

# Biost/Stat 571: Homework 2 Solutions

## Problem 1 (25 Points)

Consider the linear mixed model

$$Y = X\beta + Zb + \epsilon \quad (1)$$

where  $Y$  is a  $n \times 1$  vector of outcomes from  $m$  clusters (but note that  $m$  does not play a role in this problem since we are in the stacked matrix form),  $X$  and  $Z$  are  $n \times p$  and  $n \times q$  design matrices associated with the fixed effects and the random effects— respectively,  $\beta$  is a  $p \times 1$  vector of fixed effects,  $b$  is a  $q \times 1$  vector of random effects following  $b \sim N(0, D(\theta))$ , and the residuals  $\epsilon \sim N(0, R(\theta))$ . Denote  $V = \text{cov}(Y) = ZDZ^T + R$ .

### Part (1)

Given  $(\beta, \theta)$ , show the BLUP estimator  $\hat{b}_{BLUP} = DZ^TV^{-1}(Y - X\beta)$  is the empirical Bayes estimator  $\hat{b} = E(b|Y)$ .

The joint likelihood (up to a normalizing constant) for  $\beta$ ,  $\theta$  and  $b$  is given by

$$L(\beta, \theta, b | Y) = \frac{1}{|R|^{-\frac{1}{2}}|D|^{-\frac{1}{2}}} \exp \left\{ -\frac{1}{2}(Y - X\beta - Zb)^T R^{-1}(Y - X\beta - Zb) - \frac{1}{2}b^T D^{-1}b \right\}.$$

If we have fixed  $\beta$  and  $\theta$ , then we can view  $L(\beta, \theta, b|Y) = p(Y|b)p(b)$ , which is the posterior distribution for  $b$ , up to a normalizing constant, and where  $p(b)$  is based on the estimates from  $\beta$  and  $\theta$ , which makes the posterior mean the empirical Bayes estimator. To find the distribution of the posterior,  $p(b|Y)$ , we need to complete the square. Since  $\beta$  and  $\theta$  are now fixed, we can treat the matrices  $D$  and  $R$  as constant matrices and ignore them when finding the posterior distribution if they are alone. We have that

$$\begin{aligned} p(b|Y, \beta, \theta) &\propto \exp \left\{ -\frac{1}{2}(Y - X\beta - Zb)^T R^{-1}(Y - X\beta - Zb) - \frac{1}{2}b^T D^{-1}b \right\} \\ &= \exp \left\{ -\frac{1}{2} [Y^T R^{-1}Y - Y^T R^{-1}X\beta - Y^T R^{-1}Zb - \beta^T X^T R^{-1}Y + \beta^T X^T R^{-1}X\beta + \right. \\ &\quad \left. \beta^T X^T R^{-1}Zb - b^T Z^T R^{-1}Y + b^T Z^T R^{-1}X\beta + b^T Z^T R^{-1}Zb + b^T D^{-1}b] \right\} \\ &\propto \exp \left\{ -\frac{1}{2} [-Y^T R^{-1}Zb + \beta^T X^T R^{-1}Zb - b^T Z^T R^{-1}Y + b^T Z^T R^{-1}X\beta \right. \\ &\quad \left. + b^T Z^T R^{-1}Zb + b^T D^{-1}b] \right\} \\ &= \exp \left\{ -\frac{1}{2} [b^T D^{-1}b + b^T Z^T R^{-1}Zb - 2b^T Z^T R^{-1}Y + 2b^T Z^T R^{-1}X\beta] \right\} \\ &= \exp \left\{ -\frac{1}{2} [b^T (D^{-1} + Z^T R^{-1}Z)b - 2b^T Z^T R^{-1}(Y - X\beta)] \right\} \end{aligned}$$

Note that this is the kernel of a multivariate normal distribution. Let  $A_0 = (D^{-1} + Z^T R^{-1}Z)$  and  $A_1 = Z^T R^{-1}(Y - X\beta)$ . Then the mean of the normal distribution is  $A_0^{-1}A_1 = (D^{-1} + Z^T R^{-1}Z)^{-1}Z^T R^{-1}(Y - X\beta)$ . Using the Woodbury identity with positive definite matrices (which

works here since  $R$  and  $D$  are covariance matrices and thus, they are positive definite), we can re-write  $(D^{-1} + Z^T R^{-1} Z)^{-1} Z^T R^{-1} = D Z^T (Z D Z^T + R)^{-1}$ . So the empirical Bayes estimator, which is the mean of the posterior distribution, is given by

$$\begin{aligned}\hat{b} &= E[b \mid Y, \beta, \theta] = A_0^{-1} A_1 \\ &= (D^{-1} + Z^T R^{-1} Z)^{-1} Z^T R^{-1} (Y - X\beta) \\ &= D Z^T (Z D Z^T + R)^{-1} (Y - X\beta) \\ &= D Z^T V^{-1} (Y - X\beta),\end{aligned}$$

which is exactly the  $\hat{b}_{\text{BLUP}}$ .

## Part (2)

Consider the BLUP estimator of  $\beta$  and  $b$  under the linear mixed model (1), which jointly maximizes the joint likelihood, apart from a constant,

$$-\frac{1}{2}(Y - X\beta - Zb)^T R^{-1} (Y - X\beta - Zb) - \frac{1}{2}b^T D^{-1} b.$$

The BLUPs satisfy the normal equation,

$$\begin{pmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + D^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix} \quad (2)$$

Show that the BLUP estimators  $(\hat{\beta}, \hat{b})$  solving (2) also satisfy

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (3)$$

$$\hat{b} = D Z^T V^{-1} (Y - X\hat{\beta}),$$

where  $V = \text{cov}(Y) = Z D Z^T + R$ .

We will solve the normal equations directly to find the BLUP estimators. By multiplying the two matrices, we can re-write (2) as

$$\begin{pmatrix} X^T R^{-1} X \hat{\beta} + X^T R^{-1} Z \hat{b} \\ Z^T R^{-1} X \hat{\beta} + (Z^T R^{-1} Z + D^{-1}) \hat{b} \end{pmatrix} = \begin{pmatrix} X^T R^{-1} Y \\ Z^T R^{-1} Y \end{pmatrix}$$

Looking at just the second row, we want to get  $\hat{b}$  in terms of  $\hat{\beta}$ , we have

$$\begin{aligned}(Z^T R^{-1} Z + D^{-1}) \hat{b} &= Z^T R^{-1} Y - Z^T R^{-1} X \hat{\beta} \\ \Rightarrow \hat{b} &= (Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1} (Y - X\hat{\beta}) \\ \Rightarrow \hat{b} &= D Z^T (Z D Z^T + R)^{-1} (Y - X\hat{\beta}) \text{ using the Woodbury identity} \\ \Rightarrow \hat{b} &= D Z^T V^{-1} (Y - X\hat{\beta})\end{aligned} \quad (4)$$

using the fact that  $V = Z D Z^T + R$ . This is exactly our BLUP for  $b$ , assuming that  $\hat{\beta}$  is our BLUP for  $\beta$ , which we will show. We can plug  $\hat{b}$  into our first row since it is in terms of  $\hat{\beta}$  and solve for

$\hat{\beta}$ . Note that since  $V = R + ZDZ^T$ , then  $ZDZ^T = V - R$ , we will use this relation.

$$\begin{aligned}
& X^T R^{-1} X \hat{\beta} + X^T R^{-1} Z D Z^T (Z D Z^T + R)^{-1} (Y - X \hat{\beta}) = X^T R^{-1} Y \\
\implies & X^T R^{-1} X \hat{\beta} + X^T R^{-1} Z D Z^T V^{-1} Y - X^T R^{-1} Z D Z^T V^{-1} X \hat{\beta} = X^T R^{-1} Y \\
\implies & X^T R^{-1} X \hat{\beta} + X^T R^{-1} (V - R) V^{-1} Y - X^T R^{-1} (V - R) V^{-1} X \hat{\beta} = X^T R^{-1} Y \\
\implies & X^T R^{-1} X \hat{\beta} + X^T R^{-1} V V^{-1} Y - X^T R R^{-1} V^{-1} Y - X^T R^{-1} V V^{-1} X \hat{\beta} + X^T R^{-1} R V^{-1} X \hat{\beta} = X^T R^{-1} Y \\
\implies & X^T R^{-1} X \hat{\beta} + X^T R^{-1} Y - X^T V^{-1} Y - X^T R^{-1} X \hat{\beta} + X^T V^{-1} X \hat{\beta} = X^T R^{-1} Y \\
\implies & -X^T V^{-1} Y + X^T V^{-1} X \hat{\beta} = 0 \\
\implies & \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y.
\end{aligned}$$

So we see that our solutions to the normal equations for  $\hat{\beta}$  and  $\hat{b}$  are exactly the blups in (3) and (4).

