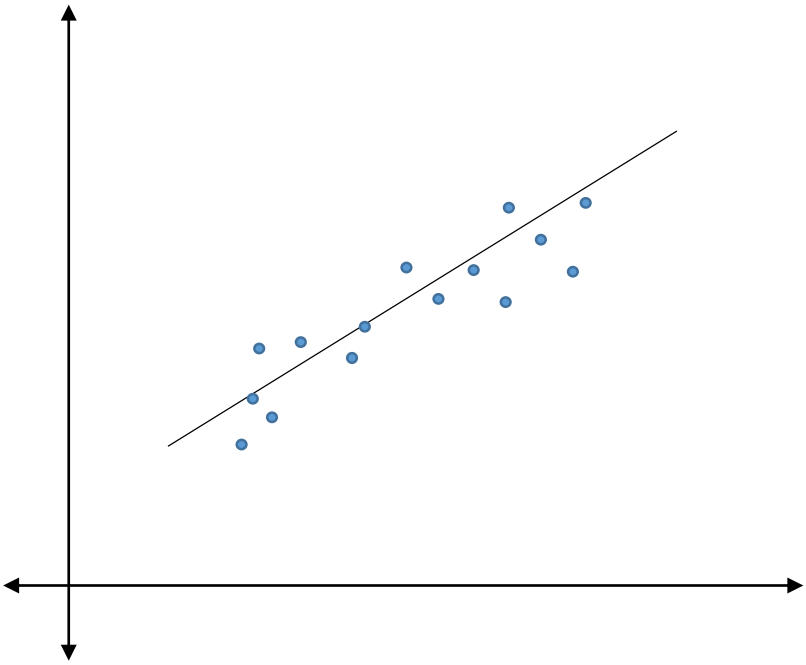
1. Explain the linear regression algorithm in detail.

Ans. Linear Regression is a machine learning algorithm based on supervised learning. Regression modelling models a target or dependent variable based on the independent variable. It is basically used to find the relationship between various variables and this is how it is different from forecasting. In forecasting we focus mostly on the output whereas in the regression modelling we try to find out the variables which help in predicting the dependent variable.



$$y = \beta_0 + \beta_1 x_1$$

Refer the above equation for the basic equation of the regression modelling. Now the main questions comes is that how that regression line is calculated or how is the perfect fit derived. Before we dive into that we need to understand the sum of the residues square or SSE. Now once a line pass through some of the points, we are going to have residues. These residues are the vertical distance of the point from the line. Thus when the square of all such residues are added we get SSE. Perfect fit is basically the line where the sum of squares of residues are the least.

One of the way regression modelling works is that it figures out the line which was lowest SSE but there are other ways too to calculate the best fit.

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

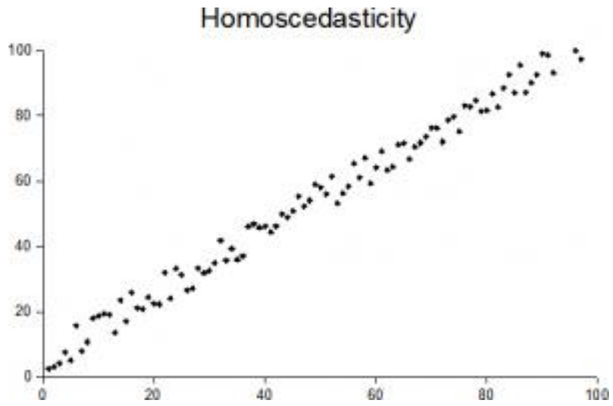
Ans. The assumptions of the linear regression are as follows,

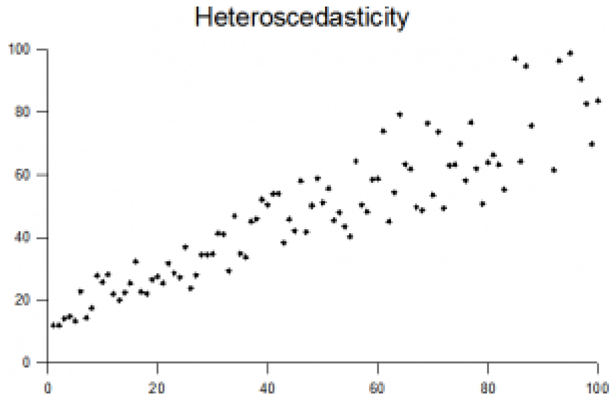
a. The mean of the residual is zero:
   The mean of all of the residual should be zero.

