Applying Best Subset Selection, Ridge Regression, and Lasso to Boston Crime Rate Prediction

Data Analysis Assignment Lesson 4

David Chen

Feburary 11th, 2020

Introduction

In this report, we attempt to build the best linear model to predict per capita crime rate from the 1978 `Boston` dataset. To determine the optimal model, we applied the best subset selection, ridge regression, and the lasso methods, validating using 10-fold cross validation. Ultimately, we decided that the lasso model with 11 variables (out of 13) was the best, having a low test MSE and a reduced number of total variables.

Data

For this report, we will be considering the `Boston` dataset, information collected by the U.S. Census Service for housing values in the suburbs of Boston (published in 1978). Included are a total of 506 observations and the following 14 variables:

1. `crim` - per capita crime rate by town
2. `zn` - proportion of residential land zoned for lots over 25,000 sq.ft
3. `indus` - proportion of non-retail business acres per town
4. `chas` - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. `nox` - nitrogen oxides concentration (parts per 10 million)
6. `rm` - average number of rooms per dwelling
7. `age` - proportion of owner-occupied units built prior to 1940
8. `dis` - weighted mean of distances to five Boston employment centres
9. `rad` - index of accessibility to radial highways
10. `tax` - full-value property-tax rate per $10,000
11. `ptratio` - pupil-teacher ratio by town
12. `black` - $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
13. `lstat` - lower status of the population (percent)
14. `medv` - median value of owner-occupied homes in $1000s

Reference: Variable information was provided by the R documentation.

Note that for all the variables are numeric, although `chas` is a dummy variable.

Exploratory Data Analysis

We begin with some preliminary analysis into the variables for any pecularities. First, we create correlation plots, detailed in the Apprendix Figures 2 and 3. Immediately, we can note that multiple variables had relatively high correlations with `rad`, the variable indicating accessibility to radial highways. Thus, we investigate further with the individual (jittered) scatterplots:

Figure 1: Rad Variable Relationships

Here, we observe that the correlations appear to be heavily influenced by the the existence of `rad` values close to 25 while the rest are much lower. Since basic transformations (squared, log) do not change this relationship, we leave the data as is and proceed with caution.

Analyses

For this analysis, we will determine the best (linear) model to predict `crim`, the per capita crime rate by town. We will use the best subset selection, ridge regression, and the lasso to determine the best model. We will use cross validation on half of the data as a training set, afterwards applying the selected model to the full dataset.

Best Subset Selection

First, we begin with best subset selection. With 10-fold cross validation, we observe that the model with the lowest cross validation error (41.381) is the ten variable model. From the full model, variables `chas`, `age`, and `tax` were removed. The remaining coefficients had the folowing coefficients:

	Coefficients
(Intercept)	16.386
zn	0.042
indus	-0.093
nox	-10.622
rm	0.448
dis	-0.991
rad	0.536
ptratio	-0.270
black	-0.008
lstat	0.131
medv	-0.198

Recall that the coefficients can be interpreted like the following: For every unit increase in `lstat`, the value of `crim` increases by 0.1308.

Post-submission comment: We needed to check for the conditions of applying a regression before continuing here. We should have transformed the response to the log-scale here for this interpretation to be valid.

Ridge Regression

We proceed forward with a ridge regression model, again validating using a 10-fold cross validation. We observe that the optimal lambda value with the lowest training MSE is 0.595, and thus we proceed forward with that value. Using that lambda, we observe a test MSE of 25.211. We can note that while all the variables are included in the model (typical of ridge regressions), the coefficients for `age`, `tax`, and `black` are extremely close to 0. These variables may not be as influential as the others.

Table 2: Ridge Regression Model Coefficients

	Coefficients
(Intercept)	8.576
zn	0.032
indus	-0.081
chas	-0.740
nox	-5.068
rm	0.327
age	0.002
dis	-0.682
rad	0.413
tax	0.004
ptratio	-0.127
black	-0.009
lstat	0.143
medv	-0.136

The Lasso

Lastly, we apply the lasso. After using 10-fold cross validation on half the data as a training set, we observe that the optimal lambda selected was 0.120. Applied to the testing set, we observe a test MSE of 25.887. Applied to the entire dataset, we observe that 11 variables were selected, `age` and `tax` being dropped from the model (coefficients set to 0). We can also note that `rm` and `black` both have coefficients extremely close to 0.