### Part (3)

Calculate the likelihood of  $Y$  as  $L(\beta, \theta) = \int L(Y|b)L(b)db$ . Show that apart from a constant, the resulting log-likelihood satisfies

$$\ell(\beta, \theta) = \frac{1}{2} \ln |V| - \frac{1}{2} (Y - X\beta)^T V^{-1} (Y - X\beta),$$

where  $\ell(\beta, \theta) = \ln\{L(\beta, \theta)\}$ .

From LMM-1 slide 20, we start with the fact that since  $L(Y|b) \sim N(X\beta + Zb, R)$  and  $b \sim N(0, D(\theta))$ , then

$$L(\beta, \theta) = \frac{1}{|R|^{\frac{1}{2}}|D|^{\frac{1}{2}}} \int \exp \left\{ -\frac{1}{2} \left( (Y - X\beta - Zb)^T R^{-1} (Y - X\beta - Zb) + b^T D^{-1} b \right) \right\} db.$$

We can expand this out, and then we will proceed by completing the square and using that to compute the value of the integral is with respect to  $b$ :

$$\begin{aligned}
L(\beta, \theta) & \propto \frac{1}{|R|^{\frac{1}{2}}|D|^{\frac{1}{2}}} \int \exp \left\{ -\frac{1}{2} \left( Y^T R^{-1} Y - Y^T R^{-1} X \beta - Y^T R^{-1} Z b \right. \right. \\
& \quad \left. \left. - (X\beta)^T R^{-1} Y + (X\beta)^T R^{-1} (X\beta) + (X\beta)^T R^{-1} Z b \right. \right. \\
& \quad \left. \left. - (Zb)^T R^{-1} Y + (Zb)^T R^{-1} X \beta + (Zb)^T R^{-1} Z b + b^T D^{-1} b \right) \right\} db \\
& = \frac{1}{|R|^{\frac{1}{2}}|D|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left( (Y - X\beta)^T R^{-1} (Y - X\beta) \right) \right\} \int \exp \left\{ -\frac{1}{2} \left( b^T (Z^T R^{-1} Z + D^{-1}) b - 2b^T Z^T R^{-1} (Y - X\beta) \right) \right\} db
\end{aligned}$$

Notice that the value inside of the integral is the kernel of a multivariate normal distribution for  $b$ . We see that the quadratic and linear coefficients are  $A_0 = Z^T R^{-1} Z + D^{-1}$  and  $A_1 = Z^T R^{-1} (Y - X\beta)$ . Then the mean of the multivariate normal distribution kernel is  $A_0^{-1} A_1$  and the variance is  $A_0^{-1}$ , and

the constant term is  $A_1^T A_0^{-1} A_1$ , which is

$$\begin{aligned}
A_1^T A_0^{-1} A_1 &= (Z^T R^{-1} (Y - X\beta))^T (Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1} (Y - X\beta) \\
&= (Y - X\beta)^T R^{-1} Z (Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1} (Y - X\beta) \\
&= (Y - X\beta)^T R^{-1} Z D Z^T (Z D Z^T + R)^{-1} (Y - X\beta) \quad \text{using Woodbury identity for PD matrices} \\
&= (Y - X\beta)^T R^{-1} Z D Z^T V^{-1} (Y - X\beta) \quad \text{noting that } V = Z D Z^T + R \\
&= (Y - X\beta)^T R^{-1} (V - R) V^{-1} (Y - X\beta) \quad \text{noting that } Z D Z^T = V - R \\
&= (Y - X\beta)^T R^{-1} (Y - X\beta) - (Y - X\beta)^T V^{-1} (Y - X\beta).
\end{aligned}$$

Now that we know the value of the constant term from completing the square, and the variance  $A_1^{-1}$ , we can compute the value of the integral:

$$\begin{aligned}
L(\beta, \theta) &\propto \frac{1}{|R|^{\frac{1}{2}} |D|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left( (Y - X\beta)^T R^{-1} (Y - X\beta) \right) \right\} \\
&\quad \exp \left\{ \frac{1}{2} \left( (Y - X\beta)^T R^{-1} (Y - X\beta) - (Y - X\beta)^T V^{-1} (Y - X\beta) \right) \right\} |(Z^T R^{-1} Z + D^{-1})^{-1}|^{\frac{1}{2}} \\
&= \frac{1}{|R|^{\frac{1}{2}} |D|^{\frac{1}{2}}} \frac{1}{|Z^T R^{-1} Z + D^{-1}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left( (Y - X\beta)^T V^{-1} (Y - X\beta) \right) \right\}.
\end{aligned}$$

Lastly, we need to take care of the determinant terms. Note that using the matrix determinant lemma

$$\det(Z^T R^{-1} Z + D^{-1}) = \det(R + Z D Z^T) \det(R^{-1}) \det(D^{-1}).$$

We know that  $R + Z D Z^T = V$ , so we have  $\det(Z^T R^{-1} Z + D^{-1}) = \det(V) \det(R^{-1}) \det(D^{-1})$ . Putting this together with what we have for the likelihood, we get

$$\begin{aligned}
L(\beta, \theta) &\propto \frac{1}{|R|^{\frac{1}{2}} |D|^{\frac{1}{2}}} \frac{1}{|V|^{\frac{1}{2}} |D|^{-\frac{1}{2}} |R|^{-\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left( (Y - X\beta)^T V^{-1} (Y - X\beta) \right) \right\} \\
&= \frac{1}{|V|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} \left( (Y - X\beta)^T V^{-1} (Y - X\beta) \right) \right\}.
\end{aligned}$$

Finally, we can take the log of this (and since we had  $\propto$  in our derivation, we will obtain the log likelihood up to a constant that doesn't depend on  $\beta$  or  $\theta$ , as the question asks),

$$\begin{aligned}
\ell(\beta, \theta) &= \log \frac{1}{|V|^{\frac{1}{2}}} - \frac{1}{2} \left( (Y - X\beta)^T V^{-1} (Y - X\beta) \right) \\
&= -\frac{1}{2} \log |V| - \frac{1}{2} (Y - X\beta)^T V^{-1} (Y - X\beta).
\end{aligned}$$

## Part (4)

Consider a random intercept and slope model for longitudinal data

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}.$$

Use the BLUPs to calculate the estimator of the subject-specific trajectory  $\mu_i(t)$  and the variance estimator of  $\hat{\mu}_i(t)$ , where  $\mu_i(t) = E[Y_i(t)|b_i]$  and  $b_i = (b_{0i}, b_{1i})^T$ .

In this model we do not assume any special structure of  $\epsilon_{ij}$  or the covariance of  $b_i$ , we will work with the general case. As we saw in part (2), the BLUPs are given by

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \quad (5)$$

$$\hat{b} = D Z^T V^{-1} (Y - X \hat{\beta}), \quad (6)$$

Let  $b = (b_{01}, b_{11}, \dots, b_{0m}, b_{1m}) \in \mathbb{R}^{2m}$ , i.e. the random effects are stacked on top of each other in a vector. Then  $Z_i = [\mathbf{1}_{n_i}, \mathbf{t}_i]$  where  $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})$  is the vector of  $t$  for the  $i^{th}$  cluster and  $n_i$  is the number of observations in the  $i^{th}$  cluster. Then our  $Z$  matrix is a block-diagonal matrix with

$$Z = \begin{pmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & Z_m \end{pmatrix}.$$

