1. Explain the linear regression algorithm in detail.

**Answer:** Regression and classification falls under supervised learning.

**In supervised learning**, the previous year’s data would be provided with labels and the model would be trained to predict data based on the trend that was followed previous years.

**Regression analysis** is part of predictive modelling that helps to find out the relationship between input and output.

Regression analysis is used for the following three types of application:

1. Finding out upcoming trends
2. Finding out effect of input variables on target variables
3. Finding out the change in target variable with the change in one or more input variable.

Linear regression is a machine learning based on both supervised learning and regression analysis.

Linear regression is a basic form of machine learning. We train the model to predict the outcome with the provided input labels based on previous year’s data. In the case of liner regression, as the name suggests linear means two variables which are on the x-axis and y-axis are linearly correlated.

Predicting future analysis has application across industries – Economics, finance, business, medicine, Sales etc.

**For example:** We can predict the increase in customer count for the sales promotion proposed by Sales department by using the data on previous promotions to see what would be the trend followed when this sales promotion is carried out

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

**Answer:**

* 1. **Normality Assumption -** It is assumed that the error term are normally distributed. If the residuals are not normally distributed then their randomness is lost and model won’t be able to explain the relation in data
  2. **Zero mean Assumption –** It is assumed that the residuals are normally distributed around mean value of zero i.e. the error terms are normally distributed around zero.
  3. **Constant Variance assumption –** The residual terms should have a constant variance. This assumption is also known as **homogeneity or homoscedasticity**
  4. **Independent error assumption –** The residual terms should be independent of each other i.e. there should be no correlation between residuals and the predicted values or among the residuals.

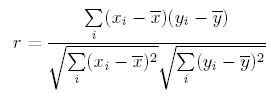
1. What is the coefficient of correlation and the coefficient of determination?

**Answer:**

* 1. **Coefficient of correlation –** Coefficient of correlation is the R value. It measures the strength and direction of a linear relationship between two variables. The formulas return a value between -1 and 1, where:
     + - 1 indicates a strong positive relationship.
       - -1 indicates a strong negative relationship.
       - A result of zero indicates no relationship at all.
       - A correlation greater than 0.8 is said to be strong correlation and less than 0.5 is said to be weak correlation

**Most commonly used formula is Pearson’s correlation coefficient.**

The formula to calculate Coefficient of correlation is:



* 1. **Coefficient of determination -** R square is also called coefficient of determination. R2 is a square of coefficient of correlation. R2 is a number which explains what portion of the given data variation is explained by the developed model.

It takes the values between 0 and 1. Overall high R-Squared the better the model fit is.

**The formula to calculate R2 = 1- RSS/TSS**

**RSS (Residual sum of squares):** In statistics, it is defined as the total sum of error across the whole sample. It is the sum difference between expected and predicted outcome. Small RSS indicates tight fit of the model.



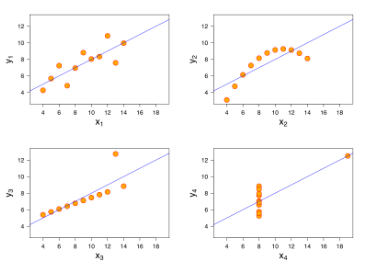
**TSS (Total sum of Squares):** Sum of errors of data points from the mean of the of theresponse variable. TSS gives us the deviation of all the points from the mean line

**TSS is given by the formula**

****

1. Explain the Anscombe’s quartet in detail.

**Answer:**



The short coming of a linear regression can be very well explained by Anscombe’s quartet. It contains 4 data sets that have nearly identical simple descriptive statistics, yet very different distribution and appear very different when graphed.

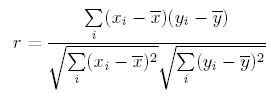
As it can be seen all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line

1. **1st graph** – The scatter plot appears to be a simple linear relationship between x and y variable
2. **2nd graph** – It shows that linear regression can only handle linear relationship and not any other data. The relationship between x and y are not linear and Pearson coefficient is not relevant in this case.
3. **3rd graph** – The distribution is linear, however the calculated regression is offset by one outlier which lowered the correlation coefficient
4. **4th graph** – Shows when one far away end point compared to other observations 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?

**Answer:** Pearson’s R correlation coefficient is designed for linear relationship between x and y variables.  It has a value between +1 and −1.

