CSE574 Introduction to Machine Learning
Programming Assignment 1
Modeling slump flow of concrete using MLE, Ridge and LASSO regression

Sravya Pidugu -- sravyapi -- 50249282

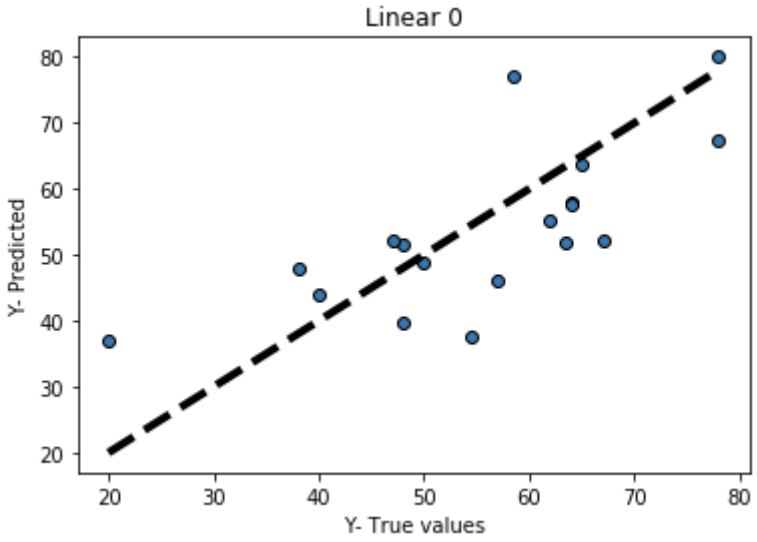
Task 0:
Infrastructure used: Python 3.6
Packages: sklearn,matplotlib,numpy,pandas

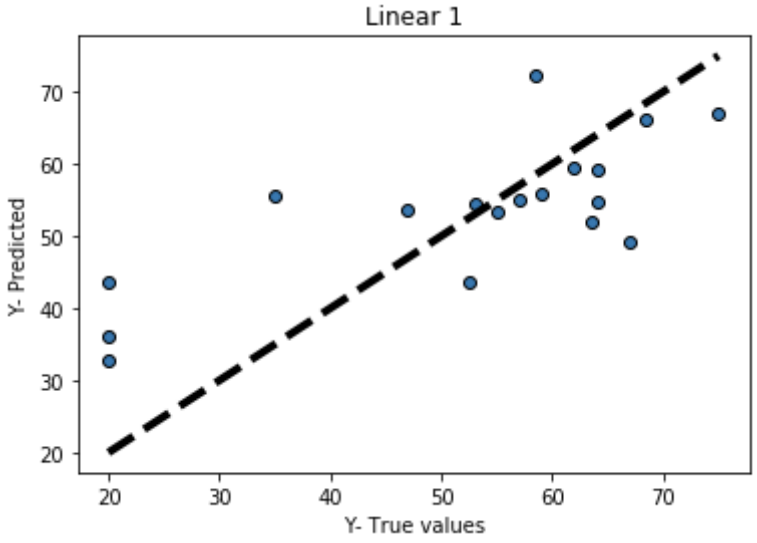
Task 1:
Perform the following regressions; saving the best model as determined by cross validation. For the three methods below, compare the average performance against test data (*not used to train or validate your models!*).
Performance scores of 3 regressions in respective best models-
1.1
AVG Mean squared error- mse: linear 159.43897532833182
AVG Mean squared error- mse: ridge 150.87596424950647
AVG Mean squared error- mse: lasso 146.9520347355642
AVG Variance score-best RSquare: linear 0.4008901173312106
AVG Variance score-best RSquare: ridge 0.4224329167301472
AVG Variance score-best RSquare: lasso 0.49424580099355677

