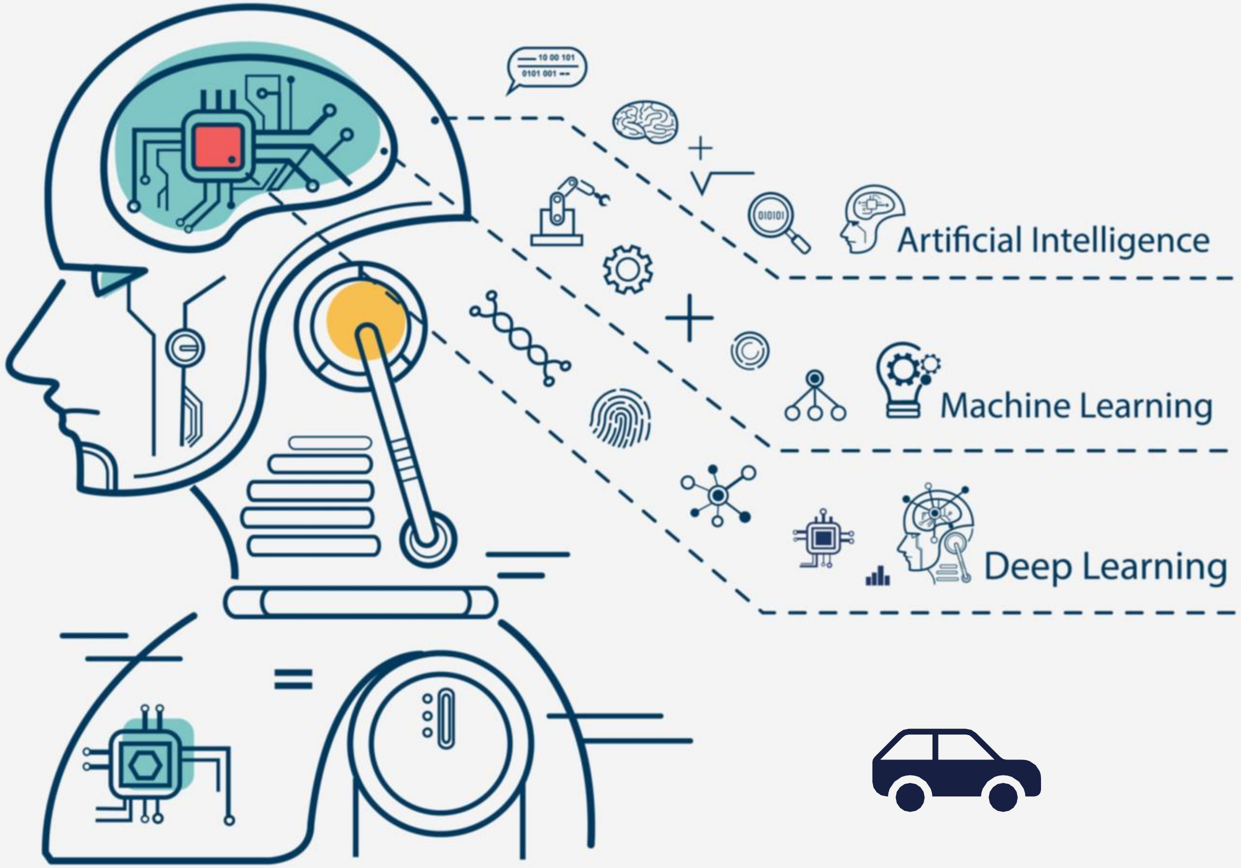
# APSSDC

## Andhra Pradesh State Skill Development Corporation

Skill AP
APSSDC




# MACHINE LEARNING USING PYTHON

# DAY2 AGENDA

Linear Regression with One variable

Linear Regression with Multiple Variables

Evaluation Metrics in Regression Models

Train/Test splitting of data

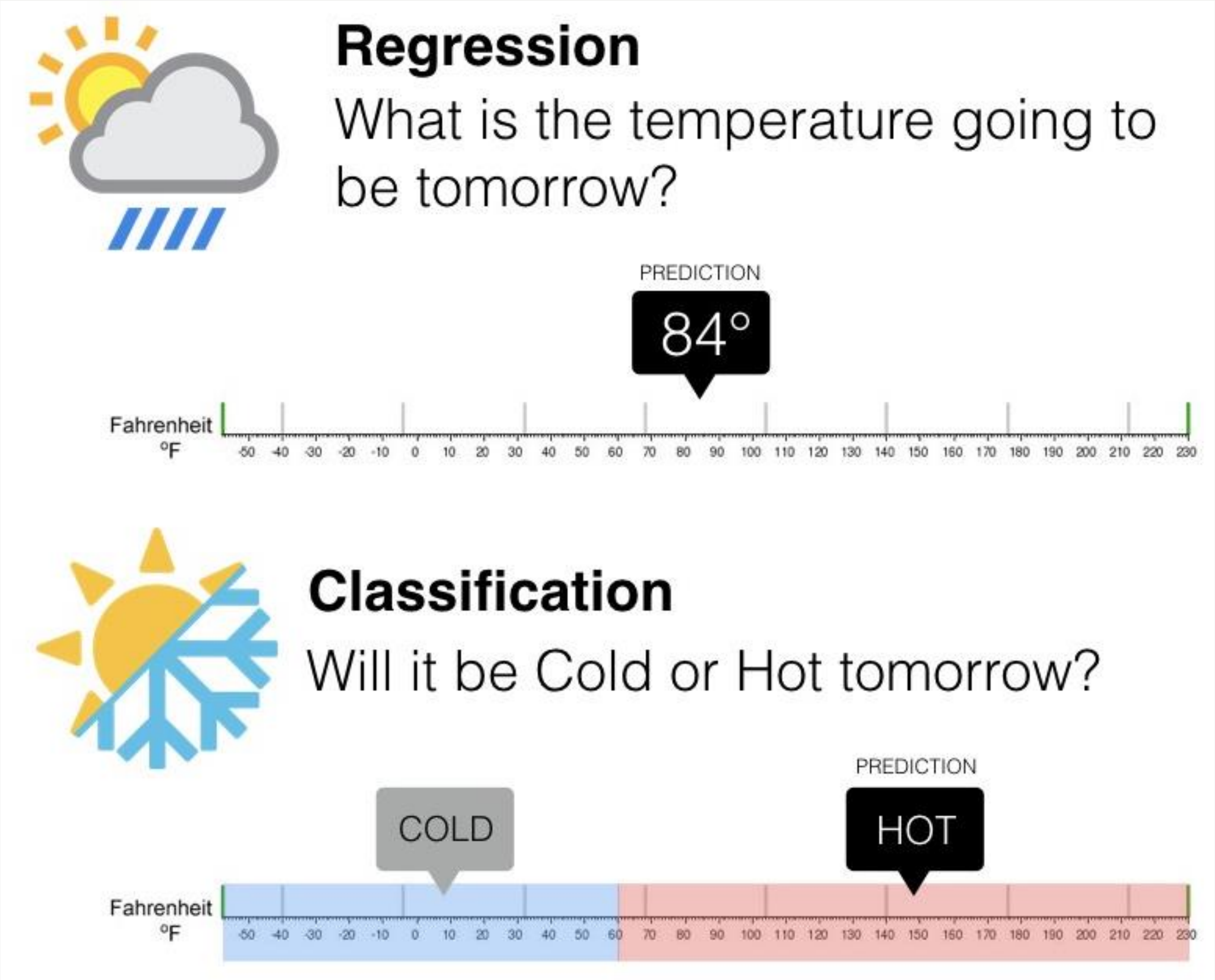
Cross Validation

- Regression
- Classification

# ML MODEL DEVELOPMENT LIFE CYCLE

1. Define Business Use Case
2. Data Exploration
3. Select Algorithm
4. Build ML Model
5. Evaluate

# REGRESSION VS CLASSIFICATION




# Linear Regression in Machine Learning

Skill AP
APSSDC

# What is Regression?

- Function: a mathematical relationship enabling us to predict what values of one variable (Y) correspond to given values of another variable (X).
- Y: is referred to as the ***dependent variable***, **the *response variable* or the *predicted variable*.**
- X: is referred to as **the *independent variable***, the *explanatory variable* or the ***predictor variable***.

Thus Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent and independent variable.

