Correlation:

- Is there a relationship between *x* and *y*?
- It indicates the degree of closeness and direction of relationship between two variables.
- What is the strength of this relationship
  - Karl Pearson's coefficient of correlation (r)

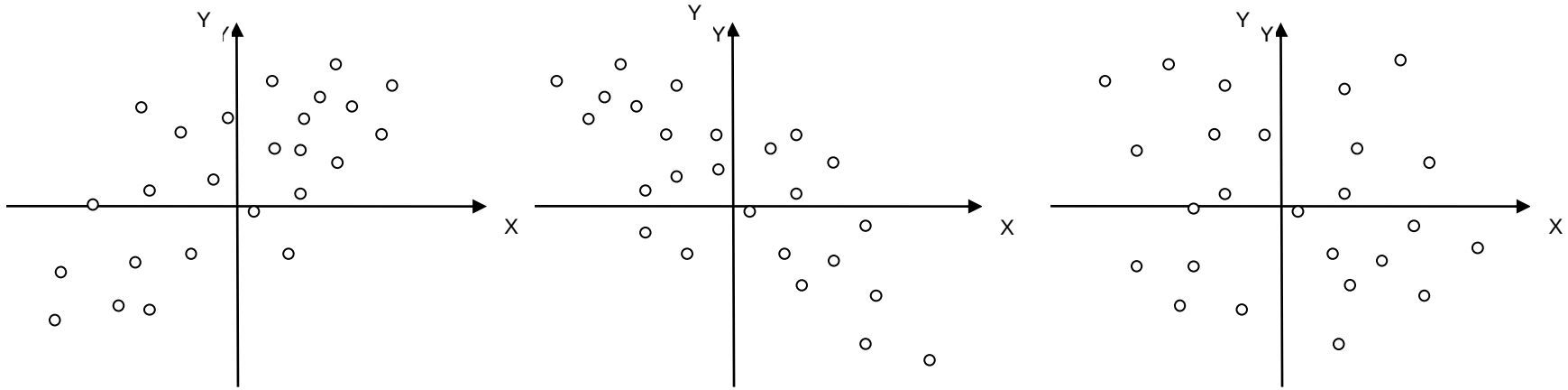
Positive:

- Age of husband and age of wife
- Price of commodity and its demand

Negative:

- Rainy days increases cold drink demand decreases
- Sales of woolen clothes and temperature increasing

Scattergrams



Positive correlation

Negative correlation

No correlation

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Given below are the monthly income and their net savings of a sample of 10 supervisory staff belonging to a firm. Calculate the correlation coefficient.

Monthly income(Rs.):780 360 980 250 750 820 900 620 650 390
Net Savings: 84 51 91 60 68 62 86 58 53 47

r= 0.78 approx.

Calculate the coefficient of correlation between the height (in inches) of father and height of son (in inches) from the following data:
Height of father: 64 65 66 67 68 69 70
Height of son : 66 67 65 68 70 68 72

r=0.81

Regression

•Can we describe this relationship and use this to predict *y* from *x*?

•how well a certain independent variable predict dependent variable?

•It provide the nature of the relationship i.e. In what form they are related.

•It describe the functional relationship which enable us to make estimates of one variable from another.

e.g. Age and blood pressure are correlated and we can estimate the expected amount of blood pressure (y) on the age of individual(x)

Regression analysis:

It is an important statistical method that allows us to examine the relationship between two or more variables in the dataset.

In this we have a dependent variable — the main factor that we are trying to understand or predict. And then we have independent variables — the factors we believe have an impact on the dependent variable.

Simple linear regression(or Univariate Regression) is a regression model that estimates the relationship between a dependent variable and an independent variable using a straight line.

Example: Age Versus Blood pressure, Income versus saving

Multiple linear regression estimates the relationship between two or more independent variables and one dependent variable. The difference between these two models is the number of independent variables.

Mean Absolute Error:

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

Mean squared Error:

$$mse = \frac{\sum_{i=1}^{n} \left(y_i - \lambda(x_i)\right)^2}{n}$$

where y_i is the true target value for test instance x_i, $\lambda(x_i)$ is the predicted target value for test instance x_i, and n is the number of test instances.

Logistic regression:

Logistic regression can be binomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types, "0" and "1" (which may represent, for example, "dead" vs. "alive" or "win" vs. "loss").