If we want to form an estimate  $\mu_i(t) = E[Y_i(t)|b_i]$ , with  $b_i = (b_{0i}, b_{1i})^T$ , we can let  $\lambda_1^T = (1, t)$ . Let  $\lambda_{2i}^T = (0, \dots, 0, 1, t, 0, \dots, 0)$  where the  $(1, t)$  component is in the  $i^{th}$  position such that  $\lambda_{2i}^T b = b_i$ . Then we have

$$\begin{aligned} \hat{\mu}_i(t) &= \hat{E}[Y_i(t)|\hat{b}_i] = \lambda_1^T \hat{\beta} + \lambda_{2i}^T \hat{b} \\ &= \lambda_1^T (X^T V^{-1} X)^{-1} X^T V^{-1} Y + \lambda_{2i}^T D Z^T V^{-1} (Y - X \hat{\beta}) \\ &= \lambda_1^T (X^T V^{-1} X)^{-1} X^T V^{-1} Y + \lambda_{2i}^T D Z^T V^{-1} (Y - X (X^T V^{-1} X)^{-1} X^T V^{-1} Y). \end{aligned}$$

In order to construct the variance,  $\text{Var}[\hat{\mu}_i(t)]$ , notice that  $\hat{\mu}_i(t)$  is a sum of two linear combinations of  $\hat{\beta}$  and  $\hat{b}$ . We need to first calculate the covariance of  $\hat{\beta}$ , the covariance matrix of  $\hat{b}$ , and the covariance of  $(\hat{\beta}, \hat{b})$ . Using the fact that  $\text{Cov}[AY] = A \text{Cov}[Y] A^T$ , and recalling that  $\text{Cov}[Y] = V$ ,

$$\begin{aligned} \text{Cov}[\hat{\beta}] &= \text{Cov}[(X^T V^{-1} X)^{-1} X^T V^{-1} Y] \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} \text{Cov}[Y] (X^T V^{-1} X)^{-1} X^T V^{-1})^T \\ &= (X^T V^{-1} X)^{-1} X^T V^{-1} V V^{-1} X (X^T V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1} V^{-1} X (X^T V^{-1} X)^{-1} \\ &= (X^T V^{-1} X)^{-1}. \end{aligned}$$

Using the same strategy for  $\hat{b}$ , we get

$$\begin{aligned}
\text{Cov}[\hat{b}] &= \text{Cov}[DZ^T V^{-1}(Y - X(X^T V^{-1} X)^{-1} X^T V^{-1})Y] \\
&= \text{Cov}[DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})Y] \\
&= DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})\text{Cov}[Y](DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1}))^T \\
&= DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})V(DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1}))^T \\
&= (DZ^T V - V - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} V)(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})^T V^{-1} Z D^T \\
&= (DZ^T - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T)(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})^T V^{-1} Z D^T \\
&= (DZ^T - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T)(I - V^{-1} X(X^T V^{-1} X)^{-1} X^T)V^{-1} Z D^T \\
&= DZ^T V^{-1} Z D^T - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T \\
&\quad - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T + DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T \\
&= DZ^T V^{-1} Z D^T - DZ^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T \\
&= DZ^T V^{-1}(I - X(X^T V^{-1} X)^{-1} X^T V^{-1})Z D^T \\
&= DZ^T P Z D
\end{aligned}$$

where  $P = V^{-1} - V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1}$ , from slide 10 of LMM-2. Finally, we use the fact that  $\text{Cov}[AY, BY] = A\text{Cov}[Y]B^T$  to find  $\text{Cov}[\hat{\beta}, \hat{b}]$ ,

$$\begin{aligned}
\text{Cov}[\hat{\beta}, \hat{b}] &= \text{Cov}[(X^T V^{-1} X)^{-1} X^T V^{-1} Y, DZ^T V^{-1}(Y - X(X^T V^{-1} X)^{-1} X^T V^{-1})Y] \\
&= (X^T V^{-1} X)^{-1} X^T V^{-1} \text{Cov}[Y](DZ^T V^{-1}(Y - X(X^T V^{-1} X)^{-1} X^T V^{-1}))^T \\
&= (X^T V^{-1} X)^{-1} X^T V^{-1} V(DZ^T V^{-1}(Y - X(X^T V^{-1} X)^{-1} X^T V^{-1}))^T \\
&= (X^T V^{-1} X)^{-1} X^T (DZ^T V^{-1}(Y - X(X^T V^{-1} X)^{-1} X^T V^{-1}))^T \\
&= (X^T V^{-1} X)^{-1} X^T (I - X(X^T V^{-1} X)^{-1} X^T V^{-1})^T V^{-1} Z D^T \\
&= (X^T V^{-1} X)^{-1} X^T (I - V^{-1} X(X^T V^{-1} X)^{-1} X^T)V^{-1} Z D^T \\
&= (X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T - (X^T V^{-1} X)^{-1} X^T V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T \\
&= (X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T - (X^T V^{-1} X)^{-1} X^T V^{-1} Z D^T \\
&= 0.
\end{aligned}$$

It is convenient that the covariance between  $\hat{\beta}$  and  $\hat{b}$  is 0, because then the variance of their sum is the sum of their variances.

$$\begin{aligned}
\text{Var}[\hat{\mu}_i(t)] &= \text{Var}[\lambda_1^T \hat{\beta} + \lambda_{2i}^T \hat{b}] \\
&= \lambda_1^T \text{Var}[\hat{\beta}] \lambda_1 + \lambda_{2i}^T \text{Var}[\hat{b}] \lambda_{2i} \\
&= \lambda_1^T (X^T V^{-1} X)^{-1} \lambda_1 + \lambda_{2i}^T DZ^T P Z D^T \lambda_{2i}.
\end{aligned}$$

## Problem 2 (25 points)

### Part (1)

Consider  $n$  independent observations  $(X_i, Y_i)$  and the classical linear model  $Y_i = X_i^T \beta + \epsilon_i$ , where  $X_i$  is a  $p \times 1$  vector and  $\epsilon_i \sim N(0, \sigma^2)$ . This is a special case of Problem 1. Using the error contrasts, show

the REML estimator of  $\sigma^2$  is

$$\widehat{\sigma^2}_{\text{REML}} = \frac{1}{n-p} \sum_{i=1}^n (Y_i - X_i^T \hat{\beta})^2,$$

where  $\hat{\beta}$  is the MLE of  $\beta$ .

Define  $A \in \mathbb{R}^{n \times (n-p)}$  such that  $A^T A = I_{n-p}$  and  $AA^T = I_n - X(X^T X)^{-1} X^T$ . Such  $A$  can be explicitly defined via the spectral decomposition of  $I_n - X(X^T X)^{-1} X^T$ : since such matrix can be decomposed as  $I_n - X(X^T X)^{-1} X^T = Q\Lambda Q^T$  where  $Q$  is orthogonal and

$$\Lambda = \begin{bmatrix} I_{(n-p) \times (n-p)} & 0 \\ 0 & 0 \end{bmatrix}.$$

Letting  $A = Q \begin{bmatrix} I_{(n-p) \times (n-p)} \\ 0 \end{bmatrix}$ , we indeed have  $A^T A = I_{n-p}$  and  $AA^T = I_n - X(X^T X)^{-1} X^T$ . Furthermore,  $A$  is orthogonal to  $X$  since  $\|A^T X\|^2 = X^T AA^T X = X^T (I_n - X(X^T X)^{-1} X^T) X = 0$ .

Let  $U = A^T Y$ , then  $U \sim N(0, \sigma^2 A^T I_n A)$  which is  $N(0, \sigma^2 I_{n-p})$ . The REML log-likelihood  $\ell_{\text{REML}}(\sigma^2; y)$  is simply given by the log-likelihood  $\ell(\sigma^2; u)$ , which is

$$\begin{aligned} \ell_{\text{REML}}(\sigma^2; y) &\propto -\frac{n-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} y^T AA^T y \\ &= -\frac{n-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (AA^T y)^T AA^T y \\ &= -\frac{n-p}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - X_i^T \hat{\beta})^2. \end{aligned}$$

Taking derivative w.r.t.  $\sigma^2$  and setting to 0, we obtain

$$\hat{\sigma^2}_{\text{REML}} = \frac{1}{n-p} \sum_{i=1}^n (y_i - X_i^T \hat{\beta})^2.$$

## Part (2)

The REML likelihood of  $\theta$  is

$$\ell_{\text{REML}}(\theta) = -\frac{1}{2} \ln |X^T V^{-1} X| - \frac{1}{2} \ln |V| - \frac{1}{2} (Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta}).$$

