Homework 2

Biost 540

General Instructions:

- When working in groups, elect one member as the leading member who will be responsible for uploading the homework assignment solutions. Please note that all members in the group are expected to equally contribute to the assignment!
- Be sure to show work for all problems! `R` code should not appear in the main body of the homework; however, the code should appear at the end of the assignment as an Appendix. It should be possible for someone to use the code to reproduce any figures or numeric results.

Part A: Cholesterol Data

[Edited FLW problem 5.1] In the National Cooperative Gallstone Study (NCGS), one of the major interests was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones. In this study, patients were randomly assigned to high-dose (750 mg per day), low-dose (375 mg per day), or placebo. We focus on a subset of the data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups. In the NCGS it was suggested that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result, serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20 and 24 months of follow-up. Many cholesterol measurements are missing because of missed visits, laboratory specimens that were lost or inadequate, or patient follow-up that was terminated. The NCGS serum cholesterol data are stored in cholesterol.csv posted on canvas. Each row of the dataset contains the following variables:

- group: a categorical variable coded 1=high dose, 2=placebo
- index: subject index
- Y1, Y2, Y3, Y4, Y5: measurements of cholesterol at baseline and times 6, 12, 20 and 24 months

Note that the data is in wide format.

1. Conduct an exploratory data analysis by using graphical methods and numerical summary statistics to describe (1) the relationship between serum cholesterol levels over time for the two groups and (2) the correlation between measurements. Write a paragraph commenting on your observations from the exploratory data analysis.
2. With baseline (month 0) and the placebo group (group 2) as the *reference group*, write out the regression model for mean serum cholesterol that corresponds to the analysis of response profiles. Provide interpretations for each coefficient.
3. Using GLS, conduct an *analysis of response profiles* (using all available measurements) with an unstructured covariance matrix and allowing for heteroscedasticity. Determine whether the patterns of change over time differ between the two groups. What do you conclude at the 5% significance level?
4. Now consider a parametric formulation of the mean profiles with only a linear trend in time. Fit the model. What is the estimated rate of increase in mean cholesterol levels for the treatment group? What is the estimated rate of increase in mean serum cholesterol levels for the placebo group? At the 5% significance level, are the patterns of change the same in the two groups?
5. Compare the models and conclusions from questions 3 and 4. Discuss the advantages and disadvantages of treating time categorically (i.e. analysis of response profiles) and treating time linearly.

Part B: Dental Growth Data

We are interested in answering the question: is there a difference in dental growth rates between males and females?

```r
library(nlme)
data(Orthodont)
```

1. Write out the regression model with dental length as the outcome to address this question, **assuming a linear model for time**. What are the interpretations for each of the parameters?
2. Create a table with estimates and standard errors from fitting each of the following models for all the parameters in the mean model:
   - OLS, model-based standard errors (homoscedasticity)
   - GLS, unstructured/symmetric correlation matrix, heteroscedasticity, REML
   - GLS, exchangeable/compound symmetric correlation matrix, homoscedasticity, REML
   - LMM, random intercepts, REML
   - LMM, random intercepts + slopes (correlated), REML
3. Comment on each of the models and the results. Things to consider:
   - What do the different models assume about the correlation structure of dental measurements? How do the estimated covariance matrices compare between the models?
   - How do the point estimates and standard errors compare between each of the models?
   - Comment on the validity of the point estimates and standard errors that OLS provides