Statistical Learning Introduction

Statistics

- ✔ Statistics is a mathematically-based field which seeks to collect and interpret data and makes smart decisions.

- ✔ Statistics is a crucial process behind how we make discoveries in science, make decisions based on data.

Types of statistics and parameters involved in statistics

Two types : 1. Descriptive
2. Inferential

- **Descriptive statistics :** if data can be described without any statistical tools then it is called descriptive statistics . ex, marks in class , height of student.
- **Inferential statistics:** if data is too big then then we use inferential statistics.
  - We take a few samples from different data and we find the average. This is called inferential statistics. The average is then applicable to all the data from where we have selected our sample

How Industry Use Statistics

1.Weather forecasting
2.Giving Insurance
3.Stock Market
4.Drug Effectiveness before releasing to the market
5.Diseased survival probability
6.Election winning and exit poll prediction
7.Loan approval and fraud detection
8.Netflix/Amazon recommendation
9.New Campaign Effectiveness

Analytics Methodology using statistics

- **Mean**= Average
- **Median**= Centre Data
  - Centre data if the sample is in an odd number.
  - If the sample is even then we add both the middle value and divide by 2.
- **Mode**= Also called frequency
  - The most number of occurrences in a sample is termed as mode.
- **Standard Deviation**
  - Number of data deviated from the given mean is called standard deviation.
- **Variance**
  - It is the Square of standard deviation.

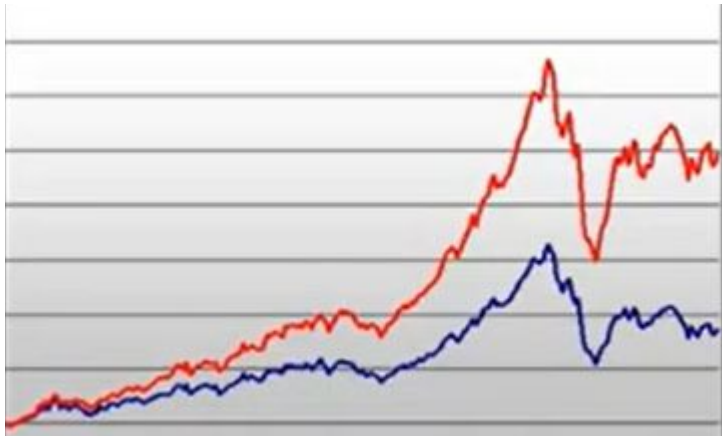
Types of Variables

- **Independent variable**
- **Dependent variables**
  - Independent variables do not depend on variables whereas dependent variables depend on various types of variables.
- **Types**
- **Continuous (numerical) Variable**
  - They include mostly numericals.
  - Numerical data are continuous and contain mostly numbers whereas discrete data can only take certain values.
- **Categorical Variables**
  - They have strings.
  - Categorical data usually have strings but they can also include numbers. There are two types of categorical data ,Ordinal and Nominal.

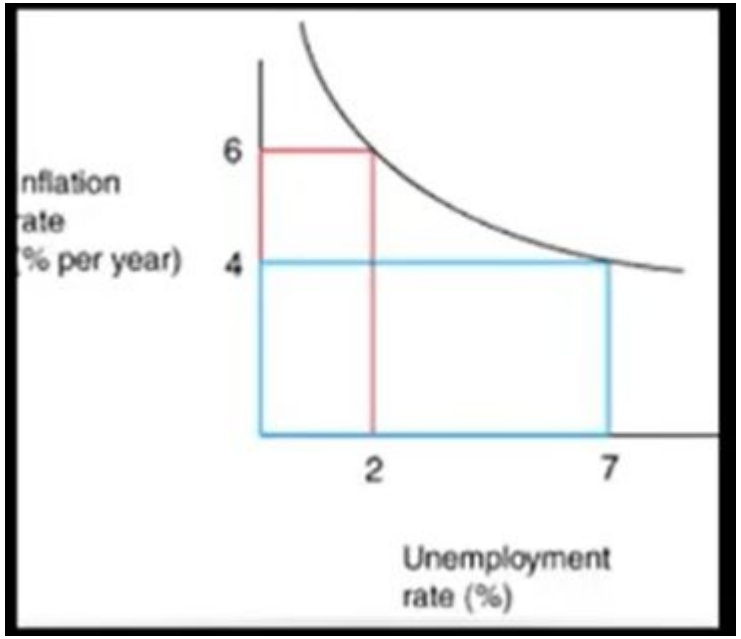
Trade off Between Prediction Accuracy and Model Interpretability

