1. **Explain the linear regression algorithm in detail.**

**Answer:** 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. 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.

**There are two types of linear regression present:**

* 1. Simple Linear Regression
* 2. Multiple Linear Regression

**1.Simple Linear Regression:**

The most elementary type of regression model is the simple linear regression which explains the relationship between the dependent variable and one independent variable using a straight line. The straight line is plotted on the scatter plot of these two points.



The standard equation of the regression line is given by the following expression:

Y = β₀ + β₁X

Where β₀ is intercept and β₁ is slope.

The strength of the linear regression model can be assessed using 2 metrics:

1. R² or Coefficient of Determination

2. Residual Standard Error (RSE)

**R² or Coefficient of Determination:**

R2 is a number which explains what portion of the given data variation is explained by the developed model. It always takes a value between 0 & 1. In general term, it provides a measure of how well actual outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model, i.e. expected outcomes. Overall, the higher the R-squared, the better the model fits your data.

Mathematically, it is represented as: R² = 1 - (RSS / TSS)

Where RSS: Residual sum of square

And TSS: Sum of errors of the data from mean

2.**Multiple Linear Regression:**

Multiple linear regression is a statistical technique to understand the relationship between one dependent variable and several independent variables. The objective of multiple regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.

Consider our previous example of sales prediction using TV marketing budget. In real life scenario, the marketing head would want to look into the dependency of sales on the budget allocated to different marketing sources. Here, we have considered three different marketing sources, i.e. TV marketing, radio marketing, and newspaper marketing. You need to consider multiple variables as just one variable alone might not be good enough to explain the feature variable, in this case, Sales.

The table below shows how adding a variable helped increase the R-squared that we had obtained by using just the TV variable.

So we see that adding more variables increases the R-squared and it might be a good idea to use multiple variables to explain a feature variable. Basically:

1. Adding variables helped add information about the variance in Y!

2. In general, we expect explanatory power to increase with increase in variables

Hence, this brings us to multiple linear regression which is just an extension to simple linear regression.

The formulation for multiple linear regression is also similar to simple linear linear regression with the small change that instead of having beta for just one variable, you will now have betas for all the variables used. The formula now can be simply given as:



Apart from the formula, a lot of other ideas in multiple linear regression are also similar to simple linear regression, such as:

1. Model now fits a ‘hyperplane’ instead of a line

2. Coefficients still obtained by minimizing sum of squared error (Least squares criterion)

3. For inference, the assumptions from from Simple Linear Regression still hold ○ Zero mean, independent, Normally distributed error terms that have constant variance

○ The inference part in multiple linear regression also, largely, remains the same.

Although, most of the ideas in simple and multiple linear regression are the same, there are a few new considerations that you need to make when moving to multiple linear regression, such as:

1. Adding more isn’t always helpful a. Model may ‘overfit’ by becoming too complex i. Model fits the train set ‘too well’, doesn’t generalize

ii. Symptoms: high train accuracy, low test accuracy

b. Multicollinearity i. Associations between predictor variables

2. Feature selection becomes an important aspect

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

**Answer: The assumption of linear regression are:**

1. **The assumption about the form of the model:** It is assumed that there is a linear relationship between the dependent and independent variables. It is know as the linear assumption.
2. **Assumptions about the residuals:**

Normality assumption: it is assumed that the error terms 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. 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 the pair-wise covariance is zero

**3. Assumptions about the estimators:**

The independent variables are measured without error.

The independent variables are linearly independent of each other i.e there is no multicollinearity in the data

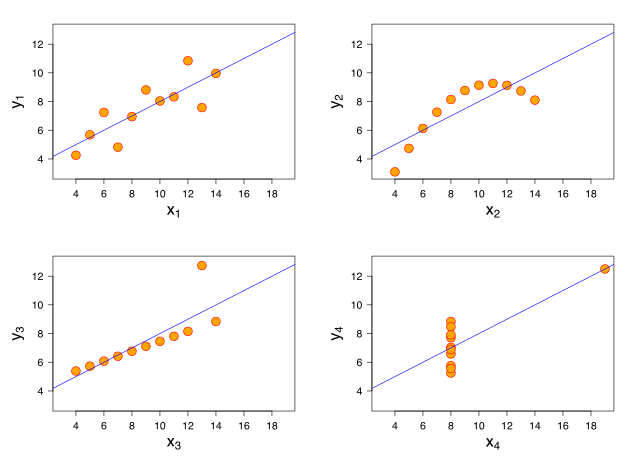
1. **What is the coefficient of correlation and the coefficient of determination?**

**Coefficient of correlation is “R”** value is the degree of relationship between two variables say x and y. It can go between -1 and 1. 1 indicates that the two variables are moving in unison. They rise and fall together and have perfect correlation. -1 means the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way.

