Estimation Via EM-Algorithm in a Bivariate Random Effects Model

Amrik Shah, Cleveland Clinic Foundation, Cleveland, Ohio
Kirk Easley, Cleveland Clinic Foundation, Cleveland, Ohio *

Abstract

This paper considers estimation in a bivariate random-effects model allowing for arbitrary measurement times and variation in the number of observations for different individuals in the context of longitudinal studies. Two different structures for the covariance matrix of measurement error are considered, uncorrelated error between responses and correlation of error terms at the same measurement times. The estimation of parameters for this model is via the EM-Algorithm. We derive the set of equations for both ML and REML estimation when the observed data consists of complete pairs. These equations are encoded in a **SAS** Macro utilizing SAS/IML for implementation of the methodology. This is illustrated with an example from AIDS clinical trials.

KEY WORDS: EM-algorithm, longitudinal data, multiple response, random-effects, REML estimates, SAS Macro.

1 Introduction

Longitudinal studies in which the response is measured at the same time for all units has been considered by many authors, including Potthoff & Roy (1964), Rao (1965), and Grizzle & Allen (1969). In the case of arbitrary measurement times and variation in the number of observations for different individuals, Laird & Ware (1982) propose a random effects approach. The **SAS** PROC MIXED procedure can be used to fit the single characteristic model for longitudinal studies. Reinsel (1984) has extended the random effects model to handle a repeated multivariate response. He deals with a complete and balanced design, i.e., there is no allowance for missing data, and presents closed form solutions for the parameter estimates. In clinical trials, measurement times may be arbitrary and dropouts will result in different number of repeated responses for patients or experimental units. In this case, it is no longer possible to obtain closed form solutions and iterative techniques such as the E-M and N-R algorithm have to be employed to obtain maximum likelihood estimates.

In this paper, we consider an extension of the EM-algorithm presented in Laird & Ware (1982) for parameter estimation in a bivariate response random-effects model. We present the algorithm for two possible types of 'missing' data structures. In the first case both characteristics are observed at each occasion, though the number and timing of observations may differ from individual to individual, i.e., the data are complete but unbalanced in number of observations per experimental unit. This case is addressed in detail with the estimating equations being derived for both ML and REML estimation. The more difficult case occurs when both characteristics may not be observed at all occasions, i.e., the data are incomplete and possibly unbalanced. For this situation we describe the data structure vis-a-vis the design matrices and briefly sketch the estimation procedure. Two possible areas of application are: AIDS clinical trials where a number of characteristics, e.g. CD4 and CD8 cell counts, are

*Amrik Shah and Kirk Easley are Biostatisticians at the Cleveland Clinic Foundation, Department of Biostatistics & Epidemiology, 9500 Euclid Avenue, Cleveland, Ohio, 44195. The authors thank *Pharmacia* for providing the data from their Rifabutin trials.

measured repeatedly over time for each patient, and systolic and diastolic blood pressure measurements. These bivariate repeated measures can be easily modeled with this approach allowing us to estimate the correlation between the slopes and intercepts of the bivariate data as will be illustrated in the example.

2 Theory and Background

2.1 Model and Assumptions

Let $\mathbf{Y}_i = [\mathbf{y}_{i1}, \mathbf{y}_{i2}]$, where each $\mathbf{y}_k$ is a column vector of dimension n_i, be the response matrix for unit i and $\mathbf{E}_i = [\mathbf{E}_{i1}, \mathbf{E}_{i2}]$ be the $n_i \times 2$ matrix of error terms. Introducing the *vec* operator, which strings out the columns of a matrix vertically, we obtain $\mathbf{y}_i = vec(\mathbf{Y}_i) = (\mathbf{y}_{i1}', \mathbf{y}_{i2}')'$ which is now a $2n_i \times 1$ column vector, and $\mathbf{e}_i = vec(\mathbf{E}_i)$. The model for each individual unit i is of form

$$\mathbf{y}_i = \mathbf{X}_i^* \beta + \mathbf{Z}_i^* \gamma_i + \mathbf{e}_i \qquad i = 1, 2, \ldots, N \tag{1}$$

where $\mathbf{X}_i^*$ is a $2n_i \times q^*$ known fixed design matrix, β is a $q^* \times 1$ matrix of fixed parameters, $\mathbf{Z}_i^*$ is a $2n_i \times r^*$ random design matrix which is usually some subset of $\mathbf{X}_i^*$, and γ_i is an $r^* \times 1$ matrix of individual random effects.

