1. Explain the linear regression algorithm in detail.

In simple terms, linear regression is a method of finding the best straight-line fitting to the given data, i.e. finding the best linear relationship between the independent and dependent variables.
Linear regression is used for finding linear relationship between target and one or more predictors. There are two types of linear regression- Simple and Multiple Linear Regression.

Simple Linear Regression:

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable

Multiple Linear Regression:

Multiple linear regression is useful for finding relationship between a continuous variable and one or more continuous and categorical variable. One is dependent variable and others are Predictor or independent variable

In technical terms, linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

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

- *Normality assumption*: It is assumed that the error terms, ε(i), are normally distributed.
- *Zero mean assumption*: It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- *Constant variance assumption*: It is assumed that the residual terms have the same (but unknown) variance, σ^2. This assumption is also known as the assumption of homogeneity or homoscedasticity.
- *Independent error assumption*: It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

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

- The Correlation Coefficient, r, is a statistical measure that calculates the strength and direction of a linear relationship between two variables. Its value lies between -1 and 1.
- The mathematical formula for computing r is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where ‘n’ is the number of pairs of data

- *Positive correlation*: Positive values indicate a relationship between x and y such that if one variable increases, other also increases. Correlation value of +1 indicates a perfect positive fit.
- *Negative correlation*: Negative values of r indicate that variables move I opposite directions i.e. one variable increase while the other decreases
- *No Correlation*: If there is no linear relationship between two variables, r is close to 0

The **Coefficient of Determination** gives proportion of variable or fluctuation of one variable that is predictable from other variable. It is a measure of how well the regression line represents the data. It lies between 0 and 1.

It is used to determine how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the lines passes through when data points and line are plotted.

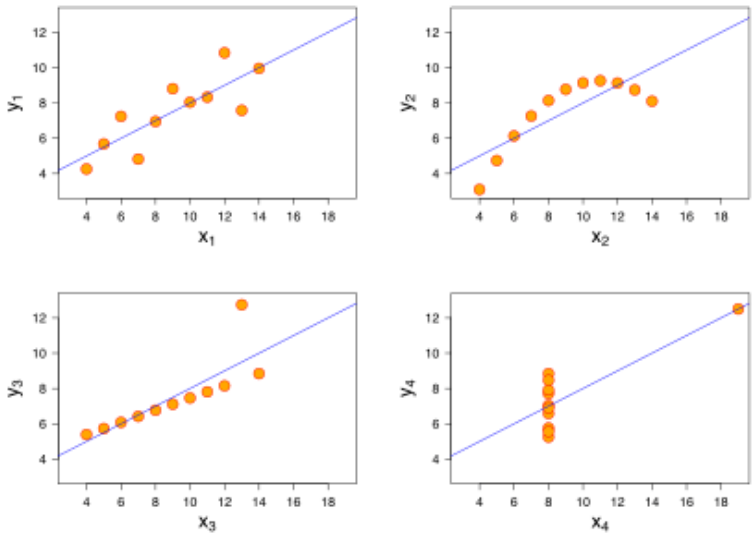
The usefulness of Coefficient of Determination is to find the likelihood of future events falling within the predicted outcome.

4. Explain the Anscombe’s quartet in detail.

Anscombe’s quartet is a collection of four datasets that are identical in statistical properties but vary considerably when graphed.

For Example, Suppose there are 4 datasets each consists of eleven (x, y) points. All four of these data sets have the same variance in x, variance in y, mean of x, mean of y and linear regression.

Anscombe’s quartet highlights the importance of graphing data before analysing it. It also demonstrates the effect of outliers on statistical properties. Thus, showing that statistics about a data set do not fully depict data in its entirety.



- The scatter plot on top left appears to be a simple linear relationship corresponding to two variables correlated and following assumption of normality.
- The graph on top right is not distributed normally but a relationship between two variables can be observed which is not linear with an irrelevant Pearson correlation coefficient.
- The graph on bottom left has a linear distribution but has a different regression line which is offset by one outlier which exerts enough influence to alter the regression line.
- The graph at the bottom right again shows that even though the relationship between two variables is not linear, one outlier is enough to produce a high correlation.

5. What is Pearson's R?

Pearson's Correlation coefficient, R is a measure of the strength of the association between two variables. It is based on the method of covariance and also gives the direction of relationship. Its value ranges from -1 to 1.

This is a good indicator if a linear relationship exists between two variables.

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

Scaling is a data pre-processing step which is applied to independent variables or features of data. Most of the times data sets contains features varying highly in magnitudes, units and range. Scaling basically helps to normalize the data within a particular range.

