Stat 432 HW 04

Name: Your Name, netID: yournetID

Summer 2024

Include the R code for this HW.

```r
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(GGally)
library(tibble)
library(dplyr)
library(knitr)
library(kableExtra)
#add mroe libraries as needed.
```

Question 1 (variable selection)

This question relates to the College data set, which can be found in the file `College.csv`. It contains a number of variables for 777 different universities and colleges in the US.

(from the previous HW) Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.

Before moving on, we're not going to use the college name, so you may remove `X` variable from data.

```r
college<-read.csv(file="College.csv",header = T,stringsAsFactors = T)
# you may need to change the directory. Modify as needed.
mycollege= subset(college, select = -c(X))


set.seed(1) # for part (d), you change the random seed.
train.id=sample(1:nrow(mycollege),trunc(0.9*nrow(mycollege)))
tr.col=mycollege[train.id,] #training data
tst.col=mycollege[-train.id,] #test data

est.id=sample(1:nrow(tr.col),trunc(0.9*nrow(tr.col)))
est.col=tr.col[est.id,] #estimation data
val.col=tr.col[-est.id,] #validation data
```

Enroll is your response variable.

(a) Using `regsubsets()` from the `leaps` package, perform the best subset selection on the estimation data. Report your model choice based on Cp criteria.

(b) Using `step()` function from the same package, perform the backward selection on the estimation data. Report your model choice based on BIC criteria.

(c) Between the models chosen from (a)-(b), which model gives lowest validation error?

(d) Using different random seeds (2,3), and see if your answer in (c) changes.

Question 2 (k-fold CV theory)

(a) Explain how k-fold cross validation is implemented.

(b) Discuss the advantages and disadvantages of k-fold cross-validation relative to:

- The validation set approach
- LOOCV

Question 3 (k-fold CV application)

In this question, we will conduct regression using `Boston` data from the `ISLR2` package.

```r
set.seed(432)
trn.idx=sample(1:nrow(ISLR2::Boston),450)
tst.boston=ISLR2::Boston[-trn.idx,] #test data
trn.boston=ISLR2::Boston[trn.idx,] #training data
```

`nox` variable is your response variable. `dis` variable is your (only) predictor variable for this question.

(a) Use the `poly()` function to fit polynomials of degree 1 to 10. Using 10-fold Cross-validation, select the optimal degree for the polynomial and explain your reason. (This step should be done to the training data.)

(b) Report test MSE of your chosen model.

Question 4 (Regularization)

This question relates to the `Boston` data set of `ISLR2` package.

```r
set.seed(432)
trn.idx=sample(1:nrow(ISLR2::Boston),450)
tst.boston=ISLR2::Boston[-trn.idx,]
trn.boston=ISLR2::Boston[trn.idx,]
```

We are splitting the data into two parts: a testing data that contains 56 observations, and the rest 450 observations as training data.

- The goal is to model `crim` (our response variable) with all the other variables in the data.

(a) Conduct linear regression with 10-fold CV. Report CV error for the chosen parameter. (RMSE or MSE, either way is ok. Just need to be consistent throughout this problem.) You may use:

(b) Conduct ridge regression with 10-fold CV. Find best tuning parameter for ridge regression, and report CV error. You need to decide your own grid of λ values.

(c) Conduct Lasso regression with 10-fold CV. Find best tuning parameter for each lasso, and report CV error.

(d) Based on (a)-(c), pick the best method and train your whole training data set using the chosen method and tuning parameter(s). Report the test MSE.