Prediction Vs. Interpretability

- Prediction



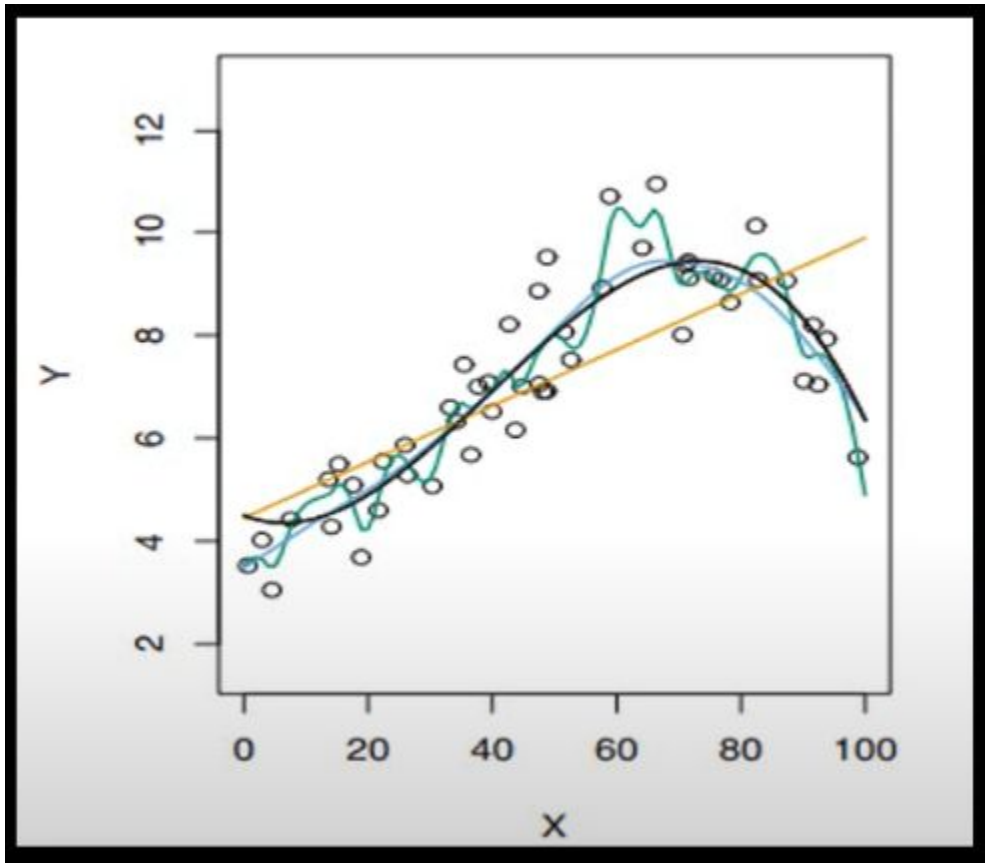
- Interpretation



Interpretation

- Linear -> Y=a+b*X
- Polynomial -> Y=a+b*X^2
- Non Linear -> Y= a+1/b*X
- Non Parametric ->

