Regression - Predicting Violent Crime in a Community

Introduction

Violent crime is a social problem. Social and economics characteristics of a community may be correlated to the level of violent crime in a population. A key to understanding why it happens is understanding where it happens. The aim of this problem is to explore the link between the various socio-economic factors and crime.

Dataset Description

The dataset contains socio-economic data from the 1990 US census for various US communities, and the number of violent crimes per capita (in the column `ViolentCrimesPerPop`). According to the dataset information file `Communities_Info.txt`, the dataset consists of a large number of variables, mainly group into two categories, socio-economic (Race, age, employment, marital status, immigration and home ownership), and Law Enforcement Management and Administrative Statistics (LEMAS) data.

Also, it is important to note that, as mentioned in the info file, all numeric data has been normalised into a range of 0.00-1.00 which preserves the ratios of values within an attribute. Outliers, as defined as 3SD above or below are also allocated to the upper end and lower end of the scale, accordingly. However, because of this normalisation, comparing between values of different attributes values are no longer meaningful. Thus, it is fair to assume that there is no multicollinearity occurred in the dataset.

Data Preparation

The dataset is loaded from a csv file `Q1/communities.csv`. The dataset contains 128 attributes and 1994 observations. The predictor variable is `ViolentCrimesPerPop`, which is a numerical variable. As mentioned, the dataset's numerical data is normalised.

Categorical Variable

First five categorical variables are removed from the dataset (`state`, `country`, `community`, `communityname string`, `fold`). Here we are interested using continuous variables to fit the models.

Missing Values

There are 23 variables containing null values which are mostly LEMAS data. Due to high percentage of missing values (>80%), 22 variables are removed from the dataset.

For variable `OtherPerCap`, which contains about 5% of missing data, we shall impute these missing data with the median of the variable.

Non-Social-Economic Attributes

Lastly, because the question is to explore the link between the various socio-economic factors and crime, the rest of the non-socio-economic attributes (namely `LandArea`, `PctUsePubTrans` and `LemasPctOfficDrugUn`) are also removed.

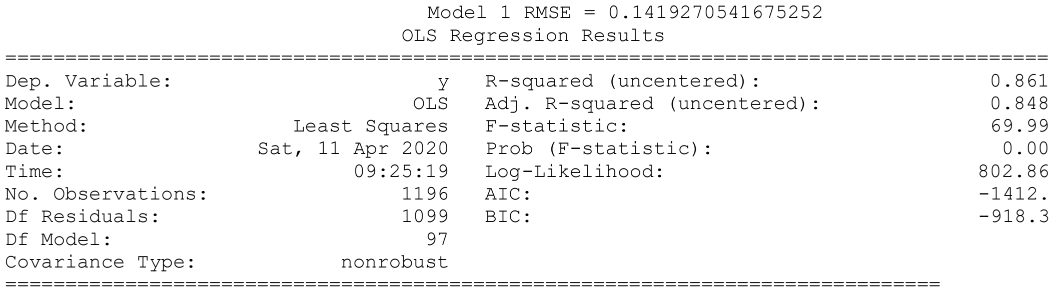
There are 98 remaining variables in the dataset.

Splitting Data

The dataset is split into training set (60%), validation set (20%) and testing set (20%). This should give us considerable amount of data to train the model and capture the variabilities sufficiently in both validation and testing sets.

Linear Regression

Linear regression is fitted as our first model. We fit a full model including all the variables. The summary of the model is included in Figure 1.1.



```
                               Model 1 RMSE = 0.1419270541675252
                            OLS Regression Results                                
=======================================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.861
Model:                            OLS   Adj. R-squared (uncentered):              0.848
Method:                 Least Squares   F-statistic:                              69.99
Date:                Sat, 11 Apr 2020   Prob (F-statistic):                        0.00
Time:                        09:25:19   Log-Likelihood:                          802.86
No. Observations:                1196   AIC:                                     -1412.
Df Residuals:                    1099   BIC:                                     -918.3
Df Model:                          97                                                  
Covariance Type:            nonrobust                                                  
==============================================================================
```

