Homework 4

Due: end of day Saturday, February 11

Submission instructions: Submit one write-up per group on gradescope.com.

IMPORTANT:

- Write names of everyone that worked on the assignment on the submission.
- Specify **every** member of the group when submitting on Gradescope (https://help.gradescope.com/article/m5qz2xsnjy-student-add-group-members)

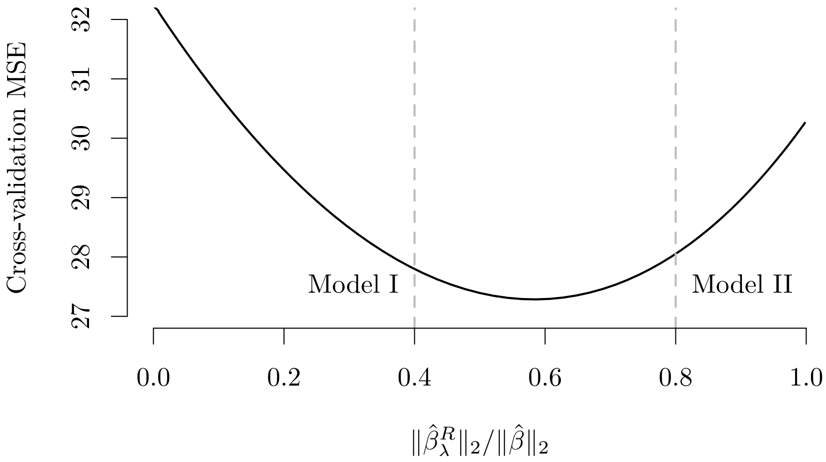
Question 1

For each statement write if it is true or false and provide a one sentence explanation. You will get points only if your explanation is correct.

1. Subset selection hierarchy:

   a. The predictors in the k-variable model identified by *forward stepwise* are a subset of the predictors in the $(k+1)$-variable model identified by *forward stepwise* selection.
   b. The predictors in the k-variable model identified by *backward stepwise* are a subset of the predictors in the $(k+1)$-variable model identified by *backward stepwise* selection.
   c. The predictors in the k-variable model identified by *best subset* are a subset of the predictors in the $(k+1)$-variable model identified by *best subset* selection.
   d. The predictors in the k-variable model identified by *best subset* are a subset of the predictors in the $(k+1)$-variable model identified by *forward stepwise* selection.
   e. The 1-variable model identified by *best subset* is the same as identified by *forward stepwise* selection.
   f. The 1-variable model identified by *backward stepwise* is the same as identified by *forward stepwise* selection.

2. The following plot shows the cross validation scores for a ridge regression model with different values of the tuning parameter λ. Note that here what is plotted as the x-axis is the ratio of the ℓ_2 norms of the ridge regression estimate $\hat{\beta}^R_\lambda$ and the least squares estimate $\hat{\beta}$. The two dashed lines correspond to two values of λ, and hence two models.



   a. Model I uses a larger λ than Model II.
   b. Model I has a higher bias than Model II.
   c. Model I has a larger in-sample RSS than Model II.
   d. Model I may have a larger test error than Model II.

Question 2

You can download files for this assignment from
http://www.kaggle.com/competitions/busn41204-winter23-homework-4

Filename	Description
`housing_train.csv`	contains data that you will use to build your model
`housing_test.csv`	contains data for which you will make predictions using your model
`sampleSubmission.csv`	example submission in the correct format
`DataDocumentation.txt`	a brief description of the variables

The data set contains prices of residential homes in Ames, Iowa. The data were cleaned and the categorical variables were converted to indicators.

Load the data using:

```
housing.train = read.csv('housing_train.csv')
housing.test = read.csv('housing_test.csv')
```

1. Create

    - a histogram of `SalePrice`;
    - a histogram of `log(Sale_Price)`;
    - a scatterplot between `SalePrice` and `Gr_Liv_Area`, the above ground living area square feet;
    - a scatterplot between `log(Sale_Price)` and `Gr_Liv_Area`.

    Comment on what you observe. Which one do you think would be more appropriate as the response of a linear regression model, `SalePrice` or `log(Sale_Price)` ?

2. Fit a linear regression on `log(Sale_Price)` using forward stepwise selection.

    - Create a plot displaying the cross validation score as a function of the number of predictors.
    - Create a plot displaying BIC scores as a function of the number of predictors.
    - How many variables get selected in the chosen model :
        - by cross validation;
        - by 1-sd rule;
        - by BIC.

3. Fit a linear regression on `log(Sale_Price)` using the lasso.

    - Create a plot displaying the cross validation score as a function of λ.
    - Create a plot displaying the coefficient values as a function of λ.
    - How many variables get selected in the chosen model:
        - by cross validation;
        - by 1-sd rule.

4. Comment on the 5 models found in parts 2 and 3. Which one seems most promising for out-of-sample prediction?

5. Build a model to predict the final sale prices in `housing_test.csv`.

    This does not need to be a linear model.

    Save your predictions to a `.csv` file that you will submit to Kaggle (see Kaggle instruction below.) Provide a write-up that explains how you went about building your model. Attach the code to create the submission `.csv` file as an appendix to your homework submission.

A sample script for generating a valid `.csv` file for submission is given below. This code builds a simple linear model for predicting `log(Sale_Price)` using all the available variables. You will easily do better than this.

```
lm.fit = lm(log(Sale_Price) ~ . , as.data.frame(housing.train))
yhat = exp(predict(lm.fit, as.data.frame(housing.test)))

#sampleSubmission is a dataframe with columns 'Id' and 'count'
sampleSubmission = data.frame(Id=1:length(yhat), Sale_Price=yhat)
write.csv(sampleSubmission,
          file = "sampleSubmission.csv",
          row.names = FALSE,
          quote = FALSE)
```

Kaggle Instructions

The link to the competition: http://www.kaggle.com/competitions/busn41204-winter23-homework-4

Submission files should contain two columns with headers: `Id` and `Sale_Price`. There should be 876 entries of predicted values in total. See code above that generates a valid submission file `sampleSubmission.csv` for reference.

Everyone needs to sign up for the competition before forming the team. Once you have signed up, you can find your teammates and form the team on the Team section of the page.

You can submit up to 5 times everyday and check your relative performance on the public leaderboard.

However, **the final ranking** will be based on the private leaderboard, which will not be released until the end of the competition.

Evaluation

You will receive points based on:

- your write-up;
- whether we can create your submission based on the code you provided; and
- your relative ranking in the class.

Submission

You need to submit two things:

- a write-up to gradescope
  - includes the name of your team
  - includes code to generate your submission file as an appendix;
- a valid submission to Kaggle.