

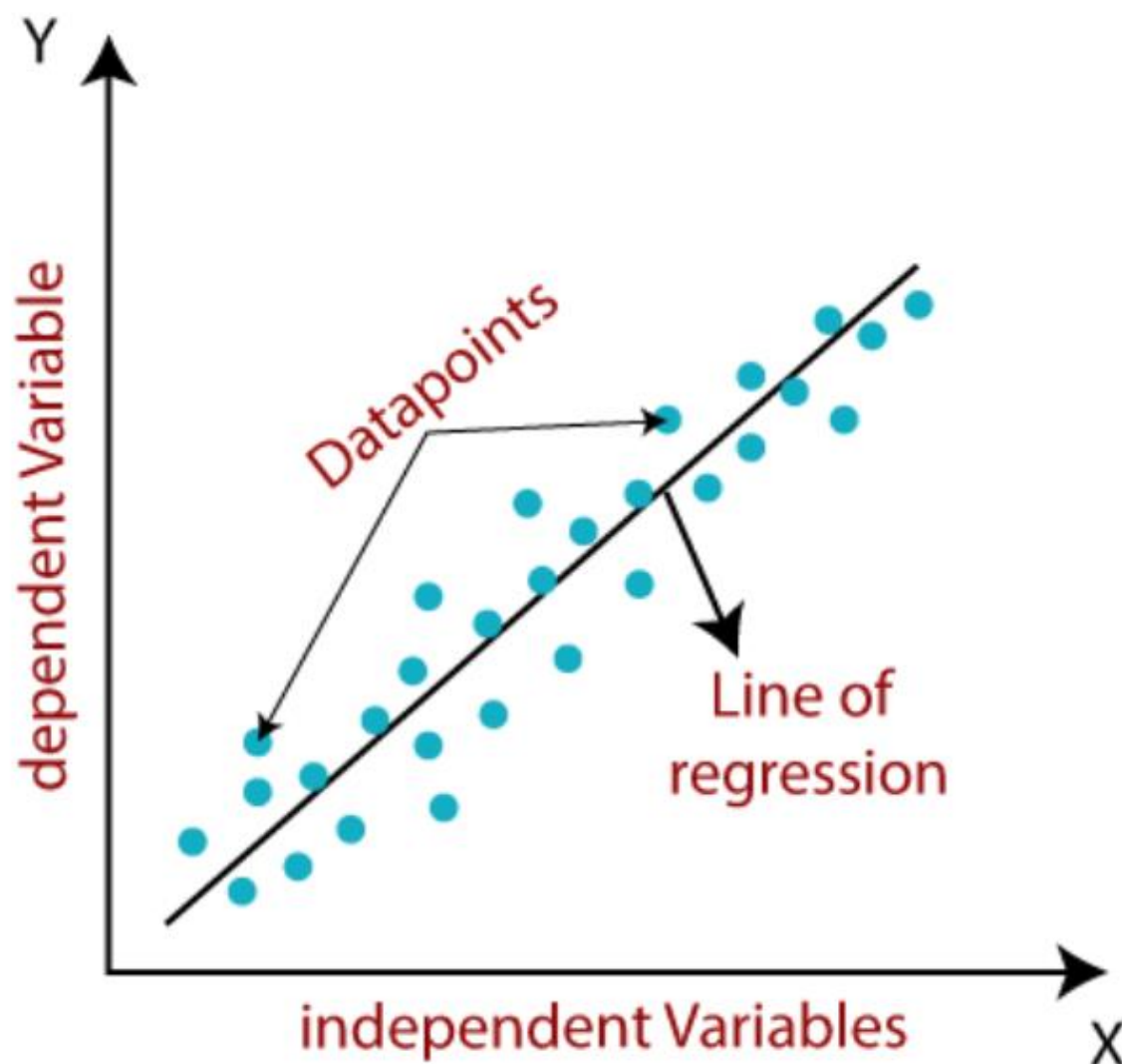
## 1. What is Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Example you want to predict weight with respect to height

So here height will be independent feature and weight will be dependent

Linear Regression Comes under the Category of Supervised Machine learning



Linear Regression can be represented by equation

$$H_0(x) = \theta_0 + \theta_1 x$$

Where  $\Theta_0$  is intercept and  $\Theta_1$  slope

If we consider error then linear regression can be represented as

$$y = a_0 + a_1x + \varepsilon$$

Y = Dependent Variable (Target Variable)

X = Independent Variable (predictor Variable)

$a_0$  = intercept of the line (Gives an additional degree of freedom)

