17 January 2018
15:53

Cross validation:

We have three methods to discuss here.

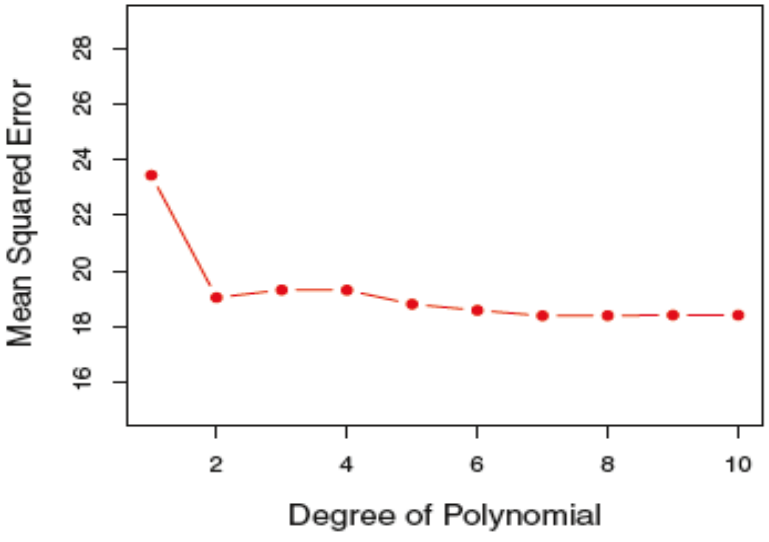
- Validation Set Approach
- Leave-One-Out-Cross validation [LOOCV]
- K-Fold Cross validation

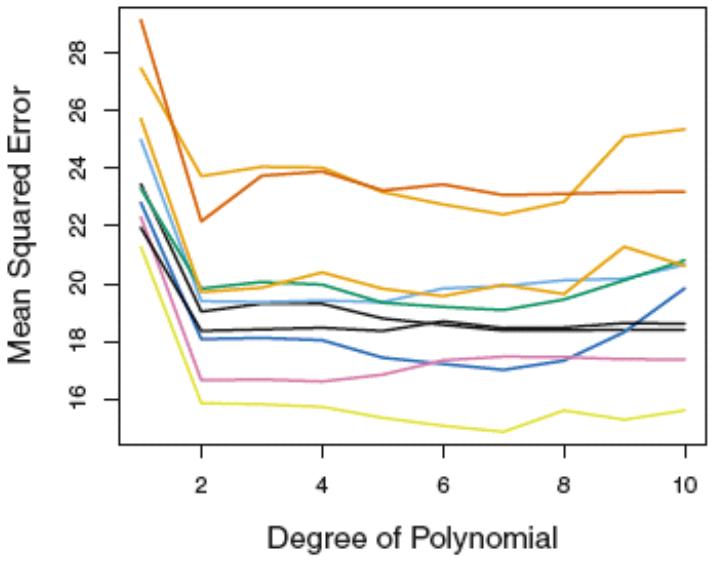
We know Test error is calculated if the test data is available. What happens when we don't have much sample data. We need to approach **Validation Set Approach**.

- Divide the dataset randomly in to two parts.

1. Training data set
2. Validation set / Hold-Out set
3. Take 10 different samples from data, different model, 10 different test errors

Ex: CV_Auto.R





Conclusion:
A model that predicts mpg using a quadratic function of horsepower performs better than a model that involves only a linear function of horsepower, and there is little evidence in favour of a model that uses a cubic function of horsepower.

Drawbacks:

- The validation set test errors is highly variable depending on precisely which observations are included in the training set and which observations are included in the validation set.

- Only subset of total dataset is used for model building. Being using of subset of data for training, statistical methods tend to perform worse when trained on fewer observations only. This suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

- Variance is less and Bias is more.

Note:
- Except in Ridge/Lasso Regression all other remaining regression methods validation and Test data considered as same.
- If we have the data on the time zone manner, we need to split first 50% as training and other 50% as validation Or else divide the data randomly in to 50 and 50 and then sort it separately.