# Linear Regression

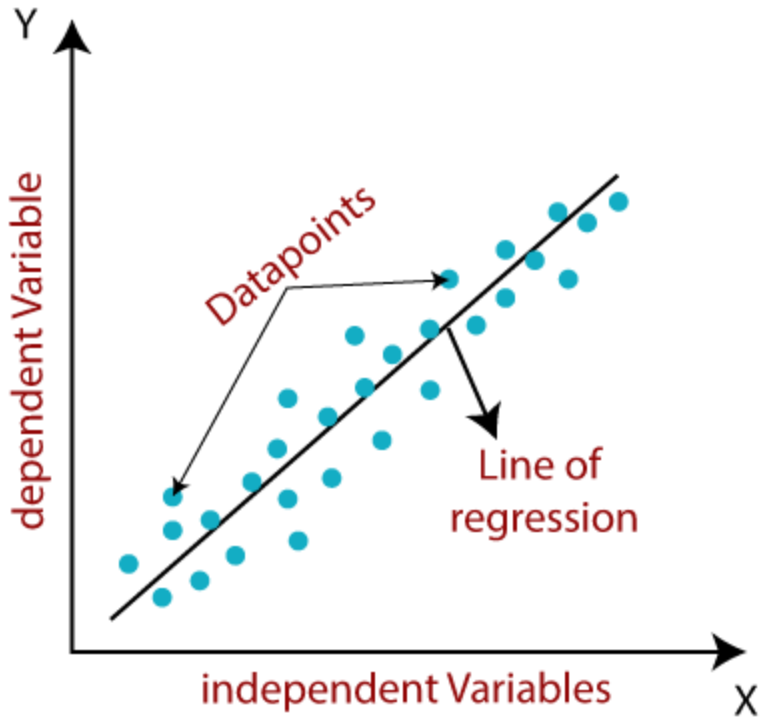
Linear learning is supervised learning

● **Linearity**: Linearity is the property of a mathematical relationship that can be graphically represented as a straight line. Linearity is closely related to proportionality.

● **Regression**: Regression is a data mining technique that is generally used for the purpose of predicting a range of continuous values in a specific data set.
For example, you might guess that there's a connection between how much you eat and how much your weight; regression analysis can help you quantify that.

**Independent Variable:**
It is a variable that stands alone and isn't changed by the other variables you are trying to measure. For example, someone's age might be an independent variable.

**Dependent Variable**:
The dependent variable is the variable that changes in response to the independent variable. For **example**, a test score could be a **dependent variable** because it could change depending on several factors such as how much you studied, how much sleep you got the night before you took the test, or even how hungry you were when you took it.
**Another example:** In a study to determine whether how long a student sleeps affects test scores, **the independent variable** is **the** length of time spent sleeping while **the dependent variable** is **the** test score.




**Linear regression:**

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc. **Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.**

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variables.

**When to use Linear Regression?:**

It is used when we want to predict the value of a variable based on the value of another variable. The variable **we** want to predict is called the dependent variable (or sometimes, the outcome variable).

It is used when we see there is a proportion relation between dependent variable and the independent variable.

For example, **you could use linear regression to understand whether exam performance can be predicted based on revision time**;

If you have two or more independent variables, rather than just one, you need to use multiple regression.

**Types of linear regression:**
1. Simple Linear Regression
2. Multiple Linear Regression

**Simple Linear Regression:**
In Simple Linear Regression, we try to find the relationship between a single independent variable (input) and a corresponding dependent variable (output). This can be expressed in the form of a straight line. Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variables. **Simple linear regression** is used to estimate the relationship between **two quantitative variables**. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

**Example:** You are a social researcher interested in the relationship between income and happiness. You survey 500 people whose incomes range from $15k to $75k and ask them to rank their happiness on a scale from 1 to 10.

Your independent variable (income) and dependent variable (happiness) are both quantitative, so you can do a regression analysis to see if there is a linear relationship between them.

If you have more than one independent variable, use multiple linear regression instead.

**Represent of Simple Linear regression**:
we can write a linear regression equation as: y=a+bx ( where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

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

Here, x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Independent variable from dataset

y = Dependent variable from dataset

n = Number of observations in the dataset

Calculate Simple Linear Regression:

| SUBJECT | AGE (X) | GLUCOSE LEVEL (Y) | XY | $X^2$ | $Y^2$ |
|---|---|---|---|---|---|
| 1 | 43 | 99 | 4257 | 1849 | 9801 |
| 2 | 21 | 65 | 1365 | 441 | 4225 |
| 3 | 25 | 79 | 1975 | 625 | 6241 |
| 4 | 42 | 75 | 3150 | 1764 | 5625 |
| 5 | 57 | 87 | 4959 | 3249 | 7569 |
| 6 | 59 | 81 | 4779 | 3481 | 6561 |
| | Σx = 247 | Σy = 486 | Σxy = 20485 | Σx2 = 11409 | Σy2 = 40022 |

From the above table, Σx = 247, Σy = 486, Σxy = 20485, Σx2 = 11409, Σy2 = 40022. n is the sample size (6, in our case).

**Find a**:

$a=((486 \times 11{,}409) – ((247 \times 20{,}485)) / 6\ (11{,}409) – 247^2)$

$484979 / 7445$

**=65.14**

**Find b**:

$b=(6(20{,}485) – (247 \times 486)) / (6\ (11409) – 247^2)$

$(122{,}910 – 120{,}042) / 68{,}454 – 247^2$

$2{,}868 / 7{,}445$

**= .385225**

y' = a + bx
**y' = 65.14 + .385225x**

Here our Model is ready. After completing the training phase, we got the value of "a = 65.14" and "b = .385225"

Now it's time to test out model:

Let, Age(X) = 60

So, glucose level (y) = 65.14+.3852245*(60)

=88.253

When we test our model 10 times, it gives the right value for 6

glucose levels out of 10.

