1. Explain the linear regression algorithm in detail.

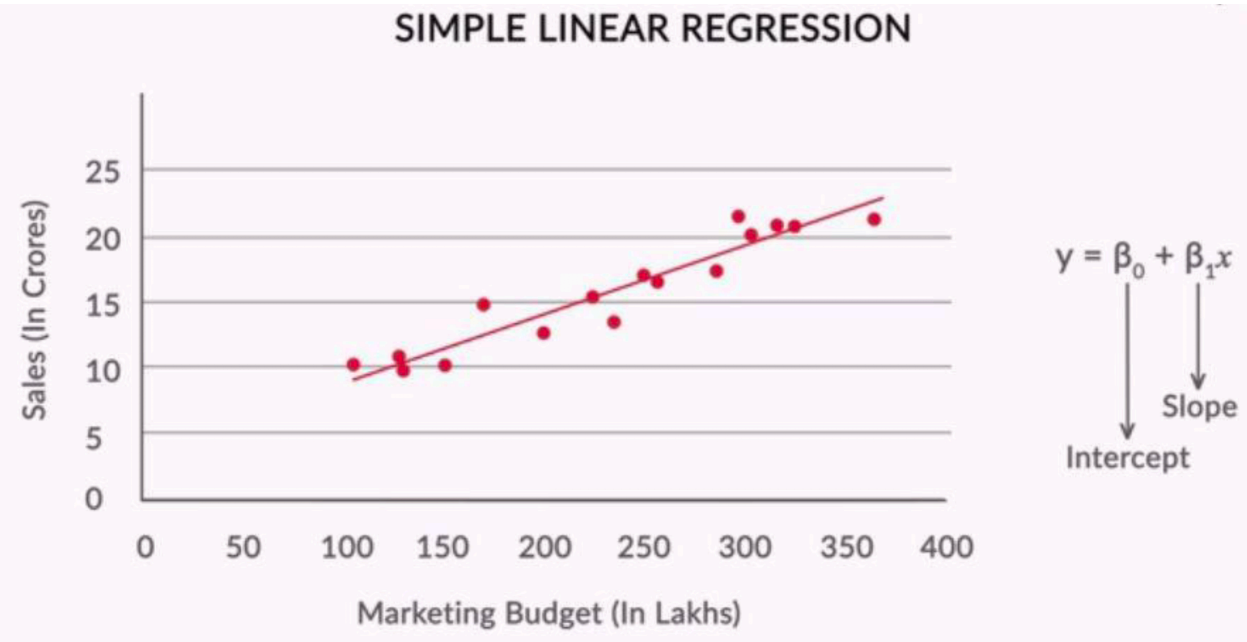
Linear Regression algorithm is a supervised machine learning algorithm where the result is predicted by the use of known parameters which are correlated with the output. It is used to predict values within a continuous range rather than trying to classify them into categories. The known parameters are used to make a continuous and constant slope which is used to predict the unknown or the result. So, It is a linear approach to modelling the relationship between a scalar response (or dependent variable or y) and one or more explanatory variables (or independent variables or x).

Broadly Linear regression model is classified into 2 categories:

- Simple linear regression
- Multiple linear regression

1. Simple linear regression

Simple linear regression explains the relationship between a dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.



The standard equation of the regression line is given by the following expression: $Y = \beta_0 + \beta_1 X$

2. Multiple linear regression

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

The formula now can be simply given as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Key-Points

1. Model now fits a 'hyperplane' instead of a line
2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)
3. For inference, the assumptions from Simple Linear Regression still hold- Zero mean, independent, Normally distributed error terms that have constant

Key Attributes

Overfitting: When we add more and more variables, for example, let's say we keep on increasing the degree of the polynomial function fitting the data, the model might end up memorizing all the data points in the training set. This will cause major problems with generalisation, i.e. now when the model runs on the test data, the accuracy will drop tremendously since, it doesn't generalise well. This is a classical symptom of overfitting.
Multicollinearity: Multicollinearity is the effect of having related predictors in the multiple linear regression model. In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated.