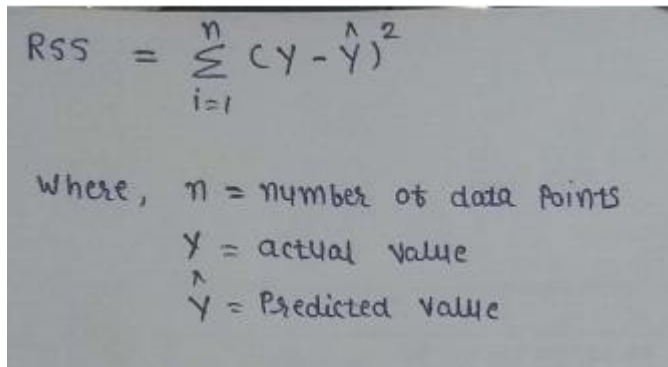
$a_1$  = Linear regression coefficient (scale factor to each input value).

$\varepsilon$  = random error

## 2. How do you calculate error

We can calculate error by following methods

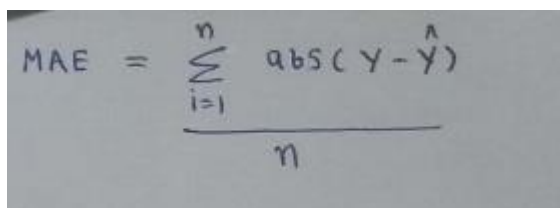
### 1. Residual Sum Of Squares : (RSS)



$$RSS = \sum_{i=1}^n (y - \hat{y})^2$$

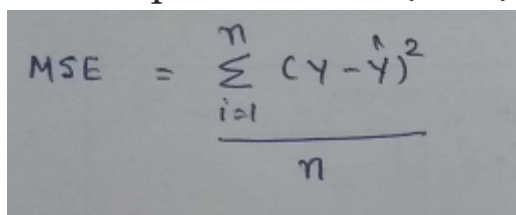
Where,  $n$  = number of data points  
 $y$  = actual value  
 $\hat{y}$  = Predicted value

### 2. Mean Absolute Error : (MAE)



$$MAE = \frac{\sum_{i=1}^n \text{abs}(y - \hat{y})}{n}$$

### 3. Mean Squared Error : (MSE)



$$MSE = \frac{\sum_{i=1}^n (y - \hat{y})^2}{n}$$

### 4. Mean Absolute percentage Error (MAPE) :

$$MAPE = \frac{\sum_{i=1}^n \text{abs}((Y - \hat{Y})/Y)}{n} \times 100$$

### 5. Mean Percentage Error : (MPE)

$$MPE = \frac{\sum_{i=1}^n ((Y - \hat{Y})/Y)}{n} \times 100$$

### 6. Root Mean Squared Error : (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y - \hat{Y})^2}{n}}$$

## 3. What is the difference between cost and loss function?

### Loss Functions

The loss function quantifies how much a model's prediction deviates from the ground truth for one particular object. So, when we calculate loss, we do it for a single object in the training or test sets.

There are many different loss functions we can choose from, and each has its advantages and shortcomings. In general, any distance metric defined over the space of target values can act as a loss function.

### Example:

For instance, let's say that our model predicts a flat's price (in thousands of dollars) based on the number of rooms, area (m<sup>2</sup>), floor, and the neighbourhood in the city (A or B). Let's suppose that its prediction for is USD 105 k. If the actual selling price is USD 96 k, then the square loss is:

$$L_{\text{square}}(105, 96) = (105 - 96)^2 = (11)^2 = 121$$

### Cost Functions

**the cost function measures the model's error on a group of objects, whereas the loss function deals with a single data instance.**

So, if  $L$  is our loss function, then we calculate the cost function by aggregating the loss over the training, validation, or test data

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n.$$

For example, we can compute the cost as the mean loss:

$$Cost(f, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n L(\hat{y}_i, y_i) \quad (\hat{y}_i = f(\mathbf{x}_i))$$

We can use various function to calculate cost like Mean Square Error, Root Mean Square Error

#### 4. What is MAE, MSE and RMSE?

MSE: Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2$$

Where,

$\hat{y}$  – predicted value of  $y$   
 $\bar{y}$  – mean value of  $y$

MAE: The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}|$$

RMSE: Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

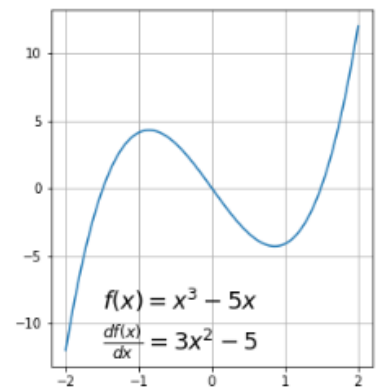
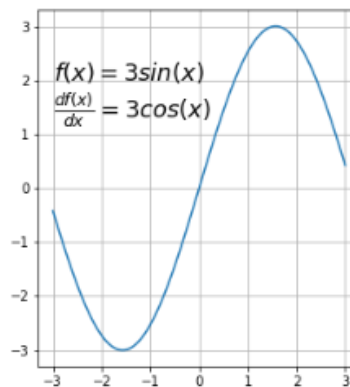
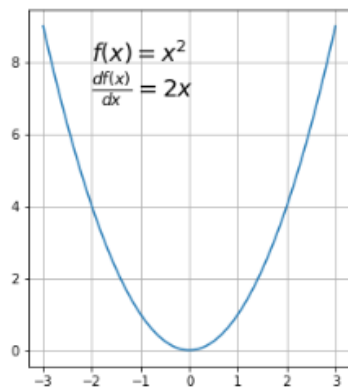
## 5. What is Gradient Descent?

**Gradient descent** (GD) is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function. This method is commonly used in *machine learning* (ML) and *deep learning* (DL) to minimise a cost/loss function (e.g., in a linear regression). Due to its importance and ease of implementation, this algorithm is usually taught at the beginning of almost all machine learning courses.

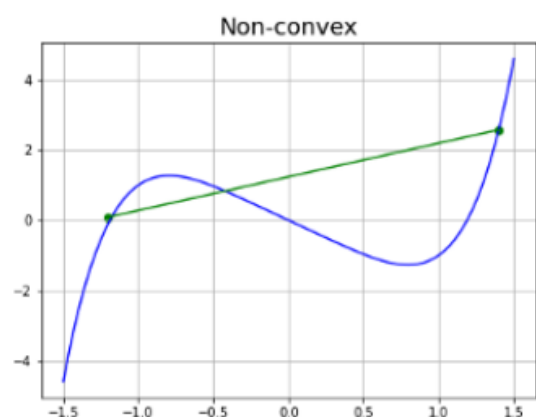
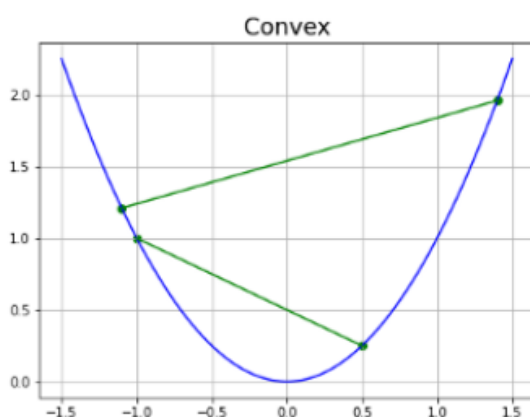
Function requirements Gradient descent algorithm does not work for all functions. There are two specific requirements. A function has to be:

