Name: Kah Meng Soh
Student ID: 5724034
Project: Boston Housing Data with Linear Methods for Regression

Summary

I have applied Ordinary Least Square (OLS), Lasso, Ridge, and Elastic-Net Linear regression on the worldwide famous Boston Housing data, with Mean Squared Error (MSE) as the loss function. As the result, Ridge is the best regression model for this dataset due to having lowest MSE, I also analyzed the distribution of the underlying data to further support the result. And lastly, came up with evidence that challenged the linearty assumption in the data.

Data Background

The dataset[1] contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It's one of the most famous datasets worldwide on data science platform like Kaggle, and have been used extensively throughout the regression studies literature. The dataset contains information about 506 houses, there are 13 predictors with MEDV (Median value of a home) as the response variable which I predicted and modeled on. The dataset is obtainable from one of python data analysis package 'sklearn.datasets'.

Ordinary Least Square (OLS), Lasso, Ridge, and Elastic-Net Methodology[2]

In Statistics, linear regression is a modelling approach where we fit a straight line between the predictors and the response. The assumptions for linear regression as follow: 1) We assume that the mean function of the response variable is a linear combination of the predictors 2) We assume constant variance – that the error is same across all observation and does not depend

[1] http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html
[2] The elements of Statistical Learning, 2nd edition, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

on the predictors value 3) Independence of errors – that the errors uncorrelated with each other's 4) Normality of errors – we assume that the errors are normally distributed with 0 mean and a constant variance as stated above. 5) Independence between predictors – or it would led to multicollinearity. OLS is a linear least square method that minimize the residual sum of square. OLS is considered as the best linear unbiased estimator, we can use Lasso (L1 norm) and Ridge (L2 norm) to sacrifice the unbiased for smaller increase in variance which might improve test accuracy, which is a shrinkage method that add penalized term to the RSS to deal with correlation between predictors with a λ (lambda) as the penalty parameter. When lambda is 0, both Ridge/Lasso will function like an OLS regression, when lambda is increasing, both Ridge/Lasso will shrink the coefficient, when lambda is approaching infinity, ridge will shrink coefficient towards 0 while Lasso will shrink coefficient to exact 0, therefore Lasso can do variable selection and is good in handling sparse model. Elastic-Net is a combination of both Lasso and Ridge utilizing their advantage in dealing with correlation between data and variable selection. The cost function for OLS, Lasso, Ridge, and Elastic Net[3] can be summarized as below:

OLS	$\mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$
Lasso	$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\lvert\beta_j\rvert$

[3] https://cran.r-project.org/web/packages/elasticnet/elasticnet.pdf

Ridge	$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$
Elastic Net	$\lambda\sum_{j=1}^{p}(\alpha\beta_j^2+(1-\alpha)\lvert\beta_j\rvert)$

Model Evaluation

I used MSE, L2 loss function to judge the accuracy of the models, with 80% of the data as training set to build model and 20% of the data as test set where we predicted to get test MSE.

Software Usage

For the data visualization, I utilized Python with the seaborn package to create correlation heat map, boxplot, and distribution plot. For statistical modelling, I used R and its library such as 'glmnet' that is built with Lasso, Ridge, and Elastic-Net regression, while R is already build in with the the OLS model with the function lm.

Exploratory Data Analysis[4]

The dataset has already been preprocessed and leave no missing data. All of the variables are numerical hence there's no need for dummy encoding. From the correlation heat map (Appendix: Figure 1), we see that there's indeed obvious correlation between the predictors ranging between -0.7 to 0.9. For example, TAX and RAD is 0.91, INDUS and DIS is -0.71, DIS and NOX is -0.77, this leave me the impression that there's indeed multicollinearity in the dataset

[4] https://github.com/PrinceRuthless95/Boston-Housing-Data-with-Regression/blob/main/Boston%20Housing%20Data%20Exploratory%20Data%20Analysis.ipynb

where OLS is poor in dealing with, and Lasso/Ridge have advantage over. There's also outlier as seen from boxplot (Appendix: Figure 2), I prefer not to filter them out for the sense of simplicity. From the distribution plot of each variables (Appendix: Figure 3), I observed some sparsity in variables such as CRIM, ZN, hence Lasso might have an edge over Ridge. It's also worth noting that CHAS is the only binary variable which gives value of either 0 or 1, for simplicity I just leave it as a numerical value.