Show  $\ell_{\text{REML}}(\theta)$  is equivalent to

$$-\frac{1}{2} \ln |X^T V^{-1} X| - \frac{1}{2} \ln |V| - \frac{1}{2} (Y - X\hat{\beta} - Z\hat{b})^T R^{-1} (Y - X\hat{\beta} - Z\hat{b}) - \frac{1}{2} \hat{b}^T D^{-1} \hat{b}.$$

It suffices to show that  $(Y - X\hat{\beta})^T V^{-1} (Y - X\hat{\beta}) = (Y - X\hat{\beta} - Z\hat{b})^T R^{-1} (Y - X\hat{\beta} - Z\hat{b}) + \hat{b}^T D^{-1} \hat{b}$ . To do this, we rewrite the RHS as

$$RHS = (Y - X\hat{\beta})^T R^{-1} (Y - X\hat{\beta}) + \hat{b}^T Z^T R^{-1} Z \hat{b} - 2\hat{b}^T Z^T R^{-1} (Y - X\hat{\beta}) + \hat{b}^T D^{-1} \hat{b} \quad (7)$$

where, recalling the normal equation in Problem 1, the [third](#) term can further be expressed as

$$\begin{aligned} -2\hat{b}^\top Z^\top R^{-1}(Y - X\hat{\beta}) &= -2(Y - X\hat{\beta})^\top V^{-1}ZDZ^\top R^{-1}(Y - X\hat{\beta}) \\ &= -2(Y - X\hat{\beta})^\top V^{-1}(V - R)R^{-1}(Y - X\hat{\beta}) \\ &= -2(Y - X\hat{\beta})^\top (R^{-1} - V^{-1})(Y - X\hat{\beta}). \end{aligned}$$

Plugging this back into [\(7\)](#), and noting that  $(D^{-1} + Z^\top R^{-1}Z)\hat{b} = Z^\top R^{-1}(Y - X\hat{\beta})$  by the normal equation in Problem 1, we obtain

$$\begin{aligned} RHS &= 2(Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) + \hat{b}^\top (D^{-1} + Z^\top R^{-1}Z)\hat{b} - (Y - X\hat{\beta})^\top R^{-1}(Y - X\hat{\beta}) \\ &= 2(Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) + (Y - X\hat{\beta})^\top (R^{-1}ZDZ^\top V^{-1})(Y - X\hat{\beta}) - (Y - X\hat{\beta})^\top R^{-1}(Y - X\hat{\beta}) \\ &= 2(Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) + (Y - X\hat{\beta})^\top [R^{-1}(V - R)V^{-1}](Y - X\hat{\beta}) - (Y - X\hat{\beta})^\top R^{-1}(Y - X\hat{\beta}) \\ &= (Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) = LHS, \end{aligned}$$

which completes our proof.

### Part (3)

*Prove the REML likelihood can be obtained by the Bayesian model assuming a flat prior for  $\beta$ , i.e.,*

$$L_{\text{REML}}(\theta) = \int L(Y; \beta, \theta) d\beta,$$

where  $\ell(Y; \beta, \theta) = \ln L(Y; \beta, \theta) = -\frac{1}{2} \ln |V| - \frac{1}{2}(Y - X\beta)^\top V^{-1}(Y - X\beta)$ .

First, we rewrite

$$\begin{aligned} (Y - X\beta)^\top V^{-1}(Y - X\beta) &= [(Y - X\hat{\beta}) + X(\hat{\beta} - \beta)]^\top V^{-1}[(Y - X\hat{\beta}) + X(\hat{\beta} - \beta)] \\ &= (Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) + (\hat{\beta} - \beta)^\top X^\top V^{-1}X(\hat{\beta} - \beta) \end{aligned}$$

where the second “=” holds because the cross term is 0, which follows from the fact that

$$\begin{aligned} (Y - X\hat{\beta})^\top V^{-1}X(\hat{\beta} - \beta) &= \left[ Y^\top - Y^\top V^{-1}X \left( X^\top V^{-1}X \right)^{-1} X^\top \right] V^{-1}X(\hat{\beta} - \beta) \\ &= Y^\top V^{-1}X(\hat{\beta} - \beta) - Y^\top V^{-1}X \left( X^\top V^{-1}X \right)^{-1} X^\top V^{-1}X(\hat{\beta} - \beta) \\ &= Y^\top V^{-1}X(\hat{\beta} - \beta) - Y^\top V^{-1}X(\hat{\beta} - \beta) = 0. \end{aligned}$$

Therefore, we obtain (apart from a constant)

$$\begin{aligned} \int L(Y; \beta, \theta) d\beta &\propto |V|^{-\frac{1}{2}} \int \exp \left[ -\frac{1}{2}(Y - X\beta)^\top V^{-1}(Y - X\beta) \right] d\beta \\ &= |V|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2}(Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) \right] \cdot \int \exp \left[ -\frac{1}{2}(\hat{\beta} - \beta)^\top X^\top V^{-1}X(\hat{\beta} - \beta) \right] d\beta \\ &\propto |V|^{-\frac{1}{2}} \exp \left[ -\frac{1}{2}(Y - X\hat{\beta})^\top V^{-1}(Y - X\hat{\beta}) \right] \left| (X^\top V^{-1}X)^{-1} \right|^{\frac{1}{2}} \\ &= L_{\text{REML}}(\theta) \end{aligned}$$

where the second “ $\propto$ ” holds because the [blue](#) term is the kernel of  $N(\hat{\beta}, (X^\top V^{-1}X)^{-1})$ .



## Part (4)

Consider the random intercept model for longitudinal data

$$Y_{ij} = X_{ij}^T \beta + b_i + \epsilon_{ij},$$

where  $b_i \sim N(0, \theta)$  and  $\epsilon_{ij} \sim N(0, \sigma^2)$ ,  $i$  indicates subject  $i$  ( $i = 1, \dots, m$ ) and  $j$  indicates the  $j$ th repeated measure ( $j = 1, \dots, n$ ). Suppose  $m=200$  and  $n = 3$ ,  $X_{ij} = (1, j)^T$ , where  $j$  indicates time,  $\beta = (1, 0.5)^T$ ,  $\theta = 1$ ,  $\sigma^2 = 1$ . Read Harville (1977, JASA, 72, pp 320–338). Write R code to simulate one data set and implement R functions using Newton-Raphson type procedure (use of optimization functions, e.g. the `optim` function, in R will only receive partial credit) to calculate the MLE of  $\beta$ , and the ML and the REML estimates of  $\theta$  and  $\sigma^2$ . Compare your results with those output from mixed models software.

We write  $\gamma = (\gamma_1, \gamma_2)^T := (\theta, \sigma^2)^T$ . Starting from an initial value  $\gamma^{(0)}$  and  $\beta^{(0)} = (X^T V^{(0)-1} X)^{-1} X^T V^{(0)-1} Y$ , where  $V^{(0)} := V(\gamma^{(0)})$  depends on  $\gamma^{(0)}$ , we iteratively update

$$\begin{aligned}\gamma^{(t)} &= \gamma^{(t-1)} + I(\gamma^{(t-1)})^{-1} U(\gamma^{(t-1)}) \\ \beta^{(t)} &= (X^T V(\gamma^{(t)})^{-1} X)^{-1} X^T V(\gamma^{(t)})^{-1} Y\end{aligned}$$

for  $t = 1, 2, \dots$  until convergence. In our expressions,

$$\begin{aligned}U(\gamma) &:= (U_1, U_2), \quad I(\gamma) := \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \quad \text{with} \\ U_j &= \frac{1}{2} \left[ -\text{tr} \left( P \frac{\partial V}{\partial \gamma_j} \right) + (Y - X\beta)^T V^{-1} \frac{\partial V}{\partial \gamma_j} V^{-1} (Y - X\beta) \right], \\ I_{ij} &= \frac{1}{2} \text{tr} \left( P \frac{\partial V}{\partial \gamma_i} P \frac{\partial V}{\partial \gamma_j} \right) \\ P &= V^{-1} - V^{-1} X (X^T V^{-1} X)^{-1} X^T V^{-1}\end{aligned}$$

for REML, and we replace  $P$  with  $V^{-1}$  for ML.

In this specific problem, we have  $V(\theta, \sigma^2) = \theta Z Z^T + \sigma^2 I$  where  $Z$  is a block matrix with each block being  $(1, 1, 1)^T$ , and further  $\partial V / \partial \gamma_1 = \partial V / \partial \theta = Z Z^T$ ,  $\partial V / \partial \gamma_2 = \partial V / \partial \sigma^2 = I$ .

