Exploratory Data Analysis and Penalized Linear Regression

Zhen Wang

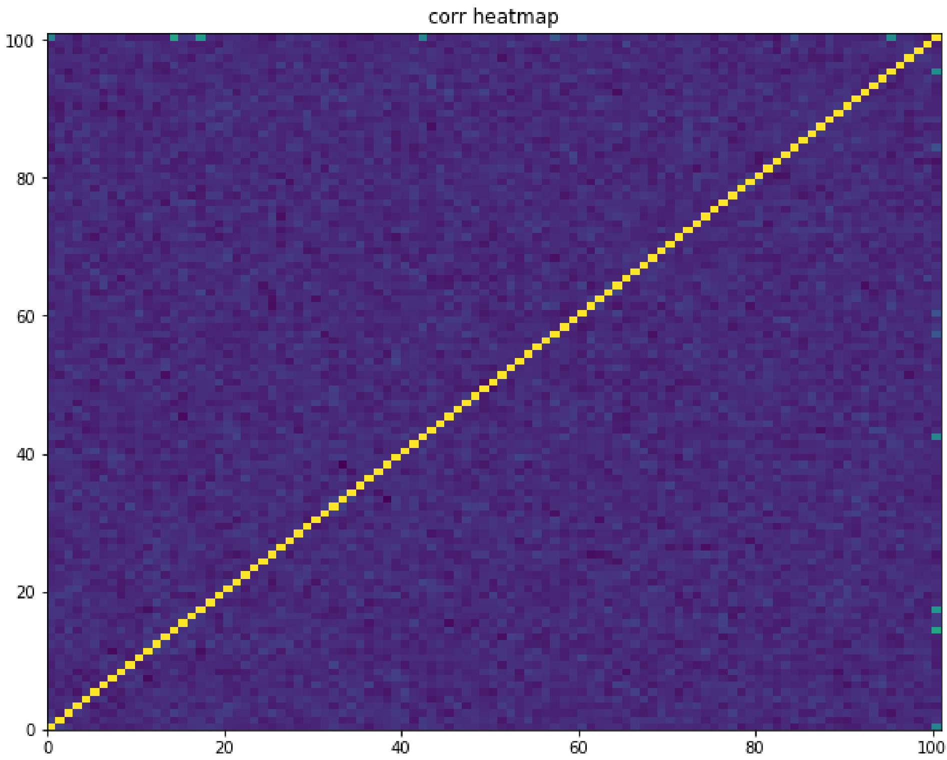
https://github.com/zhenw3/IE598_F18_HW4

Data: make_regressionX.csv make_regressionY.csv

Variable numbers: 100

Sample size: 1000

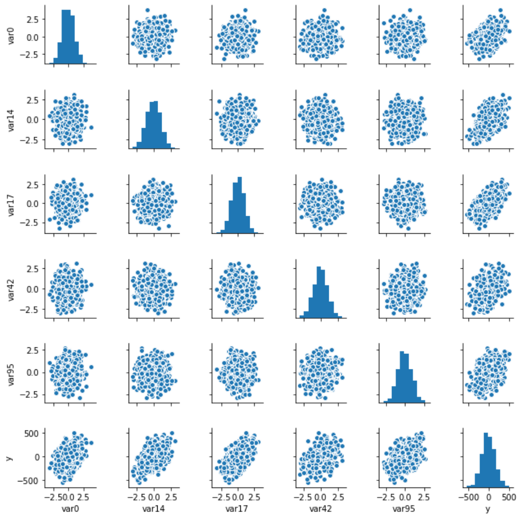
#1. Overview of correlations between 100 variables and y

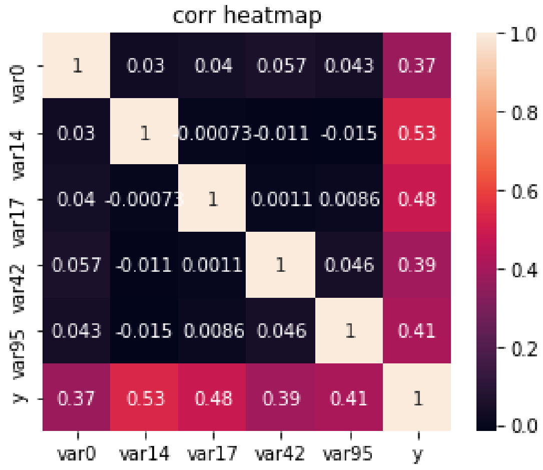


The heatmap shows the correlations between 100 variables and the dependent variable y. The picture indicates that the most correlations between variables and y are low. But because some green spots appears at the edge of the heatmap, several variables have much higher correlations with y than other variables. The fact that correlations between variables are low[1] and that some variables are linearly correlated with y[2], suggests that multiple linear regression could be applied to this dataset.

