**Q1: Explain the linear regression algorithm in detail.**

**Ans:**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

LR (Linear Regression) predicts the value of Target Variable (y) based on the given independent variables. Basically LR finds out the relationship between input and output variables.

A simple LR model can be given using below equation:

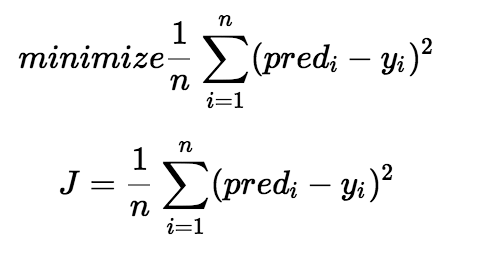
Y = b0 + b1.x

The Motive of LR is to find out the best values for b0 and b1.

The values can be found out using Cost Function and then can be updated using “Gradient Descent Function”

Cost Function:

The cost function helps us to figure out the best possible values for b0 and b1 which would provide the best fit line for the data points. Since we want the best values for b0 and b1, we convert this search problem into a minimization problem where we would like to minimize the error between the predicted value and the actual value.

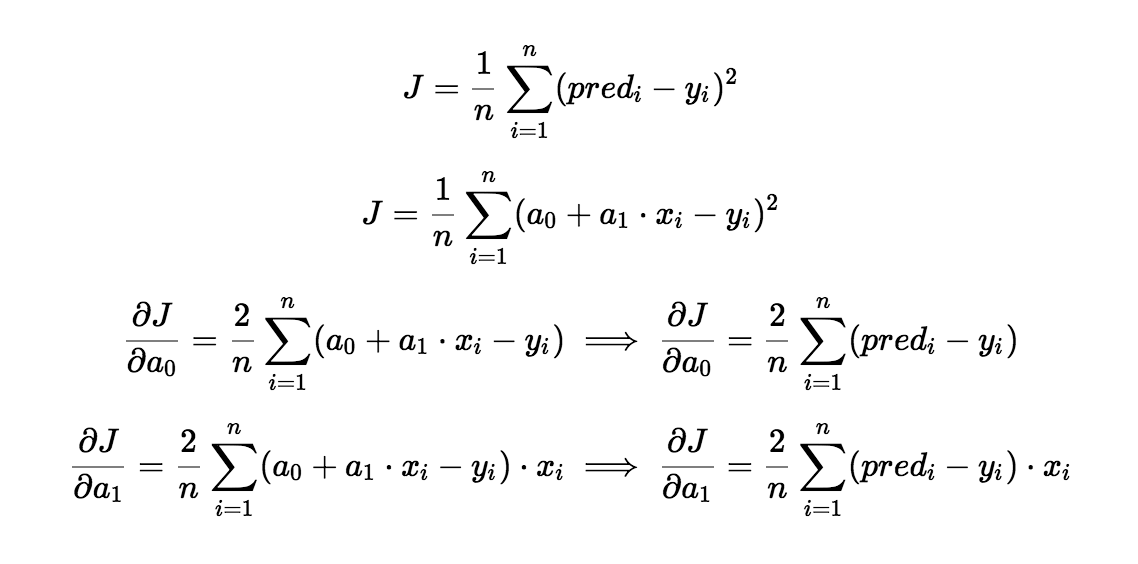


We use the above function to minimize. The difference between the predicted values and ground truth measures the error difference. We square the error difference and sum over all data points and divide that value by the total number of data points. This provides the average squared error over all the data points. Therefore, this cost function is also known as the Mean Squared Error(MSE) function. Now, using this MSE function we are going to change the values of b0 and b1 such that the MSE value settles at the minima.

Gradient Descent:

Gradient descent is a method of updating a\_0 and a\_1 to reduce the cost function(MSE). The idea is that we start with some values for a\_0 and a\_1 and then we change these values iteratively to reduce the cost. Gradient descent helps us on how to change the values.

We use the partial derivatives of the Cost Function. These are called as Gradients and use them to update the values of b0 and b1.



**Q2: What are the assumptions of linear regression regarding residuals?**

**Ans:**

Assumptions of Linear Regression regarding Residuals are as below:

1. Normality Assumption:

It is assumed that the error terms are normally distributed.

1. Zero Mean Assumption:

It is assumed that Residual have a mean value of zero i.e. the error terms are normally distributed around zero.

