Math 222 - Project 10 **Due Monday, May 1**

1. The file possum.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 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? How confident are you in the model's accuracy of this probability calculation?

2. On the 1986 General Social Survey, one of the questions asked was: *About how much time (does/did) it usually take you to travel to work - about how many minutes?* The responses are contained in the file commuting.csv under the `travel_time` variable. The file also contains two other variables: `gender` and `highest_degree` earned.

   The information is based on 1322 respondents. The data from 140 other respondents had to be removed because they didn't answer one of the questions. Another 8 respondents were removed because they said it took over 90 minutes to travel to work.

   (a) Make a two-way table and a mosaic plot showing the relationship between the two explanatory variables: `gender` and `highest_degree`. Describe any association you see. Make sure to use the `factor()` function to sort the education levels in order.

   (b) Now make interaction plots to see how the two explanatory variables affect commuting time. Give a brief description of the main effects and any interactions that are visible.

   (c) Are the main effects and interactions in part (b) statistically significant? Clearly explain what your results mean about the relationship between gender and education level was on commuting times back in 1986.