

**Project 1 of ST 755, Spring 2008**

**Due: Thursday, 2/7/2008**

Consider the following Laird-Ware model for repeated data:

$$Y_i = X_i^T \beta + Z_i^T b_i + e_i, \quad i = 1, \dots, m,$$

where  $Y_i$  is a  $n_i \times 1$  vector of data from subject  $i$ ,  $\beta$  is the fixed effects,  $b_i \stackrel{i.i.d.}{\sim} N(0, D_{k \times k})$  is the subject-specific random effects with unstructured variance matrix  $D_{k \times k}$ , and  $e_i \sim N(0, \sigma^2 I_{n_i \times n_i})$ . It is further assumed that  $b_i$  and  $e_i$  are independent.

1. Write the model in the general form of a linear mixed model

$$Y = X\beta + Zb + e.$$

Define each symbol.

2. Assume  $X$  is of full rank and define  $\tilde{X} = X(X^T X)^{-1/2}$ . Then  $\tilde{X}^T \tilde{X} = I_{p \times p}$ , where  $p$  is the dimension of  $\beta$ . Define  $A = I - \tilde{X} \tilde{X}^T = I - X(X^T X)^{-1} X^T$ . Show that there exists an orthonormal matrix  $H$  with form  $H = [H_1, \tilde{X}]$  such that

$$A = H \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} H^T.$$

3. Let  $U = H_1^T Y$ . Then the distribution of  $U$  does not depend on  $\beta$  so the REML estimation of  $D$  and  $\sigma^2$  can be based on its marginal distribution. Show that the joint distribution of  $U$  and  $b$  is normal and free of  $\beta$ . Also show that

$$\begin{aligned} E[b|U] &= GZ^T P Y \\ \text{var}(b|U) &= G - GZ^T P Z G, \end{aligned}$$

where  $P = V^{-1} - V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1}$  is the projection matrix. (Equivalently, you need to show that  $H_1(H_1^T V H_1)^{-1} H_1^T = P$ .)

4. Develop an EM algorithm for REML estimation of  $D$  and  $\sigma^2$  and hence the estimation of  $\beta$  using the above results by treating  $b$  as missing data using  $Y_i$ ,  $X_i$  and  $Z_i$  only (you need to derive the conditional distribution of  $b$  given  $U$  in terms of data  $Y$  and  $A$ ).
5. Implement the EM algorithm you developed in (4). Your implementation should be efficient and takes the feature of the repeated data into account. For example, it is not efficient to use a huge matrix to store and invert the variance matrix for the whole data vector.
6. Use your program to find the RMLE estimates of the parameters in the following model for the cholesterol data (available in the web page <http://www4.stat.ncsu.edu/~dzhang2/st755/framing.dat>. Missing data are denoted by .)

$$cholst_{ij} = \beta_0 + \beta_1 t_{ij} + \beta_2 age_i + \beta_3 sex_i + b_{i0} + t_{ij} b_{i1} + e_{ij}, \quad i = 1, \dots, m,$$

where  $cholst_{ij}$  is the  $j$ th cholesterol measurement of subject  $i$ ,  $age_i$  is the baseline age of subject  $i$ ,  $sex_i$  is the gender of subject  $i$  (1=male, 0=female),  $t_{ij}$  is the time in year since the study. Assume the random effects  $b_{i0}$  and  $b_{i1}$  have bivariate normal distribution with mean zero and unstructured variance matrix, and  $e_{ij} \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$  independent of  $b_{i0}$  and  $b_{i1}$ .