#2. Selecting features

By eliminating features that have correlation coefficient with y lower than 0.3 and larger than -0.3, the features of linear models could be reduced from 100 to 5. The correlation matrix and scatter plot of selected features (will be stated as features below) is :

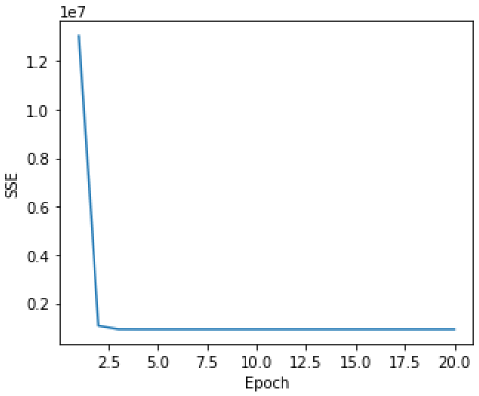




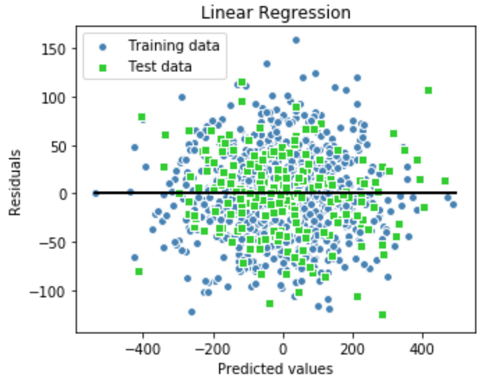
The scatter plot confirms that correlations between variables are low for the spots cluster as a noisy cloud while correlations of these five variables with y are high for the spots cluster in certain directions. Also, features and y are all distributed in a shape close to normal distribution.

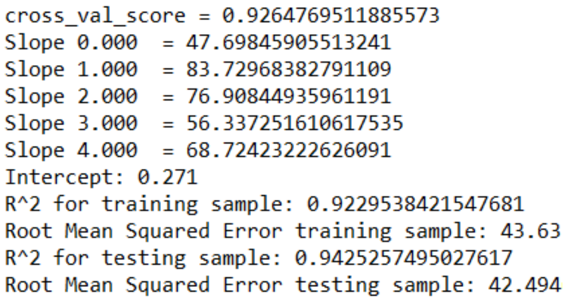
From the correlation heatmap of this 5 features and y we can also draw a similar conclusion.

#3. Linear regression



The picture shows the relation between number of iterations and sum of squared errors (SSE). As the number of iteration increases, the SSE decreases, which indicates linear regression model converges in this dataset.



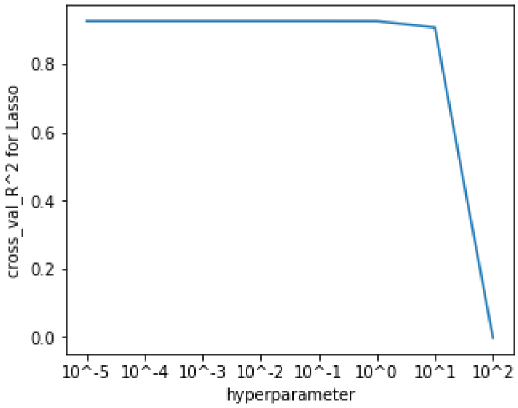


```
cross_val_score = 0.9264769511885573
Slope 0.000  = 47.69845905513241
Slope 1.000  = 83.72968382791109
Slope 2.000  = 76.90844935961191
Slope 3.000  = 56.337251610617535
Slope 4.000  = 68.72423222626091
Intercept: 0.271
R^2 for training sample: 0.9229538421547681
Root Mean Squared Error training sample: 43.63
R^2 for testing sample: 0.9425257495027617
Root Mean Squared Error testing sample: 42.494
```

