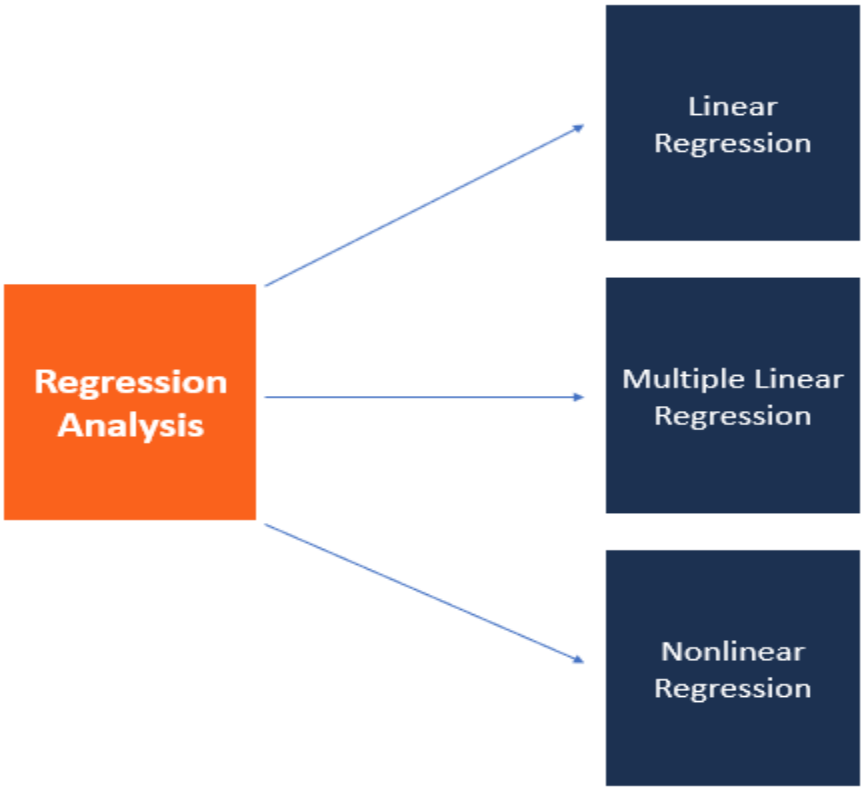
Regression Analysis

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable(Y) and one or more independent variables (X). It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

Regression analysis will provide you with an equation for a graph so that you can make predictions about your data.

For example, global warming may be reducing average snowfall in your town and you are asked to predict how much snow you think will fall this year. Looking at the following table you might guess somewhere around 10-20 inches. That's a good guess, but you could make a better guess, by using regression.

Essentially, regression is the "best guess" at using a set of data to make some kind of prediction. It's fitting a set of points to a graph.



Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear.

Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Regression analysis offers numerous applications in various disciplines.

1.1 **The basic types of regression**

The most common form of regression analysis are: simple linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis.

Linear regression analysis is based on six fundamental assumptions:

The dependent and independent variables show a linear relationship between the slope and the intercept.

The independent variable is not random.

The value of the residual (error) is zero.

The value of the residual (error) is constant across all observations.

The value of the residual (error) is not correlated across all observations.

The residual (error) values follow the normal distribution.

1.1.1 Simple linear regression

Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y,

Simple linear regression is a model that explain or predict the relationship between a dependent variable(Y) and an independent variable(X).

For example, Massie and Rose (1997) applied a simple linear regression method to predict daily maximum temperatures investigated at Nashville, Tennessee.

The simple linear model is expressed using the following equation:

$Y = a + b.X$

Where:

Y = the variable that you are trying to predict (dependent variable).

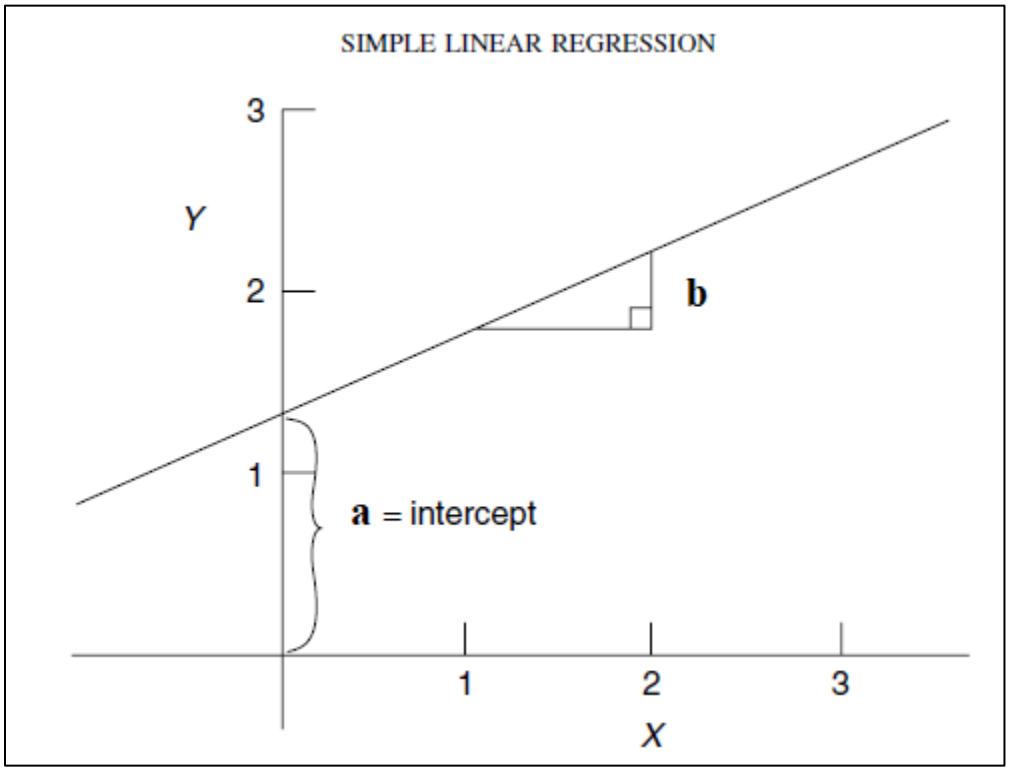
X = the variable that you are using to predict Y (independent variable).

