*Linear Regression*

*What is Linear Regression ?*

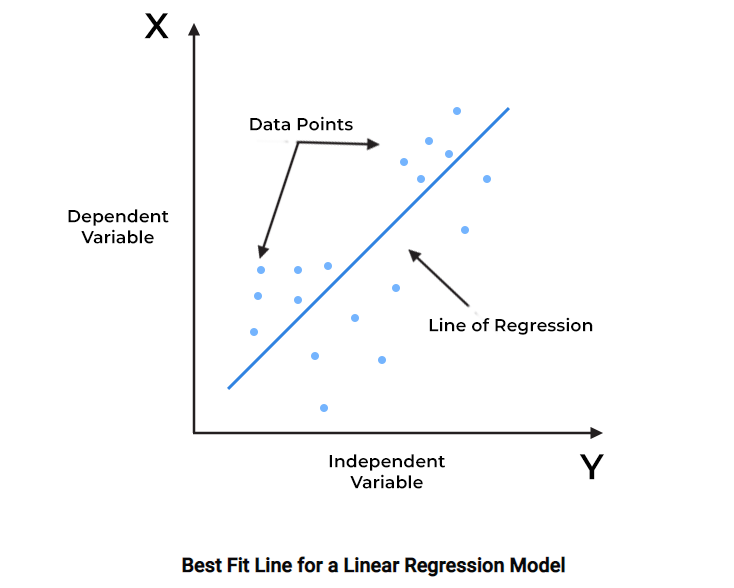
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that does not change as a result of other variables changing. The dependent variable, on the other hand, varies in response to changes in the independent variable.The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

For continuous or numerical variables like sales, salary, age, product price, etc., linear regression models a mathematical relationship between the variables and generates predictions using supervised learning techniques.

When there are at least two variables included in the data, as is the case in stock market predictions, portfolio management, scientific analysis, etc,.This analysis method is helpful.

A sloped straight line represents the linear regression model.



In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the ‘best fit line.’ The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

*Linear Regression Model Representation*

The linear function between the X and Y variables specified by the regression model best illustrates their relationship. The slant line in the above picture represents it, and the goal is to find the best regression line that matches all of the individual data points.

Mathematically these slant lines follow the following equation,

**Y = m\*X + b**

Where,

X = dependent variable (target)

Y = independent variable

m = slope of the line (slope is defined as the ‘rise’ over the ‘run’)

*Linear Regression Line:*

The least-squares method is discovered to be the most popular way for fitting a regression line in the XY plot. By minimizing the sum of the squares of the vertical deviations from each data point to the line, this procedure finds the line that fits the data the best. Linear regression determines the straight line, known as the least-squares regression line or LSRL. Suppose Y is a dependent variable and X is an independent variable, then the population regression line is given by the equation;

**Y = 0 + 1\*X + ε**

Where,

β0is the intercept of the regression line( the value of Y when X is 0)

β1is the regression coefficient line ( the change in Y for a one unit change in X)

X is the independent variable ( the variable we want to use to make predictions)

Y is the dependent variable ( the variable we want to predict

represents the error term, which captures the unexplained variation in Y.

*Properties of Linear Regression*

For the regression line where the regression parameters b0 and b1are defined, the following properties are applicable:

* The regression line reduces the sum of squared differences between observed values and predicted values.
* The regression line passes through the mean of X and Y variable values.
* The regression constant b0 is equal to the y-intercept of the linear regression.
* The regression coefficient b1 is the slope of the regression line. Its value is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X)

*Multiple Linear regression:*

In real world scenarios, relationships between variables are often more complex, and a single predictor may not be sufficient to explain the variation in the dependent variable. Multiple linear regression extends the concept of simple linear regression to accommodate multiple independent variables. The equation for multiple linear regression can be written as:

Y = β0 + β1*X1 + β2*X2 + ... + βn\*Xn + ε

Here, we have n independent variables ( X1, X2,...,Xn) with their corresponding coefficients (β1, β2, ..., βn). Each coefficient represents the change in the dependent variable associated with a one unit change in the corresponding independent variable holding other variables constant.

*Key benefits of linear regression*

Linear regression is a popular statistical tool used in data science, thanks to the several benefits it offers, such as:

1. *Easy implementation*

The linear regression model is computationally simple to implement as it does not demand a lot of engineering overheads, neither before the model launch nor during its maintenance.

1. *Interpretability*

Unlike other deep learning models (neural networks), linear regression is relatively straightforward. As a result, this algorithm stands ahead of black-box models that fall short in justifying which input variable causes the output variable to change.

1. *Scalability*

Linear regression is not computationally heavy and, therefore, fits well in cases where scaling is essential. For example, the model can scale well regarding increased data volume (big data).