The assumptions are: $\gamma_i \sim MVN(\mathbf{0}, \mathbf{D}_{r^* \times r^*})$ and $\mathbf{e}_i$ is distributed as N($\mathbf{0}$, $\mathbf{R}_i$) where $\mathbf{R}_i$ has dimensions $2n_i \times 2n_i$, and, for units i and j, where $i \neq j$, Cov($\mathbf{e}_i$, $\mathbf{e}_j$) = $\mathbf{0}$, Cov(γ_i, γ_j) = $\mathbf{0}$, and Cov($\mathbf{e}_i$, γ_i) = $\mathbf{0}$. In most cases it may be reasonable and convenient to assume a structured $\mathbf{R}_i$ for ease in estimation. One possible structure is that the row vectors of $\mathbf{E}_i$, representing the results at different measurement times, are independent and have distribution N($\mathbf{0}$, $\Sigma_{2\times2}$), where Σ is an unstructured covariance matrix. In this case Var($\mathbf{e}_i$) has structure $\Sigma \otimes \mathbf{I}_i$, where, $\mathbf{I}_i$ is an identity matrix of dimension $n_i \times n_i$. The independence of rows implies that each column of $\mathbf{E}_i$ is distributed normally with mean $\mathbf{0}$ and covariance matrix $\sigma_{kk}^2\ \mathbf{I}_i$, for $k = 1, 2$. A less likely structure is to assume Σ is diagonal. For a linear model specified by (1), β could be a vector of population intercepts and slopes for the 2 responses while the γ_i's are vectors of the individual intercepts and slopes.

2.2 Estimation via EM-Algorithm

We use the EM-algorithm for obtaining maximum likelihood(ML) or restricted maximum likelihood(REML) estimates for the parameters. It should be noted that in the case of known Σ and $\mathbf{D}$, the estimate of β is easily obtained via a closed form solution by Generalized Least Squares(GLS).

Let τ index the iterations for $\tau = 0, 1, 2, \ldots, \infty$, where $\tau = 0$ denotes the starting values. The sufficient statistics for Σ and $\mathbf{D}$ are $\sum(\mathbf{E}_i'\mathbf{E}_i)$ and $\sum(\gamma_i\gamma_i')$ respectively. Since γ_i and $\mathbf{e}_i = vec(\mathbf{E}_i)$ are unobservable, the algorithm computes the expectations of the sufficient statistics and then solves for maximum likelihood. It uses the joint density of $\mathbf{y}_i$, γ_i, $\mathbf{e}_i$ to obtain the conditional expectations of the sufficient statistics.

E-Step: Let θ be the vector of unknown parameters in Σ and $\mathbf{D}$ and $\theta^{(\tau)}$ denote their values at the end of the τ^{th} iteration. The estimate for β given values $\theta^{(\tau)}$ is

$$\beta^{(\tau)} \;=\; (\sum_{i=1}^{N} \mathbf{X}_i^{*\prime} \mathbf{P}_i^{(\tau)} \mathbf{X}_i^*)^{-1} \sum_{i=1}^{N} (\mathbf{X}_i^{*\prime} \mathbf{P}_i^{(\tau)} \mathbf{y}_i), \tag{2}$$

