Biost 540: Homework 1

General Instructions

- Students may discuss with each other but each of you will be required to submit the work in your own writing.
- Grading will be based on completion (3 pts), accuracy (3 pts), work shown (3 pts), and neatness (1 pt).
- 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.

Problem 1

The Dataset

This dataset is from a clinical trial for augmentation treatment for depression (Sanacora et al., 2004). There were 50 patients total: half were randomly assigned to the augmentation group (fluoxetine + yohimbine) and half were randomly assigned to the control group (fluoxetine + placebo). Participants in the study were followed for an additional 6 weeks and their Hamilton depression scale ratings (HDRS) and Clinical Global Impressions Scale for severity (CGI). We will focus on the participants' HDRS.

```
if (!requireNamespace("reshape2", quietly = TRUE)) {
  install.packages("reshape2")
}
library(reshape2)
aug <- read.csv("~/augmentation.csv")
aug <- aug[, c("id", "Treatment_Group", "HD_t0", "HD_t1", "HD_t2",
               "HD_t3", "HD_t4", "HD_t5", "HD_t6")]
aug_long <- melt(aug, id=c("id","Treatment_Group"))
aug_long$week <- as.numeric(gsub("HD_t","",aug_long$variable))
```

Questions

a) Summarize the distribution of HDRS at baseline and weeks 1 – 6 in the two groups. Plot the mean HDRS over time for the two groups. What do you notice?

b) Plot individual series of longitudinal observations for all subjects. Comment on what you observe, specifically in regards to differences between the two groups.

c) Characterize the correlation among the HDRS overall and separately for each group. Comment on what you notice. Is the variability in responses different at different time points?

Problem 2

The Treatment of Lead-Exposed Children (TLC) trial was a placebo-controlled, randomized study of succimer (a chelating agent) in children with blood lead levels of 20-44 micrograms/dL. These data consist of four repeated measurements of blood lead levels obtained at baseline (or week 0), week 1, week 4, and week 6 on 100 children who were randomly assigned to chelation treatment with succimer or placebo.

Each row of the data set contains the following 6 variables: ID, Treatment Group, Lead Level Week 0, Lead Level Week 1, Lead Level Week 4, Lead Level Week 6.

```
tlc <- read.csv("~/tlc.csv")
```

Questions

a) Transform the data into long format and produce spaghetti plots illustrating the progression in lead level. Facet the plot by treatment group. *Hint: the functions* `pivot_longer` *(with argument* `names_prefix`*) and* `facet_wrap` *will be useful to you.* **Be sure to label axes appropriately with units.**.

b) Compute the pearson correlation matrix between outcomes at different time points.

c) Suppose you were interested in comparing the mean difference in lead levels between treatment and control groups 6 weeks after treatment provision. Using a linear model, compare the lead levels between treatment groups at the 6 week time point. Provide and describe a point estimate for the treatment effect along with the standard error. Does treatment have a significant effect on the Week 6 lead levels at level $\alpha = 0.05$?

Hint: you may want to use the original (not long form) of the dataset for the next two parts of the problem.

d) Suppose we adopt an approach to analyze the change scores from week 0 to week 6. Using a linear model, compare the *change* in lead levels from week 0 to week 6 between treatment groups. Provide and describe a point estimate for the treatment effect along with the standard error. Does your conclusion from part c) change?

e) Suppose we wish to use an ANCOVA model to evaluate the effect of treatment on Week 6 lead levels *adjusting* for the baseline lead level. Reinterpret the treatment effect and provide an estimate of the standard error. Does your conclusion differ from part c)? *Hint: ANCOVA is a linear model which adjusts for the baseline outcome as covariate.*

f) Plot the standard errors of each of the models fit in parts c)-e). How does incorporating knowledge of the baseline response impact the precision or power of your inference on the treatment effect at 6 weeks?