Jingxia Zhu(jingxia6)

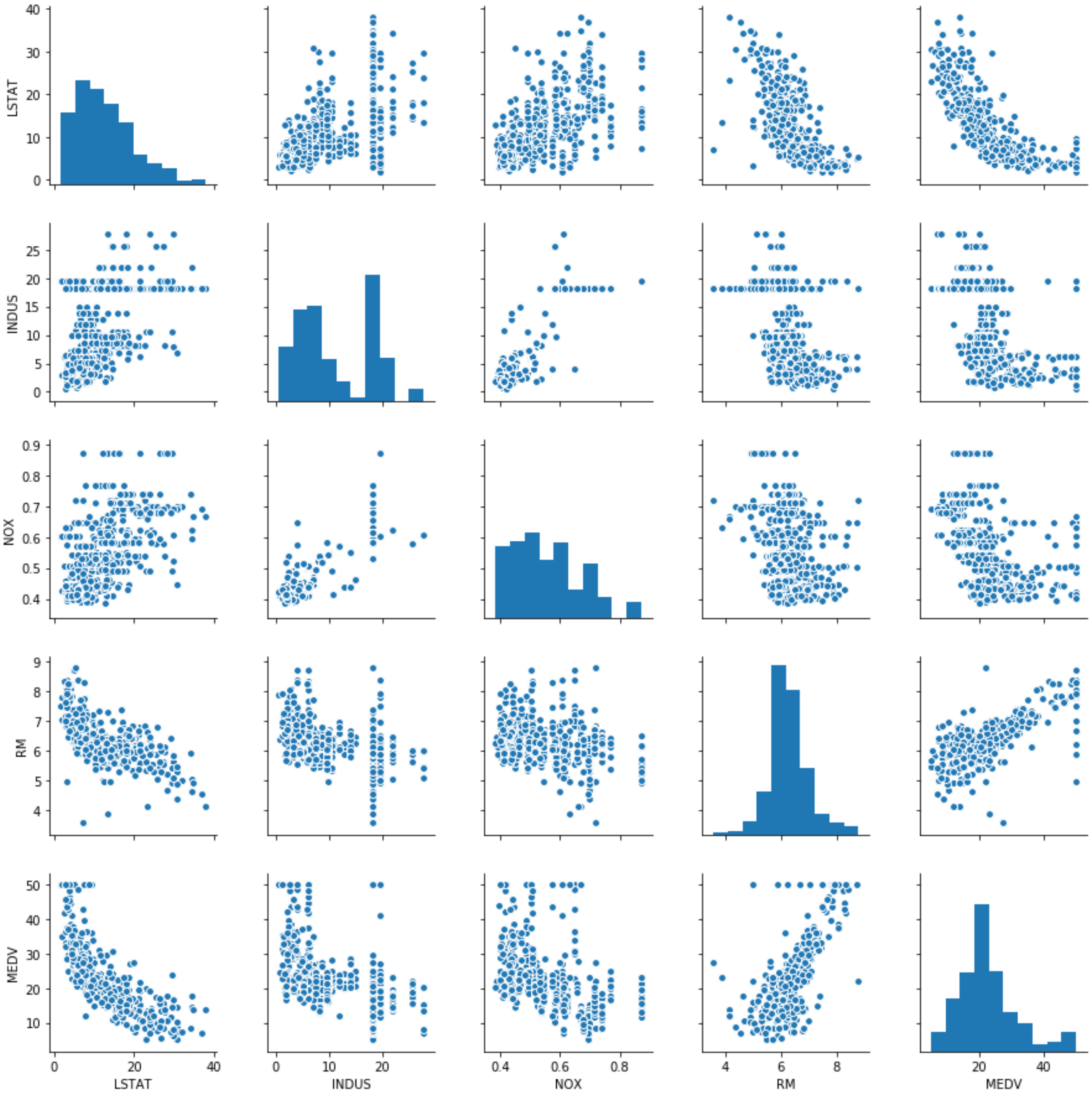
IE598 MLF F18

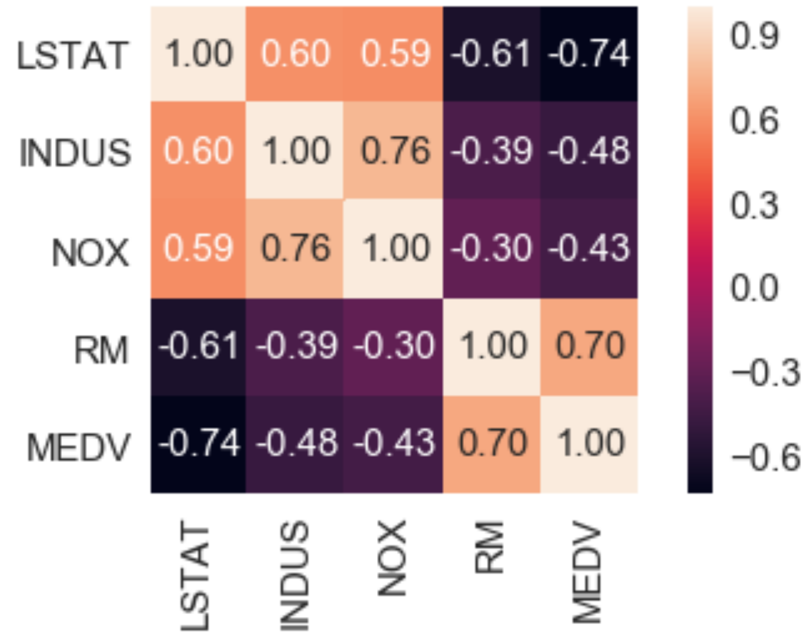
Module 4 Homework (Regression)

Part 1: Exploratory Data Analysis

Describe the data sufficiently using the methods and visualizations that we used previously in Module 3 and again this week. Include any output, graphs, tables, heatmaps, box plots, etc. Label your figures and axes. DO NOT INCLUDE CODE!

Split data into training and test sets. Use random_state = 42. Use 80% of the data for the training set. Use the same split for all models.





	LSTAT	INDUS	NOX	RM	MEDV
LSTAT	1.00	0.60	0.59	-0.61	-0.74
INDUS	0.60	1.00	0.76	-0.39	-0.48
