1. Explain the linear regression algorithm in detail.

Linear Regression is used to show linear relationship of a dependent variable with independent variables.

Here will be having a generalized model whose coefficient, we try to predict such that variation in predicted output and actual output is at minimum.

A model in linear regression is nothing but an equation.

Sequential Steps followed in the algorithm are as follows: -

Initially we will be assigning random values to the coefficients of the model.

After assigning, we predict the output of the model for a given set of inputs.

We assess how much the predicted output varies from the actual output using a cost/loss function.

Commonly used cost functions are Mean Squared Error and Mean Absolute Error.

Having the cost function value calculated we use optimization algorithms such as Gradient Descent to change the value of coefficient such that Loss function of new model is lesser than the previous model.

Optimization algorithms helps in changing the value of coefficients in correct direction.

Above process of calculating variation in actual vs predicted and changing the coefficient of the model is repeated iteratively till it settles at a minimum value for loss function i.e. till we reach a local /global minima.

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

Assumptions of Linear Regression with respect to Residuals: -

1. Variance of output error should be constant throughout the model.
   This is also known as Homoscedasticity.

   Example: - Variation of spending with income.

   Here spending will be definitely low when the income is low, variation will be much smaller. But as the income increases it is not definite that his spending also increases. Spending may or may not increases which depends on the person.

   So variation (std.dev) in spending will be high for income than compared to lower income. Such questions cannot be analyzed using linear regression.

2. Each successive Residuals should be independent of the previous one.
   Violation of this is known as auto correlation.
   These violations mainly occur in Time Series data.

   When successive Residuals depends on the previous one the residual plot will not show even spread across both sides of predicted best fit line.

   Example: - Pendulum swinging

   Pendulum swinging observation where we observe the position of pendulum at different time frame. This experiment is repeated for n number of times.

   Here we will be scatterplot points resembles a sinusoidal curve. When we try to fit a line through it we can see presence of clusters on the side of predicted best fit line.

   Such problems cannot be analyzed through linear regression.

2. Spread of Residuals for a given value should be normally distributed
   The distribution of Residuals should resemble a bell curve.

   Example: - Medical Insurance Payout

   For medical insurance payout for customer based on their age. Most of the customers won't have availed payout. Only few person based on their ages would have availed insurance.

   When we try to fit a line in such samples we won't get normally distributed Residuals.

   Such observation cannot be analyzed through linear regression.

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

Coefficient of Correlation: -

It gives the strength and direction of linear relationship between two variables.

It varies from -1 to 1.

1 represents positive correlation, whereas -1 represent negative correlation and o represents no correlation at all.

Practically it's not possible to get whole numbers like 1,-1,0 as coefficient.

Coefficient of Determination: -

It tells how predictable output given the set of inputs.

Its value varies between 0 and 1.

Here the coefficient is the measure of how well the regression line represents the sample of data.

Coefficient of Determination is the square of Coefficient of Correlation.

4. Explain the Anscombe's quartet in detail.

Anscombe's quartet is a set of four dataset that has similar statistical property but look completely different distribution when plotted.

It was used to demonstrate the importance of visualization in analysis and influence of outliers in observation.

For each samples Best fit line lies at the same spot.

Plot in each Quartet: -

First Quartet - It is a simple linearly distributed sample with normally distributed errors.

Second Quartet - Here the relationship between variable is nonlinear and when you try to fit a line through it the residuals are not distributed normally.

Third Quartet - Here a line can exactly fit through the sample but direction of the best fit is changed due to presence of an outlier.

Fourth Quartet - Here even though the samples doesn't indicate any relationship , one high leverage point causes best fit line to lie in the same spot as in other three quartets.

5. What is Pearson's R?

Pearson's R is one of the commonly used Correlation Coefficient. There are several types of Correlation Coefficient.

R is the measure of linear correlation between two variables. I.e. it measures how strong two variables are correlated.

Its value lies between -1 to 1.

Values nearer to -1 and 1 means there is a strong correlation, whereas values nearer to 0 means weak correlation.

When R value is nearer to 1 it means positive correlation, i.e. increase in value x increases the value of y.

When R value is nearer to -1 it means negative correlation, i.e. increase in value of x decreases the value of y.

Practically the possibility of getting whole value like 1,-1 and 0 is nonexistence.

Example:-

The amount of calories burned is positively correlated with the amount of time we spend on exercising.

The grade of student is negatively correlated with number of leave taken.

There is zero correlation between amount of tea drunk and level of intelligence.

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

Scaling or feature scaling is a method used to standardize the range of values for a feature. It is done for numeric features.

Scaling is performed because once scaling is done, optimization algorithm Gradient descent learns the feature much faster than with unscaled values. Also many M.L algorithms like classifiers need scaling to be done to work properly.

Difference:-

In min-max scaling, values of variable varies between 0 and 1, while standardized scaling there is no defined intervals.

In standardized scaling mean of the total samples is 0, while in min-max scaling mean is always be greater than 0.

In Standardized scaling std.dev will be 1 while in min-max scaling it is always be less than 1.

While outliers have to be handled in standardized scaling, it is not the case in min-max scaling.

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

This happens when there is a perfect correlation between samples of two different variable.

Formula for V.I.F is 1/ (1- R^2).

When there is a perfect correlation. R value is either 1 or -1.

Substituting R value in V.I.F's formula, we get VIF = 1/ (1-1) = 1/0 = infinity.

For observed data it is not rare to get perfect correlation due to presence of errors.

8. What is the Gauss-Markov theorem?

The Gauss-Markov theorem states that if your linear regression model satisfies the first six classical assumptions, then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators.

When you satisfy the classical assumptions, you can be confident that you are obtaining the best possible coefficient estimates.

The Gauss-Markov theorem does not state that these are just the best possible estimates for the OLS procedure, but the best possible estimates for any linear model estimator

9. Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm that is used to converge a model to local/global minima for given sample of inputs.

Sequential Steps followed in the algorithm are as follows: -

Initialize random values for coefficients of the model, this is the initial guess for the Gradient Descent algorithm to improve upon.

Evaluate how well the randomly initialized model fits the sample data using a loss function.

Find the slope of the loss function at a point of calculated by the above step. Slope is calculated for each coefficient.

Subtract the value proportionate to the slope from coefficient.

Iterate the above process until Residuals reaches a minimum value.

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 probability graph that is used to find kind of probability distribution a population follows.

Sequential Steps followed to plot Q-Q plot are as follows: -

Each sample point of the population is given its own quantile.

The above point is achieved by dividing the probability distribution for which we are testing into same number of quantiles as that of samples.

It is divided such a way that the probability represented in each partition is same.

Now samples is taken in one axis and quantiles of probability distribution is taken in another axis.

Intersection of smallest sample value and smallest quantile is found.

Likewise intersection for second smallest value and so on is found.

The collection of intersection points is now analyzed for how well it fits a straight line.

If the points very closely resemble a straight line then the sample follow the probability distribution for which it is tested for.

If it doesn't resemble a straight line then the sample doesn't follow the probability distribution for which it is tested for.