CSE 574 Assignment 1

Name: Rajiv Ranjan Ubid: 50249099 ubit:rajivran

Machine Learning environment setup

Data Source: The data was picked from
https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/slump/slump_test.data
and was dumped in a csv file. The csv file is sent as attachment. 108 observations in total.
Language Used: R
Tool Used: R studio Version 1.1.419
Library used: ISLR, caret, dplyr, glmnet, elasticnet

Procedure:

1) Partition the data in such a way that 85 observations are randomly selected for training the model and 18 are used for testing the accuracy of the model.
2) Among the 85 selected observations a 5-fold technique was used which subdivides data into k randomly chosen subsets (named folds) of roughly equal sizes. One subset is used to validate the model, while the remaining subsets are used for training. This process is repeated k times, such that each subset is used exactly once for validation.
3) The best model was selected among these 5.
4) Finally, this model was ran on the test data to find the accuracy of the model.
5) This was reported through various graphs and specifying different parameters. A detailed elaboration about which has been given below.

Task 1

Part1: An unregularized regression of slump flow values against the seven explanatory variables. What is the R-squared? Plot a graph evaluating each regression.

Soln: The code for this can be found in file unregularized.r

An unregularized model was trained for the training set data. The parameters of the model are as follows:

Deviance Residuals:

Min	1Q	Median	3Q	Max
-27.1476	-9.5783	-0.0982	8.9974	22.9374

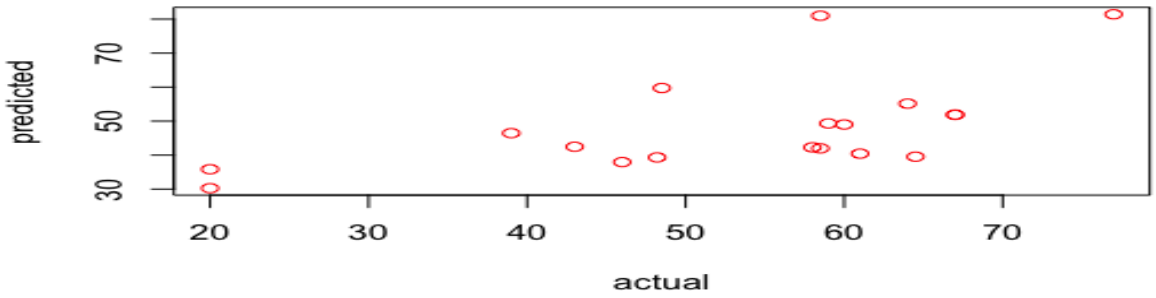
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-413.82608	377.01223	-1.098	0.2758	
Cement	0.09714	0.12295	0.790	0.4319	
Slag	0.04462	0.16895	0.264	0.7924	
Fly.ash	0.11032	0.12332	0.895	0.3738	
Water	0.94331	0.38255	2.466	0.0159	*
SP	0.81628	0.74660	1.093	0.2777	
Coarse.Aggr.	0.14055	0.14557	0.966	0.3373	
Fine.Aggr.	0.13949	0.15109	0.923	0.3588	

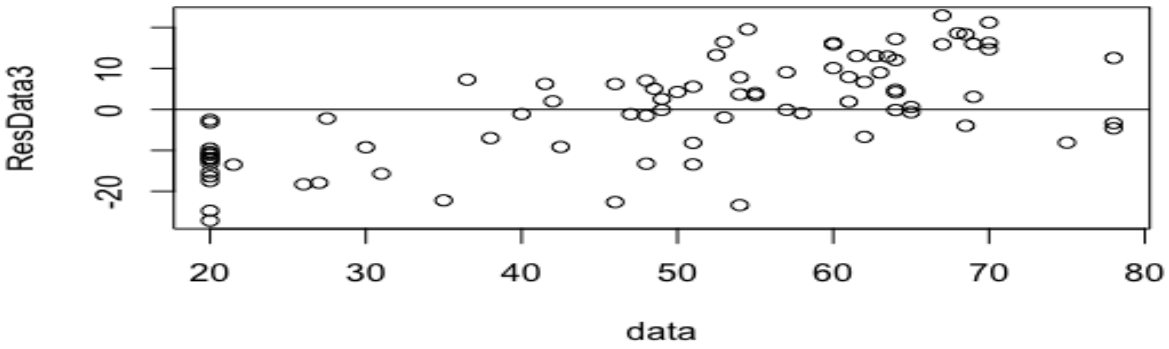
The best model was used to predict the values on the test values and following was observed:
RMSE →Root Mean Squared Error , MAE : Mean Absolute Error