Figure 1.1: Linear Regression Model Result Summary

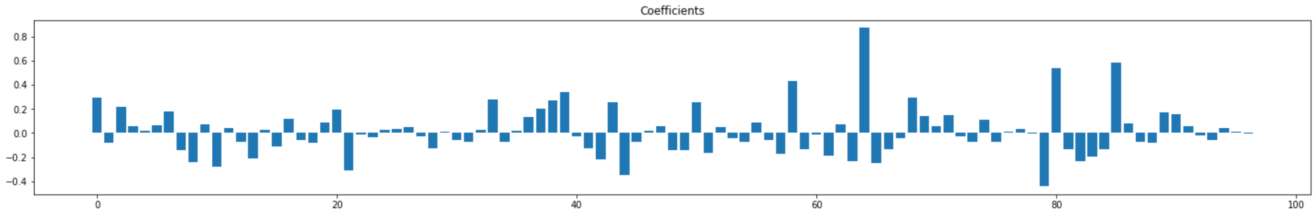


Figure 1.2: Coefficients of Linear Regression Model

With values of R^2 = 0.861 and adjusted R^2 = 0.848, the model is considered good in explaining a large amount of variations of the response variable. The range of coefficients is also considered reasonable which suggests the model is not overfitted.

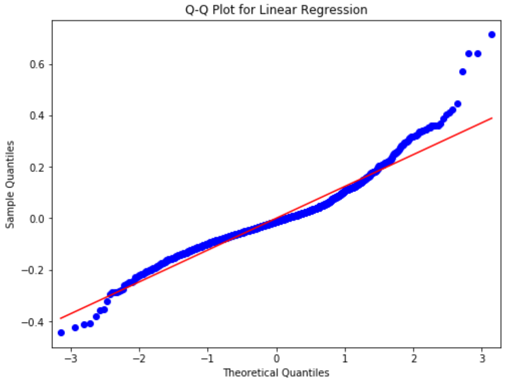


Figure 1.3: Q-Q Plot of Standardised Residuals

Lastly, from Quantile-Quantile plots for the residuals, most points are reasonably close to the line. However, there are some deviation near the ends, which *may* suggest the residuals have heavier tails than a normal distribution.

Prediction Performance (Linear Regression)

