Boston Housing Dataset

4)

* Linear Regression

In linear regression, we model the dataset on all the features and predict the target (MEDV).

Also, we perform a 10-fold cross validation, which we split the data randomly into 10 parts and each time take 90% of the data for training and intentionally regard the other 10% to have an unknown response variable for testing. We obtain the following parameters from the model with randomly-chosen dataset.

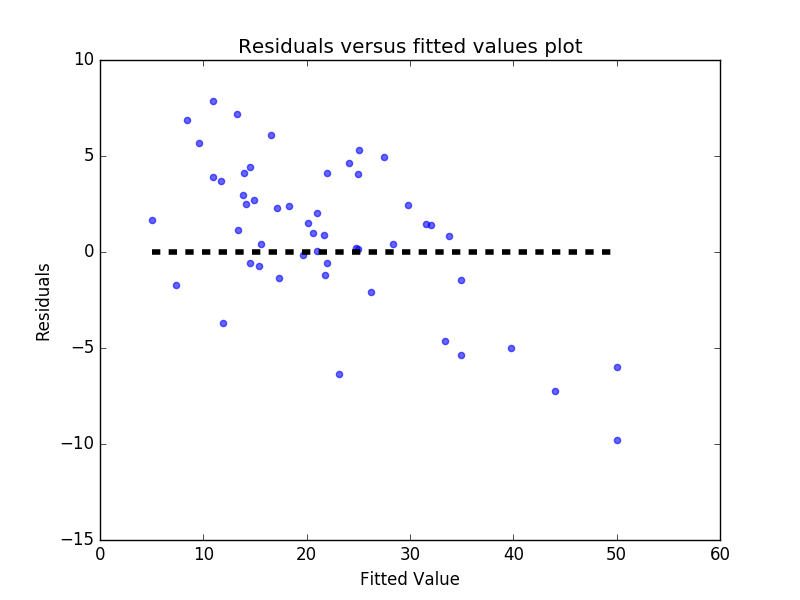
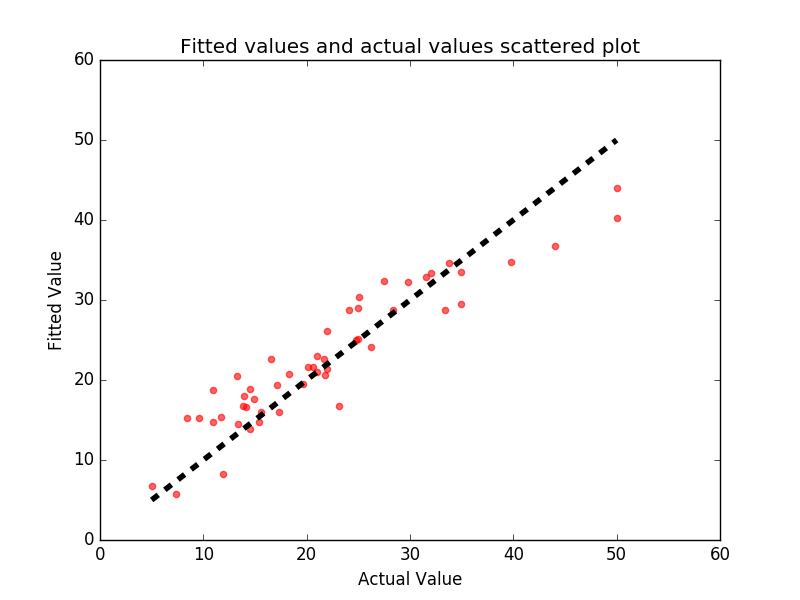
**Best RMSE: 3.926**

**Best r2\_score: 0.834**

**Optimal coefficient vector: [-0.109, 0.035, 0.009, 2.834, -18.359, 3.870, 0.000, -1.412, 0.298, -1.058, -0.965, 0.001, -0.528]**

Moreover, we analyze the significance of different variables with statistics obtained from the model I have already trained, and it returns with **p-value: 0.0099**

Finally, we evaluate how well the model fits the data by providing “Fitted values and actual values scattered plot over time”, and “residential vs fitted values plot”. For simplicity, we only plot 10% of the entire dataset, which is only the test data after 10-fold split.