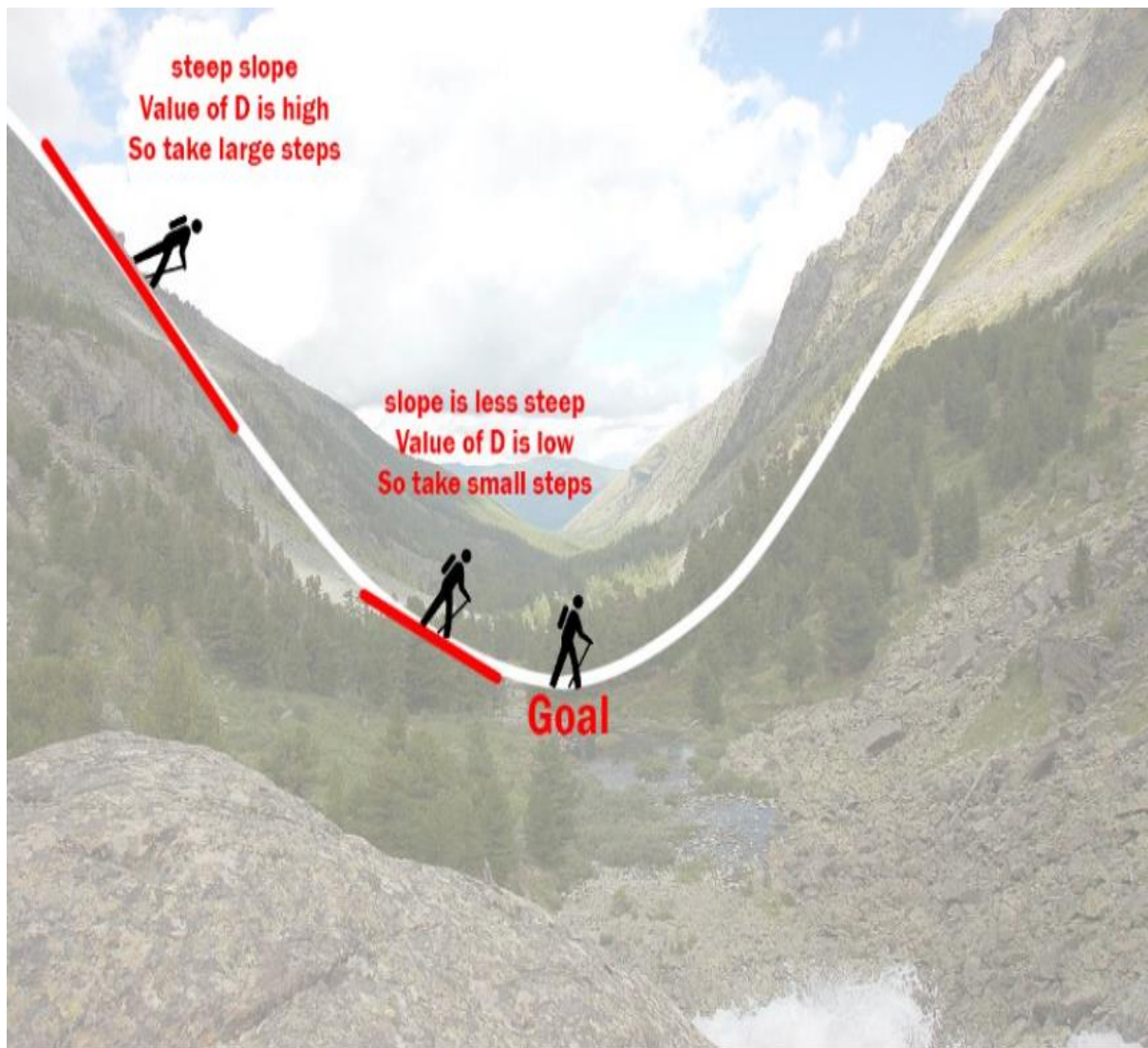
1. **Differentiable**
2. **Convex**

First, what does it mean it has to be **differentiable**? If a function is differentiable, it has a derivative for each point in its domain — not all functions meet these criteria. First, let's see some examples of functions meeting this criterion:



**function has to be convex.** For a univariate function, this means that the line segment connecting two function's points lies on or above its curve (it does not cross it). If it does it means that it has a local minimum which is not a global one.





Imagine a valley and a person with no sense of direction who wants to get to the bottom of the valley. He goes down the slope and takes large steps when the slope is steep and small steps when the slope is less steep. He decides his next position based on his current position and stops when he gets to the bottom of the valley which was his goal.

Let's try applying gradient descent to  $m$  and  $c$  and approach it step by step:

1. Initially let  $m = 0$  and  $c = 0$ . Let  $L$  be our learning rate. This controls how much the value of  $m$  changes with each step.  $L$  could be a small value like 0.0001 for good accuracy.
2. Calculate the partial derivative of the loss function with respect to  $m$ , and plug in the current values of  $x$ ,  $y$ ,  $m$  and  $c$  in it to obtain the derivative value  $D$ .

$$D_m = \frac{1}{n} \sum_{i=0}^n 2(y_i - (mx_i + c))(-x_i)$$

$$D_m = \frac{-2}{n} \sum_{i=0}^n x_i(y_i - \bar{y}_i)$$

Derivative with respect to m

$D_m$  is the value of the partial derivative with respect to  $\mathbf{m}$ . Similarly let's find the partial derivative with respect to  $\mathbf{c}$ ,  $D_c$  :