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

1. The mean of residuals is zero

Check the mean of the residuals. If it zero (or very close), then this assumption is held true for that model. This is default unless you explicitly make amends, such as setting the intercept term to zero

2. Homoscedasticity of residuals or equal variance

All the random variables have same finite variances.

3. No autocorrelation of residuals

When the residuals are autocorrelated, it means that the current value is dependent of the previous (historic) values and that there is a definite unexplained pattern in the Y variable that shows up in the disturbances.

4. The X variables and residuals are uncorrelated

Do a correlation test on the X variable and the residuals. p-value is high, so null hypothesis that true correlation is 0 can't be rejected. So, the assumption holds true for this model.

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

Coefficient of correlation is used to measure the linear relationship between two variables, while **Coefficient of determination** (R-squared) is used to measure the explained variation. Below is the detailed explanations of each of the terms.

a) Coefficient of correlation

• The quantity r, called the linear correlation coefficient, measures the strength and the direction of a linear relationship between two variables.

• The value of r is such that $-1 < r < +1$.

• **Positive correlation:** If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

• **Negative correlation:** If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

• **No correlation:** If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

• A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If r = +1, the slope of this line is positive. If r = -1, the slope of this line is negative.

• **Heuristics:** A correlation greater than 0.8 is generally described as strong, whereas a correlation less than 0.5 is generally described as weak.

b) Coefficient of determination

• The coefficient of determination, r2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

• The coefficient of determination is the ratio of the explained variation to the total variation.

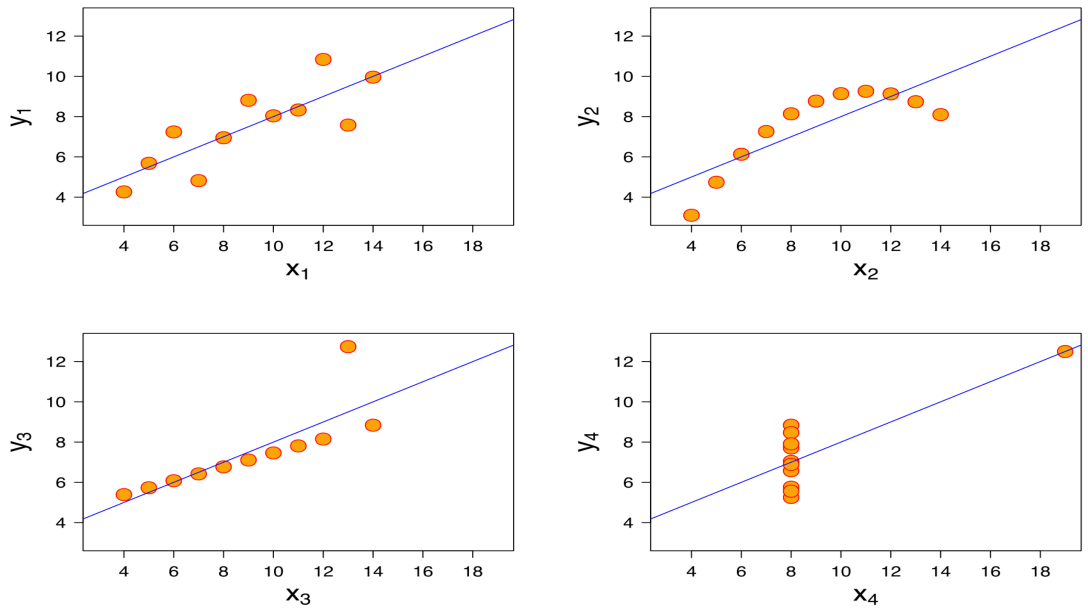
• The coefficient of determination is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

• The coefficient of determination is such that $0 < r2 < 1$, and denotes the strength of the linear association between x and y.

• The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

4. Explain the Anscombe's quartet in detail.

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.



If we visualize the above diagram, we see following insights:

- Dataset 1 appears to have clean and well-fitting linear models.
- Dataset 2 is not distributed normally.
- In Dataset 3 the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset 4 shows that one outlier is enough to produce a high correlation coefficient.