The code for this problem is in the Appendix. Table 1 compares our estimates with those from R package `lme4`, which are close to each other.

## Part (5) (Extra Credit)

Consider the random intercept model in (4). Show that the likelihood ratio test for  $H_0 : \theta = 0$  vs  $H_1 : \theta > 0$  follows  $0.5\chi_1^2 + 0.5\chi_0^2$ . Perform a simulation study to verify the results.

We first derive the asymptotic distribution of  $\hat{\theta}$  (we take ML as example but the argument for REML is similar), and in particular, argue that the distribution of  $\sqrt{m}(\hat{\theta} - \theta_0)$  converges to the mixture of a  $N(0, I^{-1}(\theta))$  distribution left-truncated at 0, and a point mass at 0, denoted as

$$\frac{1}{2} N^+(0, I^{-1}(\theta)) + \frac{1}{2} \delta_0,$$

	Our implementation		R package	
	REML	ML	REML	ML
$\beta_1$	1.1497	1.1497	1.1497	1.1497
$\beta_2$	0.4169	0.4169	0.4169	0.4169
$\theta$	1.0166	1.0108	1.0166	1.0108
$\sigma^2$	0.9023	0.9000	0.9023	0.9000

Table 1: Estimates of  $\beta, \theta, \sigma^2$

where we shall note that  $I(\theta)$  is the Fisher information for one cluster instead of  $m$  clusters.

To see why this holds, we first consider a closely related problem – testing a covariance parameter in marginal model. In particular, we consider testing  $\theta^\dagger = \theta_0^\dagger := 0$  versus  $\theta^\dagger \neq 0$  in the marginal model

$$Y_i = X_i^\top \beta + \epsilon_i, \quad i = 1, \dots, m$$

where the  $n$ -dimensional error vector  $\epsilon_i \sim N(0, V^\dagger)$  with  $V^\dagger = \sigma^2 I_n + \theta^\dagger \mathbf{1}_n$  ( $\mathbf{1}_n$  is an  $n \times n$  matrix with all entries being 1). We notice that this data generating mechanism is the same as our LMM (since the marginal variance  $V^\dagger$  has the same form as an LMM would induce), and thus the expressions of the log-likelihood  $\ell(\cdot)$ , score  $U(\cdot)$  and Fisher information  $I(\cdot)$  are also the same for these two models ( $U(\theta_0)$  and  $I(\theta_0)$  are defined via one-sided derivatives for our LMM). Furthermore, testing  $\theta^\dagger = 0$  is the same problem as testing  $\theta = 0$  **except that**  $\theta^\dagger = 0$  is in the interior of its model space (since  $V^\dagger$  is still well-defined for some  $\theta^\dagger < 0$ ) while  $\theta$  is on the boundary.

The asymptotic distribution of  $\hat{\theta}^\dagger$  follows from typical large-sample theory, i.e.

$$\sqrt{m}(\hat{\theta}^\dagger - \theta_0^\dagger) \xrightarrow{d} N(0, I^{-1}(\theta^\dagger)). \quad (8)$$

When  $\hat{\theta}^\dagger \geq 0$ , we know that such  $\hat{\theta}^\dagger$  maximizes the log-likelihood (for both the marginal model and the LMM), and thus  $\hat{\theta} = \hat{\theta}^\dagger$ . When  $\hat{\theta}^\dagger < 0$ , we know that  $U(\hat{\theta}^\dagger) = 0$  and since

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \theta^{\dagger 2}} &= -\frac{1}{2} \text{tr} \left( -V^{-1} \frac{\partial V}{\partial \theta^\dagger} V^{-1} \frac{\partial V}{\partial \theta^\dagger} \right) - \frac{1}{2} \text{tr} \left( V^{-1} \frac{\partial^2 V}{\partial \theta^{\dagger 2}} \right) \\ &\quad - 2 \times \frac{1}{2} (Y - X\beta)^T V^{-1} \frac{\partial V}{\partial \theta^\dagger} V^{-1} \frac{\partial V}{\partial \theta^\dagger} V^{-1} (Y - X\beta) + \frac{1}{2} (Y - X\beta)^T V^{-1} \frac{\partial^2 V}{\partial \theta^{\dagger 2}} V^{-1} (Y - X\beta) < 0 \end{aligned}$$

with probability approaching 1 as  $m$  grows, we have that  $\partial \ell / \partial \theta^\dagger < 0$  for all  $\theta^\dagger$ . Thus, if  $\hat{\theta}^\dagger < 0$  then in the constrained model space of our LMM, the log-likelihood is monotonically decreasing, and the ML estimator is  $\hat{\theta} = 0$ . Equivalently,

$$\hat{\theta} = \hat{\theta}^\dagger \cdot \mathbb{1} \left\{ \hat{\theta}^\dagger \geq 0 \right\}$$

where  $\mathbb{1}(\cdot)$  is the indicator function. By (8),  $\mathbb{1} \left\{ \hat{\theta}^\dagger \geq 0 \right\} \xrightarrow{d} \text{Bernoulli}(1/2)$ , and therefore

$$\sqrt{m}(\hat{\theta} - \theta_0) \xrightarrow{d} \frac{1}{2} N^+(0, I^{-1}(\theta)) + \frac{1}{2} \delta_0$$

as desired. This completes the first part of our proof (and meanwhile, implies that  $\hat{\theta}$  is consistent for  $\theta_0$ ).

Next, by a second order Taylor expansion of  $\ell$  at  $\hat{\theta}$  (when  $\hat{\theta} = 0$  this is based on single-sided derivatives), we obtain

$$\begin{aligned} -\log \lambda_m &:= -\left[\ell(\theta_0) - \ell(\hat{\theta})\right] = (\hat{\theta} - \theta_0) \frac{\partial \ell(\hat{\theta})}{\partial \theta} - \frac{(\hat{\theta} - \theta_0)^2}{2} \frac{\partial^2 \ell(\hat{\theta})}{\partial \theta^2} + o\left((\hat{\theta} - \theta_0)^2\right) \\ &= -\frac{(\hat{\theta} - \theta_0)^2}{2} \frac{\partial^2 \ell(\hat{\theta})}{\partial \theta^2} + o_P(1) \end{aligned}$$

under the null  $\theta_0 = 0$ , where the first order term cancels in the last step because we have  $\hat{\theta} - \theta_0 = 0$  when  $\hat{\theta} = 0$ , and  $\partial \ell(\hat{\theta})/\partial \theta = 0$  when  $\hat{\theta} > 0$ . Furthermore, since  $\frac{1}{m} \partial^2 \ell(\hat{\theta})/\partial \theta^2 \xrightarrow{P} -I(\theta_0)$ , and the square of  $N^+(0, I^{-1})$  and of  $N(0, I^{-1})$  are identically distributed, we have

$$-2 \log \lambda_m = m(\hat{\theta} - \theta_0)^2 I(\theta_0) + o_P(1) \xrightarrow{d} \frac{1}{2} \chi_1^2 + \frac{1}{2} \chi_0^2,$$

where  $\chi_0^2$  is just a point mass at 0.

We simulate data from the mechanism described in Part (4) (except that we let  $b = 0$  for each individual) and calculate the likelihood ratio test statistic  $-2 \log \lambda_m$ . We repeat this process for 500 times, and plot the empirical versus theoretical distribution of the test statistic in Figure 1 (code in Appendix).