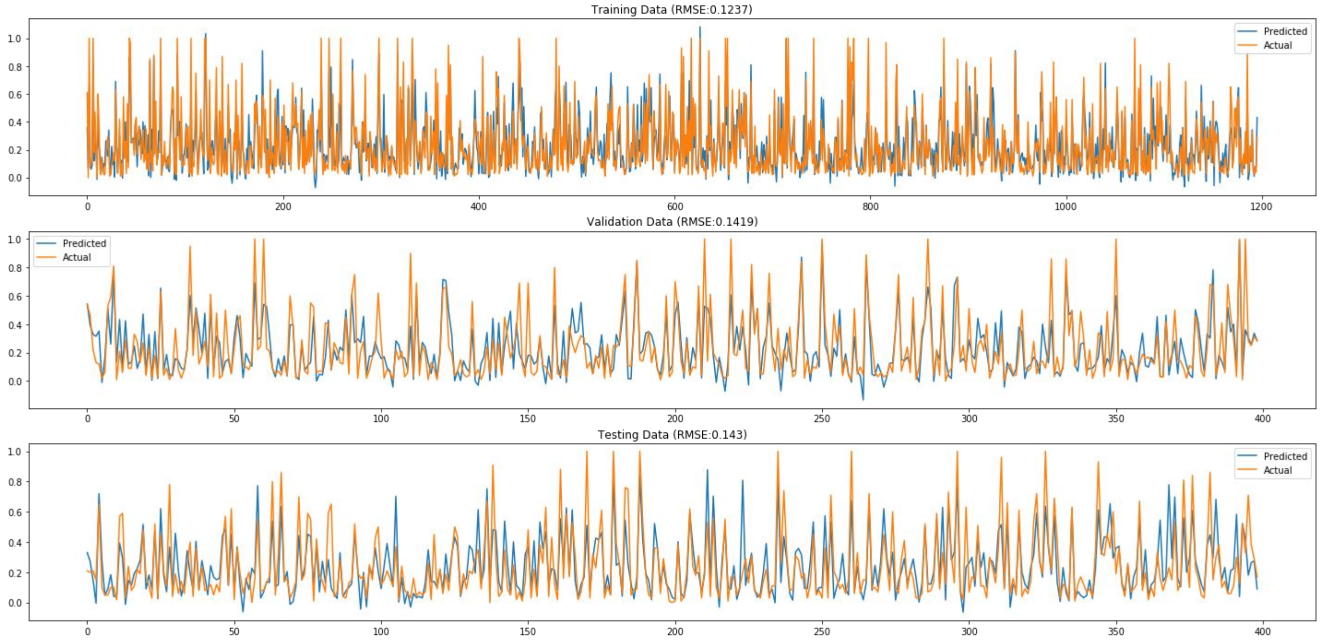


Figure 1.4: Prediction – Linear Regression Model

The linear regression model's prediction performance is very good on the training set (RMSE: 0.1237), and reasonably good on both validation set (RMSE:0.1419) and test set (RMSE:0.143). However, it is important to note that this is created from normalised data and it needs to be compared with the results from other models.

LASSO Regression

Lasso shrinks the coefficient estimates towards zero. It penalises some of the coefficient estimates to be zero when the tuning parameter lambda is sufficiently large. Thus, the lasso performs variable selection.

For a start, these are the three λ values we selected:

- a very small value, λ=0.01
- an intermediate value, λ=0.1
- a large value, λ=0.5