where $\mathbf{P}_i^{(\tau)} = \mathbf{V}_i{}^{(\tau)^{-1}}$ and $\mathbf{V}_i = Var(\mathbf{y}_i) = [\mathbf{Z}_i^*\mathbf{D}\mathbf{Z}_i^{*\prime} + \boldsymbol{\Sigma} \otimes \mathbf{I}_i]$. Letting $\mathbf{r}_i^{(\tau)} = \mathbf{y}_i - \mathbf{X}_i^*\beta^{(\tau)}$ and $\mathbf{B}^{\otimes 2} = \mathbf{B}\mathbf{B}'$, the expectations of the i^{th} term of the sufficient statistics are given by:

$$\begin{aligned}
E[(\gamma_i)(\gamma_i)'|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}] &= \{E[\gamma_i|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}]\}^{\otimes 2} + V[\gamma_i|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}] \\
E[(\mathbf{E}_{ij})'(\mathbf{E}_{ik})|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}] &= E[\mathbf{E}_{ij}|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}]'E[\mathbf{E}_{ik}|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}] \\
&\quad + tr[Cov(\mathbf{E}_{ij}, \mathbf{E}_{ik}|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)})], \quad j,k = 1,2.
\end{aligned}$$

These expectations are easily obtained using the conditional mean and covariance matrix of the multivariate normal distribution.

M-Step: In the M-step, $\boldsymbol{\Sigma}^{(\tau+1)}$ and $\mathbf{D}^{(\tau+1)}$ are found by equating them to the expected value of their sufficient statistics. For ML estimates the iterative equations are:

$$\begin{aligned}
\mathbf{D}^{(\tau+1)} &= \Big[\sum_{i=1}^{N}\Big[E[(\gamma_i)(\gamma_i)'|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}]\Big]\Big] \Big/ N, \\
\sigma_{jk}^{(\tau+1)} &= \Big[\sum_{i=1}^{N} n_i\Big]^{-1} \Big[\sum_{i=1}^{N}\Big[E[(\mathbf{E}_{ij})'(\mathbf{E}_{ik})|\mathbf{y}_i, \boldsymbol{\theta}^{(\tau)}, \boldsymbol{\beta}^{(\tau)}]\Big]\Big], \quad j,k = 1,2.
\end{aligned}$$

To obtain REML estimates, at the E-step we condition only on $\mathbf{y}_i$ given $\mathbf{D}$ and $\boldsymbol{\Sigma}$, since $\boldsymbol{\beta}$ is integrated out of the likelihood using a flat prior (Laird & Ware 1982). Using the posterior distribution of $\boldsymbol{\beta}$ and simplifying we obtain the conditional expectations of the sufficient statistics to be:

$$\begin{aligned}
E[\gamma_i\gamma_i{}'|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}] &= [\widehat{\mathbf{D}}\mathbf{Z}_i^{*\prime}\widehat{\mathbf{P}}_i(\mathbf{y}_i - \mathbf{X}_i^*\widehat{\beta})]^{\otimes 2} + \widehat{\mathbf{D}} - \widehat{\mathbf{D}}\mathbf{Z}_i^{*\prime}\widehat{\mathbf{W}}_i\mathbf{Z}_i^*\widehat{\mathbf{D}}, \\
E[\mathbf{E}_{ij}{}'\mathbf{E}_{ik}|\mathbf{y}_i, \widehat{\boldsymbol{\theta}}] &= [(\hat{\sigma}_{1j}\mathbf{I}_i|\hat{\sigma}_{j2}\mathbf{I}_i)\mathbf{P}_i^{(\tau)}\widehat{\mathbf{r}}_i]'[(\hat{\sigma}_{1k}\mathbf{I}_i|\hat{\sigma}_{k2}\mathbf{I}_i)\mathbf{P}_i^{(\tau)}\widehat{\mathbf{r}}_i] \\
&\quad + tr[\hat{\sigma}_{jk}\mathbf{I}_i - (\hat{\sigma}_{1j}\mathbf{I}_i|\hat{\sigma}_{j2}\mathbf{I}_i)\widehat{\mathbf{W}}_i(\hat{\sigma}_{1k}\mathbf{I}_i|\hat{\sigma}_{k2}\mathbf{I}_i)'], \quad j,k = 1,2. \\
\text{where} \qquad \widehat{\mathbf{W}}_i &= \widehat{\mathbf{P}}_i\Big[\mathbf{I}_i - \mathbf{X}_i^*\Big(\sum_{i=1}^{N}\mathbf{X}_i^{*\prime}\widehat{\mathbf{P}}_i\mathbf{X}_i^*\Big)^{-1}\mathbf{X}_i^{*\prime}\widehat{\mathbf{P}}_i\Big].
\end{aligned} \tag{3}$$