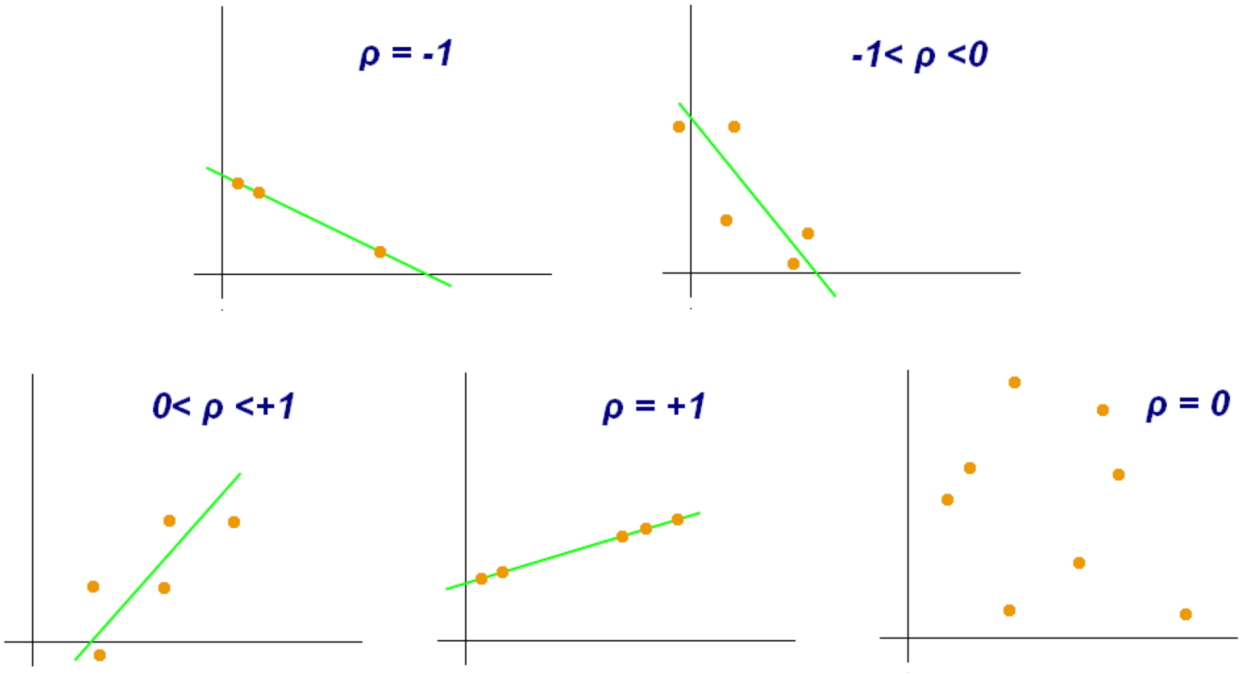
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset. Visualizing our data allows us to revisit our summary statistics and recontextualize them as needed. For example, Dataset II from Anscombe's Quartet demonstrates a strong relationship between x and y, it just doesn't appear to be linear. So a linear regression was the wrong tool to use there, and we can try other regressions. Eventually, we'll be able to revise this into a model that does a great job of describing our data and has a high degree of predictive power for future observations.

5. What is Pearson's R?

The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by *r*. Basically, a Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, *r*, indicates how far away all these data points are to this line of best fit (i.e., how well the data points fit this new model/line of best fit).

Co-efficient values:

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.



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

Most of the times we see that the data in different columns of the dataset are in different range. While some might be in the figures of thousands or millions, some might be in the figure of single or double digits. If we build the model on this dataset the column with larger range will have higher co-efficient and it will always dominant the performance of the model. To resolve the issue, we have to scale the data in such a way that they would be comparable. The is called scaling. On the other words scaling is a method used to normalize the range of independent variables or features of data.

Normalization

Normalization often also simply called **Min-Max scaling** basically shrinks the range of the data such that the range is fixed between 0 and 1 (or -1 to 1 if there are negative values). It works better for cases in which the standardization might not work so well. If the distribution is not Gaussian or the standard deviation is very small, the min-max scaler works better.