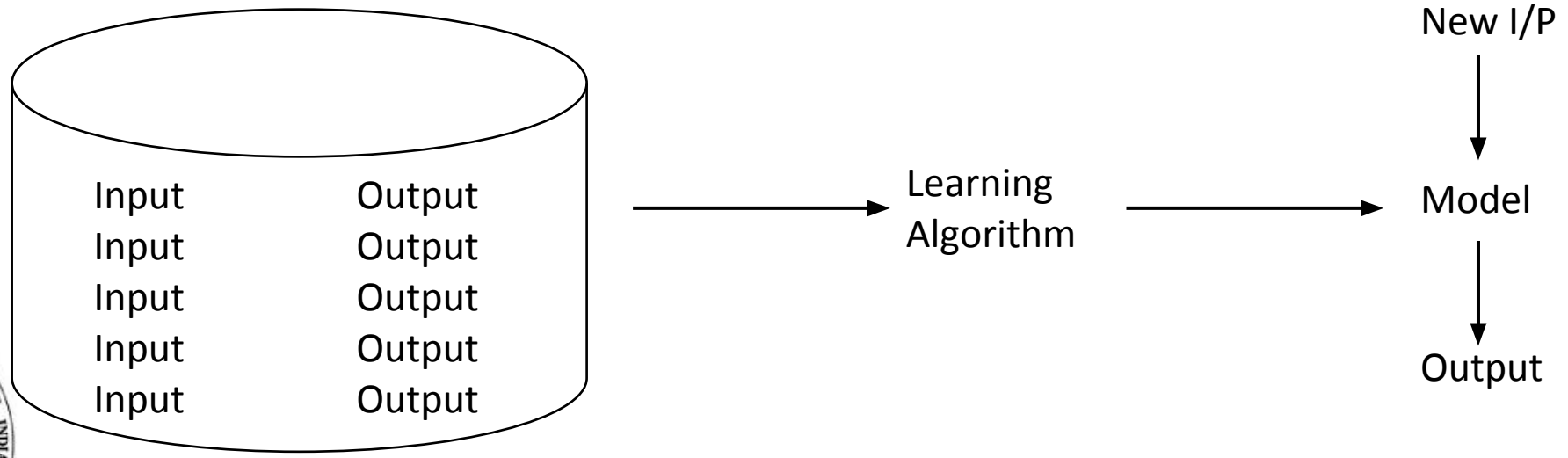
Non linear relationships



Supervised Versus unsupervised learning

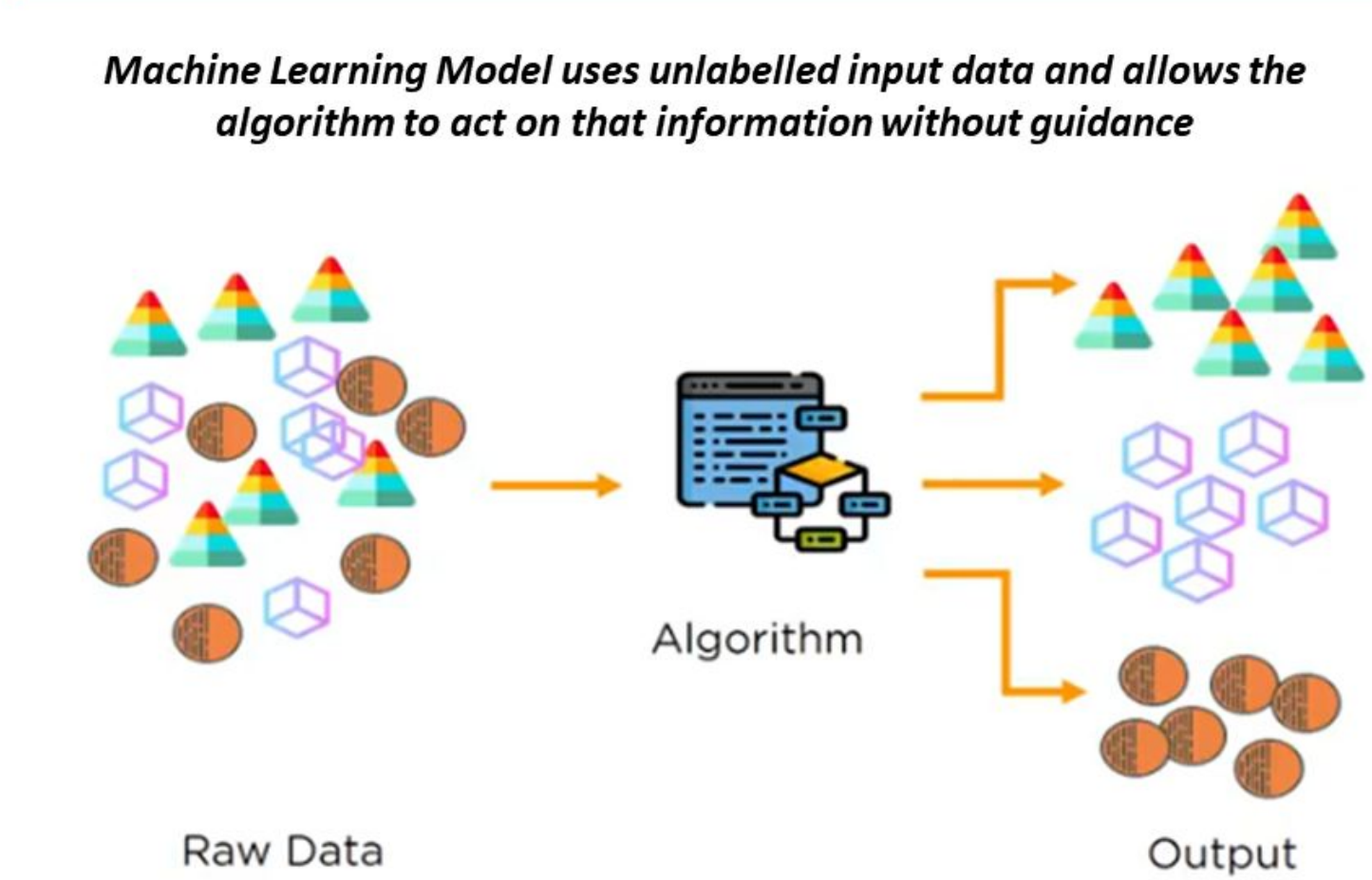
Supervised Learning

- ✔ Training Data
- ✔ Both Input and output
- ✔ Classification Problem
- ✔ Navie Bayes Algorithm



Unsupervised Learning

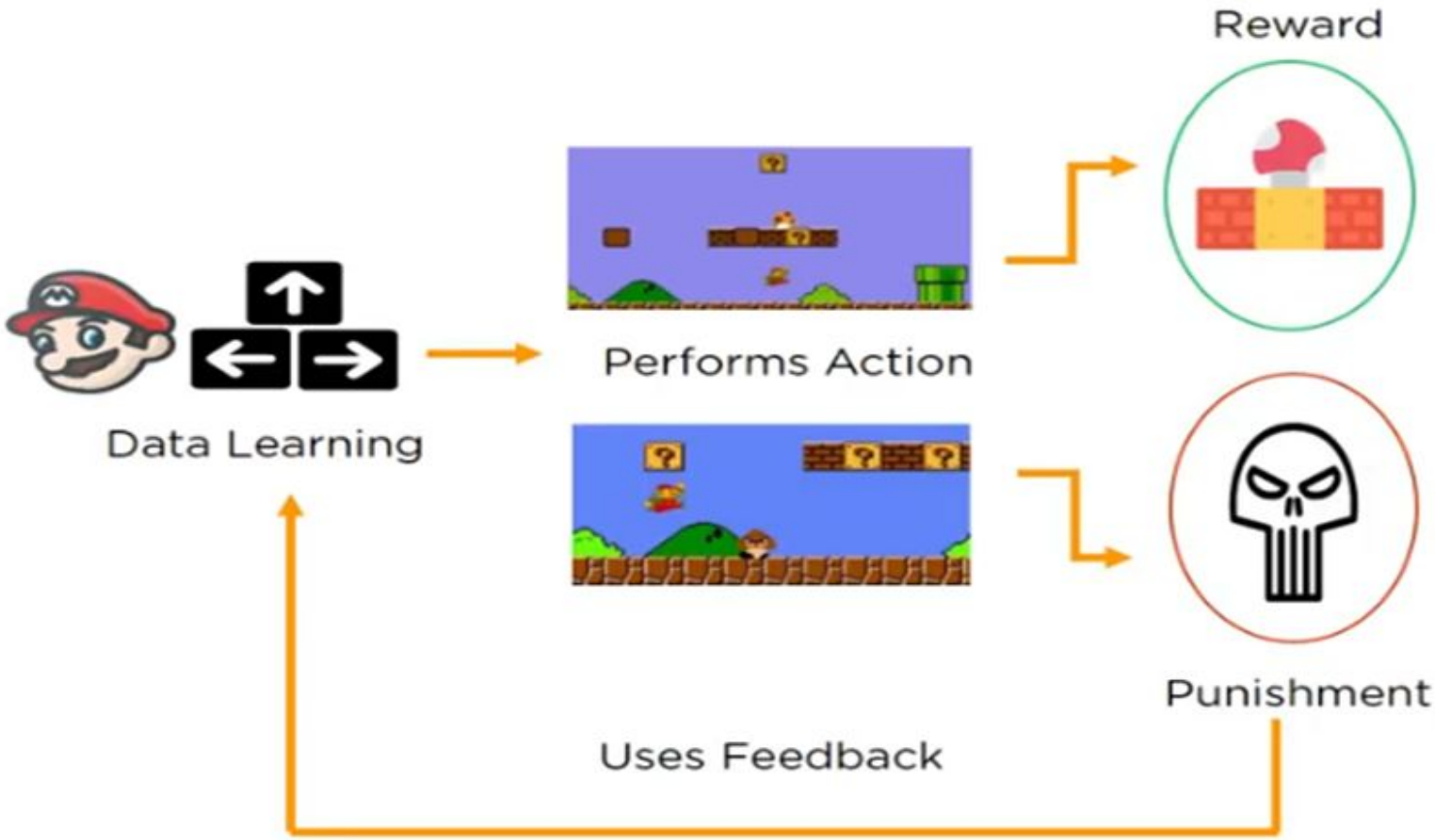
- ✔ Only Inputs
- ✔ Clustering Problem
- ✔ K-Means



