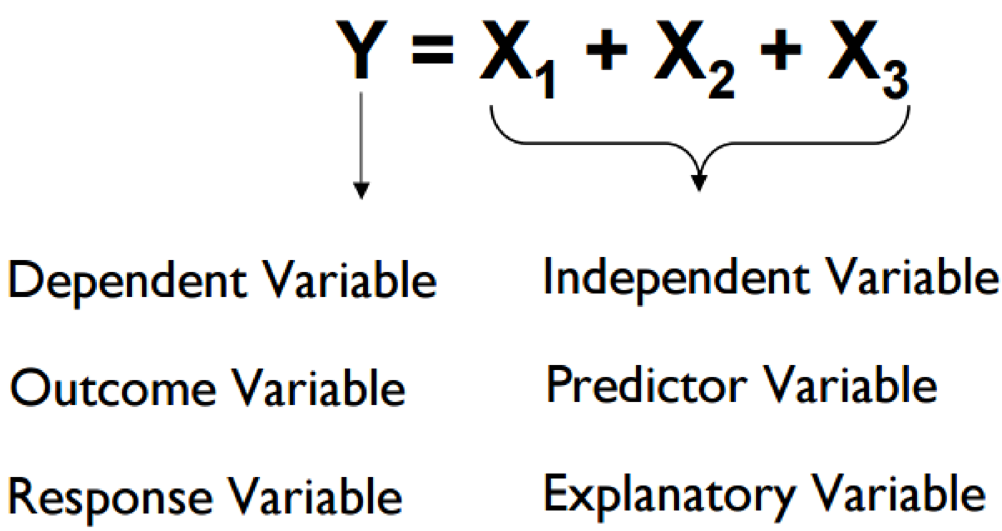
What is Linear Regression?

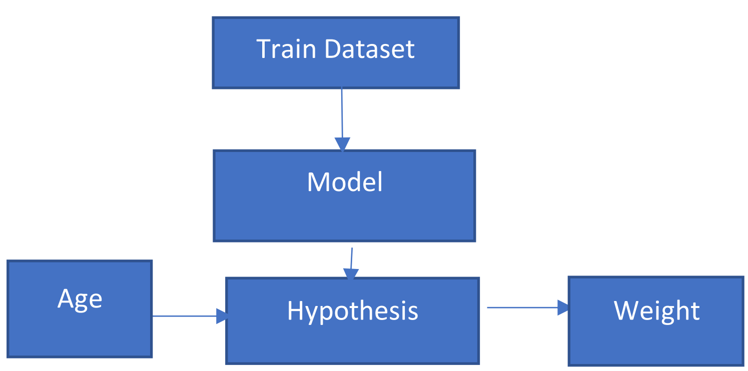
Linear Regression is a statistical supervised learning technique to predict the quantitative variable by forming a linear relationship with one or more independent features.

It helps determine:

→ If a independent variable does a good job in predicting the dependent variable.

→ Which independent variable plays a significant role in predicting the dependent variable.





Getting back to regression**, it is a supervised machine learning technique to understand the trend or relationship between 2 or more variables. Regression comes from the term** 'Regress' **which means predicting one variable from a set of variables. The variable to be predicted is called** dependent **variable and the variables used in predictions are called** independent **variables. Regression typically involves identifying the relationships or correlations between the dependent and independent variables, however it cannot explain any causal relationship between the variables. Example of regression would be one where we are trying to predict the price of houses in Florida based on different independent variables such as size of the house and number of rooms in the house.**

Assumptions of Linear Regression:

- **The *Independent variables* should be linearly related to the *dependent variables*.**
  This can be examined with the help of several visualization techniques like: Scatter plot or maybe you can use Heatmap or pairplot(to visualize every features in the data in one particular plot).

- **Every feature in the data is Normally Distributed.**
  This again can be checked with the help of different visualization Techniques,such as Q-Q plot,histogram and much more.

- **There should be little or no multi-collinearity in the data.**

The best way to check the prescence of multi-collinearity is to perform VIF(Variance Inflation Factor).

- **The mean of the residual is zero.**
  A **residual** is the difference between the observed y-value and the predicted y-value.However, Having residuals closer to zero means the model is doing great.

- **Residuals obatined should be normally distributed.**
  This can be verified using the Q-Q Plot on the residuals.

- **Variance of the residual throughout the data should be same.This is known as homoscedasticity.**
  This can be checked with the help of residual vs fitted plot.

- **There should be little or no Auto-Correlation is the data.**
  Auto-Correlation Occurs when the residuals are not independent of each other.This usally takes place in time series analysis.
  You can perform Durbin-Watson test or plot ACF plot to check for the autocorrelation.If the value of Durbin-Watson test is 2 then that means no autocorrelation,If value < 2 then there is positive correlation and if the value is between >2 to 4 then there is negative autocorrelation.