* Polynomial Regression

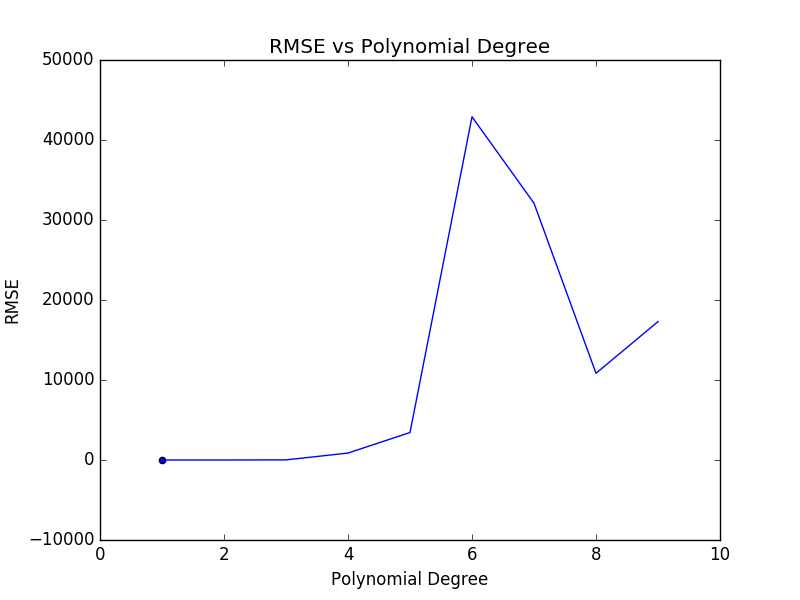
Similarly, we re-model the data into polynomial regression under 10-fold cross validation. We try increasing the degree of the polynomial to improve the fit. We cover the degree from 1 to 9, and plot the RMSE of the trained model against the degree of the polynomial respectively.

The best polynomial regression in our case is with degree 1 to 4. According to the curve below, there is a significant increase in RMSE. Hence, we can safely set the threshold to degree of 4.

**Best degree: 1**

**Best RMSE: 1.446**

**Best r2\_score: 0.855**



5)

* Ridge Regression

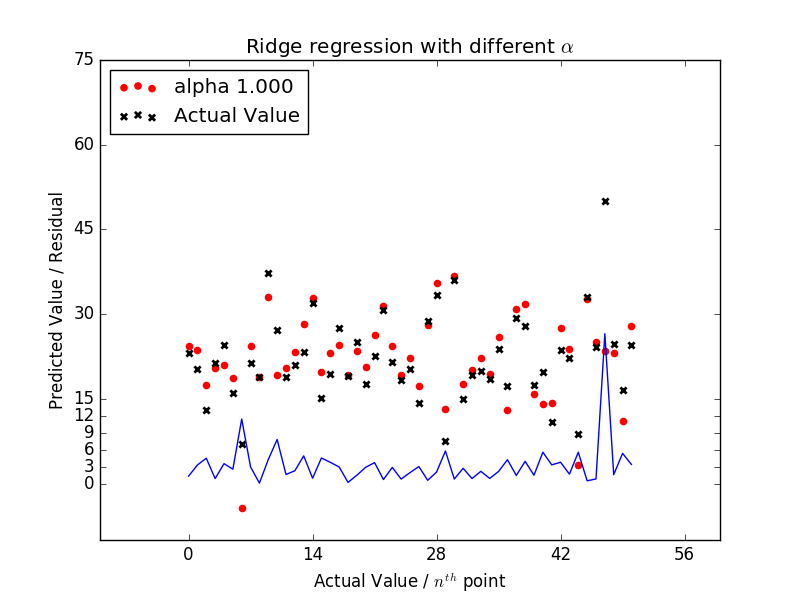
We tune the complexity parameter α of the ridge regression below in the range {1, 0.1, 0.01, 0.001} and report best RMSE via 10-fold cross validation.

We obtain the following parameters through 10-fold cross validation, and plot the figures as required only in the test case, (10% of the dataset for simplicity).

**Optimal α: 1**

**Best RMSE: 3.442**

**Best r2\_score: 0.835**



* Lasso Regression

In this scenario, we model the data with Lasso regularization, using an appropriate normalization for the range of α.

We obtain the following parameters through 10-fold cross validation, and plot the figures as required only in the test case, (10% of the dataset for simplicity).

**Optimal α: 0.001**

**Best RMSE: 5.079**

**Best r2\_score: 0.787**