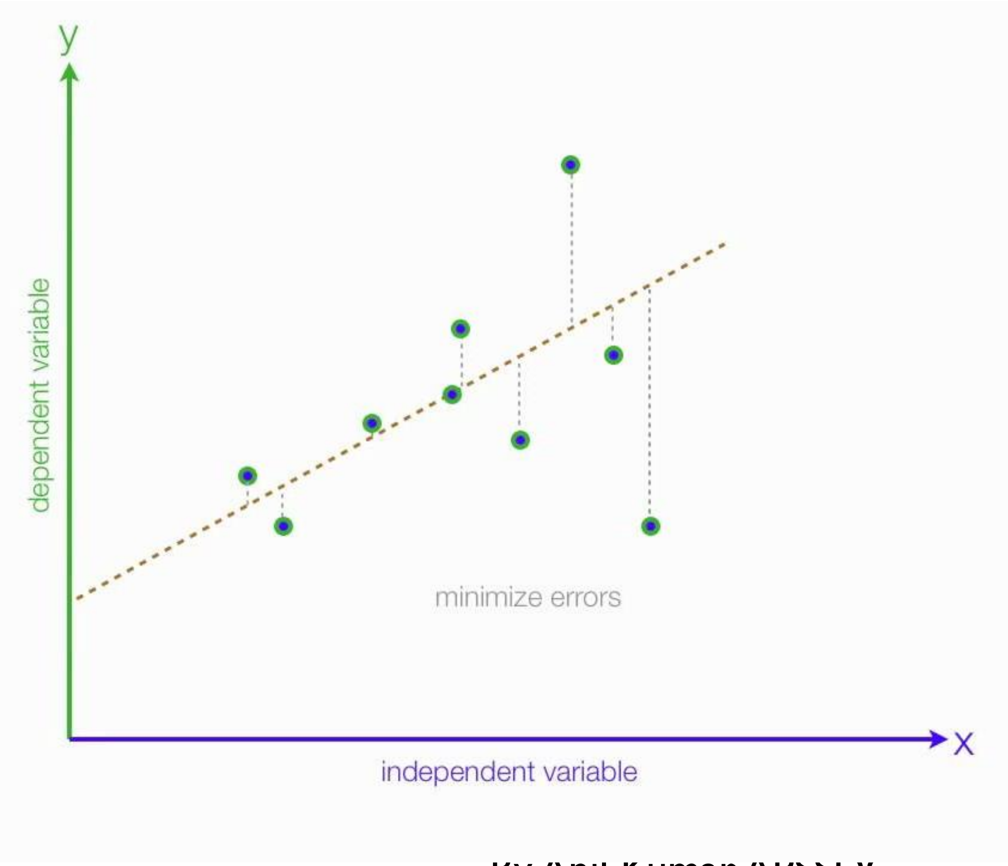
# REGRESSION

- **Linear Regression**
  - **Linear Regression with one variable**
  - **Linear Regression with multiple variable**
- Non-Linear Regression/Polynomial Regression
  - Non-Linear Regression with one variable
  - Non-Linear Regression with multiple variables
- SGD
- Ridge
- Lasso
- Elastic Net

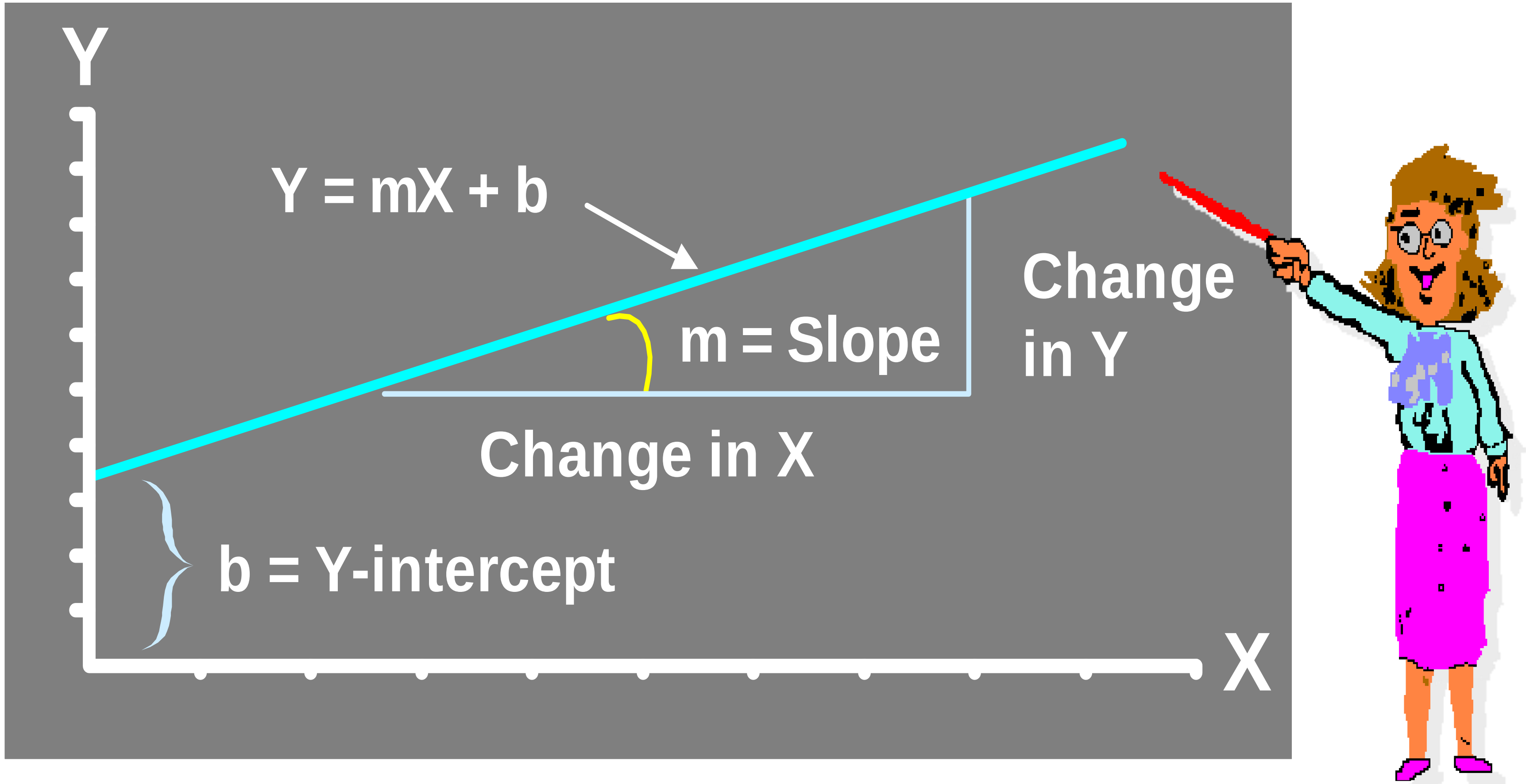
# Contd.

A typical Linear Regression model can be represented in the form :

- $y = b1x + b0$ where b1 is slope and b0 is the intercept.



# Linear Equations




By Anil Kumar APSSDC

# Linear Regression Model

**Relationship Between Variables Is a Linear Function**

Y-Intercept

Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable

Independent (Explanatory) Variable

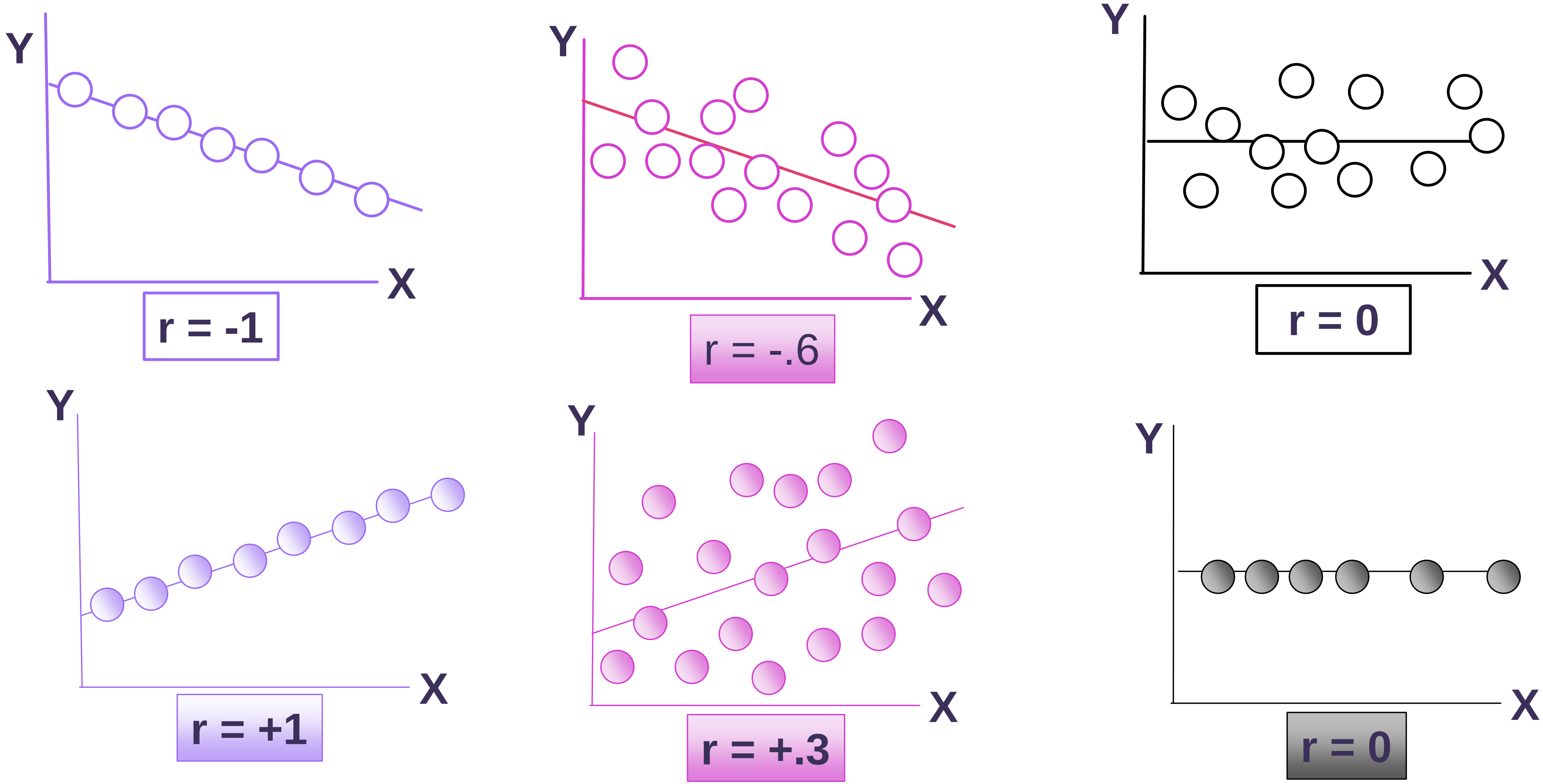
- $\beta_1 > 0 \Rightarrow$ Positive Association
- $\beta_1 < 0 \Rightarrow$ Negative Association
- $\beta_1 = 0 \Rightarrow$ No Association