$$D_c = \frac{-2}{n} \sum_{i=0}^n (y_i - \bar{y}_i)$$

Derivative with respect to c

3. Now we update the current value of  $\mathbf{m}$  and  $\mathbf{c}$  using the following equation:

$$m = m - L \times D_m$$

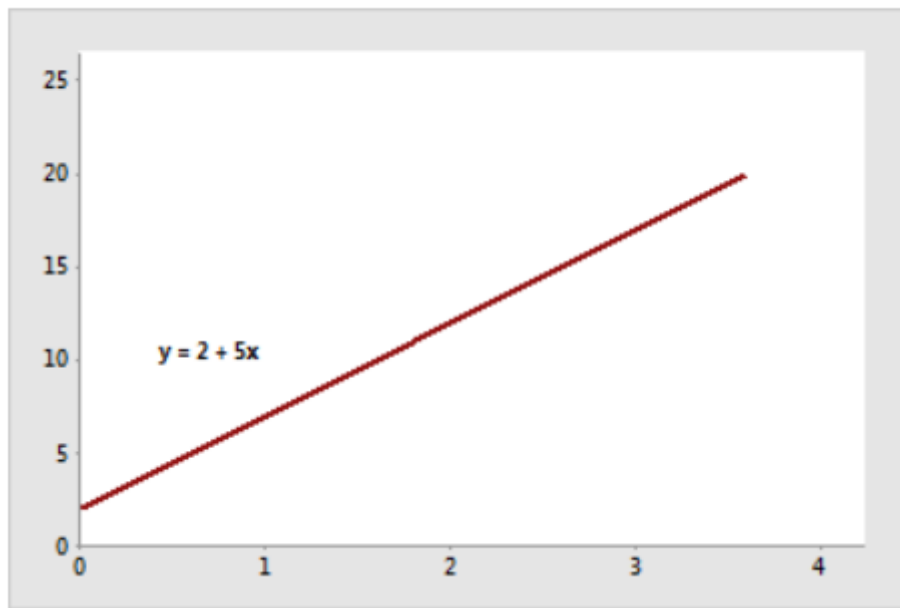
$$c = c - L \times D_c$$

We repeat this process until our loss function is a very small value or ideally 0 (which means 0 error or 100% accuracy). The value of  $\mathbf{m}$  and  $\mathbf{c}$  that we are left with now will be the optimum values.

## 6. Explain What Intercept Means?

The slope indicates the steepness of a line and the intercept indicates the location where it intersects an axis. The slope and the intercept define the linear relationship between two variables, and can be used to estimate an average rate of change. The greater the magnitude of the slope, the steeper the line and the greater the rate of change.

By examining the equation of a line, you quickly can discern its slope and y-intercept (where the line crosses the y-axis).



This follows the equation

$$Y = mx + c$$

Where m is slope c is constant or y intercept it is the point where line cuts the y axis

In our diagram  $m=5$  means when x increase by 1 unit y will be increase by 5.

And in this equation the line has cut the y axis at 2

## 7. What are all the assumption of Linear Regression?

### 1. The Two Variables Should be in a Linear Relationship

The first assumption of simple linear regression is that the two variables in question should have a linear relationship.

The example of Sarah plotting the number of hours a student put in and the amount of marks the student got is a classic example of a linear relationship.

Yes, one can say that putting in more hours of study does not necessarily guarantee higher marks, but the relationship is still a linear one. There will always be many points above or below the line of regression. These points that lie outside the line of regression are the outliers.

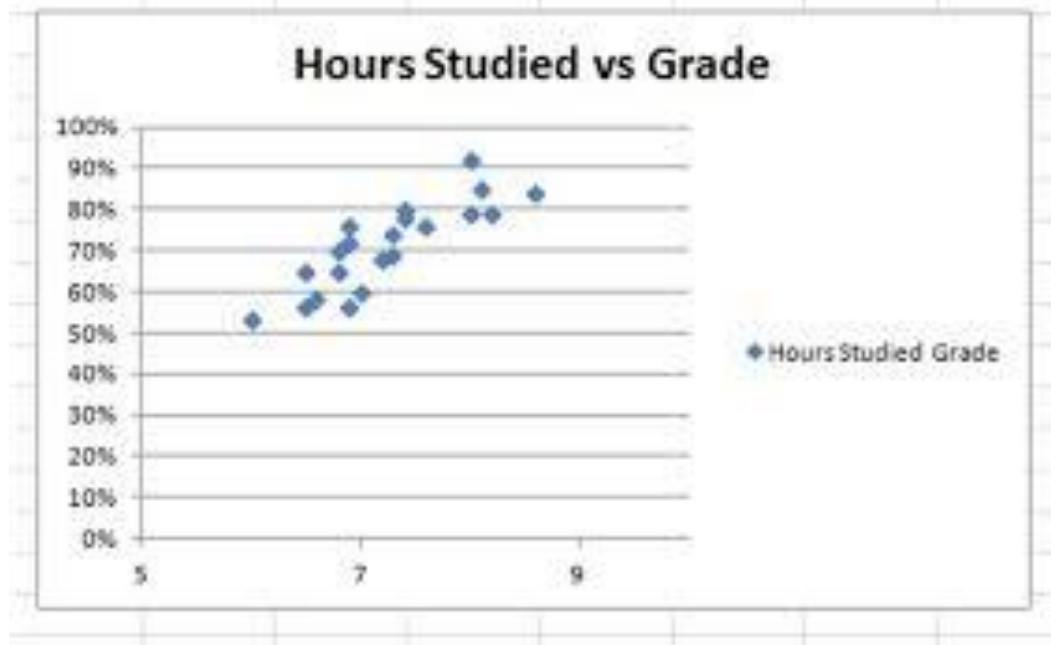
The assumption of linear regression extends to the fact that the regression is sensitive to outlier effects. This assumption is also one of the key assumptions of multiple linear regression.