**R square is called as coefficient of determination**. Multiply R times R to get the R square value. In other words, coefficient of Determination is the square of coefficient of correlation. It shows percentage variation in y which is explained by all the x variables together. Higher the better. It is always between 0 and 1. It can never be negative- since it is a squared value.

1. **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](https://en.wikipedia.org/wiki/Scatter_plot) (top left) appears to be a simple linear relationship, corresponding to two [variables](https://en.wikipedia.org/wiki/Variable_(mathematics)) correlated where y could be modelled as [gaussian](https://en.wikipedia.org/wiki/Normal_distribution) 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](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient) is not relevant. A more general regression and the corresponding [coefficient of determination](https://en.wikipedia.org/wiki/Coefficient_of_determination) would be more appropriate.

🡪 In the third graph (bottom left), the distribution is linear, but should have a different [regression line](https://en.wikipedia.org/wiki/Regression_line) (a [robust regression](https://en.wikipedia.org/wiki/Robust_regression) would have been called for). The calculated regression is offset by the one [outlier](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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.

Anscombe's Quartet is a great demonstration of the importance of graphing data to analyze it. Given simply variance values, means, and even [linear regressions](https://www.mathwarehouse.com/statistics/regressions/calculate-linear-regression.php) can not accurately portray data in its native form. Anscombe's Quartet shows that multiple data sets with many similar statistical properties can still be vastly different from one another when graphed.  
  
Additionally, Anscombe's Quartet warns of the dangers of outliers in data sets. Think about it: if the bottom two graphs didn't have that one point that strayed so far from all the other points, their statistical properties would no longer be identical to the two top graphs. In fact, their statistical properties would more accurately resemble the lines that the graphs seem to depict.

*how to analyze your data*. For example, while all four data sets have the [same linear regression,](https://www.mathwarehouse.com/statistics/regressions/calculate-linear-regression.php) it is obvious that the top right graph really shouldn't be analyzed with a linear regression at all because it's a curvature. Conversely, the top left graph probably *should* be analyzed with a[linear regression](https://www.mathwarehouse.com/statistics/regressions/calculate-linear-regression.php)because it's a [scatter plot](https://www.meta-chart.com/scatter-plot) that moves in a roughly [linear](https://www.mathwarehouse.com/algebra/linear_equation/) manner. These observations demonstrate the value in graphing your data before analyzing it.

Anscombe's Quartet reminds us that graphing data prior to analysis is good practice, outliers should be removed when analyzing data, and statistics about a data set do not fully depict the data set in its entirety.

1. **What is Pearson’s R?**

The Pearson correlation coefficient also referred to as Pearson’s r the Pearson product-moment correlation coefficient or the bivariate correlation is a measure of the linear correlation between two variables X and Y. It values varies from -1 to +1 where

r = 1 means the data is perfectly linear with a positive slope(i.e. both variables tend to change in the same direction)

r = -1 means the data is perfectly linear with a negative slope ( i.e. both variables tend to change in different directions)

r = 0 means there is no linear association

r > 0 > 5 means there is a weak association

r >5<8 means there is a moderate association

r > 8 means there is a strong association

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

Feature scaling(also know as data normalization) is the method used to standardize the range of features of data. Since the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.

**Scaling**: In scalling ( also called min-max scaling) , we transform the data such that the features are within a specific range e.g. [0,1].

**Standardization**: also called z-score normalization transforms your data such that the resulting distribution has a mean of 0 and a standard deviation of 1.

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

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity. If the VIF for a variable is 16 the associated standard error is four times as large as it would be if its VIF was 1. In such a case, the coefficient would have to be 4 times as large to be statistically significant at a given significance the level.

The VIF can be conceived as related to the R-squared of a predictor variable regressed on all other includes predictor variables.:

VIF of X1 = 1/ (1 - R-squared of X1 on all other Xs).