$$\frac{(r_1 + r_2 + r_3 + r_4 \ldots.)}{n} = 0$$

In the above equation ri is basically the residues.

b. Homoscedasticity of residuals or equal variance
Homoscedasticity basically refers to whether the residuals are equally distributed, or whether then bunch at certain values and at the other values they spread out. To visualize how homoscedastic looks is somewhat like a shotgun blast of randomly distributed data. The opposite of homoscedasticity is heteroscedasticity, the residue is in the form of the cone shape.





c. The X variables and residuals are uncorrelated

d. No autocorrelation of residuals

In linear regression, it is assumed that the residuals are independent of each other or basically not correlated with each other. We basically use the Durbin-Watson statistic to test for the presence of autocorrelation. This test is based on an assumption that errors are generated by a 1st order autoregressive process. If there are missing observations, then these are omitted from the calculations, and only non-missing observations are used.

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

Ans. The coefficient of determination, R^2, is basically used to analyse differences in one variable can be explained by a difference in a second variable. It basically measures the ability of the model to predict the outcome in regression modelling.

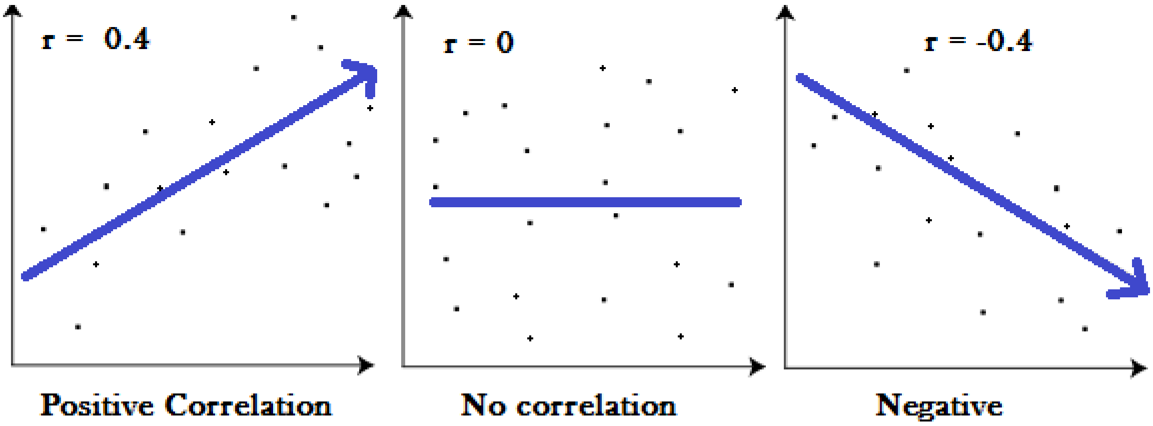
More specifically, R-squared gives you the variation in y (dependent variable) explained by x-variables (independent variable). The range of the coefficient of determination is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables.)

Correlation coefficient is basically used to find how strong a relationship is between data. For example, there are correlation coefficient can be used to find the relation between age and weight variables. The formulas return a value between -1 and 1, where:

a. Value of 1 indicates a strong positive relationship.

b. Value of -1 indicates a strong negative relationship.

c. A result of zero indicates there is no relationship between the variables.