Reinforcement Learning

- ✔ Rewards/Penalty
- ✔ Q-Learning

Reinforcement Learning involves teaching the machine to think for itself based on its past action reward.



Regression vs Classification

Regression vs Classification

✔ Classification
- Discrete Data

✔ Regression
- Continues Data

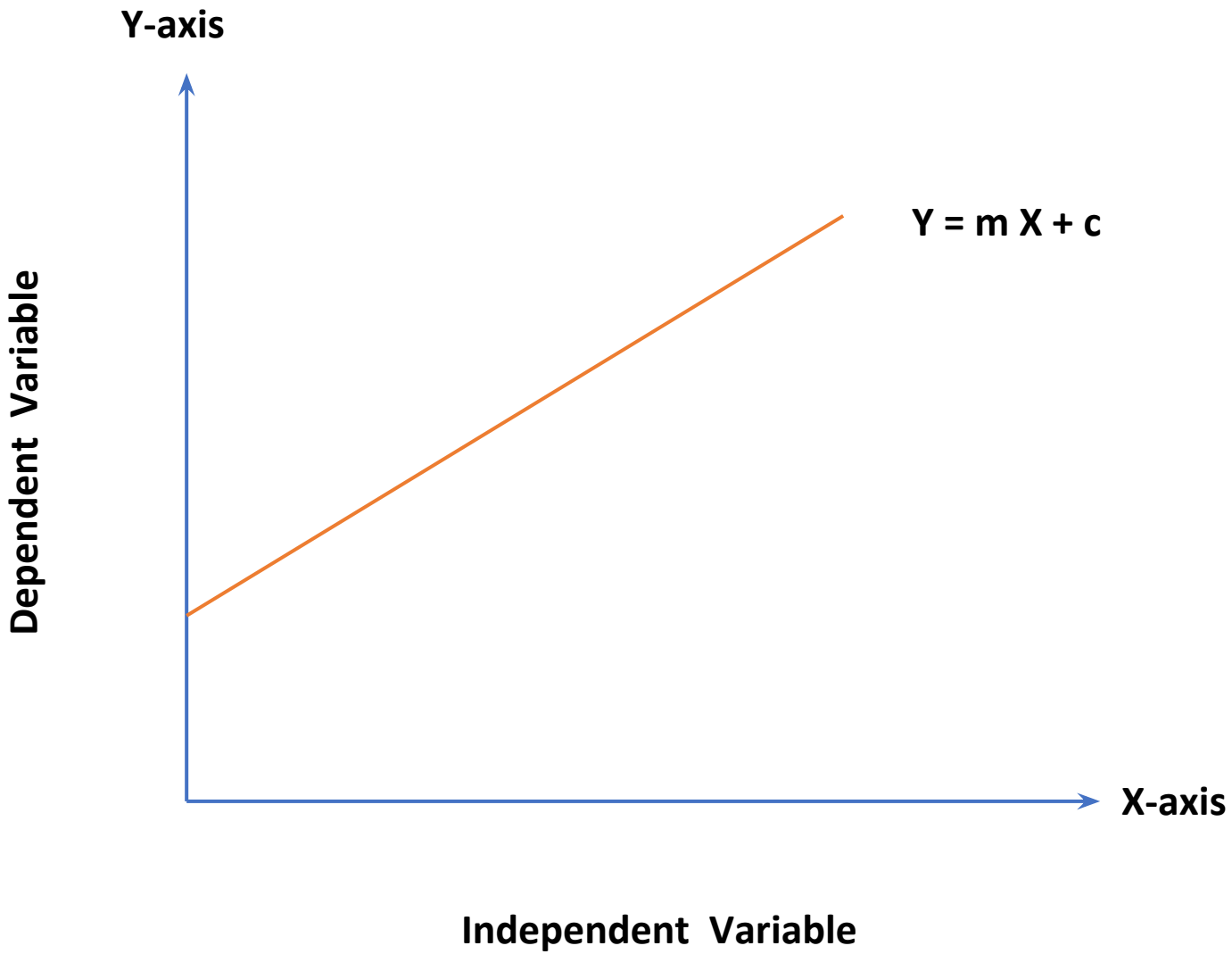
Estimating the coefficients,
Assessing the accuracy of the coefficient estimates,
Assessing the accuracy of the model

