Statistical Data Mining I
Homework 2

1) (10 points) Consider the cereal dataset in UBlearns. Suppose that you are getting this data in order to build a predictive model for nutritional rating. a) Divide the data into test and training. Fit a linear model and report the MSE.
b) With the data in (a) perform **either** forward or backwards subset selection.
c) With the data in (a) perform exhaustive subset selection.
d) Draw some conclusions through comparisons between models (a-c). Reflect on the comparative predictive accuracy, and model interpretation. Which model would you say is the "best one" based on your results?

2) (10 points) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the *zipcode* data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k. The *zipcode* data is available in the ElemStatLearn package – or the website for the text. Note that you do not have to divide the data into test and training because it is done for you.

3) (10 points) In this exercise, we will predict the number of applications received using the other variables in the College data set in the ISLR package.
*** be sure to look closely at this data, you may want to consider the multi-scale nature of the problem, and perhaps use a transformation on some of the variables.***
(a) Split the data set into a training set and a test set. Fit a linear model using least squares on the training set, and report the test error obtained.
(b) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.
(d) Fit a lasso model on the training set, with λ chosen by crossvalidation. Report the test error obtained, along with the number of non-zero coefficient estimates.
(e) Fit a PCR model on the training set. Report the test error obtained, along with justification for the choice of "k".
(f) Fit a PLS model on the training set. Report the test error obtained, along with justification for the choice of "k".
(g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches?