Simple Linear Regression

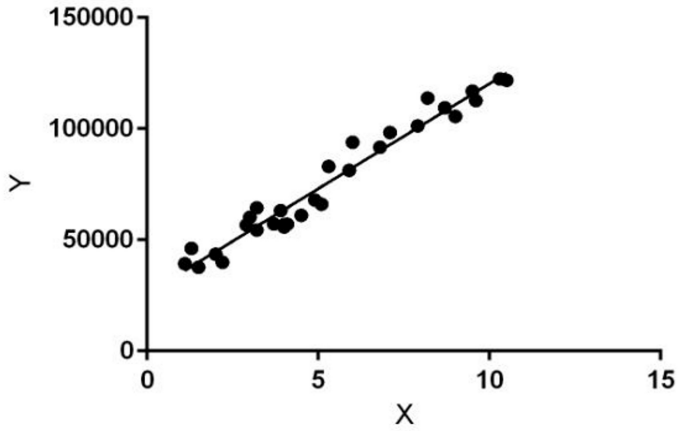
A regression model is a statistical model that estimates the relationship between one dependent variable and one or more independent variables using a line (or a plane in the case of two or more independent variables).

A regression model can be used when the dependent variable is quantitative, except in the case of logistic regression, where the dependent variable is binary.

Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**.

Simple Linear Regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line. Both variables should be quantitative.

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.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

Simple linear regression formula
The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \varepsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B0** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B1** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Error Calculation in Linear Regression Model

Linear regression most often uses mean-square error (MSE) to calculate the error of the model. MSE is calculated by:

1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. calculating the mean of each of the squared distances.

Linear regression fits a line to the data by finding the regression coefficient that results in the smallest MSE.

Assumptions of simple linear regression

Simple linear regression is a **parametric test**, meaning that it makes certain assumptions about the data. These assumptions are:

1. **Homogeneity of variance (homoscedasticity)**: the size of the error in our prediction doesn't change significantly across the values of the independent variable.

2. **Independence of observations**: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.

3. **Normality**: The data follows a normal distribution.

Linear regression makes one additional assumption:

1. The relationship between the independent and dependent variable is **linear**: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Advantages of Linear Regression

1. Linear Regression performs well when the dataset is **linearly separable**. We can use it to find the nature of the relationship among the variables.

2. Linear Regression is easier to implement, interpret and very efficient to train.

3. Linear Regression is prone to over-fitting but it can be easily avoided using some dimensionality reduction techniques, regularization (L1 and L2) techniques and cross-validation.

Disadvantages of Linear Regression

1. Main limitation of Linear Regression is the **assumption of linearity** between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. It assumes that there is a straight-line relationship between the dependent and independent variables which is incorrect many times.

2. **Prone to noise and overfitting:** If the number of observations are lesser than the number of features, Linear Regression should not be used, otherwise it may lead to overfit because is starts considering noise in this scenario while building the model.

3. **Prone to outliers:** Linear regression is very sensitive to outliers (anomalies). So, outliers should be analysed and removed before applying Linear Regression to the dataset.

4. **Prone to multicollinearity:** Before applying Linear regression, multicollinearity should be removed (using dimensionality reduction techniques) because it assumes that there is no relationship among independent variables.

Applications:

1. **Trend lines:** A trend line represents the variation in some quantitative data with passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values. However, this method suffers from a lack of scientific validity in cases where other potential changes can affect the data.
2. **Economics:** Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labour demand, and labour supply.
3. **Finance:** Capital price asset model uses linear regression to analyse and quantify the systematic risks of an investment.
4. **Biology:** Linear regression is used to model causal relationships between parameters in biological systems.