General Subjective Questions and Answer

1. Explain the linear regression algorithm in detail.

Ans: **Linear Regression Algorithm** is a machine learning **algorithm** based on supervised learning. **Linear regression** apart **regression** analysis. **Regression** analysis is a technique of predictive modelling that helps you to find out the relationship between Input and the target variable.

Regression analysis is used for three types of applications:

1. Finding out the effect of Input variables on Target variable.
2. Finding out the change in Target variable with respect to one or more input variable.
3. To find out upcoming trends.

Linear regression is one of the very basic forms of machine learning where we train a model to predict the behaviour of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.

Example for that can be let's say you are running a sales promotion and expecting a certain number of count of customers to be increased now what you can do is you can look the previous promotions and plot if over on the chart when you run it and then try to see whether there is an increment into the number of customers whenever you rate the promotions and with the help of the previous historical data you try to figure it out or you try to estimate what will be the count or what will be the estimated count for my current promotion this will give you an idea to do the planning in a much better way about how many numbers of stalls maybe you need or how many increase number of

employees you need to serve the customer. Here the idea is to estimate the future value based on the historical data by learning the behaviour or patterns from the historical data.

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward relationship.

One example for that could be that the police department is running a campaign to reduce the number of robberies, in this case, the graph will be linearly downward.

Linear regression is used to predict a quantitative response Y from the predictor variable X.

Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a\,(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line.

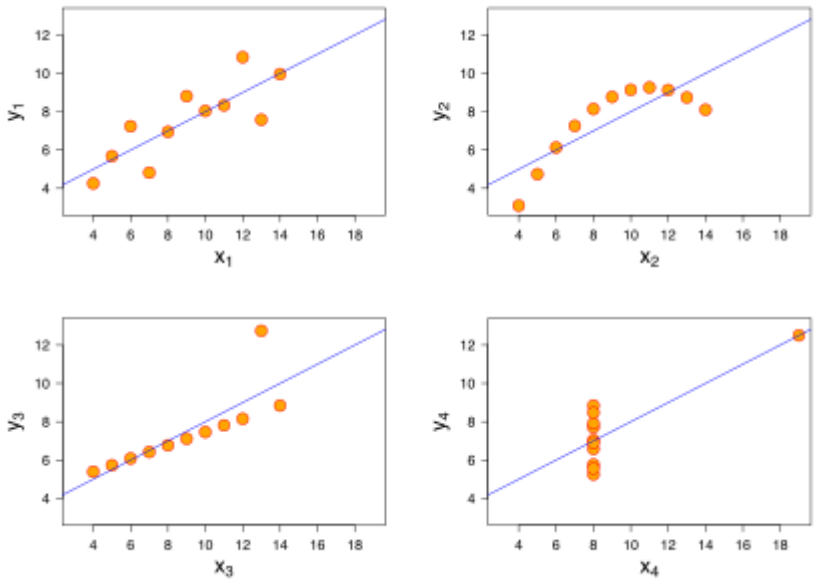
a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

2.Explain Anscombe’s quartet in detail.

Ans: **Anscombe's quartet** comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (*x*,*y*) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analysing it and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough.



For all four datasets:

Property	Value	Accuracy
Mean of *x*	9	exact
Sample variance of *x* :	11	exact

Mean of *y*	7.50	to 2 decimal places
Sample variance of *y*:	4.125	±0.003
Correlation between *x* and *y*	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression:	0.67	to 2 decimal places

1.The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated where y could be modelled as gaussian with mean linearly dependent on x.

2.The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation coefficient is not relevant. A more general regression and the corresponding coefficient of determination would be more appropriate.

3.In the third graph (bottom left), the distribution is linear, but should have a different regression line (a robust regression would have been called for). The calculated

regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

4.Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.

The quartet is still often used to illustrate the importance of looking at a set of data graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.

The datasets are as follows. The *x* values are the same for the first three datasets.

3.What is Pearson's R?

Ans: In statistics, the **Pearson correlation coefficient** (**PCC**, pronounced also referred to as **Pearson's *r***, the **Pearson product-moment correlation coefficient** (**PPMCC**), or the **bivariate correlation**, is a statistic that measures linear correlation between two variables *X* and *Y*. It has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

Naming and history

It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844. The naming of the coefficient is thus an example of Stigler's Law.

Definition

Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

The Pearson's correlation coefficient varies between -1 and +1 where:

r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
r = 0 means there is no linear association
r > 0 < 5 means there is a weak association
r > 5 < 8 means there is a moderate association
r > 8 means there is a strong association

The figure below shows some data sets and their correlation coefficients. The first data set has an r=0.996, the second has an r = -0.999 and the third has an r= -0.233

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

Ans: Scaling is the procedure of measuring and assigning the objects to the numbers according to the specified rules. In other words, the process of locating the measured objects on the continuum, a continuous sequence of numbers to which the objects are assigned is called as **scaling**.

It is **performed** during the data pre-processing to handle highly varying magnitudes or values or units. If feature **scaling** is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

In **regression**, it is often recommended to center the variables so that the predictors have mean 0. ... Another practical reason for **scaling** in **regression** is when one variable has a very large **scale**, e.g. if you were using population size of a country as a predictor.