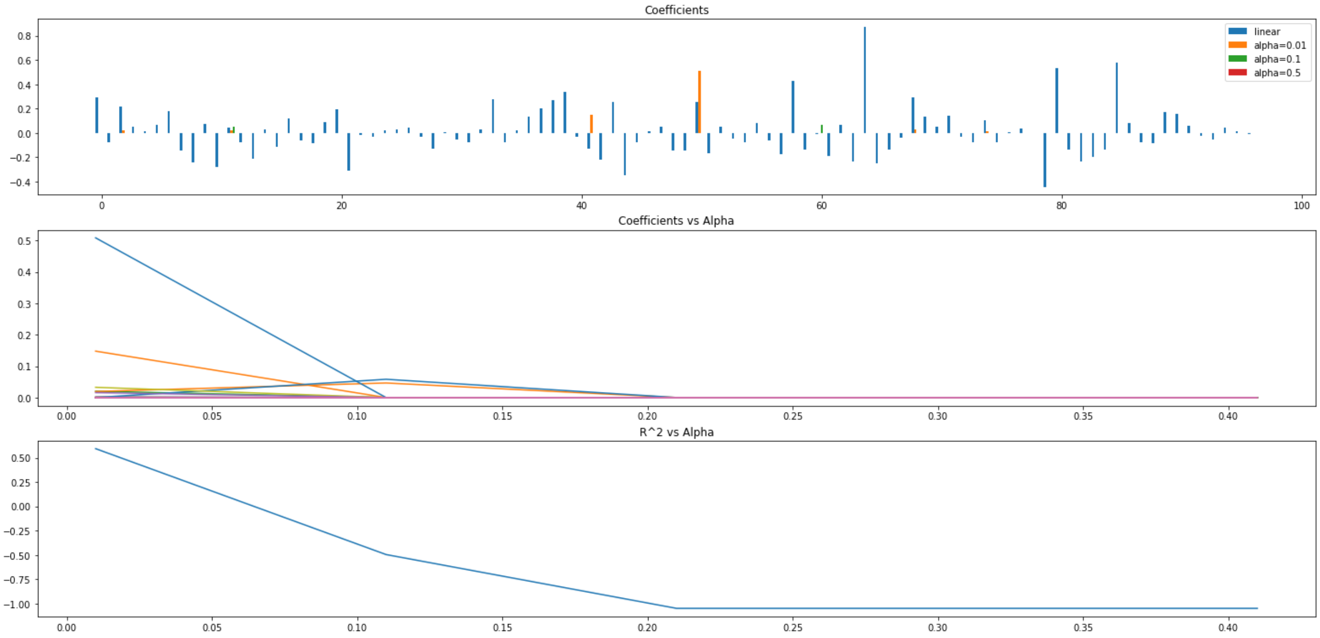


Figure 1.5: Coefficients and R^2 vs Alpha – Lasso Regression Model

From Figure 1.5, we can see that, as λ increases, R^2 decreases quickly. Most of the coefficients approach 0 after λ=0.1.

Prediction Performance (Lasso Regression)

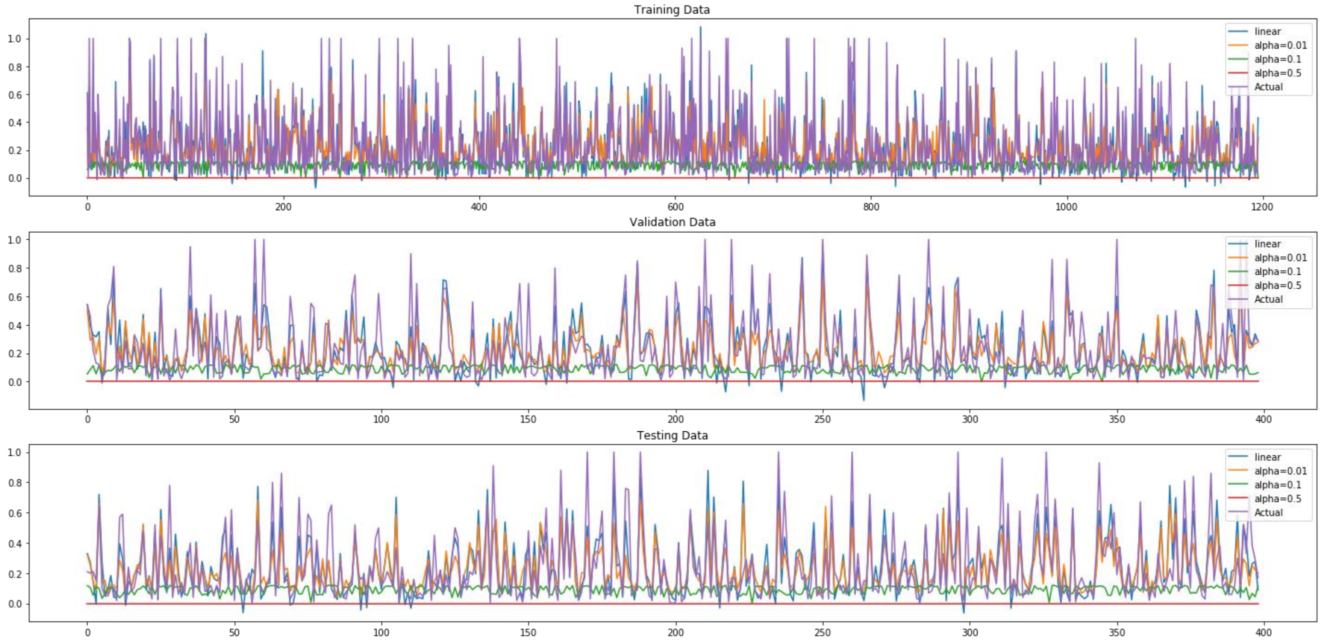


Figure 1.6: Prediction – Lasso Model (λ=0.01, 0.1, 0.5)

Clearly, λ=0.01 has the best prediction performance among the alphas and is comparable to the linear model. Next, we shall select the best value of λ from validation set (λ=0.01) and check its prediction performance closer.

