Published in final edited form as:

Stat Biosci. 2016 October ; 8(2): 220–233. doi:10.1007/s12561-015-9140-x.

Bivariate Mixed Effects Analysis of Clustered Data with Large Cluster Sizes

Daowen Zhang^{1,*}, Jie Lena Sun², and Karen Pieper²

¹Department of Statistics, North Carolina State University, Raleigh, NC 27695

²Duke Clinical Research Institute, Duke University, Durham, NC 27708

Abstract

Linear mixed effects models are widely used to analyze a clustered response variable. Motivated by a recent study to examine and compare the hospital length of stay (LOS) between patients undertaking percutaneous coronary intervention (PCI) and coronary artery bypass graft (CABG) from several international clinical trials, we proposed a bivariate linear mixed effects model for the joint modeling of clustered PCI and CABG LOS's where each clinical trial is considered a cluster. Due to the large number of patients in some trials, commonly used commercial statistical software for fitting (bivariate) linear mixed models failed to run since it could not allocate enough memory to invert large dimensional matrices during the optimization process. We consider ways to circumvent the computational problem in the maximum likelihood (ML) inference and restricted maximum likelihood (REML) inference. Particularly, we developed an expected and maximization (EM) algorithm for the REML inference and presented an ML implementation using existing software. The new REML EM algorithm is easy to implement and computationally stable and efficient. With this REML EM algorithm, we could analyze the LOS data and obtained meaningful results.

Keywords

Meta Analysis; Missing Data; Multi-center Studies

1 Introduction

For patients with coronary artery diseases, percutaneous coronary intervention (PCI) or coronary artery bypass graft (CABG) are two common procedures recommended by physicians. Among many other factors, the hospital length of stay (LOS) is an important one to evaluate the cost effectiveness of each procedure. While some patients receiving a PCI are discharged from the hospital within 24 hours after the procedure, many PCI patients require longer observation period. Compared to a PCI, the CABG is a much more invasive procedure and the patients with a CABG usually require a longer hospital stay. Because of

* dzhang2@ncsu.edu.

The authors declare no conflict of interest.

the difference in patient care and management across the world, there may be a regional difference in hospital LOS's of these two procedures.

Recently, we have a unique opportunity to examine the hospital LOS's of PCI and CABG procedures from 10 international clinical trials on patients with coronary artery diseases (See Section 7 for more details). One of the research objectives is to quantify the expected LOS's and their differences of the CABG and PCI procedures after adjusting for patients' age, smoking and diabetes statuses, for seven different representative regions across the world, while taking into account the trial-to-trial variation. A common approach used to analyze clustered data such as the LOS data is linear mixed effects models where random effects are used to model the cluster effects of trials or trial-to-trial variation. However, empirical evidence indicated that CABG and PCI LOS's exhibit very different within-trial (residual) variation, which has to be taken into account in the analysis for optimal inference. We hence propose a bivariate linear mixed model for jointly modeling CABG and PCI LOS's. Due to large number of patients in some trials, the commonly used statistical software, the MIXED procedure of SAS (SAS/STAT 9.3, 2013), for fitting linear mixed models, failed to run since it could not allocate enough memory to invert large dimensional matrices. This computational problem motivates the research of this paper.

In this paper, we considered ways to circumvent the computational problem in the maximum likelihood (ML) inference and the restricted maximum likelihood (REML) inference for the proposed bivariate linear mixed model. For the ML inference, we presented an implementation using existing software. For the REML inference, we proposed a novel expected and maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) and proved its theoretical statistical property. The new REML EM algorithm is easy to implement and computationally stable and efficient. With this REML EM algorithm, we could analyze the LOS data and obtained meaningful results.

The paper is organized as follows. In Section 2, we describe the proposed bivariate linear mixed model. In Section 3, we discuss the computational issues and solution in the inference on fixed effects and random effects with large cluster sizes. In Section 4, we discuss the ML inference on variance/covariance components and in Section 5, we discuss their REML inference and present an EM algorithm for the REML inference. We present simulation results in Section 6. The proposed REML EM algorithm was used to fit a bivariate mixed model for the LOS data in Section 7. We conclude the paper in Section 8 with some discussion.

2 Models

Suppose our sample consists of m clusters with 2 response variables. In cluster i , we have observations for response variable Y_1 from n_{1i} subjects and response variable Y_2 from n_{2i} subjects, so that the size of cluster i is $n_i = n_{1i} + n_{2i}$. At the same time, some important covariates were also measured for each subject. We consider the following Laird-Ware (Laird and Ware, 1986) linear mixed model for each response variable:

$$Y_{1ij} = X_{ij}^T \beta_1 + Z_{ij}^T b_{1i} + e_{1ij}, i=1, 2, \dots, m, j=1, 2, \dots, n_{1i},$$

$$Y_{2ij} = X_{i,n_{1i}+j}^T \beta_2 + Z_{i,n_{1i}+j}^T b_{2i} + e_{2ij}, i=1, 2, \dots, m, j=1, \dots, n_{2i}, \quad (1)$$

where X_{ij} (usually including the intercept) is the $\tilde{p}$ -dimensional vector of covariates for fixed effects β_1 and β_2 , Z_{ij} usually a subset of X_{ij} is the $\tilde{q}$ -dimensional vector of covariates for cluster specific random effects b_{1i} and b_{2i} . It is assumed that the random effects vectors b_{1i} and b_{2i} are independent across clusters and have normal distributions with mean zero and variance matrices D_{11} , D_{22} and covariance matrix D_{12} . We assume in this paper that the variance matrix D formed by D_{11} , D_{12} and D_{22} is a positive definite and unstructured matrix, an assumption commonly used in the statistical literature. It is further assumed that e_{1ij} 's and e_{2ij} 's are independent residual errors, independent of b_{1i} , b_{2i} and distributed as $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$ respectively.