a = the intercept.

b = the slope.

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y - b\sum x}{n}$$



1.1.2 Multiple linear regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables (X1, X2, X3) are used in the

model. The mathematical representation of multiple linear regression is:

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points.

$$Y = a + bX1 + cX2 + dX3$$

Where:

Y = the variable that you are trying to predict (dependent variable).

X1, X2, X3: Independent variables (the variables that you are using to predict Y)

a: Intercept

b, c, d: Slopes

For example, Paras and Mathur (2012) applied the Multiple Linear Regression (MLR) to develop a model for forecasting weather parameters. It was found that the proposed model is capable of forecasting the weather conditions for a particular station using the data collected locally.

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model:

Non-collinearity: Independent variables should show a minimum of correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Nonlinear Regression

Nonlinear regression is a form of regression analysis in which data is fit to a model and then expressed as a mathematical function.

Simple linear regression relates two variables (X and Y) with a straight line

(y = mx + b), while nonlinear regression must generate a line (typically a curve) as if every value of Y was a random variable.

Example of Nonlinear Regression

One example of how nonlinear regression can be used is to predict population growth over time.

A scatter plot of changing population data over time shows that there seems to be a relationship between time and population growth, but that it is a nonlinear relationship, requiring the use of a nonlinear regression model. A logistic population growth model can provide estimates of the population for periods that were not measured, and predictions of future population growth.

Independent and dependent variables used in nonlinear regression should be quantitative.

The Coefficient of Determination (R^2):

The coefficient of determination gives an idea of how many data points fall within the results of the line formed by the regression equation.

The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted.

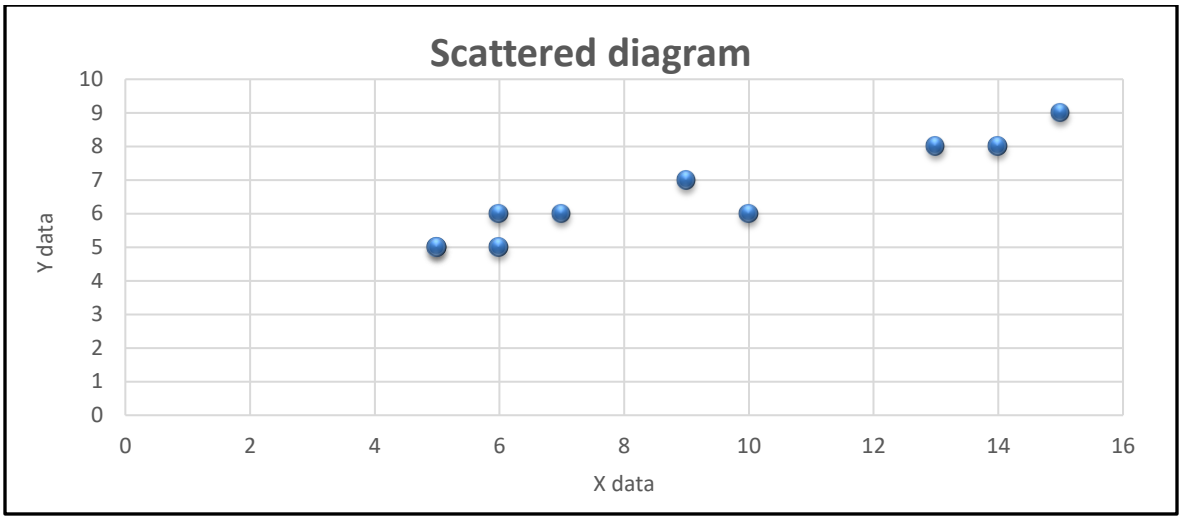
If the coefficient is 0.80, then 80% of the points should fall within the regression line.

Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coefficient is an indicator of a better goodness of fit for the observations.

Example: Find the regression equation for the following data

y	6	8	9	8	7	6	5	6	5	5
x	10	13	15	14	9	7	6	6	5	5

Solution



N	X	Y	X.Y	X^2
1	10	6	60	100
2	13	8	104	169
3	15	9	135	225
4	14	8	112	196
5	9	7	63	81
6	7	6	42	49
7	6	5	30	36
8	6	6	36	36
9	5	5	25	25
10	5	5	25	25
Total	**90**	**65**	**632**	**942**

$$b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$b = \frac{(10 * 632) - (90) * (65)}{(10 * 942) - (90)^2} = 0.36$$

$$a = \frac{\sum y - b\sum x}{n}$$

$$a = \frac{65 - (0.36 * 90)}{10} = 3.26$$

$$y = a + b.x$$

$$y = 3.26 + 0.36.x$$