Linear Models. Open Book CAT

Getrude Gichuhi

08/04/2022

Loading the DataSet and viewing it

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.1.3

df <- read_excel("Cat1.xlsx")
View(df)
```

(a) Split the data set into 75% training set and 25% test set.

```
# 75% of the sample size is the Training set
df_t <- floor(0.75 * nrow(df))

#setting the seed
set.seed(123)
training <- sample(seq_len(nrow(df)), size = df_t)

train <- df[training, ]
test <- df[-training, ]
```

Least squares

b) Fit a linear model using least squares on the training set, and report the test error obtained.

```
library(caret)

## Warning: package 'caret' was built under R version 4.1.3

## Loading required package: ggplot2

## Loading required package: lattice

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.1.3

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'readr' was built under R version 4.1.3

## Warning: package 'forcats' was built under R version 4.1.3

## -- Conflicts ---------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()

library(glmnet)

## Warning: package 'glmnet' was built under R version 4.1.3

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-3

library(dplyr)

model = lm(Response~., data = train)

summary(model)

##
## Call:
## lm(formula = Response ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.861  -7.204   0.419   7.927  24.937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -413.2188    35.1070 -11.770  < 2e-16 ***
## Var1          26.6378     0.4928  54.049  < 2e-16 ***
## Var2          -6.1812     1.2884  -4.797 1.91e-06 ***
## Var3           4.7253     1.0729   4.404 1.20e-05 ***
## Var4           4.3928     1.1677   3.762 0.000181 ***
## Var5          -4.3347     0.6446  -6.724 3.34e-11 ***
## Var6           8.6981     1.2887   6.750 2.83e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.369 on 807 degrees of freedom
```

```
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9917
## F-statistic: 1.626e+04 on 6 and 807 DF,  p-value: < 2.2e-16

#Fitting training Model on the test set
lm_pred=predict(model,new=test)

#Calculating Accuracy
LSE=mean((test$Response-lm_pred)^2)

#Print
print(LSE)

## [1] 99.11668
```

The Test error of the linear model fit is 99.116668

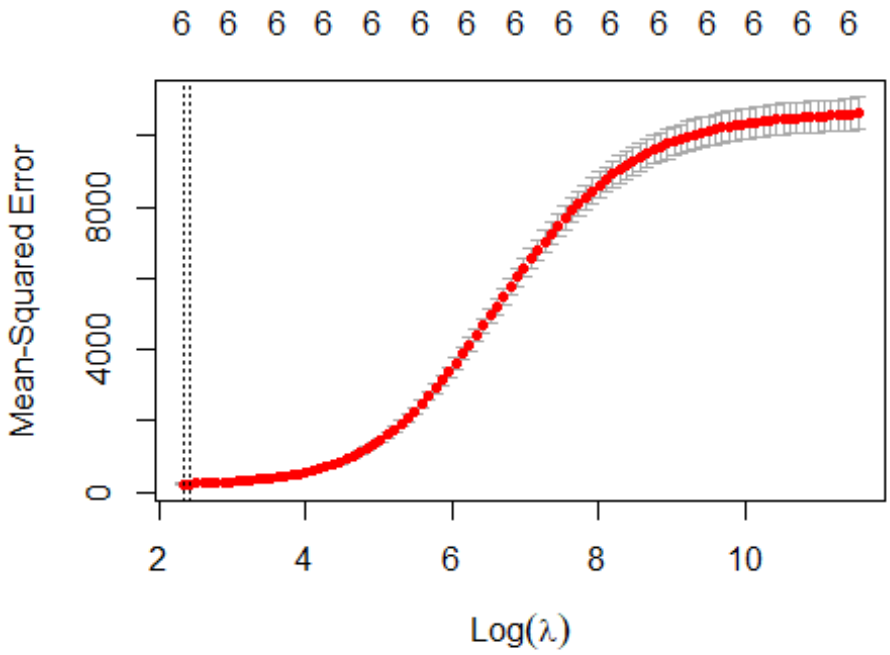
c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