Model (1) is particularly applicable for data from multi-center studies where each center is considered as a cluster, or data analyzed in meta analysis when the data from each individual study were available, in which case each study is considered as a cluster. By modeling response variables Y_1 and Y_2 jointly as in (1), we can answer many important scientific questions. For the hospital LOS data example, if we take $X_{ij} = Z_{ij} = 1$, we can then compare the overall LOS's from CABG and PCI procedures by comparing β_1 and β_2 . When we include dummy variables for different representative regions in the world as well as adjusting covariates, we can then estimate the covariate-adjusted CABG LOS's and PCI LOS's, and hence their differences for those regions.

Model (1) can be re-written as a linear mixed effects model for a single response variable. For cluster i , denote by $Y_{1i} = (Y_{1i1}, Y_{1i2}, \dots, Y_{1in_{1i}})^T$, the data vector of response variable 1 and $Y_{2i} = (Y_{2i1}, Y_{2i2}, \dots, Y_{2in_{2i}})^T$ of response variable 2. Similarly we can define the residual error vectors e_{1i} and e_{2i} . Stack $X_{ij}^T (j=1, 2, \dots, n_{1i})$ to form X_{1i} and $X_{i,n_{1i}+j}^T (j=1, 2, \dots, n_{2i})$ to form X_{2i} . Similarly we can form Z_{1i} and Z_{2i} . We then define new response vector $Y_i = (Y_{1i}^T, Y_{2i}^T)^T$, covariate matrices $X_i = \text{diag}\{X_{1i}, X_{2i}\}$, $Z_i = \text{diag}\{Z_{1i}, Z_{2i}\}$, and residual vector $e_i = (e_{1i}^T, e_{2i}^T)^T$. In matrix notation, model (1) can be re-written as a regular linear mixed effects model for the new response vector Y_i

$$Y_i = X_i \beta + Z_i b_i + e_i, i=1, 2, \dots, m, \quad (2)$$

where $\beta = (\beta_1^T, \beta_2^T)^T$ is the $p \times 1$ ($p = 2\tilde{p}$) new fixed effects vector, $b_i = (b_{1i}^T, b_{2i}^T)^T$ is the $q \times 1$ ($q = 2\tilde{q}$) new random effects vector distributed as $N(0, D)$, and $e_i \sim N(0, R_i)$ is the new residual error vector independent of b_i with $R_i = \text{diag}\{\sigma_1^2 I_{n_{1i} \times n_{1i}}, \sigma_2^2 I_{n_{2i} \times n_{2i}}\}$.

The estimation and inference for linear mixed effects models have been thoroughly studied in the statistical literature for the case where cluster sizes are small to moderate, and have been implemented in many statistical softwares such as the MIXED procedure of SAS for routine analyses of many correlated data, including clustered data. However, the cluster sizes n_i 's in the LOS data example for some clusters are so prohibitively large that a general purpose software cannot even allocate enough memory during the computation process. In this paper, we develop estimation and inference procedures for model (2) by taking advantage of the special features of the random effects and residual errors in the model.

3 Inference on Fixed Effects β and Random Cluster Effects b_i

Stack Y_i to form the response vector Y . Stack X_i to form the design matrix X for the fixed effects vector β , and define $Z = \text{diag}\{Z_1, Z_2, \dots, Z_m\}$ for the design matrix of the random effects vector $b = (b_1^T, b_2^T, \dots, b_m^T)^T$. Given D, σ_1^2 and σ_2^2 , the maximum likelihood estimate (MLE) of β is given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y = \left(\sum_{i=1}^m X_i^T V_i^{-1} X_i \right)^{-1} \sum_{i=1}^m X_i^T V_i^{-1} Y_i, \quad (3)$$

where $V = \text{var}(Y) = \text{diag}\{V_1, V_2, \dots, V_m\}$, $V_i = \text{var}(Y_i) = Z_i D Z_i^T + R_i$. For applications where $n_i \gg q$, the calculation of $\hat{\beta}$ can be facilitated by the following expressions

$$V_i^{-1} = R_i^{-1} - R_i^{-1} Z_i (D^{-1} + Z_i^T R_i^{-1} Z_i)^{-1} Z_i^T R_i^{-1}, \quad (4)$$

$$X^T V^{-1} X = \sum_{i=1}^m X_i^T R_i^{-1} X_i - \sum_{i=1}^m X_i^T R_i^{-1} Z_i (D^{-1} + Z_i^T R_i^{-1} Z_i)^{-1} Z_i^T R_i^{-1} X_i, \quad (5)$$

$$X^T V^{-1} Y = \sum_{i=1}^m X_i^T R_i^{-1} Y_i - \sum_{i=1}^m X_i^T R_i^{-1} Z_i (D^{-1} + Z_i^T R_i^{-1} Z_i)^{-1} Z_i^T R_i^{-1} Y_i. \quad (6)$$

