2	0	0
	.8	72	98	28	2	0	0
$M3(m = 50)$	0	90	99	10	1	0	0
	.2	87	98	13	2	0	0
	.5	84	98	16	2	0	0
	.8	78	95	21	3	1	2
$M3(m = 100)$	0	87	99	13	1	0	0
	.2	87	99	13	1	0	0
	.5	80	99	20	1	0	0
	.8	78	96	21	3	1	1

2. *Selection with random factors.* We now consider model selection that involves both fixed and random effects factors. Here we consider the mixed ANOVA model of (1.1), (1.2). If $\sigma_i^2 > 0$, we say that α_i is in the model; otherwise, it is not. Therefore, the selection of random factors is equivalent to simultaneously determining which of the variance components $\sigma_1^2, \dots, \sigma_s^2$ are positive. The true model can be expressed as

$$y = X\beta + \sum_{i \in b_0} Z_i \alpha_i + \epsilon, \quad (2.53)$$

where $X = (X_j)_{j \in a_0}$ and $a_0 \subseteq L$ [defined above (2.50)]; $b_0 \subseteq S = \{1, \dots, s\}$ such that $\sigma_i^2 > 0$, $i \in b_0$, and $\sigma_i^2 = 0$, $i \in S \setminus b_0$.

There are some differences between selecting the fixed covariates X_j , as we did earlier, and selecting the random effect factors. One difference is that, in selecting the random factors, we are going to determine whether the vector α_i , not a given component of α_i , should be in the model. In other words, the components of α_i are all “in” or all “out”. Another difference is that, unlike selecting the fixed covariates, where it is reasonable to assume that the X_i are linearly independent, in a linear mixed model it is possible to have $i \neq i'$ but $\mathcal{L}(Z_i) \subset \mathcal{L}(Z_{i'})$. See Example 2.17 below. Because of these features, the selection of random factors cannot be handled the same way.

To describe the basic idea, first note that we already have a procedure to determine the fixed part of the model, which, in fact, does not require knowing b_0 . In any case, we may denote the selected fixed part as $\hat{a}(b_0)$, whether or not it depends on b_0 . Now, suppose that a selection for the random part of the model (i.e., a determination of b_0) is \hat{b} . We then define $\hat{a} = \hat{a}(\hat{b})$. In other words, once the random part is determined, we may determine the fixed part using the methods developed earlier, treating the random part as known. It can be shown that, if the selection of the random part is consistent in the sense that $P(\hat{b} \neq b_0) \rightarrow 0$, and given b_0 , the selection of the fixed part is consistent; that is, $P(\hat{a}(b_0) \neq a_0) \rightarrow 0$, then $P(\hat{a} = a_0, \hat{b} = b_0) \rightarrow 1$; that is, the combined procedure is consistent.

We now describe how to obtain \hat{b} . First divide the vectors $\alpha_1, \dots, \alpha_s$, or, equivalently, the matrices Z_1, \dots, Z_s into several groups. The first group is called the “largest random factors.” Roughly speaking, those are Z_i , $i \in S_1 \subseteq S$ such that $\text{rank}(Z_i)$ is of the same order as n , the sample size. We assume that $\mathcal{L}(X, Z_u, u \in S \setminus \{i\}) \neq \mathcal{L}(X, Z_u, u \in S)$ for any $i \in S_1$, where $\mathcal{L}(M_1, \dots, M_t)$ represents the linear space spanned by the columns of the matrices M_1, \dots, M_t . Such an assumption is reasonable because Z_i is supposed to be “the largest,” and hence should have a contribution to the linear space. The second group consists of Z_i , $i \in S_2 \subseteq S$ such that $\mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus \{i\}) \neq \mathcal{L}(X, Z_u, u \in S \setminus S_1)$, $i \in S_2$. The ranks of the matrices in this group are of lower order of n . Similarly, the third group consists of Z_i , $i \in S_3 \subseteq S$ such that $\mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus S_2 \setminus \{i\}) \neq \mathcal{L}(X, Z_u, u \in S \setminus S_1 \setminus S_2)$, and so on. Note that if the first group (i.e., the largest random factors) does not exist, the second group becomes the first, and other groups also move on. As mentioned earlier [see below (2.53)], the selection of random factors cannot be treated the same way as that of fixed factors, because the design matrices Z_1, \dots, Z_s are usually linearly dependent. Intuitively, a selection procedure will not work if there is linear dependence among the candidate design matrices, because of identifiability problems. To consider a rather extreme example, suppose that Z_1 is a design matrix consisting of 0s and 1s such that there is exactly one 1 in each row, and $Z_2 = 2Z_1$. Then, to have $Z_1\alpha_1$ in the model means that there is a term α_{1i} ; whereas to have $Z_2\alpha_2 = 2Z_1\alpha_2$ in the model means that there is a corresponding term $2\alpha_{2i}$. However, it makes no difference in terms of a

model, because both α_{1i} and α_{2i} are random effects with mean 0 and certain variances. However, by grouping the random effect factors we have divided the Z_i s into several groups such that there is linear independence within each group. This is the motivation behind grouping. To illustrate such a procedure, and also to show that such a division of groups does exist in typical situations, consider the following example.

Example 2.17. Consider the following random effects model,

$$y_{ijkl} = \mu + a_i + b_j + c_k + d_{ij} + f_{ik} + g_{jk} + h_{ijk} + e_{ijkl}, \quad (2.54)$$

$i = 1, \dots, m_1$, $j = 1, \dots, m_2$, $k = 1, \dots, m_3$, $l = 1, \dots, t$, where μ is an unknown mean; a, b, c are random main effects; d, f, g, h are (random) two- and three-way interactions; and e is an error. The model can be written as

$$y = X\mu + Z_1a + Z_2b + Z_3c + Z_4d + Z_5f + Z_6g + Z_7h + e,$$

where $X = 1_n$ with $n = m_1m_2m_3t$, $Z_1 = I_{m_1} \otimes 1_{m_2} \otimes 1_{m_3} \otimes 1_t, \dots, Z_4 = I_{m_1} \otimes I_{m_2} \otimes 1_{m_3} \otimes 1_t, \dots$, and $Z_7 = I_{m_1} \otimes I_{m_2} \otimes I_{m_3} \otimes 1_t$. It is easy to see that the Z_i s are not linearly independent. For example, $\mathcal{L}(Z_i) \subset \mathcal{L}(Z_4)$, $i = 1, 2$, and $\mathcal{L}(Z_i) \subset \mathcal{L}(Z_7)$, $i = 1, \dots, 6$. Also, $\mathcal{L}(X) \subset \mathcal{L}(Z_i)$ for any i . Suppose that $m_j \rightarrow \infty$, $j = 1, 2, 3$, and t is bounded. Then, the first group consists of Z_7 ; the second group Z_4, Z_5, Z_6 ; and the third group Z_1, Z_2, Z_3 . If t also $\rightarrow \infty$, the largest random factor does not exist. However, one still has these three groups. It is easy to see that the Z_i s within each group are linearly independent.

Suppose that the Z_i s are divided into h groups such that $S = S_1 \cup \dots \cup S_h$. We give a procedure that determines the indices $i \in S_1$ for which $\sigma_i^2 > 0$; then a procedure that determines the indices $i \in S_2$ for which $\sigma_i^2 > 0$; and so on, as follows.

Group one: Write $B = \mathcal{L}(X, Z_1, \dots, Z_s)$, $B_{-i} = \mathcal{L}(X, Z_u, u \in S \setminus \{ij\})$, $i \in S_1$; $r = n - \text{rank}(B)$, $r_i = \text{rank}(B) - \text{rank}(B_{-i})$; $R = |P_B^\perp y|^2$, $R_i = |(P_B - P_{B_{-i}})y|^2$. Let \hat{b}_1 be the set of indices i in S_1 such that

$$(r/R)(R_i/r_i) > 1 + r^{(\rho/2)-1} + r_i^{(\rho/2)-1},$$

where ρ is chosen such that $0 < \rho < 2$. Let $a_{01} = \{i \in L_1 : \sigma_i^2 > 0\}$.

