
Machine Learning
Regression

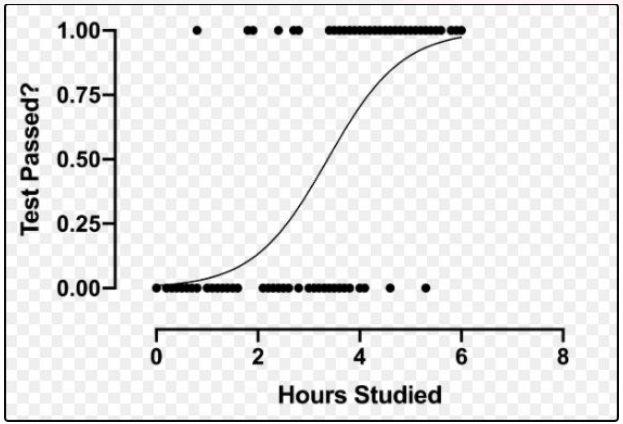
Regression

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables)

Y--outputs or responses X--inputs or predictors

Regression

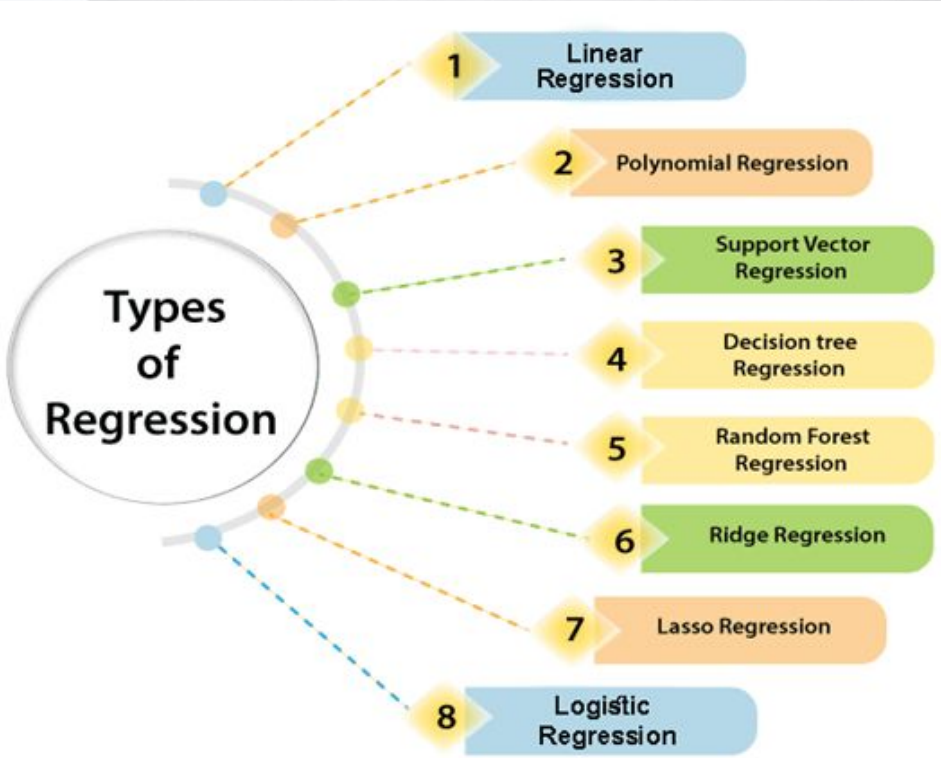
Year	Sales (Million Euro)	Advertising (Million Euro)
1	651	23
2	762	26
3	856	30
4	1,063	34
5	1,190	43
6	1,298	48
7	1,421	52
8	1,440	57
9	1,518	58



Need of regression

- To determine ***if*** **and** ***to how & what extent*** the independent variables impact or influences the dependent variable. For example, whether experience or gender impact salaries.
- Also, useful when you want **to forecast a response** using a new set of predictors. For example, predict electricity consumption of a household for the next hour given the outdoor temperature, time of day, and number of residents in that household.

