HW 4

1. For this exercise, we will use the kidney function data that can be read into R with the following command:

```
data <-
read.table("http://www.stat.ufl.edu/~rrandles/sta4210/Rclassnotes
/data/textdatasets/KutnerData/Chapter%20%209%20Data%20Sets/CH09PR
15.txt", header = FALSE)
```

The variables (in this order) are creatinine clearance (Y), serum creatinine concentration (X_1), age (X_2), and weight (X_3). The goal is to find the best linear model in terms of MS prediction error.

After each split of the data into test and training, you will have to fit *all* possible subsets of predictors. This means you must run 7 linear regressions and store the fitted models (three models with a single predictor, three with two predictors and one with all three predictors).

We now will evaluate each model based on different cross-validation splits of the data.

Part a) must be done manually with the code shown in class.

Parts b) & c) may be done with the function from the package boot as shown in class.

a) Split the data *randomly* into 20 observations training dataset and the remaining 13 observations into test dataset. Build the 7 possible models on the training dataset and compute MS prediction errors on the test dataset. Which one results in lowest MSE?

b) Repeat part a), but this time using the LOOCV (leave one out cross validation) method, meaning that there will be 33 systematic training datasets obtained with one observation deleted at a time, and prediction will be done on the left-out observation. MSE is the average square prediction error among the 33 predictions. Which model is the best?

c) Finally, repeat the above with 3-fold cross-validation, meaning that data are split into 3 equal length subsets, and each of the parts is used one at a time as a test dataset. Then the three MSE's are averaged to get one MSE. (This is similar as part a), but repeated systematically 3 times). Again, report the best fitting model.