Assignment 2
Group Submission
IFI 8420

Part 1 (20 points)

HousePrices data set is a cross-sectional data set on house prices and other features, e.g., number of bedroom, of houses in Windsor, Ontario. The data were gathered during the summer of 1987.

Use the HousePrices data to perform the following tests using Linear Regression settings:

i. Construct a summary stat for all the variables in the HousePrices data. (provide Mean, Median, Max, Min, and Std Dev of the data for all variables) (4 points)

ii. What is the percentage of houses in the data with Driveway, Gas-Heat and Air-conditioning present? (Hint: find the mean after creating dummy variables with driveway, gasheat, and aircon variables respectively). (6 points)

iii. Construct a linear regression model to test whether number of bedrooms influence house prices. Provide a summary of the linear regression model using summary() function. (5 points)

iv. Construct a multiple linear regression model by including all variables as predictors of house prices (response variable) and observe the effect on the house prices. Provide a summary of the regression model using summary() function. (5 points)

Online quiz 2B will be related to the concepts of the above model solution. Your ability to interpret the model solution will be tested. Make sure you have the code running and solutions are available while you take the quiz. It will not be under lock-down browser.

Variable description of HousePrices data: A data frame containing 546 observations on 12 variables.
price: Sale price of a house.
lotsize: Lot size of a property in square feet.
bedrooms: Number of bedrooms.
bathrooms: Number of full bathrooms.
stories: Number of stories excluding basement.
driveway: Factor. Does the house have a driveway?
recreation: Factor. Does the house have a recreational room?
fullbase: Factor. Does the house have a full finished basement?
gasheat: Factor. Does the house use gas for hot water heating?

aircon: Factor. Is there central air conditioning?
garage: Number of garage places.
prefer: Factor. Is the house located in the preferred neighborhood of the city?

Part 2: **(20 points)**

Use the Credit data to perform the following tests using Linear Regression settings:

A. Perform the following steps: (5 points)
   i. Observe the dimension of the Credit data. (1 points)
   ii. Provide a summary stat for the variables in Credit data. (1 points)
   iii. What is the percentage of Student in the Credit data? What is the percentage of Female in the Credit data? What is the percentage of Student who are Female in the Credit data? (3 points)

B. Construct a linear regression model to test the following: (15 points)

   Test how Credit Rating and Student effect Credit Card Balance. In the same model, also test whether the effect of Credit Rating on Credit Card Balance is different for students vs. non-students.

   Provide a summary of the model.

   Online quiz 2B will be related to the concepts of the above model solution. Your ability to interpret the model solution will be tested. Make sure you have the code running and solutions are available while you take the quiz. It will not be under lock-down browser.

   You do not have to write interpretation of the model solution in the python program.

Part 3: **(20 points)**

Use the Credit data to perform the following tests using Linear Regression settings: Online quiz will be based on the results of the regressions performed below.

i. Test whether Age influence Credit Card Balance on the basis of simple linear regression. (Provide a summary of the model). (2 points)

ii. Use Age and Credit Rating as predictors of Credit Card Balance (response variable) in a multiple linear regression setting. (Provide a summary of the model). (2 points)

iii. Compare effect of Age from part (i) and (ii): **Write the explanation for part (iii) in Python code. (10 points)**

iv. Observe the distribution of Age. Construct 3 dummy variables based on Age distribution: **(6 points)**
1. Age 40 and below (Age=<40)
2. Age Between 41 to 56 (41=<Age<=56)
3. Age group over 56 (Age>56)

Construct model to observe whether credit card balance is significantly different for different age group. Consider the age group over 56 as the baseline.

Your conceptual understanding of the interpretation of the model solutions for part 3 will be tested in Quiz 2B.

Part 4: **(40 points)**

i. Download monthly price data of S&P500 and a stock of your choice for the period 01/01/2005 to 12/31/2019 (or any fifteen year period)
(2 points)

ii. Compute the monthly returns for the S&P 500 and the stock. Construct one data frame to store the return series. (2 points)

iii. Construct summary statistics, histogram, correlation matrix of the return series. (5 points)

iv. Download 3 month TBill rate from Fred. Consider the TBill data for the same sample period 01/01/2005 to 12/31/2019 (or any fifteen year period that you have chosen) . (7 points)

v. Construct a matrix of return series combining Stock, S&P500, and TBill for the sample period. Construct return series with columns of excess returns of the stock (Stock return – TBill) and S&P500 (S&P500 return – Tbill). (5 points)

vi. Construct a linear regression model with the excess returns : stock excess return being the response variable and S&P500 excess return the predictor. (2 points)

vii. Find Beta for the stock based on the model constructed. Test the null hypothesis: $H_0 : \beta_1 = 0$; what do you conclude? Draw your conclusion based on p-value.
(4 points)

viii. Obtain beta of the stock from available stock report (refer the financial website that you choose). Discuss why these two measures are same or different. (2 points)

ix. Comment on model accuracy: standard error and R-square (2 points)

x. Provide the scatter plot and the fitted line for the linear regression model. (1 points)

xi. Provide interpretation of your analysis: considering return of the stock and S&P 500 and the model constructed. (8 points)

Deliverables:

1. Submit Python code electronically in iCollege in the corresponding Assignment tab.

2. Please submit **one Python program (one file)** containing **four parts of the assignment (mark/comment so that each part is separated clearly in the program).** Python code should provide comments on each sections of the assignment the code is intended for. Also indicate the names of the contributors at the beginning of the file.

3. The assignment submission grade will be based on whether you have completed each part of the analysis and whether your code run through all the parts of analysis. Your grade will be based not only on the correctness of the program but also how efficiently the program executes the tasks.

4. Note that **you do not have to provide** interpretation of the model results for Parts 1 -3 in the code (**except for part 3 (iii)**). If you do write the responses in the program - it will not be part of your grade.

5. You have to **write interpretation of the results for Part 4 in the python code** in respective sections as needed.

6. **The quiz 2B will have questions that will test your conceptual understanding of the output/results of the models for PART 1, 2, and 3. Make sure you understand the relevant concepts of the analysis in each part before you take the online quiz.**

7. Do not submit a separate word document explaining your results.