Types of Regression techniques



Linear Regression

- Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

The **simple linear regression** model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

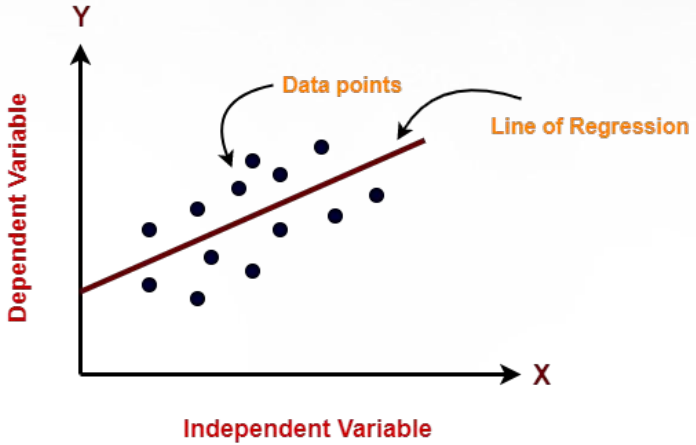
$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

Simple linear regression

- This model has single independent and single dependent variable.
- Eg: the experience impact salaries

$$y = \beta_0 + \beta_1 X$$



- This model has single independent and single dependent variable.
- Eg: the experience impact salaries

SIMPLE LINEAR REGRESSION MODEL

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

More education translates into a higher income

	Production ($ million)	Electricity usage (million kWh)
January	4.51	2.48
February	3.58	2.26
March	4.31	2.47
April	5.06	2.77
May	5.64	2.99
June	4.99	3.05
July	5.29	3.18
August	5.83	3.46
September	4.70	3.03
October	5.61	3.26
November	4.90	2.67
December	4.20	2.53

FIGURE 12.5 ●
Car plant electricity usage data set

Example

	Production ($ million)	Electricity usage (million kWh)
January	4.51	2.48
February	3.58	2.26
March	4.31	2.47
April	5.06	2.77
May	5.64	2.99
June	4.99	3.05
July	5.29	3.18
August	5.83	3.46
September	4.70	3.03
October	5.61	3.26
November	4.90	2.67
December	4.20	2.53

FIGURE 12.5 ●
Car plant electricity usage data set

$$\sum_{i=1}^{12} x_i = 4.51 + \cdots + 4.20 = 58.62$$

$$\sum_{i=1}^{12} y_i = 2.48 + \cdots + 2.53 = 34.15$$

$$\sum_{i=1}^{12} x_i^2 = 4.51^2 + \cdots + 4.20^2 = 291.2310$$

$$\sum_{i=1}^{12} y_i^2 = 2.48^2 + \cdots + 2.53^2 = 98.6967$$

$$\sum_{i=1}^{12} x_i y_i = (4.51 \times 2.48) + \cdots + (4.20 \times 2.53) = 169.2532$$

Calculate β_1 and β_0

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{(12 \times 169.2532) - (58.62 \times 34.15)}{(12 \times 291.2310) - 58.62^2} = 0.49883$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = \frac{34.15}{12} - (0.49883 \times \frac{58.62}{12}) = 0.4090$$

The fitted regression line :

$$y = \beta_0 + \beta_1 x = 0.409 + 0.499x$$

$$\hat{y}|_{5.5} = 0.409 + (0.499 \times 5.5) = 3.1535$$

If the production is $5.5 million then predict the eletricity usage

$$\hat{y}|_{5.5} = 0.409 + (0.499 \times 5.5) = 3.1535$$

Example 1:
A patient is given a drip feed containing a particular chemical and its concentration in his blood is measured, in suitable units, at one hour intervals. The doctors believe that a linear relationship will exist between the variables.

Time, x (hours)	0	1	2	3	4	5	6
Concentration, y	2.4	4.3	5.0	6.9	9.1	11.4	13.5

Calculate concentration after 3.5 hours?