NOX	0.59	0.76	1.00	-0.30	-0.43
RM	-0.61	-0.39	-0.30	1.00	0.70
MEDV	-0.74	-0.48	-0.43	0.70	1.00

Part 2: Linear regression

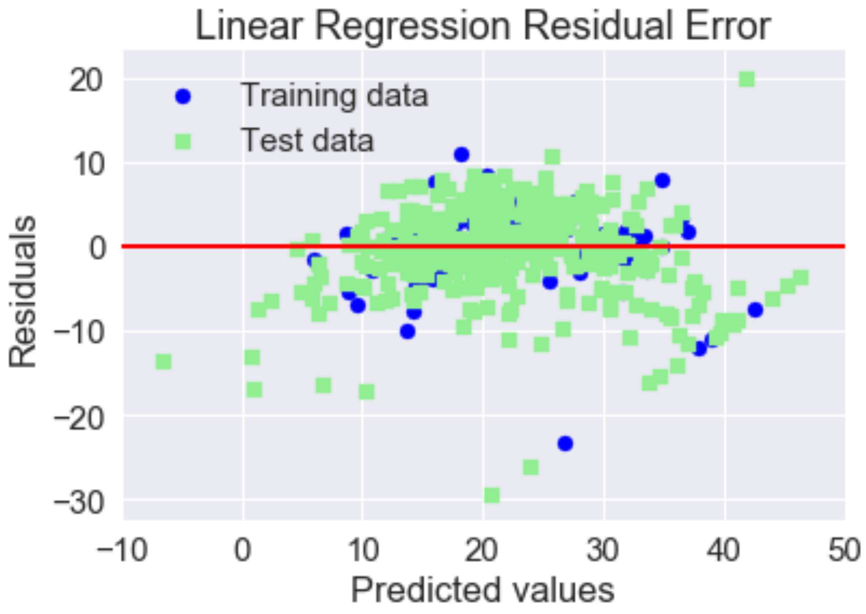
Fit a linear model using SKlearn to all of the features of the dataset. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2.

Coefficients: 9.102

Y interpret: -34.671

MSE train: 19.537, test: 25.564

R^2 train: 0.728, test: 0.708



Part 3.1: Ridge regression

Fit a Ridge model using SKlearn to all of the features of the dataset. Test several settings for alpha. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2. Which alpha gives the best performing model?

