5303 Project 3

Due by Friday Oct. 18th, 2024

Note: You need to submit two files for this assignment, a pdf report, and the original code file in R. Missing the original code file will have 50% reduction of your score.

Useful Libraries

pracma, matlib, MASS, ggthemes

Q1: Load Boston house price data in the 'Mass' package with code

For R:

```
library("Mass")
data<-Boston.
```

Note you may need to install the Mass package.

Below is some information about the dataset. Read it and use it in your questions.

The **Boston Housing dataset** contains data about housing in the Boston area, and it is often used for regression tasks.

The goal is to predict the **median value of owner-occupied homes (MEDV)** using features like per capita crime rate, pupil-teacher ratio, and property tax rate. MEDV is the label.

Features:

1. **CRIM**: per capita crime rate by town
2. **ZN**: proportion of residential land zoned for lots over 25,000 sq. ft.
3. **INDUS**: proportion of non-retail business acres per town
4. **CHAS**: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. **NOX**: nitrogen oxides concentration (parts per 10 million)
6. **RM**: average number of rooms per dwelling
7. **AGE**: proportion of owner-occupied units built prior to 1940
8. **DIS**: weighted distances to five Boston employment centers
9. **RAD**: index of accessibility to radial highways
10. **TAX**: full-value property tax rate per $10,000
11. **PTRATIO**: pupil-teacher ratio by town
12. **B**: 1000(Bk - 0.63)^2 where Bk is the proportion of Black residents by town
13. **LSTAT**: percentage of lower status of the population
14. **MEDV**: median value of owner-occupied homes in $1000's

Then answer the following questions.

a) Print a summary of the data you have.
b) Draw some plots for certain variables you choose. Histograms and QQ plots are necessary for numerical variables, and boxplots for categorical

variables. Make some interpretations about the plots you get. Please list the numerical variables and categorical variables clearly with your explanations.

c) Take the first 200 rows of the Boston data and name it 'Boston_sample'.
d) Draw the scatter plot regarding the variable ‘medv’ and the explanatory variable you choose.
e) Build a simple linear regression with the response variable 'medv' and an explanatory variable chosen in d). Use the 'Boston_sample' data to estimate the parameters in your model and interpret the estimation of your selected explanatory variable.
f) Please show the confidence interval for the parameters you get in d) with alpha=0.05. Interpret your result.
g) Please draw the regression line you get in d).
h) Randomly pick 200 rows of the Boston data and name it by ‘Boston_sample_random’. Repeat the process from d) to g). Please compare the results you get for this random dataset with that of the previous one you got.

Q2: Use the dataset in Q1.

a) Build a multiple linear regression with the response variable 'medv' and at least three explanatory variables of your choice.
b) Use the 'Boston_sample' data obtained from c) in Q1 to estimate the parameters in your model.
c) With the output from a) and write your estimated model.
d) With the output from a) to determine which explanatory variables are useful and which ones are not. Show the reasons that which one is useful, and which one is not.
e) Build a new multiple linear regression with only the useful explanatory variables. Name this model 'fit_best'. If all your variables are useful then keep the model and name it ‘fit_best’.
f) Take the 300th row from Boston data. Apply your 'fit_best' model to this row to predict 'medv' and compute the error of your prediction. Please compare the predictions you get from the model in a) and the ‘fit_best’ model. Interpret your results.

Q3: Download Advertisting.csv data and answer the following questions.

a) Make 3 scatter plots of sales vs TV, sales vs radio and sales vs newspaper. Which predictor (TV, sales, or newspaper) do you think is the most related to sales?

b) Compute the correlations between sales vs TV, sales vs radio and sales vs newspaper. Which predictor (TV, sales, or newspaper) does the correlation suggest the strongest relation to sales? Does it match with your thought from a) and why?
c) Build a simple linear regression with the most relevant predictor you picked from a) and b). Print the model output. Make a plot of residuals vs fitted values, a Q-Q plot, and a scatter plot of the predictor vs sales with fitted line of your model. With the model output and the plots, discuss whether your model fits the data well or not. (Hint: for each plot, you need to make at least one comment as well as the model output. This is an open question.)
d) Add a quadratic term of the predictor to your model in c). Repeat the analysis in c).
e) Compare the two models from c) and d). Which one do you think is the better model and why? (This is an open question. You need to have at least 3 comments.)
f) Randomly split the data into two parts with approximately 70% vs 30%. Re-train your simple linear regression model with 70% of the data and test it on the rest 30%. Compute and compare the MSEs from training set and test set. (For this question, you need to report two MSEs and discuss why one is larger than the other.)
g) Compute the average of 5-fold cross validation MSE of your simple linear regression model.

Bonus:

The model complexity can be increased by adding higher power terms of the predictor such as 2^{nd} power, 3^{rd} power and so on to the model. Choose the best model by determining how many higher power terms is the optimal choice using 5-fold cross validation.