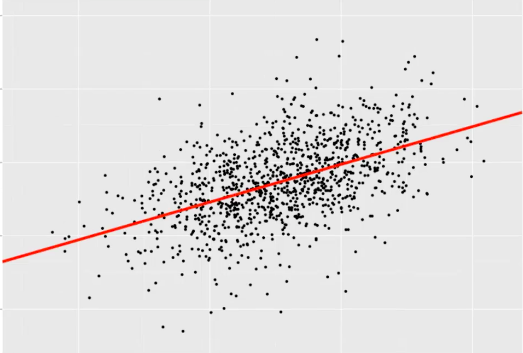
**Predictive Modelling**

|  |  |
| --- | --- |
| Roll No.: B004 | Name: Harsh Agarwal |
| Class: B.Tech. CS | Batch: B1 |

***Regression Analysis***

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variable (s) (predictor). This technique is used for forecasting, time series modelling and finding the [causal effect relationship](https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/) between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.



Regression analysis is an important tool for modelling and analysing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.  I’ll explain this in more details in coming sections.

Regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

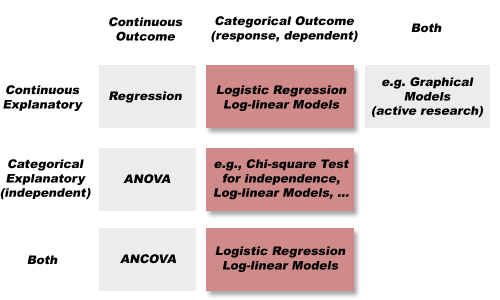
To estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.

There are multiple benefits of using regression analysis. They are as follows:

1. It indicates the **significant relationships** between dependent variable and independent variable.
2. It indicates the **strength of impact** of multiple independent variables on a dependent variable.

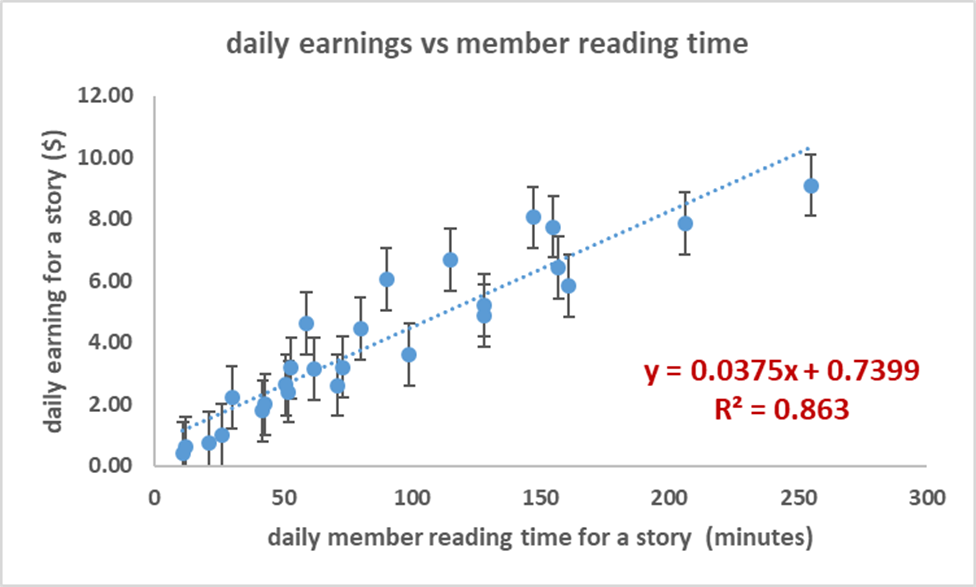
Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line).



## *Linear Regression*

It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be [continuous or discrete](https://en.wikipedia.org/wiki/Continuous_and_discrete_variables), and nature of regression line is linear. Linear Regression establishes a relationship between **dependent variable (Y)** and one or more independent variables (X) using a **best fit straight line** (also known as regression line). It is represented by an equation Y=**a+b\*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

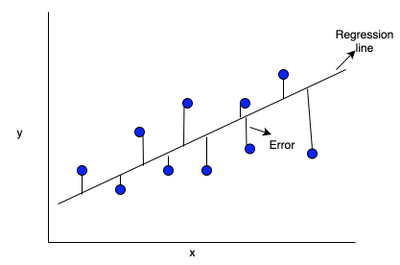


The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.  Now, the question is “How do we obtain best fit line?”.

#### **How to obtain best fit line (Value of a and b)?**

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

[least square, regression line](https://www.analyticsvidhya.com/wp-content/uploads/2015/08/Least_Square.png)



***Important Points:***

* There must be linear relationship between independent and dependent variables
* Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity.
* Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.
* Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable
* In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.