By Anil Kumar APSSDC

# Correlation

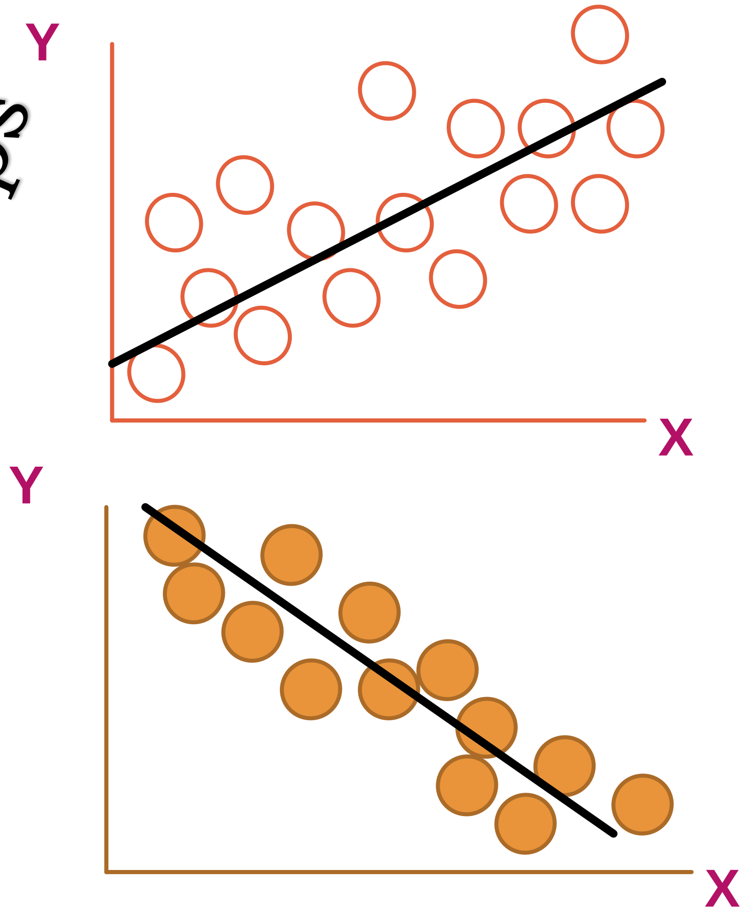
- Measures the relative strength of the *linear* relationship between two variables Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

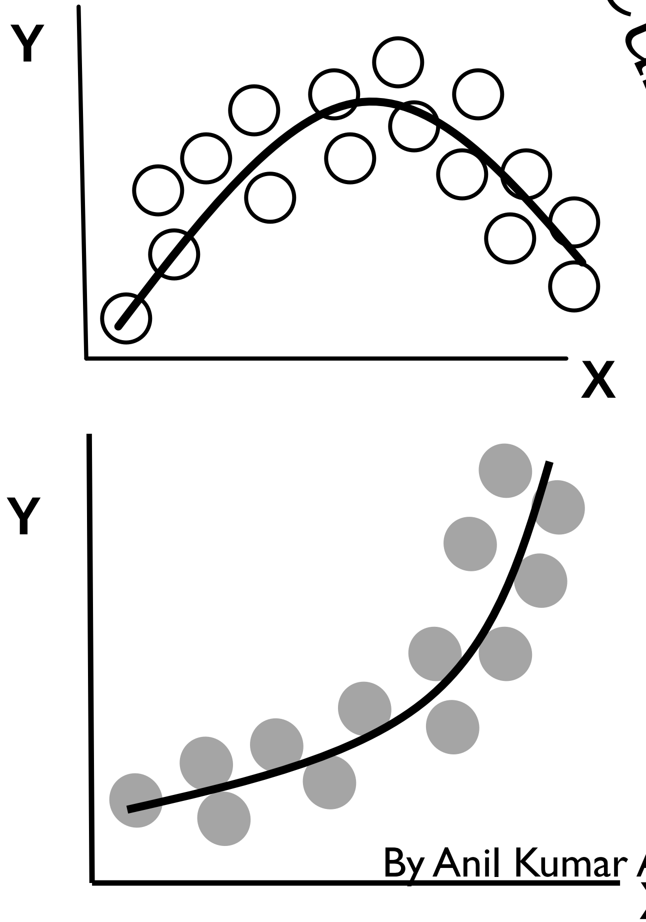
# Scatter Plots of Data with Various Correlation Coefficients



# Linear Correlation

Linear relationships



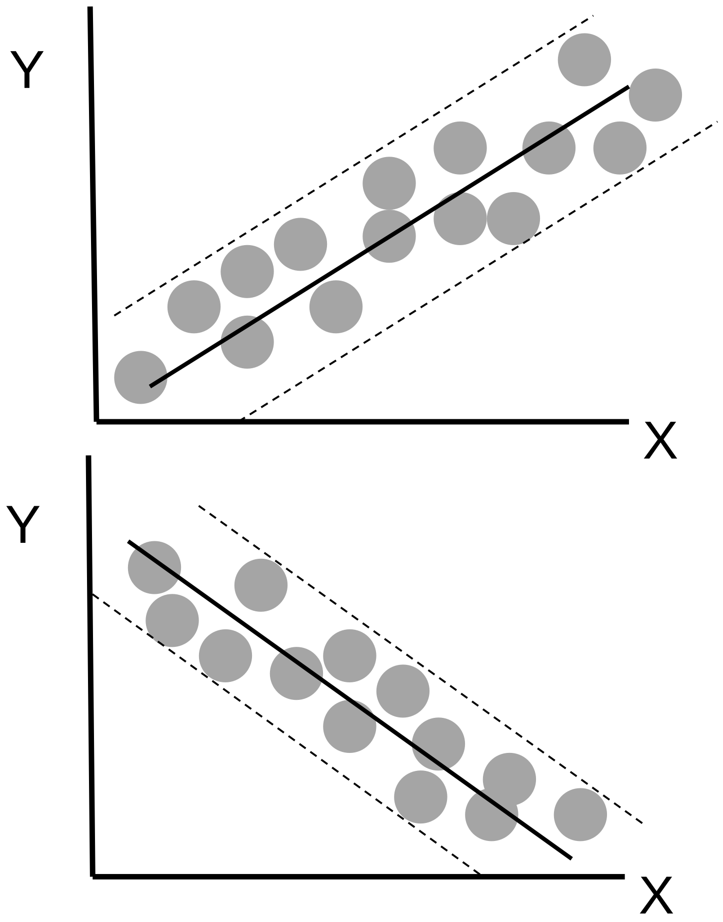




Curvilinear relationships

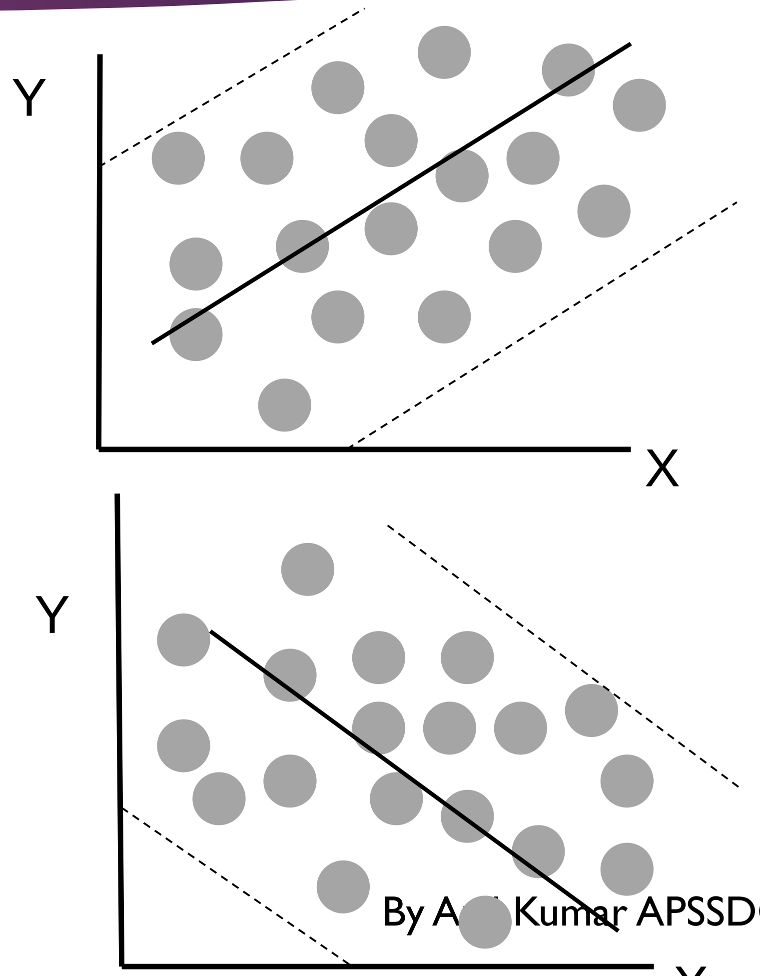
By Anil Kumar APSSDC

# Linear Correlation

Strong relationships

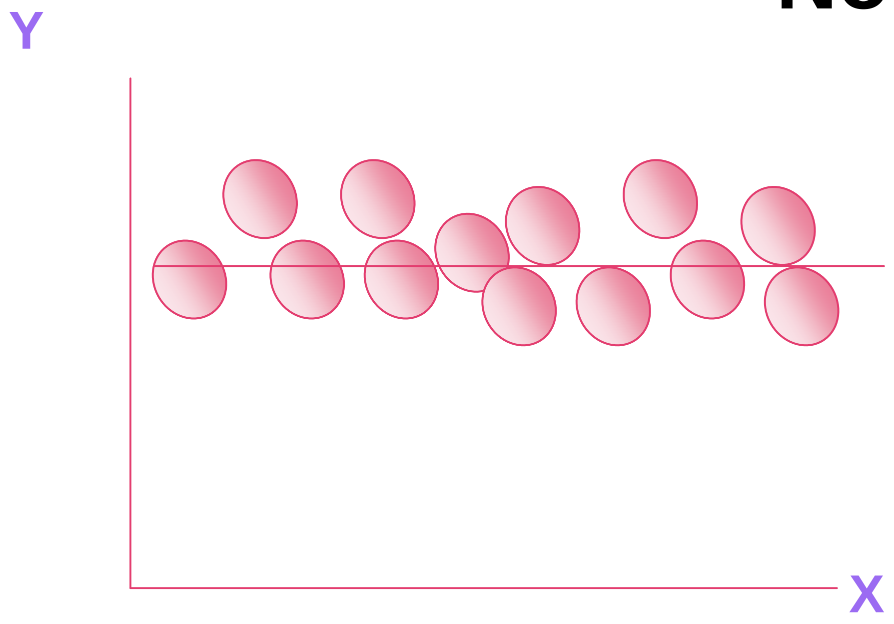


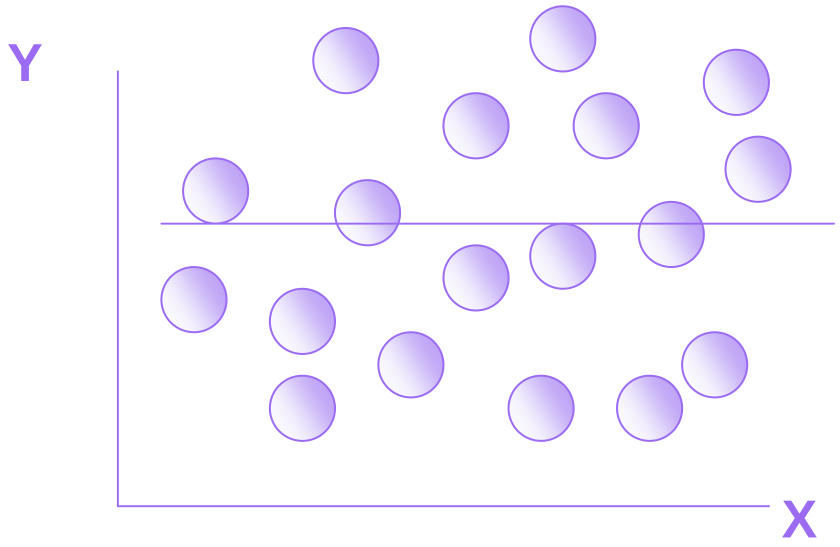

Weak relationships



# Linear Correlation

## **No relationship**



# Steps in Regression Analysis

- Examine the scatterplot of the data.
  I. Does the relationship look linear?
  II. Are there points in locations they shouldn't be?
  III. Do we need a transformation?
- Assuming a linear function looks appropriate, estimate the regression parameters.
  I. How do we do this? (Method of Least Squares)
- If there is a significant linear relationship, estimate the response, Y, for the given values of X, and compute the residuals

# Regression Analysis

- Thus we have the regression formula as :

  Y = MX + C + error(e).

Initially we calculate the value for slope and predict the values of Y for any given X values we have.

$$\text{Slope(M)} = \sum_{i=0}^{len(X)} \frac{(X_i - X_{mean}) * (Y_i - Y_{mean})}{(X_i - X_{mean})^2}$$

