This project report discusses the results and observations of the experimental analysis of Linear Regression Model for polynomial fitting for with and with out regularization scenarios. The Projects is programmed in Java using WEKA library.

1. Simple Linear Regression

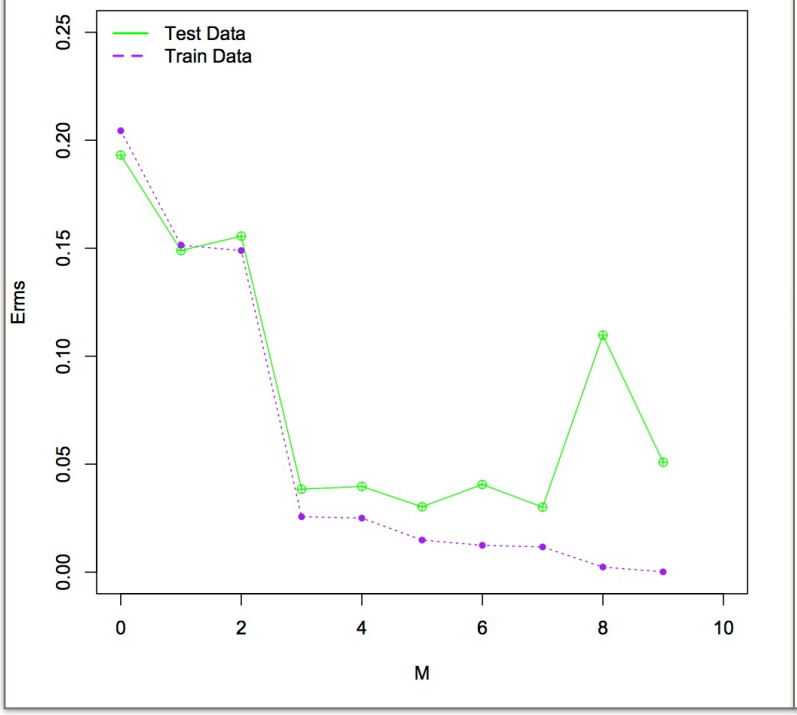
Plot A has been graphed by Training the model using Training data set and Test the model using Test data set Whereas Plot B is plotted by training the model with Training and Validation data set and Testing it on Test data set.

Observations:

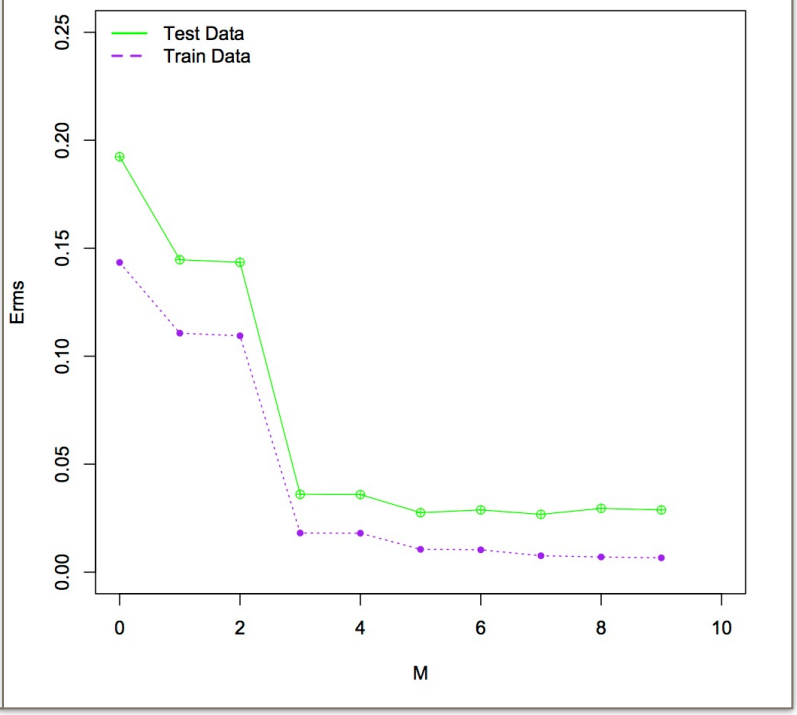
a. It is observed that increasing the amount of Training data makes the model learn the patterns more accurately and make the predictions more accurate on Test data. Its is clear that ERMS values for M(order of the polynomial) has been considerably lowered for plot B when compared to Plot A for Test and Training data, because of the addition of the validation data set to the Training data set, for training the model.