For obtaining starting values for $\widehat{\mathbf{D}}$ and $\widehat{\boldsymbol{\Sigma}}$ one may even use identity matrices for $\mathbf{D}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$. Another approach (Laird, Lange & Stram 1987) is to do OLS for each unit, or for those having the requisite minimum number of repeat measurements. The model used to fit by OLS is equivalent to the random design, $\mathbf{y}_i = \mathbf{Z}_i^*\gamma_i + \mathbf{e}_i$. This will yield $\gamma_i{}^{(0)}$ and $\mathbf{e}_i{}^{(0)} = vec[\mathbf{E}_i{}^{(0)}]$. From these one can obtain $\beta^{(0)}$ using (2), and $\mathbf{D}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$ can be obtained using the following equations

$$\mathbf{D}^{(0)} = \Big[\sum_{i=1}^{N}\Big(\gamma_i{}^{(0)} - \frac{\sum\gamma_i{}^{(0)}}{N}\Big)^{\otimes 2}\Big] \Big/ N, \quad \text{and} \quad \boldsymbol{\Sigma}^{(0)} = \Big[\sum_{i=1}^{N}(\mathbf{E}_i{}^{(0)})'(\mathbf{E}_i{}^{(0)})\Big] \Big/ \sum_{i=1}^{N} n_i.$$

2.3 Modeling with Incomplete Pairs

We define the missing situation to be one in which the rows of $\mathbf{Y}_i$ are not complete, i.e., both characteristics are not observed at all measurement times. Let $\tilde{\mathbf{X}}_{ik}$ be the design matrix for the responses of i^{th} unit on the k^{th} characteristic and $\tilde{\mathbf{Z}}_{ik}$, which is generally a subset of $\tilde{\mathbf{X}}_{ik}$, be the corresponding random design matrix. If we let

n_{ik} be the number of repeated measures on characteristic k for unit i, then the dimensions of $\tilde{\mathbf{X}}_{ik}$ are $n_{ik} \times q^*$ and $\tilde{\mathbf{Z}}_{ik}$ is $n_{ik} \times r^*$, for $i = 1, \ldots, N$. Also define $\mathbf{Y}_{ij}$ to be the response vector on the j^{th} characteristic and $\tilde{\mathbf{y}}_i = [\mathbf{Y}'_{i1}, \mathbf{Y}'_{i2}]'$. Then, the model for unit i may be represented as

$$\begin{aligned} \begin{pmatrix} \mathbf{Y}_{i1} \\ \mathbf{Y}_{i2} \end{pmatrix} &= \begin{pmatrix} \tilde{\mathbf{X}}_{i1} \\ \tilde{\mathbf{X}}_{i2} \end{pmatrix} \beta + \begin{pmatrix} \tilde{\mathbf{Z}}_{i1} \\ \tilde{\mathbf{Z}}_{i2} \end{pmatrix} \gamma_i + \begin{pmatrix} \mathbf{E}_{i1} \\ \mathbf{E}_{i2} \end{pmatrix} \\ \Rightarrow \tilde{\mathbf{y}}_i &= \tilde{\mathbf{X}}_i \, \beta + \tilde{\mathbf{Z}}_i \, \gamma_i + \mathbf{e}_i. \end{aligned}$$