The terms **normalization** and **standardization** are sometimes used interchangeably, but they usually refer to different things. **Normalization** usually means to scale a variable to have a value **between** 0 and 1, while **standardization** transforms data to have a mean of zero and a standard deviation of 1.

Normalization is useful when your data has varying scales and the algorithm you are using does not make assumptions about the distribution of your data, such as k-nearest neighbours and artificial neural networks. **Standardization** assumes that your data has a Gaussian (bell curve) distribution.

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

Ans: In statistics, the variance inflation factor (**VIF**) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least square's regression analysis.

In general, **one** starts with the selection of all variables, and proceeds by repeatedly deselecting variables showing a

high **VIF**. ... An **infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear

combination of other variables (which show an **infinite VIF** as well).

It is calculated by taking the the ratio of the variance of all a given model's betas divide by the variane of a single beta if it were fit alone. In this way, why is **Vif infinite**? If there is perfect correlation, then **VIF** = **infinity**. A large **value of VIF** indicates that there is a correlation between the variables.

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

Ans: In statistics, a **Q–Q (quantile-quantile) plot** is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.[1] First, the set of intervals for the quantiles is chosen. A point (*x*, *y*) on the plot corresponds to one of the quantiles of the second distribution (*y*-coordinate) plotted against the same quantile of the first distribution (*x*-coordinate). Thus, the line is a parametric curve with the parameter which is the number of the interval for the quantile.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line $y = x$. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to

compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions. A Q–Q plot is generally a more powerful approach to do this than the common technique of comparing histograms of the two samples, but requires more skill to interpret. Q–Q plots are commonly used to compare a data set to a theoretical model.[2][3] This can provide an assessment of "goodness of fit" that is graphical, rather than reducing to a numerical summary. Q–Q plots are also used to compare two theoretical distributions to each other.[4] Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.

The term "probability plot" sometimes refers specifically to a Q–Q plot, sometimes to a more general class of plots, and sometimes to the less commonly used P–P plot. The probability plot correlation coefficient plot (PPCC plot) is a quantity derived from the idea of Q–Q plots, which measures the agreement of a fitted distribution with observed data and which is sometimes used as a means of fitting a distribution to data.

The purpose of **Q Q plots** is to find out if two sets of data come from the same distribution. A 45-degree angle is plotted on the **Q Q plot**; if the two data sets come from a common distribution, the points will fall on that reference line. ... It's being compared to a set of data on the y-axis.

Few advantages:

a) It can be used with sample sizes also

b) Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

It is used to check following scenarios:

If two data sets —

i. come from populations with a common distribution

ii. have common location and scale

iii. have similar distributional shapes

iv. have similar tail behavior.

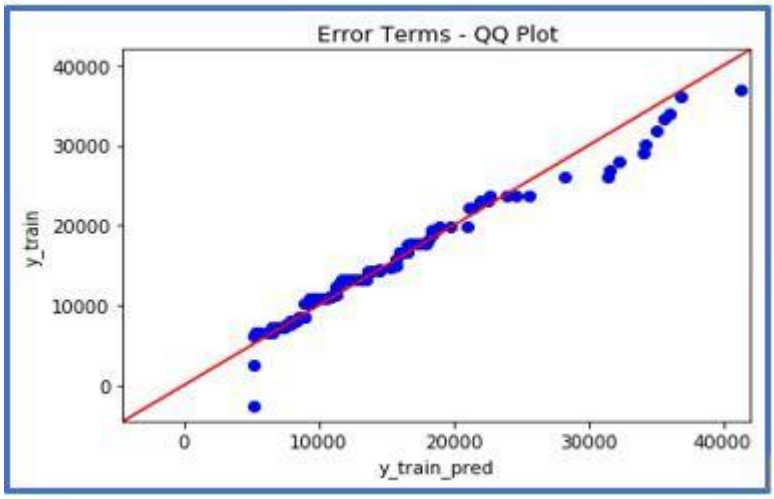
Interpretation:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

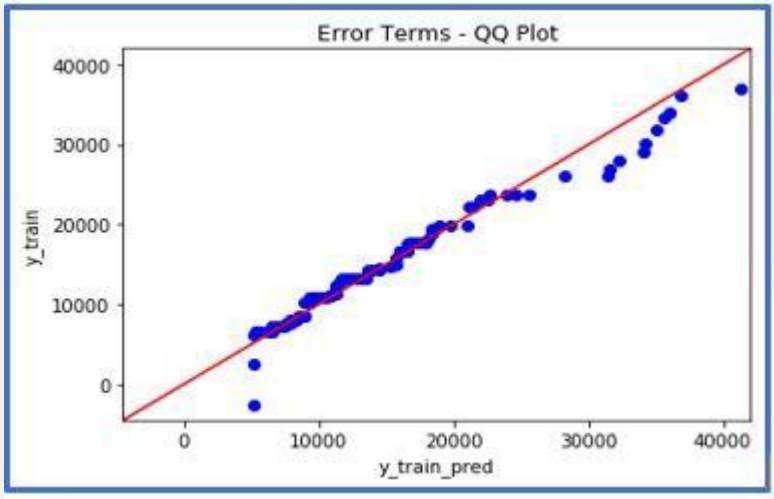
Below are the possible interpretations for two data sets.

a) **Similar distribution**: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x -axis

b) **Y-values < X-values:** If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) **Different distribution:** If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis