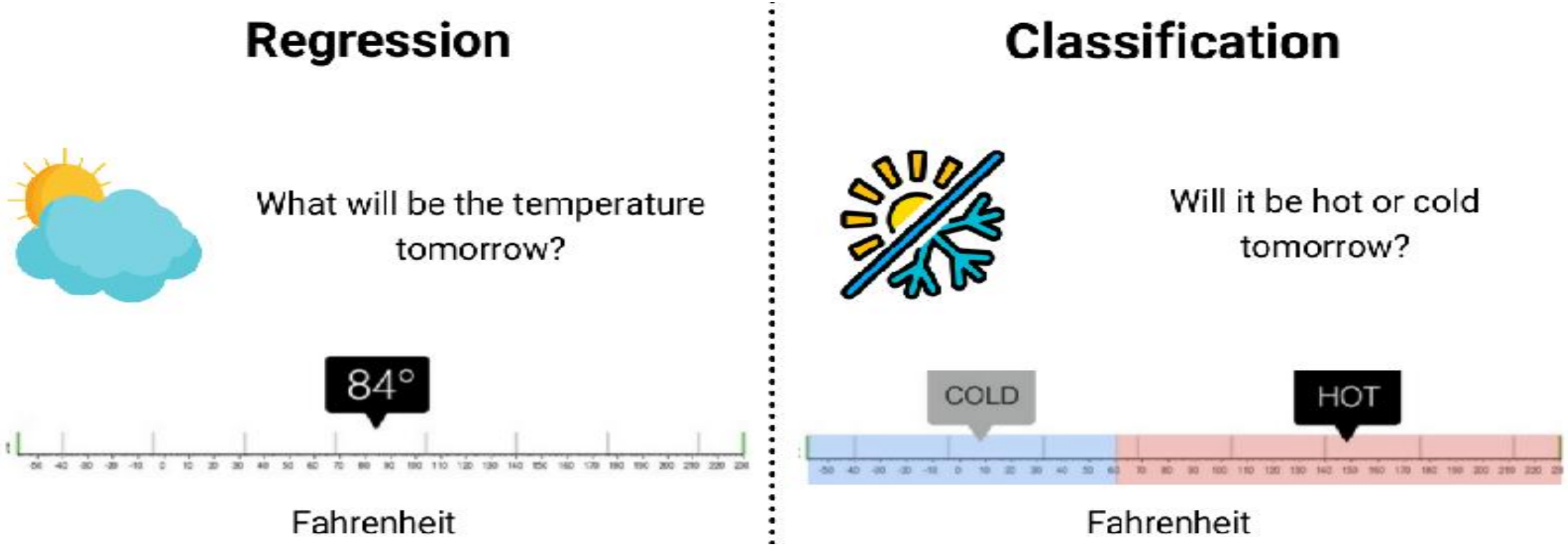
Regression Modelling Techniques

Classification Vs Regression

- Both problems deal with the mapping of the input data to output data but in a different way



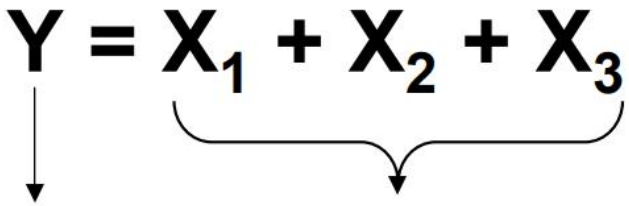
- Regression and classification both are the supervised learning methods, where regression is trained to predict real number outputs and classification is traines to identify/predict to which category the new values fall into.

Regression

- One of the fundamental task in data analysis is to find **how different variables are related to each other** and one of the central tool for learning about such relationships is regression
- Regression model exploits the relationship between two or more variables so that one can gain information about the value of one variable **(dependent/response)** based on value of another variable **(independent/explanatory)**
- Regression can be used for prediction, estimation, hypothesis testing, and modeling causal relationships
- A regression problem is when the output variable is a real or continuous value, such as “salary” or “weight”.

Types of Variable

- **Independent variables** are regarded as inputs to a system and may take on different values freely.
- **Dependent variable** are those values that change as a consequence of changes in the independent values in the system.
- The value of the dependent variable of a linear regression model is a continuous value i.e. real numbers.



Dependent Variable	Independent Variable
Outcome Variable	Predictor Variable
Response Variable	Explanatory Variable

Types of Regression

- The class from which the functions are selected to design the regression model are usually one of the following types:
  - A linear function of single independent variable x (i.e. $y = a+b\,x$) -: **Simple (univariate) linear regression**
  - A linear function of multiple independent variables x1, x2, ... xk (i.e. $y = a+b_1x_1+..b_mx_m$)-: **Multiple (multivariate) linear regression**
  - A polynomial function of independent variable x -: **Polynomial regression**
  - Any other type of function, with one or more parameters (e.g. $y = ae^{bx}$) -: **Nonlinear regression**.
  - When the dependent variable(target) is binary in nature, (0 or 1, true or false, success or failure) it is not required to have a linear relationship among independent and dependent variables. There it uses logarithmic (sigmoid) function -: **Logistic Linear Function**

The most popular types of regression are linear and logistic regressions.

Correlation Coefficient

Correlation: Correlation means association - more precisely it is a measure of the extent to which two variables are related. There are three possible results of a correlational study:

- **A positive correlation** is a relationship between two variables in which both variables move in the same direction. Therefore, when one variable increases as the other variable increases, or one variable decreases while the other decreases. An example of positive correlation would be height and weight. Taller people tend to be heavier.
- **A negative correlation** is a relationship between two variables in which an increase in one variable is associated with a decrease in the other. An example of negative correlation would be height above sea level and temperature. As you climb the mountain (increase in height) it gets colder (decrease in temperature).
- **A zero correlation** exists when there is no relationship between two variables. For example there is no relationship between the amount of tea taken and level of intelligence.

Correlation Coefficient

- A correlation can be expressed visually through a **scattergram** (or scatterplot). It indicates the strength and direction of the correlation between the co-variables (independent and dependent).
- Correlations always deals with paired scores, so the values of the 2 variables should be taken together and used to make the diagram.
- Correlation Ranges between –1 and 1

  - The closer to –1, the stronger the negative linear relationship
  - The closer to 1, the stronger the positive linear relationship
  - The closer to 0, the weaker any positive linear relationship

- A correlation of –1 indicates a perfect negative correlation, meaning that as one variable goes up, the other goes down. A correlation of +1 indicates a perfect positive correlation, meaning that as one variable goes up, the other goes up.

Correlation Coefficient

The following is the formula to estimate the correlation coefficients between two variables X and Y.

$$r = \frac{\text{cov } ariance\ (x, y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}}$$

Where $\text{cov}(x, y) = E(x, y) - E(x)E(y)$

$$= \frac{\sum xy}{n} - \frac{\sum x}{n}\frac{\sum y}{n} = n\sum xy - \sum x \sum y$$

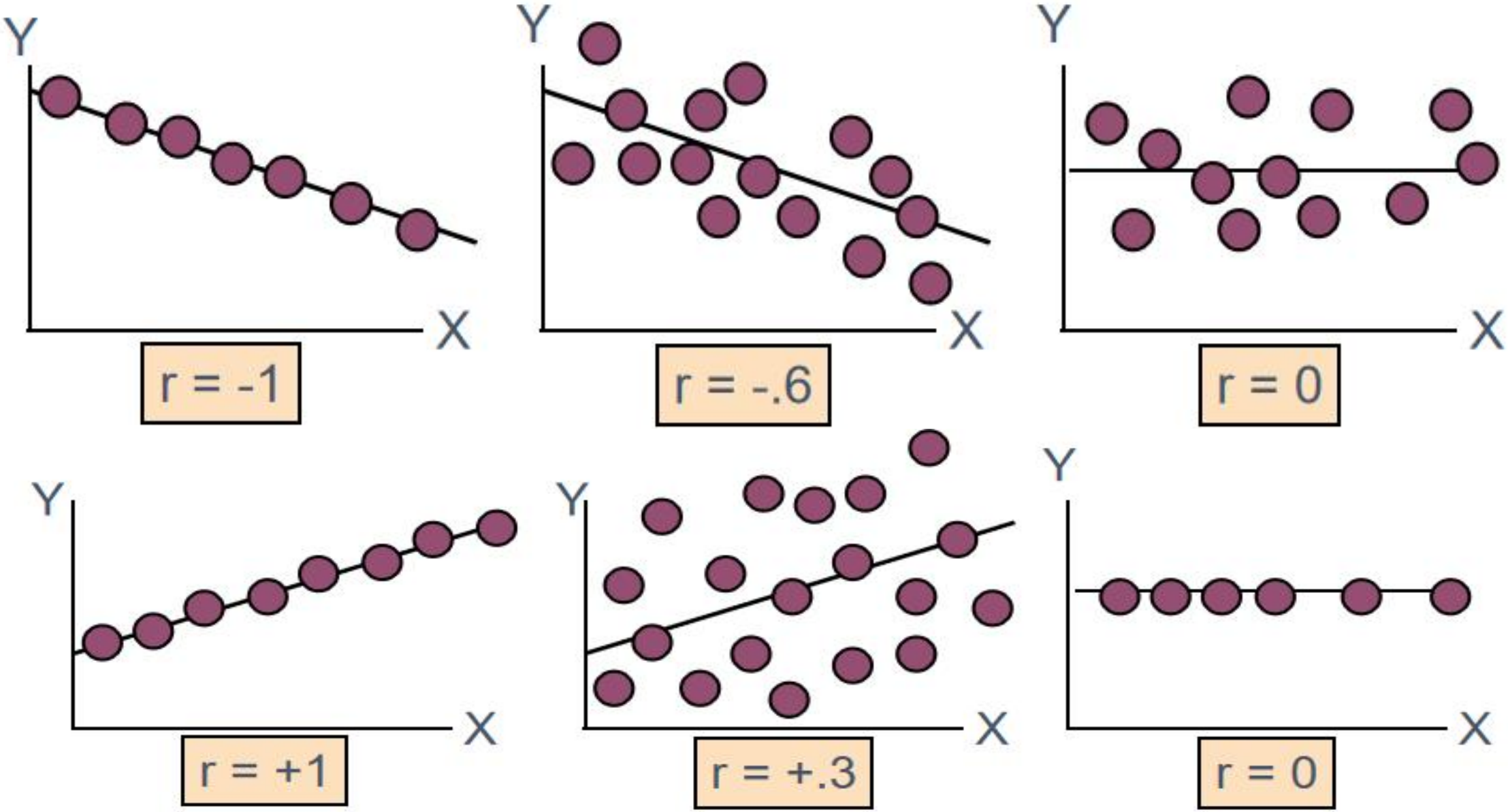
$$\text{var}(x) = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$$

$$\text{var}(y) = \sqrt{\frac{\sum y^2}{n} - \left(\frac{\sum y}{n}\right)^2}$$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

- x is the independent variable and y is the dependent variable.
- n is the number of observations
- r the computed value is known as the correlation coefficients .

Scatter Plots of Data with Various Correlation Coefficients



Correlation Coefficients Calculation

Company	Sales in 1000s (Y)	Number of agents in 100s (X)	X^2	Y^2	XY
A	25	8			
B	35	12			
C	29	11			
D	24	5			
E	38	14			
F	12	3			
G	18	6			
H	27	8			
I	17	4			
J	30	9			

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

$n = 10$

$\sum X = 80$ & $\sum Y = 255$

$\sum XY = 2289$

$\sum X^2 = 756$ & $\sum Y^2 = 7097$

$(\sum X)^2 = 6400$ & $(\sum Y)^2 = 65025$

r = 0.95

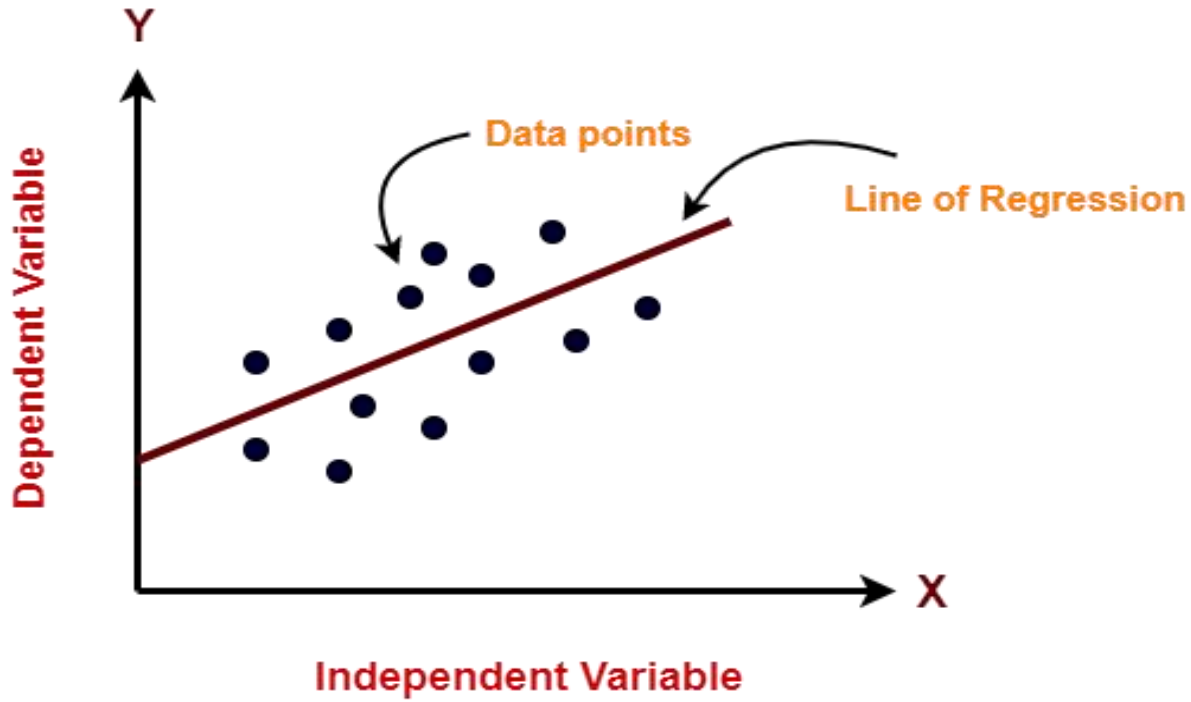
Class Exercise

Find the correlation coefficients of the below sample.

Subject	Age (x)	Glucose Level (y)
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

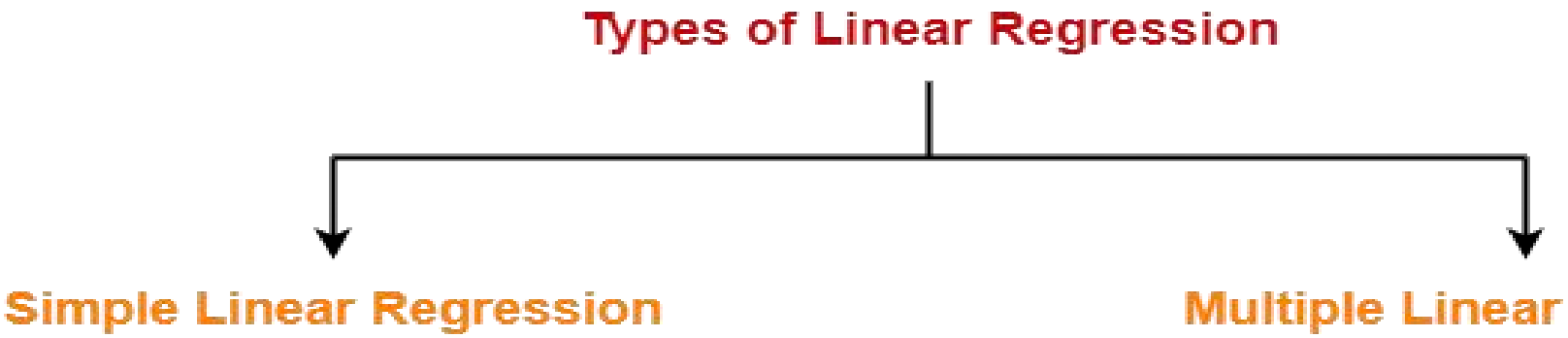
Linear Regression

- Linear Regression is a supervised machine learning algorithm.
- It tries to find the best linear relationship that describes the given data.
- It assumes that there exists a linear relationship between a dependent variable and independent variable(s).
- Linear regression model represents the linear relationship between a dependent variable and independent variable(s) via a sloped straight line.



Types of Linear Regression

Based on the number of independent variables, there are two types of linear regression-



In simple linear regression, the dependent variable depends only on a single independent variable.

For simple linear regression, the form of the model is-

$$Y = \beta_0 + \beta_1 X$$

In multiple linear regression, the dependent variable depends on more than one independent variables.

For multiple linear regression, the form of the model is-

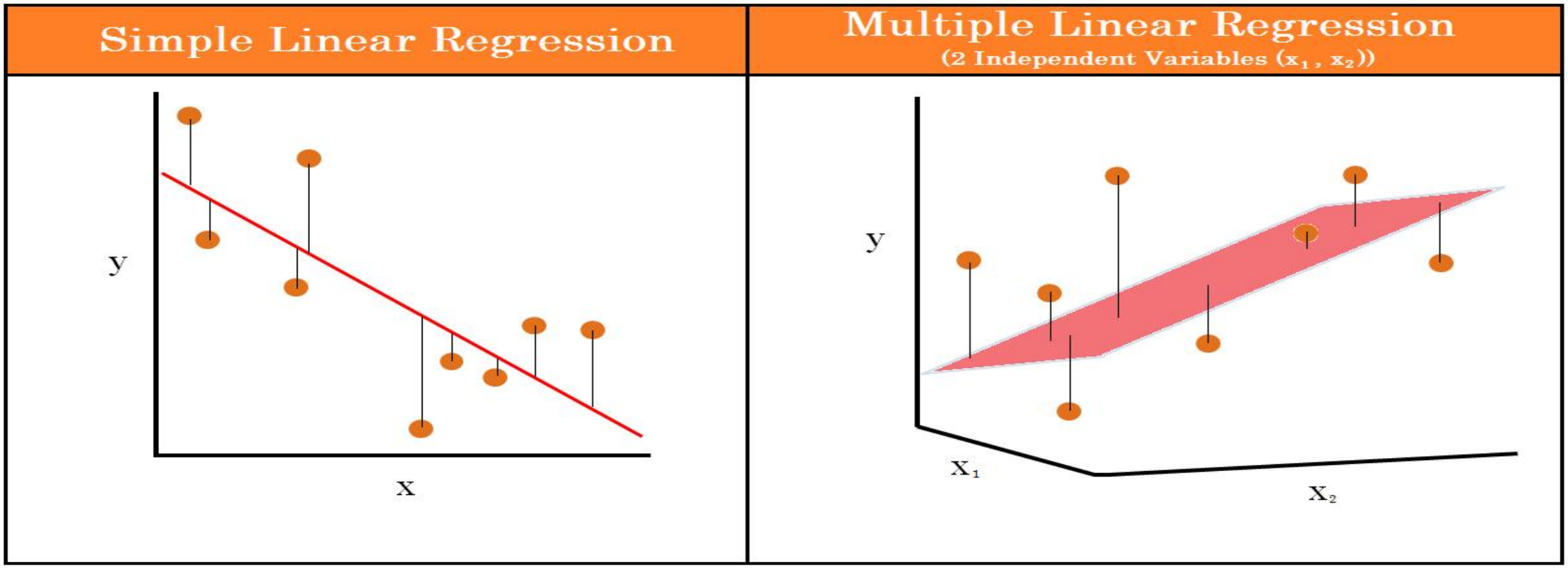
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_n X_n$$

- β_0 is the intercept or the bias that fixes the offset to a line.
- β_1 or β_j (1<=j<=n) is the slope or regression coefficient that specifies the factor by which X or Xj has an impact on Y

Types of Linear Regression …

Fit the data with the **best line** which "goes through" the points

Fit data with the **best hyper plane** which "goes through" the points



Linear Regression

- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory (independent) variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.
- Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher Aptitude scores do not cause higher college grades), but that there is some significant association between the two variables.
- If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

Linear Regression cont…

A linear regression line has an equation of the form **Y = bX +a + e**,

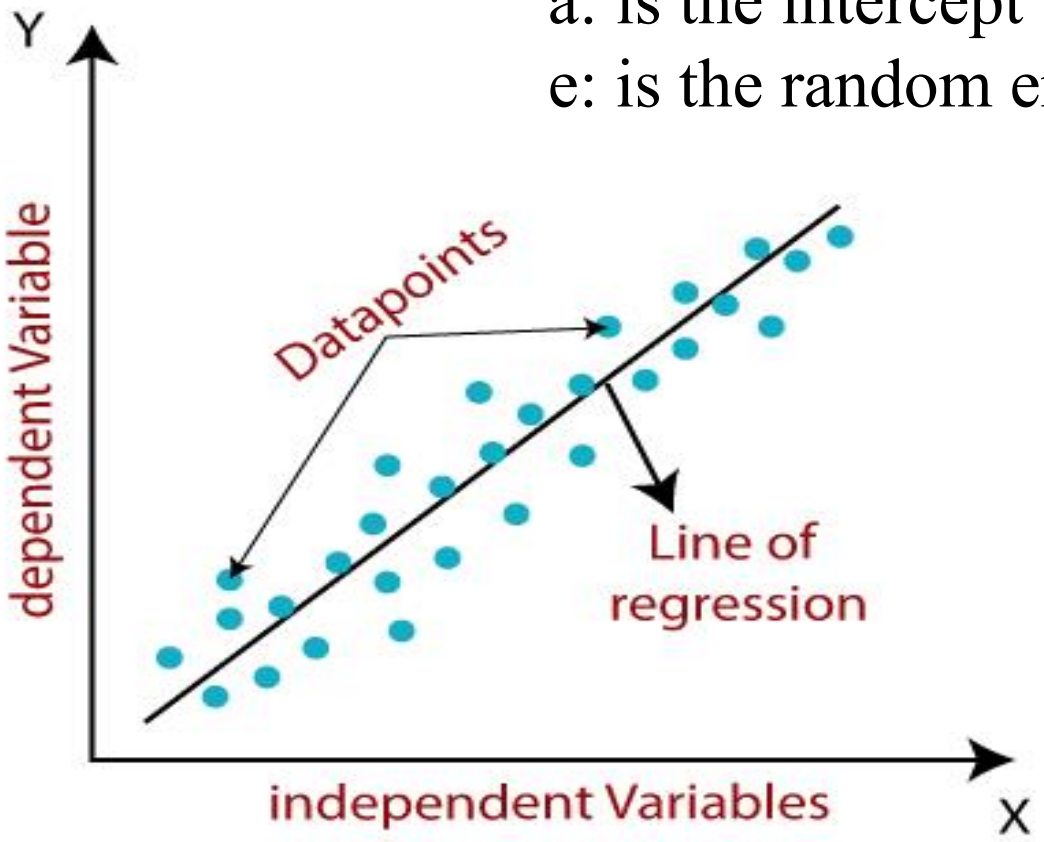