So the accuracy of our model = 6/10 = .6*100 = 60%

## Multiple Linear Regression:

In Multiple Linear Regression, **we try to find the relationship between 2 or more independent variables (inputs) and the corresponding dependent variable (output).** The independent variables can be continuous or categorical.

Multiple linear regression analysis can help us in the following ways:

- **It helps us predict trends and future values.** The multiple linear regression analysis can be used to get point estimates.
- **It can be used to forecast the effects or impacts of changes.** That is, multiple linear regression analysis can help to understand how

much will the dependent variable change when we change the independent variables.

- It can be used to identify the strength of the effect that the independent variables have on a dependent variable.

<u>The formula for multiple linear regression with k independent variables:</u>

$Y=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k+\varepsilon$

Where,

slopes= $\beta$

Y=dependent variable (unknown)

X=independent variables (It is given)

$\beta_0$=y-intercept (constant term)

$\varepsilon$=the model's error term

**Multiple Linear regression calculation:**

We have to predict the value of Y from $X_1$ and $X_2$

| Subject | Y | $X_1$ | $X_2$ |
|---|---|---|---|
| 1 | -3.7 | 3 | 8 |
| 2 | 3.5 | 4 | 5 |
| 3 | 2.5 | 5 | 7 |
| 4 | 11.5 | 6 | 3 |
| 5 | 5.7 | 2 | 1 |
| 6 | ? | 3 | 2 |

Here, 6th example is the testing example. The values of $X_1$ and $X_2$ are given to us and we need to predict the value of Y.

$$b_1 = \frac{\left(\sum x_2^2\right)\left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$b_2 = \frac{\left(\sum x_1^2\right)\left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_1 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)}$$

$$a = b_0 = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2$$

$$\sum x_1^2 = \sum X_1 X_1 - \frac{\left(\sum X_1\right)\left(\sum X_1\right)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{\left(\sum X_2\right)\left(\sum X_2\right)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{\left(\sum X_1\right)\left(\sum Y\right)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{\left(\sum X_2\right)\left(\sum Y\right)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{\left(\sum X_1\right)\left(\sum X_2\right)}{N}$$

| **Subject** | Y | $X_1$ | $X_2$ | $X_1 X_1$ | $X_2X_2$ | $X_1 X_2$ | $X_1 Y$ | $X_2Y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | -3.7 | 3 | 8 | 9 | 64 | 24 | -11.1 | -29.6 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 2 | 3.5 | 4 | 5 | 16 | 25 | 20 | 14 | 17.5 |
| 3 | 2.5 | 5 | 7 | 25 | 49 | 35 | 12.5 | 17.5 |
| 4 | 11.5 | 6 | 3 | 36 | 9 | 18 | 69 | 34.5 |
| 5 | 5.7 | 2 | 1 | 4 | 1 | 2 | 11.4 | 5.7 |
| | 19.5 | 20 | 24 | 90 | 148 | 99 | 95.8 | 45.6 |

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 10$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 32.8$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N} = 17.8$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N} = -48$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 3$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 * 17.8 - 3 * (-48)}{10 * 32.8 - 3 * 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 * (-48) - 3 * 17.8}{10 * 32.8 - 3 * 3} = -1.67$$

$$a = \overline{Y} - b_1 \overline{X}_1 - b_2 \overline{X}_2 = \frac{19.5}{5} - \frac{2.28 * 20}{5} - \frac{-1.67 * 24}{5} = 2.796$$

Final Regression equation or Model is:

$$Y = 2.796 + 2.28\,x_1 - 1.67x_2$$

$$Now\ given \quad x_1 = 3 \quad and \quad x_2 = 2 \qquad Y = ?$$

$$Y = 2.796 + 2.28 * 3 - 1.67 * 2$$
$$= 6.296$$

**Self practice:**

| Pizza Size(In inche) | Price(tk) |
|---|---|
| 6 | 350 |
| 8 | 775 |
| 12 | 1150 |
| 14 | 1395 |
| 18 | 1675 |

We have to find xy, x2 for finding slope and intercept
Here, Pizza Size is independent variable (x) and Price is dependent variable (y).

Now,

$$\bar{x} = \frac{6+8+12+14+18}{5} = 11.6$$

$$\bar{y} = \frac{350+775+1150+1395+1675}{5} = 1069$$

$$\overline{xy} = \frac{2100+6200+13800+19530+30150}{5} = 14356$$

$$(\bar{x})^2 = (11.6)^2 = 134.56$$

$$\overline{x^2} = \frac{36+64+144+196+324}{5} = 152.8$$

$$a1 = \frac{\bar{x}.\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} = \frac{(11.6 \times 1069) - 14356}{134.56 - 152.8} = \frac{-1955.6}{-18.24} = 107.21$$

$$a0 = \bar{y} - a1\bar{x} = 1069 - (107.21 \times 11.6) = -174.46$$

So The final equation will be

y = 107.21x - 174.6

This is the equation of regression line or best fit line

$y = a_0 + a_1 x$

Now it's predicting time.
What will be the price for a 17 inch's Pizza?
If we put the value of x in the equation y = 107.21x - 174.6 we will get our result.
y = 107.21 * 17 - 174.6
= 1647.96 tk