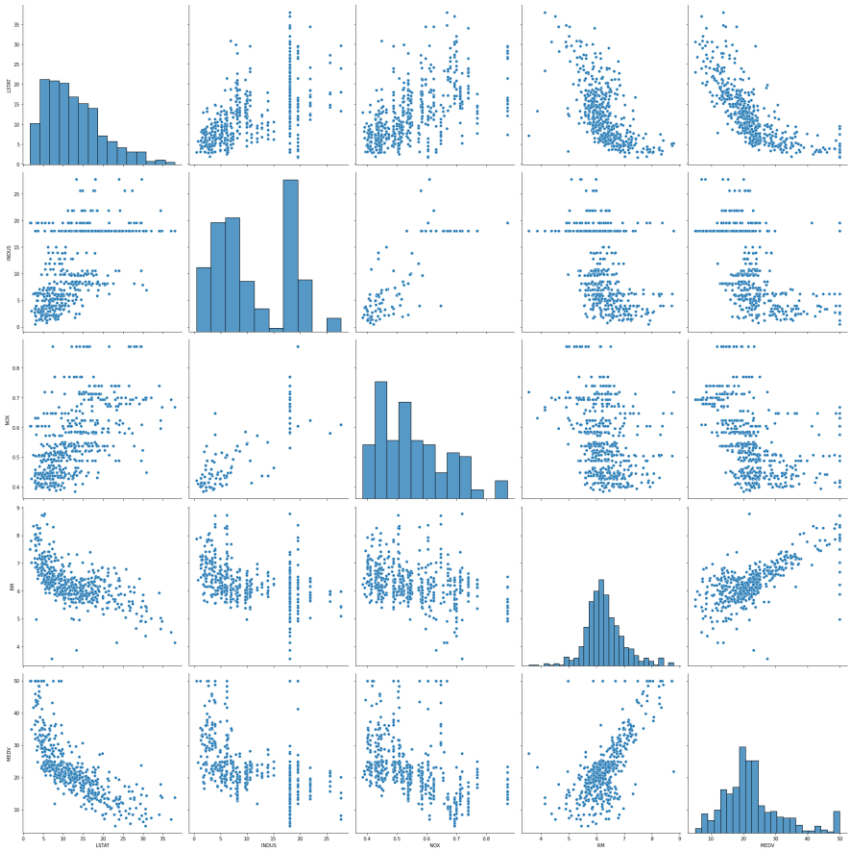
ShunTat Lam (stlam2)

IE517 S23

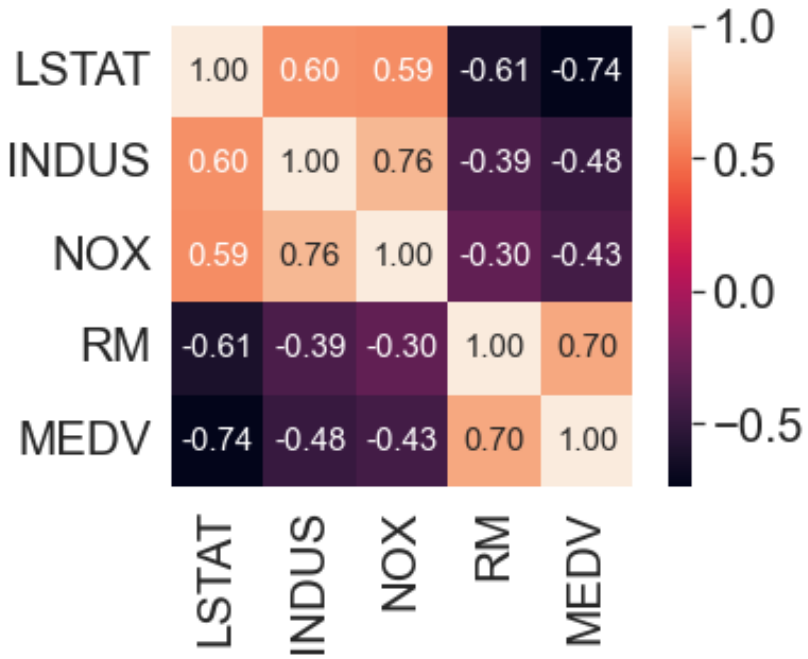
Module 4 Homework (Regression)

Part 1: Exploratory Data Analysis

The data contains 13 explanatory variables and one scalar response(MEDV)



scatter plot of features.



heatmap of correlations between features

(In EDA part only parts of the "x" are selected, since trying to run all of them ruined the laptop and resulted to a forced -shut down.)

