3.2 REGRESSION

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.

We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need regression analysis.

Regression is a **supervised learning technique** which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables.

It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given datapoints, using this plot, the machine learning model can make predictions about the data. In simple words, ***"Regression shows a line or curve that passes through all the datapoints on targetpredictor graph in such a way that the vertical distance between the datapoints and the regression line is minimum."*** The distance between datapoints and line tells whether a model has captured a strong relationship or not.

Some examples of regression can be as:

- Prediction of rain using temperature and other factors
- Determining Market trends
- Prediction of road accidents due to rash driving.

Terminologies Related to the Regression Analysis:

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

Why do we use Regression Analysis?

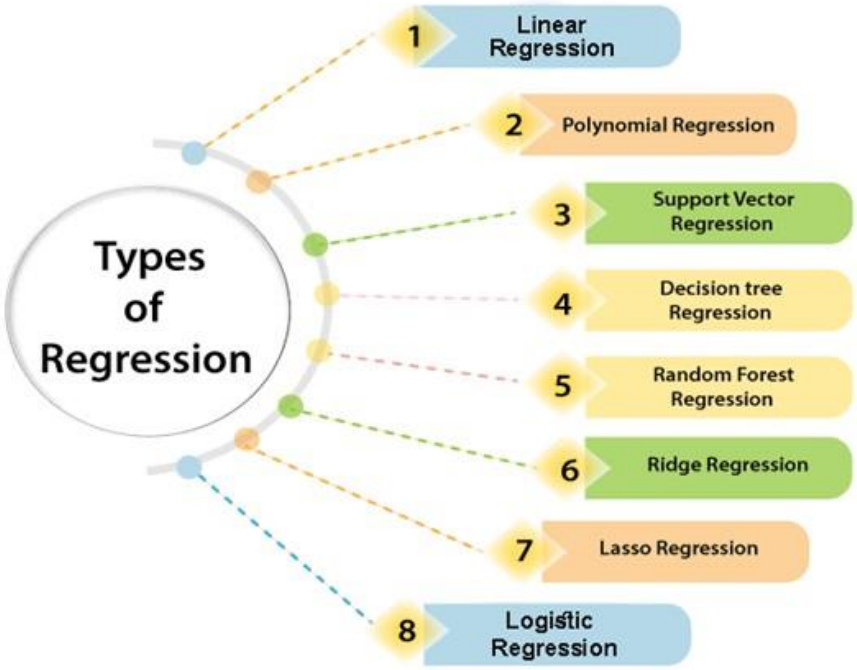
Regression analysis helps in the prediction of a continuous variable. There are various scenarios in the real world where we need some future predictions such as weather condition, sales prediction, marketing trends, etc., for such case we need some technology which can make predictions more accurately. So for such case we need Regression analysis which is a statistical method and used in machine learning and data science. Below are some other reasons for using Regression analysis:

- Regression estimates the relationship between the target and the independent variable.
- It is used to find the trends in data. ○ It helps to predict real/continuous values.
- By performing the regression, we can confidently determine the **most important factor, the least important factor, and how each factor is affecting the other factors**.

Types of Regression

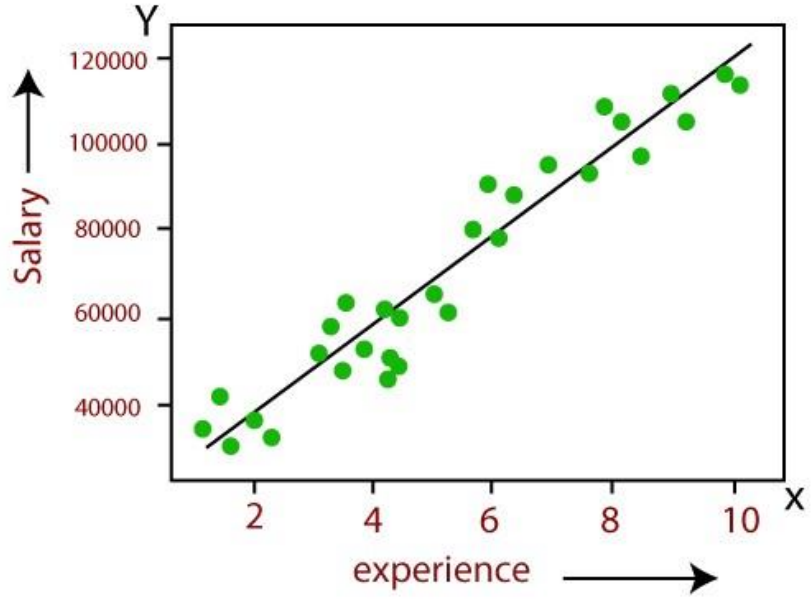
There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

- **Linear Regression ○ Logistic Regression ○ Polynomial Regression ○ Support Vector Regression ○ Decision Tree Regression ○ Random Forest Regression ○ Ridge Regression ○ Lasso Regression:**



3.2.1 LINEAR REGRESSION:

- Linear regression is a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**. ○ The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.



