## 4）Linear Regression

Similar as the linear part of Chapter 3, we also implemented 10-fold cross validation here. And the RMSEs are shown in Figure 1, and the mean of them is about 4.7403.

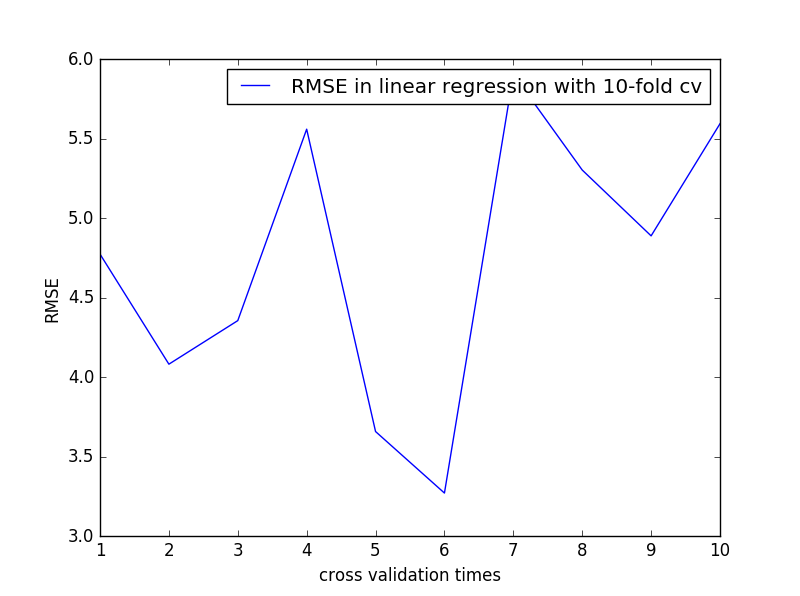


Figure RMSEs in linear regression with 10-fold cross validation

To evaluate how well our model fit the data, we provide two graphs. The first one is shown as Figure 2, which is the scattered plot of actual and fitted MEDV values. And we can see that most of them overlaps. The other graph is the residuals versus fitted values, showed in Figure 3. Since most of them are clustered around 0 of y-axis, we are confirmed that our model fits the data quite well.

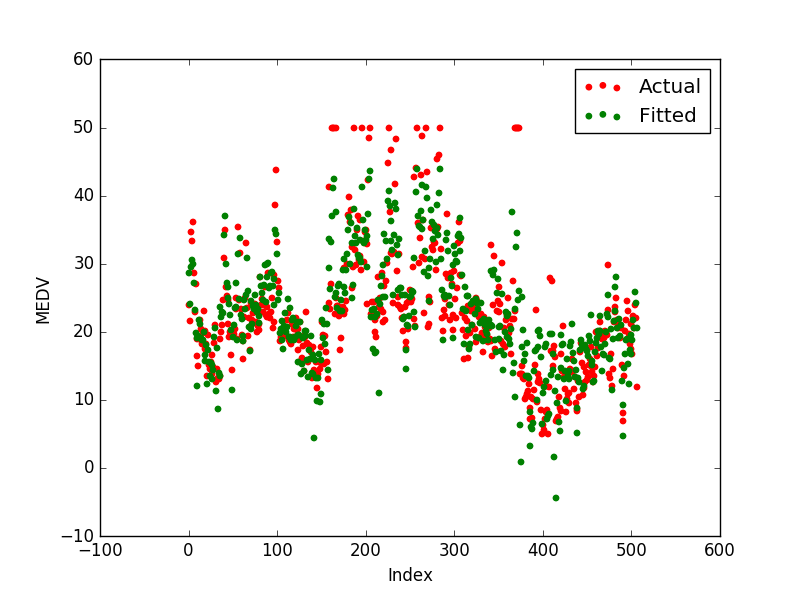


Figure Fitted values and actual values scattered plot

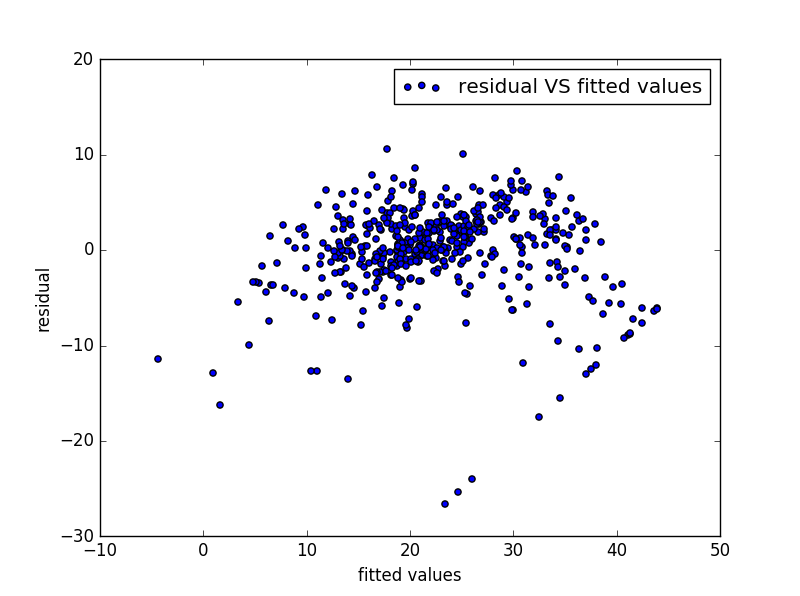


Figure Residuals versus fitted values

For the analyzation of significance of different features, we still use the p-value as before. P-value of the 13 features are shown in Figure 2, while those of 21 features are shown in Figure 3. From them, we know that the three smallest p-values are the sixth, the third last and the last, which means that the corresponding features are the most significant ones. Particularly, they are the RM, PTRATIO and LSTAT. Later, when we were using the Lasso regression model and the α parameter is set as 0.1, we noticed that the only regression coefficients remain nonzero are exactly corresponding to those mentioned above, which means only the most important features are kept.

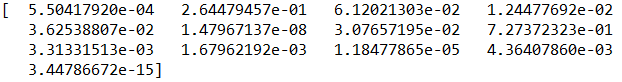


Figure p-value of 13 features

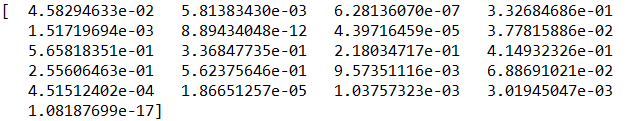


Figure p-value of 21 features