**IE 7374 Machine Learning in Engineering**

**Team 8 Lab 2 Report**

**Team Member: Yaodong Chen, Lin Zhu, Yuxiao Ma**

**Cross-Validation and Model Selection**

**Dataset description**

* Lab2.csv: There are 8 predictor variables, the response variable is y which is the weight. There are a total of 79 observations. In this project, we will perform 10-fold cross validation to choose between two models: The eight-predictor model and the two-predictors model with only x1 and x2.

**Major steps**

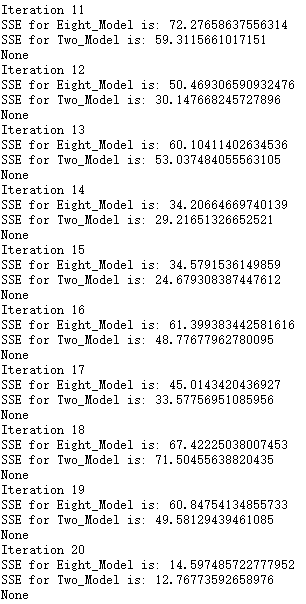
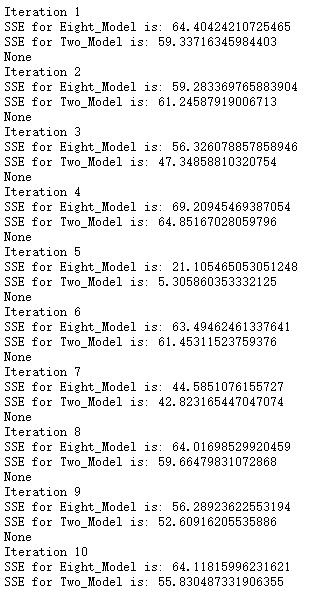
* Randomly assign all the observations into 10 parts, each part should be as balanced as possible.
  + Apply this method to both models.
  + The output data should be the same each time before cross-validation.
* Implement cross-validation to calculate the SSE value for two models and compare.
  + SSE values for the two models.
* Replicate the cross-validation process for twenty times, each time using a different random partition (apply to both models), compare the SSE values twenty times.
* Discuss the result whether or not consistent from replicate-to-replicate and explain how to reconcile the inconsistency.
* Conclusion regarding the performance between two models.

**Process**

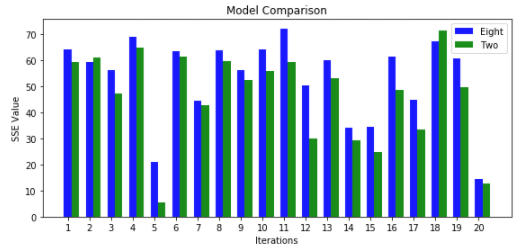
This is what we have done in this lab assignment

* We defined a linear regression model to make predictions, and the function returns the SSE values as required so that we can compare the performance of two models.
* We defined a cross validation split function to randomly divide the entire dataset into ten folds.
* We defined a k-fold cross validation function to call the cross validation split function above to do the k fold and choose one as test set and remaining as train set. The partition is the same for both models before making prediction.
* The reason why assignment should be as balanced as possible is because each part of the observations will take part in making the model, different numbers in each observation will lead to different variance. To reduce the influence of numbers in each part, we need to make the sizes of the parts as balanced as possible.
* The reason why we need to implement same partition is because different input data have a big influence on the model. Even the same model may have different accuracy when testing different dataset. Therefore, to compare the accuracy of models, we need to use the same data for our training and testing model.
* We wrote a while loop to iterate the process twenty times to observe the result.
* We randomly chose one result from the 10-fold cross validation to make visualization to make observations.

**Result**



**Visualization**

****

**Conclusion**

The results are not consistent from replicate-to-replicate. We can compare all the 20 replicates, and we can conclude which model is better by counting the number of data parts that each model fits better than the other one. In our linear regression model, from both the results and our graph above, we observe that most of the time, 8-predictor model has a higher SSE value than that of 2-predictor model. In our conclusion, 2-predictor model is better than the 8-predictor model.