The coefficients are as following:

Table 3: The Lasso Model Coefficients

	Coefficients
(Intercept)	7.625
zn	0.029
indus	-0.043
chas	-0.481
nox	-2.455
rm	0.007
dis	-0.524
rad	0.486
ptratio	-0.075
black	-0.008
lstat	0.117
medv	-0.115

Conclusion

Ultimately, we decide to proceed forward with the 11 variable model selected by the lasso. Comparing all the selected models, we observe that the best subset selection chose a ten variable model with a test MSE of 41.38097, the ridge regression had a value of 25.21093, and the lasso had a value of 25.887. While the ridge regression technically did have the lowest MSE value, the minimal difference with the lasso lead us to select the latter model with fewer variables.

Analyzing the model selected by the lasso, there are a total of 11 variables (out of the original 13 possible). The variables `tax` and `age` were both dropped, while the variables `rm` and `black` are close to 0. Thus, not all the features of the original dataset are included in the final model, indicating that not all variables were significant in predicting `crim`. We can note that while most coefficients were small, `nox`, the nitrogen oxides concentration, appeared to have the strongest relationship with the per capita crime rate.

Note that comparing the best subset, ridge, and lasso output, the variables `age`, `tax`, and `black` were all either removed or set close to 0. This inspires confidence in our results since the variables chosen were relatively universal.

Appendix

Plots

```
Boston_cor <- cor(Boston)
corrplot(Boston_cor, method = 'circle')
```

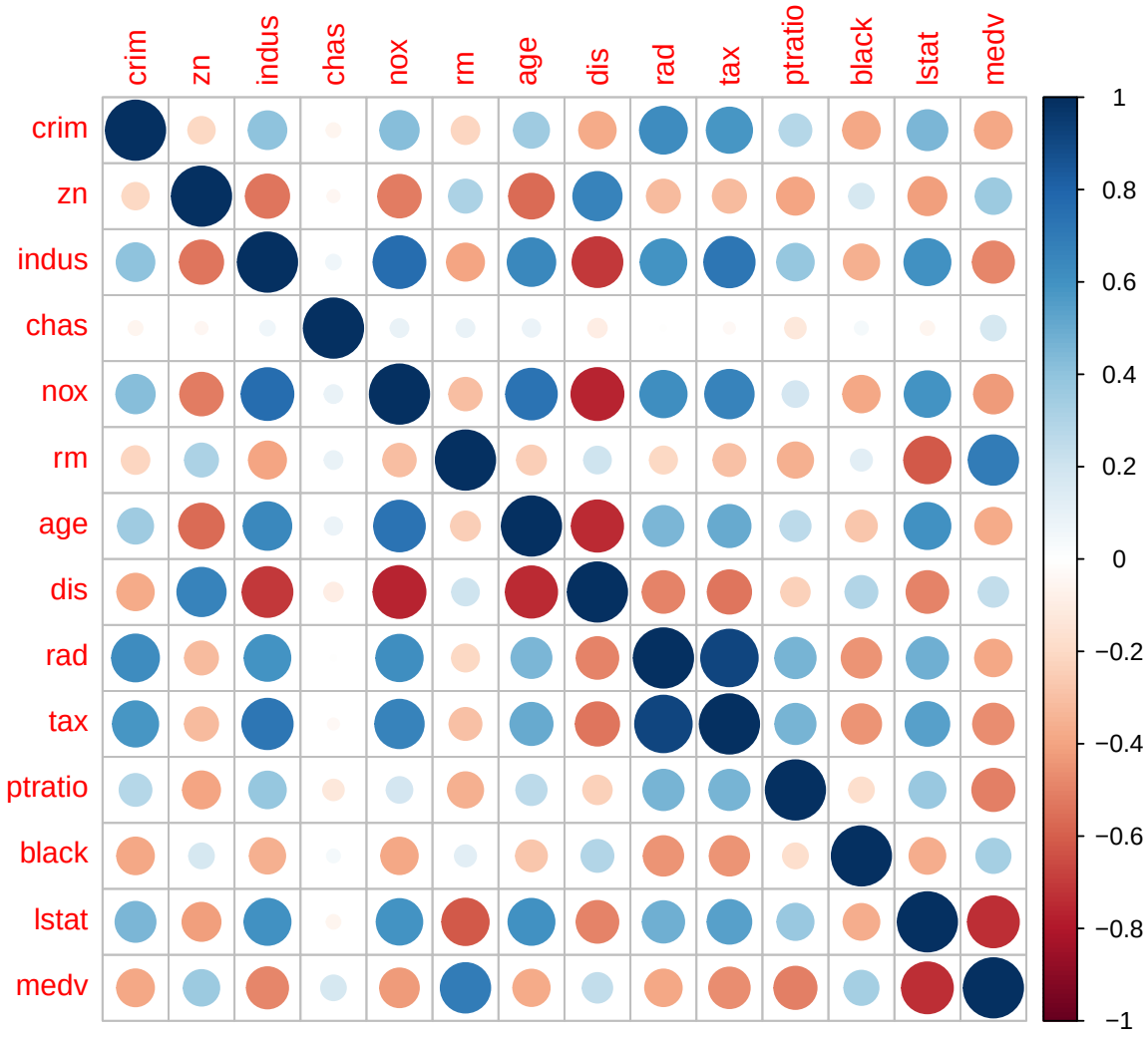

Figure 2: Circle Correlation Plot

```
corrplot(Boston_cor, method="number", type = 'upper')
```