1. Constant variance assumption (assumption of homogeneity or homoscedasticity) :

It is assumed that the residual terms have the same variance

1. Independent error assumption:

It is assumed that the residual terms are independent of each other, i.e., their pair-wise covariance is zero.

**Q3: What is the coefficient of correlation and the coefficient of determination?**

**Ans**:

**Coefficient of Correlation:**

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 that the two variables are in perfect opposites. One goes up and other goes down, in perfect negative way.

Any two variables can be said to have a correlation value. If they are not correlated then the correlation value can still be computed which would be 0.

The correlation value always lies between -1 and 1 (going thru 0 – which means no correlation at all – perfectly not related).

**Coefficient of Determination (R^2):**

is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

R^2 is derived by squaring the value of Coefficient of Correlation (R).

The coefficient of determination, R^2, is used to analyze how differences in one variable can be explained by a difference in a second variable.

R-squared gives the percentage variation in y explained by x-variables. The range is 0 to 1 (i.e. 0% to 100% of the variation in y can be explained by the x-variables.

**Q4: Explain the Anscombe’s quartet in detail.**

**Ans:**

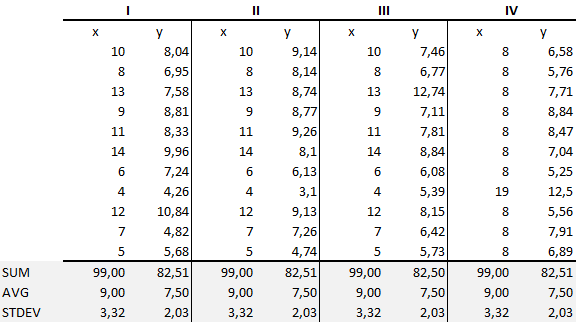
Anscombe’s Quartet was developed by statistician Francis Anscombe.

It comprises four datasets, each containing eleven (x,y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups :

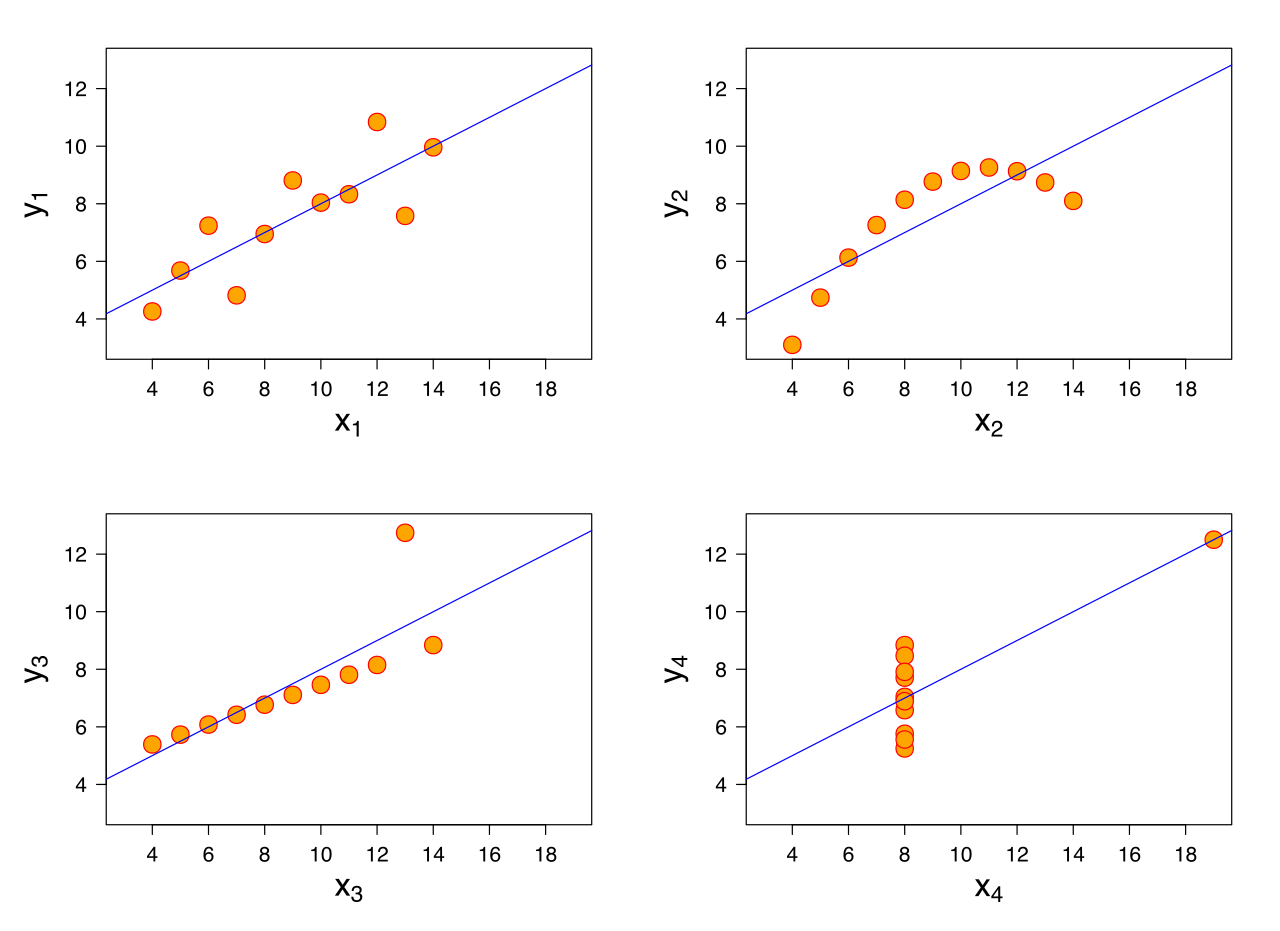
Mean of x is 9 and mean of y is 7.50 for each dataset.

Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset.

After plotting the graphs, they show the same regression line. But each data set is telling a different story.

* Dataset I appears to have clean and well-fitting linear models.
* Dataset II is not distributed normally.
* Dataset I appears to have clean and well-fitting linear models.
* Dataset II is not distributed normally.



From this we can see the importance of Visualization in Data Analysis.

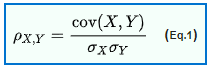
**Q5: What is Pearson’s R?**

**Ans:**

The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables.

It has values between +1 and -1 where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation

Formula:



Where

* Cov is covariance
* Sigma(x) is Standard deviation of X
* Sigma(Y) is Standard deviation of Y

Interpretation:

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

**Q6: What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

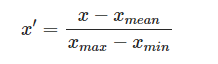
**Ans:**

Scaling is a step in data processing which is applied to independent variables or features of data. It basically helps to normalise the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should normalise when the scale is meaningful.

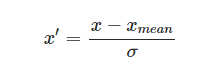
* **Normalization:**

For normalization, the maximum value we can get after applying the formula is 1, and the minimum value is 0. So all the values will be between 0 and 1.



* **Standardization:**

Standardization transforms data such that the resulting distribution has a mean of 0 and a standard deviation of 1.



**Q7: You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

**Ans:**

The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to impossible to assess accurately the contribution of predictors to a model.

VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. In order to determine VIF, we fit a regression model between the independent variables.

**How the VIF is computed:**

The standard error of an estimate in a linear regression is determined by four things:

* + The overall amount of noise (error). The more noise in the data, the higher the standard error.
  + The variance of the associated predictor variable. The greater the variance of a predictor, the smaller the standard error (this is a scale effect).
  + The sampling mechanism used to obtain the data. For example, the smaller the sample size with a simple random sample, the bigger the standard error.
  + The extent to which a predictor is correlated with the other predictors in a model.

A VIF can be computed for each predictor in a predictive model. A value of 1 means that the predictor is not correlated with other variables.

The higher the value, the greater the correlation of the variable with other variables.

Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high.

So if VIF is infinite, it means that it is highly correlated with other predictors. In other words, infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables.

**Q8: What is the Gauss-Markov theorem?**

**Ans:**

The Gauss Markov theorem says that, under certain conditions, the ordinary least squares (OLS) estimator of the coefficients of a linear regression model is the best linear unbiased estimator (BLUE), that is, the estimator that has the smallest variance among those that are unbiased and linear in the observed output variables.

We can summarize the Gauss-Markov Assumptions succinctly in algebra, by saying that a [linear regression model](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/) represented by

yi = xi‘ β + εi

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

E{εi} = 0, i = 1, . . . , N

{ε1……εn} and {x1…..,xN} are independent

cov{εi, εj} = 0, i, j = 1,…., N I ≠ j.