where X: explanatory variable

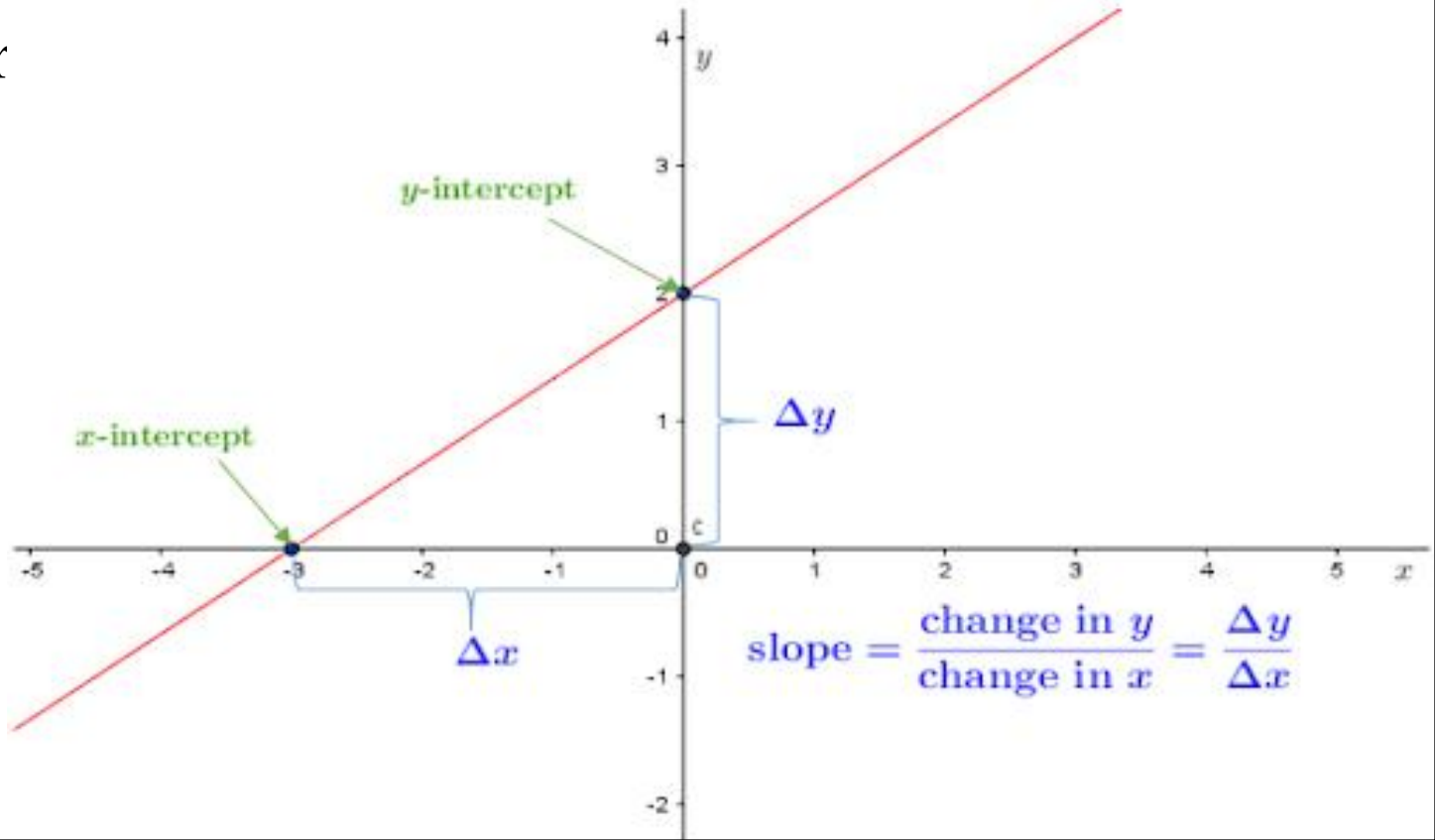
Y: is the dependent variable.