If you only have 1 X or that X is orthogonal with all the other Xs; then

VIF = 1/ (1-0) = 1 - so no variance inflation

 If two Xs are perfectly correlated

VIF = 1/ (1-1)= 1/0 = **infinity that is the estimate is as imprecise as it can be.**

1. **What is the Gauss-Markov theorem?**

* The **Gauss Markov theorem**tells us that if a [certain set of assumptions](https://www.statisticshowto.com/gauss-markov-theorem-assumptions/#assumptions) are met, the [ordinary least squares](https://www.statisticshowto.com/least-squares-regression-line/) estimate for regression coefficients gives you the *best linear unbiased estimate (BLUE)* possible.
* **Gauss Markov Assumptions**

There are five Gauss Markov assumptions (also called *conditions*):

[**Linearity**](https://www.statisticshowto.com/nonlinearity/): the [parameters](https://www.statisticshowto.com/what-is-a-parameter-statisticshowto/)we are estimating using the OLS method must be themselves linear.

**Random**: our data must have been [randomly sampled](https://www.statisticshowto.com/simple-random-sample/) from the [population](https://www.statisticshowto.com/what-is-a-population/).

[**Non-Collinearity**](https://www.statisticshowto.com/collinear/)**:** the regressors being calculated aren’t perfectly correlated with each other.

[**Exogeneity**](https://www.statisticshowto.com/exogeneity/)**:** the regressors aren’t correlated with the [error term](https://www.statisticshowto.com/error-term/).

[**Homoscedasticity**](https://www.statisticshowto.com/homoscedasticity/)**:** no matter what the values of our regressors might be, the error of the variance is constant.

🡪**Purpose of the Assumptions**

The **Gauss Markov assumptions** guarantee the [validity](https://www.statisticshowto.com/reliability-validity-definitions-examples/#Validity)of [ordinary least squares](https://www.statisticshowto.com/least-squares-regression-line/) for estimating [regression coefficients](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/#linregCoefficient).

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.

In practice, the Gauss Markov assumptions are **rarely all met perfectly**, but they are still useful as a benchmark, and because they show us what ‘ideal’ conditions would be. They also allow us to pinpoint problem areas that might cause our estimated regression coefficients to be inaccurate or even unusable.

1. **Explain the gradient descent algorithm in detail.**

**Answer**: Optimization is a big part of machine learning. Almost every machine learning algorithm has an optimization algorithm at its core.

* **Gradient Descent**: Gradient descent is an optimization algorithm used to find the values of parameters(coefficients) of a function (f) that minimized 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.
* **Gradient Descent Procedure**:

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](https://machinelearningmastery.com/learning-rate-for-deep-learning-neural-networks/) (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.

You can see how simple gradient descent is. It does require you to know the gradient of your cost function or the function you are optimizing, but besides that, it’s very straightforward. Next we will see how we can use this in machine learning algorithms.

🡪**Types of Gradient Descent algorithm:**

**There are two types of gradient descent algorithms present (**Batch gradient descent and Stochastic gradient descent)

Batch gradient descent refers to calculating the derivative from all training data before calculating an update.

Stochastic gradient descent refers to calculating the derivative from each training data instance and calculating the update immediately.

**🡪Tips for Gradient Descent**

This section lists some tips and tricks for getting the most out of the gradient descent algorithm for machine learning.

**Plot Cost versus Time:** Collect and plot the cost values calculated by the algorithm each iteration. The expectation for a well performing gradient descent run is a decrease in cost each iteration. If it does not decrease, try reducing your learning rate.

**Learning Rate:** The learning rate value is a small real value such as 0.1, 0.001 or 0.0001. Try different values for your problem and see which works best.

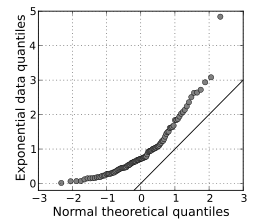
**Rescale Inputs:** The algorithm will reach the minimum cost faster if the shape of the cost function is not skewed and distorted. This can achieved this by rescaling all of the input variables (X) to the same range, such as [0, 1] or [-1, 1].

**Few Passes**: Stochastic gradient descent often does not need more than 1-to-10 passes through the training dataset to converge on good or good enough coefficients.

**Plot Mean Cost:** The updates for each training dataset instance can result in a noisy plot of cost over time when using stochastic gradient descent. Taking the average over 10, 100, or 1000 updates can give you a better idea of the learning trend for the algorithm.

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

**Answer:** The 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 the reference line.

[](https://a8h2w5y7.rocketcdn.me/wp-content/uploads/2015/08/Normal_exponential_qq.svg_.png)

A Q Q plot showing the 45 degree reference line

The (almost) straight line on the q q plot indicates the data is approximately normal.

**Use and importance of Q Q plot on linear regression:**

**Answer**: The Q Q plot helps in linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.