The Importance of Marginal Likelihood Estimation Applied to Mixed-Effects Modeling

Bradley M. Bell

Applied Physics Laboratory
Box 355640, University of Washington, Seattle, WA 98195-5640
phone: (206)-543-6855
e-mail: brad@apl.washington.edu

August 10, 2000

We consider the problem of estimating deterministic parameters in models that also have random parameters. This is often referred to as estimating fixed effects in mixed-effects models, and the random parameters are called random effects. These models lead to a direct representation of the joint likelihood of the data and the random effects. It is often difficult to obtain values of the marginal likelihood, i.e., the integral of the joint likelihood with respect to the random effects. Nonetheless, the marginal likelihood is often essential when estimating fixed effects in a mixed-effects model. We present a pharmacokinetic example to illustrate the importance of marginal likelihood estimates and their Laplace approximations when applied to mixed-effects models.

KEY WORDS: mixed effects; random effects; Laplace approximation; marginal likelihood; standard two-stage; prior distribution; NONMEM.

1. INTRODUCTION

Often when studying pharmaceutical drugs, measurements are taken from multiple individuals. In this setting, fixed effects are model parameters that have the same values for all individuals, and random effects are model parameters that have independent values for each individual. This separation of the parameters leads naturally to mixed-effects statistical models.

If the random effects were known, we could use their values when estimating the fixed effects. Unfortunately, the random effects are not known. We only have a model for their distribution, which is represented by the joint likelihood, i.e., the likelihood

1

of both the random effects and the measurement values. The marginal likelihood is the result of integrating the joint likelihood with respect to the random effects. It is the uncertainty of the values of the random effects that makes the marginal likelihood better than the joint likelihood when used as a criteria for estimating the fixed effects.

We demonstrate the power of marginal likelihood estimation for a particular pharmacokinetic example. We then show how to generalize this example. In our example, there are two fixed effects, one random effect per individual and one measurement per individual. Thus, the total number of fixed and random effects is 2 greater than the number of measurements. Estimating the fixed and random effects by optimizing the joint likelihood for this example corresponds to solving an under-determined problem by including prior information about the random effects. We show that optimizing the marginal likelihood yields a much better estimate.

Our specific example is a special case of a general Gaussian mixed-effects model with two levels of random variation. The two levels refer to between-individual and within-individual random variation, i.e., the random effects and the measurement noise. In this general setting, the marginal likelihood is expressed as an integral and can be very difficult to calculate. In this case, we suggest replacing the integral by a Laplace approximation, which is a more tractable calculation.

The standard two-stage procedure separates the estimation of the random and fixed effects into two stages. During the first stage, the fact that the individuals are related to each other is ignored, and the random effects are estimated separately for each individual. During the second stage, the estimated values for random effects are treated as if they were data, and the fixed effects are estimated. In Section 4..1, we present the result of the standard two-stage method for our particular example. The corresponding estimate of the fixed effects is biased, and the bias does not reduce as the number of individuals increases.

It is tempting to simultaneously optimize the joint likelihood with respect to both the fixed and random effects. This optimization skips the integration step in the definition of the marginal likelihood and therefore is easier to implement. This procedure is often justified on the basis of adding prior information about the random effects in order to estimate an under-determined system. We show that the corresponding estimate of the fixed effects may not even exist; i.e., there will be no minimizer of the joint likelihood that makes sense as an estimate of the fixed effects.

Our notation is similar to that in Section 4.2 of Davidian and Giltinan [1]. We

make many references to this book because it contains supporting information in a similar notation and not because it is the original source of the corresponding results. Other discussions of general Gaussian mixed-effects models can be found near Eq. (7.4.6) of ref. [2] or Eq. (3.9) of ref. [3]. The Laplace approximation for the marginal likelihood and some of its modifications are discussed in ref. [4]. Estimation methods corresponding to the Laplace approximation and some of its modifications are available using the NONMEM [5] computer program.

2. GENERAL MODEL

In this section, we present a general Gaussian mixed-effects model with two levels of random variation. A particular choice of model parameters and functions is the scientific or engineering part of the modeling process. The goal of this choice is to make the corresponding set of model assumptions a close approximation of the physical system being modeled. Once chosen, the parameters and functions can be combined with mathematical analysis and numerical computation to obtain conclusions implied by the model. These conclusions can be tested using experimentation and the scientific method.

Our notation and model structure are similar to the presentation in Section 4.2 of ref. [1]. In our notation, there are M individuals and $n(i)$ measurements corresponding to the ith individual. Fixed effects are parameters that have the same value for all individuals. For example, the average over the population of the elimination rate of a drug is a fixed effect. Random effects are parameters that have an independent value for each individual. For example, the difference between the average elimination rate and the elimination rate for a particular individual is the value of a random effect for that individual. The true, but unknown, fixed effects are the elements of the vector α^*. The true, but unknown, random effects for the ith individual are the elements of the vector b_i^*. Random effects are sometimes referred to as the inter-individual or between-individual variation (see Eq. (4.14) of ref. [1]).

The measurement vector corresponding to the ith individual is denoted by y_i. The measurement noise corresponding to the ith individual is denoted by e_i. This noise is sometimes referred to as the intra-individual or within-individual variation (see Eq. (4.13) of ref. [1]). The model for the mean of y_i is denoted by a function f_i. This function depends on both the fixed and the random effects. The model for the value

of y_i is

$$y_i = f_i(\alpha^*, b_i^*) + e_i \tag{1}$$

This model is the same as Eq. (4.2) of ref. [1] with the following exceptions: The arguments to the function f_i are the fixed and random effects instead of intermediate parameters. Known values, such as measurement time, have been included in the definition of f_i. (In Eqs. (4.12) and (4.1) of ref. [1], intermediate parameters are denoted by β_i, and known values such as measurement time are referred to as covariates and denoted by $x_{i,j}$.)

The model for the covariance of the random effects is denoted by the function D. This function is a mapping from the fixed effects into the space of real symmetric positive definite matrices. It models the covariance of b_i^* prior to knowing the measurement sequence $(y_i)_{i=1}^M$. The model for the covariance of the measurement noise for the ith individual is denoted by the function R_i. This function is a mapping from the fixed effects and random effects into the space of real symmetric positive definite matrices. It models the covariance of e_i given the value of b_i^*. Our statistical assumptions are that the random vectors $(e_i)_{i=1}^M$, $(b_i^*)_{i=1}^M$ are mutually independent and that

$$
\begin{aligned}
e_i &\sim \mathbf{N}[0, R_i(\alpha^*, b_i^*)] \\
b_i^* &\sim \mathbf{N}[0, D(\alpha^*)]
\end{aligned}
$$

These statistical assumptions are the same as those in Eqs. (4.13) and (4.14) of ref. [1]. Our model is slightly different because we have allowed f_i, R_i, and D to be arbitrary functions of a single vector containing all the fixed effects.

Table I contains the basic building blocks of our Gaussian mixed-effects model. The other terms in this model are defined using the terms in this table. The space of positive integers is denoted by $\mathbf{Z}_+$, the space of real vectors with n components is denoted by $\mathbf{R}^n$, the space of $n \times n$ real matrices is denoted by $\mathbf{R}^{n \times n}$, and an arrow points from the domain space to the range space of the corresponding function.

The values α^*, $(b_i^*)_{i=1}^M$, and $(e_i)_{i=1}^M$ are unknown. For the purposes of this paper, the rest of the information in Table I is known. In this paper, we consider methods for estimating α^* from this known information. Deciding to use a mixed-effects model with a particular choice of parameters and corresponding functions is a model-building activity and is not covered in this paper.

Table I:

$M \in \mathbf{Z}_+$	number of individuals
$q \in \mathbf{Z}_+$	number of fixed effects
$k \in \mathbf{Z}_+$	number of random effects per individual
$n(i) \in \mathbf{Z}_+$	number of measurements for ith individual
$\alpha^* \in \mathbf{R}^q$	fixed effects
$b_i^* \in \mathbf{R}^k$	random effects for ith individual
$y_i \in \mathbf{R}^{n(i)}$	measurement vector for ith individual
$e_i \in \mathbf{R}^{n(i)}$	noise vector for ith individual
$f_i : \mathbf{R}^q \times \mathbf{R}^k \to \mathbf{R}^{n(i)}$	model for expected value of y_i given b_i^*
$R_i : \mathbf{R}^q \times \mathbf{R}^k \to \mathbf{R}^{n(i) \times n(i)}$	model for covariance of y_i given b_i^*
$D : \mathbf{R}^q \to \mathbf{R}^{k \times k}$	model for covariance of b_i^* prior to knowing y_i

