Homework 3
Patricia Hoffman, PhD.

1) Once again check out wine quality data set described in the web page below:

http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names

Remember the Red Wine data set (winequality-red.csv) contains 1599 observations of 11 attributes. The median score of the wine tasters is given in the last column. Note also that the delimiter used in this file is a semi colon and not a comma. This problem is to create a linear model for this data set using the first 1400 observations. Next check the models performance on the last 199 observations. How well did the model predict the results for the last 199 observations? What measure did you use to evaluate how well the model did this prediction? Next use the model to predict the results for the whole data set and measure how well your model worked. (hint: use the r function lm and the regression example from class) Check the coefficients (use the summary function in r).

2) The objective of this problem is to see how the number of data points affects over fitting.

2a) Start with the full sonar data set (both training and test sets). Use 5 fold cross validation on the whole data set. Average the 5 training errors from each of these runs. Average the 5 test errors from each of these runs. What are these averages?

2b) Next run a series of cross validations on a series of data sets which are decreasing in size. (At a minimum you must have more data points than attributes to use lm.)

Plot both the training error and test error verses data set size on the same graph. The horizontal axis should be the number of observations in the data set and the vertical axis should be error rate.

3a) From problem 2, pick a data set size that is clearly over fit. Try to improve the result with an ensemble method. Use the small sonar data set that you have chosen as the training set and put the rest of the data into the hold out set. Generate 10 linear models using your training data set. Each of these models will incorporate a different random subset of the attributes.

To generate one of these linear models:

A) Fix n to be a number between 5 and 30. Now, choose n attributes randomly.
For example if you fixed n to be 11 then choose 11 attributes randomly
from the 60 available sonar attributes.

B) Fit the linear model to the training set using only these n attributes.

C) Use this model to make predictions on both the training set
and the hold out set.

D) Record the training error and test error.

E) Retain the predictions for both the training set and the test set
(This will become an attribute for problem 3b)

F) Rank this model. (You will have 10 models to rank.
Give the model with the lowest test error the highest rank)

3b) In this step, use linear regression to create an ensemble model. Treat the output of the 10 linear models (from step E above) as inputs to a new regression to create the ensemble model. (See the figure below.) You now have 10 new attributes for each observation (one from each of the predictions you made in step E above.) The next step is to perform the linear regression over the ensemble training set which only has the 10 new attributes. (Ignore the original 60 attributes.) Compare the performance of the ensemble model on the training set with the performance of the ensemble model on the hold out set. What are the coefficients for the 10 new attributes? Compare these coefficients with the ranks given to the models in part F) above.

Original Sonar Training Set

V1	V2	...	V60

Ensemble Method Training Set

1st Linear Model Prediction	2nd Linear Model Prediction	...	10th Linear Model Prediction

3c) Repeat Homework problem 3b with various values for n (the number of randomly chosen attributes). In part 3aA of the example above, n was fixed to be 11. Now, put n in a loop. That is in pseudo code:
for (n in seq(5,30,by=5){... create ensemble model}.
Plot the training error and test error as a function of n. How well did the new model do in comparison to the original model created by problem 2?

4) Perform a ridge regression on the wine quality data set from problem 1. Compare the coefficients resulting from the ridge regression with the coefficients that were obtained in problem 1. What conclusions can you make from this comparison?