RMSE	Rsquared	MAE
13.9910204	0.3289051	12.5948795

A plot of actual data vs the predicted data by the model has been plotted below:



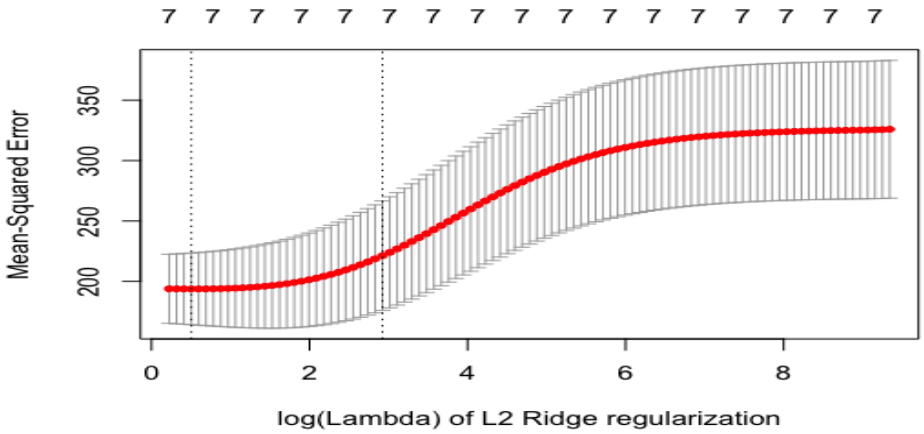
Also a plot of residuals after training the model has been plotted. A residual is a measure of how well a regression line fits an individual data point. Therefore, the model is said to fit the data well if the residuals appear to behave randomly. However, the model is clearly said to fit the data poorly if the residuals happen to display a systematic pattern. The previous figure shows that residuals appear randomly scattered around zero, with a good approximation.



The R squared value is as follows: 0.3289051
The plot has been drawn after evaluating all the methods at last.

Part2: A regression regularized by L2 (equivalently, a ridge regression). You should estimate the regularization coefficient that produces the minimum error. Is the regularized regression better than the unregularized regression?
Soln: The code is in ridge.r
In the glmnet function are used the following arguments:,alpha=0 for ridge
nlambda=100: Sets the number of lambda values (the default is 100)
lambda.min.ratio=0.0001: Sets the smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value (that is, the smallest value for which all coefficients are zero)

The model is plotted with following arguments:
plot(RidgeMod,xvar="lambda",label=TRUE) i.e. against log lambda sequence

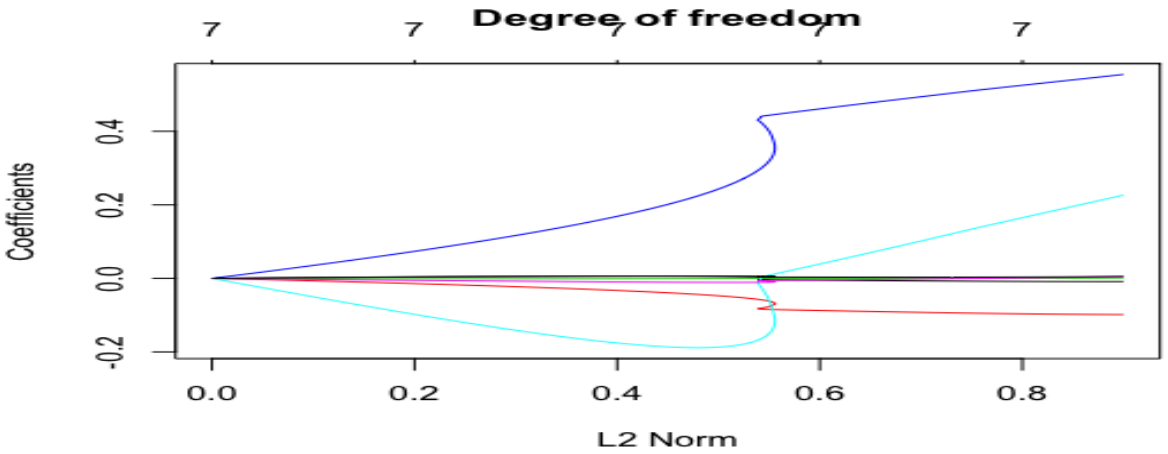


These lines then lie at two lambda values:
1)lambda.min is the value of λ that gives the minimum mean cross-validated error
2)lambda.1se, gives the most regularized model such that the error is within one standard error of the minimum
The best lambda was found to be:
Best Lambda or lambda.min ---- 1.654966
lambda.1se was reported to be 18.59063.

The value of the coefficients for best lambda or min lamda was this is where the minimum error is found by regularization coefficients

(Intercept)	-5.259289e+01
(Intercept)	.
Cement	-8.467556e-03
Slag	-9.702257e-02
Fly.ash	2.984041e-04
Water	5.312943e-01
SP	1.779519e-01
Coarse.Aggr.	2.031107e-03
Fine.Aggr.	3.874275e-03

Plotting the best Regression Model as this produces the minimum error:

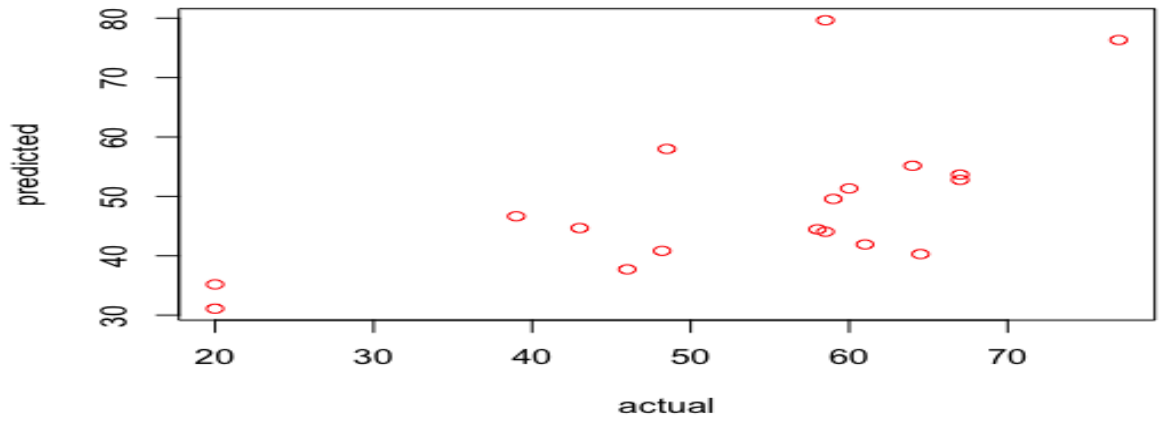


The best model was used to predict the values on the test values and following was observed:
RMSE →Root Mean Squared Error , MAE : Mean Absolute Error