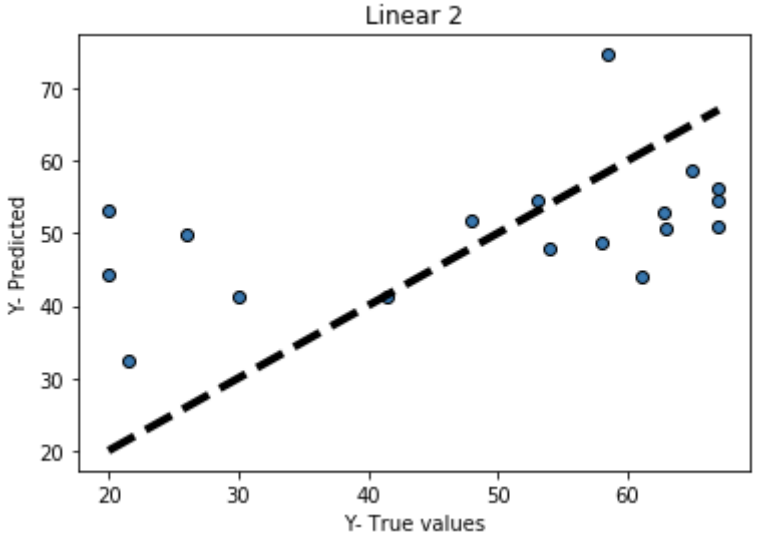
Observations-
1.The average R^2 values are least for unregularized linear regression, followed by Ridge and the highest was Lasso.
2. R2 is inversely proportional to MSE.
3.Results varied with data set but there is pattern in the output. Always linear has lesser Rsquare value than others.
1.2
R-squared values (and MSE values) Ridge regression performs a little better than unregularized regression on our data. Regularization coefficient that produces the minimum error=10
1.3
Number of variables used - 7 explanatory
response variable- 1

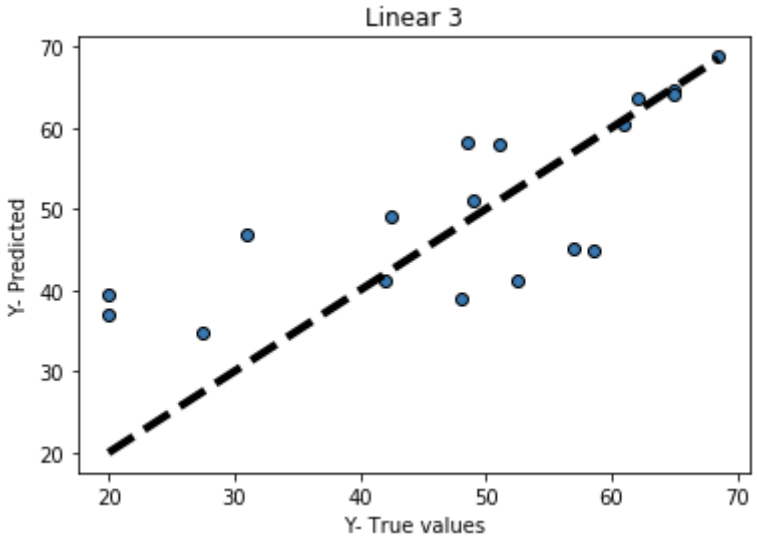
L1-regularized regression (Lasso) has better average R-squared value than unregularized regression which means it does perform better than unregularized Regression.

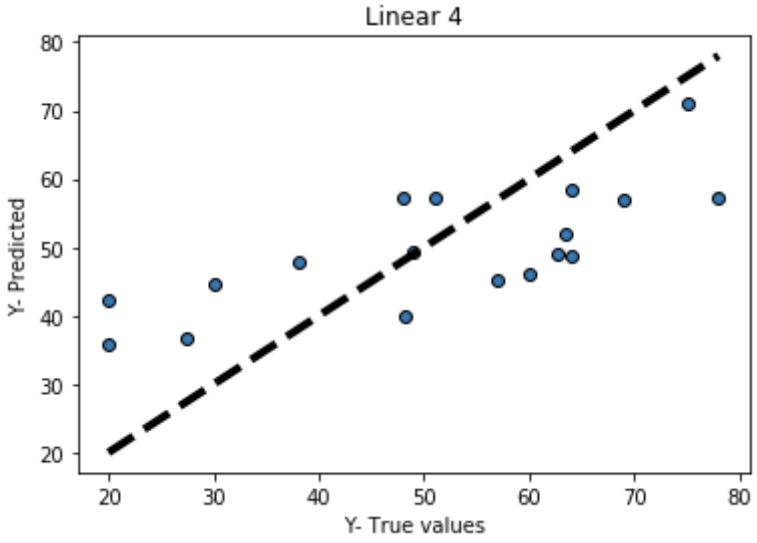
Graphs for linear regression comparing Ypredicted against Ytest in each iteration-