V{ε1 = σ2, i= 1, ….N

**Q9: Explain the gradient descent algorithm in detail.**

**Ans:**

Gradient Descent is the most common optimization algorithm in machine learning and deep learning. It is a first-order optimization algorithm. This means it only takes into account the first derivative when performing the updates on the parameters. On each iteration, we update the parameters in the opposite direction of the gradient of the objective function J(w) w.r.t the parameters where the gradient gives the direction of the steepest ascent. The size of the step we take on each iteration to reach the local minimum is determined by the learning rate α. Therefore, we follow the direction of the slope downhill until we reach a local minimum.

let’s assume that the logistic regression model has only two parameters: weight w and bias b.

1. Initialize weight w and bias b to any random numbers.

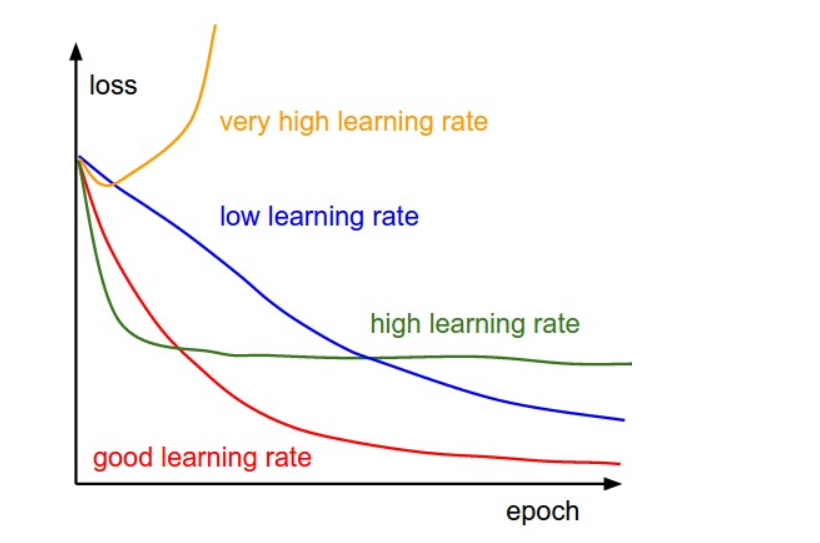
2. Pick a value for the learning rate α. The learning rate determines how big the step would be on each iteration.

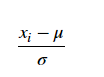
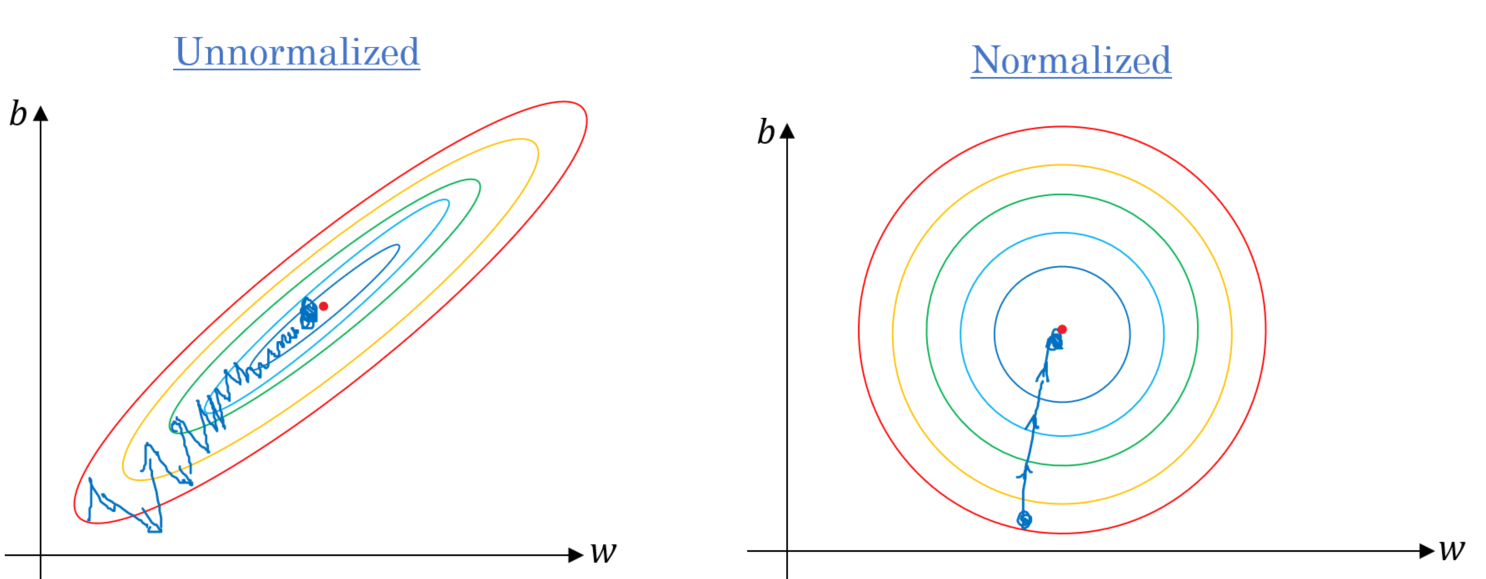
If α is very small, it would take long time to converge and become computationally expensive.

