**EE5373: Data Modeling Using R**

**Fall, 2017**

Department of Electrical and Computer Engineering

**University of Minnesota**

Lab 6: Training and testing multi-factor regression models.

Due date: See the due date shown on the class moodle page.

Goal: This lab explores in greater depth the training and testing process for multi-factor regression models.

What to do:

In the discussion of data splitting, we defined the value f to be the fraction of the complete data set used for the training set. You then test the predictive capability of your model trained on this portion of the data set using the remaining (1-f) of the data set. Comparing the predicted values with the actual values in the test set, you obtain the vector delta\_1 = (actual values) - (predicted values). This vector shows how well the model predicted the actual results for one partitioning of the data set using the sample() function. The total number of elements in delta\_1 is N= (1-f)n, where n is the total number of values in the original data set.

If you run the experiment again with the same value of f, the sample() function will allocate a different subset of the data to the testing and training sets. Call the resulting second vector of difference values delta\_2.

If you repeat this process k times, you will have k different delta\_i vectors, each with N elements. Now concatenate all of these delta\_i vectors into a single vector called D\_f. This vector will have Nk elements. The mean and confidence interval of D\_f shows how well the model predicted the actual results for k different training-testing permutations provided by the sample() function.

1. For each of the benchmark sets in CPU DB, use the process described above to plot the mean and 95 percent confidence interval for D\_f for f = {0.1, 0.2, 0.3, …, 0.9}. When you are done, you will have a graph for each benchmark showing how the mean and confidence interval of D\_f varies as f changes. Use k=100 to ensure that you have a reasonable number of data points for each confidence interval.
2. Discuss what your results show. For example, what value of f gives the best results for each benchmark? Why? What happens when f is at the extreme values? Why? What other interesting observations can you make?

What to turn in for grading:

Write a short lab report that includes the above graphs, along with the discussion of your results. Upload the pdf file with your report to moodle by the due date.