1. Explain the linear regression algorithm in detail.

Linear Regression is a way of predicting the value (dependent variables) based on independent variable also called as predictors.

Linear Regression Algo

Given a dataset

A) We identify the target variable (variable to be predicted) and set of independent variables (predictors)
B) We check if the target variable is normally distributed with the help of distribution plot. And identify there is some linear relation between the target variable and independent variable
C) We identify the correlation between the independent variable. It may happen some of the independent variable are highly correlated. One of the highly correlated variables are to be included as inclusion of both can impact the beta coefficient.
D) We normalize the data and segregate the data between train and test data set
E) On the train dataset, we include all the columns and identify the coefficient of the variables. And then checks if the variables included have high collinearity
F) If there are some variables with high collinearity we remove such variable and rebuild the model and continues the step until no features with high multicollinearity is left.
G) We check the p-value of each features and if feature has p-value > 0.05 we get rid of them and rebuild the model
H) Once the model is finalized we calculate the predicted value on train data set and see the distribution
I) On the test data we identify the predicted value and see the distribution and also calculate the residuals errors on train data set

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

    a) There should be linear relationship between dependent and independent variables.
    b) Assumption about residuals:-
        1) Error terms are normally distributed
        2) Error terms are normally distributed around 0. i.e. its mean is 0
        3) Error terms have same variance
        4) Errors are independent of each other
    c) There should be no multicollinearity between independent variables.

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

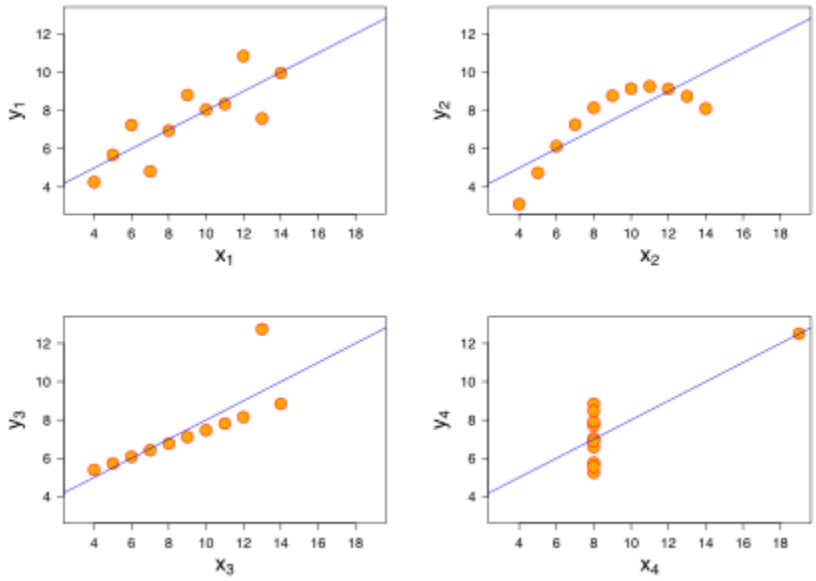
Coefficient of determination is also denoted as R-Squared (R2) and is also referred as goodness of fit. It explains how much variability in one variable can be caused by its relationship to another factor.

It's value lies between -1 to +1.

On graph, goodness of fit measures the distance between a fitted line and all of the data points scattered.

Coefficient of correlation is used to determine how strong a relationship between two variables.

4. Explain the Anscombe's quartet in detail.
It's a group of four datasets that appear to be similar when using typical summary statistics, yet tell four different stories when graphed. Each dataset consists of eleven *(x,y)* pairs.



- 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.
- 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.
- 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.

- 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.

5. What is Pearson's R?
   Pearson's R is the measure of strength of correlation between two continuous variables.
   It's value ranges from -1 to +1.
   R = -1 => data lie on perfect line with negative slope
   R = 0 => no linear relationship between the variables
   R = +1 => data lie on perfect line with positive slope
   R = ∑xy /(√∑x2 ∑y2)

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

Scaling in Machine learning is performed to standardize the independent features present in the data in a fixed range.

Why Scaling is performed?

Since the range of the values of raw data varies widely some of the machine learning algos objective functions won't work properly without normalization.

For example, many classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance.

Another reason why feature scaling is applied is that gradient descent converges much faster with feature scaling than without it

Standardization Scaling: Is the process of rescaling the features so they have the properties of gaussian distribution with mean(μ) =0 and standard deviation = 0

Standard scores are calculated as $$z = \frac{x - \mu}{\sigma}$$

It is also called as Z-score normalization

Normalization Scaling: is also called as min-max scaling.

Basically shrinks the range of data between 0 and 1

Formula is X' = (Xi – min(x)) / (max(x) – min(x))

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

Variance Inflation Factor (VIF) detects multicollinearity in regression analysis.

It is calculated as 1/(1-R2)

If there is perfect correlation between the variables it's value if infinity. Such variables should be eliminated from the model as it can impact the calculation of the coefficient of other variables.

8. What is the Gauss-Markov theorem?

The **Gauss Markov theorem** tells us 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):

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

The Gauss Markov assumptions guarantee the validity of ordinary least squares for estimating regression coefficients.
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

9. Explain the gradient descent algorithm in detail.

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

Gradient descent is best used when the parameters cannot be calculated analytically (e.g. using linear algebra) and must be searched for by an optimization algorithm.

The procedure starts off with initial values for the coefficient or coefficients for the function. These could be 0.0 or a small random value.

coefficient = 0.0

The cost of the coefficients is evaluated by plugging them into the function and calculating the cost.

cost = f(coefficient)

or

cost = evaluate(f(coefficient))

The derivative of the cost is calculated. The derivative is a concept from calculus and refers to the slope of the function at a given point. We need to know the slope so that we know the direction (sign) to move the coefficient values in order to get a lower cost on the next iteration.

delta = derivative(cost)

Now that we know from the derivative which direction is downhill, we can now update the coefficient values. A learning rate parameter (alpha) must be specified that controls how much the coefficients can change on each update.

coefficient = coefficient – (alpha * delta)

This process is repeated until the cost of the coefficients (cost) is 0.0 or close enough to zero to be good enough.

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

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.

The advantages of the q-q plot are:

a) The sample sizes do not need to be equal.
b) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.