Find how long it would be before the concentration reaches 8 units?

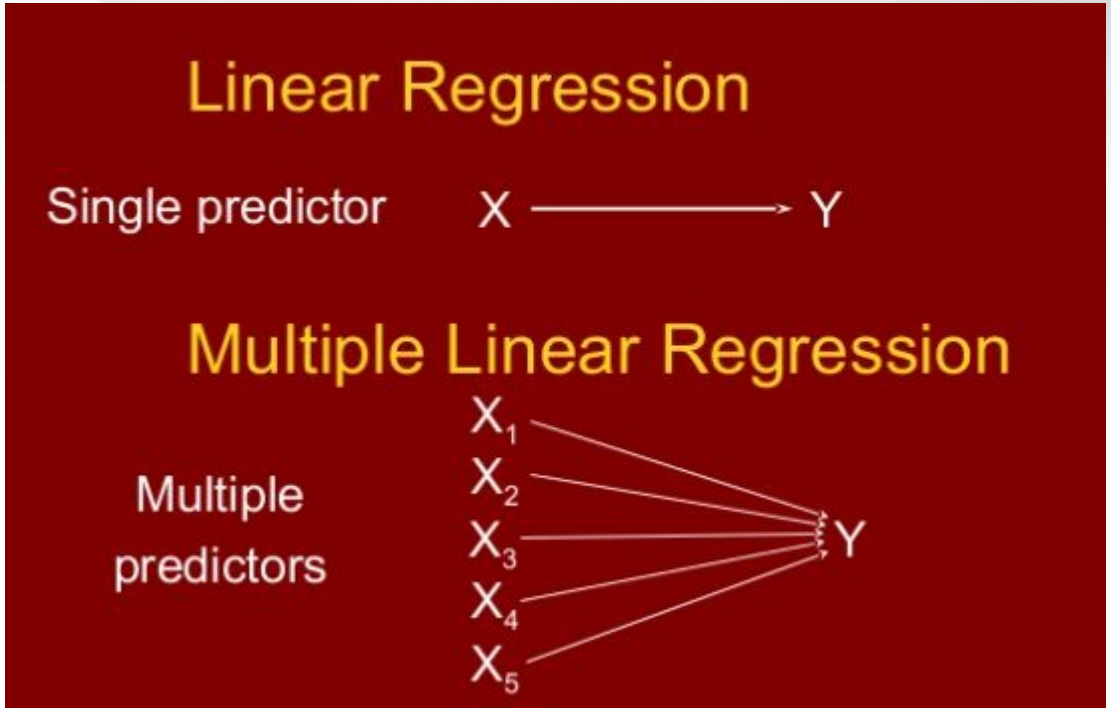
Example 2:
The heights and weights of a sample of 11 students are:

Height (m) h	1.36	1.47	1.54	1.56	1.59	1.63	1.66	1.67	1.69	1.74	1.81
Weight (kg) w	52	50	67	62	69	74	59	87	77	73	67

$[n=11 \quad \sum h=17.72 \quad \sum h^2=28.705 \quad \sum w=737 \quad \sum w^2=50571 \quad \sum hw=1196.1]$

a) Calculate the regression line of w on h.
b) Use the regression line to estimate the weight of someone whose height is 1.6m.

Multiple Linear Regression



Multiple regression simultaneously considers the influence of multiple explanatory variables on a response variable Y

