Q22

Rahul Atre

2023-10-31

Q22 [Regularization]

We will now try to predict per capita crime rate in the **Boston** data set.

1. Try out some of the regression methods explored in this chapter, such as best subset selection, the LASSO, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

```r
bostonData = read.csv("Boston.csv")
```

```r
# Method i) Least-Squares Approach
lin_modelSqr = lm(crim ~ ., data = bostonData)
summary(lin_modelSqr)
```

```
## 
## Call:
## lm(formula = crim ~ ., data = bostonData)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -9.854 -2.164 -0.363  0.993 74.935 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.709e+01  7.240e+00   2.361 0.018616 *  
## X           -1.672e-03  2.828e-03  -0.591 0.554672    
## zn           4.598e-02  1.884e-02   2.440 0.015030 *  
## indus       -6.418e-02  8.346e-02  -0.769 0.442269    
## chas        -7.310e-01  1.181e+00  -0.619 0.536327    
## nox         -1.019e+01  5.283e+00  -1.929 0.054269 .  
## rm           4.568e-01  6.149e-01   0.743 0.457889    
## age         -2.893e-04  1.818e-02  -0.016 0.987310    
## dis         -1.001e+00  2.830e-01  -3.538 0.000442 ***
## rad          6.009e-01  9.068e-02   6.626 9.07e-11 ***
## tax         -3.314e-03  5.219e-03  -0.635 0.525733    
## ptratio     -2.696e-01  1.866e-01  -1.445 0.149092    
## black       -7.484e-03  3.677e-03  -2.035 0.042347 *  
## lstat        1.241e-01  7.586e-02   1.636 0.102379    
## medv        -2.007e-01  6.064e-02  -3.310 0.001000 ** 
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

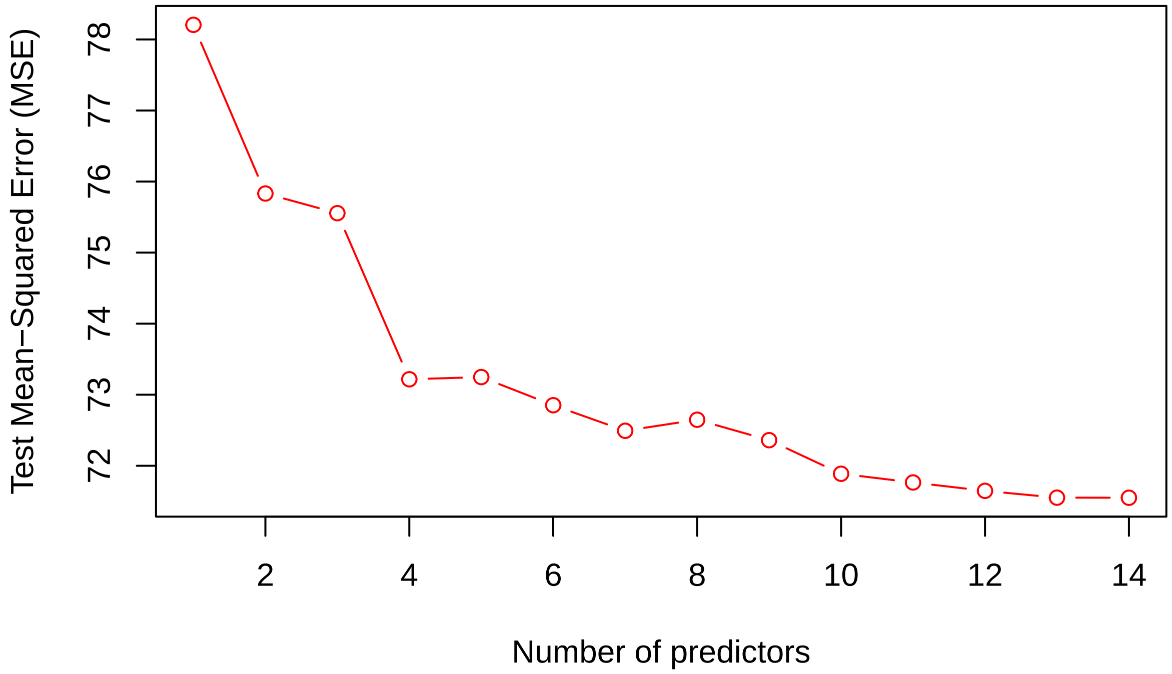
```
##
## Residual standard error: 6.443 on 491 degrees of freedom
## Multiple R-squared:  0.4544, Adjusted R-squared:  0.4388
## F-statistic: 29.21 on 14 and 491 DF,  p-value: < 2.2e-16
```

Variables that are statistically significant: zn, dis, rad, black, medv.

```r
library(leaps) #regsubsets import
```

```
## Warning: package 'leaps' was built under R version 4.3.2
```

```r
# Method 2: Best subset selection
set.seed(144) #Ensure reproducible outputs for simulation

p = 15 # Declare feature and observation values
n = 506

x = matrix(rnorm(n * p), n, p) #rnorm generates a vector which is reshaped to matrix (obs. vs. features

b = rnorm(p) # Beta vector for model
eps = rnorm(n)

y = x %*% b + eps #Create linear model from best subset selection

training_set = sample(seq(n), n*0.60, replace = FALSE)
test_set = -training_set

y.test = bostonData$crim[test_set]

test_matrix = model.matrix(crim ~ ., data=bostonData[test_set,])

best_subset <- regsubsets(crim ~ ., data = bostonData, nvmax = p)
value_errs <- rep(NA, 14) #Variable to store training MSE

for(i in 1:14){
 coefficients <- coef(best_subset, id = i) #Obtain coefficients of model from i predictors
 prediction <- test_matrix[,names(coefficients)] %*% coefficients #prediction value
 value_errs[i] <- mean((y.test - prediction)^2)
}

# Plot test MSE
plot(value_errs, xlab = "Number of predictors", ylab = "Test Mean-Squared Error (MSE)", type = "b", col
```



2. Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross- validation, or some other reasonable alternative, as opposed to using training error.

The best subset model appears to perform better than the least squares approach since it is only selecting the most significant variables. For least-squares however, all variables were chosen to predict the crime rate but only a few were significant. The best subset model has one major drawback however, that it is computationally intensive and can sometimes lead to over fitting of the data.

Due to the computational overhead of best subset selection, I would choose the least-squares model for its simplicity. From the above graph, we can see that best subset chose all the predictors due to a reduction in the Test MSE. So ultimately, it would have not made a significant difference if the least-squares approach was conducted.

3. Does your chosen model involve all of the features in the data set? Why or why not?

For the best-subset selection, all of the predictors have been selected. This is because as the number of predictors increase, the MSE decreases.