Group two: Let $B_1(b_2) = \mathcal{L}(X, Z_u, u \in (S \setminus S_1 \setminus S_2) \cup b_2)$, $b_2 \subseteq S_2$. Consider

$$C_{1,n}(b_2) = |P_{B_1(b_2)}^\perp y|^2 + \lambda_{1,n}|b_2|, \quad b_2 \subseteq S_2,$$

where $\lambda_{1,n}$ is a positive number satisfying certain conditions similar to those for λ_n in (2.50) (see Jiang and Rao 2003, Section 3.3 for details). Let \hat{b}_2 be the minimizer of $C_{1,n}$ over $b_2 \subseteq S_2$, and $b_{02} = \{i \in S_2 : \sigma_i^2 > 0\}$.

General: The above procedure can be extended to the remaining groups. In general, let $B_t(b_{t+1}) = \mathcal{L}(X, Z_u, u \in (S \setminus S_1 \setminus \dots \setminus S_{t+1}) \cup b_{t+1})$, $b_{t+1} \subseteq S_{t+1}$, $1 \leq t \leq h-1$. Define

$$C_{t,n}(b_{t+1}) = |P_{B_t(b_{t+1})}^\perp y|^2 + \lambda_{t,n}|b|_{t+1}|, \quad b_{t+1} \subseteq S_{t+1},$$

where $\lambda_{t,n}$ is a positive number satisfying certain conditions similar to those for λ_n in (2.50). Let \hat{b}_{t+1} be the minimizer of $C_{t,n}$ over $b_{t+1} \subseteq S_{t+1}$, and $b_{0t+1} = \{i \in S_{t+1} : \sigma_i^2 > 0\}$.

It can be shown that, under suitable conditions, the combined procedure is consistent in the sense that $P(\hat{b}_1 = b_{01}, \dots, \hat{b}_h = b_{0h}) \rightarrow 1$ as $n \rightarrow \infty$. One property of \hat{b}_t is that it does not depend on \hat{b}_u , $u < t$. In fact, $\hat{b}_1, \dots, \hat{b}_h$ can be obtained simultaneously, and $\hat{b} = \cup_{t=1}^h \hat{b}_t$ is a consistent selection for the random part of the model. See Jiang and Rao (2003) for details.

2.5 Bayesian Inference

A linear mixed model can be naturally formulated as a hierarchical model under the Bayesian framework. Such a model usually consists of three levels, or stages of hierarchies. At the first stage, a linear model is set up given the fixed and random effects; at the second stage, the distributions of the fixed and random effects are specified given the variance component parameters; finally, at the last stage, a prior distribution is given for the variance components. Before we further explore these stages, we briefly describe the basic elements of Bayesian inference.

Suppose that y is a vector of observations and θ a vector of parameters that are not observable. Let $f(y|\theta)$ represent the probability density function (pdf) of y given θ , and $\pi(\theta)$ a prior pdf for θ . Then, the posterior pdf of θ is given by

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}.$$

Getting the posterior is the goal of Bayesian inference. In particular, some numerical summaries may be obtained from the posterior. For example, a Bayesian point estimator of θ is often obtained as the posterior mean:

$$\begin{aligned} E(\theta|y) &= \int \theta \pi(\theta|y) d\theta \\ &= \frac{\int \theta f(y|\theta) \pi(\theta) d\theta}{\int f(y|\theta) \pi(\theta) d\theta}; \end{aligned}$$

the posterior variance, $\text{var}(\theta|y)$, on the other hand, is often used as a Bayesian measure of uncertainty.

In the first stage of a hierarchical linear model, it is assumed that, given β and α ,

$$y = X\beta + Z\alpha + \epsilon,$$

where X and Z are known matrices, and ϵ has distribution F_1 . In the second stage, it is assumed that (α, β) has a joint distribution F_2 , which depends on some parameters of variance components. Finally, in the last stage, a prior distribution F_3 is assumed for the variance components. Note that a classical linear mixed model essentially involves the first two stages, but not the last one. A hierarchical model that is used most of the time is the so-called normal hierarchy, in which it is assumed that

- (1) $\epsilon \sim N(0, R)$;
- (2) $\alpha \sim N(0, G)$, $\beta \sim N(b, B)$;
- (3) $(G, R) \sim \pi$,

where π is a prior distribution. It is often assumed that, in the second stage, α and β are distributed independently, and b and B are known. Thus, a prior for β is, in fact, given in the second stage. The following is an example.

Example 2.18. Consider the one-way random effects model (Example 1.1). A normal hierarchical model assumes that (1) given μ and α_i ($1 \leq i \leq m$), $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $j = 1, \dots, n_i$, where ϵ_{ij} s are independent and distributed as $N(0, \tau^2)$; (2) $\mu, \alpha_1, \dots, \alpha_m$ are independent such that $\mu \sim N(\mu_0, \sigma_0^2)$, $\alpha_i \sim N(0, \sigma^2)$, where μ_0 and σ_0^2 are known; and (3) σ^2, τ^2 are independent with $\sigma^2 \sim \text{Inverse-}\chi^2(a)$, $\tau^2 \sim \text{Inverse-}\chi^2(b)$, where a, b are known positive constants, and an Inverse- χ^2 distribution with parameter $\nu > 0$ has pdf $\{2^{-\nu/2}/\Gamma(\nu/2)\}x^{-(\nu/2+1)}e^{-1/2x}$, $x > 0$. Alternatively, the priors in (3) may be such that $\sigma^2 \propto 1/\sigma^2$ and $\tau^2 \propto 1/\tau^2$. Note that, in the latter case, the priors are improper.

The inference includes that about the fixed and random effects and that about the variance components. In the following we discuss these two types of inference, starting with the variance components.

2.5.1 Inference about Variance Components

First define the likelihood function under the Bayesian framework. Suppose that, given α, β , and R , $y \sim f(y|\alpha, \beta, R)$. Furthermore, suppose that, given G, α and β are independent such that $\alpha \sim g(\alpha|G)$ and $\beta \sim h(\beta|b, B)$ (b, B known). Then, the full likelihood function, or simply the likelihood, for estimating G and R , is given by

$$L(G, R|y) = \int \int f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)d\alpha d\beta, \quad (2.55)$$

where the integrals with respect to α and β may be both multivariate. Note that the difference between a likelihood and a posterior is that the prior is not taken into account in obtaining the likelihood. We now consider two special cases under the normal hierarchy.

The first case is when h is a point mass (or degenerate distribution) at β . Then, the limit of (2.55), when $b = \beta$ and $B \rightarrow 0$, reduces to

$$L(G, R|y) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} \exp \left\{ -\frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta) \right\}$$

(Exercise 2.22), where $V = ZGZ' + R$. This is simply the (normal) likelihood function given in Section 1.3.1. Under the Bayesian framework, it is also called the conditional likelihood, because a point mass corresponds to being conditional on β .

The second case is when h is a noninformative, or flat, distribution, that is, the prior for β is uniform over $(-\infty, \infty)$. Note that this is an improper prior. Nevertheless, the likelihood (2.55) does exist and has the expression

$$L(G, R|y) = \frac{1}{(2\pi)^{(n-p)/2}|A'VA|^{1/2}} \exp \left\{ -\frac{1}{2}z'(A'VA)^{-1}z \right\},$$

where $p = \text{rank}(X)$, $z = A'y$, and A is an $n \times (n - p)$ matrix such that $\text{rank}(A) = n - p$ and $A'X = 0$ (Exercise 2.23). This is exactly the (normal) restricted likelihood function defined in Section 1.3.2. Under the Bayesian framework, it is also called the marginal likelihood, because it has β integrated out with respect to the noninformative prior.

Thus, without taking the prior into account, the likelihood can be used to obtain estimators of G and R , as one does in classical situations. This method is used later to obtain empirical Bayes estimators of the effects.

If the prior is taken into account, then the posterior for G and R can be expressed as

$$\begin{aligned} & \pi(G, R|y) \\ &= \int \int \frac{f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)}{\int \int \int \int f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)d\alpha d\beta dG dR} d\alpha d\beta \\ &= \frac{L(G, R|y)\pi(G, R)}{\int \int L(G, R|y)\pi(G, R)dG dR} \end{aligned} \quad (2.56)$$

where $\pi(G, R)$ is a prior pdf for G, R . The computation of (2.56) can be fairly complicated even for a simple model (Exercise 2.24). For complex models the computation of (2.56) is typically carried out by Markov chain Monte Carlo (MCMC) methods.