Thus we calculate the C value and find out the "Line of Regression".

# Regression Analysis

- Next our job is to reduce the distance between the actual value and the predicted value or in other words reduce the error between the actual and predicted value. Thus the line with least error will be the **"Best Fit Line"**.
- In order to check it out we calculate the "Coefficient of Determination".

$$Mean\ Squared\ value\ (R^2) = \sum_{i=0}^{len(X)} \frac{(Y_{pred} - Y_{mean})^2}{(Y - Y_{mean})^2}$$

- Thus our ultimate aim is to reduce the error i.e. distance between the actual and predicted values.

# Evaluation Metrics

## 1. Mean Absolute Error

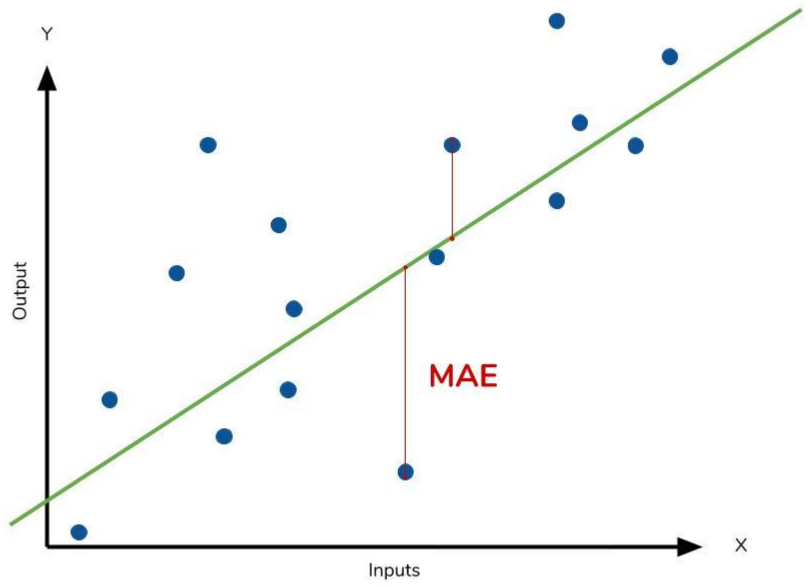
Divide by the total number of data points

Predicted output value

Actual output value

$$MAE = \boxed{\frac{1}{n}} \sum \underbrace{\left| \boxed{y} - \boxed{\hat{y}} \right|}_{\text{The absolute value of the residual}}$$

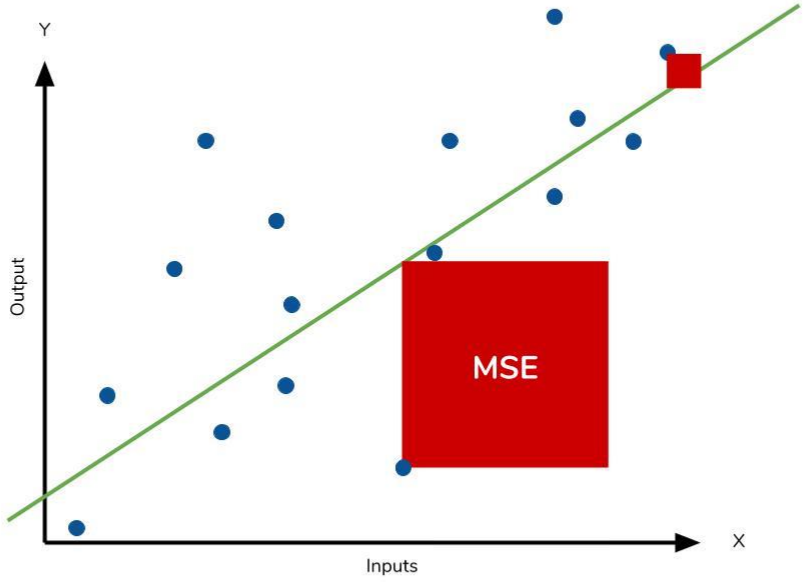
Sum of



# Contd..

## 2. Mean Square Error

$$MSE = \frac{1}{n} \Sigma \underbrace{\left( y - \widehat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$