Generally we will have

$$\begin{bmatrix} \tilde{\mathbf{X}}_{i1} \\ \tilde{\mathbf{X}}_{i2} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i1} & 0 \\ 0 & \mathbf{X}_{i2} \end{bmatrix} \quad \text{and,} \quad \begin{bmatrix} \tilde{\mathbf{Z}}_{i1} \\ \tilde{\mathbf{Z}}_{i2} \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_{i1} & 0 \\ 0 & \mathbf{Z}_{i2} \end{bmatrix},$$

be block diagonal design matrices where $\mathbf{X}_{ik}$ has dimensions $n_{ik} \times q_k$, and $\mathbf{Z}_{ik}$ has dimensions $n_{ik} \times r_k$. This is not a requirement and arbitrary structures for $\tilde{\mathbf{X}}_i$ and $\tilde{\mathbf{Z}}_i$ are feasible, provided the data permits estimation. The dimensions of $\tilde{\mathbf{X}}_i$ are $m_i \times q^*$ and $\tilde{\mathbf{Z}}_i$ has size $m_i \times r^*$ where $m_i = \sum_{k=1}^{2} n_{ik}$. For the block diagonal structure, $q^* = \sum_{k=1}^{2} q_k$ and $r^* = \sum_{k=1}^{2} r_k$.

The maximum likelihood estimate of β can be obtained in a manner similar to the complete case. The estimation of $\mathbf{D}$ also follows a similar approach but the major difference is in estimating $\boldsymbol{\Sigma}$. For the case of unstructured $\boldsymbol{\Sigma}$ and missing data, we essentially need to estimate the missing data which is considered to be the non-observable error terms corresponding to the missing parts of the responses. Further details can be found in Shah, Laird & Schoenfeld (1997).

Once we have the starting values, we cycle between the E-Step and the M-Step until we attain convergence. At convergence, we also obtain $\widehat{\beta}$, $\widehat{\gamma}_i$, $\widehat{AVar}(\widehat{\beta})$ as well as the 'robust' or sandwich estimator as given by Liang & Zeger (1986). In order to make inferences about the elements of $\widehat{\beta}$ we may use either the sandwich estimator or an estimate of its asymptotic variance based on the Fisher Information, which is, $[\sum_{i=1}^{N} \tilde{\mathbf{X}}_i{}' \widehat{\mathbf{P}}_i \tilde{\mathbf{X}}_i]^{-1}$.

3 Macro Call and Options

```
%mvemal( DATA     = data,
         METHOD   = method,
         VAR      = var,
         STVAL    = stval,
         YVAR     = yvar,
         XVARS    = xvars,
         ZVARS    = zvars,
         ID       = id,
         CONV     = conv,
         ITER     = iter) ;
```

3.1 Options

There are no default values for any of the macro parameters. All options should be in lower case and a directory with libname 'est' should be created for temporary storage of starting values and estimates.

DATA: The name of the (SAS) data set to be used.

METHOD: The two choices are *ml* and *reml* for getting maximum-likelihood or restricted maximum-likelihood estimates. As noted earlier, these differ in the calculation of $\widehat{\mathbf{W}}_i$. For ML estimates $\widehat{\mathbf{W}}_i = \mathbf{V}_i{}^{-1}$ while for REML estimation $\widehat{\mathbf{W}}_i$ is as given by equation (3).

VAR: This defines the structure of $\boldsymbol{\Sigma}$, *un* forces it to be unstructured (i.e. correlated error terms) while *sim* constrains $\boldsymbol{\Sigma}$ to be a diagonal matrix (i.e. uncorrelated error terms).

STVAL: There are three choices: *yes* asks the program to compute the starting values as described in section 2.2, *no* forces the program to use the results from a previous run as the starting values. It must be noted that this will only be possible if the random factors remain the same, and only the fixed design is varied. The third choice is *identity* which begins the estimation with $\boldsymbol{\Sigma}$ and $\mathbf{D}$ being identity matrices.

YVAR: This identifies the dependent or response variable and is numeric.