Slope: [-1.74690807e-01 1.95983720e-02 1.46574847e-01 4.76557735e+00

-1.95086828e+01 5.01603617e+00 -3.13010289e-02 -1.32551038e+00

1.50074687e-01 1.77067611e-04 -1.11013773e+00 1.19435641e-02

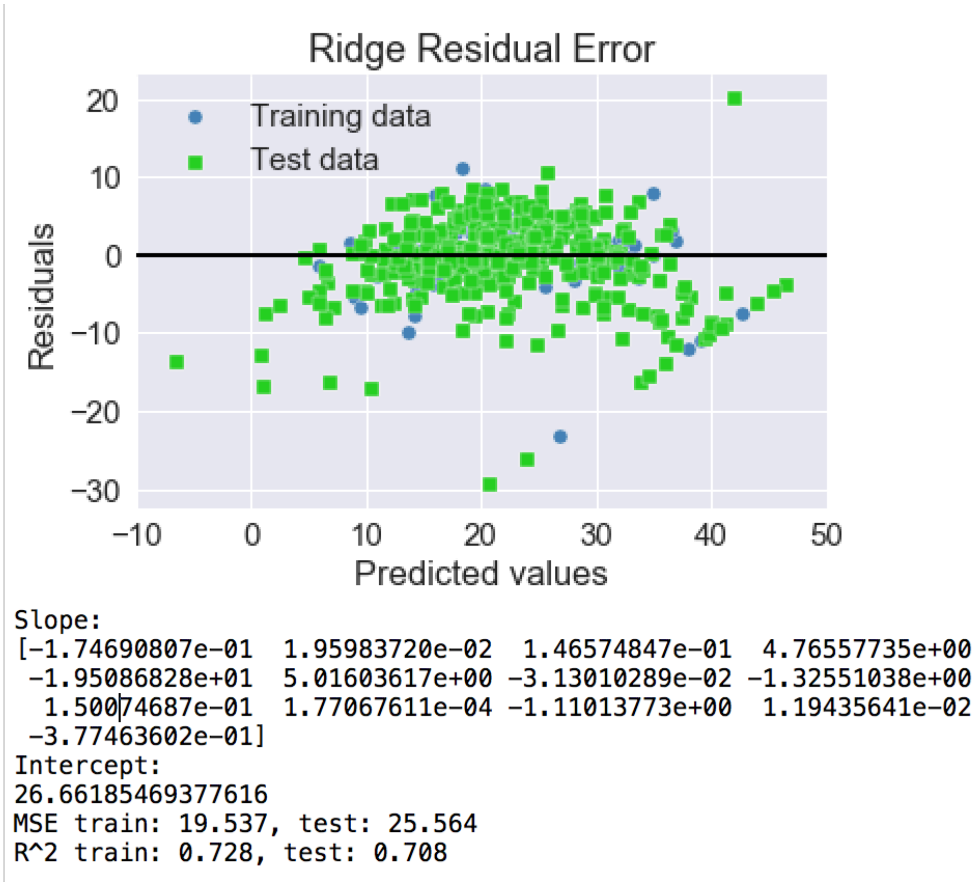
-3.77463602e-01]

Intercept: 26.66185469377616

MSE train: 19.537, test: 25.564

R^2 train: 0.728, test: 0.708

Test on alphas of 0.0001, 0.001, 0.01, 0.1, 1, and 10. Try to find the alpha that gives the lowest MSE and highest R squared. I found that alpha = 0.0001 is the best alpha among those. I also tested further for numbers less than 0.0001 and found that the alpha is performing better. Thus, an alpha that is closer to 0 is the alpha that gives the best performing model.



```
Slope:
[-1.74690807e-01  1.95983720e-02  1.46574847e-01  4.76557735e+00
 -1.95086828e+01  5.01603617e+00 -3.13010289e-02 -1.32551038e+00
  1.50074687e-01  1.77067611e-04 -1.11013773e+00  1.19435641e-02
 -3.77463602e-01]
Intercept:
26.66185469377616
MSE train: 19.537, test: 25.564
R^2 train: 0.728, test: 0.708
```

Part 3.2: LASSO regression

Fit a LASSO model using SKlearn to all of the features of the dataset. Test several settings for alpha. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2. Which alpha gives the best performing model?

