

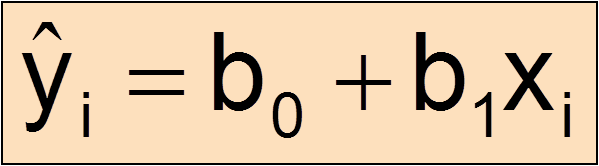
Linear Regression Algorithm :

Linear Regression Algorithm is of two types

1. Simple Linear Regression
2. Multiple Linear Regression

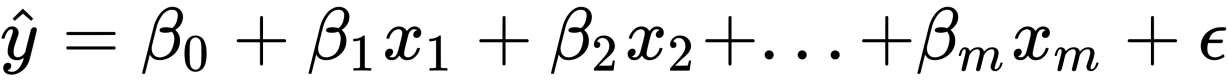
1. Simple Linear Regression

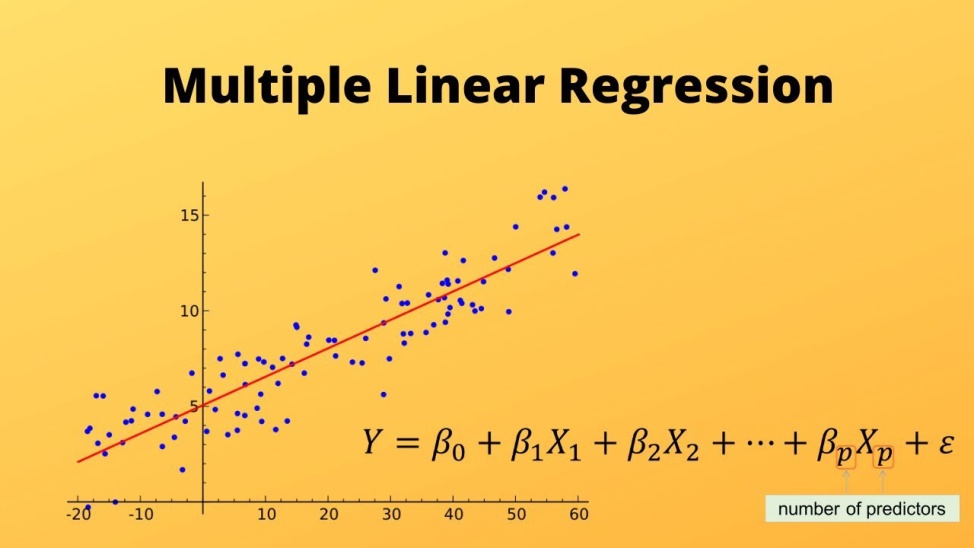
Simple Linear Regression is a regression model which defines the relationship and estimates the value of a dependent variable based on an independent variable.



2. Multiple Linear Regression

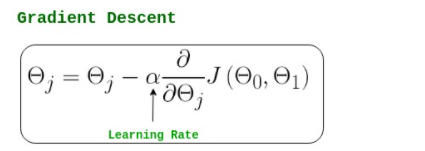
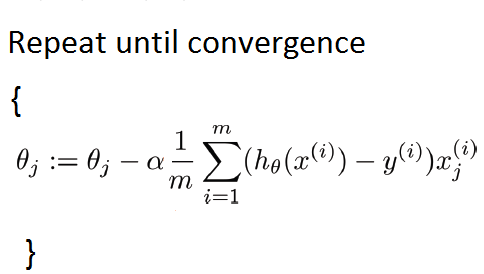
Multiple Linear Regression is a regression model which defines the relationship between one dependent variable and two or more independent variable. Here we predict the value of dependent variable based on the input features.