Normalization is typically done via the following equation:

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

Standardization

Standardization (or **Z-score normalization**) is the process of rescaling the features so that they'll have the properties of a Gaussian distribution with μ=0 and σ=1

where μ is the mean and σ is the standard deviation from the mean; standard scores (also called ***z*** scores) of the samples are calculated as follows:

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

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

As we know that VIF is calculated using the following formula

$$VIF = 1/(1-R^2)$$

When $R^2 = 1$ then VIF = 1/0 which results in infinite. This indicates that there is a perfect co-relation. Although this happens rarely.

8. What is the Gauss-Markov theorem?

The Gauss Markov theorem states that if a certain set of assumptions are met, the ordinary least squares estimate for regression coefficients gives you the best linear unbiased estimate (BLUE) possible.

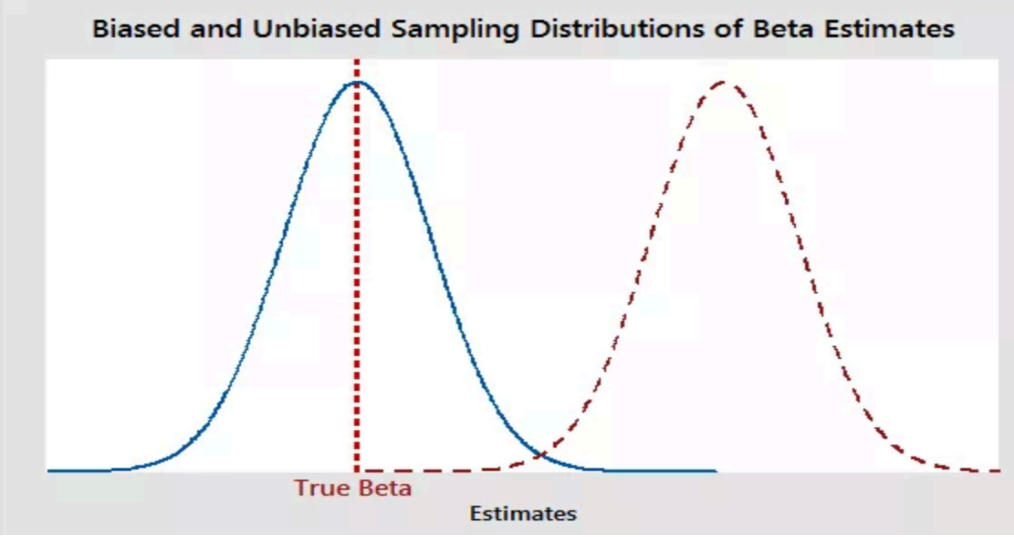
There are five Gauss Markov assumptions (also called conditions):

a) **Linearity**: the parameters we are estimating using the OLS method must be themselves linear.
b) **Random**: our data must have been randomly sampled from the population.
c) **Non-Collinearity**: the regressors being calculated aren't perfectly correlated with each other.
d) **Exogeneity**: the regressors aren't correlated with the error term.
e) **Homoscedasticity**: no matter what the values of our regressors might be, the error of the variance is constant.