* A value of +1 is total positive linear correlation
* 0 is no linear correlation, and
* −1 is total negative linear correlation

The formula to calculate Coefficient of correlation using Pearson’s R is:



However, it might not be a good fit for non-linear relationship variables.

For finding the coefficient for non-linear relationship Spearman's R is used instead. Even though Pearson’s R gives correlation for non-linear relationships it may not be reliable.

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

**Answer:** It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. It is also called as Feature scaling.When there are multiple features in the model, lot of them might be of different scales which will lead to very weird coefficient and might be difficult to interpret.

So we need to scale the features for:

1. Easy Interpretation
2. Faster convergence of gradient descent method

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

**Normalized Scaling / Min Max Scaling:**  It bring all data within range of 0 and 1. Sklearn MinMax scaling is used for implementing normalization. It is scaled using the minimum and maximum values in data



**Standardized scaling:** Standardization replaces the values by their Z-Scores. It brings all data into a standard normal distribution which has mean zero and standard deviation one

****

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

**Answer:** Variance Inflation Factor (VIF) is used to check the presence of multi-collinearity in a dataset. It is calculated as

**An infinite VIF value** indicates that the corresponding variable may be expressed exactly by a linear combination of other variables

1. What is the Gauss-Markov theorem?

**Answer:** The Gauss-Markov theorem states that OLS has lowest sampling variance within Linear Unbiased Estimators  if the errors in the linear regression model are uncorrelated, have equal variance and expectation values is zero. Ideally, if your linear [regression](https://statisticsbyjim.com/glossary/regression-analysis/) model satisfies the first six classical assumptions, then [ordinary least squares](https://statisticsbyjim.com/glossary/ordinary-least-squares/) ([OLS](https://statisticsbyjim.com/glossary/ordinary-least-squares/)) regression produces unbiased [estimates](https://statisticsbyjim.com/glossary/estimator/) that have the smallest variance of all possible linear [estimators](https://statisticsbyjim.com/glossary/estimator/).

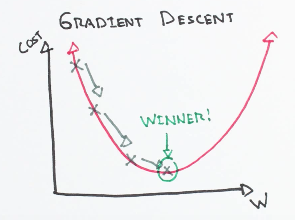
The Gauss-Markov theorem famously states that OLS is BLUE. **BLUE is an acronym for Best Linear**[**Unbiased Estimator**](https://statisticsbyjim.com/glossary/estimator/)

1. Explain the gradient descent algorithm in detail.

**Answer:** Gradient descent algorithm is one of the optimization algorithm. In linear regression, it’s used to optimize the cost function and find the value estimators corresponding to the optimized cost function.

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](https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html#glossary-parameters) of our model. Parameters refer to coefficients in [Linear Regression](https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html)

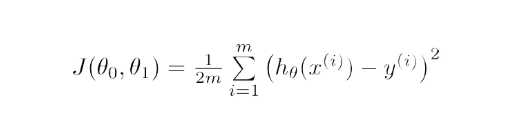
Consider we need to move from up the hill to the bottom of the hill with very low cost. Starting from the top of the hill we take a step down hill in the direction specified by the negative gradient. We keep calculating the negative gradient iteratively for taking the next step in the direction specified by it. We continue this process until we reach the bottom of the hill.



The size of each step taken is called the learning rate. If the learning rate is high we may miss the lowest slope since the slope in the hill keeps changing constantly. With low learning rate, we may not miss it since we would be re-calculating the steps over and over again. But its very time consuming.

Hence the aim of gradient function is to calculate the optimal solution and faster convergence.

The aim of gradient descent for linear regression is to find the solution of ArgMin J(Θ0,Θ1), where J(Θ0,Θ1) is the cost function of the linear regression. It is given by



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

**Answer:**  Q-Q plots (Quantile Quantile plots) are plots of two quantiles against each other. A quantile is a fraction where the values fall below it. Say Median is a quantile where 50% of the values fall below it and 50 % lies above it.

**The purpose of Q Q plots is to find out if two sets of data come from the same distribution. A normal QQ plot is one way to assess normality. Which tells us that the values are normally distributed which is one of the main assumptions of linear regression**

For example, if we run a statistical analysis that assumes our dependent variable is normally distributed, we can use a Normal Q-Q plot to check that assumption. It’s just a visual check. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that’s roughly straight. Here’s an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.