Importance: The algorithms which use Euclidian distance measure are sensitive to magnitude. Here, feature scaling helps to weigh all the features equally. If a feature in a data set is big in scale compared to others then in algorithms where Euclidian distance is measured this big scaled feature becomes dominating and needs to be normalized.

- Normalization typically means rescaling the values into a range of [0, 1]. When the data set has outliers, normalized scaling, scales the "normal" data to a very small interval. Also known as Min-Max scaling.

  Formula: $X_{new} = (X - X_{min}) / (X_{max} - X_{min})$

- Standardization means rescaling the data to have mean 0 and a unit variance (standard deviation of 1). In standardized scaling, new data isn't bounded unlike normalized scaling.

  Formula: $X_{new} = (X - \mu) / \sigma$

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

Variance inflation factors show the degree to which a regression coefficient will be affected because of the variable's redundancy with other independent variables.

As the squared multiple correlation of any predictor variable with the other predictors approaches unity, the corresponding VIF becomes infinite. This means that variable is completely redundant and can be completely defined with the help of some other variable.

Example, consider variable X and Y, and assume Y = 5 * X. This means that Y can be completely defined with help of X. Hence, they will be having perfect correlation resulting in infinite VIF.

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):

- *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.

  Linear regression model represented by

  $y_i = x_i * \beta + \varepsilon_i$

  and generated by the ordinary least squares estimate is the best linear unbiased estimate (BLUE) possible if

  - $E\{\varepsilon_i\} = 0, i = 1, . . ., N$
  - $\{\varepsilon_1 \ldots\ldots \varepsilon_n\}$ and $\{x_1 \ldots., x_N\}$ are independent
  - $cov\{\varepsilon_i, \varepsilon_j\} = 0, i, j = 1, \ldots., N\ I \neq j.$
  - $V\{\varepsilon_1 = \sigma^2, I = 1, \ldots.N\ \}$

  The first of these assumptions can be read as "The expected value of the error term is zero.".

  The second assumption is collinearity,

  the third is exogeneity, and the fourth is homoscedasticity.

9. Explain the gradient descent algorithm in detail.

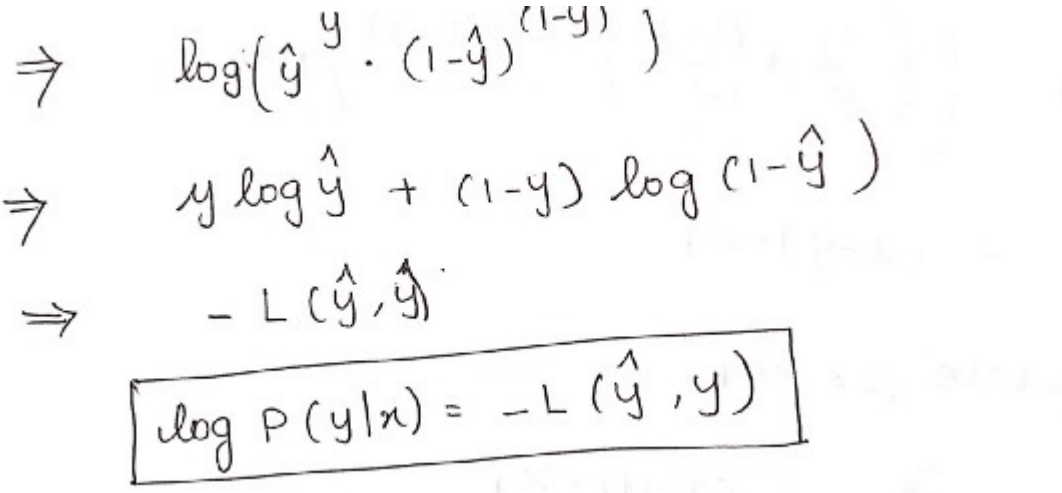
Gradient descent is by far the most popular optimization strategy, used in machine learning and deep learning now. It is used while training your model, can be combined with every algorithm and is easy to understand and implement. Therefore, everyone who works with Machine Learning should understand its concept

Let's consider the same above example. It has 'y' dependent variable and 'x' as independent variable. Let's say the Weight for 'x' is 'w' and intercept (or bias) term is 'b'

Gradient Descent tries to minimize cost function, which can be different for classification or regression. But the process it follows remains same.

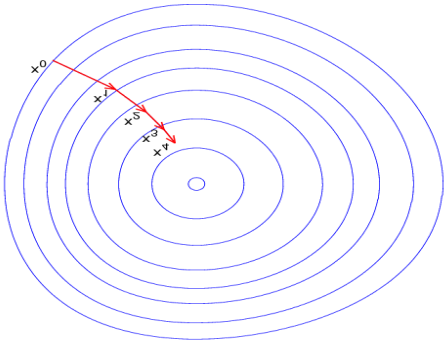
Cost Function: A mathematical formula used to predict the cost or loss associated with a certain action or a certain level of output.