b. Its clear that any model M=3 to 6 will give good results on test data with some minor errors associated with each model. So, As per Occam's Razor it is advisable to keep the model simple and accurate. So, I would take M=3 or 5 as one of the good and simple models.

Plot A(Training vs Test)



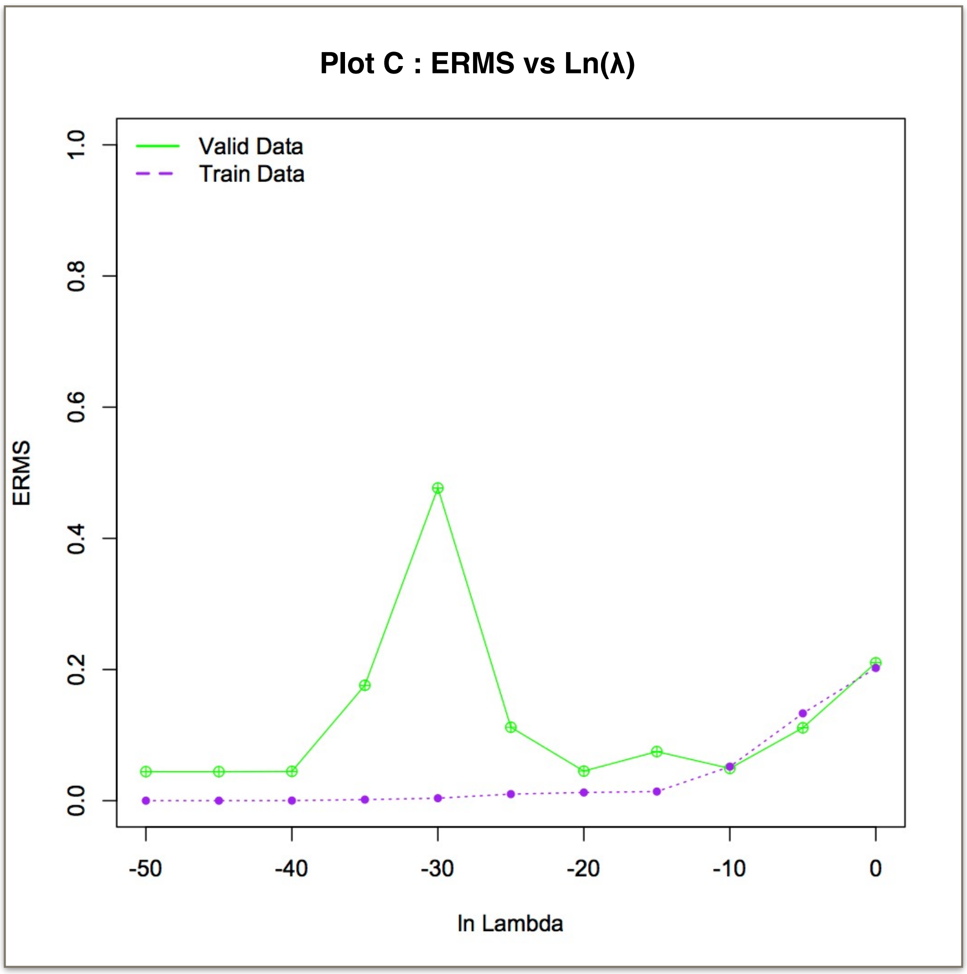
Plot B(Training+Validation vs Test Data



2. Linear Regression with Regularization

Model Selection:

