In-class Data Analysis Exercise

We will use this class as a model selection and diagnostics exercise using the water quality data set available at our website as an Excel (.xls) file. If you want to skip the Import Wizard, `PROC IMPORT` session looks like this:

```
PROC IMPORT OUT=WORK.WQ
            DATAFILE="C:\Grego\My Documents\STAT704\Ecoli.xls"
            DBMS=EXCEL REPLACE;
            RANGE="EColi$";
            GETNAMES=YES;
            MIXED=NO;
            SCANTEXT=YES;
            USEDATE=YES;
            SCANTIME=YES;
RUN;
```

As you will see when you work through the exercise, there are numerous additional steps we could take or choices we could make in the analysis of this data set. I settled on a couple topics I knew I wanted to cover, as well as some options that seemed interesting as I was in the process of constructing the exercise.

These next steps can all appear in the same `DATA` statement, though you may need to build them sequentially to make sure each is carried out correctly.

- Select only gages from the Congaree (`STATION` starts with "C-"), Savannah ("SV-", and Pee Dee ("PD-") watersheds using a WHERE statement and the SUBSTR() function.
- Create a categorical variable `Watershed` with three different levels (Congaree, Great Pee Dee, and Savannah).
- Create a numerical `Month` variable from `Collection_Date` using the MONTH() function in SAS.
- Save the log of the bacterial count variables EColi, FecalColi and Enterococci as three new variables.

Create a scatterplot matrix using the logs of the water quality variables. Comment on any patterns. Regress log(FecalColi) on log(EColi) and log(Enterococci). Are both variables significant? Note the two extreme values for Cook's D, the handful of studentized deleted residuals with absolute values greater than 5 and the leverage values greater than 0.010 (what would our rule-of-thumb for high leverage actually be in this case?). Identify the observations that generated those diagnostic statistics and comment on them. Look at DFFITS and see whether this diagnostic picked up any outliers that the other methods did not detect.

Create a scatterplot with log(Enterococci) on the x-axis and log(FecalColi) on the y-axis and overlay separate regression lines using Watershed as a group variable. What do you observe? Develop a hypothesis based on your observation and then test it in `PROC GLM` using as a full model an interaction model with log(Enterococci) as a continuous predictor and Watershed as a categorical predictor (I used a CONTRAST statement to conduct my test, but you can also construct a reduced model and build an F-test that way). Note that this is a form of data-snooping and so results from any testing should be interpreted judiciously.

Use PROC GLMSELECT to identify a best stepwise regression of log(FC) on log(Enterococci), log(EColi), Month, Watershed and all possible pairwise interactions ($\alpha_{\text{entry}} = 0.10$, $\alpha_{\text{remove}} = 0.15$). This is a large data set, so the model tends to pick up more terms than perhaps it should. Nonetheless, look at the coefficients in the final model for the two class variables, Month and Watershed, and comment on any patterns you see there in the effects.