Biostats 653 – Midterm

NAME ....

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

Time: 75 minutes (11:40am – 12:55am)

Try not to leave empty space even if you do not know the answer.

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:

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

Answer the following questions:

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

   Because this is a randomized study, so the baseline effects in the two groups are expected to be the same. This approach is typically more powerful.

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

   β_2: the expected change in the blood pressure given a unit change in time (4 weeks); or the slope for time.

   β_3: the expected increase/decrease in difference between the two treatment groups given a unit change in time (4 weeks); or slope difference for time.

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.

   Linear in time assumption. We could use LRT/AIC/BIC to test (1) if there is a time effects (i.e. compare to a reduced model with $\beta_2 = \beta_3 = 0$) and (2) if there is a linear time effects (e.g. compare to a model with quadratic time effects or linear splines). You can also use plot to visualize the pattern.

4) (**10 pt**) Suppose that with an unstructured covariance model, the covariance matrix in GLM is estimated as

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

Because of the large number of parameters in the unstructured covariance matrix $\mathbf{\Sigma}$, 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.

Heterogeneous; and banded Toeplitz with a band size of 2. You can also argue for a heterogeneous matrix with an AR(1)/exponential structure.

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?

The estimates are still consistent and asymptotically normally distributed. We just need to use the robust variance estimate for hypothesis testing (assuming that the sample size is reasonably large).

Question 2 (30 pt)

Now consider an alternative analysis based on a linear mixed model (LMM), with random intercepts, where the patient-specific log odds of moderate or severe onycholysis is modeled as follows:

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

Answer the following questions:

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

   The random intercepts are expected to vary around the population intercept with a variance estimated to be 1.6

   Subject 1 has an intercept that is estimated to be 0.5 unit higher than the population intercept.

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.

   This is equivalent to testing $\sigma_b^2 = 0$. You can use the estimate for σ_b^2 together with the standard errors for testing. Alternatively, you can use a likelihood ratio test, comparing the above full model with a reduced model with only fixed effects. Because of the boundary issue, you need to use a mixture of chisquares for testing.

3) (**10 pt**) The linear mixed model make an strong assumption about the covariance matrix $V(\boldsymbol{Y}_i)$. What is it? Based on question 1, is this a good assumption?

   Compound symmetry. Not a good assumption as detailed in Q1.4.

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 $(ZDZ^T + R)^{-1} = R^{-1} - R^{-1}Z(Z^TR^{-1}Z + D^{-1})^{-1}Z^TR^{-1}$ and $|A + uv^T| = (1 + v^TA^{-1}u)|A|$. Answer the following questions.

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

Denote $\boldsymbol{\Sigma} = \mathbf{1}_n\mathbf{1}_n^T\sigma_b^2 + \boldsymbol{I}\sigma_e^2$. Ignoring the constants, the likelihood is

$$L = \prod_{i=1}^{N} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{Y}_i - \mathbf{1}_n\mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \mathbf{1}_n\mu)}$$

Thus, the log likelihood is

$$l = \sum\nolimits_{i=1}^{N} -\frac{1}{2}log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{Y}_i - \mathbf{1}_n\mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \mathbf{1}_n\mu)$$

2) (**5 pt**) Given the current estimates $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$, derive the formula to update $\hat{\mu}$.

Taking the derivative with respect to μ and set it to zero, we have:

$$\frac{\partial l}{\partial \mu} = \sum\nolimits_{i=1}^{N} -\frac{1}{2}\left(2\ \mathbf{1}_n^T\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\mathbf{1}_n\mu - 2\ \mathbf{1}_n^T\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\boldsymbol{Y}_i\right)$$

$$\hat{\mu} = \left(N\mathbf{1}_n^T\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\mathbf{1}_n\right)^{-1}(\sum\nolimits_{i=1}^{N} \mathbf{1}_n^T\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}\boldsymbol{Y}_i)$$

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.)

You can obtain first order and second order partial derivatives from the log likelihood. With these derivatives, you can use NR algorithm to iterate. The NR algorithm uses the following updates:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{J}_f^{-1}(\boldsymbol{\theta})\boldsymbol{f}(\boldsymbol{\theta})$$

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

$$\widehat{b_\imath} = \hat{\sigma}_b^2 \mathbf{1}_n^T \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}(\boldsymbol{Y}_i - \mathbf{1}_n\hat{\mu})$$
$$\widehat{\boldsymbol{Y}_\imath} = \mathbf{1}_n\hat{\mu} + \hat{\sigma}_b^2 \mathbf{1}_n \mathbf{1}_n^T \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}}(\boldsymbol{Y}_i - \mathbf{1}_n\hat{\mu})$$

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

The EM algorithm is more stable and is guaranteed to increase the likelihood in every step. While the hybrid algorithm has a faster convergence because the use of the second derivatives, but it is sensitive to the starting values and could fail with bad initial values.