3. LAPLACE APPROXIMATION

In this section, we present the marginal likelihood and its Laplace approximation for our general model. Define

$$
\begin{aligned}
K &= kM \\
N &= n(1) + \ldots + n(M)
\end{aligned}
$$

We use the notation

$$
L : \mathbf{R}^q \times \mathbf{R}^K \times \mathbf{R}^N \to \mathbf{R}
$$

to denote the joint negative log-likelihood corresponding to a parameter value α, a random-effects sequence $(b_i)_{i=1}^M$, and a data sequence $(y_i)_{i=1}^M$. The notation $b \in \mathbf{R}^K$ and $y \in \mathbf{R}^N$ is connected to the notation $(b_i)_{i=1}^M$ and $(y_i)_{i=1}^M$ by

$$
\begin{aligned}
b &= (b_1^T, \cdots, b_M^T)^T \\
y &= (y_1^T, \cdots, y_M^T)^T
\end{aligned}
$$

Given the statistical assumptions above, and the density function for a Gaussian distribution (see Theorem 2.3.1 of ref. [6]), we obtain

$$
\begin{aligned}
L(\alpha, b; y) &= \frac{1}{2} \sum_{i=1}^M \log \det \left[2\pi R_i(\alpha, b_i) \right] \\
&+ \frac{1}{2} \sum_{i=1}^M \left[y_i - f_i(\alpha, b_i) \right]^T R_i(\alpha, b_i)^{-1} \left[y_i - f_i(\alpha, b_i) \right] \\
&+ \frac{1}{2} \sum_{i=1}^M \log \det \left[2\pi D(\alpha) \right] + \frac{1}{2} \sum_{i=1}^M b_i^T D(\alpha)^{-1} b_i \qquad (2)
\end{aligned}
$$

Note that $\exp[-L(\alpha, b^*; y)]$ is the joint probability density for the combination of a data sequence $(y_i)_{i=1}^M$ and a random-effects sequence $(b_i^*)_{i=1}^M$. We use $p : \mathbf{R}^N \times \mathbf{R}^q \to \mathbf{R}$

5

to denote the probability density for the data sequence $(y_i)_{i=1}^M$ prior to knowing the value of the random-effects sequence $(b_i^*)_{i=1}^M$. This probability density equals the marginal density of $(y_i)_{i=1}^M$ and is given in Eq. (4.20) of ref. [1], Section 2.2.2 of ref. [6], or Eq. (3.1) of ref. [3] as

$$p(y; \alpha) = \int_{-\infty}^{+\infty} \exp[-L(\alpha, b; y)] db \qquad (3)$$

We say the model has *first-order random effects* if the following conditions hold:

$$\partial_b^{(i)} R_i(\alpha, b_i) = 0 \quad \text{for } i = 1, \ldots, M, \text{ and all } \alpha, b_i$$
$$\partial_b^{(i)} \partial_b^{(i)} f_i(\alpha, b_i) = 0 \quad \text{for } i = 1, \ldots, M, \text{ and all } \alpha, b_i \qquad (4)$$

where $\partial_b^{(i)}$ denotes the partial derivative with respect to b_i. These conditions are equivalent to $R(\alpha, b_i)$ being constant with respect to b_i and changes in $f(\alpha, b_i)$ being linear with respect to changes in b_i. If the model has first-order random effects, the integral in Eq. (3) can be evaluated analytically (see pages 83–85 of ref. [1]). The example defined in Section 4. satisfies this condition.

Evaluating the integral in Eq. (3) can be difficult. One approach is to replace the integral by a Laplace approximation. The resulting approximate density $\tilde{p}$ has the same domain space and range space as p and is given in Sections 4.2 and 4.6 of ref. [7] as

$$\tilde{p}(y; \alpha) = \det \left[\partial_b^2 L(\alpha, \hat{b}(\alpha); y) / (2\pi) \right]^{-1/2} \exp \left[-L(\alpha, \hat{b}(\alpha); y) \right] \qquad (5)$$

where $\hat{b} : \mathbf{R}^q \to \mathbf{R}^K$ is defined by

$$\hat{b}(\alpha) = \text{argmin } L(\alpha, b; y) \text{ with respect to } b \qquad (6)$$

The value $\hat{b}(\alpha)$ is the mode of b^* given y where α^* has been replaced by an approximation denoted by α. On page 170 of ref. [1], this estimate of b^* is referred to as the empirical Bayes estimate. It is sometimes referred to as the maximum a posteriori (MAP) Bayesian estimate to distinguish it from the expected value of b^* given y. The dependence of $\hat{b}$ on α is explicit while its dependence on y is suppressed because it is not needed in our presentation. The Laplace marginal likelihood estimate of α^* is the value of α that maximizes $\tilde{p}(y; \alpha)$.

