STA/BST 224 Longitudinal Data Analysis

Problem Set 1
DUE: Apr. 21, 2020 (Tue)

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 software. Please include program and important results.
- The grade will be average of all problems except optional one. If you do optional problem, grade will be average of all problems.

1. **(Required for STA/BST PhD students only, optional for other students)** Prove/explain the following mentioned in lecture notes:

   (a) (Page 38-39 of Note 2A) For the **"Working" Model** for separating within-subject from between-subject variance:

   $$Y_{ij} = b_i + \epsilon_{ij}$$

   Explain why

   $$\sigma_t^2 = \text{var}(Y_{ij}) = \sigma_b^2 + \sigma_w^2$$

   (b) (Page 35 of Note 2B) In

   $$\begin{aligned}\text{var}(\hat{\Delta}) &= \text{var}(\bar{Y}_2) + \text{var}(\bar{Y}_1) - 2\text{cov}(\bar{Y}_2, \bar{Y}_1) \\ &= \frac{2}{n}\sigma^2 - 2\text{cov}(\frac{\sum_i Y_{i1}}{n}, \frac{\sum_i Y_{i2}}{n}) \\ &= \frac{2}{n}\sigma^2 - \frac{2}{n}\sigma^2\text{corr}(Y_{i1}, Y_{i2}) \\ &= \frac{2}{n}\sigma^2(1-\rho),\end{aligned}$$

   prove

   $$\text{cov}(\frac{\sum_i Y_{i1}}{n}, \frac{\sum_i Y_{i2}}{n}) = \frac{1}{n}\sigma^2\text{corr}(Y_{i1}, Y_{i2})$$

2. **(Required for all students) Background**: A study was conducted to investigate two treatments for patients suffering from multiple sclerosis (MS). 150 suffers of the disease were recruited into the study, and 75 were randomized to receive azathioprine (AZ) alone (group 1), and 75 were randomized to receive azathioprine plus methylprednisommne (AZ+MP, group 2). For each participant, a measure of auto-immunity, azathioprine AFCR, was planned at clinic visits at baseline (time 0, at initiation of treatment) and at 3, 6, 9, 12, 15, and 18 months thereafter. Multiple sclerosis affects the immune system; low values of AFCR (approaching 0) are evidence that immunity is improving, which is hopefully associated with a better prognosis for suffers of MS. Also recorded for each subject was age at entry into the study and an indicator of whether or not the subject had had previous treatment with either of the study agents (0=no, 1=yes). The average age of the men across both treatment groups was 50.45, with SD 6.69.

   The primary scientific aims of the study are to investigate whether (i) both treatments (AZ or AZ+MP) lower AFCR over the 18 month period. and (ii) whether treatment with AZ+MP results in different immune system response than does AZ alone, and, if so how it is different in terms of response over time. It was also suspected that a subject's age and prior history might be related to their AFCR level at baseline and to the rate at which AFCR changes during the 18 month period.

Response variable of interest: The square root of AFCR (square roots were taken so that the AFCR observations better satisfy the assumption of normality).

Data file: The raw data are in the file `afcr.raw`. In the file, each record corresponds to a single observation, with columns:

```
column 1 = subject id
column 2 = time (months)
column 3 = AFCR (already taken square root)
column 4 = group ( 1 = AZ alone, 2 = AZ + MP)
column 5 = prior treatment indicator
column 6 = age (yrs)
```

(a) Download the data from Canvas and read them into software. Make sure that you have them in “long” format. Describe the structure of the data set, including distribution of times (eg, range? discrete or continuous? number of unique time points?), number of subjects and number of observations (and numbers of observations in each treatment group), whether or not the covariates and response are baseline or time-varying.

(b) Ignore age and prior treatment effects for now. Using tools of your choice, explore the data with respect to the primary scientific aim of the study, and present one or two plots (or pairs of plots, by treatment group, if you wish) that illustrate the behavior of the response variable (square root of AFCR) over time and with respect to treatment group. Please do not present too much information. Write a few sentences summarizing your results to a non-statistical audience.

(c) Explore the variance of the response variable as a function of time. Then, explore the correlation structure of the response variable using correlation matrices and the sample autocorrelation function. Make sure to remove covariate effects by removing effects of age, prior treatment, time, treatment group, and interaction between time and treatment group. Describe your results.

(Hint: Use a dummy variable at each time point (ie, time is a categorical variable), since there is a small number of time points. This will be more flexible than linear time effect. You can assume a linear age effect.)