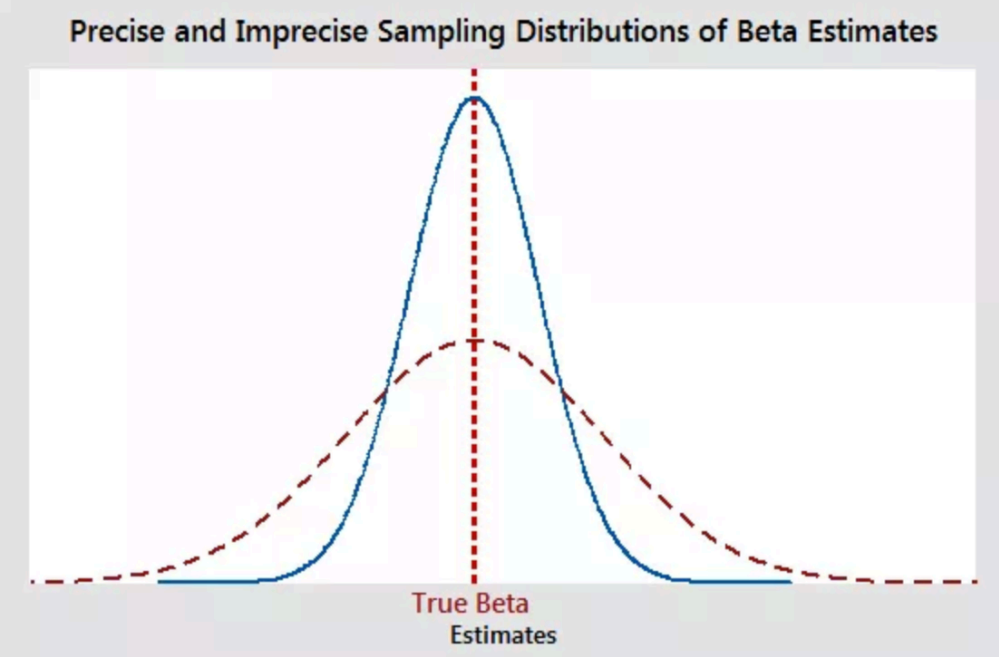
Checking how well our data matches these assumptions is an important part of estimating regression coefficients. When you know where these conditions are violated, you may be able to plan ways to change your experiment setup to help your situation fit the ideal Gauss Markov situation more closely.

