In-class Data Analysis Exercise

We will use this class to familiarize ourselves with data management in the Gambrell 003 lab in preparation for the first exam, and explore a data set. The computers in Gambrell do not provide access to our MathStat drives, so assume that you will be working from a subfolder in an available directory. I have placed the data set(s) in Blackboard, so that you can download them to the storage site of your choice (your document drive, a flash drive, etc).

Previously, we had worked with both the 2008 cohort data and follow-up data from the succeeding year. Let's include the following year as well. Refer to Class Exercise 2 to remind yourself how we match-merged data from Fall 2008 and Fall 2009. Modify the DATA step match merge to include the Fall 2010 data.

Now create a scatterplot matrix of GPA2008, GPA2009, and GPA2010. What are your initial impressions about the needs for transformation? Does it appear as though collinearity could be an issue?

Let's jump right to `PROC TRANSREG` to see whether a power transformation of the response variable GPA2010 is appropriate. Before discussing transformations, look at the LOG window–how many cases had at least one missing value in the predictors? Were additional records excluded from the analysis? If so, why might they have been excluded?

What transformation does the Box-Cox algorithm recommend? Is it reasonable to use the identity transformation? Plot GPA2010 against each of GPA2008 and GPA2009 with a regression line and loess curve to help in your decision. What effect do these plots appear to capture?

Regress GPA2010 on GPA2008 and GPA2009, and save the residuals in an output file. Comment on the least squares estimates of the slope parameters and note any interesting residuals.

Add an interaction term (`GPA2008*GPA2009`) to your output file and plot it against the residuals. What does the loess curve suggest to you?