

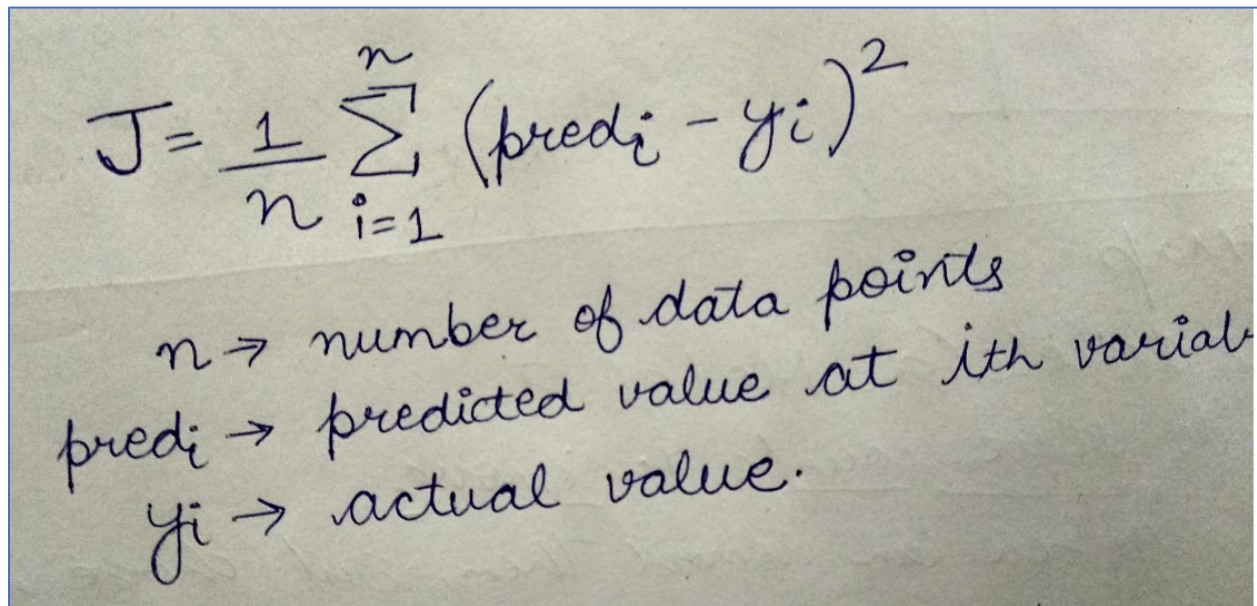
Question 1: Explain the linear regression algorithm in detail?

Answer:

When the output variable to be predicted is a continuous variable, we employ Regression techniques. As name suggests, Linear regression is one of the regression techniques based on supervised learning. It is generally classified into two types:

- Simple Linear Regression
- Multiple Linear Regression

In Linear Regression algorithm, we try to minimize the cost functions to provide the best fit line for the data points.



The image shows a handwritten formula for the Cost Function J on a piece of paper. The formula is $J = \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$. Below the formula, there are three lines of handwritten text explaining the variables: $n \rightarrow$ number of data points, $\text{pred}_i \rightarrow$ predicted value at i th variable, and $y_i \rightarrow$ actual value.

Cost Function

The difference between the predicted values and actual values makes up the error difference. Error differences are squared and sum over all data points and then is divided by the total number of data points. This calculation provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error (MSE) function.

We employ Gradient Descent for finding the minima of a graph. If we take smaller step a time, then we would reach the minima of the graph in a longer time. If we take larger steps each time, we would reach early but there is a possibility that we might overshoot the minima of the graph. In the gradient descent algorithm, the number of steps is called the learning rate.

Question 2: What are the assumptions of linear regression regarding residuals?

Answer:

Assumption 1: The mean of the residuals is always equal to 0.

Assumption 2: The sum of the residuals is always equal to 0.

Error terms are normally distributed and have constant variance.

Question 3: What is the coefficient of correlation and the coefficient of determination?