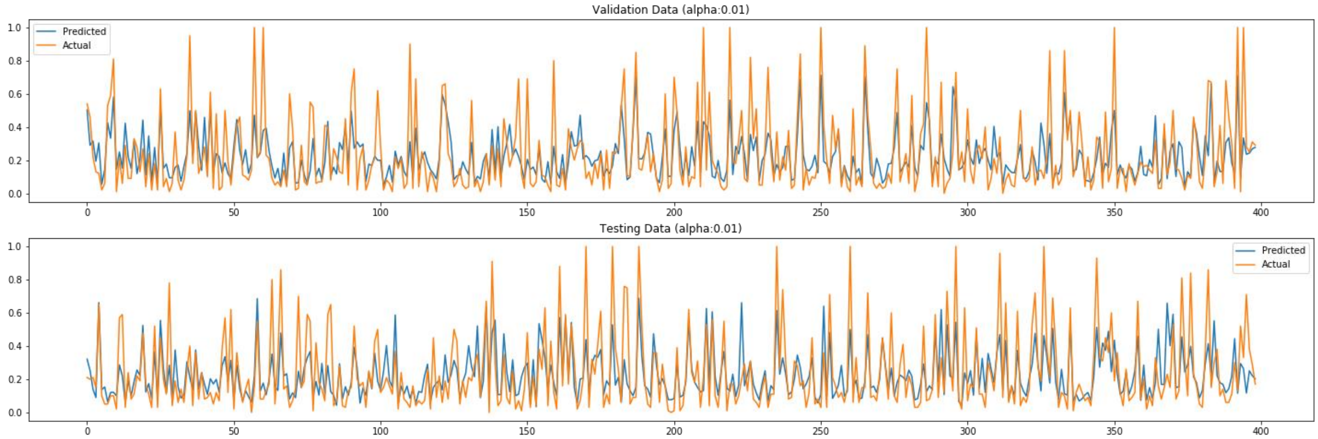


Figure 1.7: Prediction – Lasso Model (λ=0.01)

Ridge Regression

Ridge Regression uses a tuning parameter, λ, to the model to shrink the coefficients. It reduces the values of the less significant coefficients to zero.

These are the three λ values we selected:

- a very small value, λ=0.01
- an intermediate value, λ=2.5
- a large value, λ=10