XVARS: This is a list of independent variables which constitute the fixed design matrices. A blank space should separate the names. Note, an intercept variable must be created and included in this list if desired.

ZVARS: A list of variables, separated by a blank space, which make up the random design matrices. Again, an intercept variable has to be created if required.

ID: A variable which identifies the subject or experimental unit.

CONV: The convergence criterion (< 1). When the change in log-likelihood is less than the given criterion convergence is attained.

ITER: The maximum number (integer) of iterations allowed. This supersedes the convergence criterion.

4 Illustration

4.1 Model and Data

We apply this methodology to the combined data from two randomized, double-blind, multicenter trials of daily prophylactic treatment with either rifabutin or placebo. From February 1990 through January 1992, 590 patients were enrolled in study 023, and 556 patients were enrolled in study 027. All 1146 patients were symptomatic and had CD4 cell counts $\leq 200/mm^3$. The patients were scheduled to have CD4 and CD8 counts taken at baseline and every three months thereafter. The primary objective of the trial was to determine if rifabutin reduces the frequency of *Mycobacterium Avium* complex (MAC) infection. We consider in this analysis a model for the joint behavior over time of CD4 and CD8 counts. Let $\mathbf{y}_i$ be the column vector of log(CD4) and log(CD8). We assume both log(CD4) and log(CD8) change linearly over time and since CD4 and CD8 are obtained from same assay, we let $\boldsymbol{\Sigma}$ be unstructured. Assuming no missing data, the model for unit i is

$$\mathbf{y}_i = \mathbf{X}_i^* \boldsymbol{\beta} + \mathbf{Z}_i^* \gamma_i + \mathrm{e}_i. \tag{4}$$

4.2 Macro Call and Results

Since $\boldsymbol{\Sigma}$ is a matrix, **PROC MIXED** cannot be used to fit the model. A SAS macro program encoded in IML was used for the estimation. Data from patients on rifabutin were used for the analysis. Starting values were obtained by OLS and $\approx$ 65 iterations were needed for convergence which was defined to be a change less than 0.005 in the log-likelihood. The macro call statement is:
%mvemal(DATA = **exdata**, METHOD = **reml**, VAR = **un**, STVAL = **identity**, YVAR = **cd**,
XVARS = **cd4int cd4sl cd8int cd8sl**, ZVARS = **cd4int cd4sl cd8int cd8sl**, ID = **pid**,
CONV = **0.005**, ITER = **75**) ;
The data set *exdata* has the variables **cd cd4int cd4sl cd8int cd8sl pid** sorted by subject. The variable **cd** is actually log(cd4) and log(cd8) stacked together, with the cd4 vector on top. The data table for an individual with three pairs of measurement for CD4 and CD8 would look as follows:

PID	CD4INT	CD4SL	CD8INT	CD8SL	CD
1	1	1	0	0	2.56495
1	1	13	0	0	2.39790
1	1	25	0	0	1.79176
1	0	0	1	1	6.01372
1	0	0	1	13	5.75574
1	0	0	1	25	5.77144