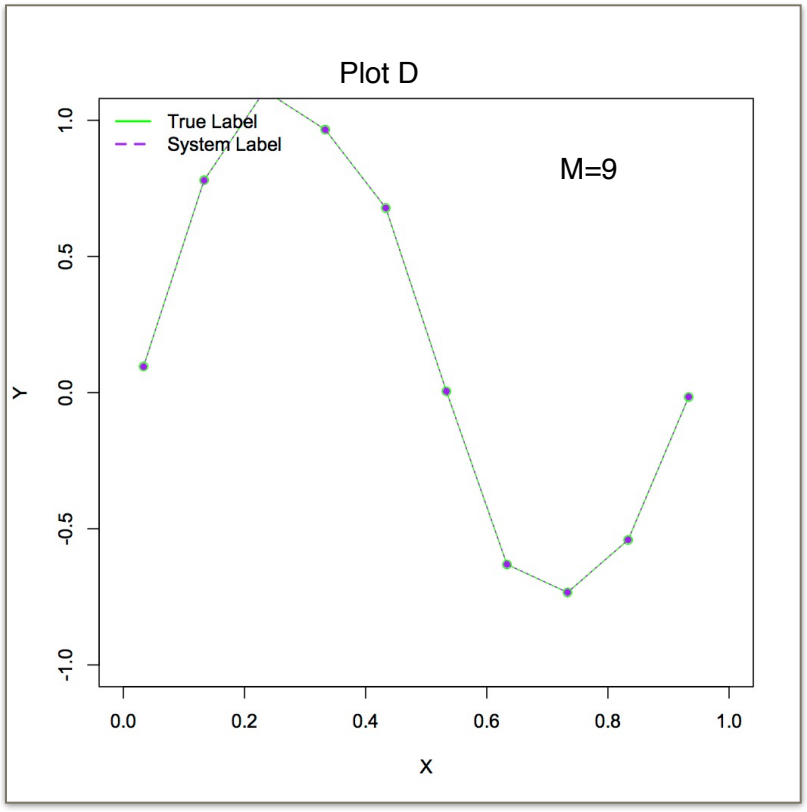
The main idea of regularization is to avoid overfitting training data of the model. Overfitting always make model give perfect results on Training data and yields poor results on test data. λ plays an important role in the regularizer. For M=9, Optimal value for λ is chosen through validation data and below **Plot C and Table 1.1** shows that ln(λ) = -50 (having lowest ERMS) would be an optimal choice for the data sets used in this experiment.



Validation Data Set											
ln(λ)	-50	-45	-40	-35	-30	-25	-20	-15	-10	-5	0
ERMS	0.044343701762335194	0.0443437039041987	0.04475672410033934	0.17601804652263486	0.4767914736367317	0.11200524795083106	0.04518160246416567	0.07491491117260687	0.04925148649189773	0.11113330087541107	0.21033871687378697

TABLE 1.1

Its also observed that with the data used in the experiment “train on training data and test on test data with M=9”, there is no change in ERMS value for Linear Regression with out regularization and with Regularization scenarios. This is because we don't see overfitting problem in linear regression with out regularization scenario as shown below in Plot D



The graph in plot D is for M=9 and linear regression with out regularization scenario. This curve is similar to sin(2*pi*x) and doesn't overfit to do bad on test data set.

ERMS Values for Linear Regression with Regularization:

ERMS Value on Test Data	
Train on Training data, Test on Test data [Mod A]	0.0508962685294462
Train on Training+Validation data, Test on Test data [Mod B]	0.02881067280512279

TABLE 1.2

It’s clear from TABLE 1.2 there is a considerable improvement in ERMS in Mod B upon addition of the validation data to the training data when compared to MOD A. This improvement is an indication that the model does very good on Test data.