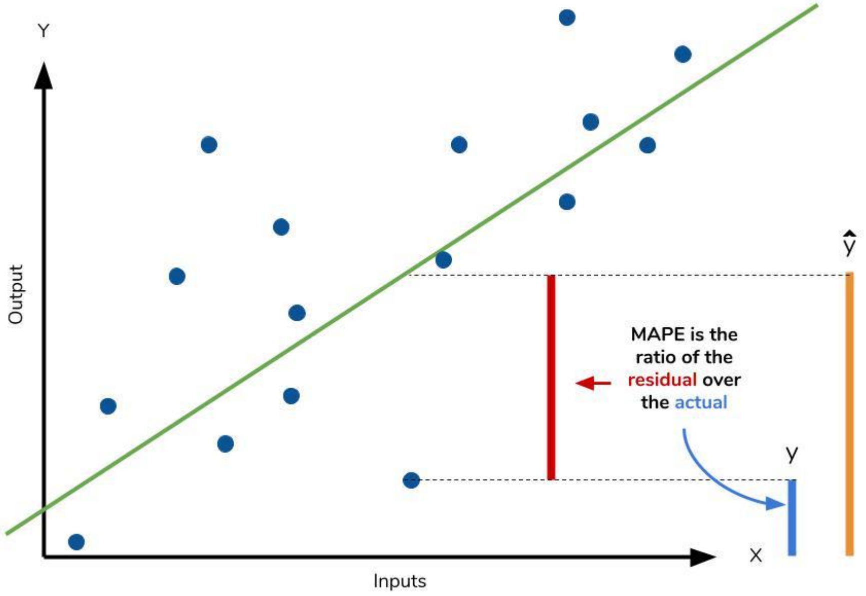
# Contd..

## 3. Mean Absolute Percentage Error

Multiplying by 100% converts to percentage

$$MAPE = \frac{\boxed{100\%}}{n}\sum\left|\frac{\overbrace{y-\hat{y}}^{\text{The residual}}}{\boxed{y}}\right|$$

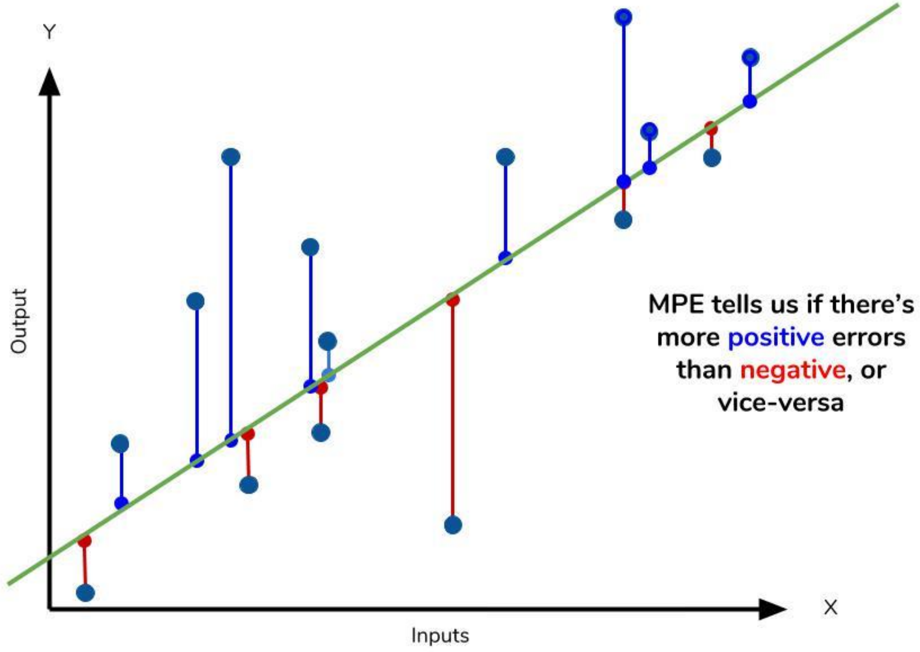
Each residual is scaled against the actual value



# Contd..

### 3. Mean Percentage Error

$$MPE = \frac{100\%}{n} \Sigma \left( \frac{y - \hat{y}}{y} \right)$$



# Conclusion

| Acroynm | Full Name | Residual Operation? | Robust To Outliers? |
|---|---|---|---|
| MAE | Mean Absolute Error | Absolute Value | Yes |
| MSE | Mean Squared Error | Square | No |
| RMSE | Root Mean Squared Error | Square | No |
| MAPE | Mean Absolute Percentage Error | Absolute Value | Yes |
| MPE | Mean Percentage Error | N/A | Yes |

By Anil Kumar APSSDC

https://www.dataquest.io/blog/understanding-regression-error-metrics/