4. EXAMPLE MODEL

Our example is a one-compartment model. We chose this model because it is simple and has been used in pharmacokinetic applications (see Figure 2-2 of ref. [8]).

In addition, we make some assumptions about our knowledge of the system to simplify the mathematical analysis.

We take one concentration measurement for each individual and we scale our units so that the measurement is taken when time equals 1. Thus, the time interval from the injection to the measurement is the same for all individuals. In addition, we use B_i to denote the amount of drug injected into the ith individual at time zero, which also equals the volume of distribution for that individual. Thus, the initial concentration is 1 for all individuals. Our goal is to estimate the expected value and variance of the rate at which the drug is leaving the system. Note that we take only one measurement per individual and our goal is to estimate two fixed effects. Figure 1 is a diagram of our model:
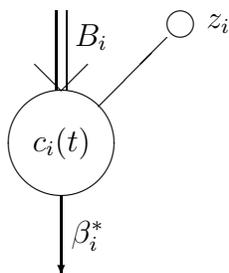
Fig. 1. Model diagram.

Here, i is the index for the individual, B_i is the amount of drug injected, $c_i(t)$ is the concentration of drug in plasma at time t, z_i is a single measurement of the drug concentration, and β_i^* is the true, but unknown, transfer rate from the plasma to the outside world. Because the initial concentration equals 1, the solution of the corresponding differential equation is

$$c_i(t) = \exp(-\beta_i^* t)$$

For the purposes of our example, the concentration measurements are log normally distributed in the following manner:

$$z_i = \exp(-\beta_i^* t_i) \exp(-e_i)$$

where $t_i = 1$ is the time of the measurement for the ith individual and

$$e_i \sim \mathbf{N}(0, 1)$$

Note that in terms of the measurements $(z_i)_{i=1}^M$, the errors in our model are multiplicative. Taking the log of both sides, we obtain the equation

$$\log(z_i) = -\beta_i^* t_i - e_i = -\beta_i^* - e_i$$

because $t_i = 1$ for all i. We make the following correspondence between this particular example and our general model:

$$
\begin{array}{lllllll}
q & = & 2 & , & \alpha_1^* & = & \text{expected value of } \beta_i^* \text{ prior to } y_i \\
k & = & 1 & , & \alpha_2^* & = & \text{variance of } \beta_i^* \text{ prior to } y_i \\
n(i) & = & 1 & , & b_i^* & = & \beta_i^* - \alpha_1^* \\
& & & & y_i & = & -\log(z_i)
\end{array}
\quad
\begin{array}{llll}
, & f_i(\alpha, b_i) & = & \alpha_1 + b_i \\
, & R_i(\alpha, b_i) & = & 1 \\
, & D(\alpha) & = & \alpha_2 \\
&&&
\end{array}
$$

Using this correspondence, Eq. (1) becomes

$$y_i = \alpha_1^* + b_i^* + e_i \tag{7}$$

As per our general model assumptions, the random variables $(e_i)_{i=1}^M$ and $(b_i^*)_{i=1}^M$ are independent. It follows directly from Eq. (7) that the random variables $(y_i)_{i=1}^M$ are independent. In addition, the distribution of y_i, prior to knowing the value of b_i^*, is given by

$$y_i \sim \mathbf{N}(\alpha_1^*, \alpha_2^* + 1) \tag{8}$$

It follows that the expected value of the sample variance of y_i satisfies the equation

$$\mathbf{E}\left[\left(y_i - \frac{1}{M}\sum_{j=1}^M y_j\right)^2\right] = \frac{M-1}{M}(\alpha_2^* + 1) \tag{9}$$

Note that the sample variance corresponds to the actual variance, $\alpha_2^* + 1$, with a correction factor for the degrees of freedom (see the discussion below Theorem 3.3.2 of ref. [6]).

4..1 Standard Two-Stage Estimation

In the standard two-stage procedure, the parameter estimate for each individual is calculated independently, and then the sample variance of that estimate is used as an estimate of the variance of the random effects. This procedure is widely used because it is the simplest way to estimate fixed effects in a mixed-effects model. Section 5.3.1 of ref. [1] describes the the standard two-stage procedure in further detail. In this section, we apply the standard two-stage procedure to the example in Section 4.. To be specific, we define

$$\beta_i^* = \alpha_1^* + b_i^*$$

8

and then we define the estimate β_i^S of β_i^* as the minimizer of the negative log-likelihood of the probability density for y_i corresponding to a value for β_i^*. For our particular example,

$$
\begin{aligned}
y_i &= \beta_i^* + e_i \\
\beta_i^S &= \operatorname{argmin} \frac{1}{2}\log(2\pi) + \frac{1}{2}(y_i - \beta_i)^2 \text{ with respect to } \beta_i
\end{aligned}
$$

