Applying PCR and PLS to the Boston Crime Dataset

Data Analysis Assignment Lesson 6

David Chen

February 25th, 2020

Introduction

Continuing our analysis of the Boston data set and prediction of the per capita crime rate by town, we attempt to perform a principal component regression (PCR) and partial least squares (PLS). These are two methods that attempt to reduce the total number of variables by creating linear combinations of the original variables, ideally explaining roughly the same amount of variance but with less predictors. However, from this analysis, we observe that the models chosen did not generate lower test MSE values than the lasso and ridge regression done previously.

Data

Repeated from Data Analysis Assignment 4

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:

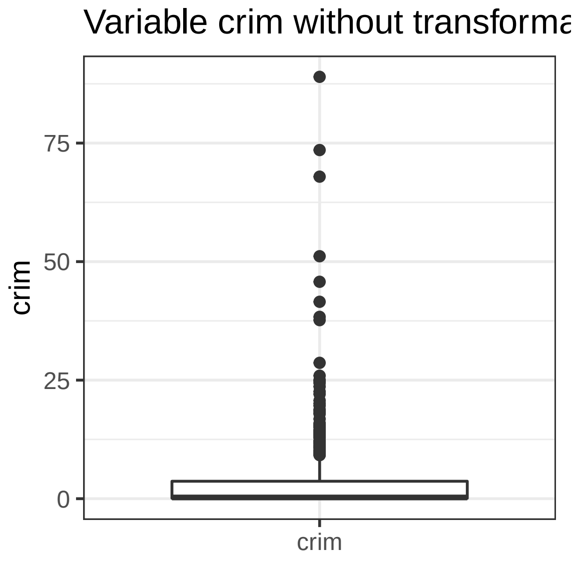
1. `crim` - per capita crime rate by town
2. `zn` - proportion of residential land zoned for lots over 25,000 sq.ft
3. `indus` - proportion of non-retail business acres per town
4. `chas` - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. `nox` - nitrogen oxides concentration (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

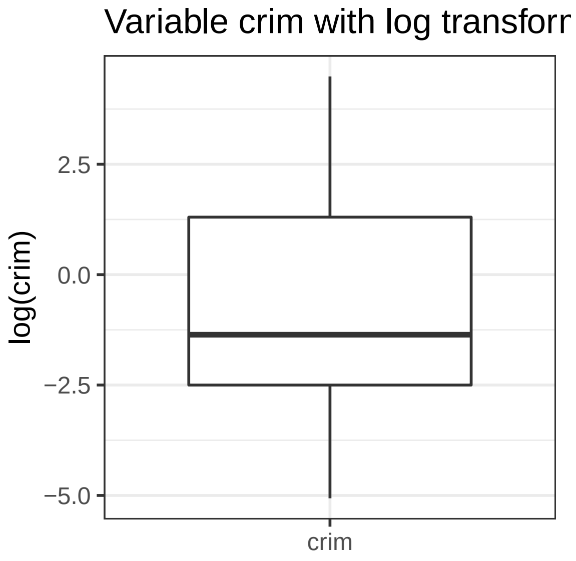
Reference: Variable information was provided by the R documentation.

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

Exploratory Data Analysis

While basic correlation plots and exploratory plots were presented in Data Analysis Assignment Lesson 4, we observe that a log transformation of `crim` vastly reduces the skewness of the data, as seen below. Thus, proceeding forward, `crim` will analyzed under a log transformation.