$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Sparse Method

- There may be uninformative variables in the data which prevent proper modeling of the problem and thus, the building of a correct regression model.
- In such cases, **a feature selection process is crucial** to select only the informative features and discard non-informative ones. This can be achieved by ***sparse methods***

```
CRIM,ZN,INDUS,CHAS,NOX,RM,AGE,DIS,RAD,TAX,PTRATIO,B,LSTAT,MEDV
0.00632,18,2.31,0,0.538,6.575,65.2,4.09,1,296,15.3,396.9,4.98,24
0.02731,0,7.07,0,0.469,6.421,78.9,4.9671,2,242,17.8,396.9,9.14,21.6
0.02729,0,7.07,0,0.469,7.185,61.1,4.9671,2,242,17.8,392.83,4.03,34.7
0.03237,0,2.18,0,0.458,6.998,45.8,6.0622,3,222,18.7,394.63,2.94,33.4
0.06905,0,2.18,0,0.458,7.147,54.2,6.0622,3,222,18.7,396.9,NA,36.2
0.02985,0,2.18,0,0.458,6.43,58.7,6.0622,3,222,18.7,394.12,5.21,28.7
```

Correlation among variables in dataset

- To study the relation among multiple variables in a dataset, there are different options.
- We can study the relationship between several variables in a dataset by using the functions **corr and heatmap**
- **Corr** allows to calculate a correlation matrix for a dataset
- The **heat map** is a matricial image which helps to interpret the correlations among variables.

Without Using LASSO Sparse Method

```
Training and testing set sizes (253, 13) (253, 13)
Coeff and intercept: [ 1.20133313 0.02449686 0.00999508
0.42548672 -8.44272332 8.87767164 -0.04850422 -1.11980855
0.20377571 -0.01597724 -0.65974775 0.01777057 -0.11480104]
-10.0174305829
Testing Score: -2.24420202674
Training MSE: 9.98751732546
Testing MSE: 302.64091133
```

Using LASSO Sparse Method

```
Coeff and intercept: [ 0. 0.01996512 -0. 0. -0. 7.69894744
-0.03444803 -0.79380636 0.0735163 -0.0143421 -0.66768539
0.01547437 -0.22181817] -6.18324183615
Testing Score: 0.501127529021
Training MSE: 10.7343110095
Testing MSE: 46.5381680949
```

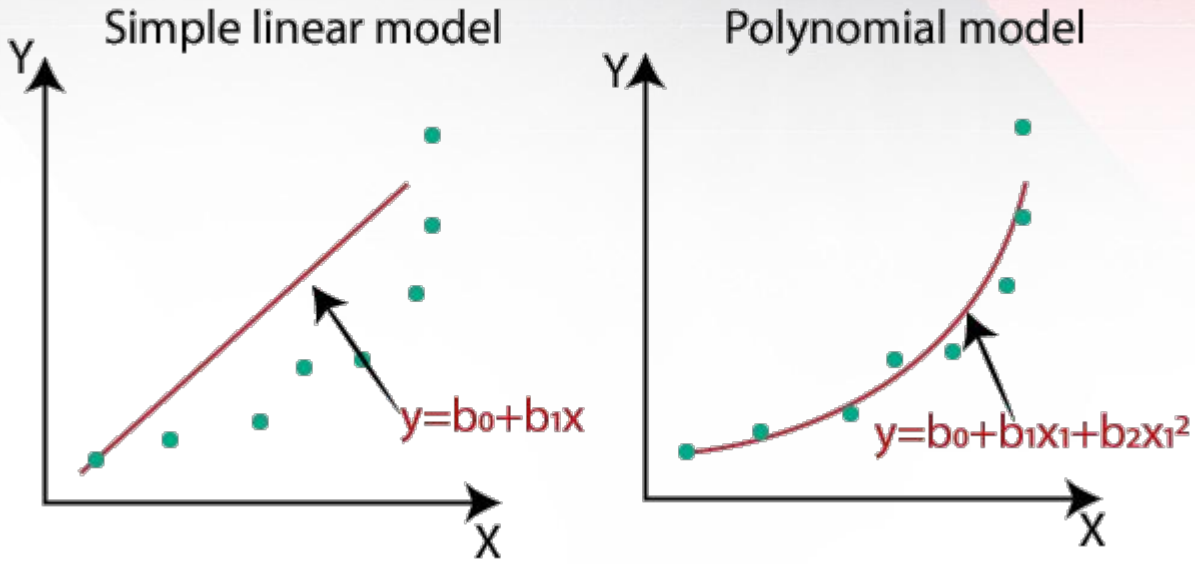
Feature Selection using LASSO method

```
Ordered variable (from less to more important): ['CRIM' 'INDUS'
'CHAS' 'NOX' 'TAX' 'B' 'ZN' 'AGE' 'RAD' 'LSTAT' 'PTRATIO' 'DIS'
'RM']
```

