Project 2: Multiple Regression Analysis

Objectives

To analyze a multivariate dataset consisting of the six variables: (Y, X_1, X_2, X_3, X_4, X_5) in order to establish a relation between the dependent variable Y and the 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.

Data Set

You will use the data set provided to you that is completely based on your student id (SI) number. The set of all data sets for all the students in the class is posted in the schedule under the name: DataSets. **Use the one that has your ID**.

Tasks

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 Calculate the correlation matrix Σ among 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.

1.3 Comment on the results.

Task 2: Linear regression

Before proceeding with the multiple regression, you will carry out a simple linear regression to estimate the parameters of the model: $Y = a_0 + a_1X + \varepsilon$, where $X = X_1$.

2.1 Determine the values for a_0, a_1, and s^2.

2.2 Check the p-values, R^2, F value to determine if the regression coefficiants are meaningfull.

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) Alternatively, draw the residuals histogram and carry out a χ^2 test that it follows the 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 in a couple of paragraphs.

3. Multivariate regression

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

3.1 Based on the p-values, R^2, F value, and correlation matrix Σ, identify which independent variables need to be left out (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). Alternatively, draw the residuals histogram and carry out a χ^2 test that it follows the N(0, s^2).

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

Note: You can do the residual analysis for each attempted model, or just for the one that you think is the best model

What to submit

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

1. The code you used for the task (It does not have to run on eos)

   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 backed by insufficient 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