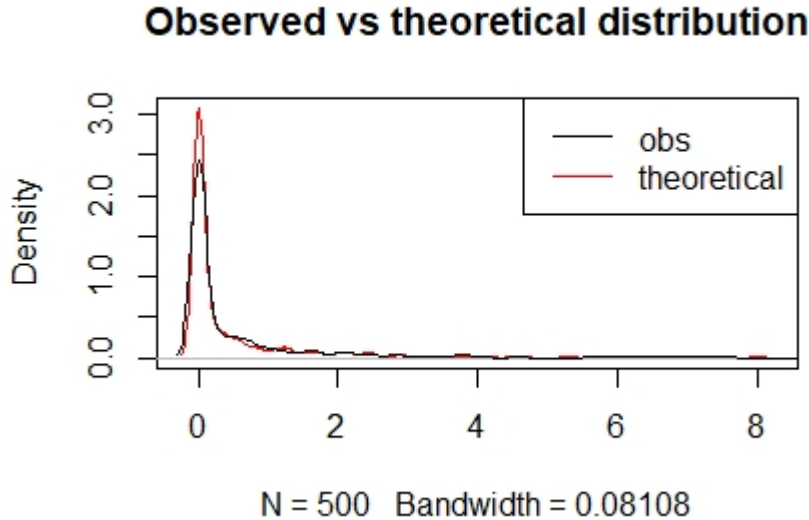


Figure 1: Density plot for empirical versus theoretical distribution of  $-2 \log \lambda_m$

### Problem 3 (25 points)

*ChatGPT has proven to be a new, potentially useful tool for a wide range of tasks. Please see the introduction to ChatGPT here: <https://openai.com/blog/chatgpt/> and sign up for a free account from OpenAI. One interesting functionality is the ability to recommend particular analytic approaches for particular data analyses. However, ChatGPT is simply a tool (just like Google) and not infallible. Please log into ChatGPT and play around. Then for this assignment, ask ChatGPT for advice on whether you should use REML or ML (carefully consider how you word the prompt).*

1. *Provide the prompt you provided to ChatGPT as well as the corresponding response. Note that this should be unique for each person in the class so don't copy someone else's response.*
2. *Assess (verify or disprove) the accuracy of what ChatGPT tells you. You may do this analytically (i.e. mathematically) or by conducting simulations.*

*Note that the system is very busy and you may get a message that their system is at capacity, in which case you may need to wait. If this continues, find a classmate who has access and you may ask them to run your prompt for you. If you continue to struggle in accessing ChatGPT, please contact the instructor. DO NOT WAIT UNTIL THE LAST FEW DAYS.*

#### Part (1)(2)

Your answer needs to contain three parts (1) the prompt and answer from ChatGPT, (2) your interpretation of the incorrect parts in the answer, and (3) an analytical or simulation study to verify your claim.

I want to fit a linear mixed effects model but I am not sure if I should use REML or ML for estimation. Could you give me some suggestions in short sentences? Thanks.

REML (restricted maximum likelihood) is generally preferred when the residual variance is unknown and needs to be estimated from the data, while ML (maximum likelihood) is used when the residual variance is known or specified. REML provides more efficient estimates of the random effects variances, but ML is computationally simpler and has a lower bias. The choice between REML and ML largely depends on the research question and the available data.

Figure 1: ChatGPT response.

Figure shows the answer I got from chatGPT. From the answer given by chatGPT, we can see that the claim 'ML has a lower bias' is not necessarily correct. We will now use a set of simulation to demonstrate the result.

We consider the following true model:

$$y_{ij} = \beta_0 + x_i\beta_1 + b_i + \epsilon_{ij}, i = 1, \dots, n, j = 1, \dots, m$$

where  $b_i \stackrel{i.i.d}{\sim} N(0, \sigma_b^2)$ ,  $\epsilon_{ij} \stackrel{i.i.d}{\sim} N(0, \sigma_e^2)$ ,  $\beta_0 = 1, \beta_1 = 2, \sigma_b^2 = \sigma_e^2 = 1$ . To generate data, we sample  $x_i$  independently from a uniform distribution on  $[-1, 1]$ . We consider combinations of  $n \in \{5, 10, 20\}$  and  $m \in \{5, 10, 25\}$ . For each combination of  $n$  and  $m$  we fit the correct model with REML and ML and record the mean and standard deviation of the estimation bias for  $\sigma_b^2$ , over 500 replications. The results are shown in Table . We observe that for most scenarios the averaged bias for  $\sigma_b^2$  under REML is lower than ML. The difference is more significant under cases when the number of data (both in terms of clusters) is small.

	$n$	$m$	$\text{bias}_{ML}(\hat{\sigma}_b^2)$	$\text{bias}_{REML}(\hat{\sigma}_b^2)$
1	5	5	-0.472(0.551)	-0.002(0.932)
2	10	5	-0.255(0.484)	-0.019(0.604)
3	20	5	-0.128(0.377)	-0.009(0.419)
4	5	10	-0.463(0.513)	-0.044(0.860)
5	10	10	-0.222(0.422)	-0.003(0.527)
6	20	10	-0.099(0.333)	0.012(0.370)
7	5	25	-0.388(0.544)	0.047(0.907)
8	10	25	-0.216(0.427)	-0.010(0.534)
9	20	25	-0.109(0.321)	-0.005(0.357)

Table 1: The average and standard deviation of estimation bias for  $\sigma_b^2$  over 500 replications for *ML* and *REML*. The standard deviation is shown in brackets.

#### Code:

```
library(lme4)
library(dplyr)
library(xtable)

# simulation setting
n_groups_choices <- c(5,10,20) # number of cluster
group_size <- c(5,10,25) # cluster sizes
pars_comb <- expand.grid(n_groups_choices, group_size)
colnames(pars_comb) <- c("n_groups", "group_size")
#pars_comb <- pars_comb[pars_comb$n_groups > pars_comb$group_size, ]

# replication number
NUM_REPLICATION <- 500
FIXED_EFFECT <- 2
RESIDUAL_SD <- 1
RANDOM_SD <- 1

n_combs <- dim(pars_comb)[1]
bias_ML_all <- bias_REML_all <- bias_ML_sd_all <- bias_REML_sd_all <- NA
set.seed(123)
for (t in 1:n_combs) {
  # Generate random intercepts and slopes
  set.seed(123)
  n_groups <- pars_comb$n_groups[t]
  group_size <- pars_comb$group_size[t]
  # biases
  bias_REML <- NA
```

```

bias_ML <- NA

for (b in 1:NUM_REPLICATION) {
  group_intercepts <- rnorm(n_groups, mean = 0, sd = RANDOM_SD)
  # Generate within-group errors
  errors <- rnorm(n_groups * group_size, mean = 0, sd = RESIDUAL_SD)
  # Generate response variable
  x <- rep(runif(n_groups,-1,1), each = group_size)
  y <- group_intercepts[rep(1:n_groups, each = group_size)] + FIXED_EFFECT * x + 1 + errors
  # Create data frame
  df <- data.frame(y = y, x = x, group = as.factor(rep(1:n_groups, each = group_size)))
  # Fit a linear mixed effects model
  fit.reml <- lmer(y ~ x + (1 | group), data = df, REML = TRUE)
  fit.ml <- lmer(y ~ x + (1 | group), data = df, REML = FALSE)

  VarCorr(fit.reml) |> data.frame() |> select(vcov)
  VarCorr(fit.ml) |> data.frame() |> select(vcov)

  bias_ML[b] <- (VarCorr(fit.ml) |> data.frame() |> select(vcov))[1,] - 1
  bias_REML[b] <- (VarCorr(fit.reml) |> data.frame() |> select(vcov))[1,] - 1
}
bias_ML_all[t] <- mean(bias_ML)
bias_REML_all[t] <- mean(bias_REML)
bias_ML_sd_all[t] <- sd(bias_ML)
bias_REML_sd_all[t] <- sd(bias_REML)
}

pars_comb$bias_ML <- bias_ML_all
pars_comb$bias_REML <- bias_REML_all
pars_comb$bias_ML_sd <- bias_ML_sd_all
pars_comb$bias_REML_sd <- bias_REML_sd_all

pars_comb |> xtable::xtable(digits = 3)

```

## Problem 4 (25 points)

*The Framingham study is one of the well known long term follow-up study to identify the relationship between various risk factors and diseases and to characterize the natural history of the chronic circulatory disease process. The data on various aspects have been and continue to be collected every two years on a cohort of individuals. It began in 1948 in Framingham, located 21 miles west of Boston, with limited goals of investigating the serum cholesterol, smoking and elevated blood pressure as the risk factors of coronary heart disease. Over the years its goal has been greatly expanded to aid in understanding the numerous etiological factors of various diseases.*

*The data `framingham.dat`, which can be downloaded from the class website, is a subset of a large data base collected in the Framingham study over years. There are 12 columns in the data file. The 1st column gives the age of the individual when they entered the study. The 2nd column provides the gender of the individual( 1-male, 2-female); the 3rd and 4th columns provide body mass index (BMI) at the baseline and at 10 years from the baseline respectively; the 5th column provides the number of cigarettes per day the individual smoked at the baseline. The columns 6 – 11 provide serum cholesterol levels at the baseline(enrollment) and then every two years through year 10. The column 12 indicates whether the individual is alive(0) or dead(1) at the end of 30 years since enrollment. That is, the data set excludes those who died during the 10 year data collection period. -9 indicates the missing data. Note that you have to convert -9 to NA or empty entries before you do any analysis.*

### Part (1)

*Fit an appropriate linear mixed model to study how cholesterol level changes over time and how it is related to age, gender, and BMI.*

Before conducting the analysis, we convert the data from wide to long format. We conduct a complete-case analysis, i.e. all observations containing NA's will be removed, which leads to a dataset with 13650 measurements in total (removed 2154 measurements).

