Math 222 - Project 5 **Due Monday, April 15**

1. Many people believe that gender, weight, drinking habits, and many other factors are much more important in predicting blood alcohol content (BAC) than simply considering the number of drinks a person consumed. Here we examine data from sixteen student volunteers at Ohio State University who each drank a randomly assigned number of cans of beer. These students were evenly divided between men and women, and they differed in weight and drinking habits. Thirty minutes later, a police officer measured their blood alcohol content (BAC) in grams of alcohol per deciliter of blood. The data is in the file bac.csv

   (a) Make a scatterplot for this data. Use it to describe the relationship between the number of cans of beer and BAC.

   (b) Write the equation of the regression line. Interpret the slope and intercept in context.

   (c) How much of the variability in BAC can be explained simply by the number of beers a student has drunk?

   (d) Make a 95% confidence interval for the slope of the regression line. Explain clearly what this confidence interval tells us about the effect of each extra beer a person drinks.

   (e) Suppose we visit a bar, ask people how many drinks they have had, and also take their BAC. Do you think the relationship between number of drinks and BAC would be as strong as the relationship found in the Ohio State study? Why or why not?

2. 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? Use algebra to rewrite the formula without logarithms. What kind of equation do you get?

3. 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 (higher numbers mean less corruption), LIFE is average life expectancy, and DEMOCRACY is a measure of civil and political liberties.

   (a) We want to see how the other four variables affect life satisfaction (LSI). Make graphs that show the relationship between each of the other four variables and LSI. Does each explanatory variable have a roughly linear relationship with LSI?

   (b) Use backwards elimination to obtain the multiple linear regression model with the best adjusted R^2 for predicting LSI. 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) Check the residuals of your model. Are they approximately normally distributed?

(d) 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.