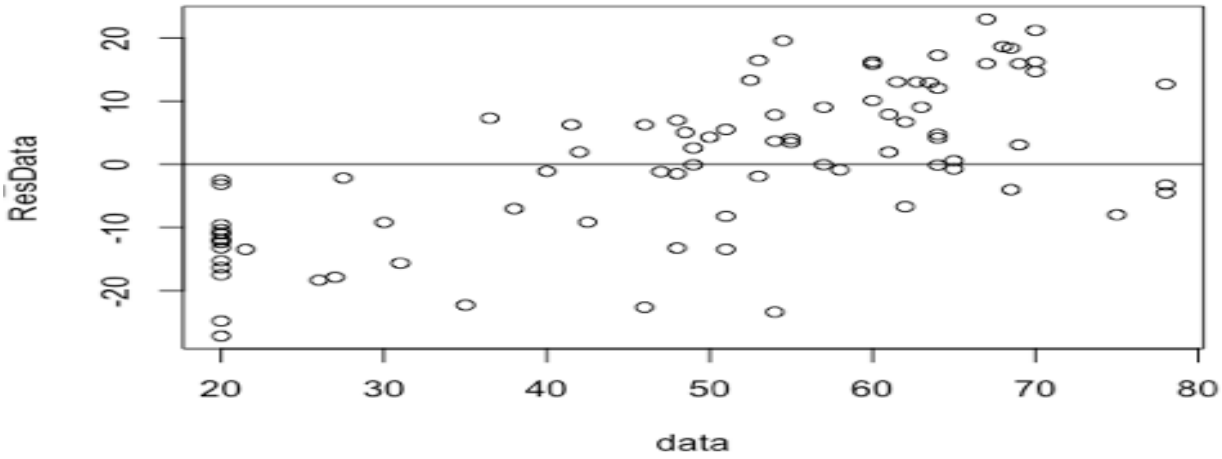
RMSE	Rsquared	MAE

12.9975220 0.3706117 11.5745868

A plot of actual data vs the predicted data by the model has been plotted below:



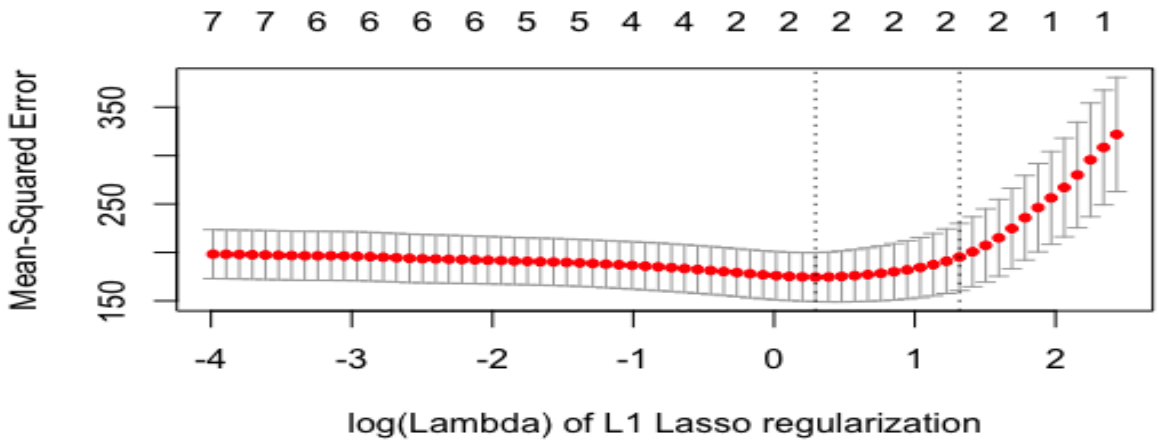
Also a plot of residuals after training the model has been plotted.



Yes, the L2 regularization is better than unregularized because
1)RMSE (L2) 12.9975220 < RMSE(Unregularized) 13.9910204
2) Rsquared (L2) 0.3706117 > Rsquared (Unregularized) 0.3289051
3) MAE(L2) 11.5745868 < MAE(Unregularized) 12.5948795

Part3: A regression regularized by L1 (equivalently, a lasso regression). You should estimate the regularization coefficient that produces the minimum error. How many variables are used by this regression? Is the regularized regression better than the unregularized regression?
Soln: The code is in lasso.r
In the glmnet function are used the following arguments:
alpha=1 for lasso
nlambda=100: Sets the number of lambda values (the default is 100)
lambda.min.ratio=0.0001: Sets the smallest value for lambda, as a fraction of lambda.max, the (data derived) entry value (that is, the smallest value for which all coefficients are zero)