Feature Selection using K highest score method k=5

```
Selected features: [(False, 'CRIM'), (False, 'ZN'), (True,
'INDUS'), (False, 'CHAS'), (False, 'NOX'), (True, 'RM'), (True,
'AGE'), (False, 'DIS'), (False, 'RAD'), (False, 'TAX'), (True,
'PTRATIO'), (False, 'B'), (True, 'LSTAT')]
```

Polynomial Regression



Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

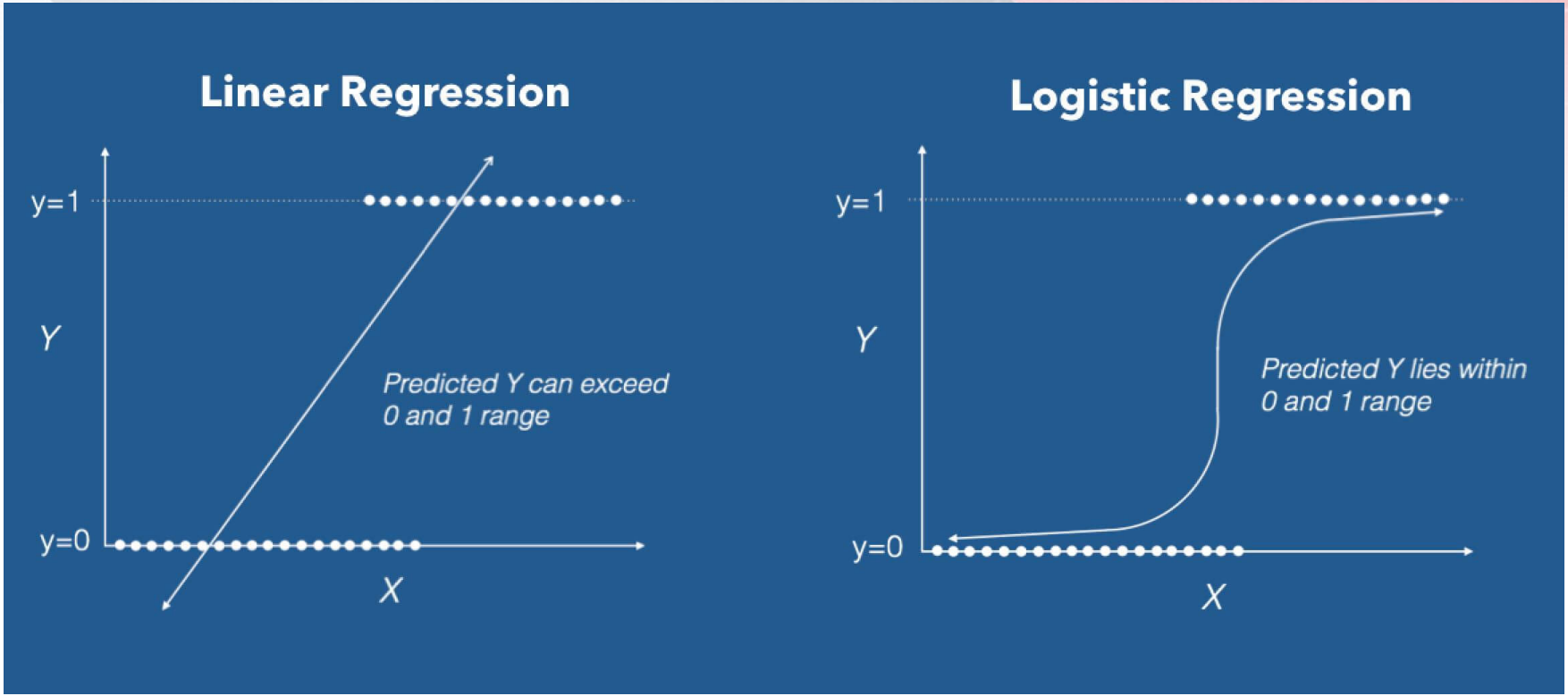
$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

https://online.stat.psu.edu/stat501/lesson/9/9.8

Logistic Regression

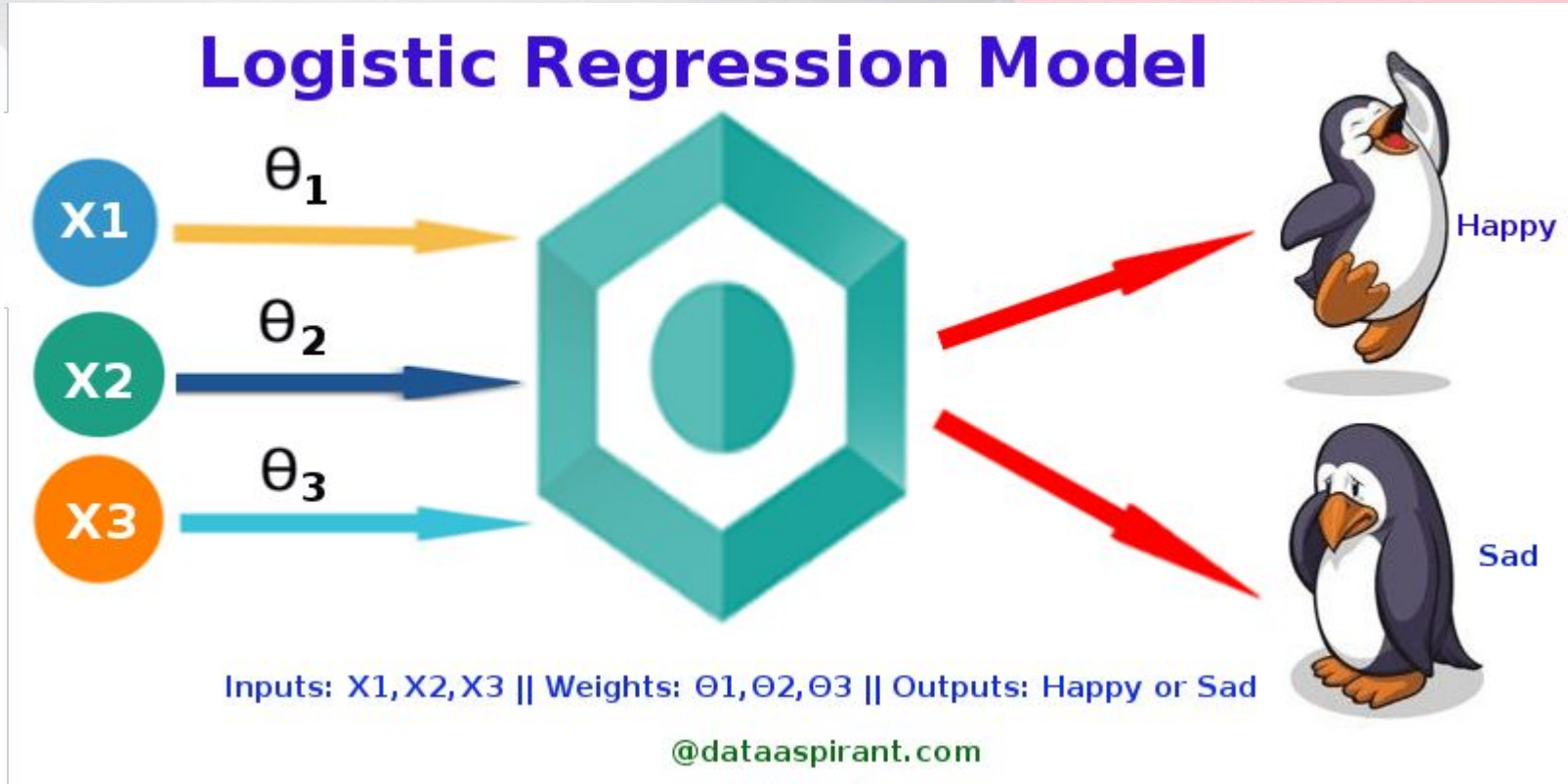


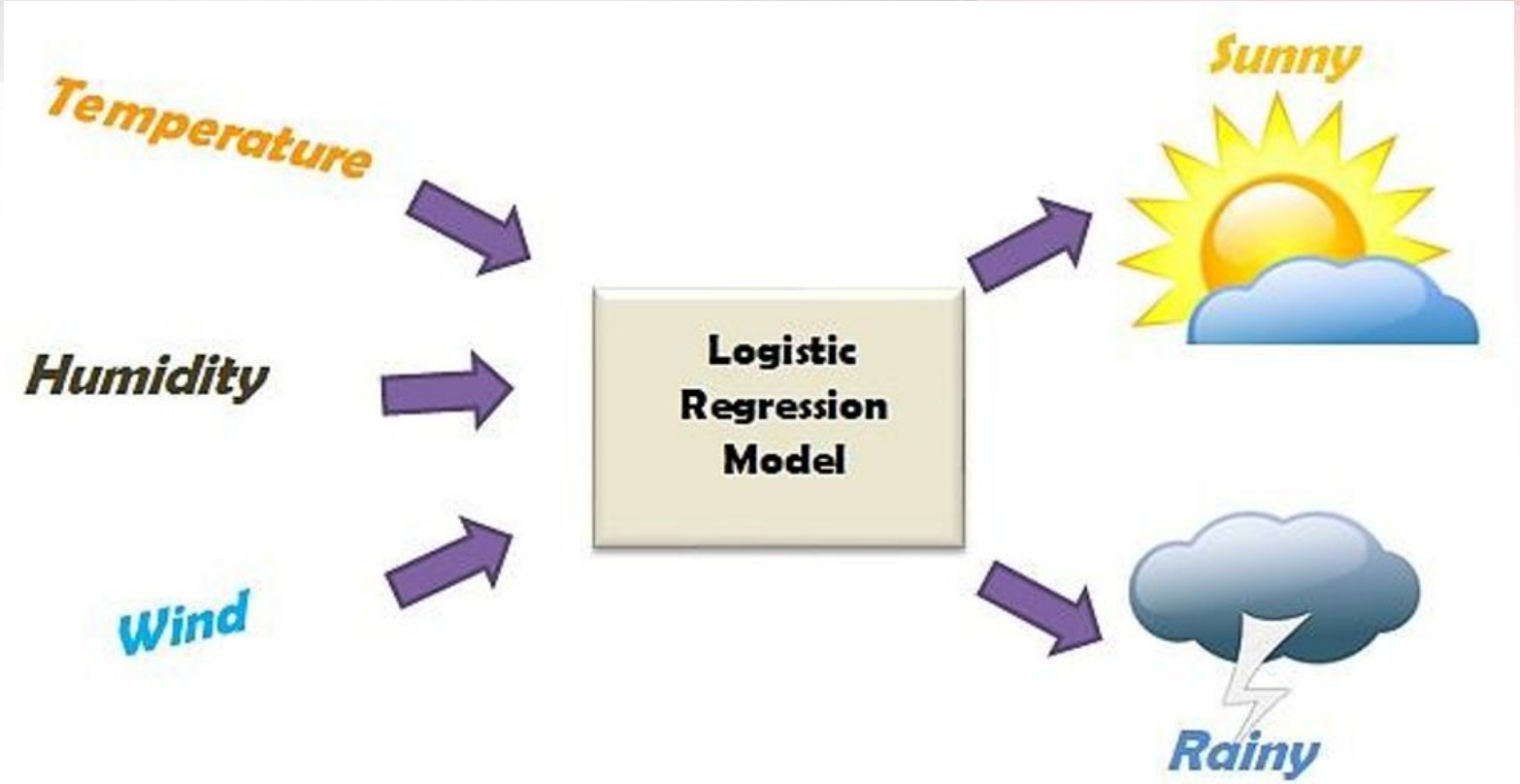
Logistic Regression

- **Logistic regression** is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.
- The logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