Answer:

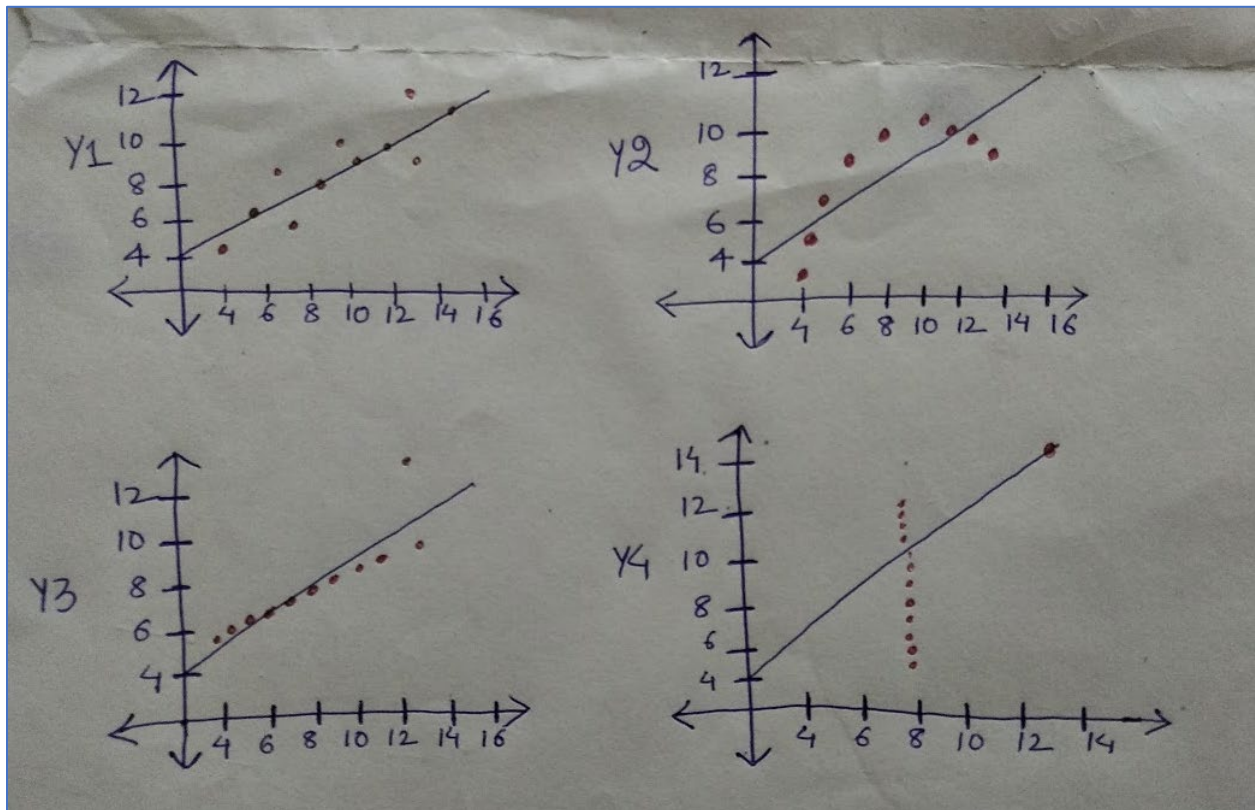
Coefficient of Correlation: It is a numerical measure of the strength of the relationship between the movements of two variables.

Coefficient of Determination: It is also known as "R-squared". It is a numerical measure which assesses how much variability of the dataset is explained by the model.

Question 4: Explain the Anscombe's quartet in detail.

Answer:

It consists of four data sets that have nearly identical simple descriptive statistics, but when they are graphed, then they have very different distributions. Each dataset consists of eleven (x, y) points particularly. It shows the importance of graphing data even though statistical parameters remain same before analyzing it and the effect of outliers and other influential observations on statistical properties.



Question 5: What is Pearson's R?

Answer:

It is a measure of the strength and direction of the linear relationship between two variables.

Question 6: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

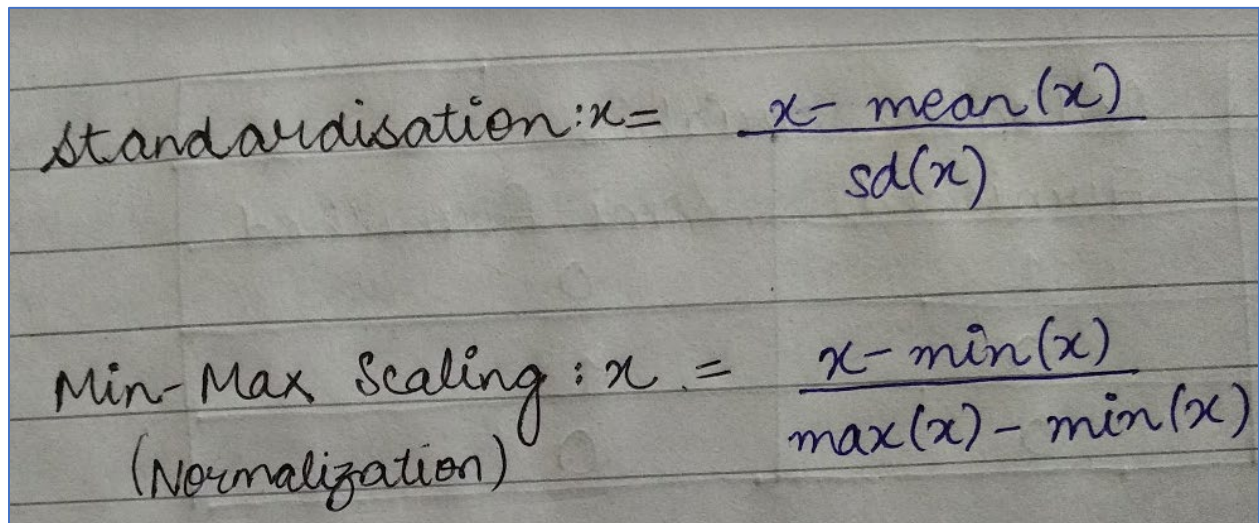
Answer:

Scaling: It is a process done to standardize the data, before we model a data. It ensures that all variables are in same scale, so that it is easy to read the coefficients of model afterwards.

Why Scale:

1. It helps with interpretation of the model.
2. Faster convergence of gradient descent.

Difference between normalized and standardized scaling:



The image shows two handwritten formulas on lined paper. The first formula is for standardization:
$$\text{standardisation: } x = \frac{x - \text{mean}(x)}{\text{sd}(x)}$$
 The second formula is for min-max scaling (normalization):
$$\text{Min-Max Scaling: } x = \frac{x - \min(x)}{\max(x) - \min(x)}$$
 Below the second formula, the word "(Normalization)" is written in parentheses.

Normalization usually means to scale a feature to have a value between 0 and 1, while standardization transforms data to have a mean of zero and a standard deviation of 1. It is also called a z-score.

Question 7: You might have observed that sometimes the value of VIF is infinite. Why does this happen?

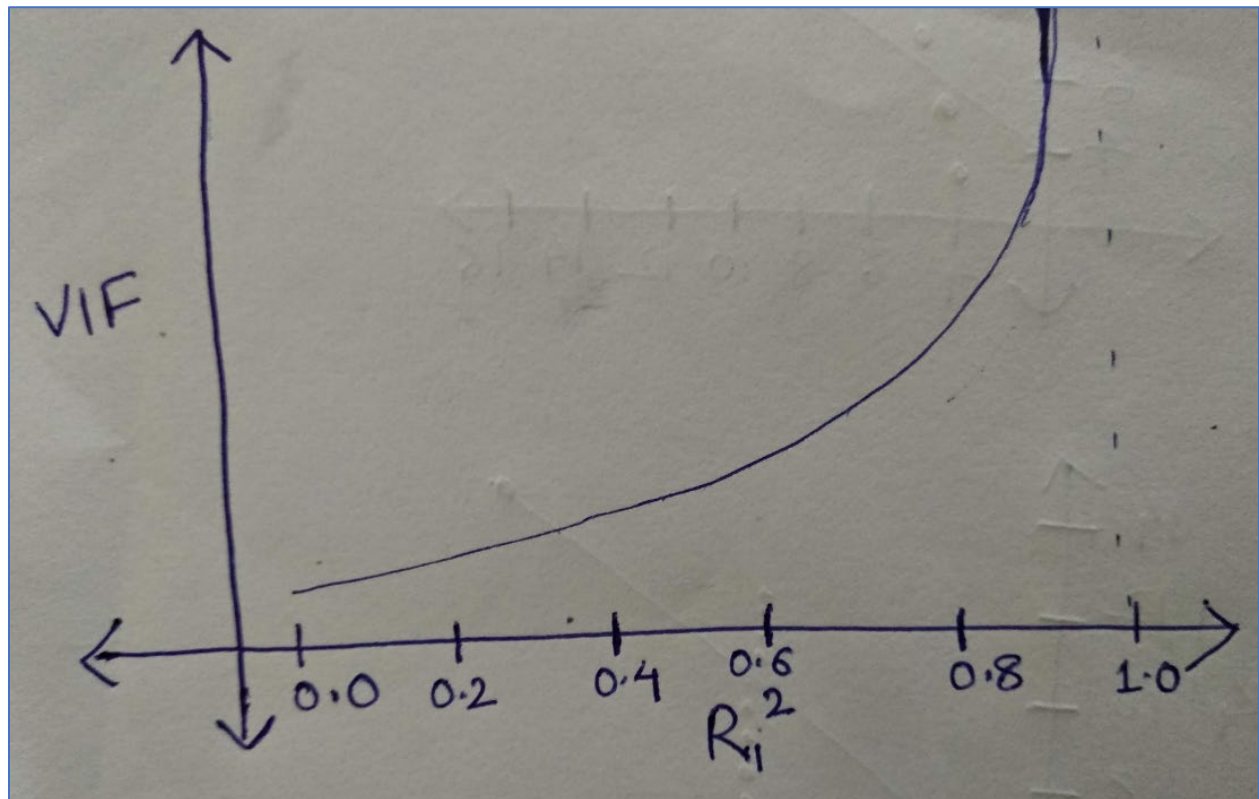
Answer:

