HOMEWORK 3
BIOSTATISTICS 755
DUE FEBRUARY 19TH, 2024

1. The dataset `MITgrowth.csv` on the course website data are from a prospective study on body fat accretion in a cohort of 162 girls from the MIT Growth and Development Study. The study was designed to look at changes in percent body fat in girls before and after menarche. All subjects had to be pre-menarche and non-obese to enter the study. Observations were taken annually until 4 years after menarche. At each observation percent body fat was measured.

   Two time-scales are included: age, and time since menarche (which can be negative). Time since menarche is the more biologically relevant time scale to use. The variables (in order) are: *Subject ID, Current Age (years), Age at Menarche (years), time relative to Menarche (years), Percent Body Fat*.

   (a) **(15 points)** Produce a spaghetti plot of the data using Age and then time relative to menarche. Which time scale appears to have a stronger relationship with percent body fat? Which time scale is best to answer the study question? Comment on the possible parametric methods that could be used to include time in the model.

   (b) **(15 points)** Fit a model with the time effect you found to be most appropriate in (a), with randomly varying intercepts and slopes.

      i. What is the estimated variance of the random intercepts?
      ii. What is the estimated variance of the random slope(s)?
      iii. What is the estimated correlation between the random intercepts and slopes?

   (c) **(10 points)** Fit a model with only randomly varying intercepts. What do you think about this model versus the previous model? Should this be used? Why?

   (d) **(20 points)** Give a full description of your findings. Include interpretations of at least two regression coefficients in your description. All descriptions should be in context of the study and the goals of the study.

2. **(40 points)** A study was conducted to investigate two treatments for patients suffering from multiple sclerosis. 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 (MS) 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 subjects 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. The square root of AFCR is the response variable of interest (square roots were taken so that the AFCR observations better satisfy the assumption of normality).

The data are in the file `afcr`, which you can download from the course website. In the file, each record corresponds to a single observation, with columns:

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

(a) Complete an analysis of this data using a linear mixed model that results in the best possible answer to the the scientific aims of the study. This should include exploratory data analysis (e.g., figures), model fit, discussion of the random effect(s) variance (or covariance), and a paragraph that accurately summarized your conclusions with interpretations of coefficients of interest. It is encouraged to give the code to replicate your findings.