The first three rows constitute the CD4 measures and the next three rows are the CD8 measures. The following output gives the REML estimates.

```
                STARTING VALUES FOR EM-ALGORITHM

     SIGHAT                  DHAT

        1       0            1       0       0       0

        0       1            0       1       0       0

                             0       0       1       0

                             0       0       0       1



          NUMBER OF INDIVIDUALS IN DATA SET =        488

                        FINAL ESTIMATES


             CONVERGENCE IN        55    ITERATIONS

                LOG-LIKELIHOOD  =   -177.011
             RESTRICTED MAXIMUM-LIKELIHOOD ESTIMATES
          ALPHAT    ASYM(VAR)     P(ASYM)    RBST(VAR)      P(RBST)


CD4INT 3.686394      0.002916           0     0.002908            0
CD4SL   -0.01612      1.82E-6           0     1.801E-6            0
CD8INT 6.233848      0.001579           0     0.001576            0
CD8SL   -0.00791     9.482E-7    4.44E-16     9.428E-7     4.44E-16

            DHAT=COVARIANCE MATRIX FOR RANDOM EFFECTS
                 correlations below the diagonal
              0.959627 -0.00691 0.360783 -0.00146
               -0.4817 0.000215 -0.00456 0.000102
              0.498037 -0.42101 0.546848 -0.00419
              -0.13244 0.619605 -0.50538 0.000126

                  SIGMA = COV MATRIX FOR ERROR
                        0.299292 0.078266
                        0.383722    0.139

                   ASYMPTOTIC VARIANCE (ALPHAT)
                    CD4INT     CD4SL   CD8INT     CD8SL
           CD4INT  0.002916 -0.00004 0.001018 -0.00001
           CD4SL   -0.00004  1.82E-6 -0.00002 6.663E-7
           CD8INT  0.001018 -0.00002 0.001579 -0.00002
           CD8SL   -0.00001 6.663E-7 -0.00002 9.482E-7

                    ROBUST VARIANCE (ALPHAT)
                    CD4INT     CD4SL   CD8INT     CD8SL
            CD4INT 0.002908 -0.00004 0.001021 -0.00001
            CD4SL  -0.00004 1.801E-6 -0.00002 6.703E-7
            CD8INT 0.001021 -0.00002 0.001576 -0.00002
            CD8SL  -0.00001 6.703E-7 -0.00002 9.428E-7
```

The correlation ($\rho \approx .38$) between the measurement errors for CD4 and CD8 probably arises because both CD4 and CD8 count are found by multiplying the percentage CD4 and CD8 cells by the total lymphocyte count. There was also a strong correlation between the intercept of CD4 and CD8 ($\rho = 0.49$) as well as between the slope estimates ($\rho = 0.62$). The robust and asymptotic variance estimates were very similar. We can use the asymptotic or robust standard errors of the fixed effect estimates to test for the slopes.

5 Discussion

The main advantage of the multiple response model lies in its ability to utilize the inherent covariance structure for a truly multivariate response, thereby resulting in more efficient estimation of the parameters. In the case of a single response, the estimation procedure as outlined reverts to the iterative equations given by Laird, Lange & Stram (1987). An area for further development would be the computational techniques involved in the estimation of parameters. Since the EM-algorithm does not yield the Hessian, alternate techniques for estimation may be needed in order to do inference on the covariance parameters. It would be useful to have a SAS procedure which would be more flexible in model fitting. It may be possible to embed equations for estimating the Hessian within the EM-algorithm itself. It would also be worthwhile to investigate potential acceleration techniques in order to enhance the rate of convergence of the algorithm.

Availability:*The SAS Macro* **mvemal** *can be obtained by contacting the first author at: amrik@bio.ri.ccf.org*

References

Grizzle, J. E. & Allen, D. M. (1969), 'Analysis of growth and dose response curves (corr: V26 p860)', *Biometrics* **25**, 357–381.

Laird, N., Lange, N. & Stram, D. (1987), 'Maximum likelihood computations with repeated measures: Application of the EM algorithm', *Journal of the American Statistical Association* **82**, 97–105.

Laird, N. M. & Ware, J. H. (1982), 'Random-effects Models for Longitudinal Data', *Biometrics* **38**, 963–974.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Potthoff, R. & Roy, S. N. (1964), 'A generalized multivariate analysis of variance model useful especially for growth curve problems', *Biometrika* **51**, 313–326.

Rao, C. R. (1965), 'The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves', *Biometrika* **52**, 447–468.

Reinsel, G. (1984), 'Estimation and prediction in a multivariate random effects generalized linear model', *Journal of the American Statistical Association* **79**, 406–414.

Shah, A., Laird, N. & Schoenfeld, D. (1997), 'A random effects model for multiple characteristics with possibly missing data', *Journal of the American Statistical Association.* In Press.