Variable crim without transforma


Variable crim with log transforr

Analysis

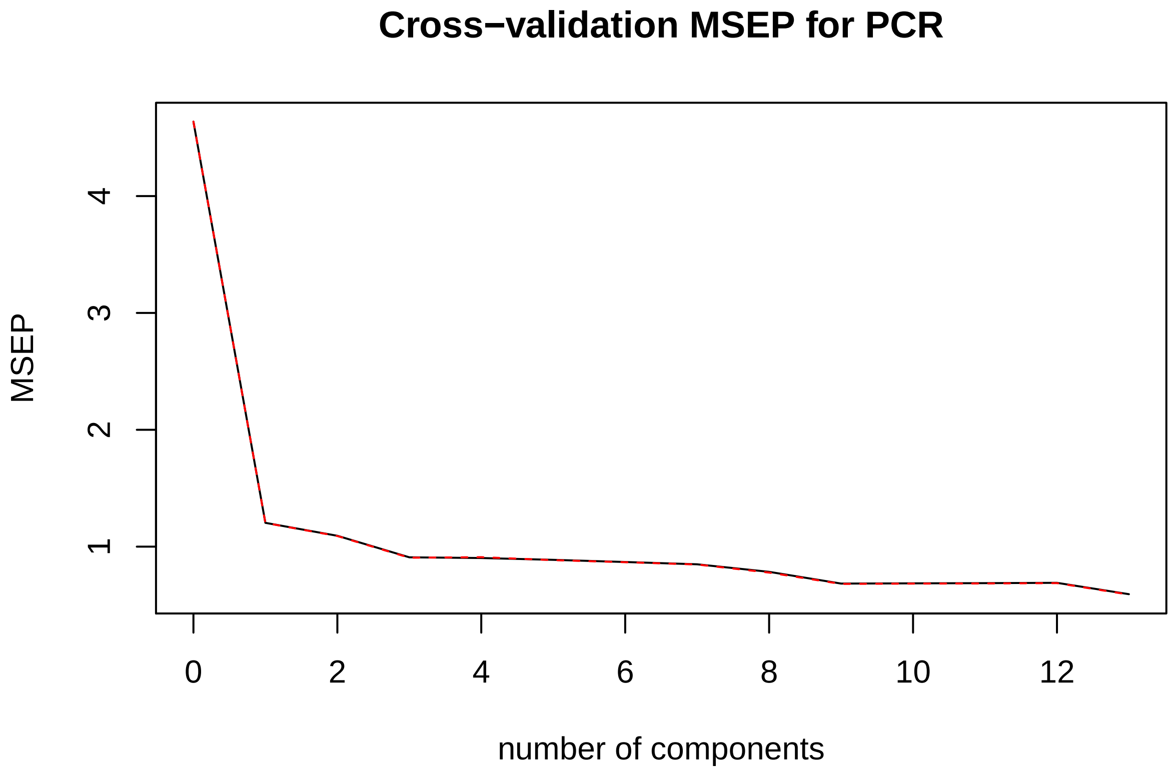
First, we will perform principal components regression (PCR) and partial least squares (PLS) to determine their predictive performance. Then, we will rerun best subset selection, the lasso, and ridge regression with the log transformation to generate compare prediction results.

For this analysis, we will evaluate model performance using validation set error (test MSE with half the original data).

Principal Components Regression

First, we begin with PCR. Recall that this method functions similarly to principal component analysis, where the original variables are transformed into linear combinations that maximize variance and are uncorrelated with each other. Here, the resulting linear combinations (principal components) are used in a regression model, ideally resulting in less total predictors than before.

Below, we observe a plot of the cross-validation MSE using the training data. We can observe that while the full model technically has the lowest training MSE, the improvement is marginal after ~9 components. While for this analysis we will stick with the best MSE (13 components), it must be understood that 9 components may perform equally well, if not better for test MSE.