Linear Regression

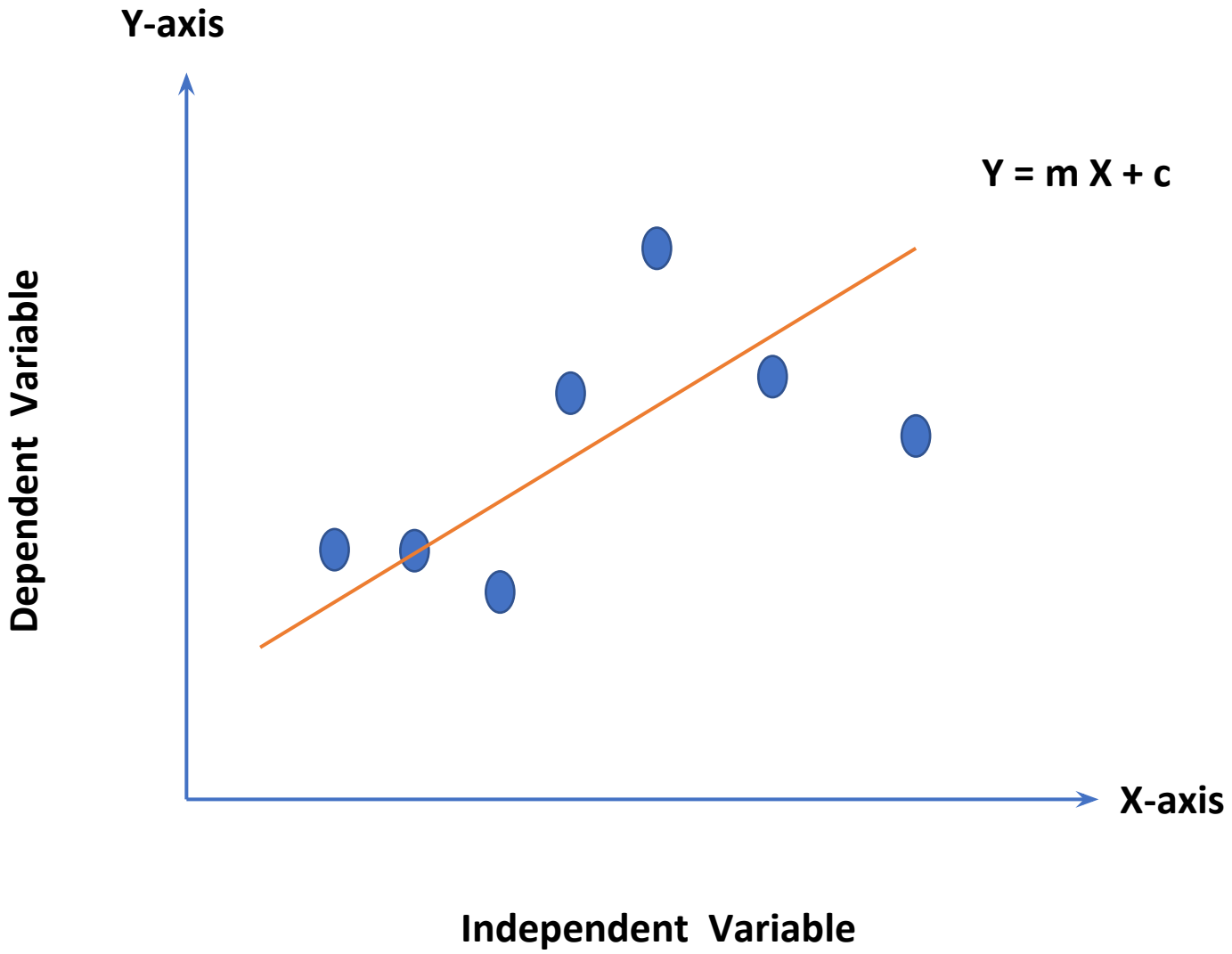
Regression Analysis is form of predictive modelling technique which investigates the relationship between a dependent and independent variables

- Determining the strength in relationship
- Forecasting an effect
- Trend forecasting