It follows that $\beta_i^S = y_i$. The fixed effect α_1^* is the expected value of β_i^* prior to knowing the value of y_i. The standard two-stage estimate for α_1^* is

$$
\alpha_1^S = \frac{1}{M}\sum_{i=1}^{M}\beta_i^S = \frac{1}{M}\sum_{i=1}^{M}y_i
$$

The fixed effect α_2^* is the variance of β_i^*. The standard two-stage estimate for α_2^* is

$$
\alpha_2^S = \frac{1}{M-1}\sum_{i=1}^{M}(\beta_i^S - \alpha_1^S)^2 = \frac{1}{M-1}\sum_{i=1}^{M}\left(y_i - \frac{1}{M}\sum_{j=1}^{M}y_j\right)^2
$$

It follows from Eq. (9) that the expected value of α_2^S is $\alpha_2^* + 1$. Thus, α_2^S a poor estimate of α_2^* unless α_2^* is large relative to 1.

4..2 Joint Likelihood

It is tempting to consider the random effects to be additional parameters, with a prior distribution specified by D, and to estimate the fixed and random effects by optimizing the joint likelihood. The joint likelihood is the probability density for a value of the measurement sequence $(y_i)_{i=1}^{M}$ and the random-effects sequence $(b_i^*)_{i=1}^{M}$. The general case for the negative log of this likelihood is given by L in Eq. (2). In this section, we obtain formulas for L and its derivatives that correspond to the particular example presented in Section 4..

The joint negative log-likelihood function corresponding our particular example is the function $L : \mathbf{R}^2 \times \mathbf{R}^M \times \mathbf{R}^M \to \mathbf{R}$, given by

$$
L(\alpha, b; y) = \frac{M}{2}\log(2\pi) + \frac{1}{2}\sum_{i=1}^{M}(y_i - \alpha_1 - b_i)^2 + \frac{M}{2}\log(2\pi\alpha_2) + \frac{1}{2\alpha_2}\sum_{i=1}^{M}b_i^2
$$

We use $\partial_b^{(i)}$ and $\partial_\alpha^{(i)}$ to denote the partial with respect to the ith component of b and α, respectively. Expressions for some partial derivatives of L are listed below so that they can be derived from the equation above. They will be used later in the text:

$$
\partial_b^{(i)} L(\alpha, b; y) = b_i(1 + 1/\alpha_2) + \alpha_1 - y_i \tag{10}
$$

9

$$\partial_\alpha^{(1)} L(\alpha, b; y) = \sum_{i=1}^{M} (\alpha_1 + b_i - y_i) \tag{11}$$

$$\partial_\alpha^{(2)} L(\alpha, b; y) = \frac{M}{2\alpha_2} - \frac{1}{2\alpha_2^2} \sum_{i=1}^{M} b_i^2 \tag{12}$$

It follows from Eq. (10) that $\partial_b^{(i)} \partial_b^{(j)} L$ is zero whenever $i \neq j$. Thus,

$$\partial_b^2 L(\alpha, b; y) = (1 + 1/\alpha_2) I_M \tag{13}$$

where I_M is the $M \times M$ identity matrix.

4..3 Joint Likelihood Estimation

In this section, we consider estimating the fixed effects in our particular example by optimizing the joint likelihood, i.e., by solving the problem

$$\text{minimize } L(\alpha, b; y) \text{ with respect to } (\alpha, b) \tag{14}$$

The function $\hat{b}$ is defined by Eq. (6). Combining Eq. (10) with the first-order necessary condition for a minimum of $L(\alpha, b; y)$ with respect to b, we obtain

$$0 = \partial_b L(\alpha, \hat{b}(\alpha); y) \tag{15}$$

$$\hat{b}_i(\alpha) = \frac{y_i - \alpha_1}{1 + 1/\alpha_2} \tag{16}$$

We define α^J to be the estimate of α^* that solves the problem

$$\text{minimize } L(\alpha, \hat{b}(\alpha); y) \text{ with respect to } \alpha \tag{17}$$