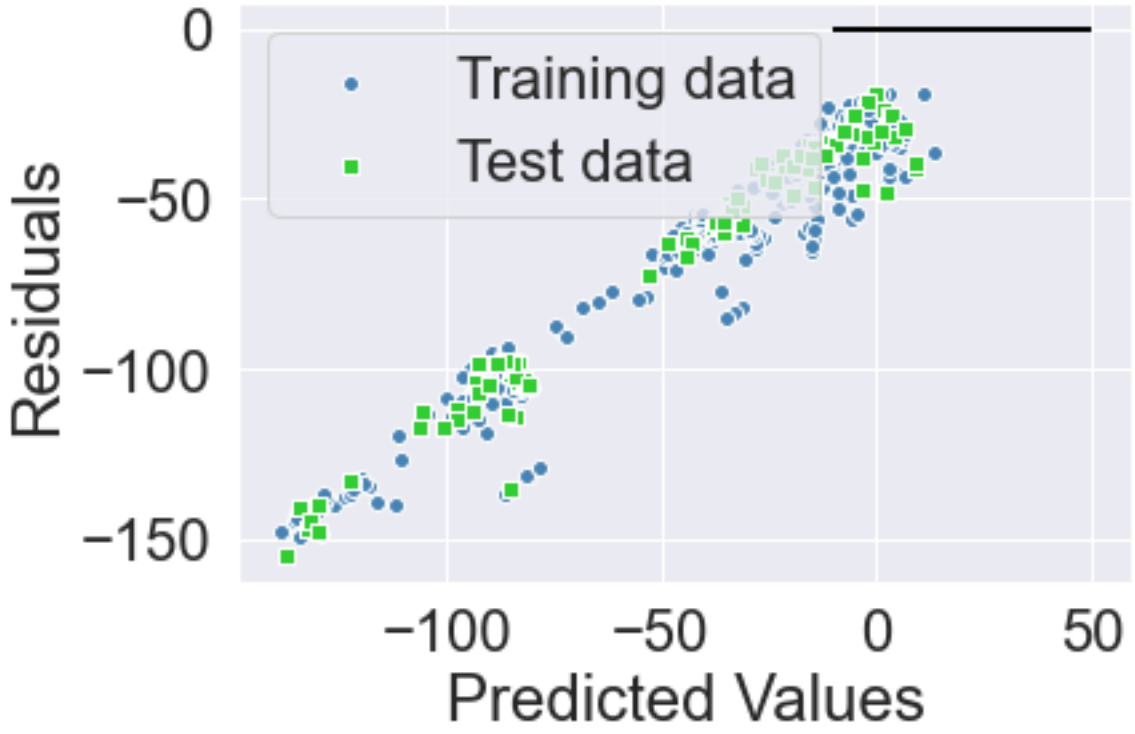
Part 2: Linear regression

The first regression trying is using all 13 x, into a regular linear regression.

The models' coefficients and y intercept are shown below

```
Slope 1 :-0.108
Slope 2 :0.075
Slope 3 :0.030
Slope 4 :0.077
Slope 5 :-0.217
Slope 6 :0.337
Slope 7 :-0.019
Slope 8 :-0.331
Slope 9 :0.242
Slope 10 :-0.190
Slope 11 :-0.219
Slope 12 :0.121
Slope 13 :-0.387
Intercept:0.000
```

The residual



The models MSE and R^2:

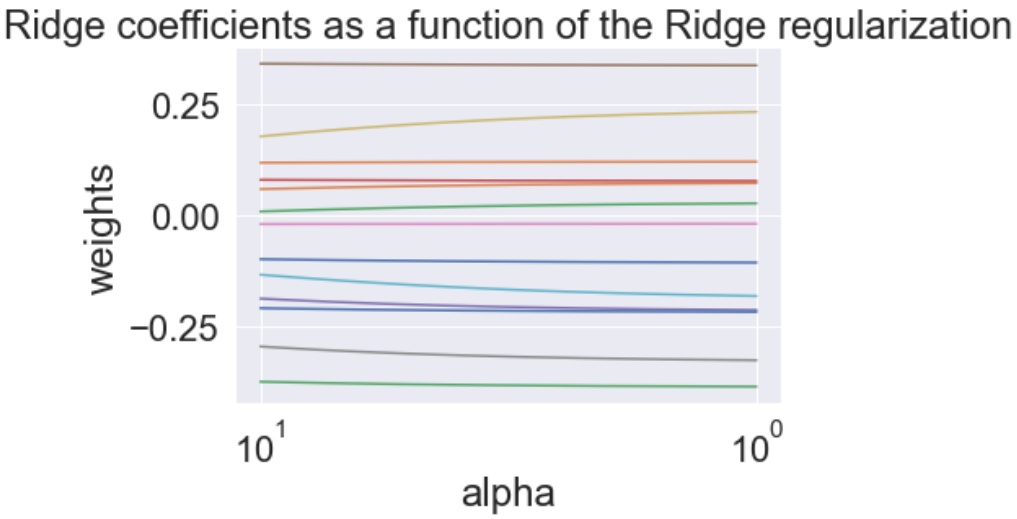
```
MSE train: 5101.026, test:5611.544
R square :-57.718
```