Example:

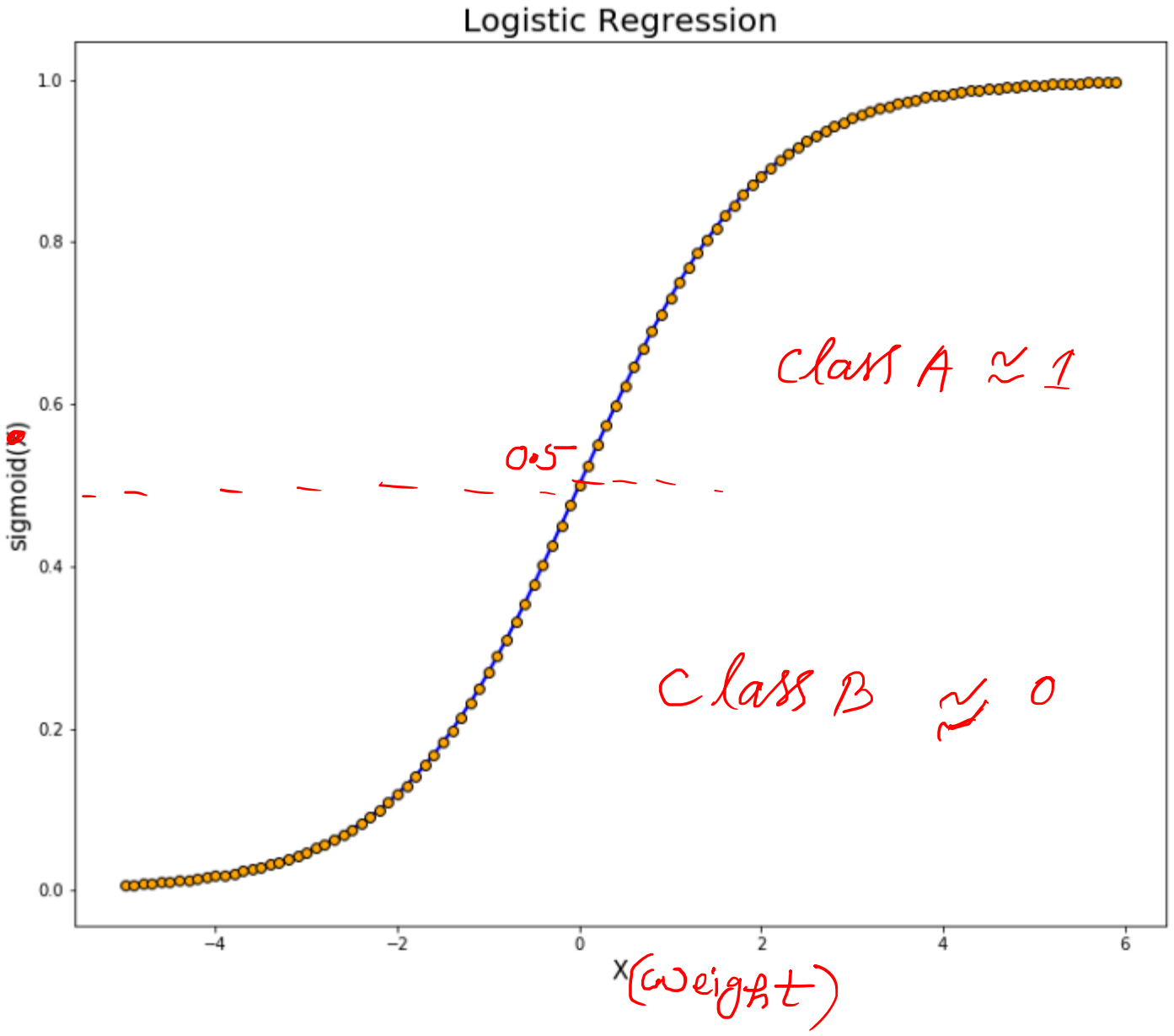
- How does the probability of getting lung cancer (yes vs. no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day?

- Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?

- Weight versus obesity

$$y = \frac{1}{1 + e^{-x}}$$

where y is dependent & x is independent variable Obesity (y)



Examples:

1. Rajesh wants to estimate the price of a house. He will collect details such as the location of the house, number of bedrooms, size in square feet, amenities available, or not. Basis these details price of the house can be predicted and how each variables are interrelated.