2.5.2 Inference about Fixed and Random Effects

Similar to (2.56), the posterior for β can be expressed as

$$\begin{aligned} & \pi(\beta|y) \\ &= \int \int \int \frac{f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)}{\int \int \int \int f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)d\alpha d\beta dG dR} \\ & \quad d\alpha dG dR, \end{aligned} \quad (2.57)$$

and the posterior for α is

$$\pi(\alpha|y) = \frac{\int \int \int \frac{f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)}{\int \int \int \int f(y|\alpha, \beta, R)g(\alpha|G)h(\beta|b, B)\pi(G, R)d\alpha d\beta dG dR} d\beta dG dR. \quad (2.58)$$

If normality is assumed, (2.57) and (2.58) may be obtained in closed forms. In fact, in the case of normal hierarchy, we have

$$\beta|y \sim N(E(\beta|y), \text{Var}(\beta|y)),$$

where $E(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}(X'V^{-1}y + B^{-1}b)$, $\text{Var}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}$; and, similarly,

$$\alpha|y \sim N(E(\alpha|y), \text{Var}(\alpha|y)),$$

where $E(\alpha|y) = (Z' LZ + G^{-1})^{-1}Z' L(y - Xb)$, $\text{Var}(\alpha|y) = (Z' LZ + G^{-1})^{-1}$ with $L = R^{-1} - R^{-1}X(B^{-1} + X'R^{-1}X)^{-1}X'R^{-1}$ (Exercise 2.25). It is interesting to note that, when $B^{-1} \rightarrow 0$, which corresponds to the case where the prior for β is noninformative, one has $E(\beta|y) \rightarrow (X'V^{-1}X)^{-1}X'V^{-1}y = \tilde{\beta}$, which is the BLUE; similarly, $E(\alpha|y) \rightarrow GZ'V^{-1}(y - X\tilde{\beta})$ (Exercise 2.26), which is the BLUP (see Section 2.3.1.2). Thus, the BLUE and BLUP may be regarded as the posterior means of the fixed and random effects under normal hierarchy and a limiting situation, or noninformative prior for β .

Note that the BLUE and BLUP depend on G and R , which are unknown in practice. Instead of assuming a prior for G and R , one may estimate these covariance matrices, which often depend parametrically on some variance components, by maximizing the marginal likelihood function introduced before (see the early part of Section 2.5.1). This is called the empirical Bayes (EB) method. Harville (1991) showed that in the special case of the one-way random effects model (see Example 1.1), the EB is identical to EBLUP (see Section 2.3.1.3). From the above derivation, it is seen that this result actually holds more generally in a certain sense. Note that when G and R in BLUE and BLUP are replaced by estimators, the results are EBLUE and EBLUP. However, as Harville noted, much of the work on EB has focused on relatively simple models, whereas EBLUP has been carried out by practitioners such as individuals in the animal breeding area and survey sampling to relatively complex models.

2.6 Real-Life Data Examples

2.6.1 Analysis of the Birth Weights of Lambs (Continued)

In this section, we revisit the example of lamb-weight data discussed in section 1.7.1, where estimates of the fixed effects and variance components were obtained.

The BLUPs of the sire effects are obtained by PROC MIXED. The results are shown in Table 2.3. Here Standard Pred Error represents the square root of the estimated mean square prediction error (MSPE) of the EBLUP of s_{ij} , the j th sire effect in line i . The estimated MSPE in PROC MIXED is obtained by substituting the REML estimates of the variance components into the formula for the MSPE assuming known variance components. This is known as the naive method of estimating the MSPE. As discussed earlier (see Section 2.3.1), the naive estimates may underestimate the true MSPE. Methods that improve the accuracy of the MSPE estimation have been proposed. See Section 2.3.1.

Table 2.3. BLUPs of the random effects

Sire	Line	Estimate	Standard Pred Error
11	1	−0.6256	0.6693
12	1	0.3658	0.6693
13	1	0.5050	0.6156
14	1	−0.2452	0.6441
21	2	0.1750	0.6701
22	2	0.1588	0.6296
23	2	−0.0423	0.6717
24	2	−0.2914	0.6457
31	3	−0.2667	0.6184
32	3	−0.2182	0.5850
33	3	0.3212	0.6397
34	3	0.1637	0.6701
41	4	−0.2015	0.6187
42	4	0.0695	0.6454
43	4	0.1319	0.6436
51	5	0.3047	0.6356
52	5	−0.2437	0.6308
53	5	−0.1177	0.6327
54	5	−0.1549	0.6656
55	5	0.3940	0.6684
56	5	−0.6311	0.6318
57	5	0.5762	0.5913
58	5	−0.1274	0.5769

2.6.2 The Baseball Example

In this section, we revisit the Efron–Morris baseball example (Example 2.11) and use it to illustrate methods of diagnostics in linear mixed models. This example is chosen because of its simplicity. The dataset has been analyzed by several authors in the past, including Efron and Morris (1975), Efron (1975), Morris (1983), Datta and Lahiri (2000), Gelman et al. (1995), Rao (2003),

and Lahiri and Li (2005), among others. Efron and Morris (1975) used this dataset to demonstrate the performance of their empirical Bayes and limited translation empirical Bayes estimators derived using an exchangeable prior in the presence of an outlying observation. They first obtained the batting average of Roberto Clemente, an extremely good hitter, from the *New York Times* dated April 26, 1970 when he had already batted $n = 45$ times. The batting average of a player is just the proportion of hits among the number at-bats. They selected 17 other major league baseball players who had also batted 45 times from the April 26 and May 2, 1970 issues of the *New York Times*. They considered the problem of predicting the batting averages of all 18 players for the remainder of the 1970 season based on their batting averages for the first 45 at-bats. This is a good example for checking the effect of an outlier on the efficiency of an EB estimation with an exchangeable prior. Gelman et al. (1995) provided additional data for this estimation problem and included important auxiliary data such as the batting average of each player through the end of the 1969 season. Jiang and Lahiri (2005b) reviewed the problem of predicting the batting averages of all 18 players for the entire 1970 season, instead of predicting the batting averages for the remainder of the 1970 season as Efron and Morris (1975) originally considered.

For the player i ($i = 1, \dots, m$), let p_i and π_i be the batting average for the first 45 at-bats and the true season batting average of the 1970 season. Note that p_i is the direct maximum likelihood (also unbiased) estimator of π_i under the assumption that conditional on π_i , the number of hits for the first n at-bats, np_i , follows a binomial distribution with number of trials n and success probability π_i , $i = 1, \dots, m$.

Efron and Morris (1975) considered the following standard arc-sine transformation,

$$y_i = \sqrt{n} \arcsin(2p_i - 1)$$

and then assumed the following model

$$y_i | \theta_i \stackrel{\text{ind}}{\sim} N(\theta_i, 1), \quad i = 1, \dots, m,$$

where $\theta_i = \sqrt{n} \arcsin(2\pi_i - 1)$. There could be a criticism about the validity of the above approximation. However, Efron and Morris (1975) and Gelman et al. (1995) noted that this is not a serious concern given the moderate sample size of 45. The data analysis by Lahiri and Li (2005) supports this conjecture. Efron and Morris (1975) assumed exchangeability of the θ_i s and used the two-level Fay–Herriot model, given in Section 2.1, without any covariate and equal sampling variances (i.e., 1).

Gelman et al. (1995) noted the possibility of an extra-binomial variation in the number of hits. The outcomes from successive at-bats could be correlated and the probability of hits may change across at-bats due to injury to the player and other external reasons not given in the dataset. However, there is no way to check these assumptions because of the unavailability of such data.