- Below is the mathematical equation for Linear regression:

$$Y= aX+b$$

Here, Y = dependent variables (target variables),
X= Independent variables (predictor variables), a
and b are the linear coefficients

Some popular applications of linear regression are:

- **Analyzing trends and sales estimates**
- **Salary forecasting**
- **Real estate prediction**
- **Arriving at ETAs in traffic.**

3.2.2 LEAST SQUARES

The **least square method** is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve. During the process of finding the relation between two variables, the trend of outcomes are estimated quantitatively. This process is termed as **regression analysis**. The method of curve fitting is an approach to regression analysis. This method of fitting equations which approximates the curves to given raw data is the least squares.

The least-squares method is a crucial statistical method that is practiced to find a regression line or a best-fit line for the given pattern. This method is described by an equation with specific parameters. The method of least squares is generously used in evaluation and regression. In regression analysis, this method is said to be a standard approach for the approximation of sets of equations having more equations than the number of unknowns.

The method of least squares actually defines the solution for the **minimization of the sum of squares of deviations or the errors in the result of each equation**. Find the formula for sum of squares of errors, which help to find the variation in observed data.

The least-squares method is often applied in data fitting. The best fit result is assumed to reduce the sum of squared errors or residuals which are stated to be the differences between the observed or experimental value and corresponding fitted value given in the model.

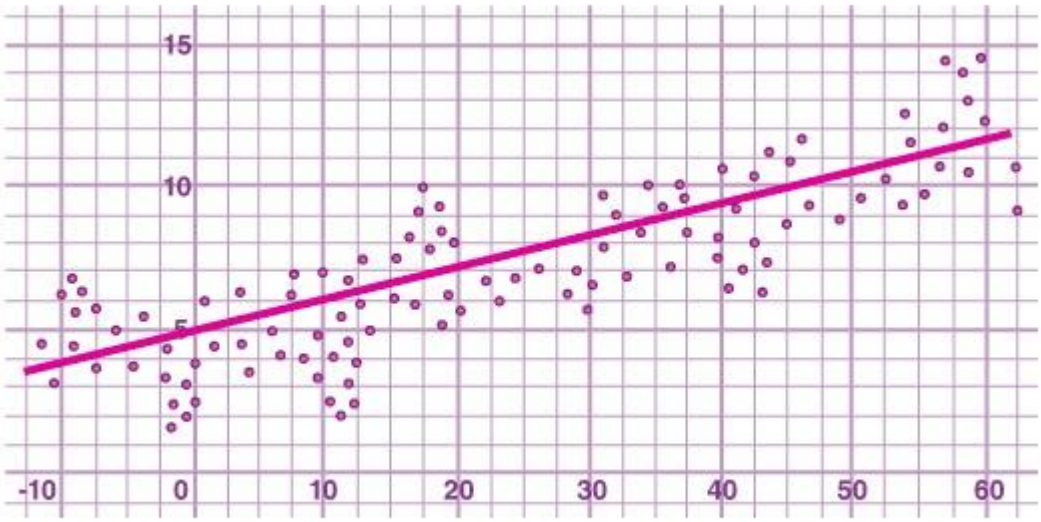
There are two basic categories of least-squares problems:

- Ordinary or linear least squares
- Nonlinear least squares

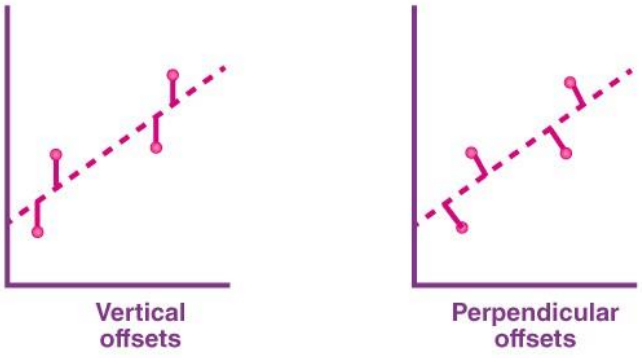
These depend upon linearity or nonlinearity of the residuals. The linear problems are often seen in regression analysis in statistics. On the other hand, the non-linear problems are generally used in the iterative method of refinement in which the model is approximated to the linear one with each iteration.

Least Square Method Graph

In linear regression, the line of best fit is a straight line as shown in the following diagram:



The given data points are to be minimized by the method of reducing residuals or offsets of each point from the line. The vertical offsets are generally used in surface, polynomial and hyperplane problems, while perpendicular offsets are utilized in common practice.