There are 2 hypotheses used in the Gauss Markov theorem

1. Unbiased Estimates: Sampling Distributions Centered on the True Population Parameter

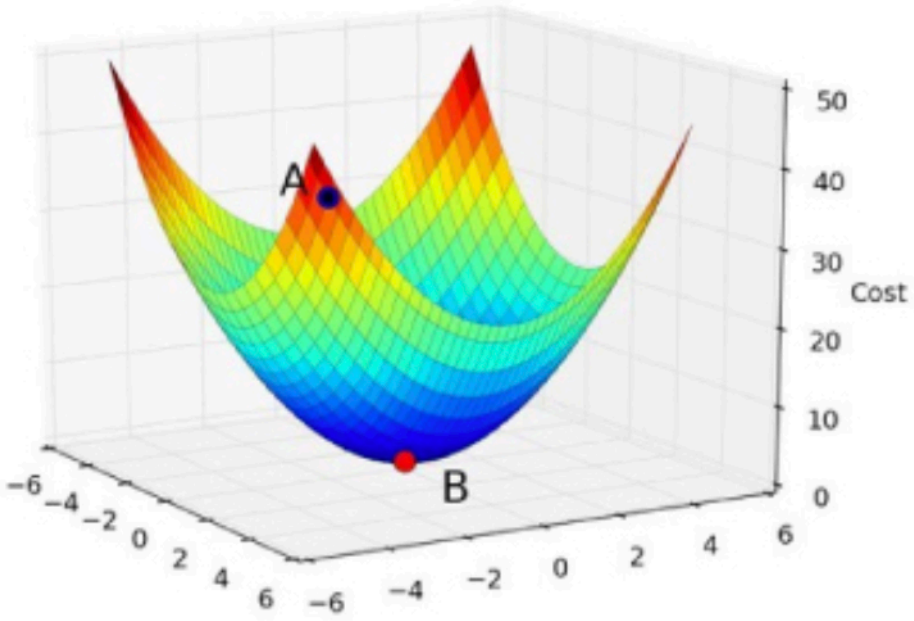


2. Minimum Variance: Sampling Distributions are Tight Around the Population Parameter

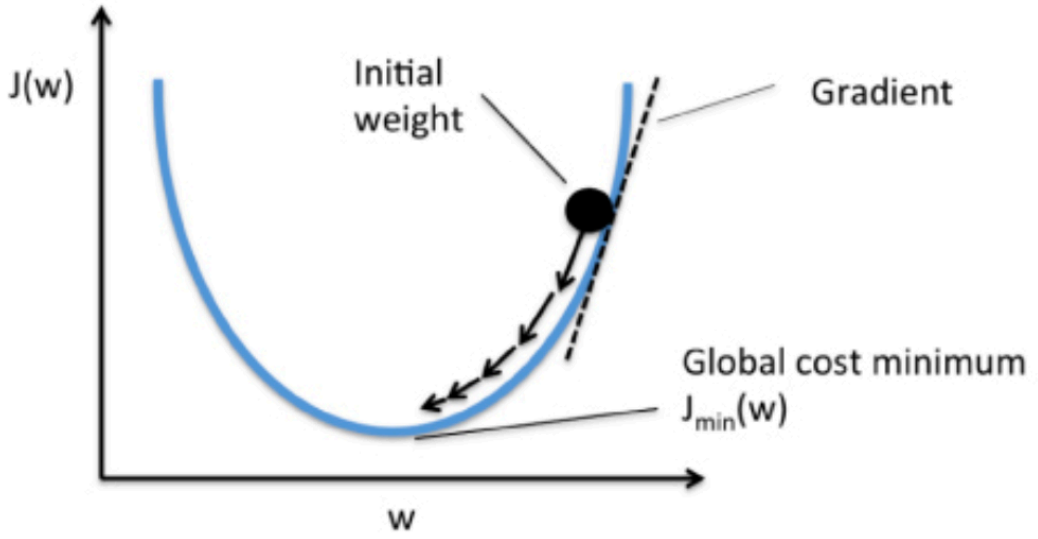


9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm which is mainly used to find the minimum of a function. Gradient descent is used to update parameters in a model. Parameters can vary according to the algorithms, such as *coefficients* in Linear Regression and weights in Neural Networks.



The inclined and/or irregular is the cost function when it is plotted, and the role of gradient descent is to provide direction and the velocity (learning rate) of the movement in order to attain the minima of the function i.e where the cost is minimum.



The objective in the case of gradient descent is to find a line of best fit for some given inputs, or X values, and any number of Y values, or outputs. A cost function is defined as “a function that maps an event or values of one or more variables onto a real number intuitively representing some “cost” associated with the event.”

A Cost Function/Loss Function tells us “how good” our model is at making predictions for a given set of parameters. The cost function has a curve and a gradient, the slope of this curve helps us to update our parameters and make an accurate model.

It is always the primary goal of any Machine Learning Algorithm to minimize the Cost Function. Minimizing cost functions will also result in a lower error between the predicted values and the actual values which also denotes that the algorithm has performed well in learning.

Types of Gradient Descent Algorithms

1. Batch Gradient Descent:

In this type of gradient descent, all the training examples are processed for each iteration of gradient descent. It gets computationally expensive if the number of training examples is large.

2. Stochastic Gradient Descent:

Stochastic Gradient Descent (SGD) samples are selected at random for each iteration instead of selecting the entire data set. When the number of training examples is too large, it becomes computationally expensive to use batch gradient descent, however, Stochastic Gradient Descent uses only a single sample, i.e., a batch size of one, to perform each iteration.

3. Mini Batch gradient descent:

This algorithm processes the data in batches in one go even if the number of training examples is large. Also, the number of iterations is lesser in spite of working with larger training samples.

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

Q-Q plot is a plot of the quantiles of the first dataset against the quantiles of the second datasets. It is a graphical technique which helps in determining if 2 of the datasets come from populations with a common distribution. Here, quantile denotes the fraction of points below the given value. E.g. .4 quantile means 40% of the data falls below the given value while 60% are the above that value.

The Q-Q plot is important for linear regression as it is used to answer the following questions:

- Do two data sets come from populations with a common distribution?
- Do two data sets have similar tail behaviour?
- Do two data sets have similar distributional shapes?
- Do two data sets have common location and scale?

Structure of Q-Q plot

- Vertical axis: Estimated quantiles from data set 1
- Horizontal axis: Estimated quantiles from data set 2

Significance of the Q-Q plot in Linear Regression:

1. The sample sizes do not need to be equal.
2. Many distributional aspects can be simultaneously tested.

E.g.