In-class Data Analysis Exercise

We will use this class to explore some of the later sections of Chapter 19 on two-factor studies. The exercise will count toward a portion of the HW 2 assignment. We will again use the water quality data set available at our website as an Excel (.xls) file. If you would like to use the course folder, the `PROC IMPORT` step looks like this:

```
PROC IMPORT OUT=WORK.WQ
            DATAFILE="/courses/ddf5e9e5ba27fe300/STAT705/EColi.xlsx"
            DBMS=XLSX
            REPLACE;
RUN;
```

I would like to set up Watershed and Month as factors for a two-factor study–we will use Congaree, Savannah and Great Pee Dee for our watersheds, and all 12 months:

```
data wqcsp; set wq;
where substr(station,1,2) in ('C-','SV','PD');
if substr(station,1,2)='PD' then watershed='Great Pee Dee';
else if substr(station,1,2)='C-' then watershed='Congaree';
else watershed='Savannah';
Month=month(collection_date);
run;
```

The possible response variables (FecalColi, EColi, Enterococci) are strongly positively-skewed. Use the Box-Cox method in PROC TRANSREG (see Chapter 18 code) to select a transformation, then transform Enterococci.

Fit an interaction model in `PROC GLM` for the transformed response variable Enterococci. Does the F test indicate an interaction is present? How strong does the evidence of interaction seem to be given the large sample size for this problem? Discuss main effects and the interaction based on an inspection of the interaction plot.

Use `PROC FREQ` to see how many observations occur in each cell of a two-way table of Watershed by Month. How unbalanced is the data set? Explain how this lack of balance is reflected in the properties of the Type III SS for Month, Watershed and Month*Watershed–do they add up to SSTR?

HW 2: Generate labels for Month using `PROC FORMAT`*. Re-run* `PROC FREQ` *with attractive labels for Month.*

Since interaction appears to be present, use the `SLICEBY` command in `PROC GLIMMIX` with Tukey adjustments to test pairwise comparisons for Watershed within Month; this should generate 12 different graphs.

HW 2: Repeat the above to study Month within Watershed. There will be only 3 graphs, each with 66 pairwise comparisons, though the pairwise comparisons tend to organize themselves by groups of months. The graphs are somewhat repetitive–discuss results for the Savannah watershed only.

it HW 2: At what confidence level can you make claims about the two sets of multiple comparisons (Month within Watershed and Watershed within Month) simultaneously? Explain.

Generate residual and diagnostic plots in `PROC GLM` if you have not already done so. Can you explain any unusual patterns you see in the plot of studentized residuals vs. $\hat{\mathbf{Y}}$? Do the residuals appear to be normally distributed?

HW 2: Save the residuals and test for normality in `PROC UNIVARIATE`.

HW 2: Order the residuals by Collection_Date and within Watershed and Station and discuss any trends over Collection_Date for the Congaree watershed only, using Station as a grouping variable.

HW 2: List 6 pairwise contrasts to test for seasonal differences (Dec-Feb=Winter, March-May=Spring, Jun-Aug=Summer, Sept-Nov=Winter) within the Congaree watershed. Test these contrasts in PROC GLIMMIX while controlling the overall error rate with a Bonferroni correction.