Inference of cluster-specific random effects b_i 's can be based on the best linear unbiased predictors (BLUPs) $\hat{b}_i = E(b_i | Y_i; \beta) |_{\beta = \hat{\beta}} = D Z_i^T V_i^{-1} (Y_i - X_i \hat{\beta})$, which can also be obtained by solving the so called "mixed model equation" (Henderson 1984) jointly for β and b :

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ b \end{bmatrix} = \begin{bmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{bmatrix}, \quad (7)$$

where $G = \text{diag}\{D, D, \dots, D\}$ and $R = \text{diag}\{R_1, R_2, \dots, R_m\}$ are the variance-covariance matrices of b and the residual vector. Note that equation (7) can also be derived as the “score equation” by treating b as a parameter vector and maximizing

$f(Y, b; \beta, D, \sigma_1^2, \sigma_2^2) = f(Y|b; \beta, \sigma_1^2, \sigma_2^2)f(b; D)$ jointly with respect to β and b . There are several advantages of using equation system (7). First, the MLE $\hat{\beta}$ and BLUPs $\hat{b}$ can be obtained without inverting any matrix whose dimension is of the magnitude of the cluster sizes n_i 's. When the total number of clusters m is small to moderate such that $p + mq$, the dimension of the equation system (7), is not too large, we can obtain the MLE $\hat{\beta}$ and BLUPs $\hat{b}$ by directly inverting the coefficient matrix of this system. Otherwise, the coefficient matrix can be inverted efficiently by recognizing that $Z^T R^{-1} Z + G^{-1}$ is a block diagonal matrix with the i th block being $Z_i^T R_i^{-1} Z_i + D^{-1}$. Second, once $\hat{\beta}$ and $\hat{b}$ are obtained, the joint inference on β and b can be made easily based on the following variance-covariance expression

$$\text{var} \begin{pmatrix} \hat{\beta} - \beta \\ \hat{b} - b \end{pmatrix} = \begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix}^{-1}. \quad (8)$$

4 Maximum Likelihood Estimation and Inference of D and (σ_1^2, σ_2^2)

In the linear mixed model literature, the inference on the variance-covariance matrix D of the cluster specific random effects b_i 's and residual variances (σ_1^2, σ_2^2) can be carried out using the maximum likelihood (ML) approach or the restricted maximum likelihood (REML) approach. We will discuss the ML implementation in this section and the REML in Section 5.

The ML approach estimates β , D and (σ_1^2, σ_2^2) by jointly maximizing the following log-likelihood function with respect to β , D and (σ_1^2, σ_2^2)

$$\ell(\beta, D, \sigma^2; Y) = -\frac{1}{2} \sum_{i=1}^m \log |V_i| - \frac{1}{2} \sum_{i=1}^m (Y_i - X_i \beta)^T V_i^{-1} (Y_i - X_i \beta), \quad (9)$$

where $|V_i|$ can be calculated using $|V_i| = \sigma_1^{2n_{1i}} \sigma_2^{2n_{2i}} |I_{q \times q} + Z_i^T R_i^{-1} Z_i D|$. The inference of D and (σ_1^2, σ_2^2) can be based on the Fisher information matrix or the observed information matrix from (9).

It is interesting to note that although the software such as the MIXED procedure of SAS routinely used to fit linear mixed model (2) may suffer from computational difficulty, the software such as the NLMIXED procedure of SAS that uses a numerical integration method for likelihood evaluation can easily implement the likelihood inference of β , D and (σ_1^2, σ_2^2) without inverting large matrices. Since $Y_{ij}|b_i$ are independent under the model specification, by definition, the log-likelihood function (9) can be equivalently re-written as

$$\ell(\beta, D, \sigma_1^2, \sigma_2^2; Y) = \sum_{i=1}^m \log \int \exp \left\{ \sum_{j=1}^{n_i} \log f(Y_{ij} | b_i; \beta, \sigma_1^2, \sigma_2^2) + \log f(b_i; D) \right\} db_i. \quad (10)$$

Again by the given model specification, the exponent

$\sum_{j=1}^{n_i} \log f(Y_{ij} | b_i; \beta, \sigma_1^2, \sigma_2^2) + \log f(b_i; D)$ inside the integration in (10) does not involve inverting any large matrix, and is a quadratic function of b_i with a negative definite second derivative matrix. Therefore, the integration can be calculated exactly by the adaptive Gaussian-Hermite quadrature method with only one quadrature point, which is the numerical integration method implemented in the NLMIXED procedure of SAS. Hence, this SAS procedure can be used to conduct the ML inference on β , D and (σ_1^2, σ_2^2) for data from large clusters.