If α is large, it may fail to converge and overshoot the minimum.

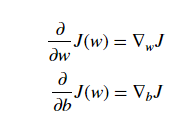
Therefore, plot the cost function against different values of α and pick the value of α that is right before the first value that didn’t converge so that we would have a very fast learning algorithm that converges .

The most commonly used rates are : 0.001, 0.003, 0.01, 0.03, 0.1, 0.3.

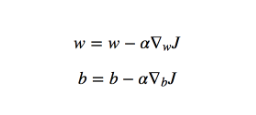
3. Make sure to scale the data if it’s on a very different scales. If we don’t scale the data, the level curves (contours) would be narrower and taller which means it would take longer time to converge 

Scale the data to have μ = 0 and σ = 1. Below is the formula for scaling each example:  

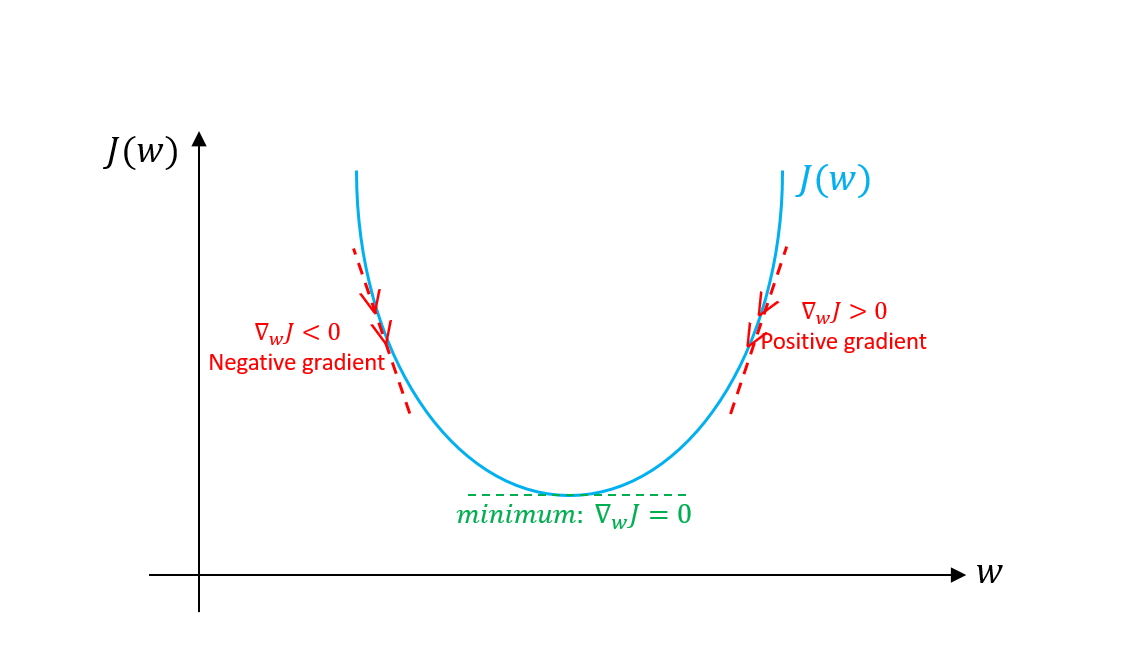
4. On each iteration, take the partial derivative of the cost function J(w) w.r.t each parameter (gradient):



The update equations are:



For the sake of illustration, let’s assume we don’t have bias. If the slope of the current value of w > 0, this means that we are to the right of optimal w\*. Therefore, the update will be negative, and will start getting close to the optimal values of w\*. However, if it’s negative, the update will be positive and will increase the current values of w to converge to the optimal values of w\*.



Continue the process until the cost function converges. That is, until the error curve becomes flat and doesn’t change.

In addition, on each iteration, the step would be in the direction that gives the maximum change since it’s perpendicular to level curves at each step.

**Q10: 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.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions. Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions