**Ch. 4 & 5 homework** – due Tuesday, November 5, 2013

As a reminder, you need to provide a word document with your answers, along with appropriate output and graphs from the R program that helps to answer each question, as well as your code. Please let me know if you have any questions.

**1.) Use data file OmniPower.** Two independent variables are considered here: the price of an OmniPower bar, as measured in cents (Price); and the monthly budget for in-store promotional expenditures, measured in dollars (Promotion). The dependent variable is the number of OmniPower bars sold in a month (Sales). Data are collected from a sample of 34 stores in a supermarket chain selected for a test-market study.

1. Run a one-factor ANOVA model, using Price as the factor.

a(i) Decide whether or not Price is an important factor in determining Sales. State your null hypothesis, alternative hypothesis and all components of the decision-making rule. Use a 5% level of significance.

a(ii) Interpret b1.

1. Run a one-factor ANOVA model, using Promotion as the factor.

b(i) Decide whether or not Promotion is an important factor in determining Sales. State your null hypothesis, alternative hypothesis and all components of the decision-making rule. Use a 5% level of significance.

b(ii) What is the point estimate of Sales when Promotion is $400?

b(iii) What is the 95% confidence interval for the point estimate when Promotion is $400?

b(iv) Interpret this interval from part b(iii).

1. If you could only pick one variable to predict Sales, which one would you choose and why?
2. Run a model using Price and Promotion to predict Sales. Interpret the y-intercept.

**2.) Use data file GCFreeRosalyn.** This data file information on homes, which includes the (Address), appraised value in $thousands (Value), 3 possible town locations (Location), property size in acres (Property), house size in square feet (House), the age of the house in years (Age), the number of rooms (Rooms), the number of bathrooms (Baths), and the number of cars that can be parked in the garage (Garage). We will not use Address as a variable, it is just an identifying column to us. Use α = 0.05 throughout this problem.

a. Run a model predicting the house Value with all the other variables.

a(i) Run a stepwise regression model. Which variable(s) are dropped?

a(ii) Run a backward regression model. Which variable(s) are dropped and what are their p-values? List these in order. Call this modelA from now on.

b(i) Make a scatterplot of property size and appraised value.

b(ii) Let’s say that you believe that it seems as when the property size gets very large, some of the appraised values are decreasing because people feel it is a detriment to maintain the property’s upkeep and thus devalue the home. This is causing what you feel is a slight curve in the scatterplot. Create a quadratic term for property size and add it to modelA, making modelB. Is this quadratic term significant? How do you know?

c. From model A, find a histogram of the standardized residuals and the leverages. Comment on these plots.

d. Identify which of the observations under model A are outliers. It should be clear to me how you have determined these.

e. Identify which of the observations under model A are high leverage points. It should be clear to me how you have determined these.

f. Rerun model A, removing any points that are outliers or high leverage points. Note the changes that this has caused in the model.

**3.) Use data file BB2009.** The following dataset is on the 30 baseball teams information in 2009 that includes information on the team name (Team); the league they belong to, where 0=American league and 1=National league (League); number of wins (Wins); the pitchers’ earned run pitching average (ERA); number of runs scored (Runs); the number of hits allowed to the other team (Hits); the number of walks allowed to the other team (Walks); the number of saves (Saves); and the number of errors committed (Errors). We will not use Team as a variable; it is just an identifying column to us. Use α = 0.05 throughout this problem.

1. Run a model predicting Wins with all the other variables, we will refer to this as model1. Then add an interaction term between Hits and Saves. Is this interaction term significant? State your null hypothesis, alternative hypothesis and use the t score in your decision-making rule.
2. Return back to model1. Determine whether the variables ERA, Walks, and Saves; which are all variables solely related to the pitcher, are jointly significant.
3. Return back to model1. Calculate the variance inflation factors and a table of correlations among only the explanatory variables. Based on these statistics, is collinearity an issue with this fitted model? Why or why not?
4. Run a stepwise and backwards regression on model1. Do these come up with the same model?
5. Continue on from the previous question and do anything you can think of to find a better model than the one in part d.