Even though we can use existing software such as the NLMIXED procedure of SAS to calculate the likelihood function $\ell(\beta, D, \sigma_1^2, \sigma_2^2; Y)$, we may still encounter numerical problems since this procedure relies on the numerical differentiation method to calculate the required derivatives in the optimization process. When the total number of parameters in the model is small, the numerical differentiation method may work well. Otherwise, non-negligible numerical errors accumulated over the optimization process will yield poor parameter estimates or cause the optimization process not to converge. To overcome these problems, we can modify the EM algorithm of Laird and Ware (1982) for the ML estimation of a linear mixed model by treating the cluster specific random effects b_i 's as missing data. It is well-known that an EM algorithm is numerically stable, will always increase the observed likelihood function during the parameter update process, and under the given model specification, parameter updates for MLE's have closed form expressions without the need to invert any large dimensional matrices. Readers are referred to Laird and Ware (1982) for more details.

5 Restricted Maximum Likelihood Estimation and Inference of D and (σ_1^2, σ_2^2)

It is well-known that the ML approach for estimating D and (σ_1^2, σ_2^2) presented in the previous section did not account for the estimation of the fixed effects β and hence will produce biased estimates of D and (σ_1^2, σ_2^2) for small to moderate sample size m . This is because when β is profiled out from the log-likelihood (9) during the maximization process, the resulting function of D and (σ_1^2, σ_2^2) alone is not a log-likelihood function of any (transformed) data. Therefore, the estimating equations for D and (σ_1^2, σ_2^2) are biased. Even though the biases in the ML estimates of D and (σ_1^2, σ_2^2) will disappear asymptotically, it may not be negligible for small to moderate sample size m , especially for the estimate of D . Since the MLE's of the fixed effects β depend on the estimates of D and (σ_1^2, σ_2^2) , the biases in these estimates usually will in turn yield more biased estimates of β . On the other hand, the restricted maximum likelihood (REML) approach will usually yield less biased estimates

of D and (σ_1^2, σ_2^2) . For example, in a regular linear regression model, the REML estimate of the residual variance is unbiased.

For the linear mixed model (2), it is well-known that the REML log-likelihood function of D and (σ_1^2, σ_2^2) is given by (Diggle et. al. 2002)

$$\ell_R(D, \sigma_1^2, \sigma_2^2; Y) = -\frac{1}{2} \log |X^T V^{-1} X| - \frac{1}{2} \log |V| - \frac{1}{2} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta}), \quad (11)$$

where $\hat{\beta}$ is the MLE of β given in (3). Compared to the maximum log-likelihood function (9), the REML log-likelihood function involves one extra term $\log |X^T V^{-1} X|$. The expression in (5) for $X^T V^{-1} X$ can be used to calculate this extra term without directly inverting individual variance-covariance matrix V_i .

Although it is computationally feasible to calculate the REML function $\ell_R(D, \sigma_1^2, \sigma_2^2; Y)$ for data with large cluster sizes, direct maximization of this function sometimes may be numerically unstable, even if the random effects dimension q is moderate. Unlike the ML estimation, we cannot adapt an existing software such as the NLMIXED procedure of SAS to implement the REML estimation. Because of the attractive properties of an EM algorithm, we present in the following sub-section an EM algorithm for maximizing the REML log likelihood function $\ell_R(D, \sigma_1^2, \sigma_2^2; Y)$.

5.1 EM Algorithm for the REML Estimation of D and (σ_1^2, σ_2^2)

It is well-known that the REML log-likelihood function $\ell_R(D, \sigma_1^2, \sigma_2^2; Y)$ can be alternatively derived using the following formula (Harville, 1974)

$$L_R(D, \sigma_1^2, \sigma_2^2; Y) = e^{\ell_R(D, \sigma_1^2, \sigma_2^2; Y)} = \int f(Y|b; \beta, \sigma_1^2, \sigma_2^2) f(b; D) d\beta db,$$

where both fixed effects β and random effects b are integrated out from the joint distribution of response Y and random effects b .

Denote by $D^{(t)}$, $\sigma_1^{2(t)}$, and $\sigma_2^{2(t)}$ the estimates of D , σ_1^2 , and σ_2^2 at the t -th iteration of the REML EM algorithm. Theorem 1 in Appendix A, the general EM algorithm for the REML estimation, indicates that the update $D^{(t+1)}$, $\sigma_1^{2(t+1)}$, $\sigma_2^{2(t+1)}$ can be obtained by maximizing the following REML Q -function with respect to D and (σ_1^2, σ_2^2)

$$\begin{aligned} Q_R(D, \sigma_1^2, \sigma_2^2 | D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}) &= E_g \log f(Y|b; \beta, \sigma_1^2, \sigma_2^2) + E_g \log f(b; D) \\ &= -\log |R| - \frac{1}{2} E_g (Y - X\beta - Zb)^T R^{-1} (Y - X\beta - Zb) \\ &\quad - \frac{m}{2} \log |D| - \frac{1}{2} \sum_{i=1}^m E_g (b_i D^{-1} b_i), \end{aligned}$$

where the expectation E_g stands for the expectation taken with respect to the distribution $g(\beta, b|Y; D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$ defined by

$$g(\beta, b|Y; D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}) = f(Y|b; \beta, \sigma_1^{2(t)}, \sigma_2^{2(t)}) f(b; D^{(t)}) / L_R(D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)}; Y).$$