### 2. All the Variables Should be Multivariate Normal

The first assumption of linear regression talks about being in a linear relationship. The second assumption of linear regression is that all the variables in the data set should be multivariate normal. In other words, it suggests that the linear combination of the random variables should have a normal distribution. The same example discussed above holds good here, as well.



The students reported their activities like studying, sleeping, and engaging in social media. Now, all these activities have a relationship with each other. If you study for a more extended period, you sleep for less time. Similarly, extended hours of study affects the time you engage in social media. Thus, this assumption of simple linear regression holds good in the example. It is possible to check the assumption using a histogram or a Q-Q plot.



All the variables should be multivariate normal

### 3. There Should be No Multicollinearity in the Data

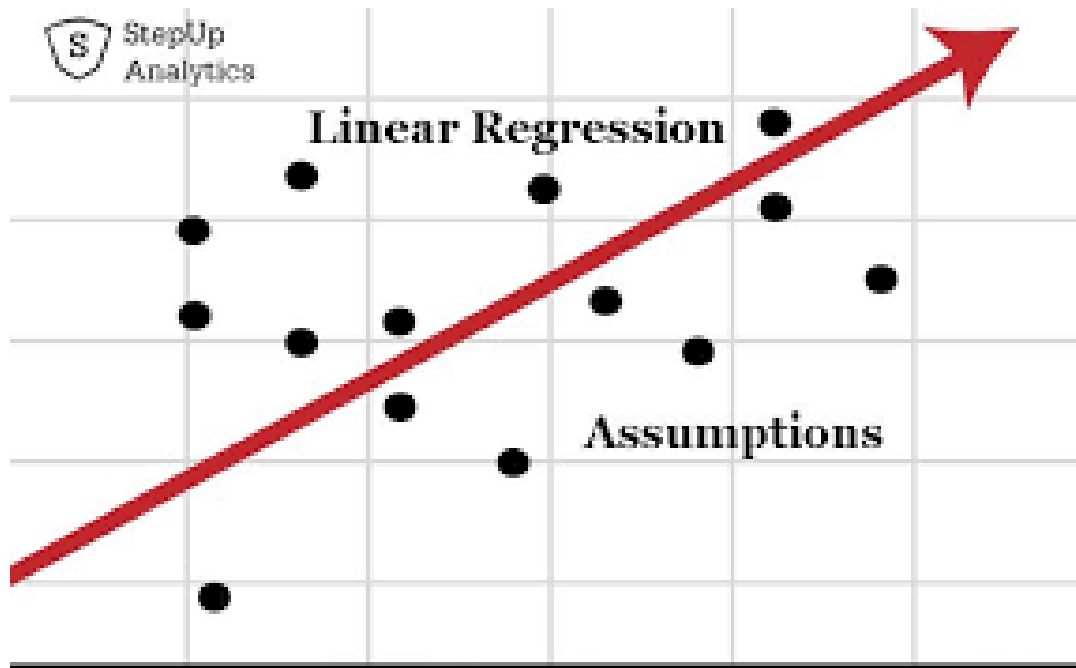
Another critical assumption of multiple linear regression is that there should not be much multicollinearity in the data. Such a situation can arise when the independent variables are too highly correlated with each other.

In our example, the variable data has a relationship, but they do not have much collinearity. There could be students who would have secured higher marks in spite of engaging in social media for a longer duration than the others.

Similarly, there could be students with lesser scores in spite of sleeping for lesser time. The point is that there is a relationship but not a multicollinear one.

If you still find some amount of multicollinearity in the data, the best solution is to remove the variables that have a high variance inflation factor.

This assumption of linear regression is a critical one.



There should be no multicollinearity in the data

#### **4. There Should be No Autocorrelation in the Data**

One of the critical assumptions of multiple linear regression is that there should be no autocorrelation in the data. When the residuals are dependent on each other, there is autocorrelation. This factor is visible in the case of stock prices when the price of a stock is not independent of its previous one.