Assuming Level 1 is reasonable, Lahiri and Li (2005) checked the validity of the above model through graphical tools. To this end, they used the following standardized residual,

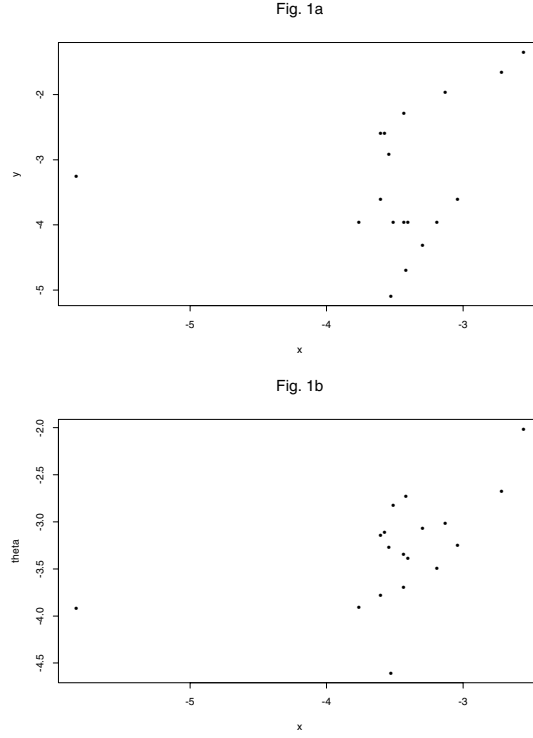
$$e_i = \frac{y_i - \bar{y}}{s},$$

where $s^2 = (m-1)^{-1} \sum_{i=1}^m (y_i - \bar{y})^2$ is the usual sample variance. Note that marginally $y_i \stackrel{iid}{\sim} N(\mu, 1+A)$. Under this marginal model, $E(e_i) \approx 0$, and $\text{var}(e_i) \approx 1+A$ for large m . Thus, if the model is reasonable, a plot of the standardized residuals versus the players is expected to fluctuate randomly around 0. Otherwise, one might suspect the adequacy of the two-level model. However, random fluctuation of the residuals may not reveal certain systematic patterns of the data. For example, Lahiri and Li (2005) noted that the residuals, when plotted against players arranged in increasing order of the previous batting averages, did reveal a linear regression pattern, something not apparent when the same residuals were plotted against players arranged in an arbitrary order. This is probably questioning the exchangeability assumption in the Efron–Morris model, a fact we knew earlier because of the intentional inclusion of an extremely good hitter.

Let p_{i0} be the batting average of player i through the end of the 1969 season and $x_i = \sqrt{n} \arcsin(2p_{i0} - 1)$, $i = 1, \dots, m$. We plot y and θ versus x in Figure 2.1 (a) and (b) respectively. This probably explains the systematic pattern of the residuals mentioned in the previous paragraph. We also note the striking similarity of the two graphs: 1(a) and 1(b). Although Roberto Clemente seems like an outlier with respect to y , θ , or x , player L. Alvarado appears to be an outlier in the sense that his current batting average is much better than his previous batting average. He influences the regression fit quite a bit. For example, the BIC for the two-level model reduced from 55 to 44 when Alvarado was dropped from the model. Further investigation shows that this player is a rookie and batted only 51 times through the end of the 1969 season compared to other players in the dataset, making his previous batting average information not very useful. The BICs for the Fay–Herriot model with and without the auxiliary data are almost the same (54.9 and 55.3, respectively), a fact not expected at the beginning of the data analysis. In spite of more or less similar BIC values and the presence of an outlier in the regression, Figure 2.2 shows that EMReg did a good job in predicting the batting averages of Clemente and Alvarado, two different types of outliers. Further details on this data analysis are given in Lahiri and Li (2005).

2.7 Further Results and Technical Notes

1. *Robust versions of classical tests.* We first state the following theorems, which also define the matrices A , B , C , and Σ introduced in Section 2.1.2.4.

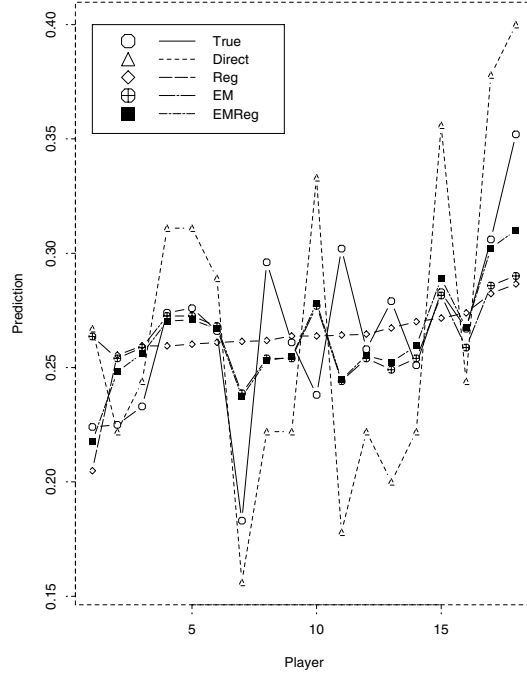
**Fig. 2.1.** Source: Adapted from Lahiri and Li (2005)

Theorem 2.1. Suppose that the following hold. (i) $l(\cdot, y)$ is twice continuously differentiable for fixed y , and $\psi(\cdot)$ is twice continuously differentiable. (ii) With probability $\rightarrow 1$, $\hat{\psi}$, $\hat{\phi}$ satisfy $\partial l / \partial \psi = 0$, $\partial l_0 / \partial \phi = 0$, respectively. (iii) There are sequences of nonsingular symmetric matrices $\{G\}$ and $\{H\}$ and matrices A , B , C with A , $B > 0$ such that the following $\rightarrow 0$ in probability,

$$\begin{aligned} & \sup_{\mathcal{S}_1} \left\| G^{-1} \left(\frac{\partial^2 l}{\partial \psi_i \partial \psi_j} \Big|_{\psi^{(i)}} \right)_{1 \leq i, j \leq q} G^{-1} + A \right\|, \\ & \sup_{\mathcal{S}_2} \left\| H^{-1} \left(\frac{\partial^2 l_0}{\partial \phi_i \partial \phi_j} \Big|_{\phi^{(i)}} \right)_{1 \leq i, j \leq p} H^{-1} + B \right\|, \\ & \sup_{\mathcal{S}_3} \left\| G \left(\frac{\partial \psi_i}{\partial \phi_j} \Big|_{\phi^{(i)}} \right)_{1 \leq i \leq q, 1 \leq j \leq p} H^{-1} - C \right\|, \end{aligned}$$

where $\mathcal{S}_1 = \{|\psi^{(i)} - \psi_0|_v \leq |\hat{\psi} - \psi_0|_v \vee |\psi(\hat{\phi}) - \psi(\phi_0)|_v, 1 \leq i \leq q\}$, $\mathcal{S}_2 = \{|\phi^{(i)} - \phi_0|_v \leq |\hat{\phi} - \phi_0|_v, 1 \leq i \leq p\}$, $\mathcal{S}_3 = \{|\phi^{(i)} - \phi_0|_v \leq |\hat{\phi} - \phi_0|_v, 1 \leq i \leq q\}$

Fig. 2

**Fig. 2.2.** Source: Adapted from Lahiri and Li (2005)

and $|a|_v = (|a_1|, \dots, |a_k|)'$ for $a = (a_1, \dots, a_k)'$;
(iv) $D(\partial l / \partial \psi)|_{\psi_0} \rightarrow 0$ in probability, where $D = \text{diag}(d_i, 1 \leq i \leq s)$ with $d_i = \|H^{-1}(\partial^2 \psi_i / \partial \phi \partial \phi')|_{\phi_0} H^{-1}\|$, and

$$G^{-1} \frac{\partial l}{\partial \psi} \Big|_{\psi_0} \longrightarrow N(0, \Sigma) \quad \text{in distribution.} \quad (2.59)$$

Then, under the null hypothesis, the asymptotic distribution of \mathcal{W} is χ_r^2 , where \mathcal{W} is defined in (2.18), and $r = \text{rank}[\Sigma^{1/2} A^{-1/2} (I - P)]$ with $P = A^{1/2} C (C' A C)^{-1} C' A^{1/2}$. In particular, if Σ is nonsingular, then $r = q - p$.

The theorem may be extended to allow the matrices A , B , and so on, to be replaced by sequences of matrices. Such an extension may be useful. For example, suppose G is a diagonal normalizing matrix; then, in many cases, A can be chosen as $-G^{-1}[E(\partial^2 l / \partial \psi \partial \psi')|_{\psi_0}]G^{-1}$, but the latter may not have a limit as $n \rightarrow \infty$.

