Project 1.

1. You will use Health data set to hone your regression skills. The data set along with description of the variables has been uploaded to UB Learns under the title “project “toy” data set”. The response variable is given in X1. Description of the variables is also given here: X1 = death rate per 1000 residents  
   X2 = doctor availability per 100,000 residents  
   X3 = hospital availability per 100,000 residents  
   X4 = annual per capita income in thousands of dollars  
   X5 = population density people per square mile
2. You will be using one of the three data sets located at <https://archive.ics.uci.edu/ml/datasets.php?format=&task=reg&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table>

This is a popular data set repository maintained by University of California at Irvine. Pick regression and attribute numerical in the left pane. The three suggested data sets are:

1. Airfoil self-noise data set
2. Combined cycle power plant
3. Real estate valuation data set

Select one of these data sets. For both parts of the project, your objective is to minimize the test error by using any suitable method. Try several methods including variants of linear regression. Use cross-validation (LOOCV, 5-fold, 10-fold).

As you can see, the “toy” set is fairly small, and the other 3 data sets are not very big either. All sets contain a relatively small number of predictors. There are papers accompanying the data sets from the UCI repository. They might provide you with a better understanding of the data set. However, you may chose to refrain from reading them before you will have completed your analysis.

You need to submit two files: R-file containing your code and pdf file containing analysis and conclusions.

Dr. Khinkis