Plotting the variables on a graph like a scatterplot allows you to check for autocorrelations if any.

Another way to verify the existence of autocorrelation is the Durbin-Watson test.



5 step workflow for multiple linear regression

### 5. There Should be Homoscedasticity Among the Data

Finally, the fifth assumption of a classical linear regression model is that there should be homoscedasticity among the data. The scatterplot graph is again the ideal way to determine the homoscedasticity. The data is said to homoscedastic when the residuals are equal across the line of regression. In other words, the variance is equal.

The Breusch-Pagan Test is the ideal one to determine homoscedasticity. The Goldfield-Quandt Test is useful for deciding heteroscedasticity.

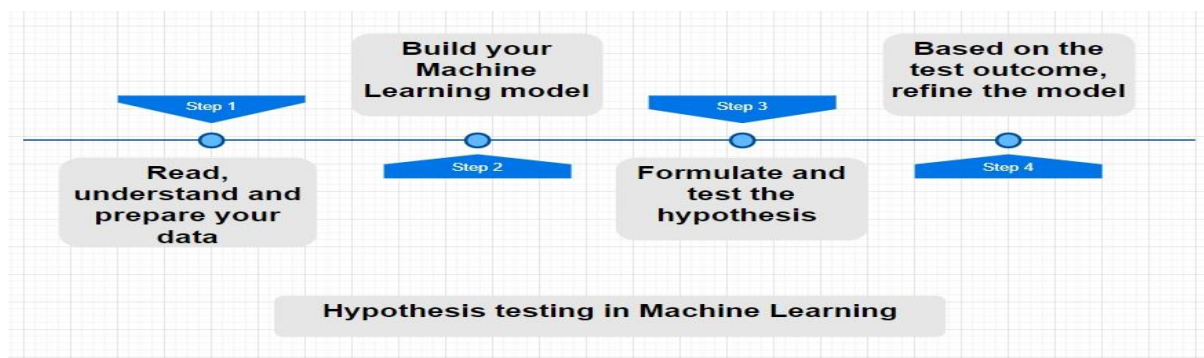


## 8. How is Hypothesis Testing Used in Linear Regression

Hypothesis testing is used to confirm if our beta coefficients are significant in a linear regression model. Every time we run the linear regression model, we test if the line is significant or not by checking if the coefficient is significant. .

### How to perform hypothesis testing in machine learning?

To trust your model and make predictions, we utilize hypothesis testing. When we will use sample data to train our model, we make assumptions about our population. By performing hypothesis testing, we validate these assumptions for a desired significance level.



Let's take the case of regression models: When we fit a straight line through a linear regression model, we get the slope and intercept for the line. Hypothesis testing is used to confirm if our beta coefficients are significant in a linear regression model. Every time we run the linear regression model, we test if the line is significant or not by checking if the coefficient is significant

Key steps to perform hypothesis test are as follows:

1. Formulate a Hypothesis
2. Determine the significance level
3. Determine the type of test
4. Calculate the Test Statistic values and the p values
5. Make Decision

### **Formulating the hypothesis**

One of the key steps to do this is to formulate the below two hypotheses:

**The null hypothesis** represented as  $H_0$  is the initial claim that is based on the prevailing belief about the population.

**The alternate hypothesis** represented as  $H_1$  is the challenge to the null hypothesis. It is the claim which we would like to prove as True

One of the main points which we should consider while formulating the null and alternative hypothesis is that the null hypothesis always looks at confirming the existing notion. Hence, it has sign  $\geq$  or  $\leq$  and  $\neq$

### **Determine the significance level also known as alpha or $\alpha$ for Hypothesis Testing**

The significance level is the proportion of the sample mean lying in critical regions. It is usually set as 5% or 0.05 which means that there is a 5% chance that we would accept the alternate hypothesis even when our null hypothesis is true

Based on the criticality of the requirement, we can choose a lower significance level of 1% as well.

**Determine the Test Statistic and calculate its value for Hypothesis Testing** Hypothesis testing uses Test Statistic which is a numerical summary of a data-set that reduces the data to one value that can be used to perform the hypothesis test.

**Select the type of Hypothesis test**

We choose the type of test statistic based on the predictor variable – quantitative or categorical.