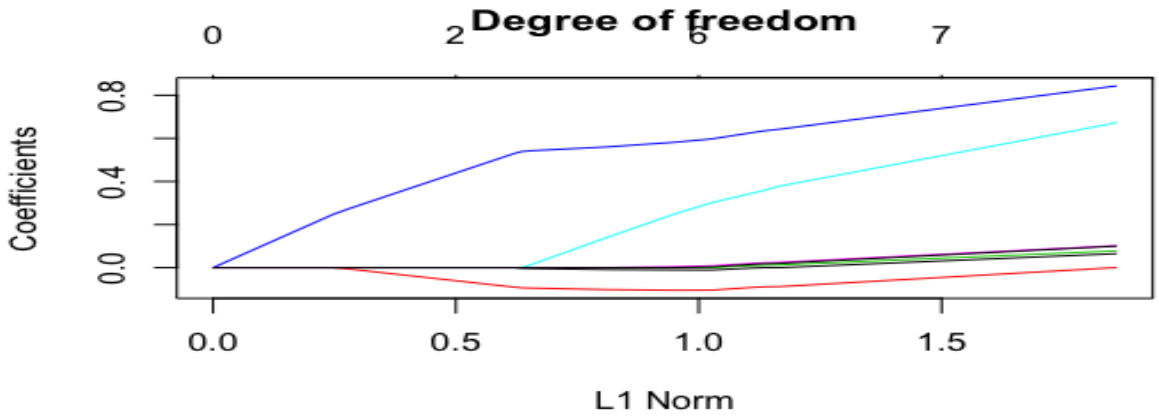
The model is plotted with following arguments:
plot(LassoMod,xvar="lambda",label=TRUE) i.e. against log lambda sequence



These lines then lie at two lambda values:
1)lambda.min is the value of λ that gives the minimum mean cross-validated error
2)lambda.1se, gives the most regularized model such that the error is within one standard error of the minimum
The best lambda was found to be:
Best Lambda or lambda.min ---- 1.342394
lambda.1se was reported to be 3.735291.

The value of the coefficients for best lambda or min lamda was this is where the minimum error is found by regularization coefficients
(Intercept) -43.23392793
(Intercept) .
Cement .
Slag -0.08018251
Fly.ash .
Water 0.49879044
SP .
Coarse.Aggr. .
Fine.Aggr. .

Plotting the best Regression Model as this produces the minimum error:

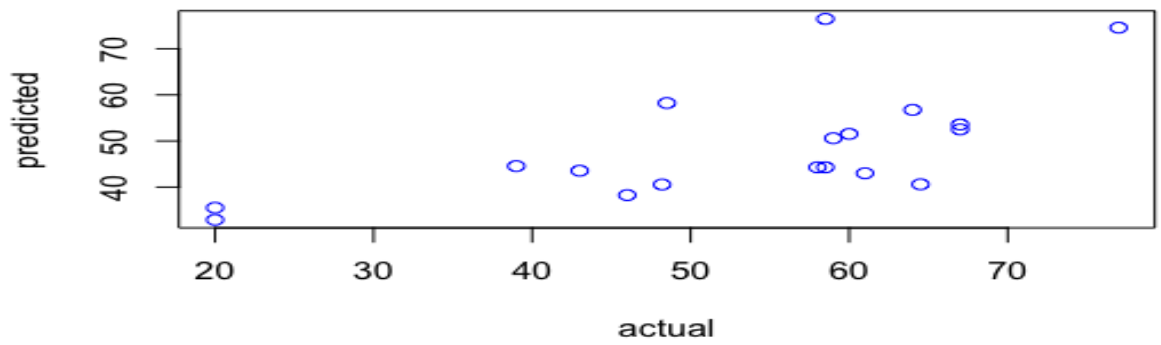


The best model was used to predict the values on the test values and following was observed:

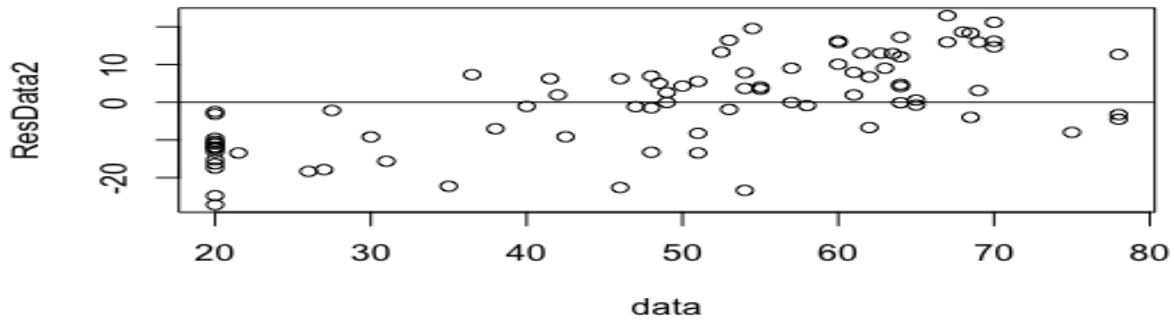
RMSE →Root Mean Squared Error , MAE : Mean Absolute Error

RMSE	Rsquared	MAE
12.5824183	0.3956832	11.2143564

A plot of actual data vs the predicted data by the model has been plotted below:



Also a plot of residuals after training the model has been plotted.

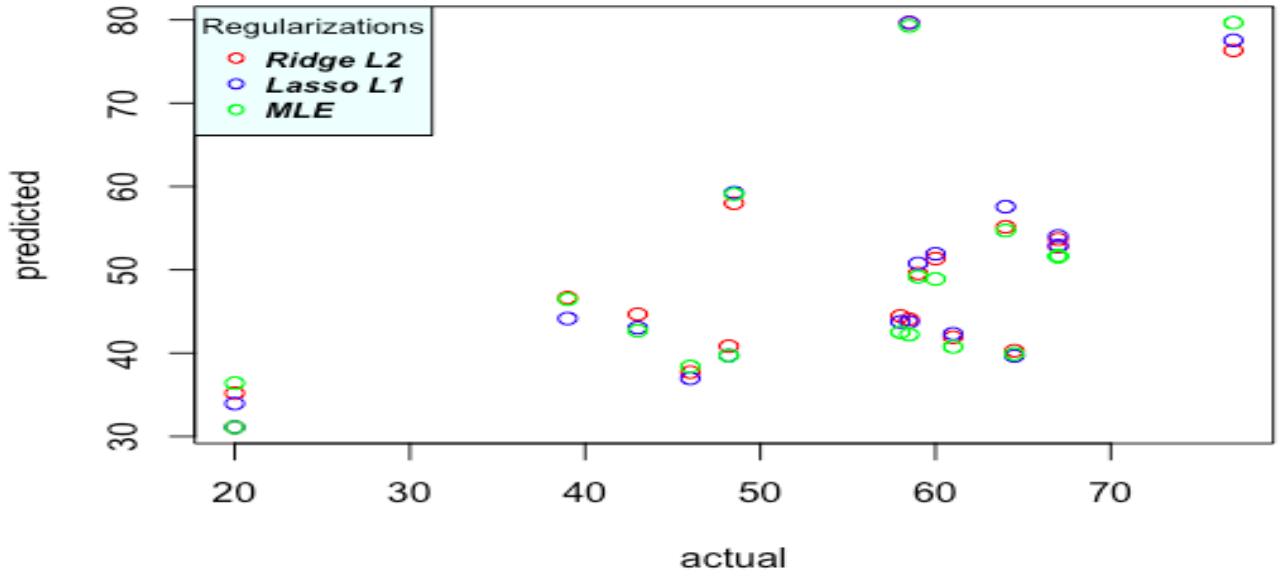


How many variables are used in this regularization?
Ans: Only two explanatory variables are used. Their coefficients are:
Slag -0.08018251
Water 0.49879044
Yes, this regularization is better than the previous 2 methods:
A comparative study of various parameters are as follows:

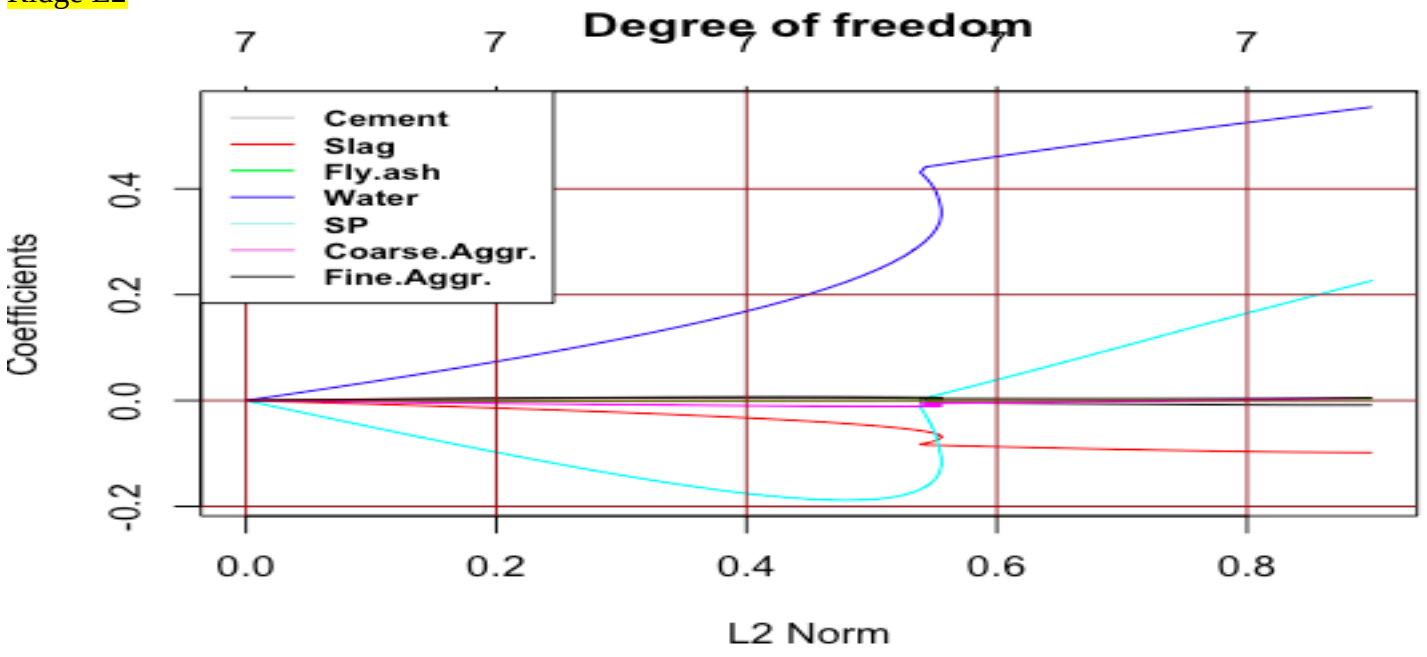
	MLE	Ridge L2	Lasso L1
R squared	0.3289051	0.3706117	0.3956832
RMSE	13.9910204	12.997522	12.5824183
MAE	12.5948795	11.5745868	11.2143564

R squared value is best and RMSE and MAE value is least among the three.