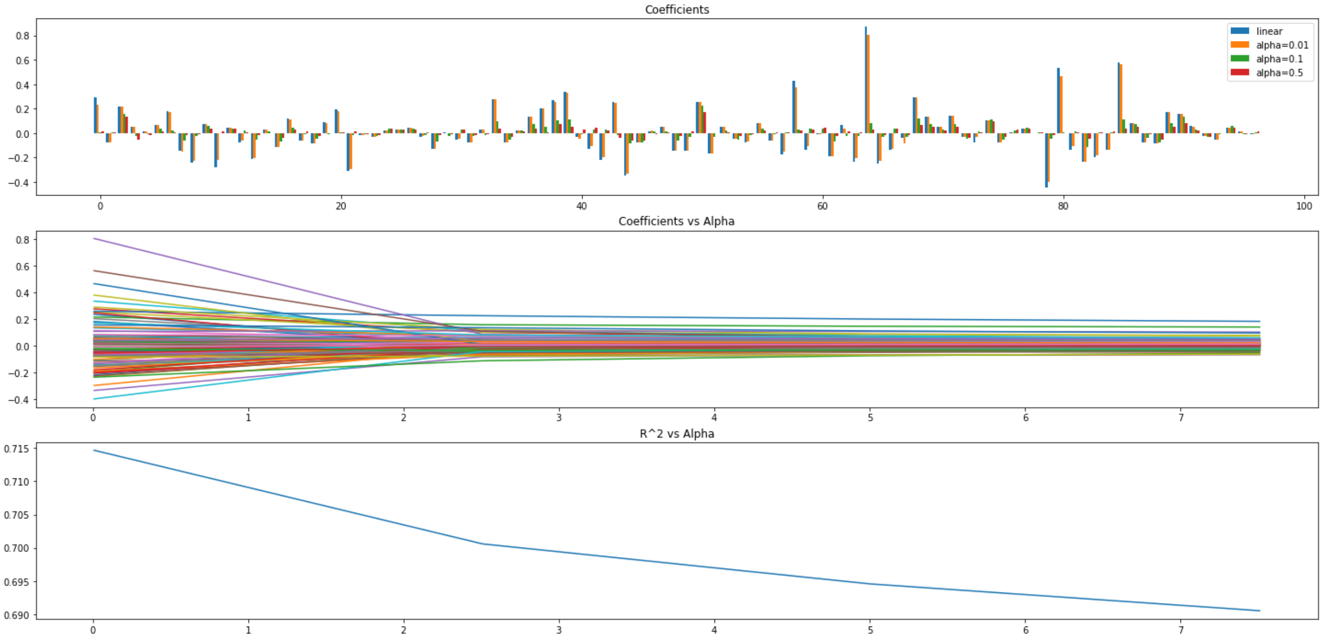


Figure 1.8: Coefficients and R^2 vs Alpha – Ridge Regression Model

From Figure 1.8, similar to Lasso, larger λ values yield lower weights. However, unlike Lasso, which penalise heavily at the start, Ridge has a more gradual drop.

Prediction Performance (Ridge Regression)

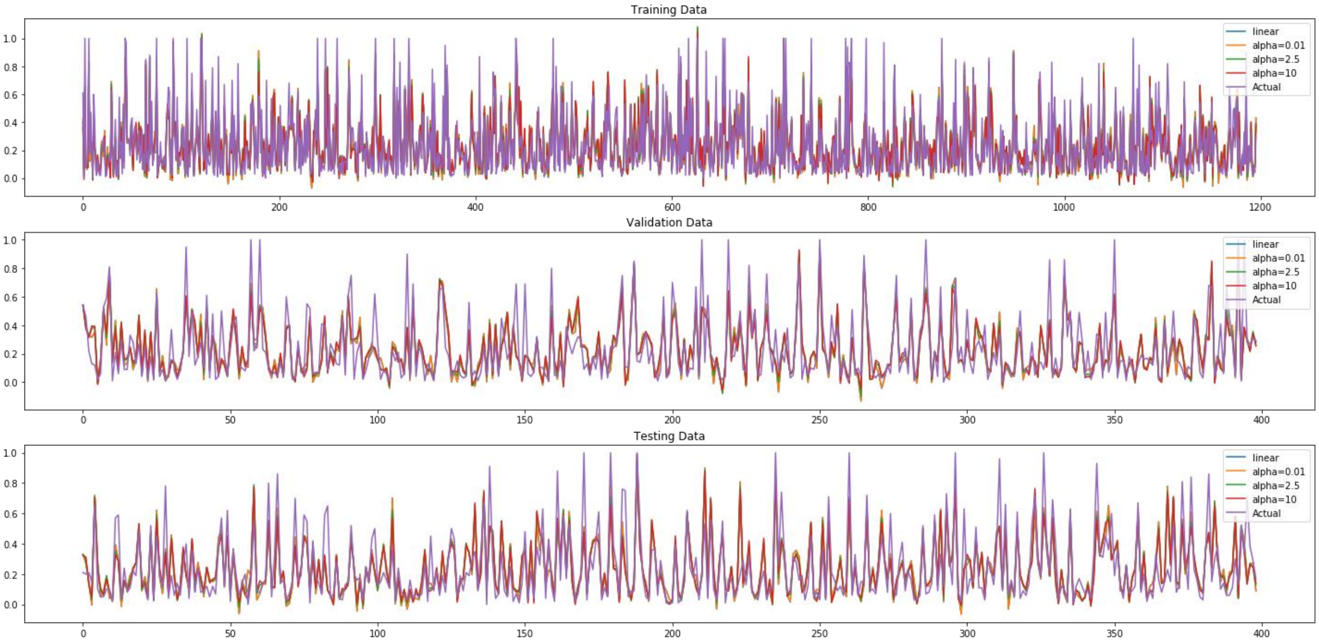


Figure 1.9: Prediction – Ridge Model (λ=0.01, 2.5, 10)

All three λs prediction performance is very close to each other. Now we shall select the best value of λ (λ =2.51) from validation set and check its prediction performance.