Slope: [-1.74601431e-01 1.96018309e-02 1.46421485e-01 4.76452165e+00

-1.94735102e+01 5.01582124e+00 -3.13322070e-02 -1.32496353e+00

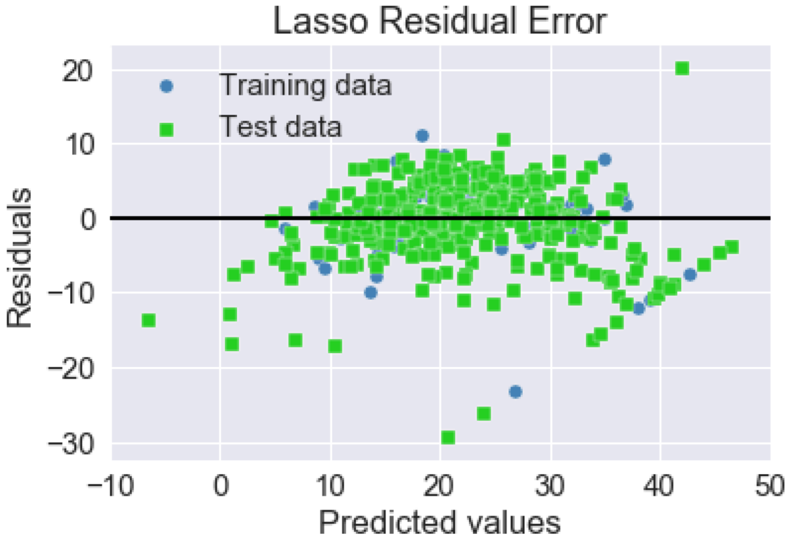
1.49887040e-01 1.75723076e-04 -1.10980528e+00 1.19433243e-02 -3.77524963e-01]

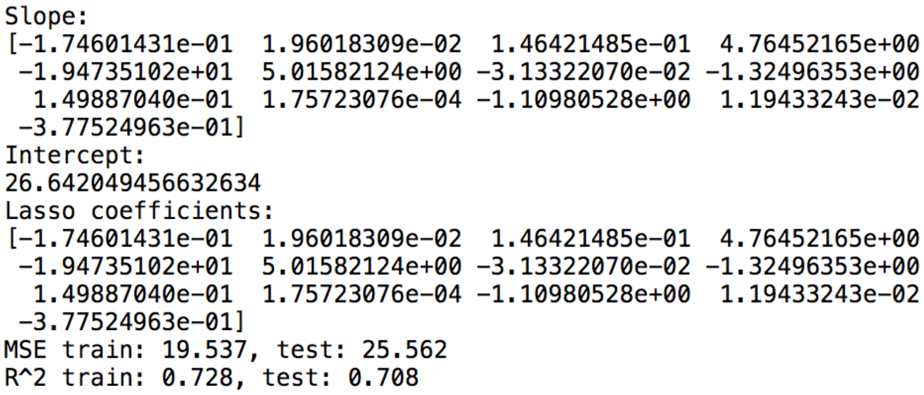
Intercept: 26.642049456632634

MSE train: 19.537, test: 25.562

R^2 train: 0.728, test: 0.708

Test on alphas of 0.0001, 0.001, 0.01, 0.1, 1, and 10. Try to find the alpha that gives the lowest MSE and highest R squared. I found that alpha = 0.0001 is the best alpha among those. I also tested further for numbers less than 0.0001 and found that the alpha is performing better. Thus, an alpha that is closer to 0 is the alpha that gives the best performing model.





```
Slope:
[-1.74601431e-01  1.96018309e-02  1.46421485e-01  4.76452165e+00
 -1.94735102e+01  5.01582124e+00 -3.13322070e-02 -1.32496353e+00
  1.49887040e-01  1.75723076e-04 -1.10980528e+00  1.19433243e-02
 -3.77524963e-01]
Intercept:
26.642049456632634
Lasso coefficients:
[-1.74601431e-01  1.96018309e-02  1.46421485e-01  4.76452165e+00
 -1.94735102e+01  5.01582124e+00 -3.13322070e-02 -1.32496353e+00
  1.49887040e-01  1.75723076e-04 -1.10980528e+00  1.19433243e-02
 -3.77524963e-01]
MSE train: 19.537, test: 25.562
R^2 train: 0.728, test: 0.708
```

Part 3.3: Elastic Net regression

Fit an ElasticNet model using SKlearn to all of the features of the dataset. Test several settings for l1_ratio. Describe the model (coefficients and y intercept), plot the residual errors, calculate performance metrics: MSE and R2. Which l1_ratio gives the best performing model?