Looking at the trajectories of cholesterol level for a subset of randomly sampled subjects (Figure 5), we notice that the trend of cholesterol level could be different for different subjects, and therefore choose to include a random intercept and a random slope for time in our model.

Also, the plot of cholesterol level versus age (baseline age plus time in study) for male and female (Figure 6) shows that the rate of change over time are different for male versus female. This motivates us to include an interaction term between gender and time in our model.

Based on this quick exploratory analysis, we fit the model

$$\mathbb{E}[Y_{ij} | b_i] = \beta_0 + \beta_1 age0_i + \beta_2 gender_i + \beta_3 BMI0_i + \beta_4 time_j + \beta_5 time_j \times gender_i + b_{0i} + b_{1i} time_j$$

where  $Y_{ij}$  is the cholesterol level of individual  $i$  at time  $j$ ,  $age_0$  represents baseline age, and  $BMI0$  represents baseline BMI. Estimates for fixed effects are summarized in Table 2.

These effect sizes are the expected difference in cholesterol level associated with one unit increment for continuous covariates, and the difference comparing female versus male for covariate “gender”. From these estimates, we observe that older age, being male (though such effect is only marginally significant),

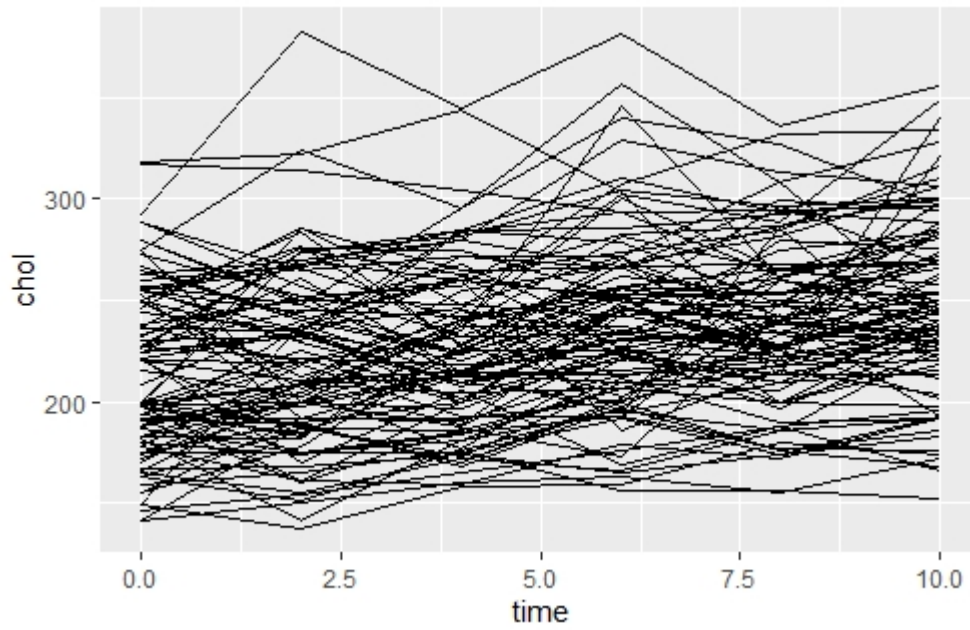


Figure 5: Trajectories of cholesterol level for 100 randomly selected subjects

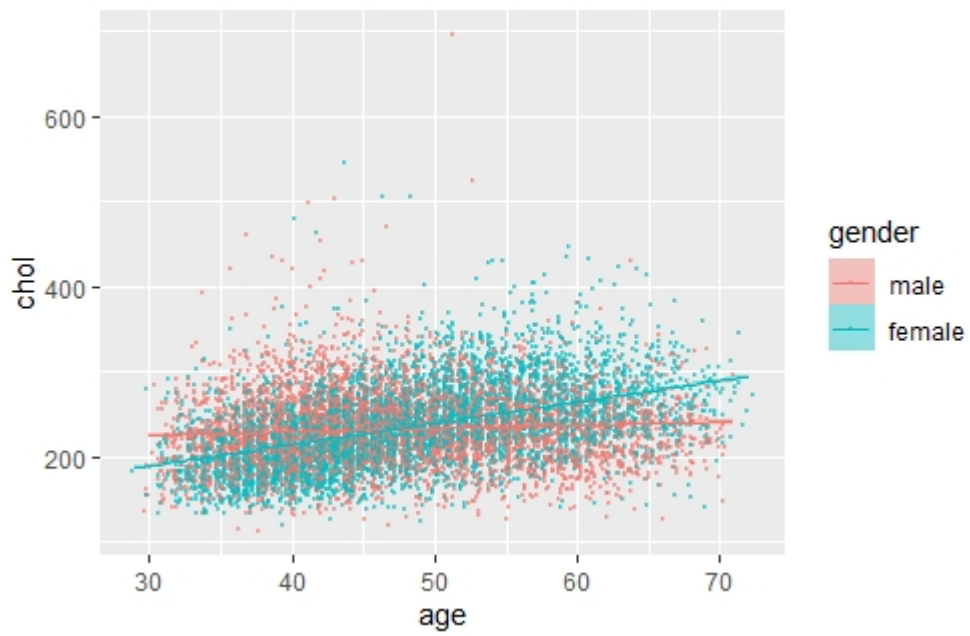


Figure 6: Change in cholesterol level for male and female



and higher baseline BMI are associated with higher cholesterol level. Also, cholesterol level is, overall, increasing over time, and the rate is on average faster for female.

The estimated standard deviation is 35.37 for the random intercept, 2.03 for the random slope and 21.11 for the error term.

	Estimate	SE	p-value
Baseline age	1.31	0.09	<.001
Gender (female)	-2.63	1.53	.086
Baseline BMI	0.94	0.17	<.001
Time	2.34	0.10	<.001
Gender:Time	1.11	0.14	<.001

Table 2: Estimated fixed effects

## Part (2)

*Use available software to conduct an analysis which allows you to investigate how baseline cholesterol level and the change rate of cholesterol level during the 10 year period affect the 30-year risk of death adjusted for age, gender, and BMI. (Hint: use the results in Problem 1).*

According to Problem 1, the change rate of cholesterol level for subject  $i$  is

$$rate_i = \beta_4 + \beta_5 gender_i + b_{1i}.$$

We fit a logistic regression model

$$\log \frac{p_i}{1 - p_i} = \gamma_0 + \gamma_1 age0_i + \gamma_2 gender_i + \gamma_3 BMI0_i + \gamma_4 chol0_i + \gamma_5 rate_i$$

where  $p_i$  is the probability of death by year 30 for individual  $i$ , and  $chol0$  is baseline cholesterol level.

Model coefficients are summarized as odds ratios, shown in Table 3. We observe that increased baseline cholesterol level is associated with little difference in the risk of death after adjusting for age, gender and baseline BMI, and such association is not statistically significant. The odds ratio of death by year 30 comparing two individuals who differ in 1 unit in their change rate of cholesterol level is 0.99 (95% CI 0.90 - 1.07) and again, such association is not significant. In addition, we found that older age at baseline (OR 1.11, 95% CI 1.10 - 1.13) and being female (OR 0.46, 95% CI 0.36 - 0.57) are significantly associated with lower risk of death by year 30.