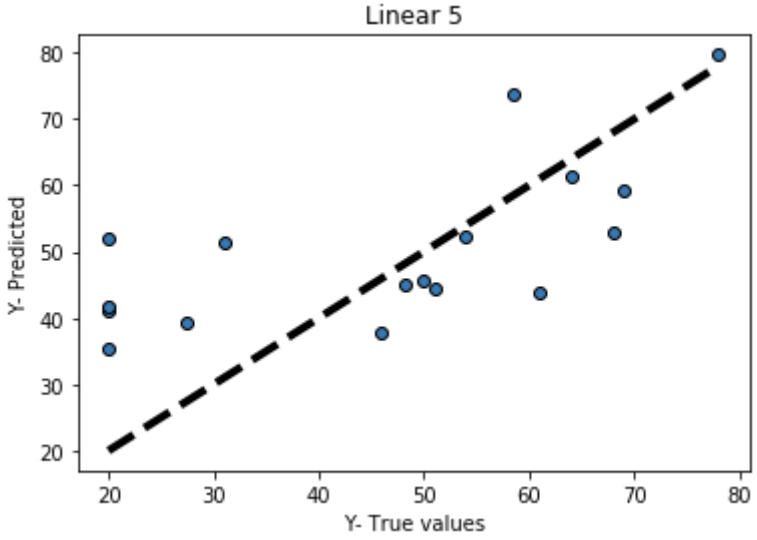

Linear 0
Y- Predicted
Y- True values
20
30
40
50
60
70
80

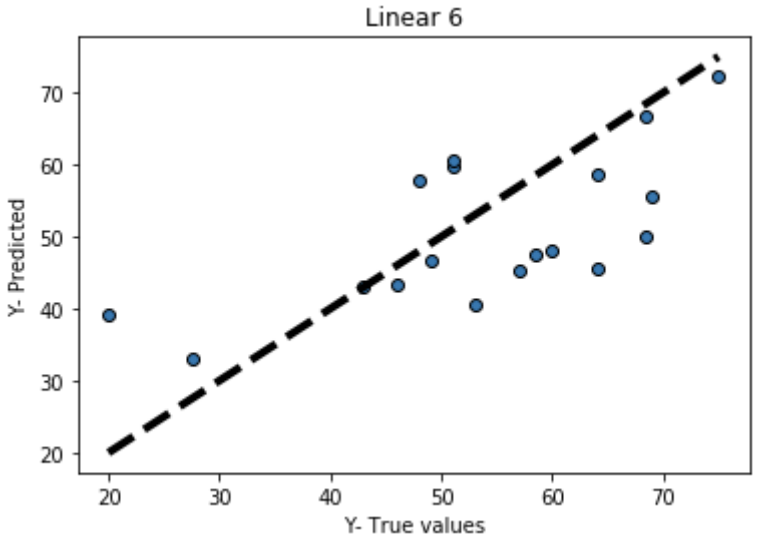

Linear 1
Y- Predicted
Y- True values
20
30
40
50
60
70

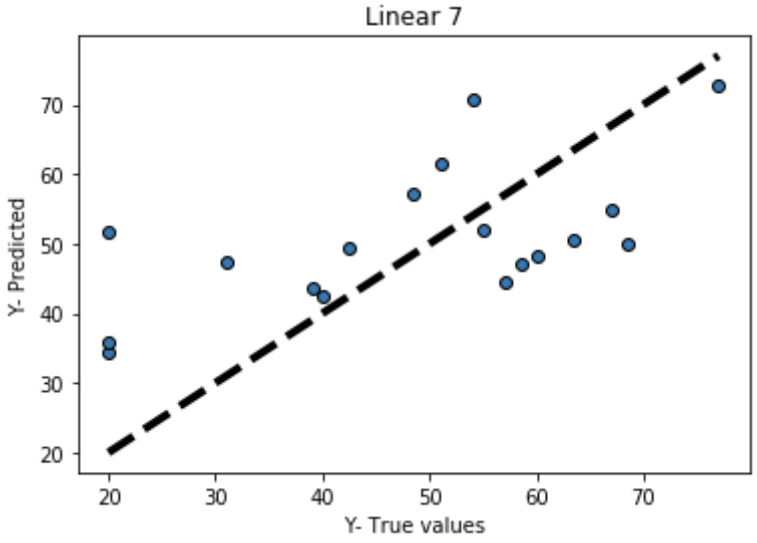

Linear 2
Y- Predicted
Y- True values
20
30
40
50
60
70

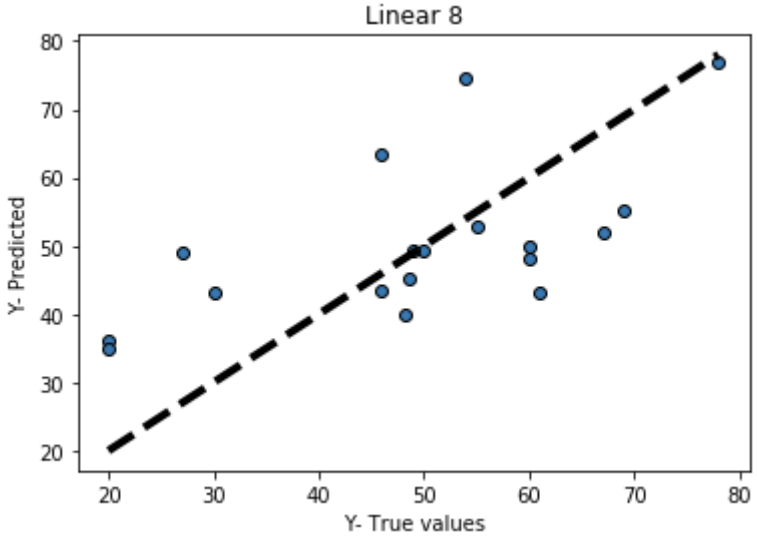

Linear 3
Y- Predicted
Y- True values
20
30
40
50
60
70

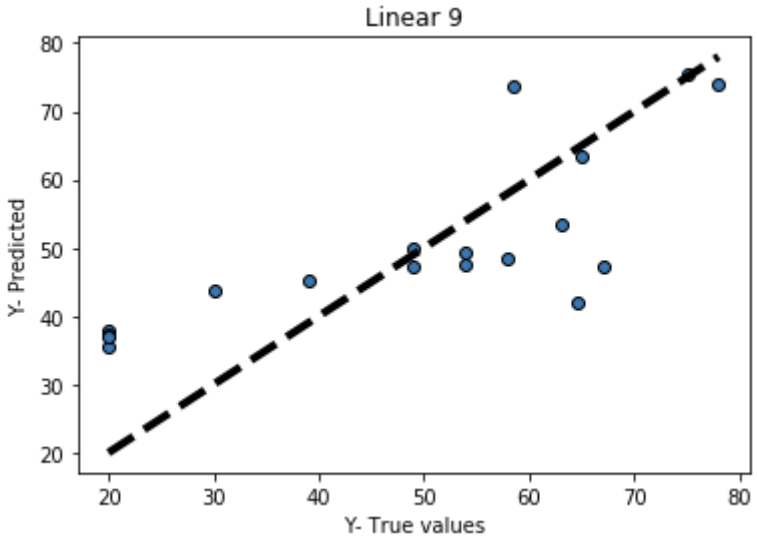

Linear 4
Y- Predicted
Y- True values
20
30
40
50
60
70
80


Linear 5
Y- Predicted
Y- True values
20
30
40
50
60
70
80


Linear 6
Y- Predicted
Y- True values
20
30
40
50
60
70

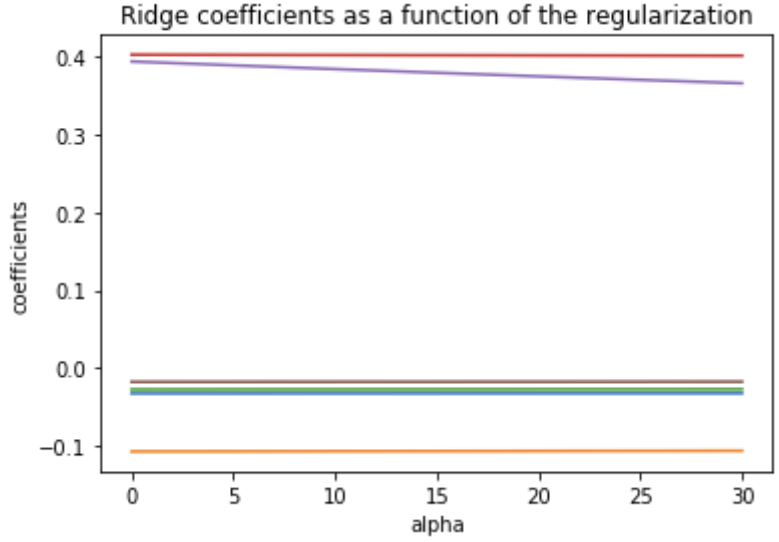

Linear 7
Y- Predicted
Y- True values
20
30
40
50
60
70

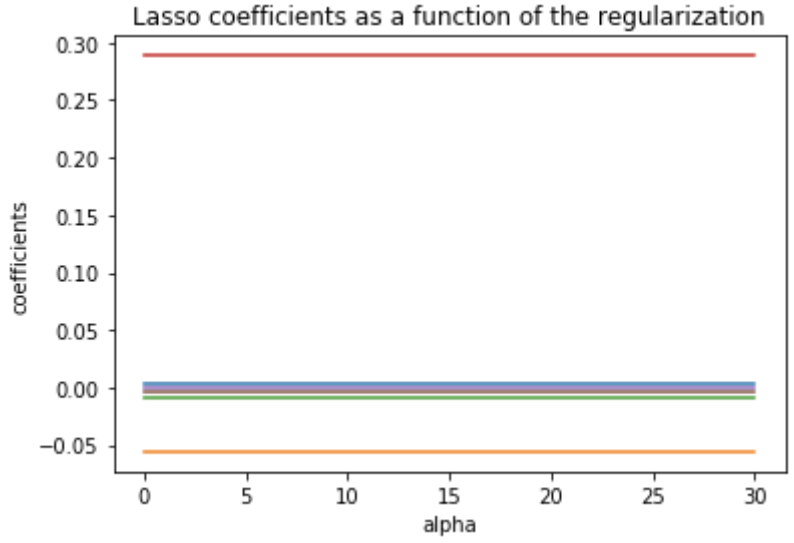

Linear 8
Y- Predicted
Y- True values
20
30
40
50
60
70
80



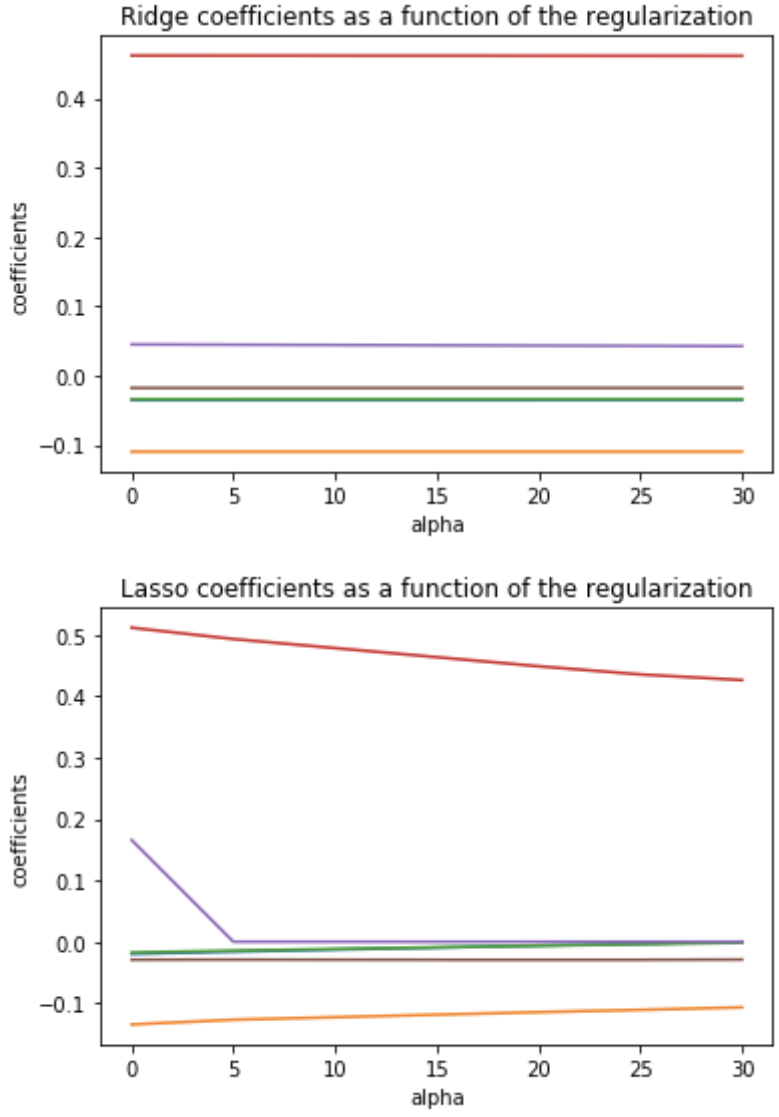
Task 2:

Regularization paths for best models-





Regularization paths for the entire dataset:



]

Time spent on the assignment: More than 20 hours.
Collaborated with: Sabreesh Iyer, Moni Pandey, Abhishek Krishna

Citations:

Approach-
https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9
https://www.analyticsvidhya.com/blog/2015/11/improve-model-performance-cross-validation-in-python-r/
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html
https://stackoverflow.com/questions/32160049/pythonscikit-different-results-for-manual-and-cross-val-score-prediction
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
Choosing alpha-
https://stats.stackexchange.com/questions/166950/alpha-parameter-in-ridge-regression-is-high

https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/
Is regularized better than unregularized-
http://blog.datadive.net/selecting-good-features-part-ii-linear-models-and-regularization/