Biost/Stat 571: Homework # 2

Due 5pm, Fri February 3 via Canvas

Note: Homework should be submitted in as a PDF document with problems clearly labeled and in order. Use of Latex is preferred, but hand-written math is acceptable. In the case of the latter, handwriting must be in clearly legible print. Illegible (unreadable by the TAs and instructors) and unclear work will not receive credit. Clarity in derivations and exposition count as these are essential in any professional setting.

Problem 1. Consider the linear mixed model

$$Y = X\beta + Zb + \epsilon, \tag{1}$$

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, β 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 $V = cov(Y) = ZDZ^T + R$.

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

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

$$-\frac{1}{2}(Y - X\beta - Zb)^TR^{-1}(Y - X\beta - Zb) - \frac{1}{2}b^TD^{-1}b.$$

The BLUPs satisfy the normal equation

$$\begin{pmatrix} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + D^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{b} \end{pmatrix} = \begin{pmatrix} X^TR^{-1}Y \\ Z^TR^{-1}Y \end{pmatrix}. \tag{2}$$

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

$$\begin{aligned} \widehat{\beta} &= (X^TV^{-1}X)^{-1}X^TV^{-1}Y \\ \widehat{b} &= DZ^TV^{-1}(Y - X\widehat{\beta}), \end{aligned}$$

where $V = cov(Y) = ZDZ^T + R$.

(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

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

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

(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 $b_i = (b_{0i}, b_{1i})^T$.

Problem 2. (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 σ^2 is

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

where $\widehat{\beta}$ is the MLE of β.

(2) The REML likelihood of θ is

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

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

$$-\frac{1}{2}\ln|X^TV^{-1}X| - \frac{1}{2}\ln|V| - \frac{1}{2}(Y - X\widehat{\beta} - Z\widehat{b})^TR^{-1}(Y - X\widehat{\beta} - Z\widehat{b}) - \frac{1}{2}\widehat{b}^TD^{-1}\widehat{b}.$$

(3) Prove the REML likelihood can be obtained by the Bayesian model assuming a flat prior for β, i.e.,

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

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

(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 $(i = 1, \cdots, m)$ and j indicates the jth repeated measure $(j = 1, \cdots, n)$. 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 β, and the ML and the REML estimates of θ and σ^2. Compare your results with those output from mixed models software.

(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.

Problem 3. 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.

4. 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.

(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.

(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).

(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)?