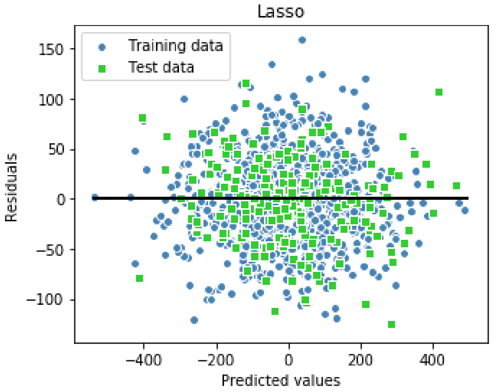
The upper left picture shows the residuals of training sample (blue) and testing sample (green). Randomly clustered residuals indicate that residuals don't vary for y, which is an important hypothesis for linear regression.

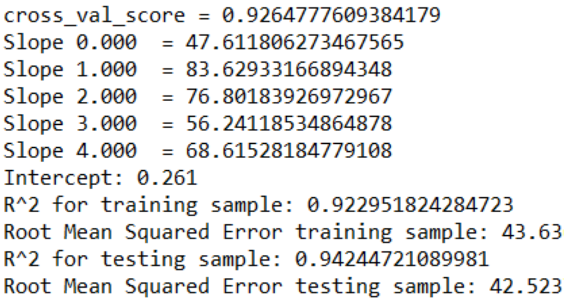
The upper right picture shows some performance indicators of the regression model. 'Cross validation score' is the average R^2 of the model in cross validation test (setting cv = 10). 92.3% is a relatively high score. Following are coefficient (or slope) of the 5 features as well as the intercept. Finally is the R^2 and mean squared errors (MSE) for training sample and testing sample. The model actually performs better in testing sample than in training sample.

#4. Lasso Regression



The picture shows the relation between the value of hyperparameter of lasso and R^2 in cross validation test. The optimized hyperparameter turns out to be 10^{-1} measured by the standard of R^2. However, when alpha equals 100, weights of each feature becomes 0 and R^2 drops dramatically from over 0.9 to 0. This indicates some limitations of lasso model.



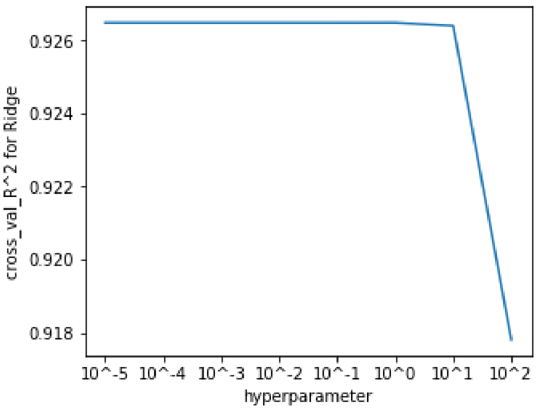


```
cross_val_score = 0.9264777609384179
Slope 0.000  = 47.611806273467565
Slope 1.000  = 83.62933166894348
Slope 2.000  = 76.80183926972967
Slope 3.000  = 56.24118534864878
Slope 4.000  = 68.61528184779108
Intercept: 0.261
R^2 for training sample: 0.922951824284723
Root Mean Squared Error training sample: 43.63
R^2 for testing sample: 0.94244721089981
Root Mean Squared Error testing sample: 42.523
```

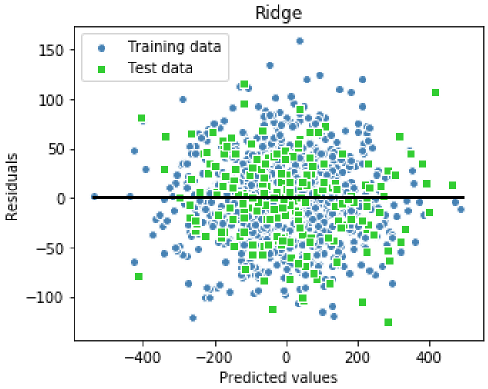
The upper left picture shows the residuals of training sample (blue) and testing sample (green).