Result[5]

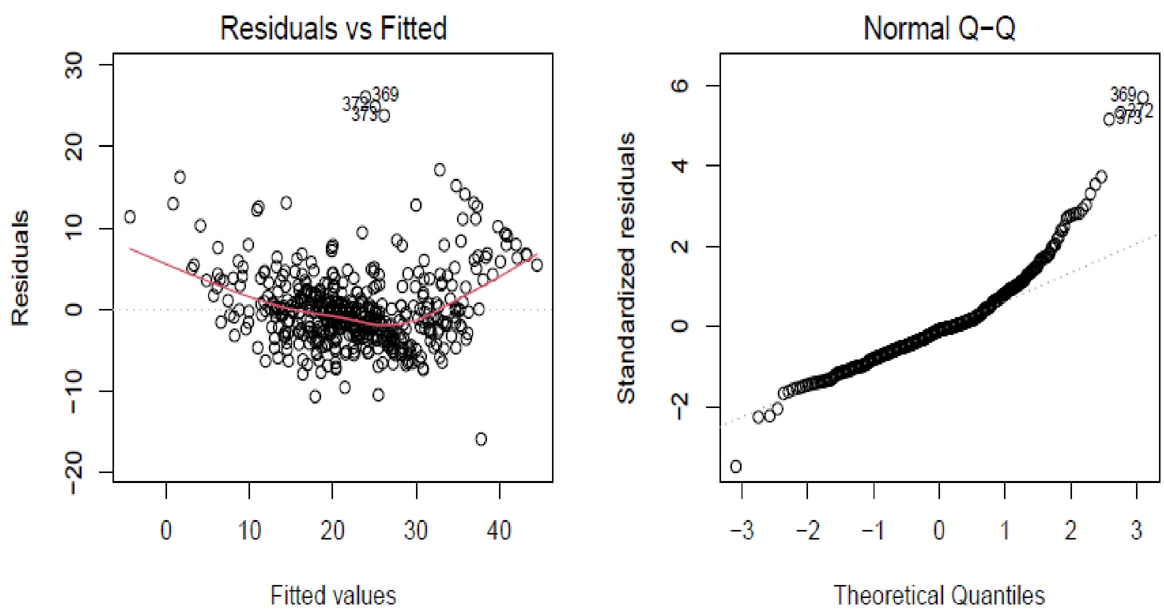
The test MSE comparison between the models are as follows, note that λ for each model is selected by cross validation which give the lowest test MSE.

Linear Model	λ Penalty	α (used in R)	Test MSE	Test RMSE
OLS	0	N/A	17.33601	4.16365
Ridge	0.7014097	0	17.13438	4.13937
Elastic-Net	0.06633903	0.25	17.28993	4.15811
Elastic-Net	0.04812329	0.5	17.30388	4.15979
Elastic-Net	0.03208219	0.75	17.29769	4.15905
Lasso	0.02640763	1	17.30783	4.16026

Based on the result, Ridge is the best model with test MSE of 17.13438, do note that as the α decrease, the test MSE improve as well. Hence the model suggested that Ridge is clearly superior over Lasso, the sparsity in the data doesn't give a lot advantage to Lasso over Ridge.

