Regression Analysis

Ankit Bhadu
2018CSB1073

Indian Institute of Technology
Ropar

Abstract

This document is based on analysis and application of various Linear regression techniques on the Boston Housing Dataset. The Boston housing dataset as the name suggests contains information about different houses in Boston. The primary objective is to predict the value of house price using given datapoints using various regression techniques and analysing there results.

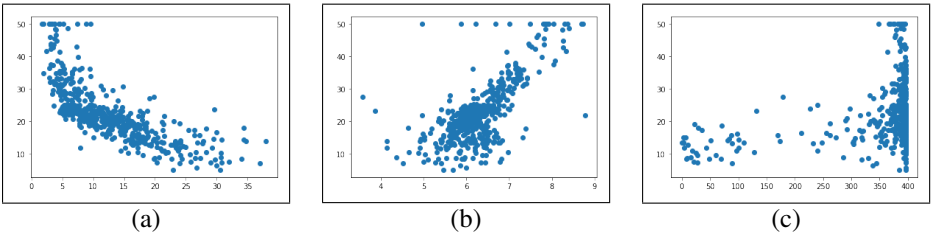
1 Introduction

Regression analysis is a statistical technique for investigating and modeling the relationship between variables especially useful when the feature to be predicted depends in a linear fashion on independent variables. Scatter plot for some of the features with the dependent features indicates that there exists a linear relationship. We perform three variants of Linear Regression namely:

1) Ordinary Least Squares(OLS) Regression
2) Ridge Regression
3) Lasso Regression

1.1 Preprocessing data

First step was to check if any data point is empty or out of the ordinary. It was found that all data points were complete. Then analysis of various features via plotting scatter plots was done and it was observed that only features like "Percentage Lower Status of population","Average number of rooms per dwelling", "1000(Bk - 0.63)exp2 where Bk is the proportion of blacks by town" displayed relationship similar to a linear one. Spilitting of training and test data was done randomly with seeding in 70:30 ratio.



(a) (b) (c)

Figure 1: Preliminary data analysis of (a)Lower Status of population,(b)Avg no of rooms, (c)Proportion of blacks per town with respect to house price.

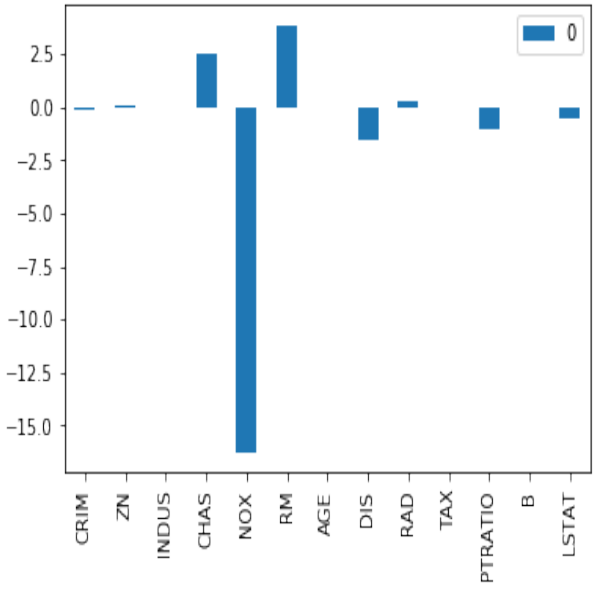


Figure 2: The values of the regression coefficients for the different predictor variables using OLS.

1.2 Ordinary Least Squares Regression

Firstly OLS Regression was performed on the dataset. Normalization is not important here as OLS is invariant. Fit Intercept parameter was kept True to fit the data bitter. The values of the regression coefficients for the different predictor variables have been plotted in Figure 2.High value of coefficient for 'NOX' represents that when it changes by one unit with other variables constant then 'PRICE' is affected the most. It doesn't mean that it has the strongest linear relationship with the independent variable as we observed in the initial processing that 'LSTAT' has probably the best linear relationship with the independent variable.The Mean Training and Test errors have been tabulated at the end of document.

1.3 Ridge Regression

ridge regression involves determining the vector of regression coefficients B= bi , whose components bi are constrained such that:

$$b0^2 + \ b1^2 + \ .. \ .. \ .. + \ bp^2 <= C^2 \tag{1}$$

While least squares produces unbiased estimates, variances can be so large that they may be wholly inaccurate. Ridge regression adds just enough bias to make the estimates reasonably reliable approximations to true population values. Ridge regression uses L2 regularization. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients.A tuning parameter ('lambda') controls the strength of the penalty term. When 'lambda' = 0, ridge regression equals least squares regression. If 'lambda' = infinity, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and infinity.

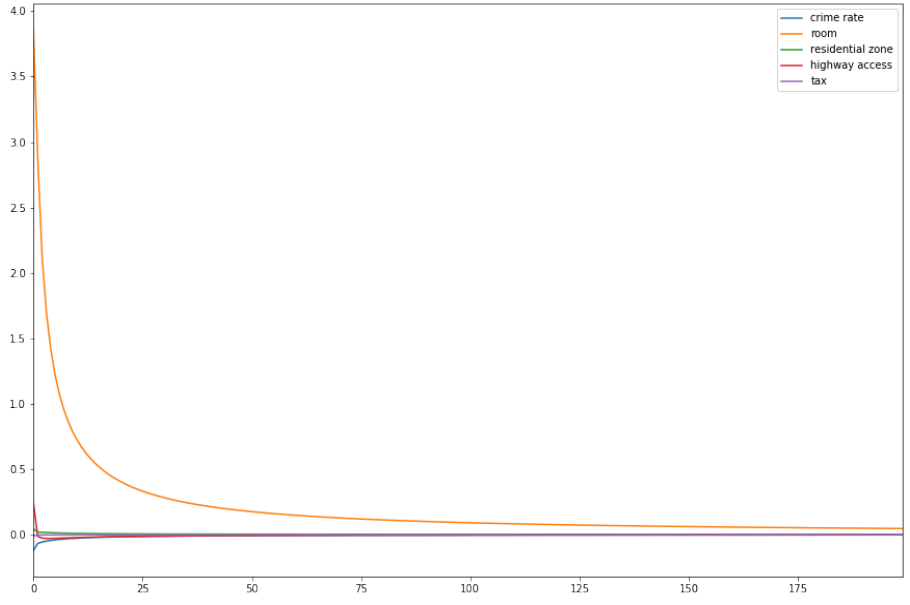


Figure 3: Variation of predictor variable coefficients with change in lambda for Ridge regression

Method	Frobnability
Theirs	Frumpy
Yours	Frobbly
Ours	Makes one's heart Frob

Table 1: Results. Ours is better.

Ridge regression for values of 'lamda' between 1 and 200 was performed and the result for some of the predictor variables is in accompanying figure.

The coefficient for predictor variable 'rooms' was the only one which had some considerable magnitude among the five chosen for analysis. This coefficient also decreases with increase in 'lambda' as expected. This is because higher penalty terms tends to minimize the number of dependent variables ie tries to make the model simpler.

Although not much change in mean squares error was observed, it did increase with increase in penalty terms from 1 to 200. Percentage change in error observed was 5 percent. It is quite significant when we want to train high accuracy models.

Since test error and training error in OLS were almost similar it can be said that there was no or little overfitting and hence when we used regularization we didn't get any advantage but rather reduced complexity of the model and thus ended up increasing error.

1.4 Lasso Regression

Lasso stands for Least absolute shrinkage and selection operator. This method is similar to ridge regression except for the way in which the regularization term is modelled. Here, the penalty term involves the sum of absolute values of the features rather than their squares.

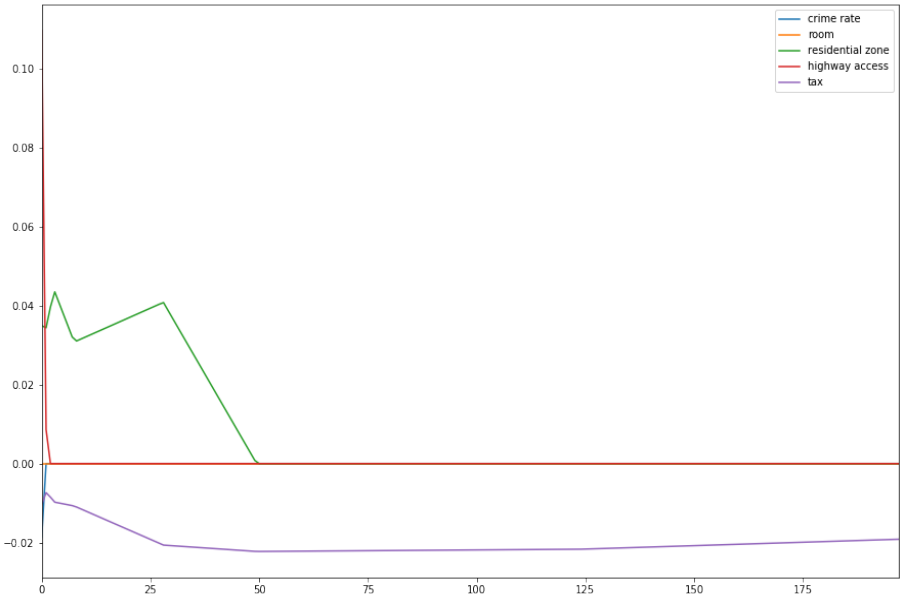


Figure 4: Variation of predictor variable coefficients with change in lambda for Lasso Regression

Regression was performed with and without normalization and results were quite similar, so we will consider the one without normalization as the graph obtained(lambda vs predictor var coefficients) without normalization seems to be more varying.

Lasso regression for values of 'lamda' between 1 and 200 was performed and the result for some of the predictor variables is in accompanying figure.

Lasso regression is basically used to reduce the no of dependent vcariable and that is what can be observed from the figure as it reduces all the coefficients to near zero values hence making the model simpler but at the cost of accuracy as the error for lambda equals 1 is almost similar to that of Ridge and OLS regressor but as the lamda is increased error changes significantly. There is a 45 percent increase in the test error and even more in the training error when lambda is increased from 1 to 200.

1.5 Visualizing residuals

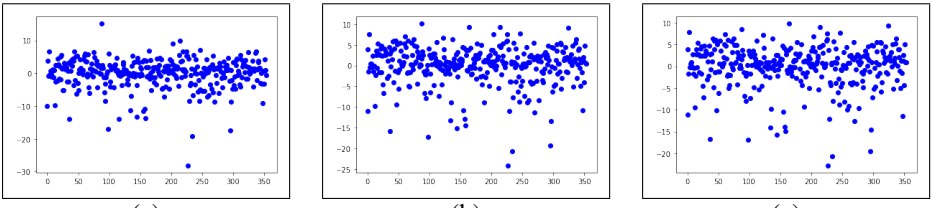
In regression analysis, the difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the residual (e). Signs that our model is a good one can be interpreted through residual plot as:

(1) they're pretty symmetrically distributed, tending to cluster towards the middle of the plot

(2) they're clustered around the lower single digits of the y-axis (e.g., 0.5 or 1.5, not 30 or 150)

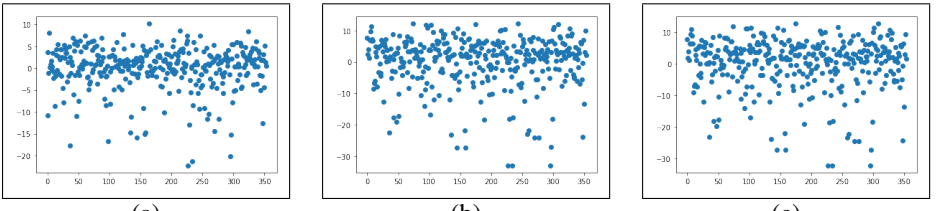
(3) in general there aren't clear patterns.

Actual residual plots for three values of lamda [1,100,200] have been shown in the accompanying figures. The values of lamba are so chosen that we get a fair idea about how plot varies with lambda as they are equally distributed in [1,200].



(a) (b) (c)

Figure 5: Residual plot for Ridge regression for lamda (a)1, (b)100, (c) 200.



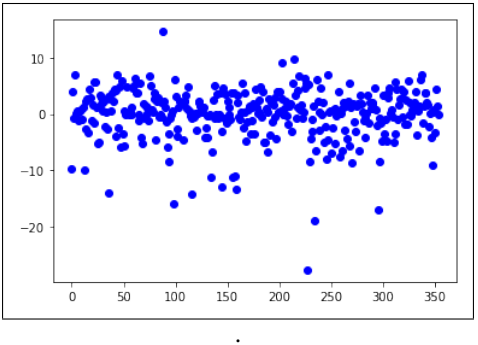
(a) (b) (c)

Figure 6: Residual plot for Lasso regression for lamda (a)1, (b)100, (c) 200.

As can be observed from the obtained residuals plots, they seem to follow all the mentioned points for a good fitting model. Residuals are distributed about y=0 randomly and there doesn't seem to be a patterns, however the spread increases as lambda increases which signifies the increase in error that was previously recorded numerically. Hence observations are in line with numerically obtained error data.

1.6 Conclusion

To conclude the above findings it can be said that penalty parameter or regularization doesn't seem to have any benefit when it is observed that OLS model is not overfitting and rather regularization may on the other hand may increase both training and test error while trying to simplify the model. Overfitting was not observed in any of the models as training and test error do not have considerable difference, atleast not so much as to declare the model overfitteing.



.

Figure 7: Residual plot for OLS regression.