apparently the model is not good. There is undesired pattern in the residual plot, the MSE is high.

(Tried multiple methods,including .score and manually calculating R^2 but always get a negative value.)

Part 3.1: Ridge regression

The next trying is to use ridge. The graph below shows the changing of weight in predictors with the changing of alpha. Looks like alpha does not affect much.



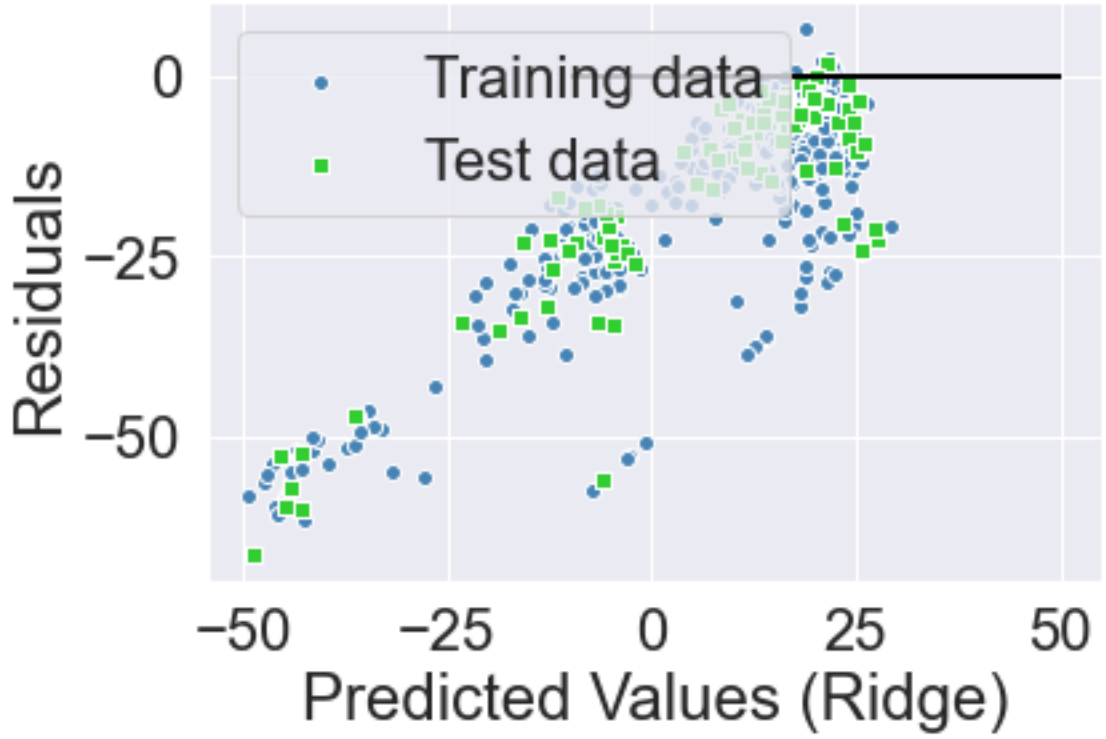
The coefficients and intercept:

```
Slope 1 :-0.074
Slope 2 :0.027
Slope 3 :-0.035
Slope 4 :0.084
Slope 5 :-0.091
Slope 6 :0.322
Slope 7 :-0.019
Slope 8 :-0.159
Slope 9 :0.044
Slope 10 :-0.054
Slope 11 :-0.170
Slope 12 :0.100
Slope 13 :-0.305
Intercept:0.000
```

MSE and R-squares:

