**Homework 4 – Ch. 5 & 6** – due Monday, November 17, 2014

As a reminder, you need to attach to an email your R code and output/answers for these problems. Please also bring a print out of the output/answers ready to be graded to class with your names on it.

**1.) Start with a model that predicts your typical Y variable with all other variables in your dataset as explanatory variables. We will call this model1.**

a(i). Do stepwise regression on model1. Write out the model that stepwise regression chooses.

a(ii). Do backwards regression on model1. Write out the model that backwards regression chooses.

b(i). Start back at model1 and identify any possible outliers.

b(ii). Start back at model1 and identify any possible high leverage points.

b(iii). Start back at model1 and identify any possible influential points with Cook’s Distance.

c(i). Remove all points identified in parts (b) above, call this model2.

c(ii). Do stepwise regression of model2.

c(iii). Do backwards regression of model2. Call this model3.

c(iv) Thoroughly compare these 3 models of parts (c) and their goodness of fits.

d(i). Build a correlation matrix of all of the explanatory variables used in model1. Do you suspect multicollinearity between any of these variables? Explain briefly.

d(ii). Find all of the variance inflation factors. Comment on these appropriately.

e(i). Is model2 significant? How do you know? State your null hypothesis, alternative hypothesis and all components of the decision-making rule. Use a 5% level of significance.

e(ii). Is the second variable in your model2 significant? How do you know? State your null hypothesis, alternative hypothesis and all components of the decision-making rule. Use a 5% level of significance.

f(i). Provide a plot of residual versus fitted values for model1. What do we hope to learn from this type of plot? Does this plot display any model inadequacies?

f(ii). Provide a plot of residual versus fitted values for model2. Does this plot display any model inadequacies? Any thoughts about the differences between this plot and the plot from model1?

g(i). Provide a plot of residual versus X1 for model1. What do we hope to learn from this type of plot? Comment on this plot appropriately.

g(ii). Provide a plot of residual versus X2 for model1. Comment on this plot appropriately.

h. Find the AIC values for model1, model2, and model23. Which model does AIC pick as the best? Did you expect this? Explain briefly.

i(i). Do you feel that any of your X variables that are in model3 suffer from possible bias from a limited sampling region? Demonstrate and comment on why or why not they may suffer from this.

i(ii). Comment on whether you feel your model3 suffers from bias due to omitted variables. Explain why or why not you think this.

**2.)** I have posted a dataset titled HW4data on Blackboard. The file contains data I have collected from students in Stat 72. There are columns for identifying a specific student by number, their homework 3 score, Test 1 score, Test 2 score, Test 3 score, Gender, Year in School, their current cumulative GPA, the number of credit hours they are currently enrolled in, whether or not they can drive a manual car (or stick shift), what row of the class they sit in counting from the front of the room, whether they prefer to drink Coke or Pepsi, the number of siblings they have, the number of countries outside of the United States they have travelled to, the number of jobs they have ever had, and whether they prefer dogs or cats.

I want you to build the best model you can, predicting Test 3. Use anything you have learned in the class to date and explain at each step what you are doing and why you are doing it.