Boston Housing Dataset

4)

* Linear Regression

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

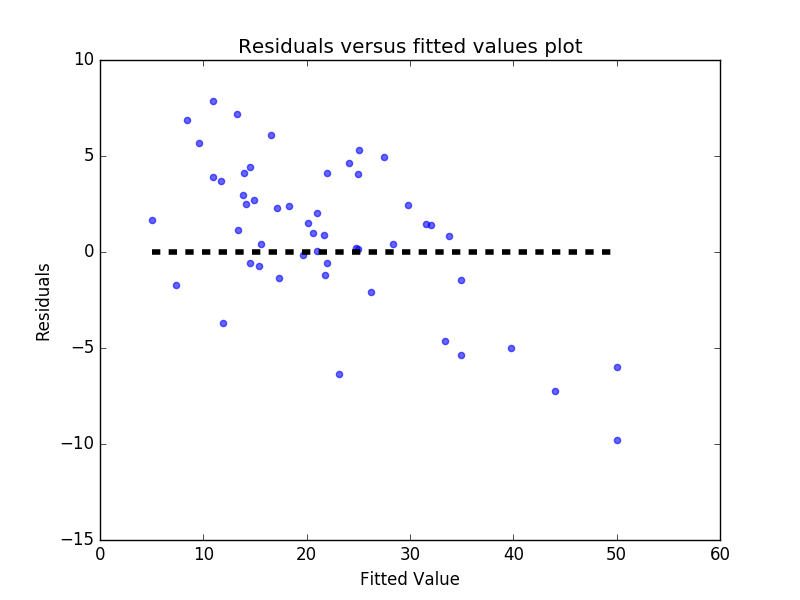
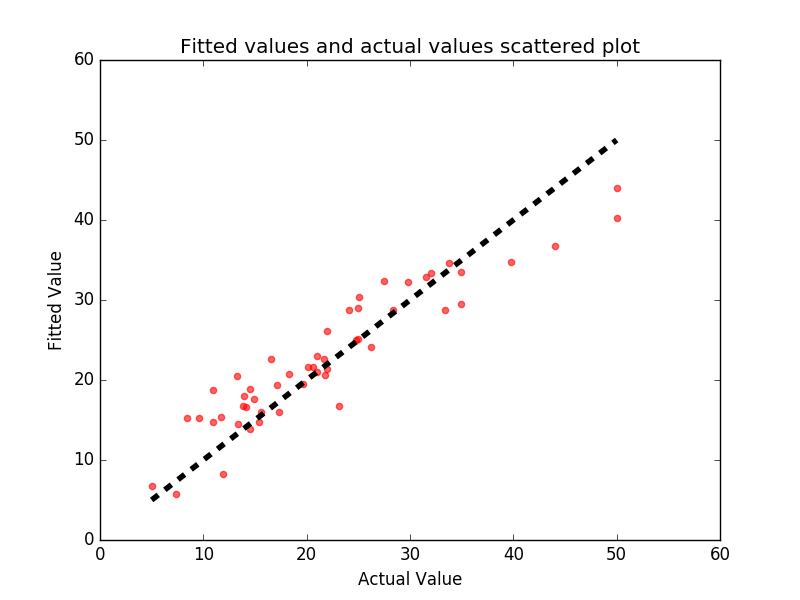
Also, we perform a 10-fold cross validation, which we split the data randomly into 10 parts and each time take 90% of the data for training and intentionally regard the other 10% to have an unknown response variable for testing. We obtain the following parameters from the model with randomly-chosen dataset.

**Best RMSE**: 3.926

**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

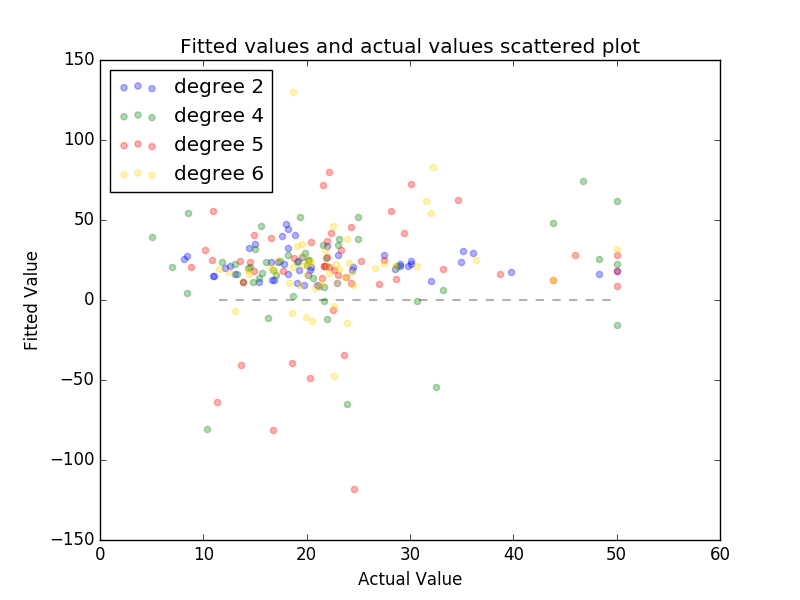
Finally, we evaluate how well the model fits the data by providing “Fitted values and actual values scattered plot over time”, and “residential vs fitted values plot”.



* 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 2 to 4, and plot the RMSE of the trained model against the degree of the polynomial respectively.

Since, polynomial regression of degree three has a much larger RMSE which can ruin the entire plot, we manually get rid of it and we summarize it into a table.



|  |  |
| --- | --- |
| degree | RMSE |
| 2 | 3.02 |
| 4 | 29.56 |
| 5 | 35.51 |
| 6 | 27.49 |

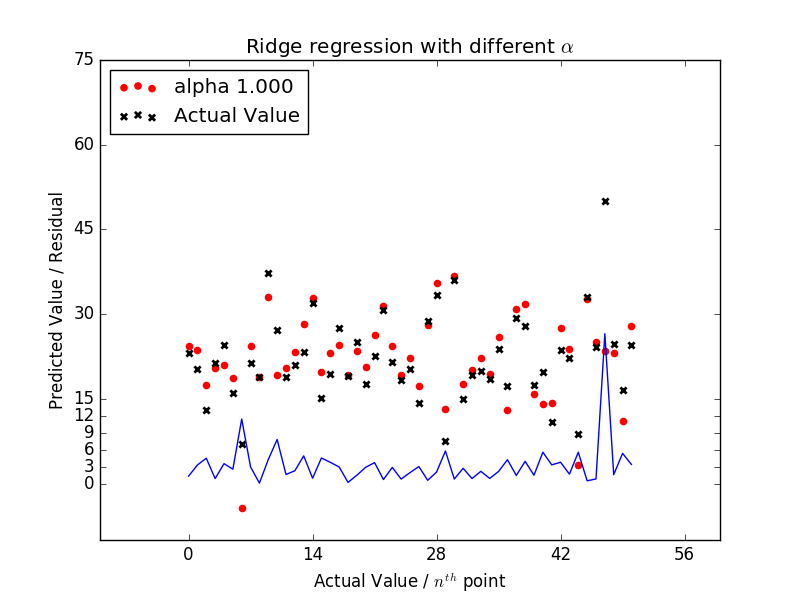
5)

* Ridge Regression

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

**Best RMSE**: 5.092

**Optimal**α:1



* Lasso Regression

**Best RMSE**: 5.079

**Optimal**α:0.001

