The Model Selection Methodology

Group 9, CHEN XU

19 novembre 2018

Group:9

Guangyue CHEN , Jiahui XU

Methodology.

Our data has two target variables to project(ASP:"Actual sales prices" and ACC:"Actual construction costs"), so we will train two models for them.

At first, we chose some models as candidates: linear regression, LASSO, RIDGE, ElasticNet. To consider the case that our dataset has a high dimension, it's better to use LASSO, RIDGE or ElasticNet regression. Because regression with regularization suit the dataset, which has high dimension and multicollinearity, very well.

To choose the model, we can compute the mean square errors of the new data, so we can choose the model with the lowest mean square error.

And also, we will use K-fold cross validation, it's a good way to compare the models.

Select the model.

We separate the data into 10 folds, and we will choose one of them as the test data. And for the rest, we will use them as the training data. So we write a for loop to get 10 models for each regression to use different folds. So that we can get the averages of 10 mean square errors for each regression. We will compare the averages to choose the regression.

Except the linear regression, we should choose the parameter(λ) for Lasso, Ridge and ElasticNet regression.

For ElasticNet regression, we use the function "cv.glmnet" 10 times with different training folds, so we get 10 models, and for each model we predict the estimators with the λ who gives minimum mean cross-validated error(for one model). Then we compute the mean square errors of these 10 models with the test data. At last, we get the average of the errors.

For Lasso regression, we use the function "lars" 10 times to get the models. For each time we predict the estimators with the λ who gives the minimum RSS. Then we compute the mean square errors of these 10 models with the test data. Finally we can compute the average of the errors.

For Ridge regression, we use the function "lm.ridge" 10 times to get the models and choose the λ. Then we do the same things.

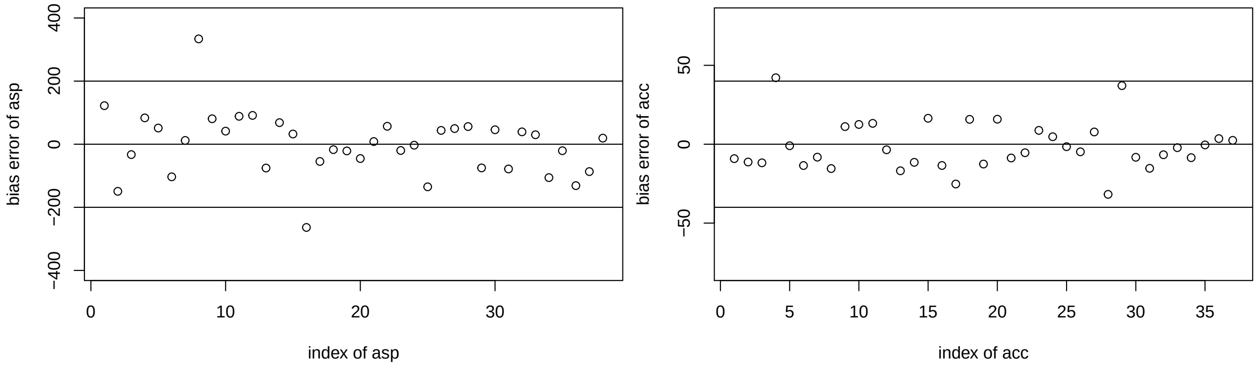
Compare all of these regression

```
## [1] "MSE for ASP[linear]:299.457427701972"

## [1] "MSE for ASP[RIDGE]:182.125637442427"

## [1] "MSE for ASP[LASSO]:62538.4859996998"

## [1] "MSE for ASP[Elastic Net]:170.513245764816"

## [1] "MSE for ACC[linear]:124.965196623734"
## [1] "MSE for ACC[RIDGE]:30.9088702761834"

## [1] "MSE for ACC[LASSO]:1526.10602497816"

## [1] "MSE for ACC[Elastic Net]:29.9210603086819"
```

We can see that for “Actual sales prices”, Elastic Net regression has a small average of errors. And for “Actual construction costs”, RIDGE regression is better.

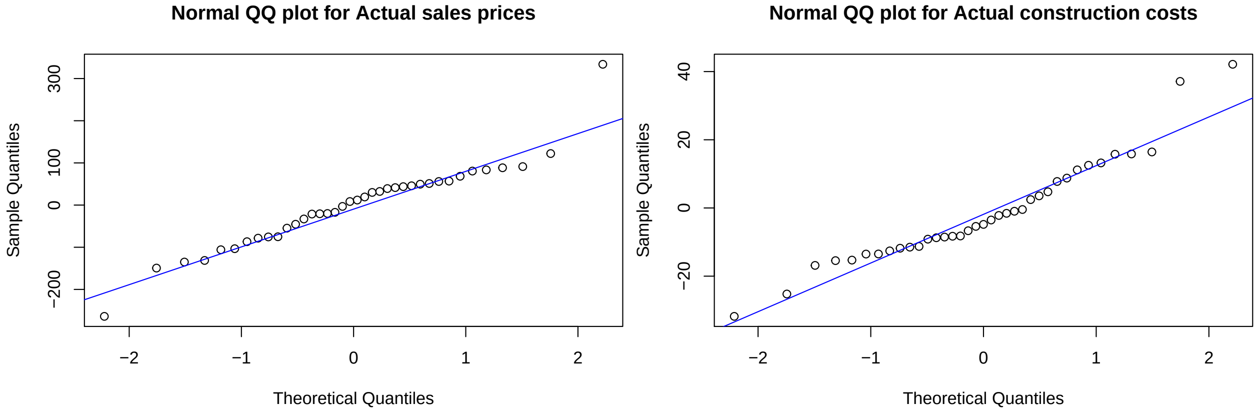
So we will choose the one , which has the lowest mean square error, of the 10 Elastic Net model to predict “Actual sales prices”. And RIDGE regression for predicting “Actual construction costs”.

Test the Final Model



The resaults of the model will have the bias errors, for the colomn “Actual sales prices”, the prediction equals the real value ±200, and for the colomn “Actual construction costs”, the prediction equals the real value ±40.

Then we test it with QQ plot.



So we see that the Normal Q-Q Plot is almost a straight line , although there are some differences on the two sides of the blue line, that’s the result we mostly expect to see. So we can summarize that they are good models.