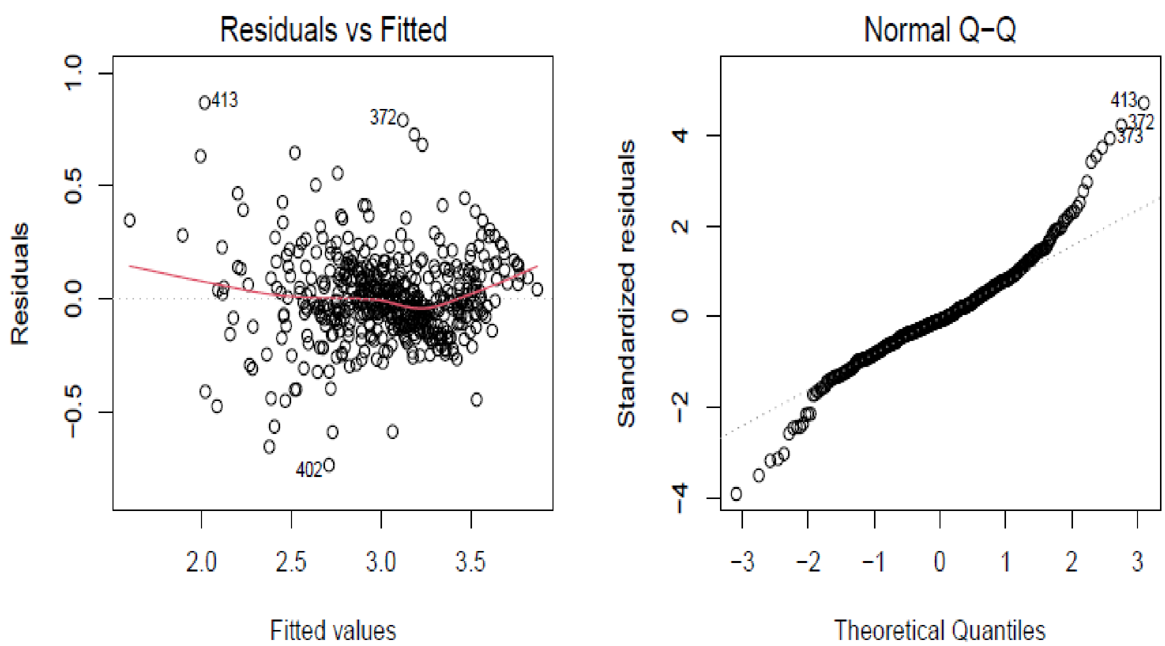
[5] https://github.com/PrinceRuthless95/Boston-Housing-Data-with-Regression/blob/main/Boston-Housing.pdf

We also see that Ridge is better than OLS, this might be due to OLS inability to handle multicollinearity. To challenge the linearity assumption of the data, the test Root Mean Squared Error (RMSE) of OLS, 4.16365 is very high compared to the sample mean of response variable MEDV, 22.532806, which suggested that the result from our model is poor, regardless of which model we selected. This might be the assumption of linear model being violated or the underlying data is not linear. From the residual plot of OLS below:



we see that the errors shaping like a banana with imbalance density. This could be due to the violation of constant variance (Heteroscedasticity) or non-linearity of the underlying data, to further support that, I also carried out a boxcox transformation which suggested log transformation. The transformed residual plot of OLS did not improved like below:

Name: Kah Meng Soh
Student ID: 5724034
Project: Boston Housing Data with Linear Methods for Regression



Hence, I have enough evidence that the poor test MSE is due to the non-linearity of underlying data where linear regression method works poorly. Therefore, we could also try other regression methods like SVM, tree-based, KNN, and non-linear. We could also try L1 loss – Mean Absolute Error as our data consist some outlier as observe from boxplot.

Conclusion

For linear method, Ridge with lambda penalty of 0.7014097 is best to predict the Boston Housing data, the linear model coefficient predicted by Ridge with full datasets are as follow.

```
  (Intercept)          CRIM            ZN         INDUS          CHAS
 27.832905126  -0.087189969   0.032417614  -0.038828587   2.902056950
          NOX            RM           AGE           DIS           RAD
-11.787030343   4.012856001  -0.003811209  -1.109846262   0.151202898
          TAX       PTRATIO             B         LSTAT
 -0.005662260  -0.852615617   0.009063426  -0.470965108
```

Name: Kah Meng Soh
Student ID: 5724034
Project: Boston Housing Data with Linear Methods for Regression

