LINEAR REGRESSION ASSIGNMENT – MURALI MOHAN R (ML C-34)

ASSIGNMENT BASED SUBJECTIVE QUESTIONS

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A. The following inferences can be drawn on analysis of Categorical variables:-

   a. Season 3 leads to significant increase in Bike Hires. Season 2 and 4 are average and Season 1 proves to be the least significant contributor towards Bike rentals.

   b. The average bike rentals are significantly higher in the second year compared to the first year. The demand tends to increase over time.

   c. The middle of the year, from April - October contribute to majority of bike rentals compared to the other months.

   d. Weather situation 1 contributes to huge number of Bike rentals, with Situation 2 also contributing a significant sum. Weather situation 3 proves to be a time during which the demand is extremely less.

2. Why is it important to use drop_first=True during dummy variable creation?

A. Whilst creating dummy variables, the number of levels that the categorical variable holds (no. of possible values) is of significant importance. If a variable has ‘n’ levels, we need to create ‘n-1’ dummy variables because ‘n-1’ variables will be able to capture the information held by the categorical variables, without any data loss. If all the dummy variables hold a 0 value, it implies that that the record belongs to the category which is omitted. It is always important to reduce the number of variables wherever possible in order to ensure that the model’s complexity is kept to minimum.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A. On observing pairplot of numerical variables, the independent variables ‘registered’ and ‘casual’ have the highest correlation. Especially, the majority of the users hiring bikes are registered users and hence the company should target to increase their registered user base. However, these two variables are mere distribution of the target variable, hence there is no surprise in observing significant correlation between them. Apart from these two independent variables, ‘temp’ and ‘atemp’ are the two independent variables that exhibit a positive linear relationship with the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A. The following analysis have been made in order to validate the assumptions of Linear Regression:-

   a. The residual terms obtained from actual and predicted values of the target variable have been plotted, and it has been observed that they follow a normal distribution with 0 as the centre.
   b. On observing Variance Inflation Factor for all the variables in the final model, it has been ascertained that there is no Multicollinearity.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

A. The three most significant variables are ‘workingday’, ‘casual’ and ‘Season4’.

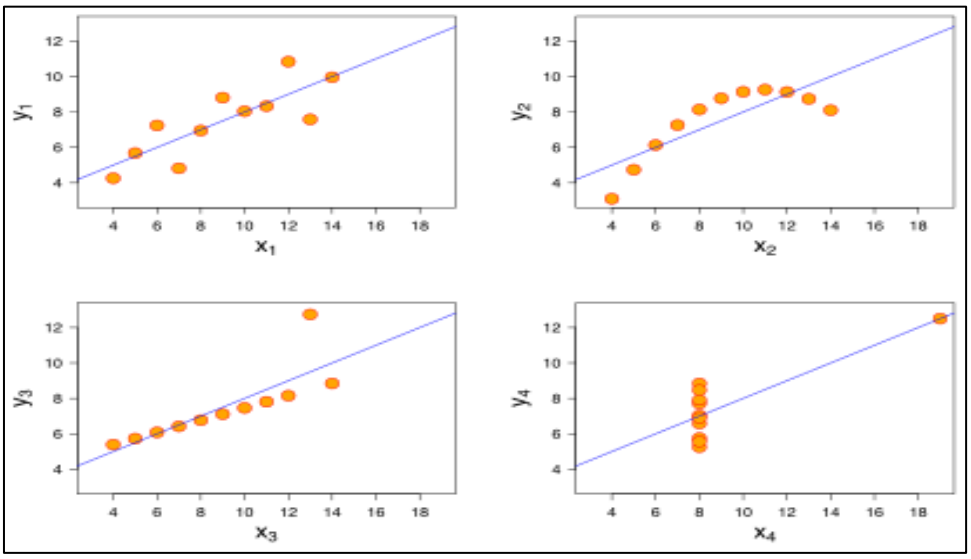
GENERAL SUBJECTIVE QUESTIONS

1. Explain the linear regression algorithm in detail.

A. The linear regression algorithm is used in prediction of a target variable on the basis of one or more independent variables where a linear relationship is present between them. There can be many factors that significantly influence the target variable and these variables play a part in general of the Linear Regression model. All the relevant variables are associated with coefficients on the basis on their importance in predicting the target variable.

2. Explain the Anscombe’s quartet in detail.

A. Ansombe’s Quartet comprises of four different datasets (each dataset comprising of two variables ‘x’ and ‘y’) that have the same descriptive summary statistics (Mean and Standard Deviation / Variance) for respective ‘x’ and ‘y’ variables in all four datasets. In addition, the parameters such as Correlation, Regression Line and R Square are also same for all four datasets. However, when these datasets are plotted in the two dimensional graph, they are completely different. The visualization is as follows:-



3. What is Pearson's R?

A. Pearson's R also known as Correlation Coefficient is a measure of linear correlation between the independent variables. It is the ratio between the covariance of two variables and the product of their standard deviations; thus it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1. Any value closer to 1 indicates a positive linear relationship and a value closer to -1 indicates a negative linear relationship. A value closer to 0 indicates that there is no linear relationship between the variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

A. Scaling is a measure in which a variable's values are scaled to a particular range. Scaling is important because when there are more than one independent variables, the variables can have different ranges where one variable ranges from 1 Cr. to 50 Cr, and another variable may range between 1 and 35. In such cases, the coefficients associated with these variables as part of regression, will swing wildly and in turn make the analysis of the variables difficult thereby misleading the decision making process.
Normalized scaling is a method in which the values of a variable are scaled between 0 and 1, with 0 representing the Min value and 1 representing the Max value. In Standardized Scaling, the values are centred around 0, with 0 representing the mean value and the values distributed with a unit standard distribution.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A. In case of an infinite value for VIF, we may infer that the Multicollinearity is extremely high. It also means that there is perfect correlation between the independent variables used in regression.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A. A Quantile-Quantile plot is used to determine whether the sample data that is used, has come from a normal distribution. From the perspective of linear regression where we have two sets of data namely training and test datasets, in case if we have received these two datasets separately, we can ascertain that these datasets have been derived from a common population with same distribution, with the help of Q-Q plot.