Linear Regression Assignment Subjective Questions

1. What are the assumptions of linear regression regarding residuals?

Answer –

Assumptions of linear regression regarding residuals are:

a.) It is assumed that error terms are normally distributed.

b.) It is assumed that residuals have mean equal to zero.

c.) It is assumed that residual terms have same variance.

d.) It is assumed that residual terms are independent of each other.

2. What is the coefficient of correlation and the coefficient of determination?

Answer-

Coefficient of correlation:-

Coefficient of correlation is the measure of the strength between two variables. It measure the relationship between the two variables. The most common Coefficient of correlation is the Pearson's Coefficient of correlation(R). Its value range between -1 and 1 The positive correlation means that if the value of one variable increases so will the other and vice-a-versa. And negative correlation means that if the value of one variable decreases the other one increases and vice-a-versa. And higher the value of correlation stronger is the correlation.

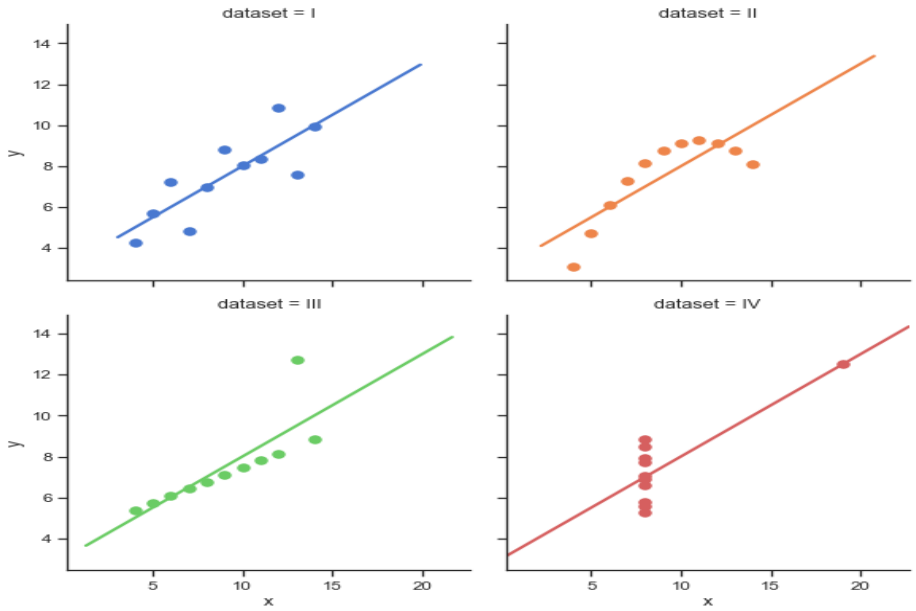
Coefficient of determination:-

Coefficient of determination is the square of the coefficient of correlation (R-Squared). It is a statistical measure of how close the regression line fitted into the data. It tells us how much the data variance can be explained by the regression line. e.g., if R-squared is 0.80, it means that 80% of the variance in the data can be explained by the regression line.

3. Explain the Anscombe's quartet in detail.

Answer-

Anscombe's quartet is developed by statistician Francis Anscombe. It contains four data set, each having 11 pairs of (x,y). Each of the four data set shows similar summary statistics, that if we plot a linear regression each of the four data set we will get same regression line, but when we plot the data points, we will see the data points are highly different.



4. What is Pearson's R?

Answer-

Pearson's R is the Pearson's correlation coefficient, it is the measure of strength of linear relationship between two variables. Pearson's correlation coefficient is represented by p when measured for population and it is represented by R when measured for sample. Its value lies in between -1 and 1. When its value is positive then it means that the variables is positively correlated i.e., if the value of one variable increases so does the other and vice-a-versa, and when its value is negative then it means that variables are negatively correlated i.e., if value of one variable increases the other one decreases and vice-a-versa. And higher is the value more correlated the variables are.

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

Answer-

Scaling is a method used to normalize the data. It is generally performed in data processing stage. In scaling we bring all the data in a particular range which increases the speed of algorithm calculations. Because in the data we have a lot of variables which are in different range and that slows down the algorithm and its also difficult for the interpretation when the variables have high difference in their range, and scaling solves those problems.

Normalized scaling:-

Normalized scaling also known as Min-Max Scaling brings all the data in range of 0 to 1.

The formula used in Normalized scaling:-

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardized Scaling:-

Standardized scaling brings all the data into a standard normal distribution with mean equal to 0 and standard deviation equal to 1.

The formula used in Standardized scaling :-

$$z = \frac{x_i - \mu}{\sigma}$$

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

Answer-

When VIF is infinite that means that we have perfect collinearity, that is we have redundant variables. This happens when we have variables which can be completely explained by other variables. Since VIF = 1/(1-R^2), when VIF is infinity that means that the R = 1 for that variable, i.e., that variable is redundant that is it can be fully explained by other variables, and we should drop that variable.