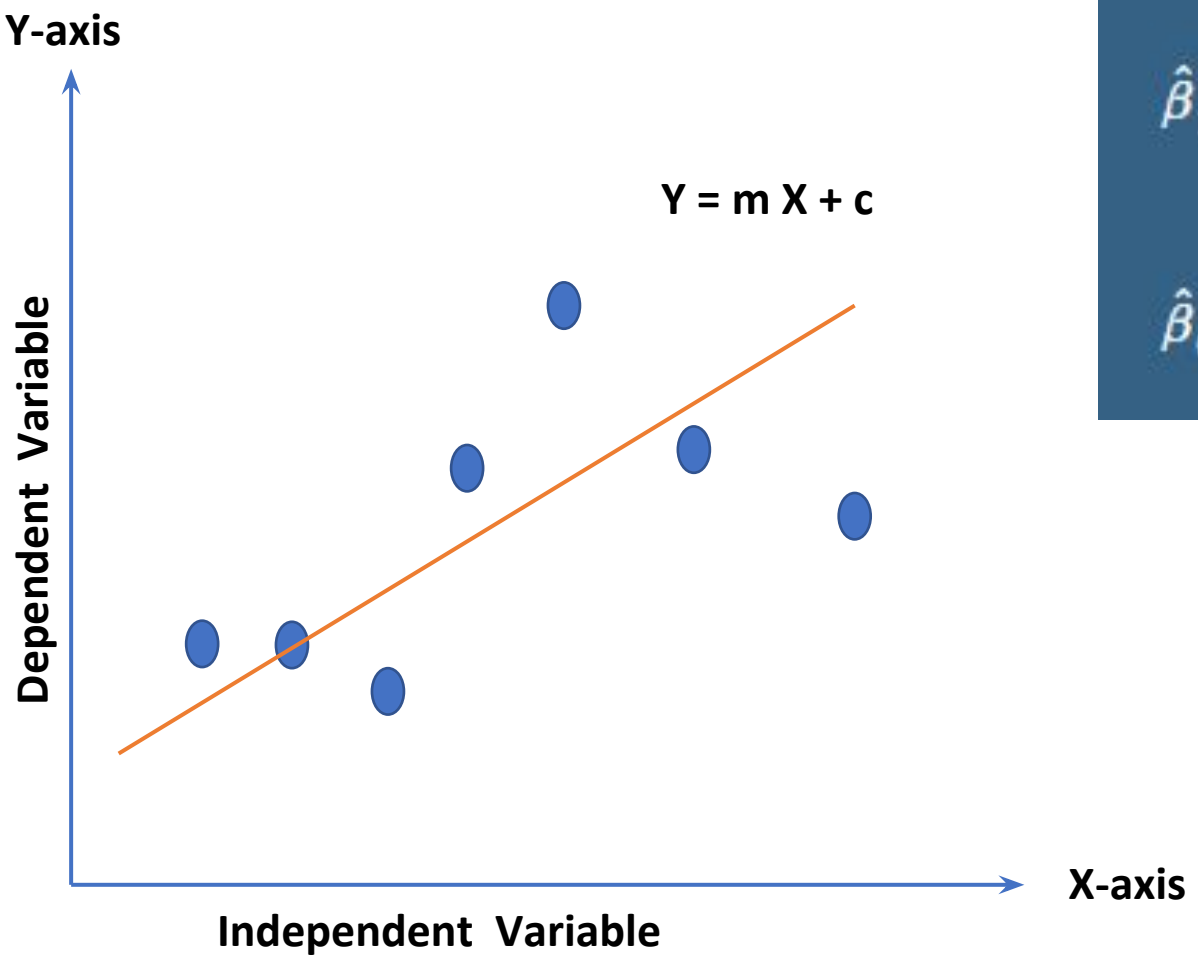
Linear Regression



Linear Regression



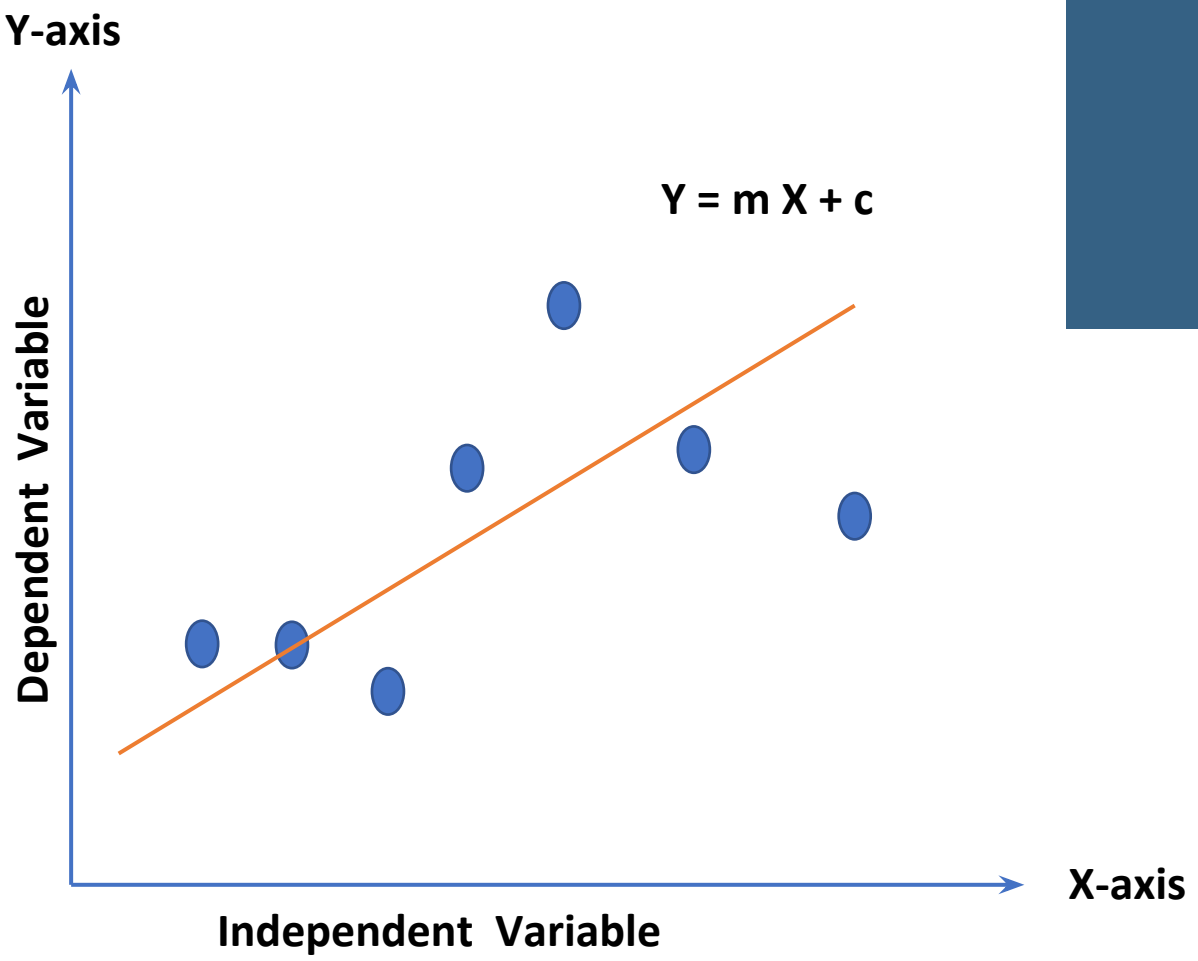
Linear Regression



$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

