NAME ....

Date: February 24, 2016 *Instructor: Xiang Zhou*

Time: 75 minutes (11:40am – 12:55am)

Question 1 (40 pt)

Clinicians carried out a randomized, double-blind, parallel-group, multi-center study comparing two treatments (denoted as A and B) for hypertension. In the study, 294 patients were measured for their blood pressure at the baseline (week 0) and at weeks 4, 8, 12, and 24 thereafter. The outcome blood pressure is treated as a continuous variable. The main objective of the analysis is to compare the effects of treatments A and B on changes in blood pressure over the duration of the study.

First consider a marginal model for blood pressure. Using the general linear model, fit a model that assumes linear trends over time, with common intercept for the two treatment groups, but different slopes:

$$Y_{ij} = \beta_1 + \beta_2 Week_{ij} + \beta_3 Treatment_i * Week_{ij} + \epsilon_{ij}, \boldsymbol{\epsilon}_i \sim MVN(0, \boldsymbol{\Sigma})$$

Answer the following questions:

1) (**5 pt**) Explain the rationale for including the interaction term $Treatment_i * Week_{ij}$ without the main effect term $Treatment_i$ in the model.

2) (**5 pt**) Interpret the meaning of both parameters β_2 and β_3 in the above model.

3) (**10 pt**) The above GLM based analysis has made a strong assumption about changes in the blood pressure over the duration of the study. What is this modeling assumption? Suggest appropriate statistical approaches to evaluating this assumption.

4) (**10 pt**) Suppose that with an unstructured covariance model, the covariance matrix $\boldsymbol{\Sigma}$ in GLM is estimated as

$$\widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 10.28 & 5.18 & -0.78 & -0.30 & 0.44 \\ 5.18 & 11.17 & 7.16 & 1.39 & 1.49 \\ -0.78 & 7.16 & 17.92 & 11.05 & 0.12 \\ -0.30 & 1.39 & 11.05 & 16.90 & 6.70 \\ 0.44 & 1.49 & 0.12 & 6.70 & 18.80 \end{pmatrix}$$

Because of the large number of parameters in the unstructured covariance matrix $\boldsymbol{\Sigma}$, you decided to use a different covariance structure to model the data. What covariance model/structure would you choose? Explain your reasons for your choice.

5) (**10 pt**) Suppose that you decided to use a compound symmetry covariance matrix for estimation. However, you made a mistake in your REML estimation algorithm: instead of updating the off-diagonal elements ρ, you fixed it to an initial value of 0. Do you still trust your $\boldsymbol{\beta}$ estimates and can you still perform the hypothesis test $H_0: \beta_3 = 0$? Why or why not?

Question 2 (30 pt)

Now consider an alternative analysis for Question 1 based on a linear mixed effects model with random intercepts as follows:

$$Y_{ij} = \tilde{\beta}_1 + b_i + \tilde{\beta}_2 Week_{ij} + \tilde{\beta}_3 Treatment_i * Week_{ij} + \epsilon_{ij}, b_i \sim N(0, \sigma_b^2), \epsilon_{ij} \sim N(0, \sigma^2)$$

with usual modeling assumption. Answer the following questions:

1) (**10 pt**) Interpret the estimates $\hat{\sigma}_b^2 = 1.6$ and $\hat{b}_1 = 0.5$.

2) (**10 pt**) How would you test the hypothesis $H_0: b_1 = \cdots = b_N = 0$ (where N is the sample size)? Explain it to a scientist.

3) (**10 pt**) The linear mixed effects model make a strong assumption about the covariance matrix $V(\boldsymbol{Y}_i)$. What is it? Explain whether this is a good assumption based on Question 1.4.

Question 3 (30 pt)

Consider the following random intercept model:

$$\boldsymbol{Y}_i = \mathbf{1}_n \mu + \mathbf{1}_n b_i + \boldsymbol{\epsilon}_i, b_i \sim N\left(0, \sigma_b^2\right), \epsilon_{ij} \sim N(0, \sigma_e^2)$$

where $i = 1, \cdots, N$; $\boldsymbol{Y}_i$ is an n-vector of repeated measurements; μ is the population-specific intercept; b_i is the subject-specific random intercept; ϵ_{ij} is the measurement error and $\mathbf{1}_n$ denotes an n-vector of 1s. Note that $(ZDZ^T + R)^{-1} = R^{-1} - R^{-1}Z(Z^T R^{-1} Z + D^{-1})^{-1} Z^T R^{-1}$ and $|A + uv^T| = (1 + v^T A^{-1} u)|A|$. Answer the following questions.

1) (**5 pt**) Write down the likelihood and the log-likelihood (you can ignore the constants).

2) (**10 pt**) Given the current estimates $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$, derive the formula to update $\hat{\mu}$. (Use the blank pages if needed.)

3) (**5 pt**) Given the current estimates $\hat{\mu}$, describe how you would use Newton Raphson's algorithm to update $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$. (You will get 5 bonus points if you are able to obtain the first derivatives. Use the blank pages if needed. But finish sub-questions 4 and 5 first before trying it.)

4) (**5 pt**) Given the estimates $\hat{\sigma}_b^2$, $\hat{\sigma}_e^2$, and $\hat{\mu}$, compute the empirical BLUP to predict $\boldsymbol{Y}_i$.

5) (**5 pt**) Compared with the EM algorithm described in the lecture, what is the advantage and dis-advantage of this hybrid algorithm?