Math 222 - Project 5 **Due Friday, April 24**

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:

   `http://people.hsc.edu/faculty-staff/blins/StatsExamples/PEclass.csv`

   (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 (exponential, logarithmic, or power function)?

2. The data set

   `http://people.hsc.edu/faculty-staff/blins/classes/spring17/math222/data/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.

3. The file

   `http://people.hsc.edu/faculty-staff/blins/StatsExamples/possum2.csv`

   contains data on 104 brushtail possums from two regions in Australia, where the possums may be considered a random sample from the population. The first region is Victoria, which is in the eastern half of Australia and traverses the southern coast. The second region consists of New South Wales and Queensland, which make up eastern and northeastern Australia. We use logistic regression to differentiate between possums in these two regions. The outcome variable,

population (`pop`), takes value 1 when a possum is from Victoria and 0 when it is from New South Wales or Queensland. We consider five predictors: head length (`headL`) in millimeters, `sex`, skull width (`skullW`) in millimeters, total length (`totalL`) in centimeters, and tail length (`tailL`) in centimeters.

(a) Make a logistic regression model to predict the value of the population variable from the other four variables. Which variables in the model appear to have statistically significant coefficients in the model?

(b) Use backward elimination to find the model with the lowest AIC. Which variables, if any, do you end up removing?

(c) Suppose we see a brushtail possum at a zoo in the US, and a sign says the possum had been captured in the wild in Australia, but it doesn't say which part of Australia. However, the sign does indicate that the possum is male, its skull is about 63 mm wide, its tail is 37 cm long, and its total length is 83 cm. What is the reduced model's computed probability that this possum is from Victoria?