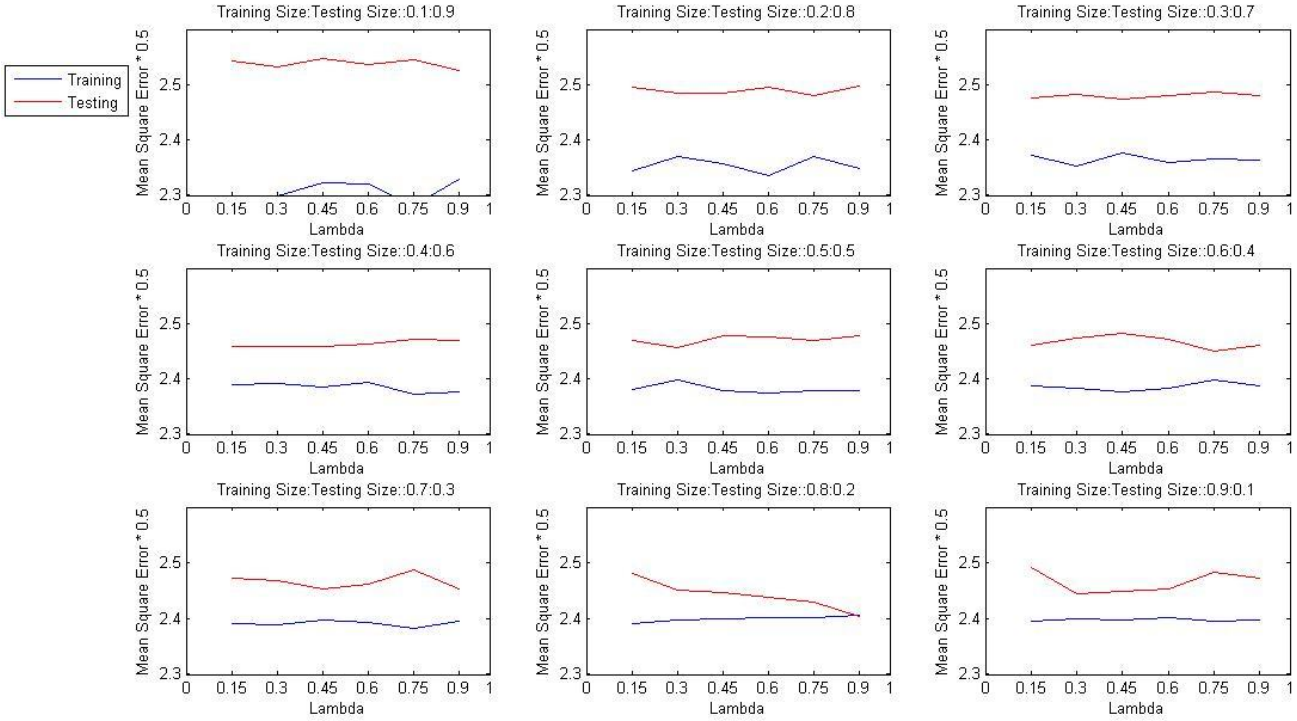
Prakhar Asthnana(2011cs1027)

Q1

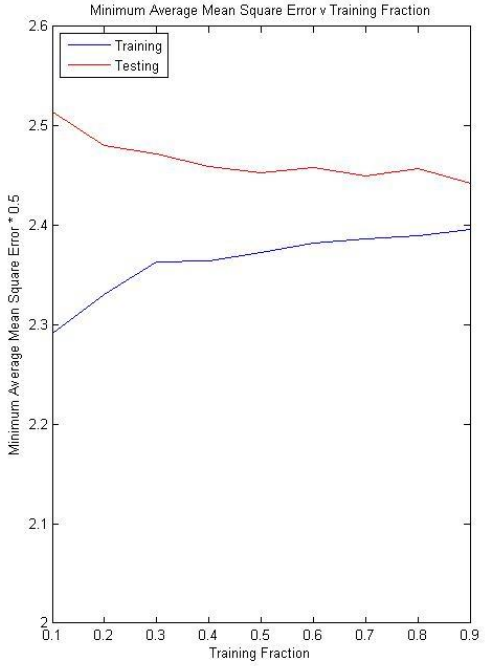
1. Program was run for over 100 times for the lambda value=0.75 and training split=0.8. Least significant attribute was generally the male/female/infant information of the abalone. The most significant attribute was generally the 6th attribute (in the original dataset-without encoding of male/female/infant).