Extension of Theorem 2.1. Suppose that, in Theorem 2.1, A , B , C are replaced by sequences of matrices $\{A\}$, $\{B\}$, and $\{C\}$, such that A , B are symmetric,

$$0 < \liminf[\lambda_{\min}(A) \wedge \lambda_{\min}(B)] \leq \limsup[\lambda_{\max}(A) \vee \lambda_{\max}(B)] < \infty,$$

and $\limsup \|C\| < \infty$. Furthermore, suppose that (2.59) is replaced by

$$\Sigma^{-1/2} G^{-1} \left. \frac{\partial l}{\partial \psi} \right|_{\psi_0} \longrightarrow N(0, I) \quad \text{in distribution,} \quad (2.60)$$

where $\{\Sigma\}$ is a sequence of positive definite matrices such that

$$0 < \liminf \lambda_{\min}(\Sigma) \leq \limsup \lambda_{\max}(\Sigma) < \infty,$$

and I is the p -dimensional identity matrix. Then, the asymptotic distribution of \mathcal{W} is χ_{q-p}^2 .

The proofs are given in Jiang (2002). According to the proof, one has $G[\hat{\psi} - \psi(\hat{\phi})] = O_P(1)$, hence

$$\begin{aligned} \hat{\mathcal{W}} &= [\hat{\theta} - \theta(\hat{\phi})]' G[Q_w^- + o_P(1)] G[\hat{\theta} - \theta(\hat{\phi})] \\ &= \mathcal{W} + o_P(1). \end{aligned}$$

Thus, by Theorem 2.1, we conclude the following.

Corollary 2.1. Under the conditions of Theorem 2.1, the asymptotic distribution of $\hat{\mathcal{W}}$ is χ_r^2 , where r is the same as in Theorem 2.1. Thus, in particular, if Σ is nonsingular, $r = q - p$. Under the conditions of Extension of Theorem 2.1, the asymptotic distribution of $\hat{\mathcal{W}}$ is χ_{q-p}^2 .

We now consider the asymptotic distribution of the \mathcal{S} -test defined in (2.19).

Theorem 2.2. Suppose that the conditions of Theorem 2.1 are satisfied with the following changes: (1) in (ii), that $\hat{\psi}$ satisfies $\partial l / \partial \psi = 0$ with probability $\rightarrow 1$ is not required; and (2) in (iii), the supremum for the first quantity (involving A) is now over $|\psi^{(i)} - \psi_0|_v \leq |\psi(\hat{\phi}) - \psi(\phi_0)|_v$, $1 \leq i \leq q$. Then, under the null hypothesis, the asymptotic distribution of \mathcal{S} is χ_r^2 , where r is the same as in Theorem 2.1. In particular, if Σ is nonsingular, then $r = q - p$.

In exactly the same way, we have the following.

Extension of Theorem 2.2. Suppose that, in Theorem 2.2, A , B , and C are replaced by $\{A\}$, $\{B\}$, and $\{C\}$, and (2.59) by (2.60), where the sequences of matrices $\{A\}$, $\{B\}$, $\{C\}$, and $\{\Sigma\}$ satisfy the conditions of the Extension of Theorem 2.1. Then, the asymptotic distribution of \mathcal{S} is χ_{q-p}^2 .

Corollary 2.2. Under the conditions of Theorem 2.2, the asymptotic distribution of $\hat{\mathcal{S}}$ is χ_r^2 , where r is the same as in Theorem 2.1. Thus, in particular, if Σ is nonsingular, $r = q - p$. Under the conditions of the Extension of Theorem 2.2, the asymptotic distribution of $\hat{\mathcal{S}}$ is χ_{q-p}^2 .

Finally, we consider the asymptotic distribution of the L -test. It is seen that the asymptotic distributions for the W - and \mathcal{S} -tests are both χ^2 . However,

the following theorem states that the asymptotic distribution for the L -test is not χ^2 but a “weighted” χ^2 (e.g., Chernoff and Lehmann 1954). Recall that Q_l is defined near the end of Section 2.1.2.4.

Theorem 2.3. Suppose that the conditions of Theorem 2.1 are satisfied except that the third quantity in (iii) (involving C) $\rightarrow 0$ in probability is replaced by $G[(\partial\psi/\partial\phi)|_{\phi_0}]H^{-1} \rightarrow C$. Then, under the null hypothesis, the asymptotic distribution of $-2\log R$ is the same as $\lambda_1\xi_1^2 + \cdots + \lambda_r\xi_r^2$, where r is the same as in Theorem 2.1; $\lambda_1, \dots, \lambda_r$ are the positive eigenvalues of Q_l ; and ξ_1, \dots, ξ_r are independent $N(0, 1)$ random variables. In particular, if Σ is nonsingular, then $r = q - p$.

Again, the proofs are given in Jiang (2002). It should be pointed out that if $L(\theta, y)$ is, indeed, the likelihood function, in which case the L -test is the likelihood-ratio test, the asymptotic distribution of $-2\log R$ reduces to χ^2 see Weiss 1975).

Let \hat{Q}_l be a consistent estimator of Q_l . Then, by Weyl’s eigenvalue perturbation theorem (see Appendix B), the eigenvalues of \hat{Q}_l are consistent estimators of those of Q_l , and therefore can be used to obtain the asymptotic critical values for the L -test.

We now specify the W -, S -, and L -tests under the non-Gaussian mixed ANOVA model (see Section 1.2.2) with the additional assumption that

$$E(\epsilon_1^3) = 0, \quad E(\alpha_{i1}^3) = 0, \quad 1 \leq i \leq s. \quad (2.61)$$

As it turns out, this assumption is not essential but simplifies the results considerably. First define

$$\begin{aligned} A_1 &= (\text{tr}(V^{-1}V_i)/2\lambda\sqrt{nm_i})_{1 \leq i \leq s}, \\ A_2 &= (\text{tr}(V^{-1}V_iV^{-1}V_j)/2\sqrt{m_im_j})_{1 \leq i, j \leq s}, \\ A &= \begin{pmatrix} X'V^{-1}X/\lambda n & 0 & 0 \\ 0 & 1/2\lambda^2 & A'_1 \\ 0 & A_1 & A_2 \end{pmatrix}. \end{aligned} \quad (2.62)$$

Let $b = (I \sqrt{\gamma_1}Z_1 \cdots \sqrt{\gamma_s}Z_s)$, $B_0 = b'V^{-1}b$, $B_i = b'V^{-1}V_iV^{-1}b$, $1 \leq i \leq s$. Furthermore, we define

$$\begin{aligned} D_{0,ij} &= \sum_{l=1}^n B_{i,l}B_{j,l}, \\ D_{1,ij} &= \sum_{l=n+1}^{n+m_1} B_{i,l}B_{j,l}, \\ &\vdots \\ D_{s,ij} &= \sum_{l=n+m_1+\cdots+m_{s-1}+1}^{n+m_1+\cdots+m_s} B_{i,l}B_{j,l}, \end{aligned}$$

where $B_{i,kl}$ is the (k, l) element of B_i , $0 \leq i, j \leq s$. The kurtoses of the errors and random effects are defined by $\kappa_0 = (E\epsilon_1^4/\sigma_0^4) - 3$, and $\kappa_i = (E\alpha_{i1}^4/\sigma_i^4) - 3$, $1 \leq i \leq s$. Let $\Delta_1 = (\Delta_{0i}/\sqrt{nm_i})_{1 \leq i \leq s}$, $\Delta_2 = (\Delta_{ij}/\sqrt{m_i m_j})_{1 \leq i, j \leq s}$, and

$$\Delta = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \Delta_{00}/n & \Delta_1' \\ 0 & \Delta_1 & \Delta_2 \end{pmatrix}, \quad (2.63)$$

where $\Delta_{ij} = \{4\lambda^{1(i=0)+1(j=0)}\}^{-1} \sum_{t=0}^s \kappa_t D_{t,ij}$, $0 \leq i, j \leq s$. Let

$$W = b'V^{-1}X(X'V^{-1}X)^{-1/2},$$

and W'_l be the l th row of W , $1 \leq l \leq n + m$, where $m = m_1 + \dots + m_s$.

Theorem 2.4. Suppose that the following hold.

- (i) $\theta(\cdot)$ is three-times continuously differentiable and satisfies (2.21), and $\partial\theta_{i_k}/\partial\phi_k \neq 0$, $1 \leq k \leq d$.
 - (ii) $E\epsilon_1^4 < \infty$, $\text{var}(\epsilon_1^2) > 0$, $E\alpha_{i1}^4 < \infty$, $\text{var}(\alpha_{i1}^2) > 0$, $1 \leq i \leq s$, and (2.61) holds.
 - (iii) $n \rightarrow \infty$, $m_i \rightarrow \infty$, $1 \leq i \leq s$, $0 < \liminf \lambda_{\min}(A) \leq \limsup \lambda_{\max}(A) < \infty$, and $\max_{1 \leq l \leq n+m} |W_l| \rightarrow 0$;
- Then, for $l(\theta, y)$ there exist $\hat{\theta}$ and $\hat{\phi}$ such that the conditions of the Extensions of Theorems 2.1 and 2.2 are satisfied with

$$\begin{aligned} G &= \text{diag}(\sqrt{n}, \dots, \sqrt{n}, \sqrt{m_1}, \dots, \sqrt{m_s}) \\ &= \text{diag}(g_i, 1 \leq i \leq q), \end{aligned}$$

$H = \text{diag}(g_{i_k}, 1 \leq k \leq a)$, A is given by (2.62), $C = \partial\theta/\partial\phi$, $B = C'AC$, and $\Sigma = A + \Delta$, where Δ is given by (2.63). Therefore, the asymptotic null distribution of both $\hat{\chi}_w^2$ and $\hat{\chi}_s^2$ is χ_{q-d}^2 . The same conclusion holds for $l_R(\theta, y)$ as well.

Note that the i th row of $\partial\theta/\partial\phi$ is $\partial\theta_i/\partial\phi'$, which is $(0, \dots, 0)$ if $i \notin \{i_1, \dots, i_a\}$, and $(0, \dots, 0, \partial\theta_{i_k}/\partial\phi_k, 0, \dots, 0)$ (k th component nonzero) if $i = i_k$, $1 \leq k \leq a$ under (2.21).

Theorem 2.5. Suppose that the conditions of Theorem 2.4 are satisfied except that, in (iii), the condition about A is strengthened to that $A \rightarrow A_0$, where $A_0 > 0$, and $\Sigma \rightarrow \Sigma_0$. Then, the conditions of Theorem 2.3 are satisfied with $A = A_0$, $\Sigma = \Sigma_0$, and everything else given by Theorem 2.4. Therefore, the asymptotic null distribution of $-2\log R$ is the same as $\sum_{j=1}^r \lambda_j \xi_j^2$, where $r = \text{rank}\{\Sigma^{1/2}A^{-1/2}(I - P)\}$, evaluated under H_0 with $P = A^{1/2}C(C'AC)^{-1}C'A^{1/2}$; λ_j s are the positive eigenvalues of Q_l given by (2.20), again evaluated under H_0 ; and ξ_j s are independent $N(0, 1)$ random variables. In particular, if Σ is nonsingular under H_0 , then $r = q - d$. The same conclusion holds for $l_R(\theta, y)$ as well.

The proof of Theorems 2.1–2.5 can be found in Jiang (2005c).

It is seen from (2.63) that Δ , and hence Σ , depends on the kurtoses κ_i , $0 \leq i \leq s$, in addition to the variance components σ_i^2 , $0 \leq i \leq s$. One already has consistent estimators of σ_i^2 , $0 \leq i \leq s$ (e.g., the ML or REML estimators). As for κ_i , $0 \leq i \leq s$, they can be estimated by the empirical method of moments (EMM) of Jiang (2003b).

The extension of Theorem 1 and Theorem 2 without assuming (2.61) is fairly straightforward, although the results will not be as simple. Note that Theorems 2.1–2.3 (and their extensions) do not require (2.61). However, there is a complication in estimating the additional parameters involved in Σ . This is because, without (2.61), the matrix Δ also involves the third moments of the random effects and errors (on the off-diagonal). In such a case, the EMM of Jiang (2003b) is not directly applicable. Alternatively, Σ can be consistently estimated by the POQUIM method (see Sections 1.4.2 and 1.8.5), which does not require (2.61).

2. Existence of moments of ML and REML estimators. Jiang (2000a) established the existence of moments of ML and REML estimators under non-Gaussian linear mixed models (see Section 1.4.1) as an application of a matrix inequality. Let A_1, \dots, A_s be nonnegative definite matrices. Then, there are positive constants depending on the matrices such that for all positive numbers x_1, \dots, x_s ,

$$A_i \leq \frac{c_i}{x_i^2} \left(I + \sum_{j=1}^s x_j A_j \right)^2, \quad 1 \leq i \leq s.$$

Now consider a non-Gaussian mixed ANOVA model (see Section 1.2.2.1), where $y = (y_i)_{1 \leq i \leq n}$. The ML and REML estimators are defined in Sections 1.3.1 and 1.3.2, respectively, and EBLUE and EBLUP in Sections 2.2.1.4 and 2.3.1.3, respectively.

Theorem 2.6. The k th moments ($k > 0$) of the ML or REML estimators of $\sigma_1^2, \dots, \sigma_s^2, \tau^2$ are finite, provided that the $2k$ th moments of y_i , $1 \leq i \leq n$ are finite.

3. Existence of moments of EBLUE and EBLUP. In the same paper, Jiang (2000a) established the existence of moments of EBLUE and EBLUP as another application of the same matrix inequality. Again, no normality assumption is made. Note that here the only requirement for the variance components estimators is that they are nonnegative. In the following theorem, the abbreviations EBLUEs and EBLUPs stand for the components of EBLUE and EBLUP, respectively.

Theorem 2.7. The k th moments ($k > 0$) of EBLUEs and EBLUPs are finite, provided that the k th moments of y_i , $1 \leq i \leq n$ are finite, and the variance components estimators are nonnegative.

Because it is always assumed that the second moments of the data are finite, we have the following conclusion.

Corollary 2.3. The means and MSEs of EBLUE and EBLUP exist as long as the variance components estimators are nonnegative.

Note 1. Kackar and Harville (1984) showed that the EBLUE and EBLUP remain unbiased if the variance components are estimated by nonnegative, even, and translation-invariant estimators (see Section 2.3.1.3). In deriving their results, Kackar and Harville avoided the existence of the means of EBLUE and EBLUP. Jiang (1999b) considered a special case of linear mixed models corresponding to $s = 1$ in (1.2) and proved the existence of the means. The above corollary has solved the problem for the general case.

Note 2. The ML and REML estimators are nonnegative by their definitions (see Section 1.4.1). However, for example, the ANOVA estimators may take negative values (see Section 1.5.1).

4. *The definition of $\Sigma_n(\theta)$ in Section 2.4.1.2.* First consider the case $s = 0$, that is, the case of linear regression. In this case, we have $y_i = x'_i\beta + \epsilon_i$, $i = 1, \dots, n$, where x'_i is the i th row of X , which has full rank p , and ϵ_i s are i.i.d. errors with mean 0, variance τ^2 , and an unknown distribution $G(\cdot|\tau)$. Thus, in this case, $\theta = (\beta', \tau^2)'$. The matrix $\Sigma_n(\theta)$ is defined as $n^{-1} \sum_{i=1}^n \text{Var}(h_i)$, where

$$h_i = (1_{(y_i \in E_k)} - p_{ik}(\theta))_{1 \leq k \leq M} - \left(\sum_{i=1}^n \frac{\partial p_i(\theta)}{\partial \beta'} \right) (X'X)^{-1} x_i \epsilon_i \\ - \frac{1 - x'_i(X'X)^{-1}x_i}{n-p} \left(\sum_{i=1}^n \frac{\partial p_i(\theta)}{\partial \tau^2} \right) \epsilon_i^2$$

with $p_i(\theta) = (p_{ik}(\theta))_{1 \leq k \leq M}$ and $p_{ik}(\theta) = P_\theta(y_i \in E_k)$. Jiang (2001) gives a more explicit expression of $\Sigma_n(\theta)$. On the other hand, it may be more convenient to compute $\hat{\Sigma}_n = \Sigma_n(\hat{\theta})$ by a Monte Carlo method, where $\hat{\theta} = (\hat{\beta}', \hat{\tau}^2)'$ with $\hat{\beta}$ being the least squares estimator and $\hat{\tau}^2 = |y - X\hat{\beta}|^2/(n-p)$.