The form of the logistic function is:

$$f(x) = \frac{1}{1 + e^{-\lambda x}}$$


Logistic Regression Model
X1
X2
X3
θ1
θ2
θ3
Happy
Sad
Inputs: X1, X2, X3 || Weights: Θ1, Θ2, Θ3 || Outputs: Happy or Sad
@dataaspirant.com


Temperature
Humidity
Wind
Logistic Regression Model
Sunny
Rainy

- **The performance of the regression model**

- The Mean Square Error, Mean Absolute Error, Root Mean Square Error, and R-Squared or Coefficient of determination metrics are used to evaluate the performance of a model in regression analysis.

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

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

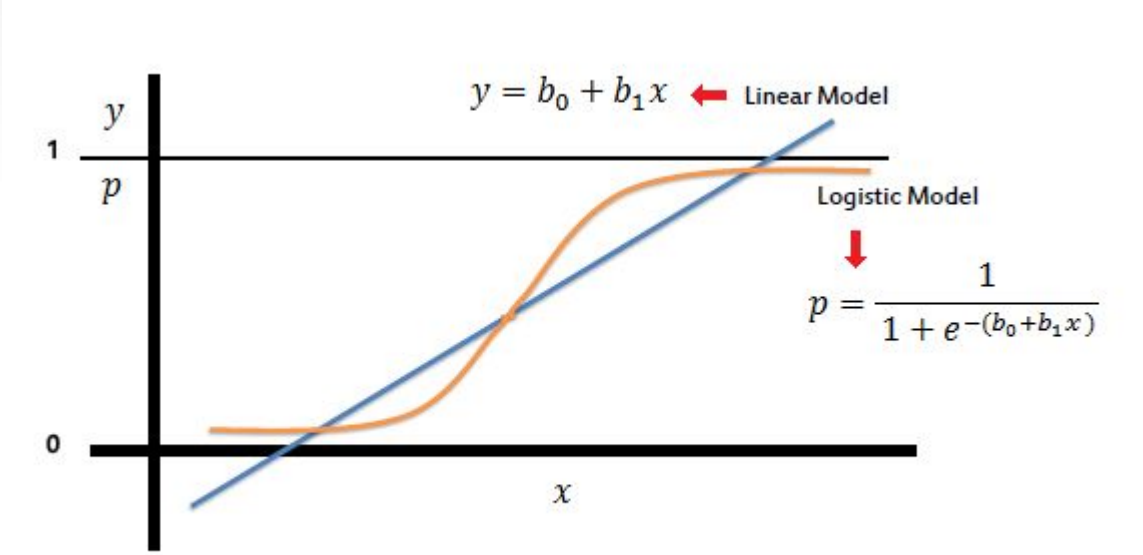
$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y


$y = b_0 + b_1 x$
Linear Model
y
1
p
Logistic Model
$p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$
0
x

1. From the data given below

Marks in Economics:	25	28	35	32	31	36	29	38	34	32
Marks in Statistics:	43	46	49	41	36	32	31	30	33	39

Find (a) The two regression equations, (b) The coefficient of correlation between marks in Economics and statistics, (c) The mostly likely marks in Statistics when the marks in Economics is 30.

2. The heights (in cm.) of a group of fathers and sons are given below

Heights of fathers:	158	166	163	165	167	170	167	172	177	181
Heights of Sons :	163	158	167	170	160	180	170	175	172	175

Find the lines of regression and estimate the height of son when the height of the father is 164 cm.