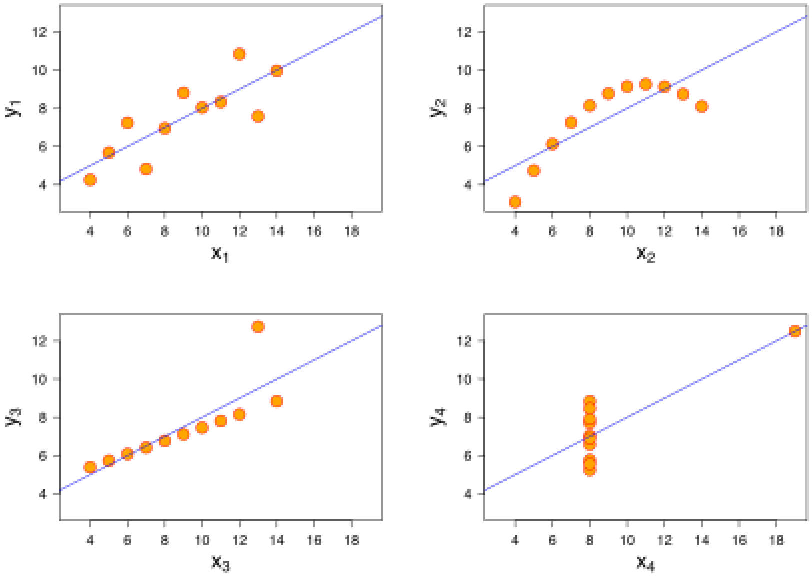
4. Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is basically set of the four datasets which have same summary statistics but when these datasets are visualized we can vividly see the difference between all the four datasets. Each of the four datasets basically consists of 11 datapoints. They were constructed in 1973 by the statistician *Francis Anscombe* to demonstrate that how important it is to visualize the data and not just analyse the data based on the summary statistics. While it is important to rely on the summary statistics, we cannot rule out the importance of visualization of the data.

- 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 normally distributed; (we can see that the relationship between the variables may be polynomial in nature) while a relationship between the two variables is obvious, it is not linear, and the Pearson correlation (R^2) coefficient is not relevant.
- In the third graph (bottom left), the distribution is linear, we can see that the spread of the data points are very different from graph 1.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.

The quartet is still often used to illustrate the inadequacy of statistic properties for describing realistic datasets and also look at the importance of looking at a set of data graphically before starting to analyse.