Like for binary Classification problem we use cost function given below:

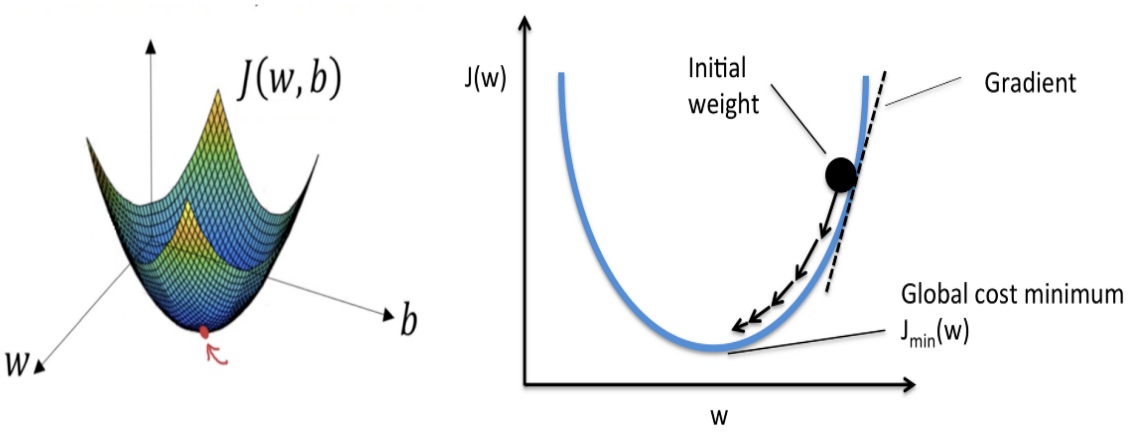


Consider the below example first:

Imagine a blindfolded man who wants to climb a hill, with the fewest steps possible. He just starts climbing the hill by taking big steps in the steepest direction, which he can do, if he is not close to the top. As he comes further to the top, he will do smaller and smaller steps, since he doesn't want to overshoot it. This process can be described mathematically, using the gradient.



So, in our example, Cost function will be dependent on Weights (w) and bias (b). Hence, we need to find the optimal values of 'w' and 'b' which minimizes the cost function. Let's denote Cost function by J(w, b).



Let's now go step by step to understand the **Gradient Descent algorithm:**

Step 1: Initialize the weights (w) and bias (b) with random values and calculate Cost Function.

Step 2: Calculate the gradient i.e. change in Cost when the 'w' and 'b' are changed by a very small value from their original randomly initialized value. This helps us move the values of 'w' and 'b' in the direction in which Cost function is minimized.

Step 3: Adjust the weights with the gradients to reach the optimal values where Cost function is minimized

Step 4: Use the new weights for prediction and to calculate the new Cost function

Step 5: Repeat steps 2 and 3 till further adjustments to weights doesn't significantly reduce the Error

General Equation for both weights and bias becomes:

$W = W - \alpha * (d J(W,b) / dw)$

$b = b - \alpha * (d J(W,b) / db)$

where $d J(W,b) / dw$ = partial derivative of cost with respect to weights

α = learning rate

Why do we use partial derivative in the equation?

Partial derivatives represent the rate of change of the functions as the variable change.

10. 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.

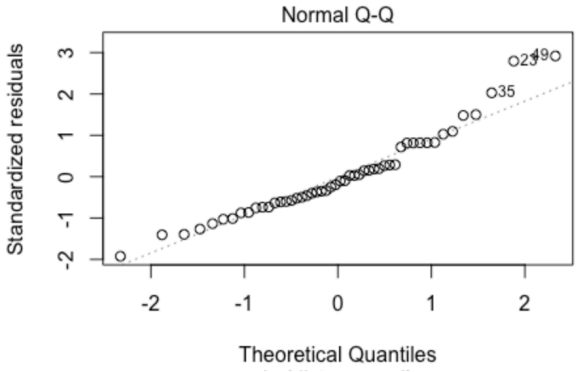
For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it.

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.

Use in Linear Regression

As per the assumptions of Residuals of Linear Regression, Residuals from Linear Regression model should normally distributed with mean 0 and standard deviation of 1.
The Residuals are plotted and against standard normal distribution reference line. If all the points of residuals fall on reference line, it shows that Residuals are normally distributed.

In simple way: fit a linear regression model, check if the Residual points lie approximately on the line, and if they don't, your residuals aren't Gaussian and thus your errors aren't either.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
-1
0
1
2
3
23
49
35