Project 3 **Math 222**
Regression **Due Friday, March 4**

1. 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.
   (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.

2. The file possums.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: `sex`, head length (`headL`) in millimeters, skull width (`skullW`) in millimeters, total length (`totalL`) in centimeters, and tail length (`tailL`) in centimeters.
   (a) Make histograms to display each variable. Are there any outliers that are likely to have a very large influence on the logistic regression model?
   (b) Make a logistic regression model to predict the value of the population variable from the other five variables. Which variables in the model appear to have statistically significant coefficients in the model?
   (c) Use backward elimination to remove the variables with the largest p-values corresponding to their coefficients. Repeat until all of the remaining variables in the model have p-values of less than 5%.
   (d) Write down the equation for the logistic regression model using the estimates for the coefficients given by R.
   (e) 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 doesnt 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? How confident are you in the models accuracy of this probability calculation?