We now consider another special case, the case $s = 1$, such that $y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij}$, $i = 1, \dots, m$, $j = 1, \dots, k_i$, where the α_i s are i.i.d. with mean 0, variance σ^2 , and an unknown distribution $F(\cdot|\sigma)$, ϵ_{ij} s are i.i.d. with mean 0, variance τ^2 , and an unknown distribution $G(\cdot|\tau)$, and α, ϵ are independent. In other words, we consider the nested error regression model (2.48). Write the model in the standard form $y = X\beta + Z\alpha + \epsilon$. Let $\theta = (\beta', \tau^2, \gamma)'$, where $\gamma = \sigma^2/\tau^2$. Define

$$\Sigma_n(\theta) = a_n^{-1} \left\{ \sum_{i=1}^m \text{Var}(h_i) + 2\Phi'(\mathcal{I} - \mathcal{R})\Phi \right\},$$

where \mathcal{I} is defined in Section 1.8.3, and h_i, Φ, \mathcal{R} are defined as follows. Recall the notation introduced in Section 1.8.3. Redefine $p_1 = [\text{tr}\{(Z'V(\gamma)Z)^2\}]^{1/2}$.

Recall $p_0 = \sqrt{n-p}$. Let $\rho = \text{tr}\{Z'V(\gamma)Z\}/p_0p_1$. Let $P_{ij}(\theta)$ be the $M \times (p+2)$ matrix whose (k, r) element is

$$\frac{\partial}{\partial \theta_r} \int \{G(c_k - x'_{ij}\beta - u|\tau) - G(c_{k-1} - x'_{ij}\beta - u|\tau)\} dF(u|\sigma)$$

(θ_r is the r th component of θ). Let $P_{ij}[r](\theta)$ be the r th column of $P_{ij}(\theta)$, and $P_{ij}[1, p](\theta)$, the matrix, consist of the first p columns of $P_{ij}(\theta)$. Define

$$\begin{aligned} \Phi &= \frac{1}{1-\rho^2} \begin{pmatrix} \tau^4 & -\tau^2\rho \\ -\tau^2\rho & 1 \end{pmatrix} \begin{pmatrix} p_0^{-1} \sum_{i,j} P_{ij}[p+1](\theta)' \\ p_1^{-1} \sum_{i,j} P_{ij}[p+2](\theta)' \end{pmatrix} \\ &= \begin{pmatrix} \Phi'_0 \\ \Phi'_1 \end{pmatrix}, \\ \Psi &= \tau b(\gamma) V_\gamma^{-1} X (X' V_\gamma^{-1} X)^{-1} \sum_{i,j} P_{ij}[1, p](\theta)' \\ &= (\Phi'_l)_{1 \leq l \leq m+n}, \end{aligned}$$

where $V_\gamma = V/\tau^2$. Let $S_i = \{l : \sum_{i' < i} k_{i'} + 1 \leq l \leq \sum_{i' \leq i} k_{i'}\} \cup \{n+i\}$. Write $\omega(i) = (\omega_l)_{l \in S_i}$, $V_j(i, i') = (V_j(\gamma)_{l, l'})_{l \in S_i, l' \in S_{i'}}$, $j = 0, 1$, $\Psi(i) = (\Phi'_l)_{l \in S_i}$. Let

$$\begin{aligned} h_i &= \left(\sum_{j=1}^{k_i} \{1_{(y_{ij} \in E_k)} - p_{ijk}(\theta)\} \right)_{1 \leq k \leq M} - \Psi(i)' \omega(i) \\ &\quad - \sum_{j=0}^1 \frac{\omega(i)' V_j(i, i) \omega(i)}{\tau^{2(1-j)} p_j} \Phi_j, \end{aligned}$$

where $p_{ijk}(\theta) = P_\theta(y_{ij} \in E_k)$. Finally, let $\mathcal{R} = (r_{j,j'})_{j,j'=0,1}$, where

$$r_{j,j'} = \frac{\sum_{i=1}^m \text{tr}\{V_j(i, i) V_{j'}(i, i)\}}{\tau^{2(2-j-j')} p_j p_{j'}}.$$

Finally, in the case of multiple random effect factors, that is, $s \geq 2$, $\Sigma_n(\theta)$ is defined in a similar way; that is, $\Sigma_n(\theta) = a_n^{-1} \{\sum_{l=1}^L \text{Var}(h_l) + 2\Phi'(\mathcal{I} - \mathcal{R})\Phi\}$. We omit the definitions of h , Φ , and \mathcal{R} here and refer the details to Jiang (2001, Section 4) (\mathcal{I} is the same as before).

2.8 Exercises

2.1. Derive explicit expressions of the test statistic (2.3) (in terms of the y_{ijk} s) for the two cases considered in Example 2.1 where the exact F -test applies: (i) testing $\sigma_1^2 = 0$ under the model without interaction; and (ii) testing $\sigma_3^2 = 0$ under the model with interaction.

2.2. Consider the following random effects model,

$$y_{ijkl} = \mu + f_i + g_j + u_{ij} + v_{jk} + w_{ijk} + e_{ijkl}$$

(see, e.g., Searle 1971, for notation), $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, c$, $l = 1, \dots, d$, where μ is an unknown mean, e_{ijkl} is an error, and all the others are random effects. Assume that the random effects and errors are independent such that $f_i \sim N(0, \sigma_1^2)$, $g_j \sim N(0, \sigma_2^2)$, $u_{ij} \sim N(0, \sigma_3^2)$, $v_{jk} \sim N(0, \sigma_4^2)$, $w_{ijk} \sim N(0, \sigma_5^2)$, and $e_{ijkl} \sim N(0, \tau^2)$. Do exact or optimal tests exist for testing $H_0: \sigma_2^2 = 0$? Please explain. (Hint: Consider Result 2 of Mathew and Sinha (1988) described in Section 2.1.1.2).

2.3. Derive an expression for $-2\log \mathcal{R}$, where \mathcal{R} is the likelihood ratio (2.6), under the one-way random effects model of Example 2.3 for testing $H_0: \sigma^2 = 0$. What is the asymptotic distribution of the likelihood-ratio test, that is, the asymptotic distribution of $-2\log \mathcal{R}$? Study empirically the (asymptotic) size of the likelihood-ratio test and compare it with the nominal levels. For the empirical study, let the true parameters be $\mu = 0.5$ and $\tau^2 = 1.0$; and consider sample sizes $m = 50, 100, 200$ and $k_i = 5$ for all i in all cases.

2.4. Suppose that X_1, \dots, X_n are i.i.d. observations from a population with mean μ and variance σ^2 , and the problem of interest is to estimate μ . A well-known estimator is the sample mean, $\hat{\mu} = \bar{X}$. However, because $\text{var}(\bar{X}) = \sigma^2/n$, in order to evaluate the precision of $\hat{\mu}$, one needs knowledge about σ^2 . Show that an EMM estimator of σ^2 is given by $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, which is the same as the ML estimator when the data are normal.

2.5. Consider a linear regression model

$$y_i = x_i' \beta + \epsilon_i, \quad i = 1, \dots, n,$$

where $x_i = (x_{i1}, \dots, x_{ip})'$ is a vector of known covariates; β is a vector of unknown regression coefficients that are of main interest; and $\epsilon_1, \dots, \epsilon_n$ are i.i.d. errors with mean 0 and variance σ^2 . The model can be expressed as $y = X\beta + \epsilon$, where the i th row of X is x_i' . Assume that $\text{rank}(X) = p$. Then, the least squares (LS) estimator of β is given by

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Although β is of main interest, because $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, to find the standard errors of the estimators one needs knowledge about σ^2 . Show that an EMM estimator of σ^2 is $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (y_i - x_i' \hat{\beta})^2$, which, again, is the ML estimator when normality is assumed.

2.6. Show that the estimating function $M(\beta, \sigma^2, \kappa, y)$ defined above (2.16) is unbiased in the sense that $E\{M(\beta, \sigma^2, \kappa, y)\} = 0$ when β, σ^2, κ correspond to the true parameters.

2.7. Show that the EMM estimators derived in closed form in Example 2.2 (continued) below Lemma 2.1 are consistent, provided that $m \rightarrow \infty$ and $k \geq 2$. You may assume that $\hat{\sigma}_0^2$ and $\hat{\sigma}_1^2$ are the REML estimators and that they are consistent.