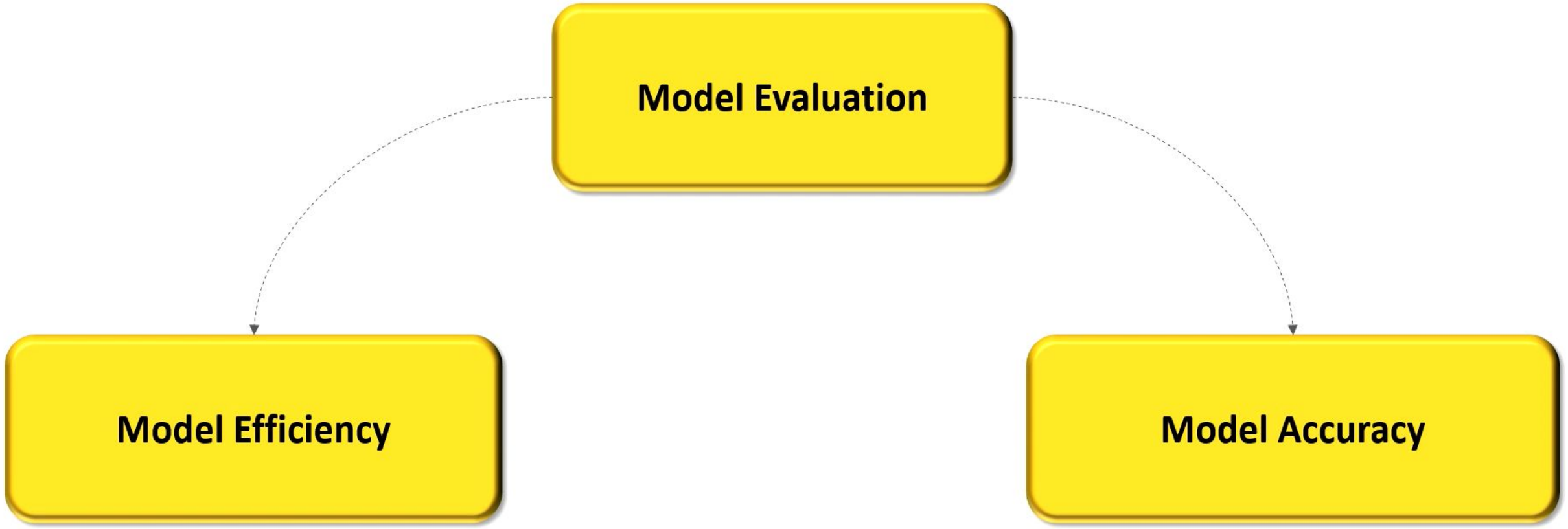
Linear Regression



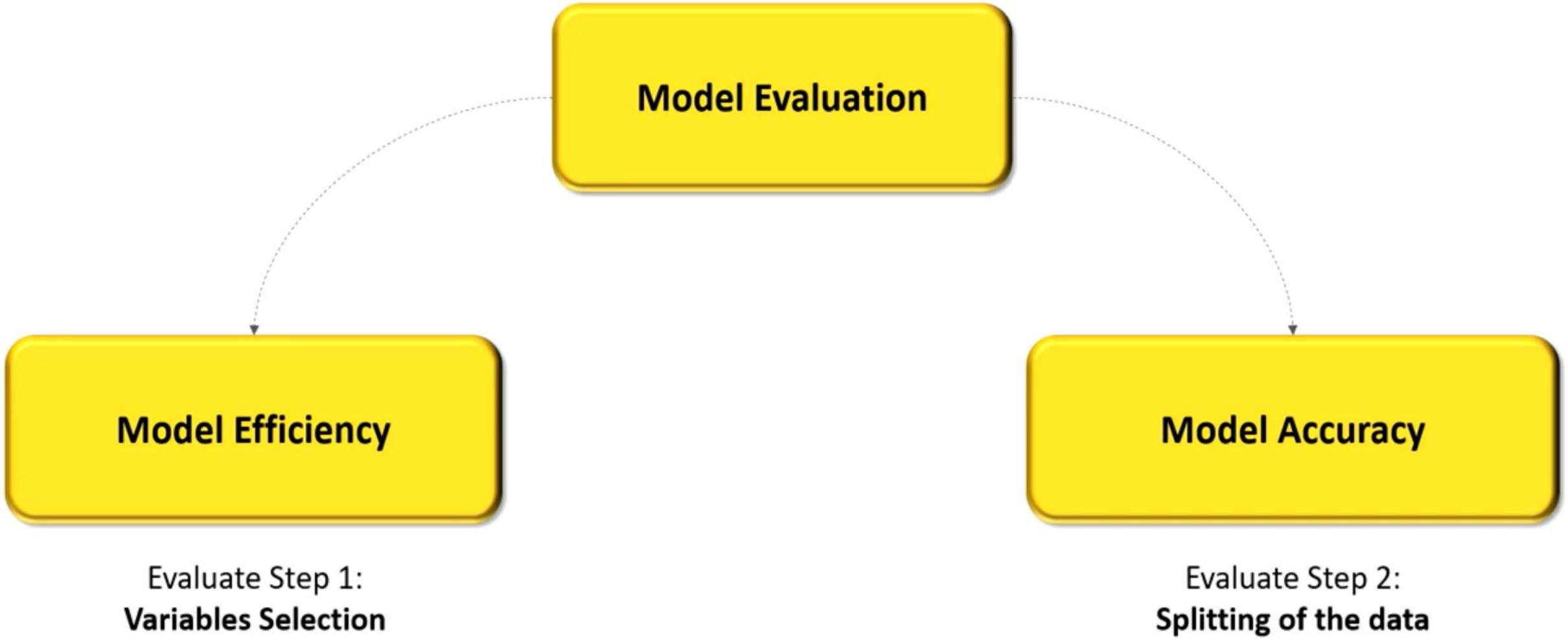
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,$$

