Project 2: Regression Project

The objective of this project is to develop a linear multivariable regression to establish a relation between the dependent variable Y and a 5-tuple of independent variables X_1, X_2, X_3, X_4 and X_5. For this analysis you can use any statistical package, such as MatLab, R, SAS, or use Python with all the available statistical functions. You may also use multiple packages for different tasks.

Data Set

You will use the data set provided to you that was generated based on your student id (SI) number. The set of all data sets for all the students in the class is posted in the folder entitled "Individual data files". **Use the one that has your ID**.

In each csv file, the first five columns give the data for the five independent variables X_1, X_2, X_3, X_4 and X_5. The dependent variable Y is in the last column.

If your data set does not appear to be good, let us know and we will provide you with another one.

Task 1. Basic Statistics Analysis

1.1. For each variable X_i, i.e., column in the data set corresponding to X_i, calculate the following: histogram, mean, variance.
1.2 Use a box plot or any other function to remove outliers (do not over do it!). This can also be done during the model building phase (see tasks 2 and 3).
1.3 Calculate the correlation matrix for all variables, i.e., Y, X_1, X_2, X_3, X_4 and X_5. Draw conclusions related to possible dependencies among these variables.
1.4 Comment on your results.

Task 2: Simple Linear Regression

Before proceeding with the multivariable regression, carry out a simple linear regression to estimate the parameters of the model: $Y = a_0 + a_1X_1 + \varepsilon$.
2.1 Determine the estimates for a_0, a_1, and σ^2.
2.2 Check the p-values, R^2, and F value to determine if the regression coefficients are significant.
2.3 Plot the regression line against the data.
2.4 Do a residuals analysis:
   a. Do a Q-Q plot of the pdf of the residuals against N(0, s^2) In addition, draw the residuals histogram and carry out a χ^2 test that it follows the normal distribution N(0, s^2).
   b. Do a scatter plot of the residuals to see if there are any correlation trends.
2.7 Use a higher-order polynomial regression, i.e., $Y = a_0 + a_1X + a_2X^2 + \varepsilon$, to see if it gives better results.
2.8 Comment on your results.

Task 3. Linear Multivariable Regression

3.1 Carry out a multivariable regression on all the independent variables, and determine the values for all the coefficients, and σ^2.

3.2 Based on the *p*-values, R^2, *F* value, and correlation matrix, identify which independent variables need to be removed (if any) and go back to step 3.1.

3.3 Do a residuals analysis:

a. Do a Q-Q plot of the pdf of the residuals against N(0, s^2). In addition, draw the residuals histogram and carry out a χ^2 test that it follows the normal distribution N(0, s^2).

b. Do a scatter plot of the residuals to see if there are any trends.

3.4 Comment on your results.

What to submit

1. For each task 1, 2, and 3 submit the following:

   The code you used for the task. It does not have to run on eos, and it may be a number of different pieces of code from different packages.

   Sharing code is not allowed and constitutes cheating, in which case both students (the one that aids and the one that receives) will get a zero for the project and will be reported to the student conduct office.

2. Your results (graphs, tables, etc) and your conclusions.

   You will receive a bad grade if you submit results without substantive conclusions, or conclusions that are not backed by sufficient results.

Grading

The TA will first verify that your code works and produces the results you submit. The break down of the grades will be as follows:

Task 1: 15 points
Task 2: 35 points
Task 3: 50 points

Remember that you will be graded mostly on your ability to interpret the results