Least Square Method Formula

The least-square method states that the curve that best fits a given set of observations, is said to be a curve having a minimum sum of the squared residuals (or deviations or errors) from the given data points. Let us assume that the given points of data are (x_1, y_1), (x_2, y_2), (x_3, y_3), …, (x_n, y_n) in which all x's are independent variables, while all y's are dependent ones. Also, suppose that f(x) is the fitting curve and d represents error or deviation from each given point.

Now, we can write:

$d_1 = y_1 - f(x_1)$ d_2

$= y_2 - f(x_2)$ $d_3 =$

$y_3 - f(x_3)$ …..

$d_n = y_n - f(x_n)$

The least-squares explain that the curve that best fits is represented by the property that the sum of squares of all the deviations from given values must be minimum, i.e:

$$S = \sum_{i=1}^{n} d_i^{\,2}$$

$$S = \sum_{i=1}^{n} [y_i - f_{x_i}]^2$$

$$S = d_1^{\,2} + d_2^{\,2} + d_3^{\,2} + \cdots + d_n^{\,2}$$

Sum = Minimum Quantity

Suppose when we have to determine the equation of line of best fit for the given data, then we first use the following formula.

The equation of least square line is given by Y = a + bX

Normal equation for 'a':

$\sum Y = na + b\sum X$

Normal equation for 'b':

$\sum XY = a\sum X + b\sum X^2$

Solving these two normal equations we can get the required trend line equation.

Thus, we can get the line of best fit with formula $y = ax + b$

3.2.3 MULTIPLE REGRESSION

Regression models are used to describe relationships between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable**. You can use multiple linear regression when you want to know:

1. How strong the relationship is between two or more independent variables and one dependent variable (e.g. how rainfall, temperature, and amount of fertilizer added affect crop growth).
2. The value of the dependent variable at a certain value of the independent variables (e.g. the expected yield of a crop at certain levels of rainfall, temperature, and fertilizer addition).

Multiple linear regression example. You are a public health researcher interested in social factors that influence heart disease. You survey 500 towns and gather data on the percentage of people in each town who smoke, the percentage of people in each town who bike to work, and the percentage of people in each town who have heart disease.

Because you have two independent variables and one dependent variable, and all your variables are quantitative, you can use multiple linear regression to analyze the relationship between them.

How to perform a multiple linear regression

Multiple linear regression formula

The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

- y = the predicted value of the dependent variable
- B_0 = the y-intercept (value of y when all other parameters are set to 0)
- $B_1 X_1$ = the regression coefficient (B_1) of the first independent variable (X_1) (a.k.a. the effect that increasing the value of the independent variable has on the predicted y value)
- … = do the same for however many independent variables you are testing
- $B_n X_n$ = the regression coefficient of the last independent variable
- ϵ = model error (a.k.a. how much variation there is in our estimate of y)

3.2.4 BAYESIAN LINEAR REGRESSION

Bayesian linear regression allows a useful mechanism to deal with insufficient data, or poor distributed data. It allows user to put a prior on the coefficients and on the noise so that in the absence of data, the priors can take over. A prior is a distribution on a parameter.

If we could flip the coin an infinite number of times, inferring its bias would be easy by the law of large numbers. However, what if we could only flip the coin a handful of times? Would we guess that a coin is biased if we saw three heads in three flips, an event that happens one out of eight times with unbiased coins? The MLE would overfit these data, inferring a coin bias of p=1

A Bayesian approach avoids overfitting by quantifying our prior knowledge that most coins are unbiased, that the prior on the bias parameter is peaked around one-half. The data mast overwhelms this prior belief about coins.

Bayesian methods allow us to estimate model parameters, to construct model forecasts and to conduct model comparisons. Bayesian learning algorithms can calculate explicit probabilities for hypotheses.

Bayesian classifiers use a simple idea that the training data are utilized to calculate an observed probability of each class based on feature values.

When Bayesian classifier is used for unclassified data, it uses the observed probabilities to predict the most likely class for the new features.

Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.

Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. In Bayesian learning, prior knowledge is provided by asserting a prior probability for each candidate hypotheses and a probability distribution over observed data for each possible hypothesis.

Bayesian methods can accommodate hypotheses that make probabilistic predictions. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

Uses of Bayesian classifiers are as follows:

1. Used in text-based classification for finding spam or junk mail filtering.
2. Medical diagnosis.
3. Network security such as detecting illegal intrusion.

The basic procedure for implementing Bayesian Linear Regression is

i) Specify priors for the model parameter.

ii) Create a model mapping the training inputs to the training outputs iii) Have a Markov Chain Monte Carlo (MCMC) algorithm draw samples from the posterior distributions for the parameters