It should be noted that β and b are both treated as random variables in the distribution $g(\beta, b|Y; D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$. By the given model specification, it is easy to see that $g(\beta, b|Y; D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$ is a normal distribution with the mean given by the solution of (7) and the variance matrix given by (8), both of which are evaluated at the current estimates $(D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$. Denote the mean vector of this distribution by $(\hat{\beta}^T, \hat{b}^T)^T$ and the variance matrix by Σ (we suppress the obvious dependence of these quantities on t for a cleaner presentation). Then maximizing $Q_R(D, \sigma_1^2, \sigma_2^2 | D^{(t)}, \sigma_1^{2(t)}, \sigma_2^{2(t)})$ with respect to D and (σ_1^2, σ_2^2) leads to the following updates:

$$D^{(t+1)} = \frac{1}{m} \sum_{i=1}^m \{ \hat{b}_i \hat{b}_i^T + \Sigma_i \},$$

$$\sigma_1^{2(t+1)} = \frac{1}{n_1} \left[\sum_{i=1}^m (Y_{1i} - X_{1i} \hat{\beta}_1 - Z_{1i} \hat{b}_{1i})^T (Y_{1i} - X_{1i} \hat{\beta}_1 - Z_{1i} \hat{b}_{1i}) + \text{tr}\{\Sigma W_1^T W_1\} \right],$$

$$\sigma_2^{2(t+1)} = \frac{1}{n_2} \left[\sum_{i=1}^m (Y_{2i} - X_{2i} \hat{\beta}_2 - Z_{2i} \hat{b}_{2i})^T (Y_{2i} - X_{2i} \hat{\beta}_2 - Z_{2i} \hat{b}_{2i}) + \text{tr}\{\Sigma W_2^T W_2\} \right],$$

where Σ_i is the block of Σ corresponding to the random effects b_i , W_1 and W_2 are sub-matrices of (X, Z) corresponding to responses Y_1 and Y_2 respectively, and

$n_1 = \sum_{i=1}^m n_{1i}$, $n_2 = \sum_{i=1}^m n_{2i}$, total numbers of patients for responses Y_1 and Y_2 . We can take advantage of the special structure of Σ described in Section 3 to facilitate the calculation of the updates $D^{(t+1)}$, $\sigma_1^{2(t+1)}$ and $\sigma_2^{2(t+1)}$.

It is clear that this EM algorithm for the REML estimation is easy to implement since it yields explicitly closed form parameter updates at each iteration. It is also computationally efficient since it does not involve inverting any matrix that is of the same magnitude of the cluster sizes. Another advantage of this EM algorithm is that the variance-covariance matrix (8) is readily available at the convergence of the algorithm for the joint inference of fixed effects β and random effects b .

6 Simulation

In this section, we conducted a simulation study to demonstrate the superior performance of the EM algorithm for the REML estimation presented in the previous section over the ML

inference using the NLMIXED procedure of SAS, especially in terms of computational time. In the simulation, we set $m = 20$ and conducted 100 simulation runs. In each simulation run, we generated n_{1i} and n_{2i} from $\text{Bin}(2000, 0.5)$ for $i = 1, 2, \dots, m = 20$. Here the expected cluster size is 2000, so large that existing software such as the MIXED procedure of SAS cannot allocate enough memory during the computation. We consider the situation of one covariate x , which was generated from $N(0, 1)$. Then clustered responses Y_{1ij} and Y_{2ij} were generated according to the following bivariate linear mixed model:

$$Y_{1ij} = \beta_{10} + x_{ij}\beta_{11} + b_{1i} + e_{1ij}, i=1, 2, \dots, m, j=1, 2, \dots, n_{1i},$$

$$Y_{2ij} = \beta_{20} + x_{i,n_{1i}+j}\beta_{21} + b_{2i} + e_{2ij}, i=1, 2, \dots, m, j=1, 2, \dots, n_{2i}, \quad (12)$$

where $\beta_{10} = 2, \beta_{11} = 3, \beta_{20} = 3, \beta_{21} = 2, b_{1i}, b_{2i}$ were generated from a bivariate normal distribution with mean zero and variance matrix $\{\sigma_{ij}\}$ with $\sigma_{11} = 2, \sigma_{12} = \sigma_{21} = 1$ and $\sigma_{22} = 5$, and the variances of the residual errors e_{1ij} and e_{2ij} were set as $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$. The generated data set was analyzed by the proposed REML EM algorithm and the ML method implemented using the NLMIXED procedure of SAS. The simulation results summarized over 100 simulation runs are presented in Table 1.

From Table 1, we observe that both REML and ML estimates of the fixed effects are virtually unbiased with the same empirical standard deviations up to 3 decimal points. However, the ML approach produced slightly smaller estimated standard errors for two intercept estimates $\hat{\beta}_{10}$ and $\hat{\beta}_{20}$, resulting in lower than nominal level coverage probability of a 95% confidence interval (CI) of β_{20} . We also observe that the REML estimates of all variance/covariance components are virtually unbiased. However, there is a sizable bias in the ML estimate of σ_{22} . Although it is well-known that ML estimates of variance/covariance components are somewhat biased since the ML approach does not take into account the estimation of the fixed effects, it is possible that the bias in the ML estimate of σ_{22} and the low empirical coverage of a 95% CI of β_{20} is more a consequence of the computation instability. This is because the numerical differentiation method in calculating required derivatives used in the NLMIXED procedure of SAS may not be stable. The significance of the REML EM algorithm compared to the ML approach is the savings in the computational time. The simulation study was performed on a Linux platform built on Intel(R) (8)Core(TM) i7 CPU at 2.93GHz with 8 GB RAM. On average the REML EM took only 6 seconds to analyze a data set, while the ML approach took 210 seconds to analyze the same data set.

7 Application to the LOS Data

We applied the bivariate linear mixed model (1) and the REML EM algorithm developed in this paper to analyze the LOS data. The data are from ten international clinical trials for patients with acute coronary syndromes (ACS): EARLY ACS, GUSTON IIB, GUSTO IV, PARAGON A, PARAGON B, PRISM, PRISM PLUS, PURSUIT, SUNERGY and TRACER. Each trial was conducted in several of the following representative regions in the world:

Asia, Australia/New Zealand, Europe, Latin America, Middle East, North America and South Africa. For more description of these trials, please see Chan, et. al. (2012). In this paper, we only included patients receiving a CABG or a PCI, resulting following cluster sizes (numbers of patients) for the ten clinical trials: 6469 for EARLY ACS, 1651 for GUSTON IIb, 1952 for GUSTO IV, 402 for PARAGON A, 1525 for PARAGON B, 1072 for PRISM, 1027 for PRISM PLUS, 4013 for PURSUIT, 6385 for SYNERGY and 8689 for TRACER. One of the research objectives is to estimate the expected LOS's of the CABG and PCI procedures, and their differences in different regions adjusting for patients' age, smoking and diabetes statuses, while taking into account the study-to-study variation. To address this objective, we considered the following model

$$Y_{1ij} = D_{1ij}^T \beta_1 + C_{1ij}^T \gamma_1 + b_{1i} + e_{1ij}, i=1, 2, \dots, 10, j=1, 2, \dots, n_{1i},$$

$$Y_{2ij} = D_{2i, n_{1i}+j}^T \beta_2 + C_{2ij}^T \gamma_2 + b_{2i} + e_{2ij}, i=1, 2, \dots, 10, j=1, \dots, n_{2i}, \quad (13)$$

where Y_{1ij} is the CABG LOS of patient j in trial i , and Y_{2ij} is the PCI LOS of the $(n_{1i} + j)$ th patient in the same trial, D_{1ij} , $D_{2i, n_{1i}+j}$ are 7×1 vectors of dummy variables for regions with corresponding effects β_1 , β_2 , C_{1ij} , $C_{2i, n_{1i}+j}$ are 3×1 vector of covariates representing patients' age in year (centered at 60 years, the approximate sample mean), smoking status (1 for ever/current smoking, 0 for never-smoking) and diabetes status (1 for diabetes and 0 for non-diabetes). Therefore, β_1 , β_2 are vectors representing the expected LOS's of CABG and PCI procedures respectively for the patient populations (called target populations) who are 60 years old, non-smoking and free of diabetes in the seven regions in the world, and γ_1 , γ_2 are the effects of those covariates on two LOS's. We used normally distributed random effects $(b_{1i}, b_{2i})^T$ with mean zero and variance matrix $\{\sigma_{ij}\}$ to account for the study-to-study variation in the CABG LOS's and PCI LOS's, as well as their correlation. The residual errors e_{1ij} 's and e_{2ij} 's are assumed to be independent with residual variances σ_1^2 and σ_2^2 respectively.

Because of the large cluster sizes in this application, the commercial software such as the MIXED procedure of SAS could not fit the bivariate linear mixed model (13) since it failed to allocate enough memory during the optimization process. Using the REML EM algorithm we developed, the fitting of model (13) was completed in seconds. The estimated expected LOS's of the CABG and PCI procedures and their differences for the target populations in the seven regions are presented in Table 2. Also presented in this table are contrasts of these parameters of a region to the reference region: South Africa. From this table, we observe that patients' CABG and PCI LOS's in Asia and Europe are much longer than those in other regions, that the expected CABG LOS is about 4 or 5 days longer the expected PCI LOS in any region, with Asia having the biggest difference. This result may reflect the difference in patient management and quality of care in different regions.

Even though the estimated expected CABG LOS is much greater than the expected PCI LOS in any region, the estimated variances of random trial effects for the CABG LOS and PCI

LOS are very close ($\hat{\sigma}_{11} = 0.36$, $\hat{\sigma}_{22} = 0.43$). The covariance of random trial effects for CABG LOS and PCI LOS is estimated to $\hat{\sigma}_{12} = 0.20$, indicating studies with longer CABG LOS's tend to have longer PCI LOS's. The estimated residual variances of patient CABG LOS's and PCI LOS's are $\hat{\sigma}_1^2 = 31$ and $\hat{\sigma}_2^2 = 11$ respectively, indicating that there is much greater within-trial variation in patients' CABG LOS's than in patients' PCI LOS's in any given study and region.