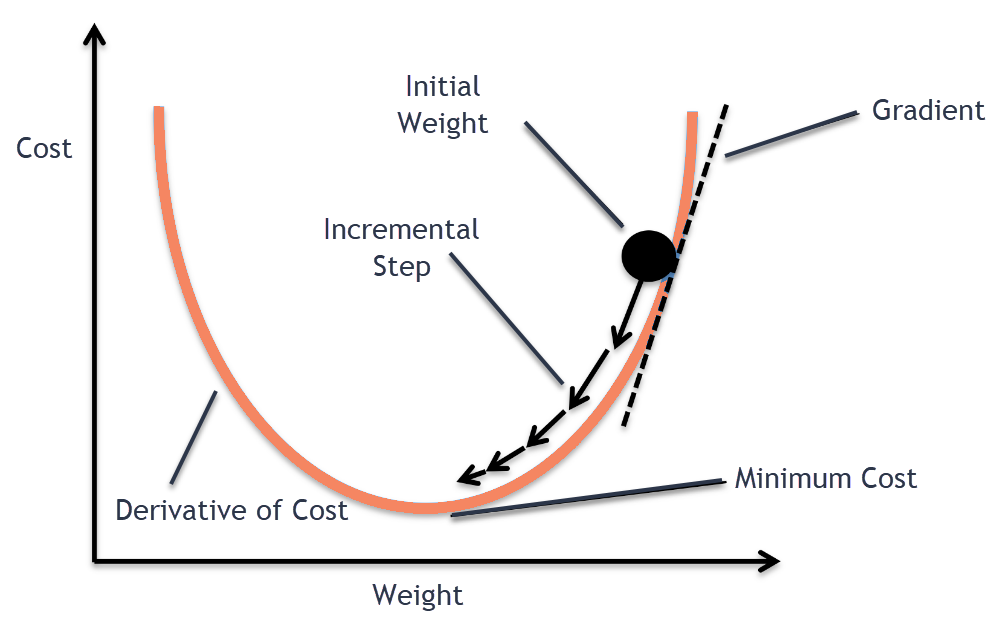




Convergence Algorithm

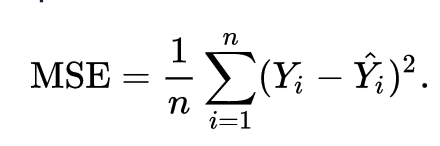
It is the Algorithm to which we iterate multiple times until we get a specific value. To get the value we complete some iterative steps, in each step we get closer to the value (each step size is getting smaller with iterations).



Cost Function

Cost function helps you to find out the point when your Machine Learning Model is most accurate by finding the relationship between input & output and how badly your model is predicting. It is the summation of average of quadratic equation of difference between actual value and predicted value.



|  |  |
| --- | --- |
| **Advantages** | **Disadvantage** |
| i. This equation is differentiable. | i. This equation is not robust to outliers. |
| ii. This equation also has one global minima. | ii. Unit is changing |

Mean Absolute Error

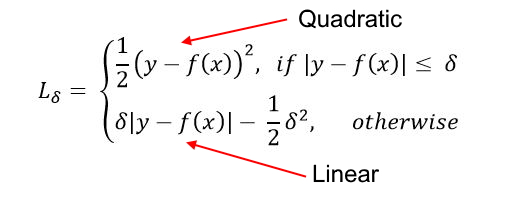
Mean Absolute Error is the measure of average absolute value of difference between actual value and absolute value.

|  |  |
| --- | --- |
| **Advantages** | **Disadvantage** |
| i. Robust to outliers. | i. Convergence usually takes more time. Optimization is a complex task. |
| ii. It will be in same unit. | ii. Time consuming. |

Huber Loss

Huber function is a loss function that is less sensitive to outliers in the dataset compared to MSE. It handles the data which are not handled by MAE(with outliers) and MSE(without outliers).

Huber loss provide function by balancing MAE & MSE



This function states that for the loss values less than **δ** use quadratic equation and for the loss values greater than **δ** use linear equation.

**RSME**

RSME (Root Mean Squared Error) is a model to get the difference between predicted values and actual values(residuals). It gets calculated by square root of Mean Square Error(MSE).



RMSE is the square root of variance, so called as standard deviation.

The lower the RSME the better the model is able to fit a dataset. The range of the dataset you are working with is important in determining whether your RSME value is low or not.

Performance Matrix

It is of two types

1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. R-squared
5. Adjusted R-squared

R-squared

R-squared score defines the performance of your model, not the absolute loss.

MAE and MSE depend on the context, where as R-squared score is independent of context.

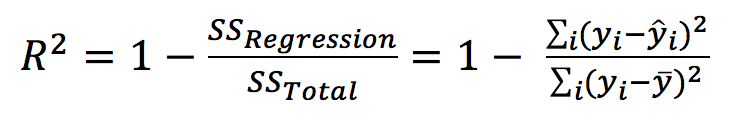
So in R-square value we have a baseline model to compare. The same we have in classification problems where threshold is fixed at 0.5.

Basically R-squared value calculates how much regression line is better than the mean line.

R-squared value ranges between 0 to 1.Which means the accuracy of your model.

If R-squared value is negative then our model is considered as very bad model.

If R-squared value is 1, then all the points fit in the line and we get the best fit line.



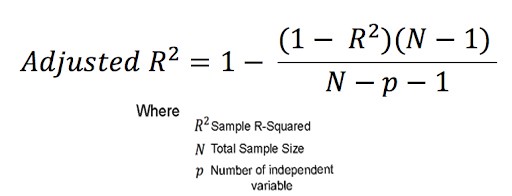
Drawbacks :

When there are multiple features in the model and features are highly correlated, the increase in accuracy will be very high, later if there will be irrelevant feature while adding them there will be less increase as there is no direct correlation.

Adjusted R-squared

Adjusted R-square came into picture to overcome the drawbacks of R-squared. It determines the accuracy of your model based on important features. (No biasness)

In adjusted R-square if there is any feature which is not highly correlated then it’ll decrease the accuracy of your model.



Over-fitting and Under-fitting

Over fitting occurs when your model fits exactly against your training data by becoming closely aligned to a limited set of data points. This leads the algorithm not to perform well against unseen data.

Under fitting means your model won’t perform well in training dataset and it may perform well in test dataset.

Training data information – Bias

Test data information – Variance

Bias : Bias is defined as the error rate of training data. When error value is high it is called as high bias and when the error value is low it is called as low bias.

Variance : The difference between the error rate of training data and test data is called variance. If the difference is high then it is called high variance and if the difference is low it is called low variance

Training data – Very Good accuracy – Low bias

* Bad Accuracy – High bias

Test data – Very Good Accuracy – Low variance

* Bad Accuracy – High variance

High bias and high variance – Under-fitting

High bias and low variance – Under-fitting

Low bias and high variance – Over-fitting

Low bias and low variance – good model