If the features in the dataset are not normally distributed try out different transformation

techniques to transform the distribution of the features present in the data.

Understanding the Slope and intercept in the linear regression model:

What is a slope?

In a regression context, the slope is very important in the equation because it tells you how much you can expect *Y* to change as *X* increases.

It is denoted by m in the formula $y = mx+b$.

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

It can also be calculated by the formula,

m = r(Sy/Sx),*

Where r is the correlation co-efficient.

Sy and Sx is the standard deviation of x and y.

And r can be calculated as

$$r = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

What is Intercept?

The *y*-intercept is the place where the regression line $y = mx + b$ crosses the *y*-axis (where $x = 0$), and is denoted by *b*.

Formula to calculate the intercept is:

$$Slope \quad a = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

$$intercept \quad b = \overline{y} - a\overline{x}$$

$$\_\_where \quad overbar \quad denotes \quad average$$

Now,Put this slope and intercept into the formula (y = mx +b) and their you have the description of the best fit line.

Cost Function of Linear Regression

Cost Function is a function that **measures the performance of a Machine Learning model** for given data.

Cost Function is basically the calculation of the error between predicted values and expected values and **presents it in the form of a single real number**.

Many people gets confused between **Cost Function** and **Loss**

Function,
Well to put this in simple terms **Cost Function** is the average of error of n-sample in the data and **Loss Function** is the error for individual data points.In other words,**Loss Function** is for one training example,**Cost Function** is the for the entire training set.

So,When it's clear what cost function is Let's move on.

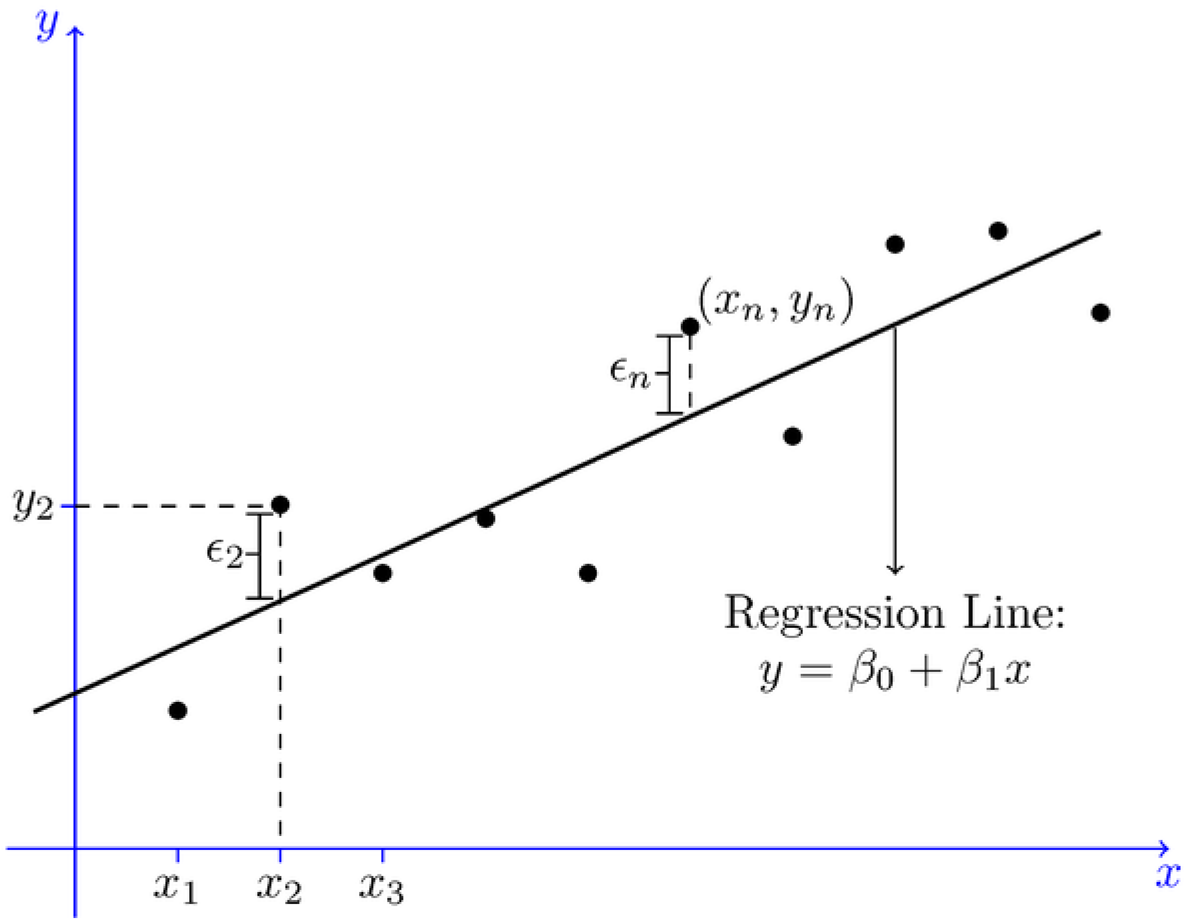
The Cost Function of a linear Regression is taken to be Mean Squared Error.
some.People may also take Root Mean Square Error.Both are basically same,However adding a Root significantly reduces the value and makes it easy to read.We,take Square here so that we don't get values in negative.

