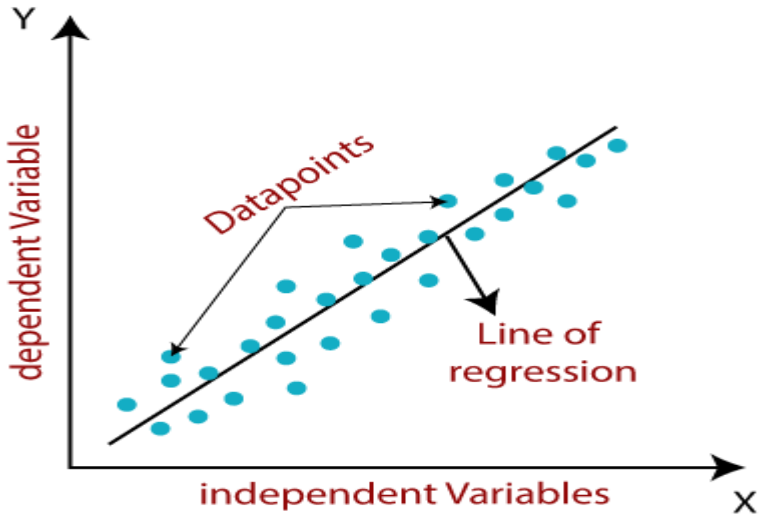
Linear Regression

- Linear regression is one of the easiest and most popular Supervised 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 as 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.
- The linear regression model provides a sloped straight line representing the relationship between the variables.
- Consider the below image:



- Mathematically, we can represent a linear regression as: $\mathbf{y= b_0+b_1x+ \varepsilon}$
- Here,
  y= Dependent Variable (Target Variable)
  x= Independent Variable (predictor Variable)
  b_0= intercept of the line (Gives an additional degree of freedom)
  b_1 = Linear regression coefficient (scale factor to each input value).
  ε = random error
- The values for x and y variables are training datasets for Linear Regression model representation.

Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

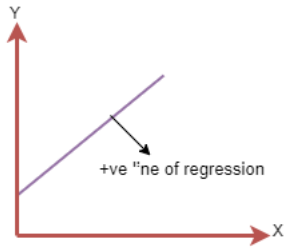
1. **Simple Linear Regression**: If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

2. **Multiple Linear regression**: If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression Line

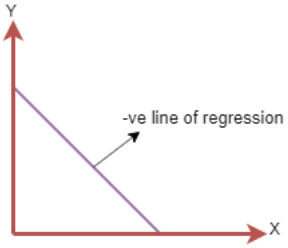
A linear line showing the relationship between the dependent and independent variables is called a **regression line**. A regression line can show two types of relationship:

- **Positive Linear Relationship:** If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line of equation will be: $y = b_0 + b_1x$

- **Negative Linear Relationship:** If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $y = -b_0 + b_1x$

Finding the best fit line:

- When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.
- The different values for weights or the coefficient of lines (b_0, b_1) gives a different line of regression, so we need to calculate the best values for b_0 and b_1 to find the best fit line, so to calculate this we use cost function.

Cost function

- The different values for weights or coefficient of lines (b_0, b_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.

- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.
- For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (b_1 x_i + b_0))^2$$

where,
N=Total number of observation
y_i = Actual value
$(b_1x_i+b_0)$= Predicted value.

Residuals

- The distance between the actual value and predicted values is called residual.
- If the observed points are far from the regression line, then the residual will be high, and so cost function will high.
- If the scatter points are close to the regression line, then the residual will be small and hence the cost function.

Gradient Descent

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Scatter Plot

- A scatter plot is a set of points plotted on a horizontal and vertical axes.
- Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables).
- If no correlation exists between the variables, the points appear randomly scattered on the coordinate plane.
- If a large correlation exists, the points concentrate near a straight line.
- Scatter plots are useful data visualization tools for illustrating a trend.
- Besides showing the extent of correlation, a scatter plot shows the sense of the correlation:
- If the vertical (or y-axis) variable increases as the horizontal (or x-axis) variable increases, the correlation is positive.
- If the y-axis variable decreases as the x-axis variable increases or vice-versa, the correlation is negative.

- If it is impossible to establish either of the above criteria, then the correlation is zero.
- The maximum possible positive correlation is +1 or +100%, when all the points in a scatter plot lie exactly along a straight line with a positive slope.
- The maximum possible negative correlation is -1 or -100%, in which case all the points lie exactly along a straight line with a negative slope.