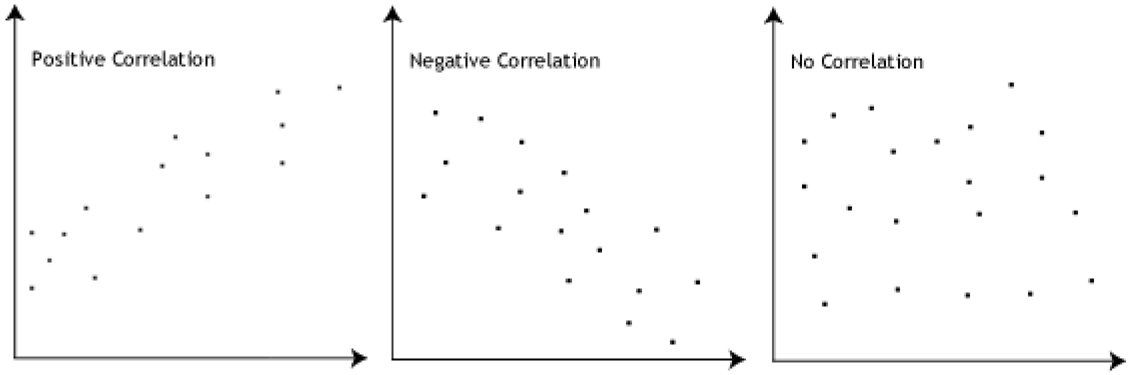


5. What is Pearson’s R?

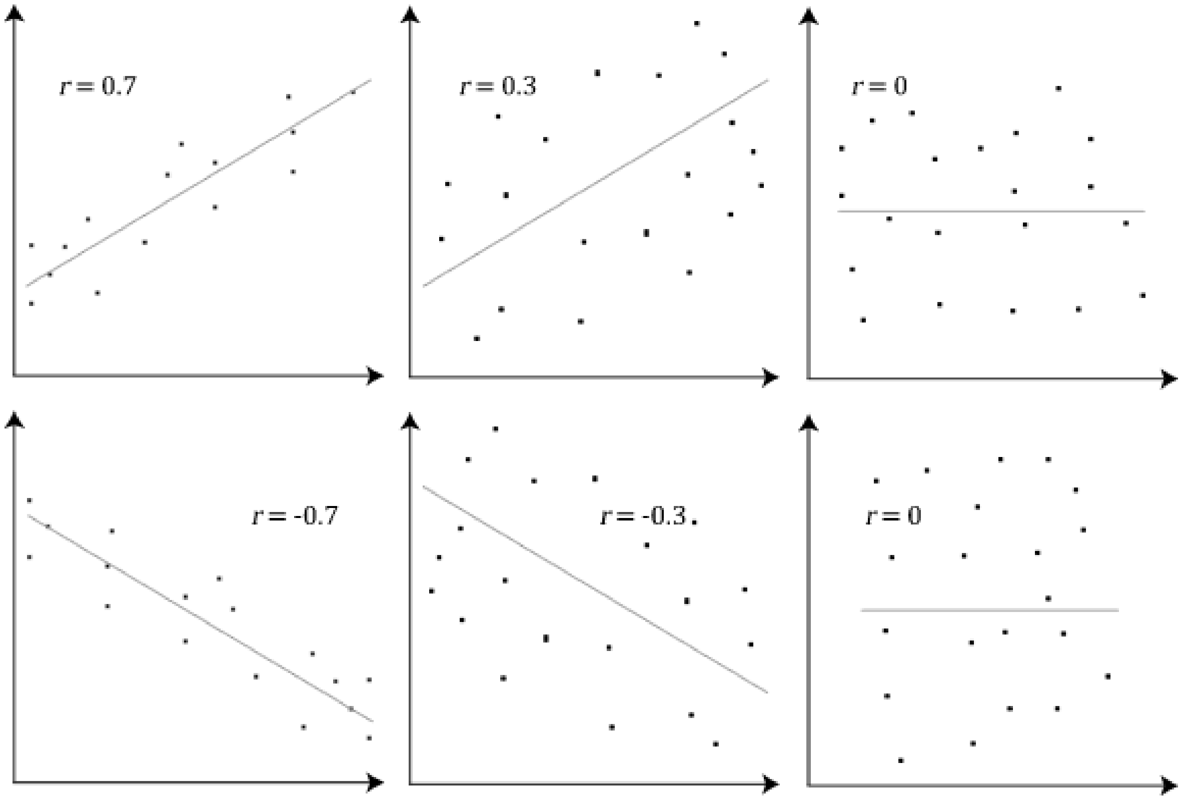
Ans. The Pearson correlation coefficient, r can take the values between -1 and 1.

$$-1 < r < 1$$

Now there are different implication for the each of the values that r assumes. For example when the value of the r is positive we can conclude that there is positive type of association between both the variables thus when the value of one of the variables increase we observe that the value of the corresponding variable also increase. If the value of r is 0, it implies that there is no association between two variables and so the changes in one of the variable doesn’t affect the values 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. This is shown in the diagram below:



We observe that stronger the relationship of association between the two variables the value of r is changed accordingly. For example is the value of the r is 0.7 for two variables then we can conclude that the association between the two variable is quite strong. If the value of r is either equal to 1 or -1 then it indicated that the association is represented by a single line. For any value of r between +1 and -1, as observed in the graph below, it indicates that there is variation around the line of best fit. If the value of the coefficient r is closer to 0 then we can see higher variation across the line of best fit. Refer the diagram below to understand the concept better,



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

Ans. Feature Scaling is a technique to scale down the independent features present in the data in a fixed range. It is performed before the regression modelling to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Consider the below example that one of the variable is measure in 1000's like meters and the other variable is being capture in Kilometers. And thus when the regression fit is done we will observe that the coefficient of one of the variables is greater than the other which might not help narrow down the variable which has the major impact on the dependent variable.

