Linear Regression

Introduction to Linear Regression

Last updated date: 01st Feb 2023 • Total views: 202.2k • Views today: 1.94k

Linear regression is used to predict the relationship between two variables by applying a linear equation to observed data. There are two types of variable, one variable is called an independent variable, and the other is a dependent variable. Linear regression is commonly used for predictive analysis. The main idea of regression is to examine two things. First, does a set of predictor variables do a good job in predicting an outcome (dependent) variable? The second thing is which variables are significant predictors of the outcome variable? In this article, we will discuss the concept of the Linear Regression Equation, formula and Properties of Linear Regression.

Examples of Linear Regression

The weight of the person is linearly related to their height. So, this shows a linear relationship between the height and weight of the person. According to this, as we increase the height, the weight of the person will also increase. It is not necessary that one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to simplify the strength of the relationship between the variables. If there is no relation or linking between the variables then the scatter plot does not indicate any increasing or decreasing pattern. In such cases, the linear regression design is not beneficial to the given data.

Linear Regression Equation

The measure of the relationship between two variables is shown by the correlation coefficient. The range of the coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data between two variables.

Linear Regression Equation is given below:

$Y=a+bX$

where X is the independent variable and it is plotted along the x-axis

Y is the dependent variable and it is plotted along the y-axis

Here, the slope of the line is b, and a is the intercept (the value of y when x = 0).

Linear Regression Formula

As we know, linear regression shows the linear relationship between two variables. The equation of linear regression is similar to that of the slope formula. We have learned this formula before in earlier classes such as a linear equation in two variables. Linear Regression Formula is given by the equation

Y= a + bX

We will find the value of a and b by using the below formula

a= $$\frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum x^2) - (\sum x)^2}$$

b= $$\frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum x^2) - (\sum x)^2}$$

Simple Linear Regression

Simple linear regression is the most straight forward case having a single scalar predictor variable x and a single scalar response variable y. The equation for this regression is given as y=a+bx

The expansion to multiple and vector-valued predictor variables is known as multiple linear regression. It is also known as multivariable linear regression. The equation for this regression is given as Y = a+bX. Almost all real-world regression patterns include multiple predictors. The basic explanations of linear regression are often explained in terms of multiple regression. Note that, in these cases, the dependent variable y is yet a scalar.

Least Square Regression Line or Linear Regression Line

The most popular method to fit a regression line in the XY plot is found by using least-squares. This process is used to determine the best-fitting line for the given data by reducing the sum of the squares of the vertical deviations from each data point to the line. If a point rests on the fitted line accurately, then the value of its perpendicular deviation is 0. It is 0 because the variations are first squared, then added, so their positive and negative values will not be cancelled. Linear regression determines the straight line, known as the least-squares regression line or LSRL. Suppose Y is a dependent variable and X is an independent variable, then the population regression line is given by the equation;

$Y= B_0+B_1X$

Where

B_0 is a constant

B_1 is the regression coefficient

When a random sample of observations is given, then the regression line is expressed as;

$\hat{y} = b_0+b_1x$

where b_0 is a constant

b_1 is the regression coefficient,

x is the independent variable,

$\hat{y}$ is known as the predicted value of the dependent variable.

Properties of Linear Regression

For the regression line where the regression parameters b_0 and $b_1$are defined, the following properties are applicable:

- The regression line reduces the sum of squared differences between observed values and predicted values
- The regression line passes through the mean of X and Y variable values.
- The regression constant b_0 is equal to the y-intercept of the linear regression.
- The regression coefficient b_1 is the slope of the regression line. Its value is equal to the average change in the

dependent variable (Y) for a unit change in the independent variable (X)

Regression Coefficient

The regression coefficient is given by the equation :

$Y= B_0+B_1X$

Where

B_0 is a constant

B_1 is the regression coefficient

Given below is the formula to find the value of the regression coefficient.

$B_1=b_1 = \sum[(x_i-x)(y_i-y)]/\sum[(x_i-x)^2]$

Where x_i and y_i are the observed data sets.

And x and y are the mean value.

Importance of Regression Line

A regression line is used to describe the behaviour of a set of data, a logical approach that helps us study and analyze the relationship between two different continuous variables. Which is then enacted in machine learning models, mathematical analysis, statistics field, forecasting sectors, and other such quantitative applications. Looking at the financial sector, where financial analysts use linear regression to predict stock prices and commodity prices and perform various stock valuations for different securities. Several well-renowned companies make use of linear regressions for the purpose of predicting sales, inventories, etc.

Key Ideas of Linear Regression

- Correlation explains the interrelation between variables within the data.
- Variance is the degree of the spread of the data.
- Standard deviation is the dispersion of mean from a data set by studying the variance's square root.
- Residual (error term) is the actual value found within the dataset minus the expected value that is predicted in linear regression.

Important Properties of Regression Line

- Regression coefficient values remain the same because the shifting of origin takes place because of the change of scale. The property says that if the variables x and y are changed to u and v respectively u= (x-a)/p v=(y-c) /q, Here p and q are the constants.Byz =q/p*bvu Bxy=p/q*buv.
- If there are two lines of regression and both the lines intersect at a selected point (x', y'). The variables x and y are considered. According to the property, the intersection of the two regression lines is (x`, y`), which is the solution of the equations for both the variables x and y.
- You will understand that the correlation coefficient between the two variables x and y is the geometric mean of both the coefficients. Also, the sign over the values of correlation coefficients will be the common sign of both the coefficients. So, if according to the property regression coefficients are byx= (b) and bxy= (b') then the correlation coefficient is r=+-sqrt (byx + bxy) which is why in some cases, both the values of coefficient are negative value and r is also negative. If both the values of coefficients are positive then r is going to be positive.
- The regression constant (a_0) is equal to the y-intercept of the regression line and also a_0 and a_1 are the regression parameters.

Regression Line Formula:

A linear regression line equation is written as-

$Y = a + bX$

where X is plotted on the x-axis and Y is plotted on the y-axis. X is an independent variable and Y is the dependent variable. Here, b is the slope of the line and a is the intercept, i.e. value of y when x=0.

Multiple Regression Line Formula: $y= a +b_1x_1 +b_2x_2 + b_3x_3 +…+ b_tx_t + u$

Assumptions made in Linear Regression

- The dependent/target variable is continuous.
- There isn't any relationship between the independent variables.
- There should be a linear relationship between the dependent and explanatory variables.
- Residuals should follow a normal distribution.
- Residuals should have constant variance.
- Residuals should be independently distributed/no autocorrelation.

Solved Examples

1. Find a linear regression equation for the following two sets of data:

x	2	4	6	8
y	3	7	5	10

Sol: To find the linear regression equation we need to find the value of Σx, Σy, Σx

2

2

and Σxy

Construct the table and find the value

x	**y**	**x²**	**xy**
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
Σx = 20	Σy = 25	Σx² = 120	Σxy = 144

The formula of the linear equation is y=a+bx. Using the formula we will find the value of a and b

a= $\frac{(\sum Y)(\sum X^2)-(\sum X)(\sum XY)}{n(\sum x^2)-(\sum x)^2}$

Now put the values in the equation

$$a=\frac{25\times 120-20\times 144}{4\times 120-400}$$

a= $\frac{120}{80}$

a=1.5

b= $\frac{n(\sum XY)-(\sum X)(\sum Y)}{n(\sum x^2)-(\sum x)^2}$

Put the values in the equation

$$b=\frac{4\times 144-20\times 25}{4\times 120-400}$$

b=$\frac{76}{80}$

b=0.95

Now put the value of a and b in the equation

Hence equation of linear regression is y = 1.5 + 0.95x