2.8. Show that, in the balanced one-way random effects model with the Hartley–Rao form of variance components, the POQUIM estimator of the

asymptotic variance of the REML estimator of γ , that is, the diagonal element of the POQUIM estimator of the asymptotic covariance matrix of the REML estimator corresponding to $\hat{\gamma}$, is given by $\hat{\Sigma}_{R,11}$ in Example 2.2 (Continued) in Section 2.1.2.2.

2.9. This and the next three exercises concern Example 2.2 (Continued) in Section 2.1.2.4. Verify the expression for the Gaussian log-likelihood, $l(\psi, y)$, given there. Show that $E(\text{MSA}) = 1 + k\gamma$, therefore, under the null hypothesis, the probability approaches one as $m \rightarrow \infty$, so that the estimator $\hat{\phi}_2$ is well defined.

2.10. Continuing with the previous exercise, verify that the W-test statistic for $H_0: \lambda = 1$ and $\gamma > 1$ is given by

$$\hat{\chi}_w^2 = \left(\frac{2k}{k-1} + \hat{\kappa}_0 \right)^{-1} mk(\text{MSE} - 1)^2,$$

where $\hat{\kappa}_0$ may be chosen as the EMM estimator of κ_0 given in Example 2.2 (Continued) below Lemma 2.1. Also show that $2k/(k-1) + \kappa_0 > 0$ unless ϵ_{11}^2 is a constant with probability one.

2.11. Continuing with the previous exercise, show that the S-test statistic is identical to the W-test statistic in this case.

2.12. Continuing with the previous exercise, show that the L-test statistic is equal to

$$-2 \log R = m(k-1)\{\text{MSE} - 1 - \log(\text{MSE})\}$$

in this case. Furthermore, show that the asymptotic null distribution of the test statistic is $\lambda_1 \chi_1^2$, where $\lambda_1 = 1 + \{(k-1)/2k\}\kappa_0$, which is estimated by $1 + \{(k-1)/2k\}\hat{\kappa}_0$. Note that the asymptotic null distribution is χ_1^2 if the errors are normal but regardless of the normality of the random effects. (Hint: Use Theorem 2.5.)

2.13. Consider the balanced one-way random effects model of Example 2.2. Consider the Hartley–Rao form of variance components $\lambda = \tau^2$ and $\gamma = \sigma^2/\tau^2$. Suppose that one is interested in constructing an exact confidence interval for γ . Consider the following quantity

$$F = \frac{\text{MSA}}{(1 + k\gamma)\text{MSE}},$$

where $\text{MSA} = \text{SSA}/(m-1)$ and $\text{MSE} = \text{SSE}/m(k-1)$. Show that, under normality, F has an F-distribution with $m-1$ and $m(k-1)$ degrees of freedom. Furthermore, show that, given ρ ($0 < \rho < 1$), an exact $(1 - \rho)\%$ confidence interval for γ is

$$\left[\frac{1}{k} \left(\frac{R}{F_U} - 1 \right), \frac{1}{k} \left(\frac{R}{F_L} - 1 \right) \right],$$

where $R = \text{MSA}/\text{MSE}$, $F_L = F_{m-1, m(k-1), 1-\rho/2}$ and $F_U = F_{m-1, m(k-1), \rho/2}$.

2.14. Consider the one-way random effects model of Example 2.3. Let c_{ij} , $1 \leq j \leq k_i$ be constants such that $\sum_{j=1}^{k_i} c_{ij} = 0$ and $\sum_{j=1}^{k_i} c_{ij}^2 = 1 - 1/k_i$. Define $u_i = \bar{y}_i + \sum_{j=1}^{k_i} c_{ij} y_{ij}$, $1 \leq i \leq m$. Prove the following.

a. The random variables u_1, \dots, u_m are independent and normally distributed with mean μ and variance $\sigma^2 + \tau^2$.

b. The quantity $\chi^2 = \sum_{i=1}^m (u_i - \bar{u})^2 / (\sigma^2 + \tau^2)$ is distributed as χ_{m-1}^2 .

2.15. In Exercise 2.14, find an exact confidence interval for τ^2 , the variance of the error ϵ_{ij} .

2.16*. In the balanced one-way random effects model of Example 2.2, it is known that a UMVU estimator of $\zeta = c\lambda_1 + \lambda_2$ is $\hat{\zeta} = cS_1^2 + S_2^2$, where S_1^2 and S_2^2 are MSA and MSE, respectively, defined in Example 1.1 (continued) in Section 1.5.1.1.

a. Show that S_j^2 is a consistent estimator of λ_j , $j = 1, 2$.

b. Show that $(\hat{\zeta} - \zeta) / \sqrt{\text{var}(\hat{\zeta})}$ converges in distribution to the standard normal distribution.

2.17. Show that, in Example 2.8, the BLUE is given by (2.25) and (2.26) and its covariance matrix is given by (2.27). How do these formulae compare with the corresponding expressions under a linear regression model, that is, those for the least squares estimators? and when do the former reduce to the latter?

2.18. Show that, in Section 2.3.1.2, the logarithm of the joint pdf of α and y can be expressed as (2.36). Furthermore, derive Henderson's mixed model equations (2.37).

2.19. For the following linear mixed models determine the order of d_* above (2.39).

a. One-way random effects model (Example 1.1)

b. Two-way random effects model (Example 1.2)

c. Example 2.8, which is a special case of the nested error regression model

2.20. In Example 2.3 (continued) in Section 2.4.1.1, let the true parameters be $\mu = -0.5$, $\sigma^2 = 2.0$, and $\tau^2 = 1.0$. Also, let $m = 100$ and $k_i = 5$, $1 \leq i \leq m$. In the following, the errors are always generated from a normal distribution.

a. Generate the random effects from a normal distribution. Make a Q-Q plot to assess normality of the random effects, using REML estimators of the parameters.

b. Generate the random effects from a double-exponential distribution (with the same variance). Make a Q-Q plot to assess normality of the random effects, using REML estimators of the parameters.

c. Generate the random effects from a centralized-exponential distribution (with the same variance). Here a centralized-exponential distribution is the distribution of $\xi - E(\xi)$, where ξ has an exponential distribution. Make a Q-Q plot to assess normality of the random effects, using REML estimators of the parameters.

d. Compare the plots in a, b, and c. What do you conclude?

2.21. Show that, in Example 2.15, $\rho_n \sim k$ and $\nu_n \sim mk$ as $m \rightarrow \infty$ (k may or may not go to ∞). Also show that, in Example 2.15 (continued) below (2.37), $\eta_n \sim mk$.

2.22. Show that, in Section 2.5.1, under normal hierarchy and when $b = \beta$ and $B \rightarrow 0$, the likelihood (2.55) reduces to the normal likelihood of Section 1.3.1 when the prior for β is a point mass at β .

2.23. Show that, in Section 2.5.1, under normal hierarchy the likelihood (2.55) reduces to the normal restricted likelihood of Section 1.3.21 when the prior for β is noninformative.

2.24. Consider Example 2.18. Let the priors be such that $\sigma^2 \propto 1/\sigma^2$, $\tau^2 \propto 1/\tau^2$, and σ^2, τ^2 independent. Derive the likelihood (2.55) and posterior (2.56). Is the posterior proper (even though the priors are improper)?

2.25. Show that, under normal hierarchy, the posterior of β is multivariate normal with $E(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}(X'V^{-1}y + B^{-1}b)$ and $\text{Var}(\beta|y) = (X'V^{-1}X + B^{-1})^{-1}$. Similarly, the posterior of α is multivariate normal with $E(\alpha|y) = (Z' LZ + G^{-1})^{-1}Z' L(y - Xb)$ and $\text{Var}(\alpha|y) = (Z' LZ + G^{-1})^{-1}$, where $L = R^{-1} - R^{-1}X(B^{-1} + X'R^{-1}X)^{-1}X'R^{-1}$.

2.26. Show that, under normal hierarchy and when $B^{-1} \rightarrow 0$, which corresponds to the case where the prior for β is noninformative, one has $E(\beta|y) \rightarrow (X'V^{-1}X)^{-1}X'V^{-1}y = \tilde{\beta}$, which is the BLUE; similarly, $E(\alpha|y) \rightarrow GZ'V^{-1}(y - X\tilde{\beta})$.



<http://www.springer.com/978-0-387-47941-5>

Linear and Generalized Linear Mixed Models and Their
Applications

Jiang, J.

2007, XIV, 257 p., Hardcover

ISBN: 978-0-387-47941-5