Slope: [-1.36059645e-01 3.12989496e-02 6.84574920e-02 2.38879659e+00

-0.00000000e+00 4.07857097e+00 -4.42775002e-02 -1.08485379e+00

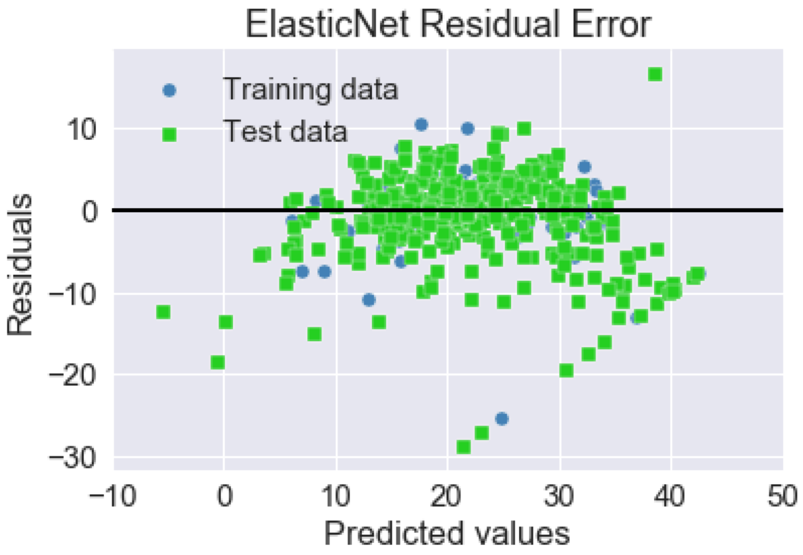
8.80233659e-02 -2.61156057e-03 -9.45487490e-01 1.08164533e-02 -4.72944221e-01]

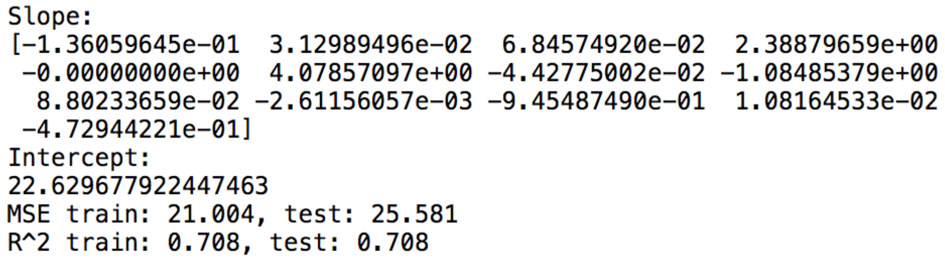
Intercept: 22.629677922447463

MSE train: 21.004, test: 25.581

R^2 train: 0.708, test: 0.708

Set alpha = 0.1 and then test r1_ratios of 0.0001, 0.001, 0.01, 0.1, 1, 0.5, 0.7, 0.9. Try to find the r1_ratio that gives the lowest MSE and highest R squared. I found that alpha = 0.7 is the best alpha among those. I first look at the test data's MSE, limit the choices of r1_ratio to 0.5 and 0.7 since they both have the lowest MSE of 25.581 and a largest R squared of 0.708. Then I compare these two's train dataset's MSE and R squared. I found that when r1_ratio = 0.7, it gives a lower MSE and higher R squared. Thus, alpha = 0.7 is the alpha that gives the best performing model.





```
Slope:
[-1.36059645e-01  3.12989496e-02  6.84574920e-02  2.38879659e+00
 -0.00000000e+00  4.07857097e+00 -4.42775002e-02 -1.08485379e+00
  8.80233659e-02 -2.61156057e-03 -9.45487490e-01  1.08164533e-02
 -4.72944221e-01]
Intercept:
22.629677922447463
MSE train: 21.004, test: 25.581
R^2 train: 0.708, test: 0.708
```

Part 4: Conclusions

Among these models, the best performing model is the Simple Linear Regression model. As we can see that when the alpha is getting closer and closer to 0, Ridge and Lasso is performing better and better. And when alpha = 0, Ridge and Lasso are actually the same as Simple Linear Regression model. Also, comparing the MSE and R-Squared of Simple Linear Regression, Ridge Regression, Lasso Regression, and Elastic Net Regression, we still see that Simple Linear Regression is performing the best on this given data set.

Part 5: Appendix

Link to github repo: https://github.com/jzhuuhzj/IE598_F18_HW4.git