```
set.seed(1)

#Matrices
train_mat = model.matrix(Response~., data = train)
test_mat = model.matrix(Response~., data = test)

#Choose the lambda using cross-validation

cv = cv.glmnet(train_mat, train$Response, alpha=0)
plot(cv)
```




```
lam = cv$lambda.min

lam

## [1] 10.24674

#Fitting the ridge regression

ridge_mod = glmnet(train_mat,train$Response, alpha =0)

#Make Predictions
ridge_pred = predict(ridge_mod, s=lam,newx = test_mat)

#Calculating test error
mean((ridge_pred - test$Response)^2)

## [1] 224.7245
```

The test error of the ridge regression fit with lambda chosen by cross-validation is 224.7245, which is higher that the linear model error.

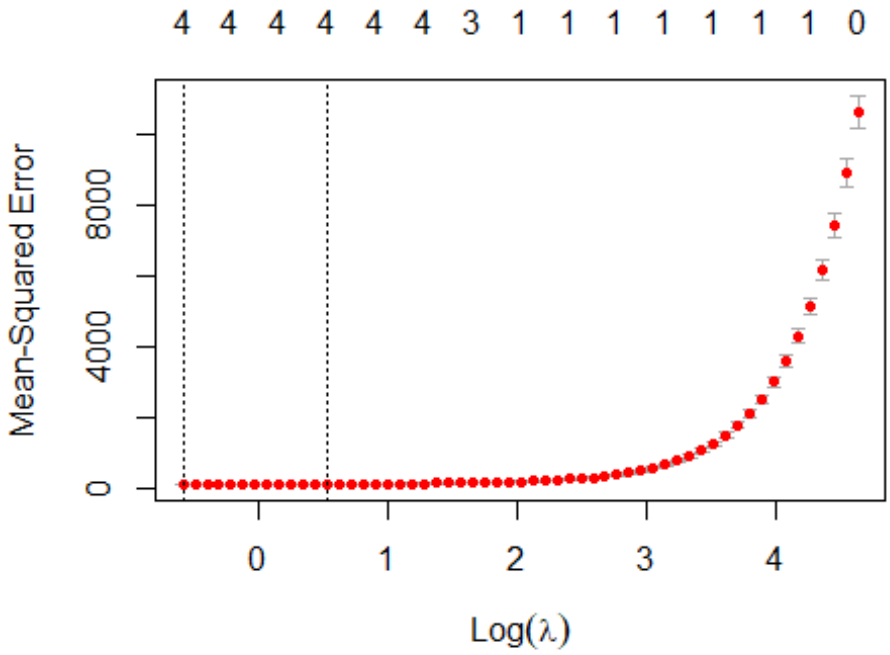
d) Fit a lasso model on the training set, with λ chosen by cross validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
#Choosing the lambda to be used for the cross-validation

set.seed(1)
```

```
cv2 = cv.glmnet(train_mat, train$Response, alpha=1)

plot(cv2)
```




```
lam2 = cv2$lambda.min
lam2

## [1] 0.5597055

#Fitting the Lasso model

lasso = glmnet(train_mat, train$Response, alpha =1)

lasso_1= predict(lasso, s=lam2, newx=test_mat)

mean((lasso_1 - test$Response)^2)

## [1] 106.7259
```

The test error of the lasso model fit with a lambda chosen by cross-validation is 106.7259. This error is between the least square error slightly higher but lower than the ridge regression error.

e) Comment on the results obtained. How accurately can we predict the response variable? Is there much difference among the test errors resulting from these three approaches? Present and discuss results for the approaches

The Model performance are as follows i) Linear Model using least square error is 99.11668 ii) Ridge Regression with lambda chosen by cross-validation is 224.7245 iii) Lasso model with lambda chosen by cross validation is 106.7259

Therefore lasso model performs the best, while ridge regression model performs the worst.