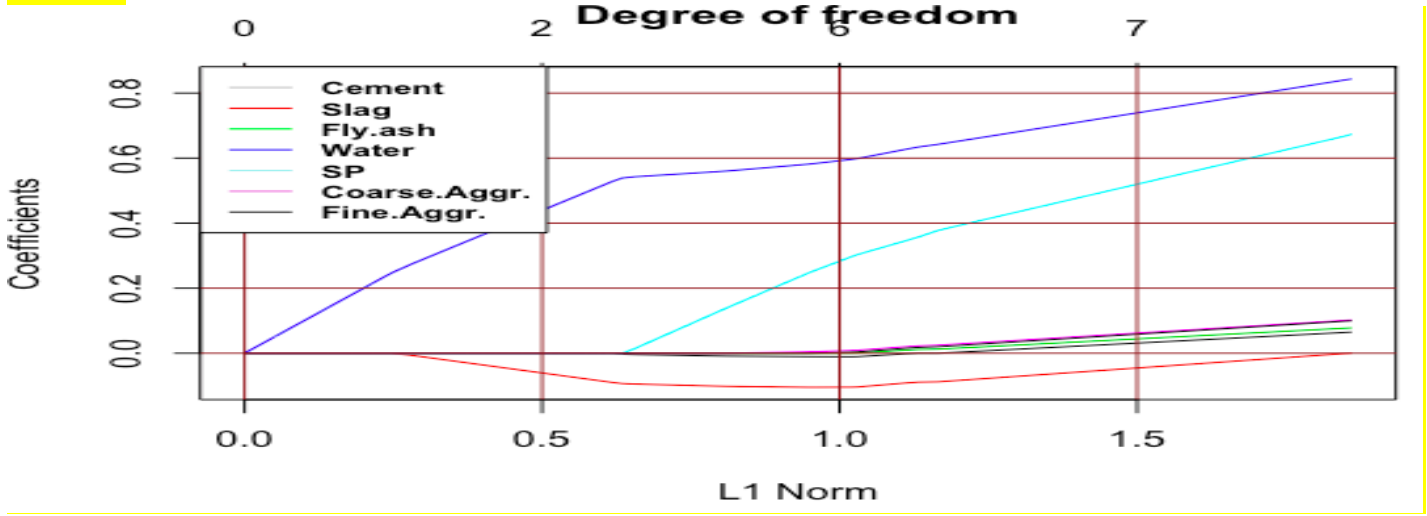
Plot a graph evaluating each regression.
Soln: A combined graph showing the actual vs predicted values by the 3 models i.e. MLE unregularized, ridge and lasso regularized has been shown below. The code is in combined.r



Task 2
Graph the regularization paths, as illustrated in Murphy figure 13.7, below (ridge left, lasso right)
Soln: The code is in task2.r The output is below.
Ridge L2



Lasso L1



TASK3:

Time taken: 1 week Collaboration: None

External Resources: Book: Regression Analysis with R by Giuseppe Ciaburro and Stack Overflow

TASK * extra EXTRA CREDIT

The code for implementation is in extra.r file

The performance of the model can be improved by using model of higher orders for example polynomial models of higher order yields better results.

This can be seen below:

Performance of Polynomial Regression of various orders:

	RMSE	Rsquared	MAE
5 fold linear model	13.9910204	0.3289051	12.5948795
polynomial model of degree 2	12.7784587	0.4153541	12.003616
polynomial model of degree 3	12.7849602	0.4801239	11.9378826
polynomial model of degree 4	12.2260392	0.5139255	10.8528654
polynomial model of degree 5	14.9423103	0.2967855	13.2206259

Analysis of the result:

1) As we can see clearly that there is a significant improvement in the model when we used a polynomial model of degree 2 as the Rsquared value increases from 0.32 to 0.41 apart from this the RMSE and MAE value also decreases.
2) As we increase the degree of the model from 2 to 3, we get an increase in Rsquared value which now shoots up to, 0.48 RMSE and MAE values are almost consistent.
3) Now we fit the test data onto a model which has been trained on a polynomial model of degree 4. An increase in Rsquared value is seen. The value becomes 0.51. Also there is a decrease in RMSE and MAE values.
4) Now if we use a polynomial model of degree 5 then, Rsquared value goes down. Also RMSE and MAE value increases.

Conclusion: The best model for the training the data would be a polynomial model of degree 4, if we go beyond that the model fails.

An actual vs predicted value graph is shown below for model of Polynomial degree 4.