STA/BST 224 Longitudinal Data Analysis

Problem Set 4
DUE: Jun. 4, 2020 (Thu)

Instruction:

- Please submit it through Canvas. You can scan or take picture of hand-written solution as long as it is clear.
- You can use either R/SAS/Stata or other software. Please include program and important results.
- The grade will be average of all problems. Due to policy of Graduate Studies, all students need to work on the same problems. You can work together on the homework and ask TA for help. However, each of you is responsible for your own statistical programming and for writing-up your solutions in your own words.

1. In a study of exercise therapies, 37 patients were assigned to one of two weightlifting programs (Freund at al., 1988). In the first program (treatment 1), the number of repetitions was increased as subjects became stronger. In the second program (treatment 2), the number of repetitions was fixed but the amount of weight was increased as subjects became stronger. Measures of strength were taken at baseline (day 0), and on day 2, 4, 6, 8, 10, and 12 (Y_1 to Y_7 in data).

   The raw data can be downloaded from course website (`exercise.raw`). Each row contains following nine variables:

   ID Treatment Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7

   (Note: The categorical variable Treatment is coded 1 = Program 1, 2 = Program 2.)

   (a) Read the data and put the data in a "long" format, with 7 "records" per patient. You may need create the (`timedays`) variable 0, 2, 4, 6, 8, 10, and 12, for the 1, ..., 7 measures, respectively. We will use this `timedays` as time in following model fitting.

   (b) Fit a linear model with randomly varying intercepts and slopes (with respect to time) by REML. The response variable is Y. Your model should include treatment effect (create a dummy variable for it instead of considering it as continuous), linear time trend, and their interaction as fixed effects.

      i. Write down the model and indicate the random effect.
      ii. Present the fitted model, and report the estimated variances of random intercept and random slope.

   (c) Is the model with only randomly varying intercept defensible? Do ReML-based LRT to test it. State null and alternative hypotheses, and draw conclusions.

   (d) Plot the predicted trajectory for the subject with ID=24 using model in (b), as well as its observed trajectory.

2. The dataset "leprosy.csv" consists of count data from a placebo-controlled clinical trial of 30 patients with leprosy at the Eversley Childs Sanitorium in the Philippines. Participants in the study were randomized to either of two antibiotics (denoted treatment drug A and B) or to a placebo (denoted treatment drug C). Prior to receiving treatment, baseline data on the number of leprosy bacilli at six sites of the body where the bacilli tend to congregate were recorded for each patient. After several months of treatment, the number of leprosy bacilli at six sites of the body were recorded a second time. The outcome variable is the total count of the number of leprosy bacilli at the six sites. In this study,

the question of main scientific interest is whether treatment with antibiotics (drugs A and B) reduces the abundance of leprosy bacilli at the six sites of the body when compared to placebo (drug C). The variables include: Drug (A, B, or C=placebo), ID, y(Bacilli Count), visit (=0 if visit is Pre-Treatment, =1 if visit is Post-Treatment).

(a) To check overdispersion, list sample mean and variance of the bacilli count for each treatment group for baseline and post-treatment visits. Is an overdispersion parameter necessary for describing this data set if a Poisson model is considered? (Hint: Poisson model assumes that the variance is the same as the mean if no overdispersion.)

(b) Fit a marginal model (GEE) that uses a log function as a link function for the mean, and includes drug (use placebo group as the reference group, ie, create dummy variables for treatment A and B), visit, and the drug-visit interaction as covariates. Use working exchangeable correlation as the correlation structure.

i. Write down the model (including mean model and "working" variance-covariance-correlation model for V_i).

ii. Fit the GEE model and present estimated β.

iii. Test whether there is a treatment group difference at the baseline (ie, at least two treatments are different at baseline). State null hypothesis and alternative hypothesis, and write down L matrix. Do test and draw conclusion.

iv. Refit model with only the covariates visit and drug-visit interaction. Use this model for questions (iv) and (v). Show the parameter estimates for the covariates, and show both robust standard errors and model-based standard errors. Are the model-based standard errors different from the robust standard errors? If they diverge, which one do you prefer and why? Interpret the parameter estimates based on robust standard errors and draw conclusions about effects of two drugs.

v. What is the estimate of the dispersion parameter? What is the estimated working correlation? What conclusions can you draw from these estimates?

(c) Consider a generalized linear mixed effect model (GLMM) with randomly varying intercepts for the subject-specific log rate of the bacilli count:

$$\log\{E(Y_{ij}|U_i)\} = (\beta_1 + U_i) + \beta_2 \text{Visit}_{ij} + \beta_3 \text{TrtA}_i \text{Visit}_{ij} + \beta_4 \text{TrtB}_i \text{Visit}_{ij},$$

where $Y_{ij}|U_i$ is assumed to have a Poisson distribution. Assume that $U_i \sim N(0, \nu^2)$. Fit the generalized linear mixed effect model. Compare these parameter estimates β_1-β_4 to their corresponding parameters in the marginal model. Do you expect agreement or disagreement between the estimates in marginal and conditional models? Why?