Variance Inflation Factor (VIF): Pairwise correlations may not be enough to identify all multicollinearity variables. A variable can be explained by multiple variables together. So, to identify such variables, VIF is required.

Variation Inflation Factor:

$$VIF_i = \frac{1}{1 - R_i^2}$$

common heuristic: while >10 is definitely high;
 >5 shouldn't be ignored



As R^2 (parameter which explains variability) approaches 1 for a model, then VIF approaches infinity.

Question 8: What is the Gauss-Markov theorem?

Answer:

Gauss-Markov theorem states that in a linear regression model in which the errors are uncorrelated, have equal variances and expectation value of zero, the best linear unbiased estimator (BLUE) of the coefficients is given by the ordinary least squares (OLS) estimator, if it exists. "Best" means giving the lowest variance of the estimate, as compared to other unbiased, linear estimators. The errors do not need to be normal, nor do they need to be independent and identically distributed.

Question 9: Explain the gradient descent algorithm in detail.

Answer:

Gradient descent is an optimization algorithm used to find the values of parameters (coefficients) of a function (f) that minimizes a cost function (cost).

The size of these steps to achieve optimization is called the learning rate.

The image shows handwritten notes on lined paper. The first line defines the hypothesis as $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$. The second line lists the parameters as $\theta_0, \theta_1, \dots, \theta_n$. The third line is the heading Cost function. The fourth line shows the cost function formula $J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$. Below this, an arrow points from the parameters in the formula to the text 'Refer to this as θ vector'. The fifth line shows the simplified formula $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$.

Hypothesis: $h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$

Parameters: $\theta_0, \theta_1, \dots, \theta_n$

Cost function

$$J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

↓ Refer to this as θ vector

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

Gradient Descent:

Repeat = L

$$\rightarrow \theta_j := \theta_j - \alpha \frac{\partial J(\theta_0, \dots, \theta_n)}{\partial \theta_j}$$

(simultaneously update for every $j = 0, \dots, n$)

$$\rightarrow \theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

Question 10: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Answer:

The Q-Q plot, or quantile-quantile plot, is a visualization tool which assesses if a set of data came from some theoretical distribution such as a Normal or exponential. For example, if statistical analysis is run that assumes dependent variable is Normally distributed, Normal Q-Q plot can be used to check that assumption. It's just a visual check.

Usage, and Importance of QQ plot in linear regression:

Fit a linear regression model and note down the residuals. Since Linear Regression assumption is that residuals are normally distributed, we can use QQ plot to verify the same. Q-Q plots checks that the data meet the assumption of normality. It compares the distribution of data to a normal distribution by plotting the quartiles of data against the quartiles of a normal distribution.