We also implemented the ML approach using the NLMIXED procedure of SAS. Unfortunately, the program produced a Hessian matrix with some negative eigen-values during the optimization process and failed to provide valid estimates, especially for the variance/covariance parameters. This further demonstrates the advantage of the proposed REML EM algorithm for fitting a linear mixed model for clustered data with large cluster sizes.

8 Discussion

In this paper, we discussed joint modeling of two clustered response variables, hospital length of stay (LOS) for patients with acute coronary syndromes (ACS) who received CABG and PCI surgeries from several international clinical trials. A bivariate linear mixed model with separate fixed effects and random effects was proposed for the joint modeling where random effects were used to model the study-to-study variation. Due to large cluster sizes of these clinical trials, commercial software such as SAS could not fit the proposed model. We proposed a computational solution to the ML inference and the REML inference. Specially, we proposed an EM algorithm for the REML inference and provided an implementation of the ML inference using an existing procedure of SAS. Simulation studies indicated that compared to the ML approach implemented with SAS, the proposed REML EM algorithm is computationally stable and efficient. By applying the proposed REML EM algorithm to the LOS data, we are able to make inference on the expected LOS's and their difference of CABG and PCI surgeries for different regions in the world, after adjusting for important covariates.

With the current random effects structure, when a large number of candidate covariates are available, we may use traditional model selection methods such as forward, backward and step-wise selection, or more advanced penalized likelihood methods with sparsity penalties on potential fixed effects. In this case, we need to calculate the likelihood function of the potential fixed effects. For this purpose we can develop an ML EM algorithm similar to that of Laird and Ware (1982) for our bivariate linear mixed model. The current random effects structure uses two correlated response specific random intercepts to model the cluster-to-cluster variation of the responses. If prior information indicates that some "covariate effects" on each response variable vary from cluster to cluster, we can expand the current random effect structure to include these "covariate effects" as additional (correlated) random effects for both responses. Information criterion such as Bayesian Information Criterion together with the penalized likelihood approach may be used to select a final random effects structure and fixed effects.

The proposed bivariate linear mixed model is more appropriate for jointly modeling two continuous clustered responses. For discrete clustered responses such as binary responses, a more attractive model would be the bivariate generalized linear mixed model. It is straightforward to extend the idea to this model for the case of large cluster sizes. For example, the ML inference can be easily implemented with the NLMIXED procedure of SAS. However, the EM algorithm of the REML-like inference in the bivariate generalized linear mixed model does not have all the attractive features of the bivariate linear mixed model. For example, there is no closed form expression for updated parameters during the EM iteration, and a numerical method has to be used to obtain the parameter update. Despite this problem, we think that the EM algorithm will still be computationally more stable and efficient compared to the ML inference implemented with the NLMIXED procedure of SAS. It will be an interesting future project to evaluate their performance for the case of large cluster sizes.

Acknowledgments

The work of D. Zhang was supported by NIH grant R01 CA85848-12. The work of J. L. Sun and K. Pieper was supported through a grant from the Duke Clinical research Institute. The authors are grateful to Dr. Eric Peterson for his institutional financial support as well as providing the LOS data, without which this research would not have been possible.

References

- Chan M, Sun J, Newby L, Lokhnygina Y, White HD, Moliterno DJ, Throux P, Ohman EM, Simoons ML, Mahaffey KW, Pieper KS, Giugliano RG, Armstrong PW, Califf RM, Van de Werf F, Harrington RA. Trends in clinical trials of non-ST-segment elevation acute coronary syndromes over 15 years. *International Journal of Cardiology*. 2012
- Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*. 1977; 39:138.
- Diggle, PJ.; Heagerty, P.; Liang, k-Y.; Zeger, SL. *Analysis of Longitudinal Data*. 2nd. Oxford University Press; 2002.
- Harville DA. Bayesian inference for variance components using only error contrasts. *Biometrika*. 1994; 61:383–385.
- Henderson, CR. *Applications of Linear Models in Animal Breeding*. University of Guelph; 1984.
- Laird NM, Ware JH. Random effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- SAS/STAT 9.3 User's Guide. Cary, NC: SAS Institute Inc.; 2013.

Appendix A

Properties of the REML EM Algorithm

The following theorem states the general EM algorithm for the REML estimation.

Theorem 1

Suppose $f(y|b; \beta, \theta)$ is the conditional probability density function of y given random effects b and $f(b; \theta)$ is the probability density function of the random effects b . Assume the following “REML” likelihood

$$L_R(\theta; y) = \int f(y|b; \beta, \theta) f(b; \theta) d\beta db$$

exists, so that $g(\beta, b|y; \theta) = f(y|b; \beta, \theta) f(b; \theta) / L_R(\theta; y)$ is a probability density function. Given estimate $\theta^{(t)}$ at the t th iteration, define the Q -function for the REML algorithm as follows:

$$Q_R(\theta|\theta^{(t)}) = E_g[\log\{f(y|b; \beta, \theta) f(b; \theta)\}],$$

where E_g stands for the expectation taken with respect to $g(\beta, b|y; \theta^{(t)})$. Denote $\theta^{(t+1)}$ the update of θ obtained by maximizing $Q_R(\theta|\theta^{(t)})$ with respect to θ , then

$$\ell_R(\theta^{(t+1)}; y) \geq \ell_R(\theta^{(t)}; y), \text{ for all } t,$$

where $\ell_R(\theta; y) = \log L_R(\theta; y)$.

