Explain the linear regression algorithm in detail.

It is used to show the linear relationship between a dependent variable and one or more independent variables.

Step 1: *Make a chart of your data, filling in the columns in the same way as you would fill in the chart if you were finding the Pearson's Correlation Coefficient.*

Step 2: Use the following equations to find a and b.

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

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

Step 3: *Insert the values into the equation*.
y' = a + bx

The function for Linear Regression is y = β0 + β1.x

y is our output variable which we are trying to predict

x is our independent variable

When we perform training of the model basically what the model is doing, is it is trying to fit the best line to predict the y variable for given input x variable/s, and this is done by finding out the best values of β0 (intercept) and β1 (coefficient of x). Once we get our optimized β0 & β1 values we get our best fit line and then this line/coefficients are used finally in prediction of the y target value for a given value x.

This best-fit line is obtained by minimizing the expression of RSS (Residual Sum of Squares) which is equal to the sum of squares of the residual for each data point in the plot (i.e. the difference between the actual y value and the predicted y value).

This process of reducing the RSS to achieve the ideal β0 & β1 values is an iterative process which is called Gradient Descent method.

Another way of checking our linear regression model is find out what is called as R2 value (Coefficient of Determination). The formula for R2 = 1 – (RSS/TSS). R2 value ranges from 0 to 1 and closer your R2 value to 1 the model is fitting the data that much well. R2 value tells you how much variance in the y is explained by the model. We already know what RSS is, here TSS means total sum of squares

What are the assumptions of linear regression regarding residuals?

The regression has five key assumptions:

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

❖ Error terms are normally distributed: This can be visualized by drawing a histogram of the residuals,and then examine the normality of the residuals. If the residuals are not skewed, that means the assumption is satisfied, i.e. mean of the errors should be close to 0 and std. deviation close to 1.

❖ Error terms are independent of each other.

❖ Error terms have constant variance (homoscedasticity): Here we check for visible patterns in the error terms in order to determine that these terms have a constant variance. If the errors show a certain pattern then we know that error terms are not having constant variance, whereas if the error terms are randomly distributed around 0 we can say that the error terms have constant variance.

❖ Linearity of Error terms: Here we can draw a scatter plot of the errors and y values. y values are taken on the vertical axis, and the errors are plotted on the horizontal axis. If the scatter plot shows a linear pattern that shows that linearity assumption is met.

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

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 $-1 < r < +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively. A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined

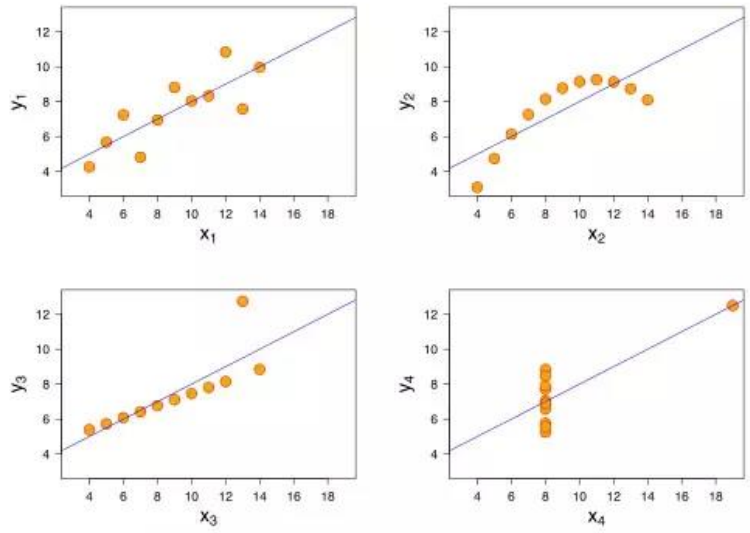
The *coefficient of determination,* r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.

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

The *coefficient of determination* is such that $0 < r^2 < 1$, and denotes the strength of the linear association between *x* and *y*. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in *y* can be explained by the linear relationship between *x* and *y* (as described by the regression equation). The other 15% of the total variation in *y* remains unexplained.

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.



The first scatter plot (top left) looks like a simple linear relationship.

The second graph (top right) is not distributed normally; while a relationship between the two variables is obvious, it is not linear.

In the third graph (bottom left), the distribution is linear, but should have a different regression line. The calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient.

In the fourth graph (bottom right) is 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.

What is Pearson's R?

the Pearson correlation coefficient ,also referred to as Pearson's *r*, the Pearson product-moment correlation coefficient (PPMCC) or the bivariate correlation,[1] is a measure of the linear correlation between two variables *X* and *Y*. According to the Cauchy–Schwarz inequality it has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

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

Scaling also called as feature scaling becomes very important, when you have a lot of independent variables in a model and you may notice that a lot of them might be on very different scales, and the outcome will be a model with very weird coefficients that might be difficult to interpret. So the process of bringing all variables to a comparable scale is called feature scaling. The need to scale features are mainly:

1. For ease of interpretation of the coefficients in the model.

2. Faster convergence for gradient descent methods.

There are multiple ways of scaling but two very frequently used are:

Normalised and Standardised scaling:

- Standardized Scaling: The variables are scaled in a way that their mean is zero and standard deviation is one.
- Normalized Scaling: The variables are scaled in a way that all the values lie between zero and one. It is also called as Min-Max scaling since it uses the maximum and the minimum values in the data to perform this operation.

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

VIF is a measure that explains the relationship of one independent variable with all the other independent variables. It is possible that one variable might not totally explain the other variable but there is a possibility that when two or three variables put together might explain completely the other variable. In this case the other variable becomes totally redundant and we must drop that variable.

An infinite VIF means that there is perfect correlation between the variables and one of them should be dropped as it is redundant to the model.

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, provided it exists.

There are five major assumptions:

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

Explain the gradient descent algorithm in detail.

Gradient descent is an optimization algorithm used to minimize some function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. In machine learning, we use

gradient descent to update the parameters of our model. Parameters refer to coefficients in Linear Regression and weights in neural networks.

Running gradient descent using our new cost function. There are two parameters in our cost function we can control: mm (weight) and bb (bias). Since we need to consider the impact each one has on the final prediction, we need to use partial derivatives. We calculate the partial derivatives of the cost function with respect to each parameter and store the results in a gradient.

To solve for the gradient, we iterate through our data points using our new mm and bb values and compute the partial derivatives. This new gradient tells us the slope of our cost function at our current position (current parameter values) and the direction we should move to update our parameters. The size of our update is controlled by the learning rate.

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

Q Q Plots (Quantile-Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. 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.

Step 1: Order the items from smallest to largest.

Step 2: Draw a normal distribution curve

Step 3: Find the z-value (cut-off point) for each segment

Step 4: Plot your data set values (Step 1) against your normal distribution cut-off points