Standard Error of Estimate

- The standard error of estimate is a measure of the accuracy of predictions.
- It is given by:

$$\sigma_{est} = \sqrt{\frac{\sum(y - y')^2}{N}}$$

where, σ_{est} is the standard error of estimate, y is an actual score, y' is a predicted score and N is the number of pairs of scores.

Methods to solve Linear Regression Problems

1. Least Square Method
2. Karl Correlation Coefficient Method

1. Least Square Method

We have linear regression equation as **$y = b_0 + b_1 x$**

Using least square method,

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

2. Karl Pearson's Correlation Coefficient Method

We have linear regression equation as **$y = b_0 + b_1 x$**

Using Karl Correlation Coefficient method,

$$b_1 = \frac{n \times \sum xy - \sum x \sum y}{n \times \sum x^2 - (\sum x)^2} \quad and \; b_0 = \frac{\sum y - b_1 \times \sum x}{n}$$

Q1. Find predicted value of y for one epoch and RMSE using Linear Regression. [May – 19]

X	2	3	4	5	6	7	8	9	10
y-Actual	1	3	6	9	11	13	15	17	20

Solution:

	x	y-Actual	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
	2	1	-4	-9.56	16	38.22
	3	3	-3	-7.56	9	22.67
	4	6	-2	-4.56	4	9.11
	5	9	-1	-1.56	1	1.56
	6	11	0	0.44	0	0.00
	7	13	1	2.44	1	2.44
	8	15	2	4.44	4	8.89
	9	17	3	6.44	9	19.33
	10	20	4	9.44	16	37.78
Sum =	**54**	**95**			**60**	**140**
Average =	**6**	**10.56**				

We have linear regression equation as $\mathbf{y= b_0+b_1x}$

Using least square method,

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{140}{60} = 2.33$$

$$b_0 = \bar{y} - b_1\bar{x} = 10.56 - 2.33 \times 6 = -3.42$$

∴ The linear regression line becomes: $\boldsymbol{-3.42 + 2.33 \times x}$

To find the predicted value of y for one epoch of x:

Let x = 3

$$\therefore \boldsymbol{y = -3.42 + 2.33 \times 3 = 3.57}$$

We have $MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (b_1x_i + b_0))^2$

For the above epoch when x = 3,

$$MSE = (3 - 3.57)^2 = 0.3249$$

$$\therefore \boldsymbol{RMSE = \sqrt{MSE} = \sqrt{0.3249} = 0.57}$$

Q2. Given the following data for the sales of car of an automobile company for six consecutive years. Predict the sales for next two consecutive years. [Dec – 2019]

Years	2013	2014	2015	2016	2017	2018
Sales	110	100	250	275	230	300

Solution:

We have linear regression equation as $\mathbf{y= b_0+b_1x}$

	Years (x)	Sales (y)	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	2013	110	-2.5	-100.83	6.25	252.08
	2014	100	-1.5	-110.83	2.25	166.25
	2015	250	-0.5	39.17	0.25	-19.58
	2016	275	0.5	64.17	0.25	32.08
	2017	230	1.5	19.17	2.25	28.75
	2018	300	2.5	89.17	6.25	222.92
Sum =	**12093**	**1265**			**17.5**	**682.5**
Average =	**2015.5**	**210.83**				

Using least square method,

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{682.5}{17.5} = 39$$

$$b_0 = \bar{y} - b_1\bar{x} = 210.83 - 39 \times 2015.5 = -78393.7$$

∴ The linear regression line becomes: $\mathbf{-78393.7 + 39 \times x}$

The sales for next two consecutive years:

$$\boldsymbol{When\ x = 2019, y = -78393.7 + 39 \times 2019 = 347.33 \approx 347}$$

$$\boldsymbol{When\ x = 2020, y = -78393.7 + 39 \times 2020 = 386.33 \approx 386}$$

Q3. The following table shows the midterm and final exam grades obtained for students in Database course. Use the method of least squares using regression to predict the final exam grade of a student who received 86 on the midterm exam. [May – 2017]

Midterm Exam (x)	72	50	81	74	94	86	59	83	65	33	88	81
Final Exam (y)	84	63	77	78	90	75	49	79	77	52	74	90

Solution:

	Midterm Exam (x)	Final Exam (y)	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	72	84	-0.17	10	0.03	-1.67
	50	63	-22.17	-11	491.36	243.83
	81	77	8.83	3	78.03	26.50
	74	78	1.83	4	3.36	7.33
	94	90	21.83	16	476.69	349.33
	86	75	13.83	1	191.36	13.83
	59	49	-13.17	-25	173.36	329.17
	83	79	10.83	5	117.36	54.17
	65	77	-7.17	3	51.36	-21.50
	33	52	-39.17	-22	1534.03	861.67
	88	74	15.83	0	250.69	0.00
	81	90	8.83	16	78.03	141.33
Sum	**866**	**888**			**3445.67**	**2004**
Average	**72.17**	**74**				

We have linear regression equation as **$y = b_0 + b_1 x$**

Using least square method,

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{2004}{3445.67} = 0.58$$

$$b_0 = \bar{y} - b_1\bar{x} = 74 - 0.58 \times 72.17 = 32.14$$

∴ The linear regression line becomes: $\boldsymbol{y = 32.14 + 0.58 \times x}$

The final exam grade of a student who received 86 on the midterm exam:

$$\boldsymbol{y = 32.14 + 0.58 \times 86 = 82.02 \approx 82\ marks}$$

Q4. The values of independent variable x and dependent variable y are given below. Find the least square line y = ax + b. Estimate the value of y when x is 10. [May – 2018]

x	0	1	2	3	4
y	2	3	5	4	6

Solution:

	x	**y**	$(\boldsymbol{x - \bar{x}})$	$(\boldsymbol{y - \bar{y}})$	$\boldsymbol{(x - \bar{x})^2}$	$\boldsymbol{(x - \bar{x})(y - \bar{y})}$
	0	2	-2	-2	4	4
	1	3	-1	-1	1	1
	2	5	0	1	0	0
	3	4	1	0	1	0
	4	6	2	2	4	4
Sum	**10**	**20**			**10**	**9**
Average	**2**	**4**				

We have linear regression equation as **$y = b_0 + b_1 x$**

Using least square method,

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{9}{10} = 0.9$$

$$b_0 = \bar{y} - b_1\bar{x} = 4 - 0.9 \times 2 = 2.2$$

∴ The linear regression line becomes: $\boldsymbol{y = 2.2 + 0.9 \times x}$

The value of y when x is 10:

$$\boldsymbol{y = 2.2 + 0.9 \times 10 = 11.2}$$

Q5. The rent of the property is related to its area. Given the area in square feet and rent in rupees in table below. Find the relation between area and rent using the concept of linear regression. Also predict the rent for a property of 790 ft^2.

Sr. No.	1	2	3	4	5	6
Area (ft^2)	340	1080	640	880	990	510
Rent (₹)	500	1700	1100	800	1400	500

Solution:

1. **Using least square method:**

Sr. No.	Area (ft^2) (x)	Rent (₹) (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	340	500	-400	-500	160000	200000
2	1080	1700	340	700	115600	238000
3	640	1100	-100	100	10000	-10000
4	880	800	140	-200	19600	-28000
5	990	1400	250	400	62500	100000
6	510	500	-230	-500	52900	115000
Sum	**4440**	**6000**			**420600**	**615000**
Average	**740**	**1000**				

We have linear regression equation as **$y= b_0+b_1x$**

$$b_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{615000}{420600} = 1.4622$$

$$b_0 = \bar{y} - b_1\bar{x} = 1000 - 1.4622 \times 740 = -82.028$$

∴ **The linear regression line becomes: $\boldsymbol{y = -82.028 + 1.4622 \times x}$**

The value of rent when area is 790 ft^2:

$$\boldsymbol{y = -82.028 + 1.4622 \times 790 = 1073.11 \approx 1073}$$

2. Using Karl Correlation Coefficient method:

Sr. No.	Area (ft^2) (x)	Rent (₹) (y)	x^2	$x \times y$
1	340	500	115600	170000
2	1080	1700	1166400	1836000
3	640	1100	409600	704000
4	880	800	774400	704000
5	990	1400	980100	1386000
6	510	500	260100	255000
Sum	**4440**	**6000**	**3706200**	**5055000**

We have linear regression equation as **$y= b_0+b_1x$**

Using Karl Correlation Coefficient method,

$$b_1 = \frac{n \times \sum xy - \sum x \sum y}{n \times \sum x^2 - (\sum x)^2} = \frac{6 \times 5055000 - 4440 \times 6000}{6 \times 3706200 - 4440^2} = 1.4622$$

$$b_0 = \frac{\sum y - b_1 \times \sum x}{n} = \frac{6000 - 1.4622 \times 4440}{6} = -82.028$$

∴ **The linear regression line becomes: $\boldsymbol{y = -82.028 + 1.4622 \times x}$**

The value of rent when area is 790 ft^2:

$$\boldsymbol{y = -82.028 + 1.4622 \times 790 = 1073.11 \approx 1073}$$

Logistic Regression

- Logistic regression is basically a supervised classification algorithm.
- In a classification problem, the target variable (or output), y, can take only discrete values for given set of features (or inputs), X.
- **In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, Infectious, etc.) or 0 (FALSE, failure, non-infectious, etc.).**
- The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables.
- Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of dependent variable:

  $\boldsymbol{logit(p)=b_0+b_1X_1+b_2X_2+b_3X_3+...+b_kX_k}$

  where p is the probability of presence of the dependent variable.
- The logit transformation is defined as the logged odds, where

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristics}{probability\ of\ absence\ of\ characteristics}$$

  and $logit(p) = \ln(\frac{p}{1-p})$
- Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.
- Logistic regression's ability to provide probabilities and classify new samples using continuous and discrete measurements makes it a popular machine learning method.
- The **logistic regression** technique involves dependent variable which can be represented in the binary (0 or 1, true or false, yes or no) values, means that the outcome could only be in either one form of two. For example, it can be utilized when we need to find the probability of successful or fail event.
- Here, the same formula as of linear regression is used with the additional sigmoid function, and the value of Y ranges from 0 to 1.
- Logistic regression equation:

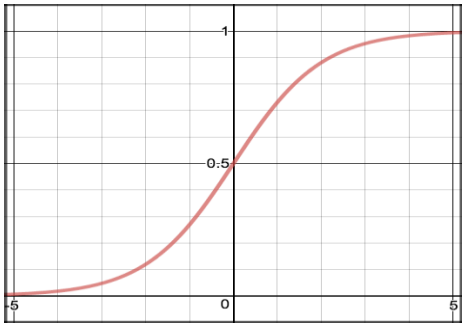
  Linear regression $Y=b_0+b_1\times X_1+b_2\times X_2+...+b_k\times X_k$

  Sigmoid Function $P = \frac{1}{1+e^{-Y}}$

  By putting Y in Sigmoid function, we get the following result.

$$\ln\left(\frac{p}{1-p}\right)=b_0+b_1\times X_1+b_2\times X_2+...+b_k\times X_k$$

- The logistic function has asymptotes at 0 and 1, and it crosses the y-axis at 0.5 as shown in figure below.



- A prediction function in logistic regression returns the probability of our observation being positive, True, or "Yes". We call this class 1 and its notation is *P*(*class*=1).
- In order to map this to a discrete class (true/false, cat/dog), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 2.
- p≥0.5, class=1
- p<0.5, class=0

Q1. Say we're given data on student exam results and our goal is to predict whether a student will pass or fail based on number of hours slept and hours spent studying. We have two features (hours slept, hours studied) and two classes: passed (1) and failed (0).

Studied (x_1)	**Slept (x_2)**	**Passed (y)**
4.85	9.63	0
8.62	3.23	1
5.43	8.23	0
9.21	6.34	1

Let b_0 = -0.406, b_1 = 0.8525, b_2 = -1.105. Determine the predicted value of y.

For 1st input tuple, x_1 = 4.85 and x_2 = 9.63,

$$y = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}} = \frac{1}{1+e^{-(-0.406+0.8525*4.85-1.105*9.63)}} = 0.0022$$

For 2nd input tuple, x_1 = 8.62 and x_2 = 3.23,

$$y = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}} = \frac{1}{1+e^{-(-0.406+0.8525*8.62-1.105*3.23)}} = 0.96$$

For 3rd input tuple, x_1 = 5.43 and x_2 = 8.23,

$$y = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}} = \frac{1}{1+e^{-(-0.406+0.8525*5.43-1.105*8.23)}} = 0.0076$$

For 4th input tuple, x_1 = 9.21 and x_2 = 6.34,

$$y = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2)}} = \frac{1}{1+e^{-(-0.406+0.8525*4.85-1.105*9.63)}} = 0.6082$$

The predicted and actual value are shown in table below:

Studied (x_1)	**Slept (x_2)**	**Passed (y)**	**Fitted Value**	**Prediction**
4.85	9.63	0	0.0022	0
8.62	3.23	1	0.96	1
5.43	8.23	0	0.0076	0
9.21	6.34	1	0.6082	1

Confusion Matrix: Helps classify the values that were correctly predicted using the model built.

	Predicted 0	Predicted 1
Actual 0	True Negative (TN)	False Positive (FP)
Actual 1	False Negative (FN)	True Positive (TP)

From the table above for the example:

	Predicted 0	Predicted 1
Actual 0	2 (TN)	0 (FP)
Actual 1	0 (FN)	2 (TP)

$$Accuracy = \frac{TP + TN}{N} = \frac{2+2}{4} = \frac{4}{4} = 1\ i.e. 100\%$$

$$Precision\ (Positive\ Predictive\ Rate) = \frac{TP}{TP + FP} = \frac{2}{2+0} = 1$$

$$(Negative\ Predictive\ Rate) = \frac{TN}{TN + FN} = \frac{2}{2+0} = 1$$

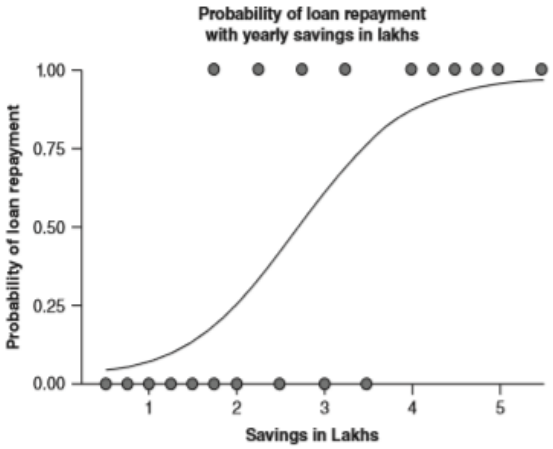
$$Sensitivity = \frac{TP}{TP + FN} = \frac{2}{2+0} = 1$$

$$Specificity = \frac{TN}{TN + FP} = \frac{2}{2+0} = 1$$

Q2. Table below shows the amount of saving of individual customers (in lakhs) and whether they are loan non-defaulters (0 indicates loan defaulters and 1 indicates loan non-defaulters). Determine the predicted value of each savings using Logistic Regression. Let b_0 = -4.07778, b_1 = 1.5046. Also calculate the accuracy.

Amount in Savings (in lakhs)	Loan Non-Defaulter
0.50	0
0.75	0
1.00	0
1.25	0
1.50	0
1.75	0
1.75	1
2.00	0
2.25	1
2.50	0
2.75	1
3.00	0
3.25	1
3.50	0
4.00	1
4.25	1
4.50	1
4.75	1
5.00	1
5.50	1

Solution:



Given b_0 = -4.07778, b_1 = 1.5046

Using Logistic Regression, probability of customer being loan non-defaulter is given by

$$y = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}} = \frac{1}{1 + e^{-(-4.0778 + 1.5046 * savings)}}$$

Amount in Savings (in lakhs)	Loan Non-Defaulter/Defaulter	Fitted Value	Prediction
0.50	0	0.035	0
0.75	0	0.049	0
1.00	0	0.071	0
1.25	0	0.100	0
1.50	0	0.139	0
1.75	0	0.191	0
1.75	1	0.191	0
2.00	0	0.256	0
2.25	1	0.334	0
2.50	0	0.422	0
2.75	1	0.515	1
3.00	0	0.607	1
3.25	1	0.693	1
3.50	0	0.766	1
4.00	1	0.874	1
4.25	1	0.910	1
4.50	1	0.937	1
4.75	1	0.956	1
5.00	1	0.969	1
5.50	1	0.985	1

From the table above for the example, the confusion matrix is:

	Predicted 0	Predicted 1
Actual 0	8 (TN)	2 (FP)
Actual 1	2 (FN)	8 (TP)

$$Accuracy = \frac{TP + TN}{N} = \frac{8 + 8}{20} = \frac{16}{20} = 0.8\ i.e.\ 80\%$$

Linear Regression Vs Logistic Regression

Linear Regression	Logistic Regression
Linear regression is used to predict the continuous dependent variable using a given set of independent variables.	Logistic Regression is used to predict the categorical dependent variable using a given set of independent variables.
Linear Regression is used for solving Regression problem.	Logistic regression is used for solving Classification problems.
In Linear regression, we predict the value of continuous variables.	In logistic Regression, we predict the values of categorical variables.
In linear regression, we find the best fit line, by which we can easily predict the output.	In Logistic Regression, we find the S-curve by which we can classify the samples.
Least square estimation method is used for estimation of accuracy.	Maximum likelihood estimation method is used for estimation of accuracy.
The output for Linear Regression must be a continuous value, such as price, age, etc.	The output of Logistic Regression must be a Categorical value such as 0 or 1, Yes or No, etc.
In Linear regression, it is required that relationship between dependent variable and independent variable must be linear.	In Logistic regression, it is not required to have the linear relationship between the dependent and independent variable.
In linear regression, there may be collinearity between the independent variables.	In logistic regression, there should not be collinearity between the independent variable