Math 222 - Project 8 **Due Friday, April 6**

1. A PE teacher at one middle school wanted to know if there was a correlation between the number of push-ups seventh-graders could do and their mile-run times. She collected the data in the following file: PEClass.txt.

   (a) Make a scatterplot showing the relationship between the number of push-ups and mile-run times. Use push-ups as the explanatory variable.

   (b) Make three more scatterplots by plotting:

   - log(PushUp) vs. MileTime.
   - PushUp vs. log(MileTime).
   - log(PushUp) vs. log(MileTime).

   Which graph has the strongest correlation? *Since you can't take a logarithm of zero, you need to remove the two students who could not do any push-ups from the data. You can use the subset command to do this.*

   (c) For the graph with the strongest linear correlation, find the best fit regression line. What is the formula for this model? Can you use algebra to get rid of the logarithms in the formula?. What kind of equation do you get?

   (d) Now consider a multiple regression model for predicting mile-run times based on both the gender and the number of push-ups that a student can do. Compute the R-squared and the adjusted R-squared for such a model. Does it have more explanatory power than the model that ignored gender?

2. The data set happiness.csv compares five variables for countries around the world: LSI stands for life satisfaction index and measures happiness, GINI measures inequality, CORRUPT measures the level of corruption in government, LIFE is average life expectancy, and DEMOCRACY is a measure of civil and political liberties.

   (a) Graphically show the distributions of each of these five variables, then graphically show the relationship between each pair of variables. Briefly describe your findings, focusing on whether the conditions for linear regression are likely to be satisfied.

   (b) We will now build a regression model to predict the life satisfaction index based on the other variables. Start with a full model, and use backwards elimination to obtain the model with the best adjusted R^2. Write a brief description of the steps as you perform the backwards elimination, and explain which variables you remove and why. At the end, clearly describe which subset of variables is best for predicting life satisfaction levels, and describe what percent of the variability in the response variable is explained by the model.

   (c) Using your regression model, make a 95% prediction interval for the LSI of a country where GINI = 30, CORRUPT = 2, LIFE = 80, and DEMOCRACY = 5.