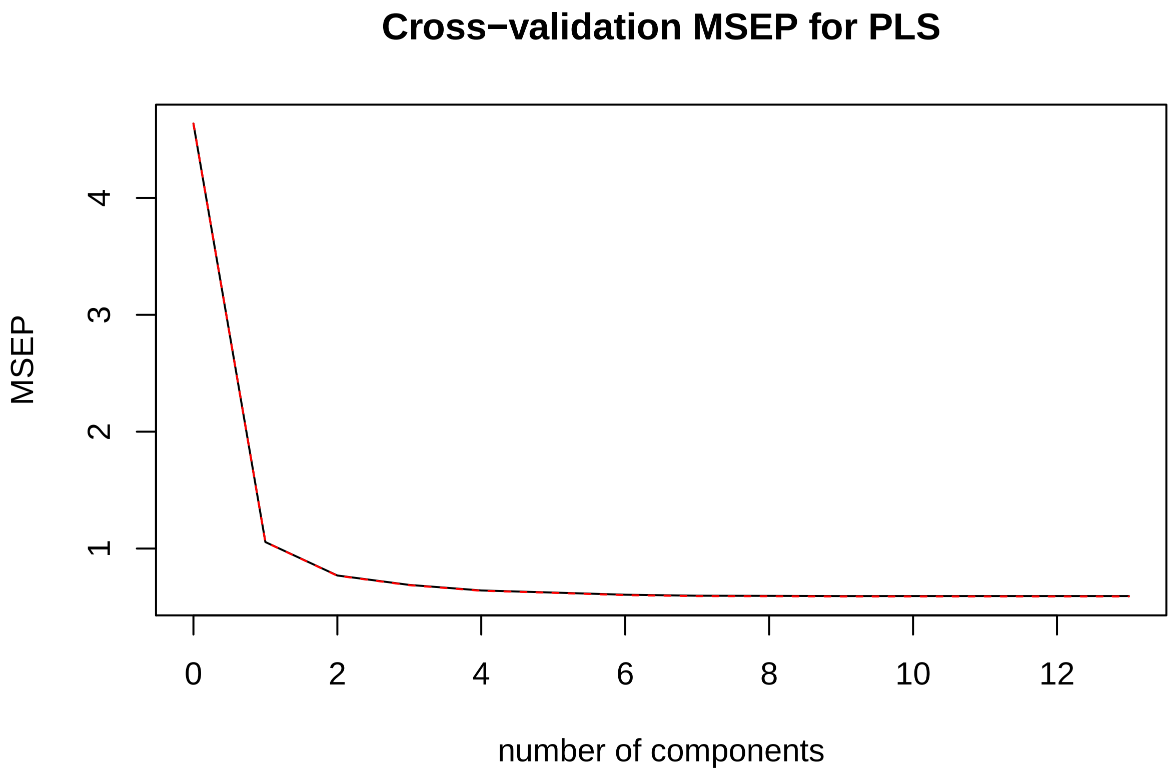
Using 13 components, we then determine the test MSE and the full model statistics. We observe that the test MSE is 0.715, with 87.54% of the total variance in log(crim) explained by the model. Note that since all variables are included, this is the same result as a normal linear regression.

Partial Least Squares

Next, we continue with a PLS. This model function similarly to the PCR, but instead emphasizes predictor's relationships with the response (`crim`) more than PCR.

Conducting the PLS, we observe that the number of components with the smallest training MSEP is 9 components, although the difference between many of the models is extremely small. From the plot, one can reasonably argue that a model with as few as 4 components could be reasonable, but again for this analysis we will remain with the lowest MSE.

Checking the test MSE with 9 components, we observe a value of 0.713. We can note that 87.54% of the total variance in log(crim) can be explained with this model, a this value only marginally increases with each added component.



Repeating DA4 Analysis

While this analysis has already been done in DA4, previously I did not conduct a log transformation on `crim`, reducing the predictive abilities. Thus, briefly we will repeat these analyses with the new transformation. We will also determine the test MSE of all the results to compare with the PCR and PLS results.

Much of the technical details are removed here to emphasize the results. For a more in-depth discussion of the cross validation or model foundations, please refer to the previous assignment.

Best Subset Selection

Conducting the best subset selection method, we observe that the model with all 13 variables has the lowest training MSE. Thus, when applied to the test set, we observe the test MSE of 0.715, the same as the PCR.

The Lasso

Applying the lasso, we observe that 2 variables (`chas` and `tax`) were removed from the final model. With 11 variables, the model had a test MSE value of 0.665.

Ridge Regression

Lastly, we conduct the ridge regression. We observe that the optimal lambda with the training data is 0.186, resulting in a test MSE of 0.629.

All Results

The test MSE of all the methods are shown in this table:

	MSE
PCR	0.715
PLS	0.713
Best Subset	0.715
Lasso	0.665
Ridge	0.629

Conclusion

Examining our models, we observe that both PCR and PLS performed worse than the lasso and ridge regression results. This is likely due to the fact that both PCR and PLS were unable to effectively reduce the total number of variables. Thus, we can observe that our rule of choosing the lowest training MSE, regardless of marginal gains, was likely not the optimal selection method for number of components. If we chose less components for each method, we may have observed higher MSE values.