It follows that the pair $(\alpha^J, \hat{b}(\alpha^J))$ solves Problem 14. If we could solve Problem 17, the solution would satisfy the first-order necessary conditions

$$\begin{aligned}
0 &= \partial_\alpha L(\alpha^J, \hat{b}(\alpha^J); y) + \partial_b L(\alpha^J, \hat{b}(\alpha^J); y) \partial_\alpha \hat{b}(\alpha^J) \\
&= \partial_\alpha L(\alpha^J, \hat{b}(\alpha^J); y)
\end{aligned} \tag{18}$$

Note that the second equality above follows from Eq. (15). Using Eq. (11) and the $\partial_\alpha^{(1)}$ component of the equation above, we conclude that

$$0 = \sum_{i=1}^{M} \alpha_1^J + b_i(\alpha^J) - y_i = \sum_{i=1}^{M} \frac{\alpha_1^J - y_i}{\alpha_2^J + 1}$$

$$\alpha_1^J = \frac{1}{M} \sum_{i=1}^{M} y_i \tag{19}$$

10

Using Eq. (12) and the $\partial_\alpha^{(2)}$ component of Eq. (18), we conclude that

$$0 = \frac{M}{2\alpha_2^J} - \frac{1}{2(\alpha_2^J)^2}\sum_{i=1}^{M}\hat{b}_i(\alpha^J)^2 = M - \frac{1}{\alpha_2^J}\sum_{i=1}^{M}\left(\frac{y_i - \alpha_1^J}{1 + 1/\alpha_2^J}\right)^2$$

$$\frac{1}{M}\sum_{i=1}^{M}(y_i - \alpha_1^J)^2 = \alpha_2^J(1 + 1/\alpha_2^J)^2 = \alpha_2^J + 2 + 1/\alpha_2^J \tag{20}$$

Only positive values of α_2^J make sense because it is an estimate of α_2^*, which is a variance. In addition, the right-hand side of Eq. (20) is always greater than 4 (for positive values of α_2^J). Using Eqs. (9) and (19), we conclude that

$$\mathbf{E}\left[(y_i - \alpha_1^J)^2\right] = (1 - 1/M)(\alpha_2^* + 1)$$

Thus, by the law of large numbers, if $\alpha_2^* < 3$, the probability that there is a value of α_2^J that satisfies Eq. (20) goes to zero as M goes to infinity. In other words, it is unlikely that there is a solution to either Problem 14 or Problem 17.

4..4 Marginal Likelihood Estimation

Our particular example has first-order random effects; i.e., the conditions in Eq. (4) hold. It follows from Lemma 2 in the appendix that

$$-\log[\tilde{p}(y;\alpha)] = \frac{1}{2}\sum_{i=1}^{M}\log\det[2\pi V_i(\alpha)] + [y_i - f_i(\alpha, 0)]^T V_i(\alpha)^{-1}[y_i - f_i(\alpha, 0)]$$

where $V_i(\alpha) = 1 + \alpha_2$. We used the general definition of V_i in Eq. (A1) and the specific definitions of R_i, f_i and D for our example (see Section 4.). Further simplifying using the specific forms for f_i and $V_i(\alpha)$, we obtain

$$-\log[\tilde{p}(y;\alpha)] = \frac{M}{2}\log[2\pi(1 + \alpha_2)] + \frac{1}{2(1 + \alpha_2)}\sum_{i=1}^{M}(y_i - \alpha_1)^2 \tag{21}$$

Note that, by Eq. (8) and the independence of the random variables $(y_i)_{i=1}^{M}$, the right hand side of Eq. (21) is the likelihood of $(y_i)_{i=1}^{M}$ prior to any knowledge of the sequence $(b_i^*)_{i=1}^{M}$. The Laplace marginal likelihood estimate $\hat{\alpha}$ for α^* is the value of α that maximizes $\tilde{p}(y;\alpha)$. It follows from simple calculus (or Theorem 3.2.1 of ref. [6]) that

$$\hat{\alpha}_1 = \frac{1}{M}\sum_{i=1}^{M}y_i$$

$$\hat{\alpha}_2 + 1 = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{\alpha}_1)^2$$

5. CONCLUSION

For our example, all three of the methods presented above have the same estimate for α_1^*:

$$\alpha_1^S = \alpha_1^J = \hat{\alpha}_1 = \frac{1}{M}\sum_{j=1}^{M} y_j$$

The standard two-stage estimate for α_2^* is given by

$$\alpha_2^S \;=\; \frac{1}{M-1}\sum_{i=1}^{M}\left(y_i - \frac{1}{M}\sum_{j=1}^{M} y_j\right)^2$$