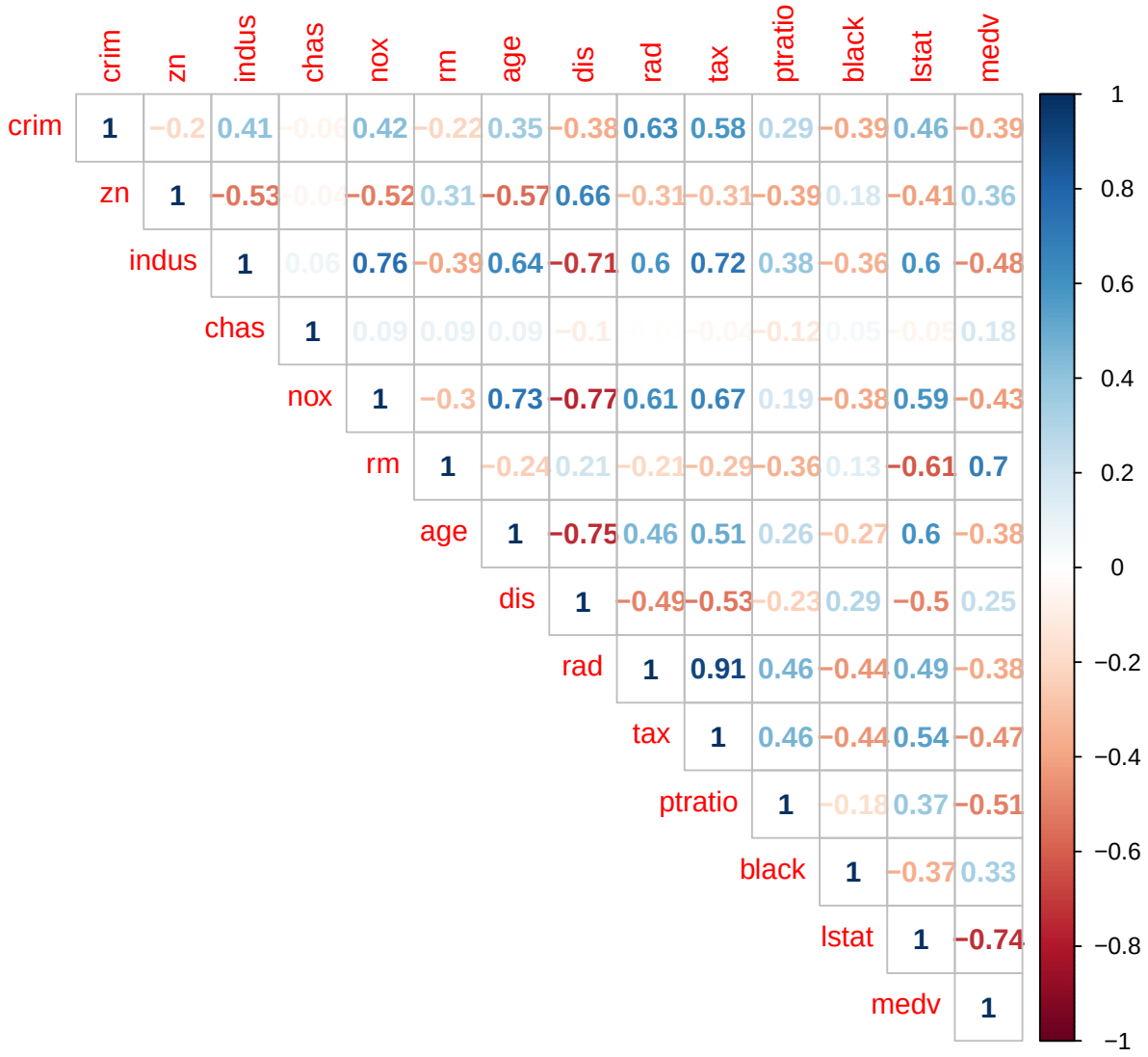

Figure 3: Numeric Correlation Plot