Regression Analysis of Boston House Price Dataset

Praful Gupta
2018CSB1112

Indian Institute Of Technology
Ropar , Punjab

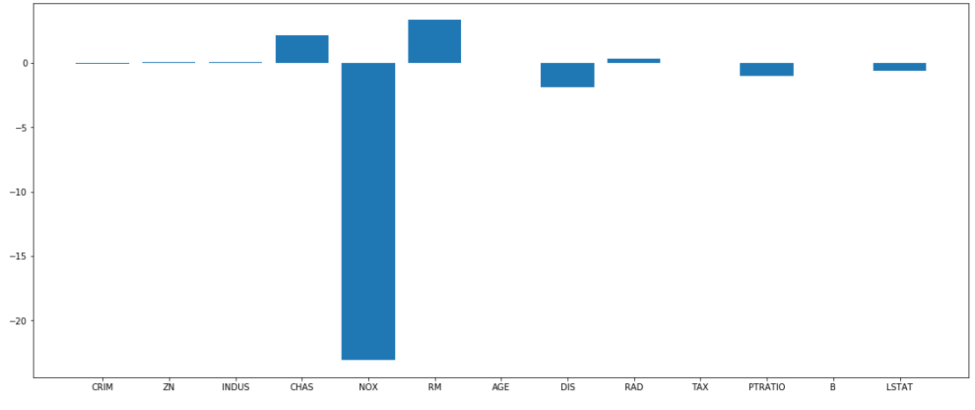
Abstract

This report analyses coefficient of different Linear Regression under different constraint on loss function and its coefficients. Also there is a comparison between different linear regression model trained on the same data of Boston House Price Dataset .

1 Task 1 - OLS Linear Regression

After loading the Boston Dataset from sklearn.dataset library into Pandas Dataframe , the Dataframe was 30:70 randomly split into test and train sets using train_test_split from sklearn.model_selection.
From sklearn.linear_model, LinearRegression model was imported and fitted into the training data. The regression coefficients were plotted in bar graph and were found to be ranging from approximately -20 to 5. However, the OLS model is prone to overfitting, implying that it could result in low bias and high variance.

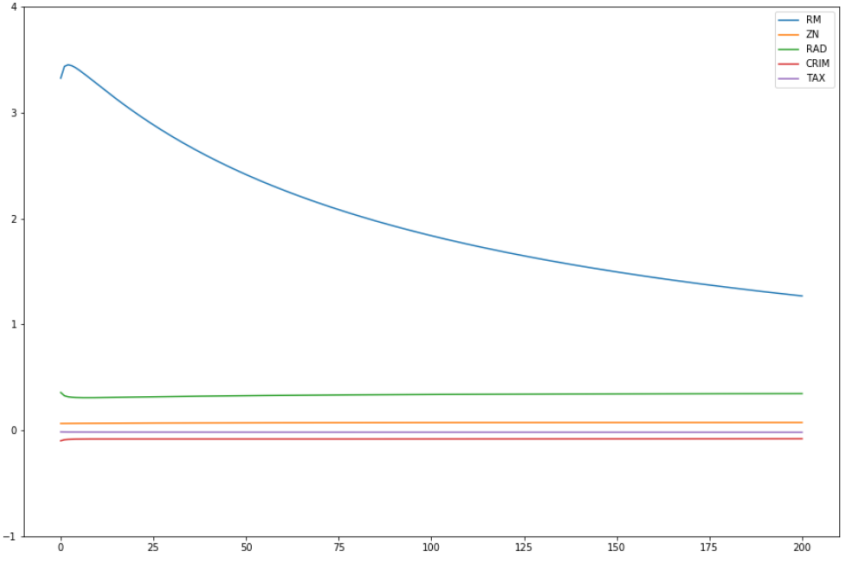
Figure 1: OLS Linear Regression Coefficients



2 Task 2 - Ridge Regression

To ensure that the regression model does not overfit, ridge regression involves determining the vector of regression coefficients B= βi whose components βi are constrained.
λ also known as the regularization coefficient, that is designed to synthesize a 'smoother' line fit. λ was varied from 0 to 200 and appropriate Ridge regression model were obtained. It was observed that some of the coefficients reduced almost exponentianaly as λ increased from 0 to 200 while others were almost constant. The ridge regression model generally makes better predictions than the OLS model. Indirectly, this ridge regression gives higher importance to more informative features, while not dropping unimportant features.