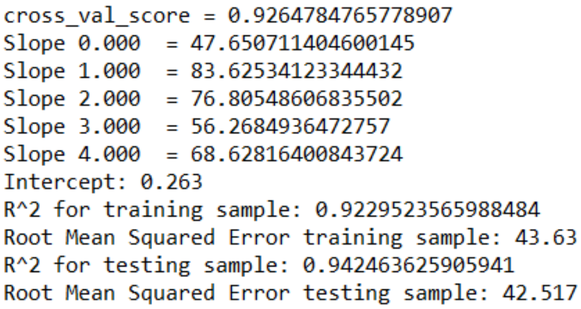
The upper right picture shows R^2 of lasso regression (using the optimized alpha 0.1) in cross validation test, training sample and training sample, the MSE and coefficients. Linear regression actually slightly outperforms lasso both in training sample and testing sample, while lasso outperforms linear regression in cross validation. Both models perform better in testing sample than in training sample.

#5. Ridge regression



The picture shows the relation between the value of hyperparameter of ridge and R^2 in cross validation test. The optimized hyperparameter turns out to be 1. Though the shape is similar to that of lasso, it should be noticed that R^2 only drops to around 91.8% rather than to 0.



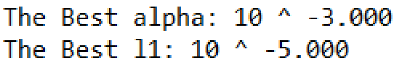


```
cross_val_score = 0.9264784765778907
Slope 0.000  = 47.650711404600145
Slope 1.000  = 83.62534123344432
Slope 2.000  = 76.80548606835502
Slope 3.000  = 56.2684936472757
Slope 4.000  = 68.62816400843724
Intercept: 0.263
R^2 for training sample: 0.9229523565988484
Root Mean Squared Error training sample: 43.63
R^2 for testing sample: 0.942463625905941
Root Mean Squared Error testing sample: 42.517
```

The upper left picture shows the residuals of training sample (blue) and testing sample (green).

The upper right picture shows R^2 of ridge regression (using the optimized alpha 1) in cross validation test, training sample and training sample, the MSE and coefficients. Linear regression actually slightly outperforms ridge both in training sample and testing sample, while ridge outperforms linear regression in cross validation. Both models perform better in testing sample than in training sample. Also, ridge strictly outperforms lasso.

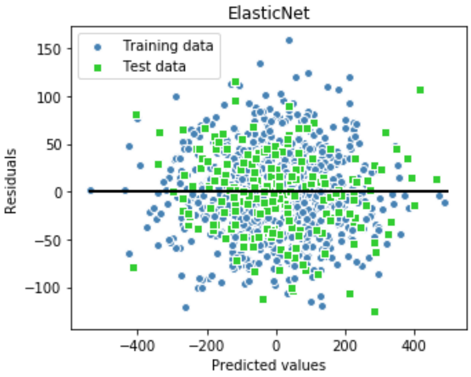
#6. Elastic Net

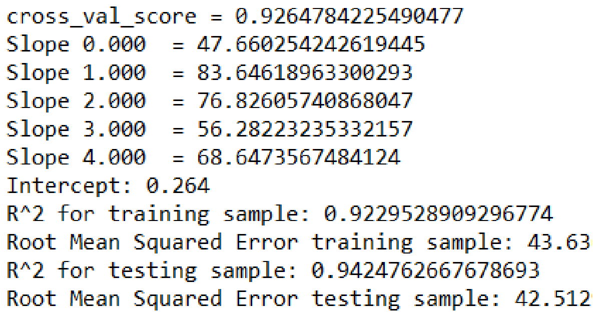


```
The Best alpha: 10 ^ -3.000
The Best l1: 10 ^ -5.000
```

After searching the value iteratively ranging from 10^{-5} to 10 by the standard of average R^2 for cross validation (cv = 10), the optimized alpha and l1_ratio are 10^{-3} and 10^{-5}.

Use the optimized value of alpha and l1_ratio to train elastic net model.





```
cross_val_score = 0.9264784225490477
Slope 0.000  = 47.660254242619445
Slope 1.000  = 83.64618963300293
Slope 2.000  = 76.82605740868047
Slope 3.000  = 56.28223235332157
Slope 4.000  = 68.6473567484124
Intercept: 0.264
R^2 for training sample: 0.9229528909296774
Root Mean Squared Error training sample: 43.63
R^2 for testing sample: 0.9424762667678693
Root Mean Squared Error testing sample: 42.512
```

The upper left picture shows the residuals of training sample (blue) and testing sample (green).

The upper right picture shows R^2 of elastic net regression in cross validation test, training sample and training sample, the MSE and coefficients.

#6. Conclusion

Linear regression, lasso regression, ridge regression and elastic net regression actually return similar result (similar R^2, similar coefficients) in this dataset. This might be due to the feature selection process before running regressions, during which irrelevant features have been eliminated. They all fit the data relatively well.