Linear Regression

Models use machine learning algorithms, during which the machine learns from the data just like humans learn from their experiences. Machine learning models can be broadly divided into two categories based on the learning algorithm which can further be classified based on the task performed and the nature of the output.

1. **Supervised learning methods:** It contains past data with labels which are then used for building the model.

- **Regression**: The output variable to be predicted is *continuous* in nature, e.g. scores of a student, diamond prices, etc.
- **Classification**: The output variable to be predicted is *categorical* in nature, e.g.classifying incoming emails as spam or ham, Yes or No, True or False, 0 or 1.

2. **Unsupervised learning methods:** It contains no predefined labels assigned to the past data.

- **Clustering**: No predefined labels are assigned to groups/clusters formed, e.g. customer segmentation.

Linear Regression is a supervised learning algorithm in machine learning that supports finding the *linear* correlation among variables. The result or output of the regression problem is a real or continuous value.

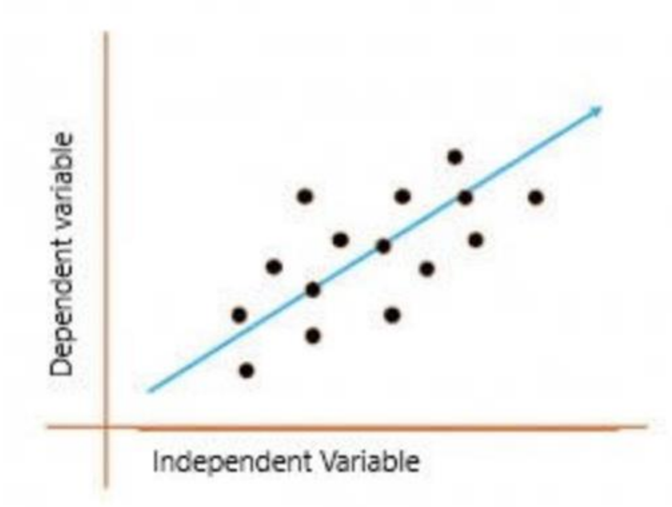
What is Linear Regression?

Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and the dependent variable, and aims to find the best-fitting line that describes the relationship. The line is determined by minimizing the sum of the squared differences between the predicted values and the actual values.

Simple Linear Regression

In a simple linear regression, there is one independent variable and one dependent variable. The model estimates the slope and intercept of the line of best fit, which represents the relationship between the variables. The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.

Linear regression is a quiet and the simplest statistical regression method used for predictive analysis in machine learning. Linear regression shows the linear relationship between the independent(predictor) variable i.e. X-axis and the dependent(output) variable i.e. Y-axis, called linear regression. If there is a single input variable **X**(independent variable), such linear regression is called ***simple linear regression***.



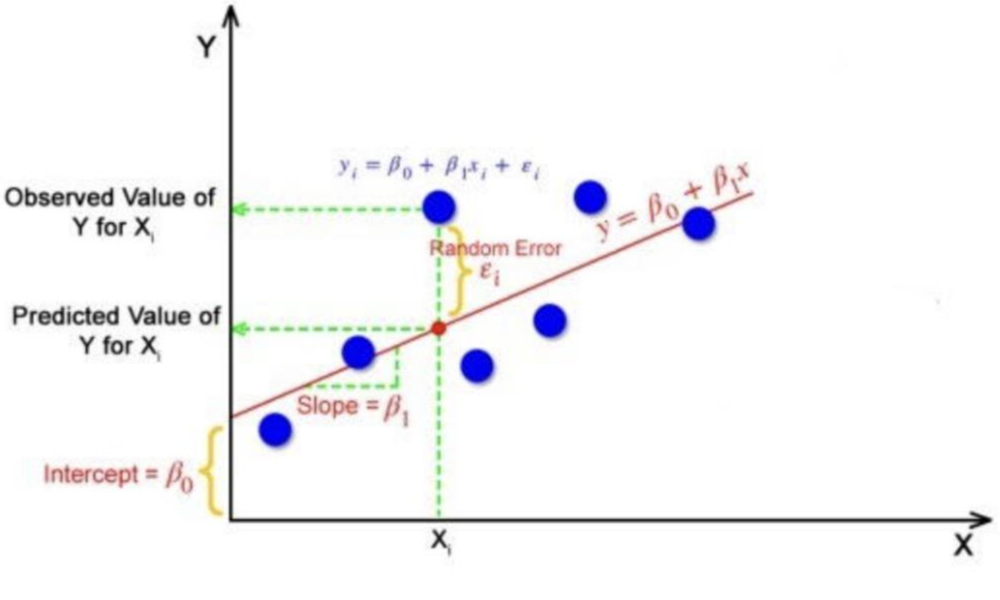
The graph above presents the linear relationship between the output(y) and predictor(X) variables. The blue line is referred to as the *best-fit* straight line. Based on the given data points, we attempt to plot a line that fits the points the best.

*T*o calculate best-fit line linear regression uses a traditional slope-intercept form which is given below,

$$\mathbf{Y_i = \beta_0 + \beta_1 X_i}$$

where Y_i = Dependent variable, β_0 = constant/Intercept, β_1 = Slope/Intercept, X_i = Independent variable.

This algorithm explains the linear relationship between the dependent(output) variable y and the independent(predictor) variable X using a straight line Y= $B_0 + B_1$ X.


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
$y = \beta_0 + \beta_1 x$
Random Error ε_i
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

But how the linear regression finds out which is the best fit line?

The goal of the linear regression algorithm is to get the **best values for B_0 and B_1** to find the best fit line. The best fit line is a line that has the least error which means the error between predicted values and actual values should be minimum.

