# Analysis of various forms of Linear Regression

Jasnoor Singh
2018eeb1154@iitrpr.ac.in

Indian Institute of Technology
Ropar, PB

**Abstract**

This document contains the report as well as the graphical visualizations of variations of the linear regression(OLS,Ridge,Lasso) applied on the BOSTON HOUSING PRICE DATA-SET. It also includes multiple inferences pertaining to the various tasks performed during the assignment.

## 1 Introduction

Linear Regression is a linear approach to model the relationship between a scalar response(i.e. dependent variable) and one or multiple predictor variables(i.e. independent variables). There are various mathematical models to preform a linear regression varying from the standardized Ordinary Least Squares to the models involving regularization terms for better performance. In this assignment we'll perform and discuss three basic aforementioned variants of linear regression using various in python.

In order to simulate and visualize various mathematical models, we'll be using Sci-kit learn(Sklearn), Matplotlib, Pandas and Numpy libraries in Python.

## 2 Dataset Used

The Boston Dataframe has 506 rows and 14 columns. Each record in the database describes a Boston suburb or town. The data was drawn from the Boston Standard Metropolitan Statistical Area (SMSA) in 1970. The attributes are defined as follows:

CRIM: Per capita crime rate by town
ZN: Proportion of residential land zoned for lots over 25,000 sq. ft
INDUS: Proportion of non-retail business acres per town
CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX: Nitric oxide concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property tax rate per $10,000$
PTRATIO: Pupil-teacher ratio by town
B: $1000(Bk - 0.63)^2$, where Bk: Black Population ratio by town

LSTAT: Percentage of lower status of the population
MEDV: Median value of owner-occupied homes in 1000*s*

# 3 Tasks

## 3.1 Task 1: OLS Regression

First, the dataset was loaded and split into training and testing sets in a 7:3 ratio using Test-TrainSplit method of Sklearn. OLS regression is applied and the regression coefficients for all predictor variables are plotted using a bar graph representation.

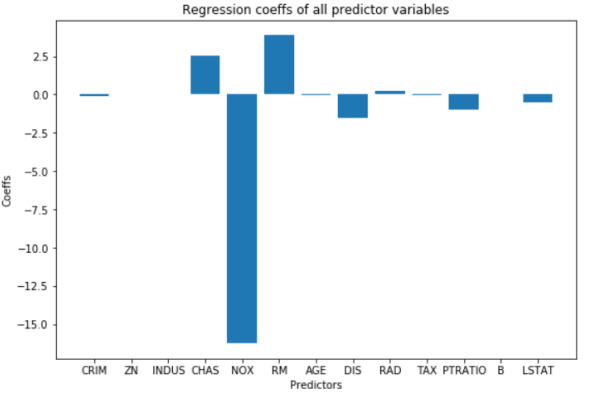


Figure 1: Bar graph plot of all regression coefficients(Normalized)

**Inferences:**

- By default, there's no NORMALIZATION in sklearn's LinearRegression model.
- No learning rate or optimization algorithm was asked for by sklearn, hence we can conclude that computes the regression coefficients using closed-form equations.
- There exists NO direct relation between the coefficients assigned to a parameter ,and its importance given the parameters are not NORMALIZED.
- For NORMALIZED parameters, the weights of coefficients tell us about the 'importance' of the respective parameter in the regression model.
- The sign of regression coefficients convey whether there's a positive or a negative correlation between the respective parameters.

## 3.2 Task 2: RIDGE Regression

Ridge Regression model used a penalty term:

$$\lambda[|B|]^2 \tag{1}$$

, where lambda is the normalization coeff. The ridge regression model generally makes better predictions than the OLS model as indirectly, this ridge regression gives higher importance to more informative features, while not dropping unimportant features.

The regression coefficients ( estimates with ridge regression for predictors: room, residential zone, highway access, crime rate and tax) as alpha varies from 0-200 plotted separately:

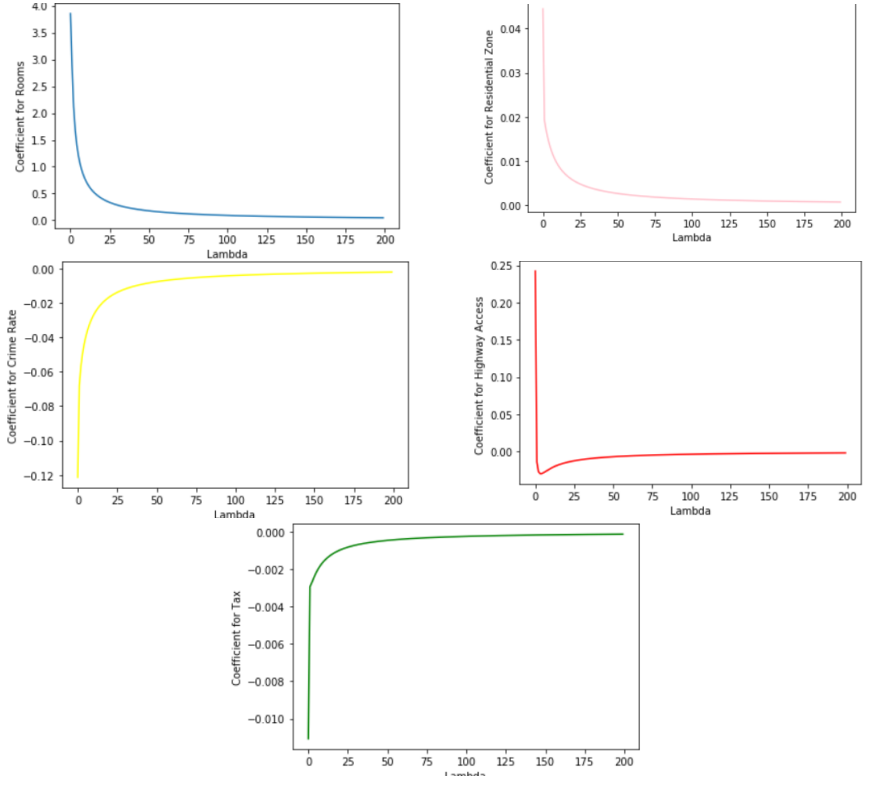


Figure 2: Variations of various coefficients with change in lambda

**Inferences:**

- As the regularization parameter lambda is increases, the regression coefficients of all the parameters tends to zero due to the penalty term.
- Higher the alpha, more the model moves towards under fitting, as the curve gets smoother and simpler. The 'importance' of the parameters is reduced depending on the value of alpha.

## 3.3 Task 3: Lasso Regression

Lasso Regression model used a penalty term:

$$\lambda[|B|] \tag{2}$$

, where lambda is the normalization coeff. Below are the regression coefficient estimates from lasso regression for given predictors: room, residential zone, highway access, crime rate, tax as lambda(overfitting coeff) varies from 0-200.

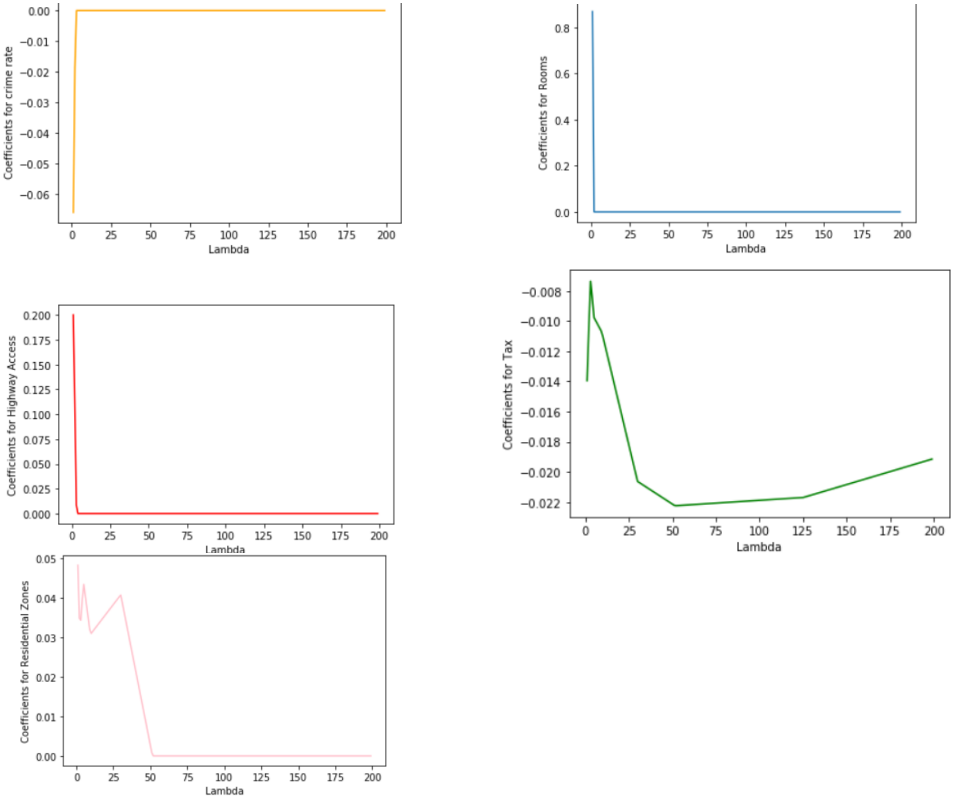


Figure 3: Variations of various coefficients with change in lambda

**Inferences:**

- The regression coefficients decrease in value with increase in lambda, same as we observed in ridge.
- The change in the slopes is much more prominent in Lasso regression as compared to ridge(due to linear penalty term).
- The smaller weights are driven down equally as the larger weights resulting in the coefficients of smaller weight becoming gradually negligible.

## 3.4 Task 4: Residual Plotting

Residuals are the differences between the actual values and the output values predicted by our model. They are an excellent way to know about the effectiveness of the model. In this task, we're going to plot the respective residuals for all the test data.
For all the above models,we plot the residuals obtained for the training data. For ridge and lasso regression, we've chosen three different values of lambda: 0.1,10 and 100.

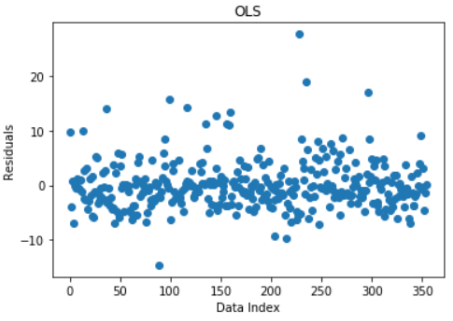


Figure 4: Scatter plot of residuals for OLS Regression

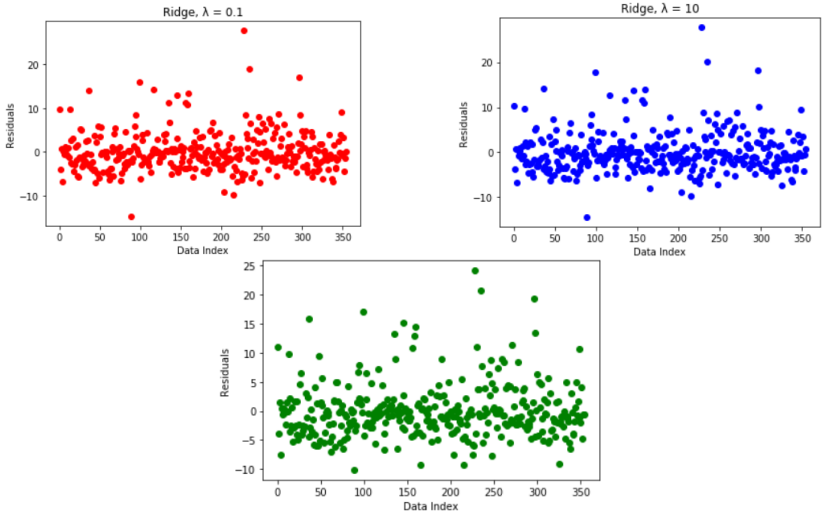


Figure 5: Scatter plot of residuals for Ridge Regression for various values of lambda

**Inferences:**

- As the value of lambda increases, we higher number of outliers in the plot which is a clear indication of presence of under-fitting in the model.
- Ideally, all residuals should be small and randomized, this would then mean that our model has been successfully able to predict the essential part of variation in the dependent variable. If not structured randomly, the residuals may predict presence of a considerable bias in the model which needs to be sorted out.
- Overlooked outliers will show up as, larger residuals. If the relationship is not linear, some structure will appear among the residuals.
- Overall, residuals are a very strong tool to diagnose the effectiveness of the linear reg. model.

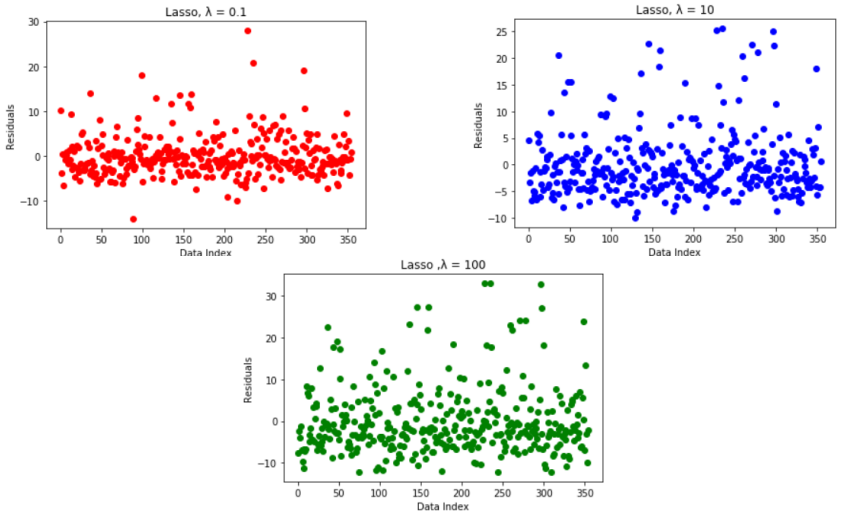


Figure 6: Scatter plot of residuals for Lasso Regression for various values of lambda

## 3.5 Task 5:Mean training and test errors

Mean squared error calculated, by averaging five distinct random shuffling of the test set keeping lambda=1 for both ridge and lasso regression.

| | Regression_Type | Training_error | Test_error |
|---|---|---|---|
| 0 | OLS | 65.271815 | 66.246352 |
| 1 | Ridge(1) | 21.744247 | 24.156750 |
| 2 | Lasso(1) | 26.506363 | 28.978511 |

**Inferences:**

- Best performance was achieved by Ridge regression, slightly better than Lasso regression.
- OLS gave the worst performance in the group, most probably due to over to overfitting.
- Model started to show underfitting as the value of lambda was increased.
- According to the results, ridge regression would be the best model for this application.

# 4 References:


- https://scikit-learn.org/stable/tutorial/index.html
- https://matplotlib.org/contents.html
- https://www.kaggle.com/c/boston-housing