Appendix

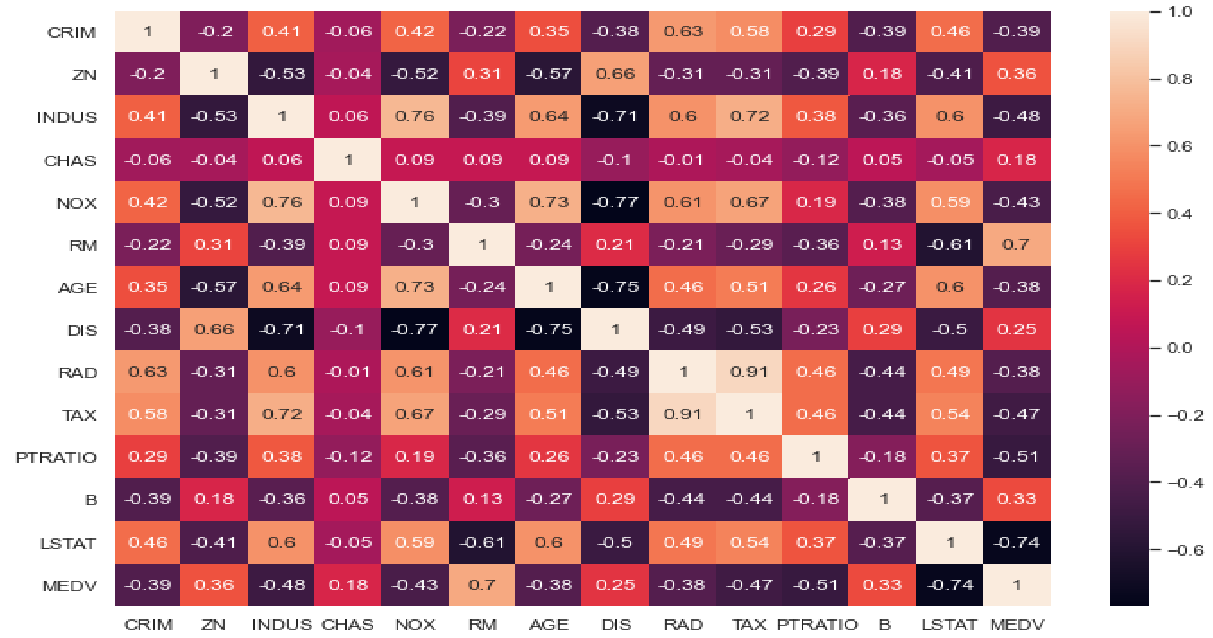


Figure 1: Correlation heatmap for Boston Housing Dataset

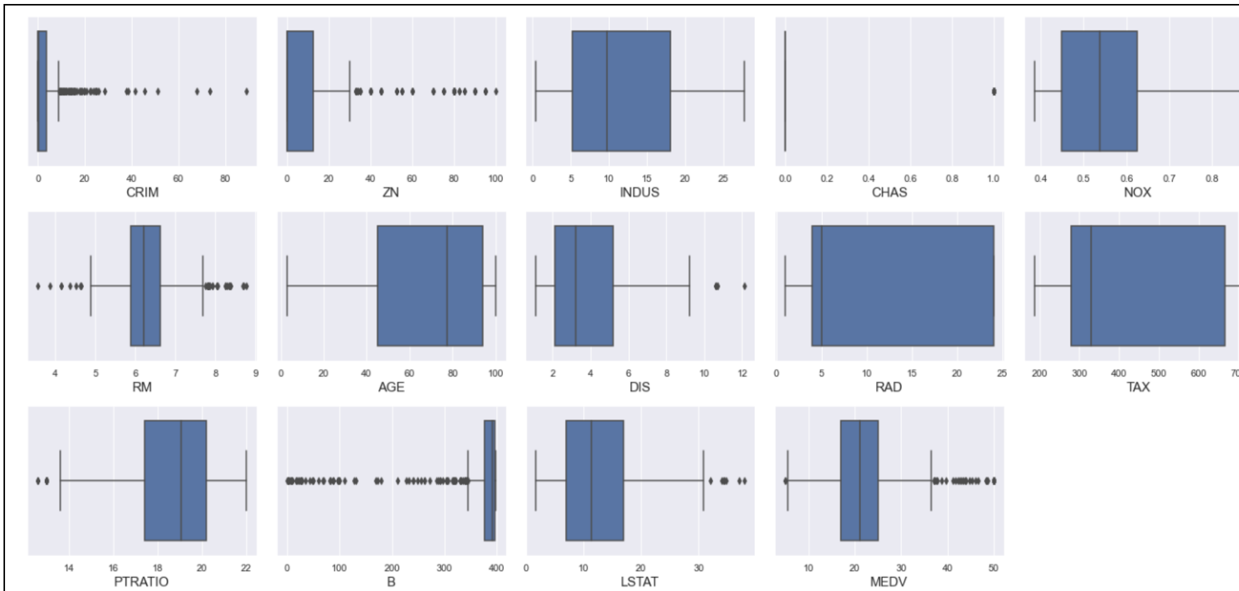


Figure 2: Boxplot for Boston Housing Dataset

Name: Kah Meng Soh
Student ID: 5724034
Project: Boston Housing Data with Linear Methods for Regression

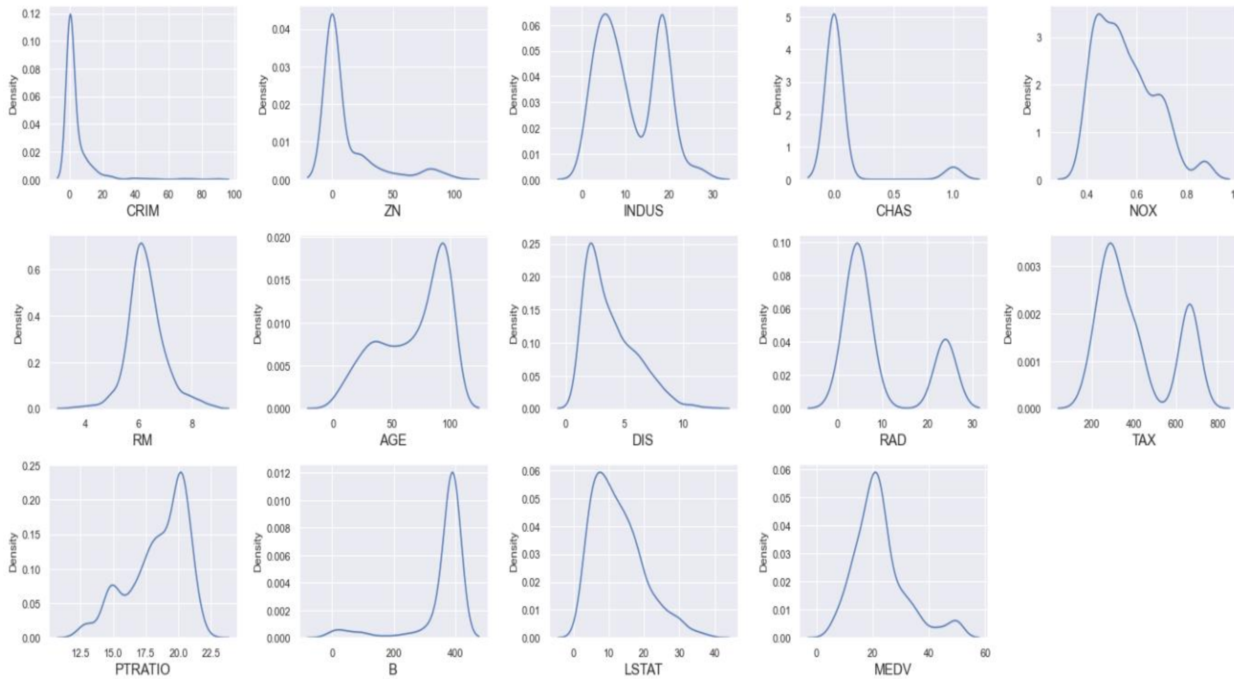


Figure 3: Boxplot for Boston Housing Dataset

Code

https://github.com/PrinceRuthless95/Boston-Housing-Data-with-Regression