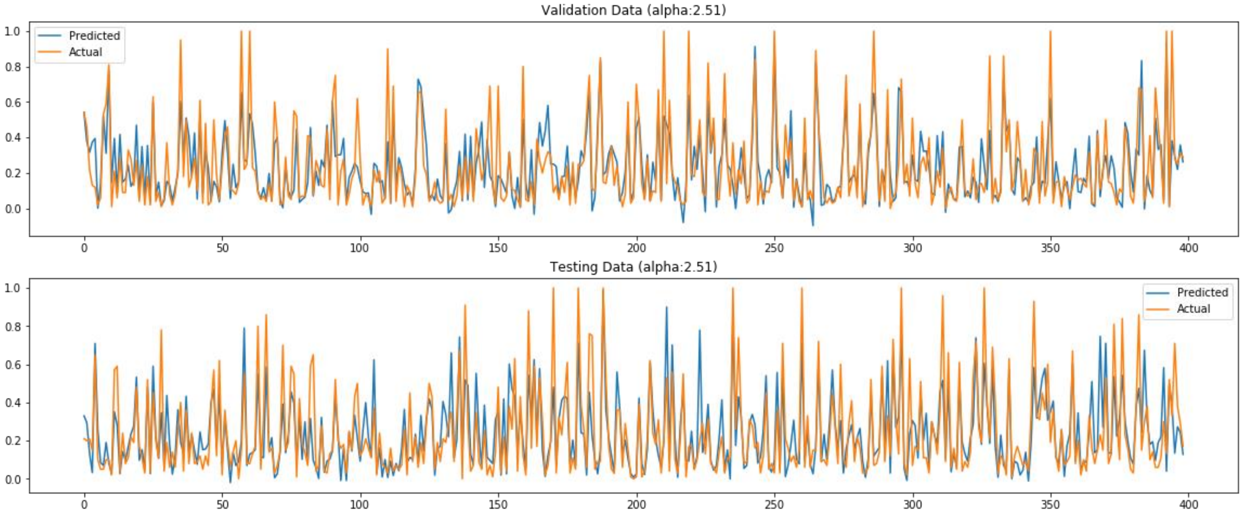


Figure 1.10: Prediction – Lasso Model (λ=2.51)

Model Evaluation

The performance of the models on testing set is summarized below:

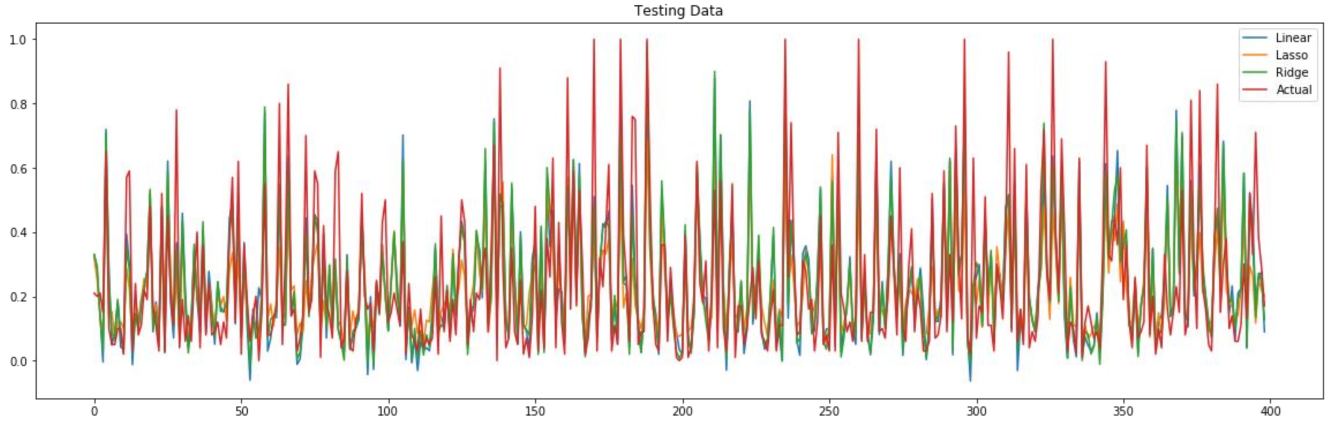
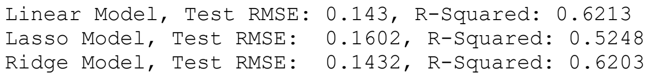