What is the best fit line?

In simple terms, the best fit line is a line that fits the given scatter plot in the best way. Mathematically, the best fit line is obtained by minimizing the Residual Sum of Squares(RSS).

Cost Function for Linear Regression

The cost function helps to work out the optimal values for B_0 and B_1, which provides the best fit line for the data points.

In Linear Regression, generally **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the $\mathbf{y}_{\mathbf{predicted}}$ and $\mathbf{y}_{\mathbf{i}}$.

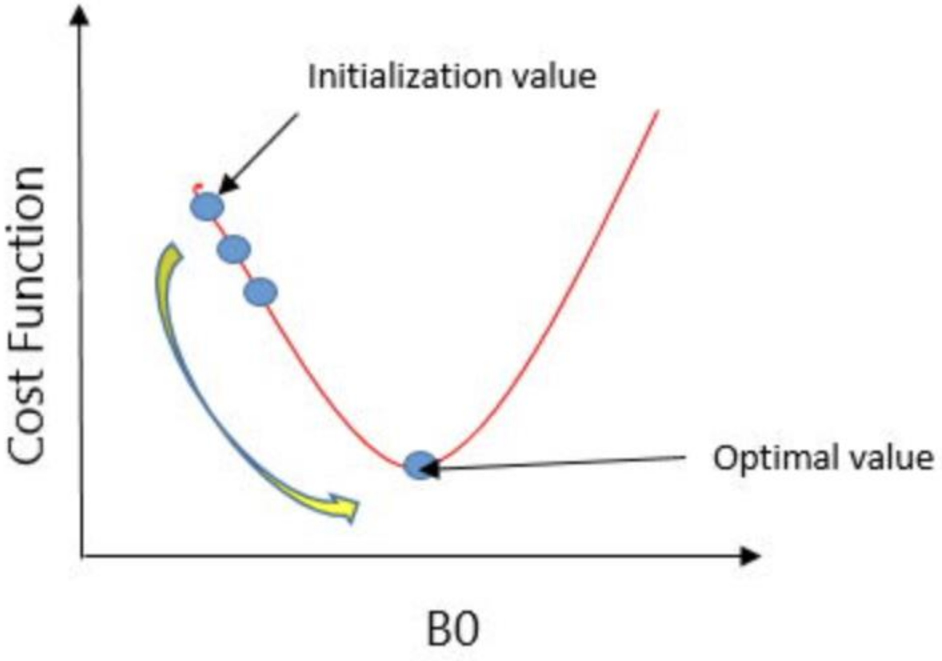
We calculate MSE using simple linear equation y=mx+b:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (\mathrm{B1}x_i + \mathrm{B0}))^2$$

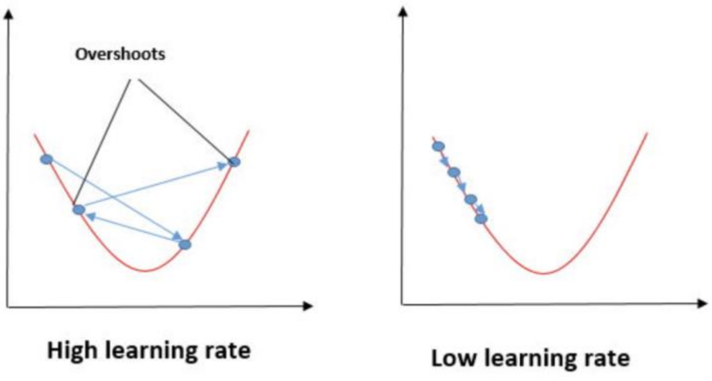
Gradient Descent for Linear Regression

Gradient Descent is one of the optimization algorithms that optimize the cost function(objective function) to reach the optimal minimal solution. To find the optimum solution we need to reduce the cost function(MSE) for all data points. This is done by updating the values of B_0 and B_1 iteratively until we get an optimal solution.

A regression model optimizes the gradient descent algorithm to update the coefficients of the line by reducing the cost function by randomly selecting coefficient values and then iteratively updating the values to reach the minimum cost function.



In the gradient descent algorithm, the number of steps you're taking can be considered as the **learning rate**, and this decides how fast the algorithm **converges** to the minima.



Evaluation Metrics for Linear Regression

The strength of any linear regression model can be assessed using various evaluation metrics. These evaluation metrics usually provide a measure of how well the observed outputs are being generated by the model.

The most used metrics are,

1. Coefficient of Determination or R-Squared (R2)
2. Root Mean Squared Error (RSME) and Residual Standard Error (RSE)

Multiple Linear Regression

Multiple linear regression is a technique to understand the relationship between a *single* dependent variable and *multiple* independent variables.

The formulation for multiple linear regression is also similar to simple linear regression with

the small change that instead of having one beta variable, you will now have betas for all the variables used. The formula is given as:

$$Y = B_0 + B_1X_1 + B_2X_2 + \ldots + B_pX_p + \boldsymbol{\varepsilon}$$

Considerations of Multiple Linear Regression

All the four assumptions made for Simple Linear Regression still hold true for Multiple Linear Regression along with a few new additional assumptions.

1. **Overfitting**: When more and more variables are added to a model, the model may become far too complex and usually ends up memorizing all the data points in the training set. This phenomenon is known as the

overfitting of a model. This usually leads to high training accuracy and very low test accuracy.

2. **Multicollinearity**: It is the phenomenon where a model with several independent variables, may have some variables interrelated.
3. **Feature Selection:** With more variables present, selecting the optimal set of predictors from the pool of given features (many of which might be redundant) becomes an important task for building a relevant and better model.