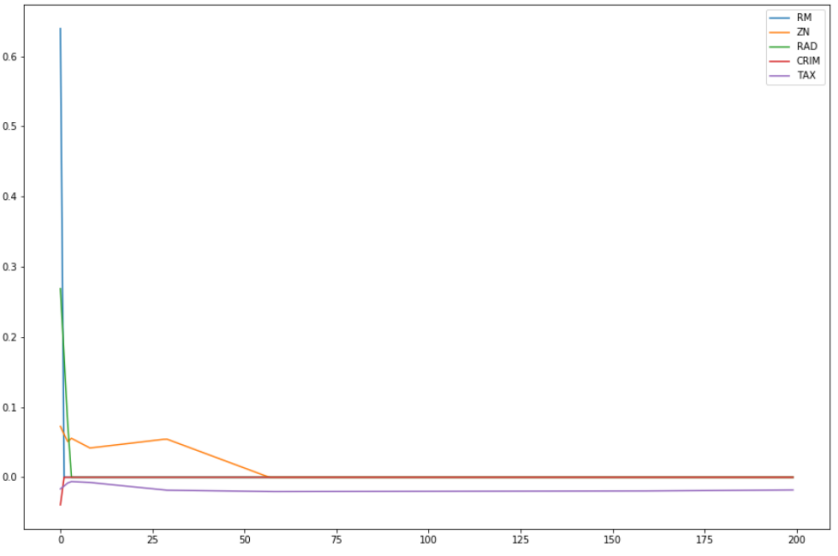
Figure 2: Ridge Regression Coefficients



3 Task 3 - Lasso Regression

Another variant of the ridge regression formulation is the lasso regressor, where 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.
This type of regularization (L1) can lead to zero coefficients i.e. some of the features are completely neglected for the evaluation of output. So Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.
It was observed that the almost all regression coefficients immediately tend to 0 as λ increased from 0 to 200 and became constant.

Figure 3: Lasso Regression Coefficients



4 Task 4 - Residual Plots

Residuals were plotted for Regression models with λ values [1,50,200] for Ridge and Lasso Regression and Simple OLS model.

1. It was observed that the Residual plot of Simple OLS regression was almost near to the 0 line but the lowess of graph showes a quardatic curve.
2. The residual plot of Ridge regression for λ values [1,50,200] almost remained constant with varying λ .
3. The residual plot of Lasso regression for λ values [1,50,200] clustered near range of 20 - 30 as λ increased.

Figure 4: Ridge Regression Residual Plot

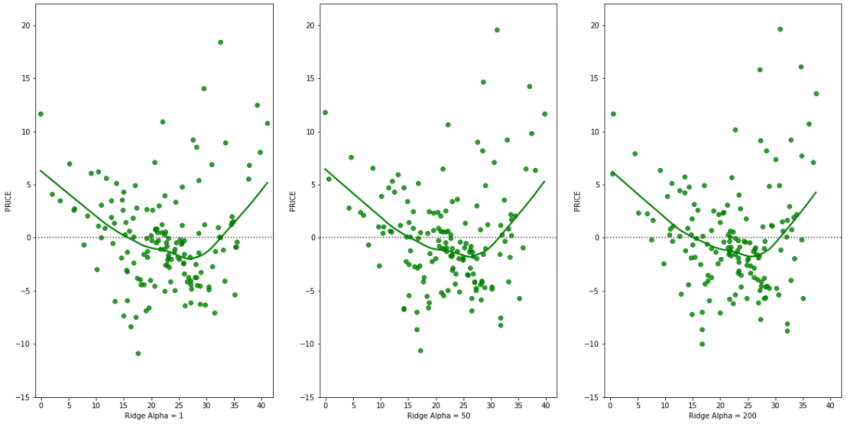


Figure 5: Lasso Regression Residual Plot

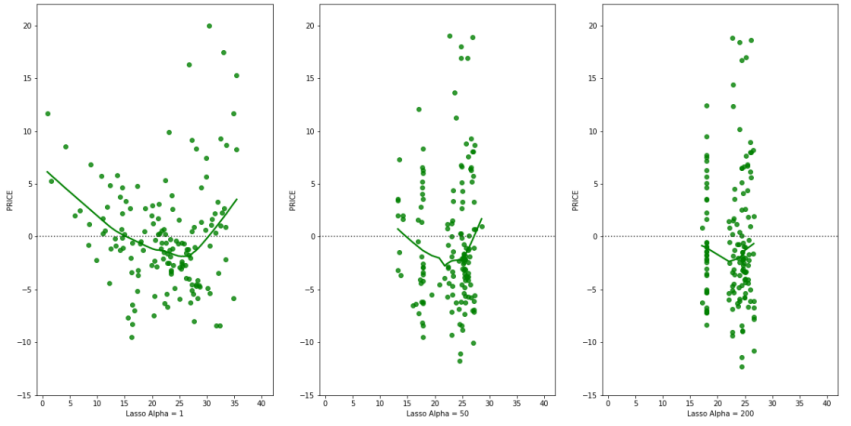
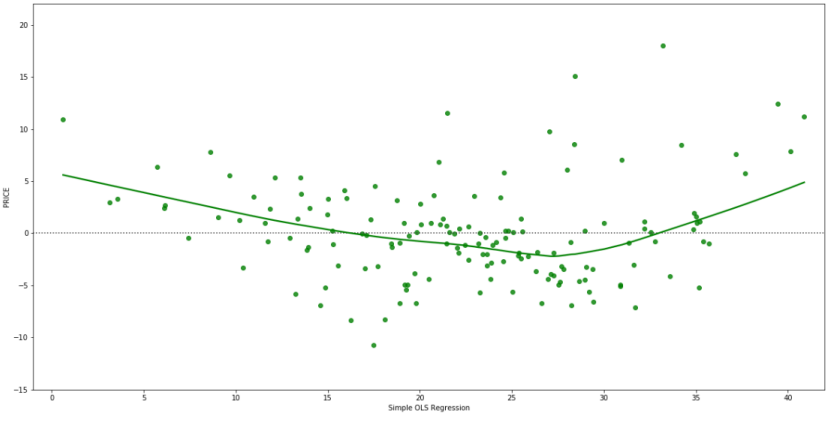


Figure 6: OLS Linear Regression Residual Plot



5 Task 5 - Training and Test Errors

Mean Squared Errors of Training and Test Error and r2 Scores were tabulated and it was found that Simple OlS Linear Regression preformed the best in with respect to Training Error (MSE 22.188 , r2 0.749) while Redge Regression with λ as 1 preformed best With respect to Test Error (MSE 21.935 , r2 0.705) . Lasso Regression with λ as 200 preformed the worst in both Training Error (MSE 72.323 , r2 0.184) and Test Error (MSE 56.239 , r2 0.245)

Figure 7: Training and Test Errors

	Training Error (Mean Squared Error)	R2 Score Training	Test Error (Mean Squared Error)	R2 Score Test
Linear Regression	22.188644	0.749711	22.776030	0.694488
Ridge Regression alpha = 1	22.565231	0.745463	21.935497	0.705763
Ridge Regression alpha = 50	24.250017	0.726458	23.169601	0.689209
Ridge Regression alpha = 200	25.778923	0.709212	24.510473	0.671223
Lasso Regression alpha = 1	27.407455	0.690842	26.100340	0.649897
Lasso Regression alpha = 50	67.635475	0.237068	53.044094	0.288481
Lasso Regression alpha = 200	72.323730	0.184184	56.239243	0.245622

6 Conclusion

With increasing regularization coefficient λ the preformance of both Ridge and Lasso models are decreasing. Thus we obtain the best model as Ridge Regression with λ as 1 having Train Error (MSE 21.935 , r2 0.705) and Training Error (MSE 22.565 , r2 0.705)

7 Learnings

1. We may consider Ridge Regression insted of simple OLS to avoid overfitting.
2. R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
3. Experimenting with Different parameters for Ridge and Lasso Regression might yield a better model than simple OLS linear model.