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

Here, n is the total number of data in the dataset.

You must be wondering where does the slope and intercept comes into play here!!

```
J = 1/n*sum(square(pred - y))Which, can also be written as :J =
1/n*sum(square(pred-(mx+b)))  i.e, y = mx+b
```



We want the Cost Function to be 0 or close to zero,which is the best possible outcome one can get.

Interpreting the results of Linear Regression:

You have cleaned the data and passed it on to the model,Now the question arises,***How do you know if your Linear Regression Model is Performing well?***

For That we Use the Statsmodel Package in Python and After fitting the data we do a **.Summary()** on the model,which gives result as shown in the pic below.(P.S. — I used a pic from google images)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     6.646
Date:                Sun, 18 Feb 2018   Prob (F-statistic):            0.00157
Time:                        15:18:50   Log-Likelihood:                -12.978
No. Observations:                  32   AIC:                             33.96
Df Residuals:                      28   BIC:                             39.82
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4639      0.162      2.864      0.008       0.132       0.796
x2             0.0105      0.019      0.539      0.594      -0.029       0.050
x3             0.3786      0.139      2.720      0.011       0.093       0.664
const         -1.4980      0.524     -2.859      0.008      -2.571      -0.425
==============================================================================
Omnibus:                        0.176   Durbin-Watson:                   2.346
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
Skew:                           0.141   Prob(JB):                        0.920
Kurtosis:                       2.786   Cond. No.                         176.
==============================================================================
```

Now,If you Look at the Pic Carefully you will see a bunch of different Statistical test.
I Suppose you are familiar with R-Squared and Adjusted R-Squared shown on the top right of the image,If you don’t no worries read my blog about R-Squared and P-value.

Here we will see what the lower block of the image interprets.

```
==============================================================================
Omnibus:                        0.176   Durbin-Watson:                   2.346
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
Skew:                           0.141   Prob(JB):                        0.920
Kurtosis:                       2.786   Cond. No.                         176.
==============================================================================
```

- **Omnibus/Prob(Omnibus)**:It is statistical test that tests the skewness and Kurtosis of the residual.
  A value of Omnibus close to 0 show the normalcy(normally distributed) of the residuals.
  A value of Prob(Omnibus) close to 1 show the probability that the residuals are normally distributed.

- **Skew:** It is a measure of the symmetry of data,values closer to 0 indicates the residual distribution is normal.

- **Kurtosis:** It is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.
  Greater Kurtosis can be interpreted as a tighter clustering of residuals around zero, implying a better model with few outliers.

- **Durbin-Watson:**It is a statistical test to detect any auto-correlation at a lag 1 present in the residuals.
  → value of test is always between 0 and 4
  → IF value = 2 the there is no auto-correlation
  → IF value greater than(>) 2 then there is negative auto-correlation,which means that the positive error of one observation increases the chance of negative error of another observation and vice versa.
  → IF value less than(<) 2 the there is positive auto-correlation.

- **Jarque-Bera/Prob(Jarque-Bera):**It is a Statistical test which test a goodness of fit of whether the sample data has skewness and kurtosis matching the normal distribution. Prob(Jarque-Bera) indicates normality of the residuals.

- **Condition Number:**This is a Statistical Test that measures the sensitivity of a function's output as compared to its input. When there is multicollinearity present , we can expect much higher fluctuations to small changes in the data,So the value of the test will be very high.
  A lower value is expected,something below 30,or more specifically value closer to 1.

What is R squared –

R-Squared statistic

— As RSE is measured in the units of Y we are never sure of what value is a good RSE. But R-squared is measured as proportion of variability in Y that can be explained using X and always will be range of 0 to 1 unlike RSE.

— Formula of **R-squared** is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\mathrm{TSS} = \sum (y_i - \bar{y})^2$$

— Total Sum of Squares measures the total variance or the inherent variance present in the response variable Y before the regression was performed.

— A value near 0 implies the model is unable to explain variance and a value close to 1 says model is able to capture the variability. **A good performing model would have the R2 score close to 1**

Tow ways to do

1) Ordinary least square
2) Gradient descent

   $Y = b0+b1x$

   $Y = m+cx$