Figure 1.10: Prediction – Linear, Lasso and Ridge Model vs Acutal



```
Linear Model, Test RMSE: 0.143, R-Squared: 0.6213
Lasso Model, Test RMSE:  0.1602, R-Squared: 0.5248
Ridge Model, Test RMSE:  0.1432, R-Squared: 0.6203
```

Figure 1.11: RMSE and R^2 on Test Data – Linear, Lasso and Ridge Model

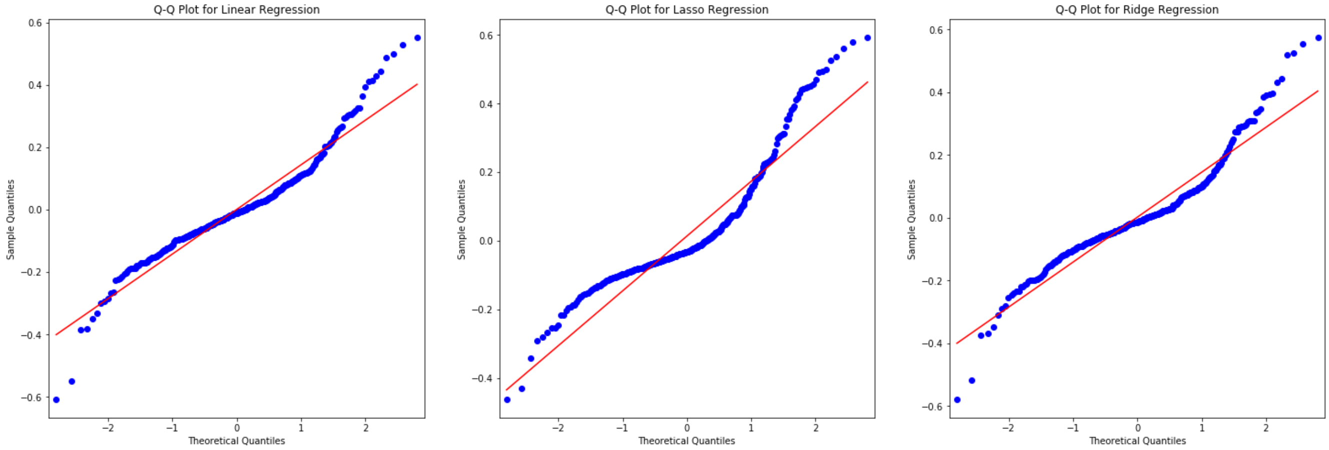


Figure 1.12: Q-Q plot on Residuals for Test Data – Linear, Lasso and Ridge Model

The performance of models is evaluated using two metrics – R-squared value and Root Mean Squared Error (RMSE). Ideally, lower RMSE and higher R-squared values are indicative of a good model.

As the values of R^2 suggest, both ridge and linear models fitted are considered to have a moderate predictive power. In fact, both models have almost identical R^2 and RMSE values. But Lasso is clearly far worse than the two.

Nevertheless, all models' Q-Q plots for the residuals for test data are considered quite bad, which may suggest further work can be done to improve model accuracy and validity.