	OR	CI lower	CI upper	p-value
Baseline age	1.11	1.10	1.13	0.00
Gender (female)	0.46	0.36	0.57	0.00
Baseline BMI	1.02	0.99	1.04	0.16
Baseline cholesterol level	1.00	1.00	1.00	0.14
Change rate of cholesterol level	0.99	0.90	1.07	0.74

Table 3: Estimated odds ratios and p-values

## Part (3)

*Suppose you have time to develop your own program, what is a better and more systematic approach (or model) would you like to propose to answer the question in (2)?*

A potentially better and more systematic approach would be to consider a survival model, particularly if we could narrow down the time when death happened or if we had information on those individuals who died during the 10-year data collection period.

The question in Part (2) was phrased more like a causal question, while we were only able to model association given limited information. It would be helpful if we had data on other potential confounders, and/or to adopt causal inference techniques.

Finally, instead of conducting a complete-case analysis which assumes the missingness happened complete at random, we could have a more careful treatment of the missing values, e.g. using multiple imputation, pattern mixture model or selection model (to be covered later this quarter!).

## Appendix: R code

### Problem 2

#### Part (4)

```
library(Matrix)
library(lme4)
library(mvtnorm)
set.seed(2333)

m = 200
n = 3
theta = 1
sigsq = 1
bt = c(1, 0.5)
b = rnorm(m, mean = 0, sd = sqrt(theta))
eps = rnorm(m*n, mean = 0, sd = sqrt(sigsq))
X = cbind(1, rep(1:n, m))
Y = X %*% bt + rep(b, each = n) + eps
dat = data.frame(x = X[,2], y = Y, id = rep(1:m, each = n))

tr = function(m)
  sum(diag(m))

score.info = function(bt, theta, sigsq, method){
  # uncomment the following four lines when using this function for Q2 Part (5)
  #n = parent.frame()$n
```

```

#m = parent.frame()$m
#Y = parent.frame()$Y
#X = parent.frame()$X

Z = bdiag(rep(list(matrix(rep(1,n), ncol = 1)) ,m))
V = theta * Z %*% t(Z) + sigsq * diag(m*n)
Vinv = solve(V)
X_Vinv_X_inv = solve(t(X) %*% Vinv %*% X)
if (method == "ML"){
  P = Vinv
} else if (method == "REML"){
  P = Vinv - Vinv %*% X %*% X_Vinv_X_inv %*% t(X) %*% Vinv
}
Vinv_res = Vinv %*% (Y - X%*% bt)
U1 = as.numeric(1/2 * (-tr(P %*% Z%*%t(Z)) + t(Vinv_res) %*% Z%*%t(Z) %*% Vinv_res))
U2 = as.numeric(1/2 * (-tr(P) + t(Vinv_res) %*% Vinv_res))
I11 = as.numeric(1/2 * (tr(P%*%Z%*%t(Z)%*%P%*%Z%*%t(Z))))
I22 = as.numeric(1/2 * (tr(P%*%P)))
I12 = as.numeric(1/2 * (tr(P%*%Z%*%t(Z)%*%P)))
return(list(score = matrix(c(U1, U2), ncol = 1),
  info = matrix(c(I11, I12, I12, I22), ncol = 2),
  Vinv = Vinv))
}

soln = function(method){
  tol = 1e-6
  max.iter = 50
  stop = FALSE
  theta.est = 10
  sigsq.est = 10
  bt.est = c(10, 10)
  iter = 0

  while(!stop){
    iter = iter + 1
    bt.old = bt.est
    theta.old = theta.est
    sigsq.old = sigsq.est

    lst = score.info(bt.est, theta.est, sigsq.est, method)
    bt.est = as.vector(solve(t(X)%*%lst$Vinv%*%X, t(X)%*%lst$Vinv%*%Y))
    theta.sigsq = c(theta.est, sigsq.est) + solve(lst$info, lst$score)
    theta.est = theta.sigsq[1]
    sigsq.est = theta.sigsq[2]

    if (max(abs(c(bt.old - bt.est, sigsq.old - sigsq.est, theta.old - theta.est))) < tol

```

```

        | iter > max.iter)
    stop = TRUE
}
return(list(iter = iter, bt.est = bt.est, theta.est = theta.est, sigsq.est = sigsq.est))
}

reml = soln("REML")
ml = soln("ML")

reml.mod = lmer(y ~ x + (1|id), data = dat, REML = TRUE)
ml.mod = lmer(y ~ x + (1|id), data = dat, REML = FALSE)

```

## Part (5)

```

sim.one = function(seed){
  set.seed(seed)
  m = 200
  n = 3
  theta = 1
  sigsq = 1
  bt = c(1, 0.5)
  X = cbind(1, rep(1:n, m))
  Y = X %*% bt + rnorm(m*n, mean = 0, sd = sqrt(sigsq))

  tol = 1e-6
  max.iter = 50
  stop = FALSE
  theta.est = 5
  iter = 0

  while(!stop){
    iter = iter + 1
    theta.old = theta.est

    lst = score.info(bt, theta.est, sigsq, "ML")
    theta.est = theta.est + lst$score[1]/lst$info[1,1]

    if (abs(theta.est - theta.old) < tol | iter > max.iter)
      stop = TRUE
  }
  #theta.est = max(0, theta.est)

  if (theta.est > 0 & iter < max.iter){
    V0 = diag(sigsq, m*n)
    Z = bdiag(rep(list(matrix(rep(1,n), ncol = 1)), m))
  }
}

```

```

V.est = theta.est * Z %*% t(Z) + sigsq * diag(m*n)
mu = X %*% bt
test.stat = -2 * (dmvnorm(c(Y), mean = c(mu), sigma = V0, log = TRUE, checkSymmetry = FALSE)
                  - dmvnorm(c(Y), mean = c(mu), sigma = V.est, log = TRUE, checkSymmetry = FALSE))
} else if (iter < max.iter) {
  test.stat = 0
} else {
  test.stat = NA
}
c(test.stat, theta.est)
return(test.stat)
}

ts = sapply(1:500, sim.one)
ts0 = rbinom(500, 1, .5) * rchisq(500, df = 1)
plot(density(ts0), col = 'red', main = "Observed vs theoretical distribution")
lines(density(ts))
legend("topright", c("obs","theoretical"), lty = c(1,1), col = c("black", "red"))

```

## Problem 4

```
library(ggplot2)
library(nlme)
rm(list = ls())

set.seed(2333)
dat.wide = read.table("data/framingham.dat",
                      col.names = c("age0", "gender", "bmi0", "bmi10", "cigarette",
                                     "chol_0", "chol_2", "chol_4", "chol_6", "chol_8",
                                     "chol_10", "death"))

dat.wide$id = 1:nrow(dat.wide)
dat.wide$gender = as.factor(dat.wide$gender)
levels(dat.wide$gender) = c("male", "female")
dat = reshape(dat.wide, idvar = c("id", "age0", "gender", "bmi0", "bmi10", "cigarette", "death"),
              varying = paste("chol_", seq(0, 10, 2), sep = ""),
              direction = "long", v.names = "chol",
              times = seq(0, 10, 2))

dat[dat == -9] = NA
dat = dat[complete.cases(dat),]

ggplot(data = dat[dat$id %in% sample(unique(dat$id), 100),],
       aes(x = time, y = chol, group = id)) +
  geom_line()

ggplot(data = dat, aes(x = time + age0, y = chol, color = gender)) +
  geom_point(position = "jitter", size = .1, alpha = .5) +
  geom_smooth(method = lm, aes(fill = gender), size = .5) +
  labs(x = "age", y = "chol")

mod = lme(chol ~ age0 + gender + bmi0 + time + time*gender,
          random = ~ 1 + time|id, data = dat)
```

```

summary(mod)

dat.wide = dat.wide[dat.wide$id %in% unique(dat$id), ]
dat.wide$rate = summary(mod)$coefficients$fixed['time'] +
  (dat.wide$gender == "female")*summary(mod)$coefficients$fixed['genderfemale:time'] +
  summary(mod)$coefficients$random$id[,2]

mod2 = glm(death ~ age0 + gender + bmi0 + chol_0 + rate,
           family = 'binomial', data = dat.wide)
tab.out = data.frame(summary(mod2)$coefficients[-1,-3])
tab.out$lower = exp(tab.out$Estimate - 1.96*tab.out$Std..Error)
tab.out$upper = exp(tab.out$Estimate + 1.96*tab.out$Std..Error)
tab.out$OR = exp(tab.out$Estimate)
tab.out = tab.out[, c(6,4,5,3)]
xtable::xtable(tab.out)

```