The expected value of this estimate is given by

$$\mathbf{E}[\alpha_2^S] \;=\; \alpha_2^* + 1$$

The marginal likelihood estimate for α_2^* is given by

$$\hat{\alpha}_2 \;=\; -1 + \frac{1}{M}\sum_{i=1}^{M}\left(y_i - \frac{1}{M}\sum_{j=1}^{M} y_j\right)^2$$

The expected value of this estimate is given by

$$\mathbf{E}[\hat{\alpha}_2] \;=\; \alpha_2^* - \alpha_2^*/M$$

In addition, we showed that the joint likelihood estimate α_2^J will often not even exist when α_2^* is less than 3. We note that as the number of individuals increases, the absolute bias $|\alpha_2^* - \mathbf{E}[\hat{\alpha}_2]|$ decreases. On the other hand, the absolute bias $|\alpha_2^* - \mathbf{E}[\alpha_2^S]|$ is independent of the number of individuals.

Our analysis of the example demonstrates that the Laplace marginal likelihood estimate is superior to the standard two-stage and joint likelihood estimates. The standard two-stage estimate α_2^S is biased because it does not account for the difference between the estimate β_i^S and the true value β_i^*. The Global Two Stage method defined in Section 5.3.2 of ref. [1] uses an approximation for

$$\mathbf{E}\left[(\beta_i^S - \beta_i^*)(\beta_i^S - \beta_i^*)^T\right] \tag{22}$$

to account for this difference. This approximation is exact and this variance is constant with respect to b for our example because it has first-order random effects; i.e., the conditions in Eq. (4) hold. In general nonlinear cases, the Global Two Stage approximation for Expression (22) is not constant with respect to $\hat{b}(\alpha)$, and the corresponding method becomes an iterative procedure. This procedure can be very slow

and may not even converge because the change in the approximation is not accounted for when deciding how to change the current value of α. The EM method defined in Section 5.3.2 of ref. [1] and the Lindstrom Bates method defined in Section 6.3.2 of ref. [1] may be slow to converge for similar reasons. One intuitive way to view these methods is that the total derivative with respect to α of the expression

$$\det \left[\partial_b^2 L(\alpha, \hat{b}(\alpha); y)/(2\pi) \right]^{-1/2} \tag{23}$$

is not computed before deciding how to change α. (This expression is part of the formula for $\tilde{p}(\alpha; y)$ in Eq. (5).) Not computing this total derivative makes these methods easier to implement, but they are slower to converge, and sometimes even fail to converge, for general nonlinear cases. The first-order objective in Eq. (6.5) of ref. [1] replaces the expression above with

$$\det \left[\partial_b^2 L(\alpha, 0; y)/(2\pi) \right]^{-1/2}$$

and also approximates f_i as linear with respect to b_i and R_i as constant with respect to b_i. (These approximations are exact when the model has first-order random effects.) Thus, it also does not take the total derivative of the expression in Eq. (23) into account during its optimization procedure.

Formulas that account for the total derivative of the expression in Eq. (23) are available in ref. [4]. Estimation procedures that account for these derivatives are available using the NONMEM [5] computer program.

APPENDIX

In this appendix, we consider the general model defined in Section 2. and do not restrict our attention to the example presented in Section 4.. We show that the results in Section 4..4 hold for any model that has first-order random effects, i.e., when the conditions in Eq. (4) hold.

The following lemma can be proven using the argument on page 83 of ref. [1]:

Lemma 1 *Suppose that $B \in \mathbf{R}^{k \times k}$, $R \in \mathbf{R}^{K \times K}$, $F \in \mathbf{R}^{K \times k}$, and $w \in \mathbf{R}^K$ where B and R are symmetric and positive definite. Define $h : \mathbf{R}^k \to \mathbf{R}$ by*

$$
\begin{aligned}
h(x) &= \frac{1}{2} \log \det(2\pi R) + \frac{1}{2} \log \det(2\pi B) \\
&\quad + \frac{1}{2} x^T B^{-1} x + \frac{1}{2} (w - Fx)^T R^{-1} (w - Fx)
\end{aligned}
$$

It follows that

$$\int_{-\infty}^{+\infty} \exp[-h(x)] dx = \exp \left[-\frac{1}{2} \log \det(2\pi V) - \frac{1}{2} w^T V^{-1} w \right]$$

13

where $V = FBF^T + R$.