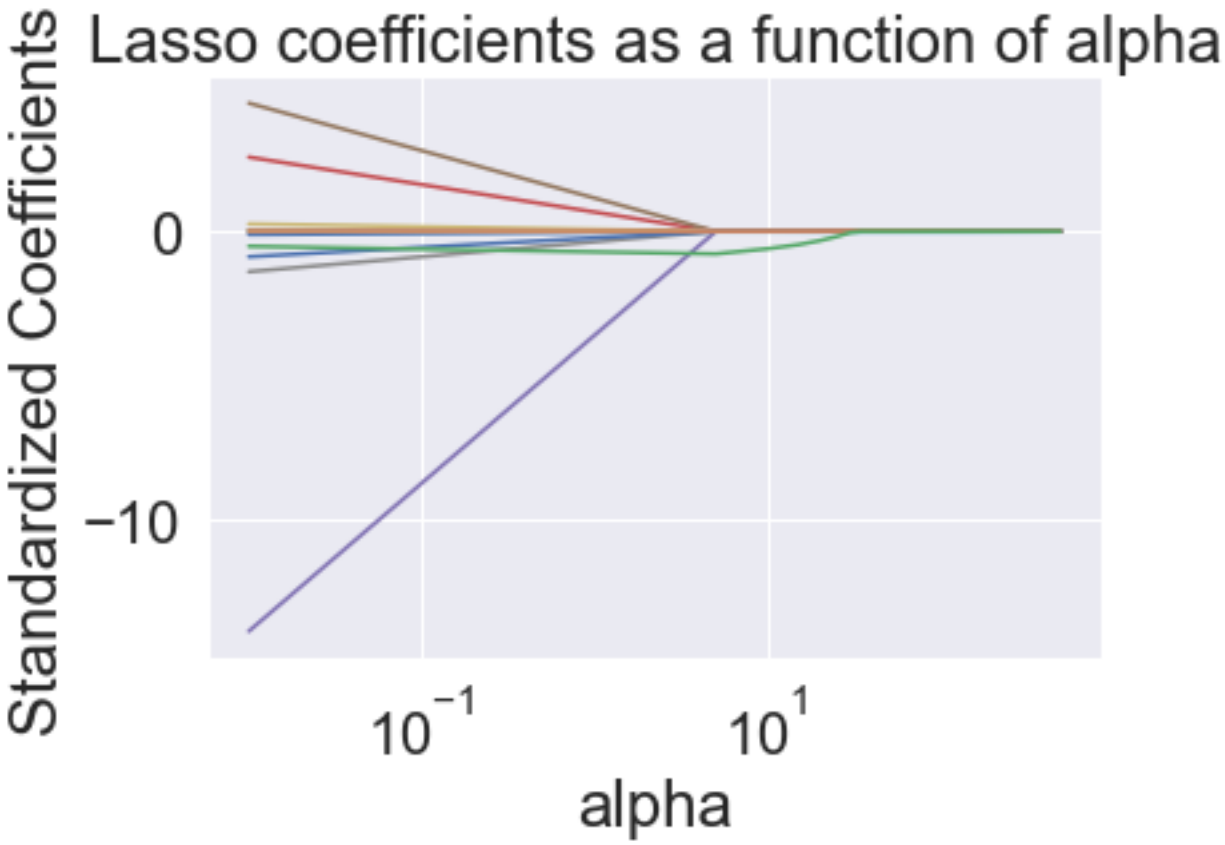
```
MSE train: 451.152, test:471.666
R square :-4.193
```

The residual:



Based on the MSE, ridge performs better.

Part 3.2: LASSO regression



The issue of lasso is it seems like the alpha setting results to set all predictors' weight to 0.

```
Slope 1 :-0.000
Slope 2 :0.000
Slope 3 :-0.000
Slope 4 :0.000
Slope 5 :-0.000
Slope 6 :0.000
Slope 7 :-0.000
Slope 8 :0.000
Slope 9 :-0.000
Slope 10 :-0.000
Slope 11 :-0.000
Slope 12 :0.000
Slope 13 :-0.000
Intercept:0.000
```

```
MSE train: 24.060, test:23.466
R square :0.723
```

So, just use the ridge

Part 4: Conclusions

When dealing with multiple x, ridge performs much better than the regular regression with no penalty. For lasso, it seems like alphas are hard to deal with, sometimes they just result to all weights = 0

Part 5: Appendix

Link to github repo

https://github.com/stlam2/IE517_S2023/blob/main/IE517_S23_HW4/Homework4.py

link to reference:

https://scikit-learn.org/stable/auto_examples/linear_model/plot_ridge_path.html#sphx-glr-auto-examples-linear-model-plot-ridge-path-py

https://www.kirenz.com/post/2019-08-12-python-lasso-regression-auto/