**Math 222 - Project 6** **Due Friday, April 25**

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) Are there any outliers in the data 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?

2. Dansinger, Griffith, Gleason et al. (2005) report on a randomized, comparative experiment in which 160 subjects were randomly assigned to one of four popular diet plans: Atkins, Ornish, Weight Watchers, and Zone (40 subjects per diet). These subjects were recruited through newspaper and television advertisements in the greater Boston area; all were overweight or obese with body mass index values between 27 and 42. Among the variables measured were
   - Which diet the subject was assigned to
   - Whether or not the subject completed the twelve-month study
   - The subject's weight loss after two months, six months, and twelve months (in kilograms, with a negative value indicating weight gain)
   - The degree to which the subject adhered to the assigned diet, taken as the average of 12 monthly ratings, each on a 1-10 scale (with 1 indicating complete non-adherence and 10 indicating full adherence)

   Data for the 93 subjects who completed the 12-month study are in the file ComparingDiets.csv. Some of the questions that the researchers studied are:
   (a) Do the average weight losses after 12 months differ significantly across the four diet plans?
   (b) Is there a significant difference in the completion/dropout rates across the four diet plans?
   (c) Is there a significant positive association between a subject's adherence level and his/her amount of weight loss?
   (d) Is there strong evidence that dieters actually tend to lose weight on one of these popular diet plans?

   For each of these research questions, first identify the explanatory variable and the response variable, and classify each as categorical or quantitative. Then use graphical and numerical summaries to investigate the question, and summarize your findings. Next, identify the inference technique that can be used to address the question, and apply that technique. Be sure to include all aspects of the procedure, including a check of its technical conditions. Finally, summarize your conclusions for each question. Write a paragraph summarizing your findings from these four analyses. [Hint: To determine the completion rate for each diet, count how many of the 93 subjects who completed the study are in each diet group and compare those counts to the 40 that were originally assigned to each diet.]