Below are a few of the commonly used test statistics for quantitative data

Type of predictor variable	Distribution type	Desired Test	Attributes
Quantitative	Normal Distribution	Z – Test	<ul style="list-style-type: none"> <li>• Large sample size</li> <li>• Population standard deviation known</li> </ul>
Quantitative	T Distribution	T-Test	<ul style="list-style-type: none"> <li>• Sample size less than 30</li> <li>• Population standard deviation unknown</li> </ul>
Quantitative	Positively skewed distribution	F – Test	<ul style="list-style-type: none"> <li>• When you want to compare 3 or more variables</li> </ul>
Quantitative	Negatively skewed distribution	NA	<ul style="list-style-type: none"> <li>• Requires feature transformation to perform a hypothesis test</li> </ul>
Categorical	NA	Chi-Square test	<ul style="list-style-type: none"> <li>• Test of independence</li> <li>• Goodness of fit</li> </ul>
			<ul style="list-style-type: none"> <li>•</li> </ul>

*Z-statistic – Z Test*

Z-statistic is used when the sample follows a normal distribution. It is calculated based on the population parameters like mean and standard deviation.

One sample Z test is used when we want to compare a sample mean with a population mean

Two sample Z test is used when we want to compare the mean of two samples

*T-statistic – T-Test*

T-statistic is used when the sample follows a T distribution and population parameters are unknown.

T distribution is similar to a normal distribution, it is shorter than normal distribution and has a flatter tail.

If the sample size is less than 30 and population parameters are not known, we use T distribution. Here also, we can use one Sample T-test and a two-sample T-test.

### *F-statistic – F test*

For samples involving three or more groups, we prefer the F Test. Performing T-test on multiple groups increases the chances of Type-1 error. ANOVA is used in such cases.

*Analysis of variance (ANOVA)* can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

F-statistic is used when the data is positively skewed and follows an F distribution. F distributions are always positive and skewed right.

$F = \text{Variation between the sample means} / \text{variation within the samples}$

For negatively skewed data we would need to perform feature transformation

### *Chi-Square Test*

For categorical variables, we would be performing a chi-Square test.

Following are the two types of chi-squared tests:

1. Chi-squared test of independence – We use the Chi-Square test to determine whether or not there is a significant relationship between two categorical variables.
2. Chi-squared Goodness of fit helps us determine if the sample data correctly represents the population.
- 3.

### **The decision about your model**

Test Statistic is then used to calculate P-Value. A P-value measures the strength of evidence in support of a null hypothesis. If the P-value is less than the significance level, we reject the null hypothesis.

if the **p-value**  $< \alpha$ , then we have statistically significant evidence against the null hypothesis, so we reject the null hypothesis and accept the alternate hypothesis

if the p-value  $> \alpha$  then we do not have statistically significant evidence against the null hypothesis, so we fail to reject the null hypothesis.

As we make decisions, it is important to understand the errors that can happen while testing.

## 9. How would you decide the importance of variable for the multivariate regression?

When building a linear or logistic regression model, you should consider including:

1. Variables that are already proven in the literature to be related to the outcome
2. Variables that can either be considered the cause of the exposure, the outcome, or both
3. Interaction terms of variables that have large main effects

However, you should watch out for:

1. Variables that have a large number of missing values or low variability
2. Variables that are highly correlated with other predictors in the model (causing a collinearity problem)
3. Variables that are not linearly related to the outcome (in case you're running a linear regression)

## 10. What is the difference between $R^2$ and Adjusted $R^2$

### R-squared ( $R^2$ )

It measures the proportion of the variation in your dependent variable explained by all of your independent variables in the model. It assumes that every independent variable in the model helps to explain variation in the dependent variable. In reality, some independent variables (predictors) don't help to explain dependent (target) variable. In other words, some variables do not contribute in predicting target variable.

Mathematically, R-squared is calculated by dividing sum of squares of residuals (**SSres**) by total sum of squares (**SStot**) and then subtract it from 1. In this case, SStot measures total variation. **SSreg** measures explained variation and SSres measures unexplained variation.

As  $SSres + SSreg = SStot$ ,  $R^2 = \text{Explained variation} / \text{Total Variation}$



$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}} \equiv 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$$\downarrow$$

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$$

R-Squared is also called **coefficient of determination**. It lies between **0%** and **100%**. A r-squared value of 100% means the model explains all the variation of the target variable. And a value of 0% measures zero predictive power of the model. **Higher R-squared value, better the model.**

### Adjusted R-Squared

It measures the proportion of variation explained by only those independent variables that really help in explaining the dependent variable. It penalizes you for adding independent variable that do not help in predicting the dependent variable.

Adjusted R-Squared can be calculated mathematically in terms of sum of squares. The only difference between R-square and Adjusted R-square equation is degree of freedom. Adjusted R-squared value can be calculated based on value of r-squared, number of independent variables (predictors), total sample size.

$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

$R^2$  = sample R-square

p = Number of predictors

N = Total sample size.