CSC 591/ECE 592 IoT Analytics - Project 3
- Regression and Forecasting

Objectives

The objective of this project is to develop a linear multi-variable regression to establish a relation between the dependent variable Y and a (3-5)-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 same data set you obtained in Project 2. If your data set **does not appear to be good** (cannot yield good fitting model), additional data samples may be collected.

Task 1. Basic Statistic 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). Here the objective is to observe the correlation.

1.3 Calculate the correlation matrix for all variables, i.e., Y, X_1, X_2, X_3, X_4 and X_5 (if you have 5 variables). Note that the correlation matrix is a matrix with correlation coefficients for a pair of variables, which can be obtained by either coding or a statistical package.

1.4 Draw conclusions related to possible dependencies among these variables. Comment on your results, that is, whether these is a strong or weak correlation (r is close to 1 or not), positive or negative correlation, etc.

Task 2. Simple Linear Regression

Before proceeding with the multi-variable regression, carry out a simple linear regression to estimate the parameters of the model: $Y = a_0 + a_1 X_1 + \epsilon$.

2.1 Determine the estimates for a_0, a_1, and σ^2.

2.2 Check the p-values, R-squared, and adjusted R-squared 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.

2.5 Use a higher-order polynomial regression, i.e., $Y = a_0 + a_1 X_1 + a_2 X_1^2 + \epsilon$ to see if it gives better results.

2.6 Comment on your results.

Task 3: Multi-variable Linear Regression

3.1 Carry out a multi-variable 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-square, adjusted R-square, 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. Submit a report with detailed procedures of your coding or the packages you use, or the AI tools that assist you for all three Tasks.

   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, snapshots, tables, etc) and your conclusions.

Grading:

- Task 1: 25 points.
- Task 2: 35 points.
- Task 3: 40 points.

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