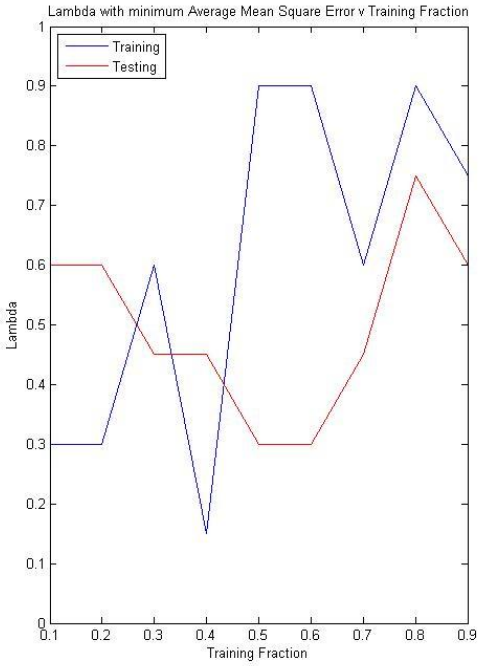
   After removal of least significant attributes, average mean squared error increases for both training(from 4.7872 to 4.9044) and testing (from 4.9428 to 5.0096).

2. Program was run for the lambda values between 0 – 1 and split fraction from 0.1 to 0.9 for 100 times. Then the average mean square error was calculated for both testing and training data.



3. Also minimum average mea squared was calculated for varying lambda for varying training split

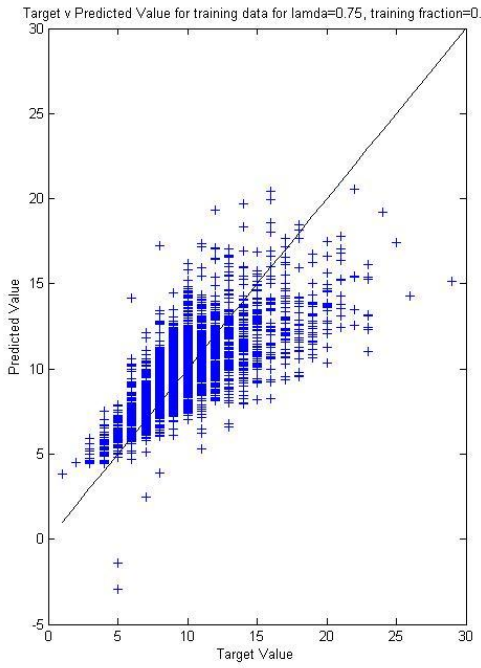


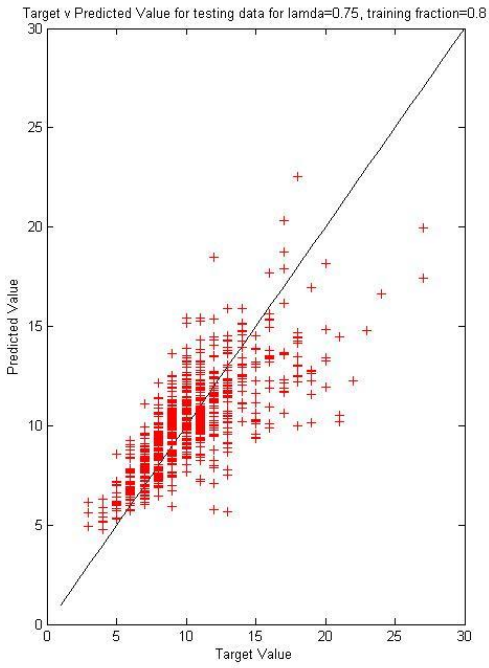


From plots, we can see that

- lambda value at which minimum testing average mean squared error is obtained, decreases in general with increasing fraction of training data
- lambda value at which minimum testing average mean squared error is obtained, increases in general with increasing fraction of training data with a steep rise as training fraction increases beyond 0.5.

4. Plot of predicted vs actual value for lambda=0.75 and training fraction=0.8





Q2

a. RSS for quartic will be less than RSS for linear. Because quartic will fit better
b. RSS for Linear will be less in case of test data as quartic regression results in overfitting.
c. RSS for linear will be less in case of quartic as it can fit better than linear in training set due to its more flexible nature
d. We can not say anything in this case as it depends upon the closeness between model and actual pattern of data.