b: regression coefficient or slope of the line

a: is the intercept

e: is the random err

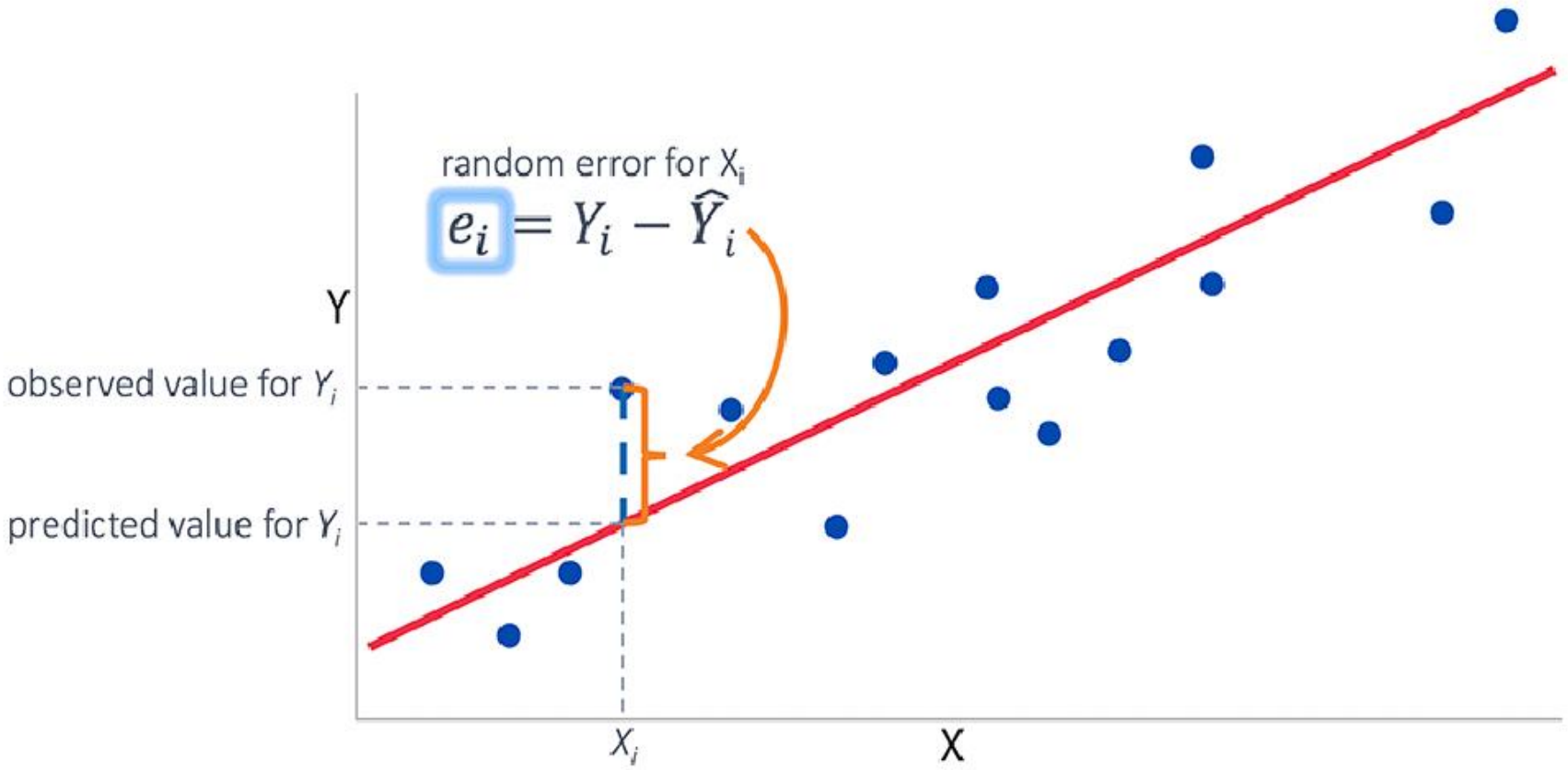




Linear Regression cont…

The random error in the following linear equation of line:

$\hat{Y}_i$ is predicted values of Y_i



Linear Regression cont…

- The calculation of b and a is as follows:

$$b = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad a = \frac{\Sigma Y}{n} - b\,\frac{\Sigma X}{n}$$

- If b > 0, then x(predictor) and y(target) have a positive relationship. That is increase in x will increase y.
- If b < 0, then x(predictor) and y(target) have a negative relationship. That is increase in x will decrease y.
- Error e can be determined using any of the following method:
  - Mean Square Error (MSE)
  - Root Mean Square Error (RMSE)
  - Mean Absolute Percentage Error (MAPE)

Linear Regression cont…

Company	Sales in 1000s (Y)	Number of agents in 100s (X)
A	25	8
B	35	12
C	29	11
D	24	5
E	38	14
F	12	3
G	18	6
H	27	8
I	17	4
J	30	9

$$b = \frac{10 \times 2289 - (80 \times 255)}{[10 \times 756 - (80)^2]} = 2.1466;$$

$$a = \frac{255}{10} - 2.1466\frac{80}{10} = 8.3272$$

Linear Regression cont…

- The linear regression will thus be Predicted **Y = 2.1466 X + 8.3272**
- The above equation can be used to predict the volume of sales for an insurance company given its agent number. Thus if a company has 1000 agents (10 hundreds) the predicted value of sales will be around ?
- **In summary, linear regression consists of the following steps:**
  - Collection of sample of independent and dependent variable.
  - Compute b (slope) and a(intercept).
  - Use these values to formulate the linear regression equation.
  - Given the new values for X predict the value of Y.
- Larger and better the sample of data, more accurate would be the regression model and would lead to more accurate forecasts.

Performance Evaluation of Regression: Regression Metrics

- What makes a good regression? Of course, a good regression is an accurate regression.
- A regression "error" is the difference between an observed value and its predicted value.

  The "error" does not mean a mistake, **it means the unpredictable part of an observation.**
- Error measure plays an important role in calibrating and refining regression model and helps the analyst to improve forecasting method.

- The popular and highly recommended error measures are
  A. Mean Square Error (MSE)
  B. Root Mean Square Error (RMSE)
  C. Mean Absolute Percentage Error (MAPE)

A. Mean Square Error (MSE)

MSE is defined as mean or average of the square of the difference between actual and estimated values. Mathematically it is represented as:

$$\text{MSE} = \frac{\sum_{j=1}^{N} (\text{observation } (j) - \text{prediction } (j))^2}{N}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Value	42	45	49	55	57	60	62	58	54	50	44	40
Predicted Value	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56 and **MSE = 56 / 12 = 4.6667**

B. Root Mean Square Error (RMSE)

It is just the square root of the mean square error. Mathematically it is represented as:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{j=1}^{N}(\text{observation }(j)-\text{prediction }(j))^2}{N}}$$

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Value	42	45	49	55	57	60	62	58	54	50	44	40
Predicted Value	44	46	48	50	55	60	64	60	53	48	42	38
Error	-2	-1	1	5	2	0	-2	-2	1	2	2	2
Squared Error	4	1	1	25	4	0	4	4	1	4	4	4

Sum of Square Error = 56, MSE = 56 / 12 = 4.6667, **RMSE = SQRT(4.667) = 2.2**

C. Mean Absolute Percentage Error (MAPE)

The formula to calculate MAPE is as follows:

$$\text{MAPE} = (100/n) \times \sum_{i=1}^{n} \frac{(|X'(t) - X(t)|}{X(t)}$$

Here, X'(t) represents the predicted data value of point t and X(t) represents the actual data value of point t. Calculate MAPE for the below dataset.

Month	1	2	3	4	5	6	7	8	9	10	11	12
Actual Value	42	45	49	55	57	60	62	58	54	50	44	40
Predicted Value	44	46	48	50	55	60	64	60	53	48	42	38

- MAPE is commonly used because it's easy to interpret and easy to explain. For example, a MAPE value of 11.5% means that the **average difference between the predicted value and the actual value is 11.5%.**
- The lower the value for MAPE, the better a model is able to predict values e.g. a model with a **MAPE of 2% is more accurate** than a model with a MAPE of 10%.