2. An agriculture scientist wants to predict the total crop yield expected for the summer. He collected details of the expected amount of rainfall, fertilizers to be used, and soil conditions. By building a Multivariate regression model scientists can predict his crop yield. With the crop yield, the scientist also tries to understand the relationship among the variables.

3. If an organization wants to know how much it has to pay to a new hire, they will take into account many details such as education level, number of experience, job location, has niche skill or not. Basis this information salary of an employee can be predicted, how these variables help in estimating the salary.

4. Economists can use Multivariate regression to predict the GDP growth of a state or a country based on parameters like total amount spent by consumers, import expenditure, total gains from exports, total savings, etc.

5. A company wants to predict the electricity bill of an apartment, the details needed here are the number of flats, the number of appliances in usage, the number of people at home, etc. With the help of these variables, the electricity bill can be predicted.

Mathematical equation:

The simple regression linear model represents a straight line meaning y is a function of x. When we have an extra dimension (z), the straight line becomes a plane.

Here, the plane is the function that expresses y as a function of x and z. The linear regression equation can now be expressed as:

y = m1.x + m2.z+ c

Here y is the dependent variable, that is, the variable that needs to be predicted.

x is the first independent variable. It is the first input. m1 is the slope of x1. It lets us know the angle of the line (x).

z is the second independent variable. It is the second input. m2 is the slope of z. It helps us to know the angle of the line (z).

c is the intercept.

The equation for a model with two input variables can be written as:

$y = \beta 0 + \beta 1.x1 + \beta 2.x2$

What if there are three variables as inputs? Human visualizations can be only three dimensions. In the machine learning world, there can be n number of dimensions. The equation for a model with three input variables can be written as:

$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + \beta 3.x3$

Below is the generalized equation for the multivariate regression model-

$y = \beta 0 + \beta 1.x1 + \beta 2.x2 + \ldots\ldots + \beta n.xn$

What is Cost Function?

The cost function is a function that allows a cost to samples when the model differs from observed data. This equation is the sum of the square of the difference between the predicted value and the actual value divided by twice the length of the dataset. A smaller mean squared error implies a better performance. Here, the cost is the sum of squared errors.

Cost of Multiple Linear regression:

$$\text{MSE}=\frac{1}{2m}\Sigma\left(h_\theta(\text{x})^{(i)}-y^i\right)^2$$

By **the method of least square** the following normal equations are obtained

The regression line of y on x is given by

$y = a + bx$

The two normal equations are

$\Sigma y = na + b\Sigma x$

$\Sigma xy = a \Sigma x + b \Sigma x^2$

The following data relate to number of days and weight gain. Find regression of y on x

Time(days):	3	4	5	6	7	8	9	10
Weight(y)	1.4	1.5	2.2	2.4	3.1	3.2	3.2	3.9

x	y	xy	x^2
3	1.4	4.2	9
4	1.5	6.0	16
5	2.2	11.0	25
6	2.4	14.4	36
7	3.1	21.7	49
8	3.2	25.6	64
9	3.2	28.8	81
10	3.9	39.0	100
$\Sigma x=52$	$\Sigma y=20.9$	$\Sigma xy=150.7$	$\Sigma x^2 =380$

Putting the values in above equations

$20.9 = 8a + 52b$

$150.7 = 52a + 380b$

$b = 0.35$

$a = 0.34$

$y = 0.34 + 0.35x$

The decrease in heart rate (beats/min) due to different concentrations of a drug are given below

Dose of drug(mg)	1.0	1.25	1.50	1.75	2.0	2.25	2.50	2.75	3.0
Decrease in heart beat	10	12	12	14	16	17	20	18	21

Conduct regression analysis

Soln. $y = 4.7 + 5.48x$

Q.1 Logistic regression is used when you want to:

1.Predict a dichotomous variable from continuous or dichotomous variables.
2.Predict a continuous variable from dichotomous variables.
3.Predict any categorical variable from several other categorical variables.
4.Predict a continuous variable from dichotomous or continuous variables.

Q.2 Logistic regression assumes a:

1.Linear relationship between continuous predictor variables and the outcome variable.
2.Linear relationship between continuous predictor variables and the logit of the outcome variable.
3.Linear relationship between continuous predictor variables.
4.Linear relationship between observations.

Q.3 In binary logistic regression:

1.The dependent variable is continuous.
2.The dependent variable is divided into two equal subcategories.
3.The dependent variable consists of two categories.
4.There is no dependent variable.