Techniques to perform Feature Scaling
Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

$$X_{new} = \frac{X_i - \min(X)}{\max(x) - \min(X)}$$

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

$$X_{new} = \frac{X_i - X_{mean}}{\text{Standard Deviation}}$$

An important point to note here is that we use only transform method on test dataset and not fit transform. This will cause the normalization value to sometimes increase more than 1 or either less than 1.

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

Ans. VIF is basically an index that provides a measure of how much variance of an estimated regression coefficient increases due to collinearity. Collinearity basically inflates the amount of the impact a variable has on the dependent variables. In order to calculate VIF, we fit a regression model between the independent variables.

VIF is calculated using the below formula,

$$VIF = \frac{1}{1 - R^2}$$

Now VIF basically increases as the collinearity increases. In the case that VIF is infinity it basically means that the two variables are perfectly collinear.

If all the independent variables are orthogonal to each other, then VIF = 1.0. If there is perfect correlation, then VIF may be equal or tending to infinity

8. What is the Gauss-Markov theorem?

Ans. The Gauss-Markov theorem states that if 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.

The Gauss-Markov theorem famously states that ordinary least squares is BLUE. BLUE is an abbrevation for ‘Best Linear Unbiased Estimator’.

In this context, the definition of “best” refers to the minimum variance or the narrowest sampling distribution. More specifically, when your model satisfies the assumptions, OLS coefficient estimates follow the tightest possible sampling distribution of unbiased estimates compared to other linear estimation methods.

Regression analysis is like any other inferential methodology. Our goal is to draw a random sample from a population and use it to estimate the properties of that population. In multivariate analysis, the coefficients within the equation are estimates of the particular population parameters.

The notation for the model of a population is the following:

$$Y = \beta_0 + \beta_{1X_1} + \beta_{2X_2} + \cdots + \beta_{kX_k}$$

The betas (β) represent the population parameter for each term in the model. Epsilon (ε) represents the random error that the model doesn't explain. Unfortunately, we'll never know these population values because it is generally impossible to measure the entire population. Instead, we'll obtain estimates of them using our random sample.

Typically, statisticians consider estimates to be useful when they are unbiased (correct on average) and precise (minimum variance). To apply these concepts to parameter estimates and the Gauss-Markov theorem, we'll need to understand the sampling distribution of the parameter estimates.

9. Explain the gradient descent algorithm in detail.

Ans. Gradient Descent is an optimization algorithm used for minimizing the cost function in various machine learning algorithms. It is basically used for updating the parameters of the learning model.

Types of gradient Descent:

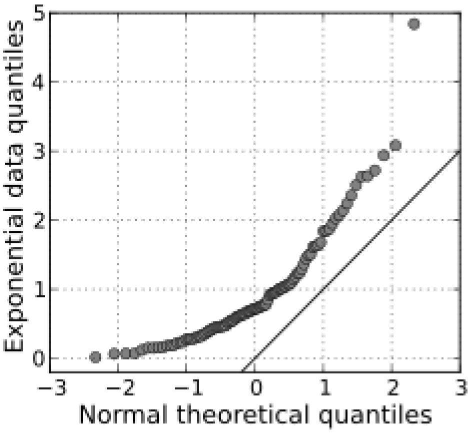
Batch Gradient Descent: this is often a kind of gradient descent which processes all the training examples for every iteration of gradient descent. But if the amount of coaching examples is large, then batch gradient descent is computationally very expensive. Hence if the amount of coaching examples is large, then batch gradient descent isn't preferred. Instead, we like better to use stochastic gradient descent or mini-batch gradient descent.

Stochastic Gradient Descent: this is often a kind of gradient descent which processes 1 training example per iteration. Hence, the parameters are being updated even after one iteration in which only a single example has been processed. Hence this is often quite faster than batch gradient descent. But again, when the number of training examples is large, even then it processes only one example which can be additional overhead for the system as the number of iterations will be quite large.

Mini Batch gradient descent: This is a type of gradient descent which works faster than both batch gradient descent and stochastic gradient descent. Here b examples where b<m are processed per iteration. So even if the number of training examples is large, it is processed in batches of b training examples in one go. Thus, it works for larger training examples and that too with lesser number of iterations.

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

Ans. 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.



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.