Lasso Regression

Tran Bui

6/18/2021

a. Divide the data set into training and testing sets, where the training data consists of 80% of the total data.

b. Use cross-validation on the training data to find the best value of the hyperparameter λ.

c. Fit the lasso regression model with the best value of hyperparameter λ.

Lasso

$$C(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

$\lambda \geq 0$ is a tuning parameter (or hyperparameter).

Packages

Load the data

```
summary(df)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.70   Min.   :4.000   Min.   : 65.68   Min.   : 47.72  
##  1st Qu.:15.11   1st Qu.:4.000   1st Qu.:127.00   1st Qu.: 94.95  
##  Median :19.23   Median :6.000   Median :194.66   Median :116.37  
##  Mean   :20.17   Mean   :6.188   Mean   :234.19   Mean   :144.07  
##  3rd Qu.:22.44   3rd Qu.:8.000   3rd Qu.:333.71   3rd Qu.:191.46  
##  Max.   :35.65   Max.   :8.000   Max.   :518.72   Max.   :345.13  
##       drat             wt             qsec             vs        
##  Min.   :2.640   Min.   :1.572   Min.   :15.24   Min.   :0.0000  
##  1st Qu.:3.040   1st Qu.:2.733   1st Qu.:16.40   1st Qu.:0.0000  
##  Median :3.606   Median :3.343   Median :17.96   Median :0.0000  
##  Mean   :3.556   Mean   :3.268   Mean   :18.21   Mean   :0.4375  
##  3rd Qu.:3.912   3rd Qu.:3.712   3rd Qu.:19.89   3rd Qu.:1.0000  
##  Max.   :4.624   Max.   :5.794   Max.   :23.78   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

→ Identify missing values (if any)

```
with(df, sum(is.na(mpg)))
```

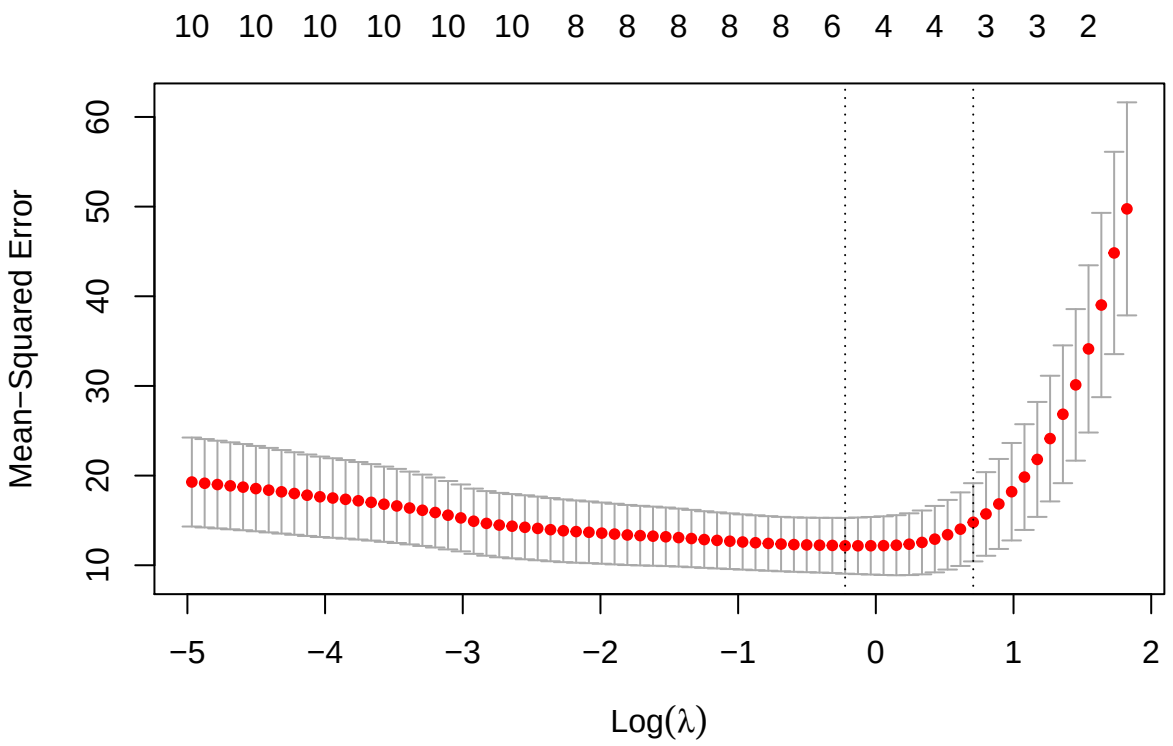
```
## [1] 0
```

⇒ The variable mpg has no NA values.

```
# Predict the mpg variable as a function of all other variables.
# Remove all constant coefficients
x <- model.matrix(mpg ~., df)[,-1]

#select the first column of mpg
y <- df$mpg
```

1. What is the best value of the hyperparameter λ?

```
#Number of observations
n <- length(df$mpg)

#Training data consisting of 80% of the total data.
set.seed(12)
train <- sample(n,round(.8*n))

#Do cross validation using cv.glmnet(alpha = 1)
cv_lasso <- cv.glmnet(x[train,], y[train], alpha = 1)
plot(cv_lasso)
```




```
# return the best lambda value
best_lam <- cv_lasso$lambda.min
best_lam
```

```
## [1] 0.8001868
```

The best value of the hyperparameter λ is 0.8001868

2. What is the root-mean-square error (RMSE) in predicting the mpg in the test data using the model?

```
# fit the function on the training data
lasso_mod <- glmnet(x[train,], y[train], alpha = 1, lambda = best_lam)

# try the predict function on the testing set
lasso_pred <- predict(lasso_mod, s = best_lam, newx = x[-train,])

# calculate the RMSE
sqrt(mean((lasso_pred-y[-train])^2))
```

```
## [1] 3.780886
```

The root-mean-square error (RMSE) used in predicting the mpg in the test data using the model is 3.780886