Proof—By the definition of $L_R(\theta; y)$, we have

$$\log\{f(y|b; \beta, \theta) f(b; \theta)\} = \log g(\beta, b|y; \theta) + \ell_R(\theta; y).$$

Taking expectation with respect to $g(\beta, b|y; \theta^{(t)})$ in both sides of the above equation leads to

$$Q_R(\theta|\theta^{(t)}) = E_g \log g(\beta, b|y; \theta) + \ell_R(\theta; y).$$

Denote $H(\theta) = E_g \log g(\beta, b|y; \theta)$. Then

$$\ell_R(\theta^{(t+1)}; y) - \ell_R(\theta^{(t)}; y) = Q_R(\theta^{(t+1)}|\theta^{(t)}) - Q_R(\theta^{(t)}|\theta^{(t)}) - \{H(\theta^{(t+1)}) - H(\theta^{(t)})\}.$$

By the definition of $H(\theta)$, E_g and Jensen's inequality, we have

$$H(\theta^{(t+1)}) - H(\theta^{(t)}) = E_g \log \left\{ \frac{g(\beta, b|y; \theta^{(t+1)})}{g(\beta, b|y; \theta^{(t)})} \right\} \leq \log E_g \left\{ \frac{g(\beta, b|y; \theta^{(t+1)})}{g(\beta, b|y; \theta^{(t)})} \right\} = \log(1) = 0.$$

Since $\theta^{(t+1)}$ maximizes $Q_R(\theta|\theta^{(t)})$, it follows that $Q_R(\theta^{(t+1)}|\theta^{(t)}) \geq Q_R(\theta^{(t)}|\theta^{(t)})$. Therefore

$$\ell_R(\theta^{(t+1)}; y) \geq \ell_R(\theta^{(t)}; y), \text{ for all } t.$$

Table 1

Simulation results comparing REML and MLE. “Bias”, “SD”, “SE” and “CP” are the bias, empirical standard deviation, estimated standard error and empirical coverage probability of a 95% CI of a parameter estimate based on 100 Monte Carlo runs.

Parameter	REML				ML			
	Bias	SD	SE	CP	Bias	SD	SE	CP
β_{10}	-0.015	0.316	0.318	0.950	-0.015	0.316	0.310	0.950
β_{11}	0.000	0.007	0.007	0.930	0.000	0.007	0.007	0.930
β_{20}	-0.054	0.510	0.496	0.910	-0.054	0.510	0.483	0.890
β_{21}	0.001	0.015	0.014	0.910	0.001	0.015	0.014	0.910
σ_{11}	0.071	0.614	-	-	-0.032	0.584	-	-
σ_{12}	0.052	0.754	-	-	0.000	0.717	-	-
σ_{22}	0.048	1.670	-	-	-0.205	1.588	-	-
σ_1^2	-0.001	0.001	-	-	-0.001	0.010	-	-
σ_2^2	0.002	0.037	-	-	0.002	0.037	-	-

Table 2

Estimated expected CABG LOS ($\hat{\mu}_C$), PCI LOS ($\hat{\mu}_P$) and their difference $\hat{\mu}_C - \hat{\mu}_P$ for the target populations in the seven regions of the world. Also presented are contrasts (denoted by $\hat{\mu}_C - \hat{\mu}_P$) of these parameters of a region to the referee region: South Africa. Numbers inside parentheses are the estimated standard error (SE) for the estimated parameters. AU/NZ stands for Australia/New Zealand.

Region	CABG			PCI		CABG - PCI	
	$\hat{\mu}_C$	$\hat{\mu}_C$	$\hat{\mu}_P$	$\hat{\mu}_P$	$\hat{\mu}_C - \hat{\mu}_P$	$(\hat{\mu}_C - \hat{\mu}_P)$	
Asia	12.77(0.58)	3.08(0.70)	6.29(0.24)	2.23(0.23)	6.48(0.60)	0.85(0.74)	
AU/NZ	9.30(0.42)	-0.39(0.58)	3.83(0.25)	-0.23(0.24)	5.46(0.45)	-0.17(0.63)	
Europe	11.95(0.24)	2.26(0.46)	6.07(0.21)	2.00(0.20)	5.88(0.25)	0.23(0.50)	
Latin America	10.46(0.35)	0.78(0.53)	5.23(0.24)	1.17(0.23)	5.24(0.37)	-0.39(0.57)	
Middle East	8.70(0.42)	-0.98(0.58)	4.80(0.23)	0.73(0.22)	3.91(0.44)	-1.71(0.62)	
North America	8.10(0.23)	-1.58(0.46)	3.77(0.22)	-0.29(0.20)	4.33(0.25)	-1.29(0.50)	
South Africa	9.68(0.49)	-	4.06(0.29)	-	5.62(0.54)	-	