Suppose that we have a general model as defined in Section 2. for which the conditions in Eq. (4) hold. We can use Eq. (1) to conclude that

$$y_i = f_i(\alpha^*, 0) + \partial_b^{(i)} f_i(\alpha^*, 0)b_i^* + e_i$$

We define $V_i : \mathbf{R}^q \to \mathbf{R}^{n(i) \times n(i)}$ by

$$V_i(\alpha) = R_i(\alpha, 0) + \partial_b^{(i)} f_i(\alpha, 0)D(\alpha)\partial_b^{(i)} f_i(\alpha, 0)^T \tag{A1}$$

It follows that the distribution of y_i, prior to any knowledge of b_i^*, is

$$y_i \sim N\left[f_i(\alpha, 0), V_i(\alpha)\right] \tag{A2}$$

Equation (A2) corresponds to the result in Eq. (8) for the arbitrary case of a model with first-order random effects.

The following lemma establishes that, for models with first-order random effects, the Laplace approximation for the marginal likelihood equals the likelihood of an independent sequence $(y_i)_{i=1}^M$ with distribution given by Eq. (A2).

Lemma 2 *We are given a model of the form defined in Section 2. that satisfies the first-order random-effects conditions defined in Eq. (4). The Laplace approximation, $\tilde{p}(y; \alpha)$, defined by Eq. (5) satisfies the following equations:*

$$
\begin{aligned}
\tilde{p}(y; \alpha) &= p(y; \alpha) \\
-\log[\tilde{p}(y; \alpha)] &= \frac{1}{2}\sum_{i=1}^M \log \det[2\pi V_i(\alpha)] + [y_i - f_i(\alpha, 0)]^T V_i(\alpha)^{-1}[y_i - f_i(\alpha, 0)]
\end{aligned}
$$

where $p(y; \alpha)$ is given by Eq. (3) and $V_i(\alpha)$ is given by Eq. (A1).

Proof: The first-order random-effects conditions in Eq. (4) imply that $L(\alpha, b; y)$ is quadratic with respect to b. It follows that the marginal density p equals its Laplace approximation $\tilde{p}$; i.e., the first assertion of the lemma is true (see Section 4.6 of ref. [7]).

It follows from Eqs. (2) and (3) that

$$
\begin{aligned}
p(y; \alpha) &= \int_{-\infty}^{+\infty} \exp[-L(\alpha, b; y)]db \\
&= \prod_{i=1}^M \int_{-\infty}^{+\infty} \exp[-L_i(\alpha, b_i; y)]db_i
\end{aligned}
$$

14

where

$$L_i(\alpha, b_i; y) = \frac{1}{2} \log \det \left[2\pi R_i(\alpha, b_i)\right] + \frac{1}{2} \log \det \left[2\pi D(\alpha)\right] + \frac{1}{2} b_i^T D(\alpha)^{-1} b_i$$
$$+ \frac{1}{2} \left[y_i - f_i(\alpha, b_i)\right]^T R_i(\alpha, b_i)^{-1} \left[y_i - f_i(\alpha, b_i)\right]$$

The conclusion of this lemma now follows from a direct application of Lemma 1 with

$$w = y_i - f_i(\alpha, 0) , \ F = \partial_b^{(i)} f_i(\alpha, 0) , \ x = b_i , \ R = R_i(\alpha, 0) , \ B = D(\alpha)$$

Note that we have used the first-order random-effects conditions to make the replacements:

$$f_i(\alpha, b_i) = f_i(\alpha, 0) + \partial_b^{(i)} f_i(\alpha, 0) b_i$$
$$R_i(\alpha, b_i) = R_i(\alpha, 0)$$

ACKNOWLEDGMENT

References

[1] M. Davidian and D.M. Giltinan. *Nonlinear Model for Repeated Measurement Data,* Chapman & Hall, New York, 1995.

[2] E.F. Vonesh and V.M. Chinchilli. *Linear and Nonlinear Models for the Analysis of Repeated Measurements,* Marcel Dekker, New York, 1997.

[3] P.C. Carlin and T.A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis,* Chapman & Hall, New York, 1996.

[4] B.M. Bell. Approximating the marginal likelihood estimate for models with random parameters, *Appl. Math. Comput.,* in press, 2000.

[5] S.L. Beal and L.B. Sheiner (eds.). NONMEM users guide - Part VII: Conditional estimation methods, NONMEN project group, University of California, San Francisco, CA (1998).

[6] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis,* John Wiley & Sons, New York, 1958.

[7] N.G. De Bruijn. *Asymptotic Methods in Analysis,* North Holland, Amsterdam, 1961.

[8] W. Simon. *Mathematical Techniques for Biology and Medicine,* Dover Publications, New York, 1986.

Fig. 1. Model diagram.