$$\hat{u}_i = Y_i - \hat{Y}_i.$$

Model Evaluation

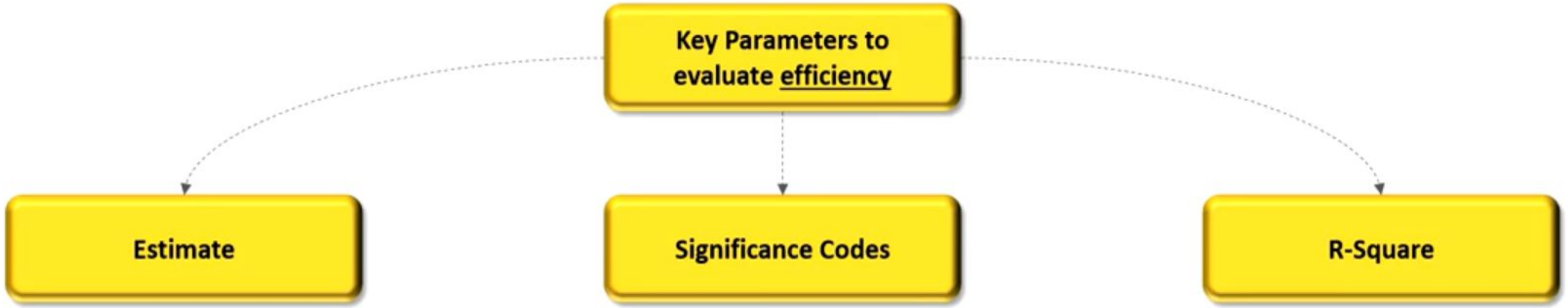


Model Evaluation



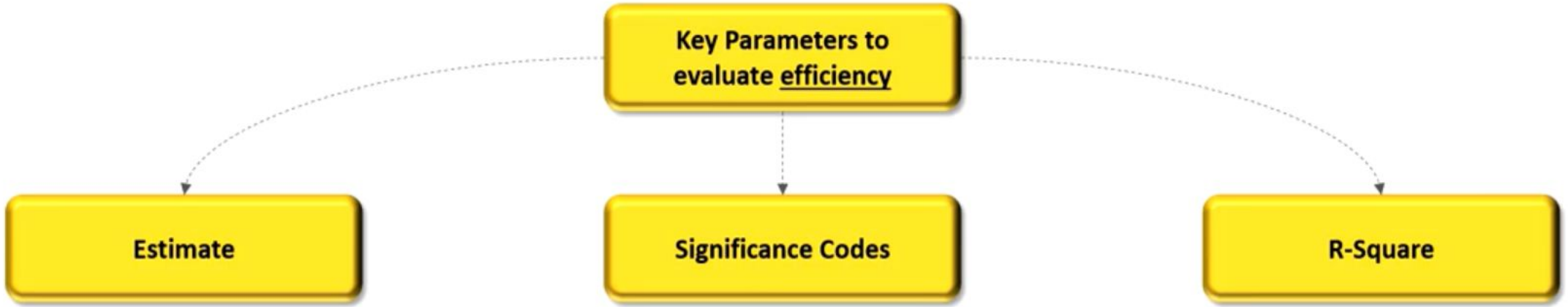
Evaluating Model Efficiency

Predicted Variable (Y)	Predictor (X1)
Height	Weight



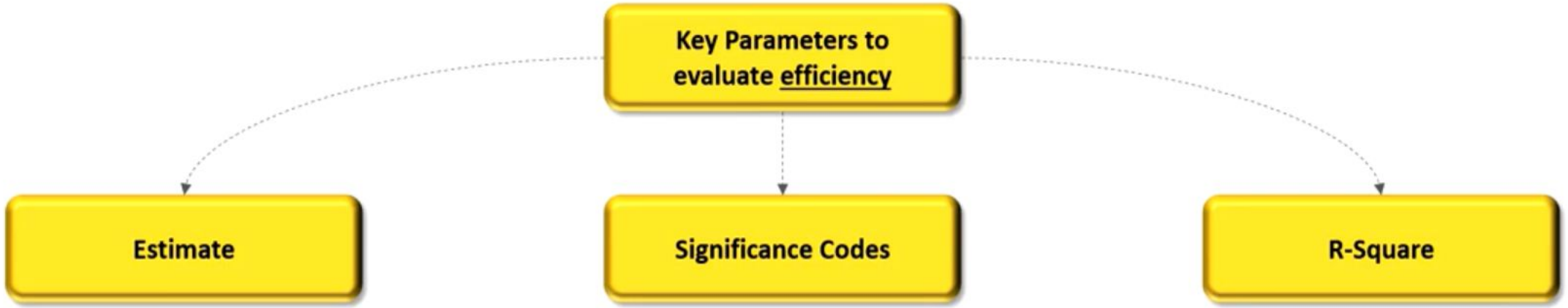
Evaluating Model Efficiency

Predicted Variable (Y)	Predictor (X1)
Height	Weight



Evaluating Model Efficiency

Predicted Variable (Y)	Predictor (X1)
Height	Weight



Evaluating Model Efficiency

Predicted Variable (Y)	Predictor (X1)	Predictor (X2)
Height	Weight	Age
	R-Square = 60%	R-Square = 80%

Key Parameters to evaluate efficiency

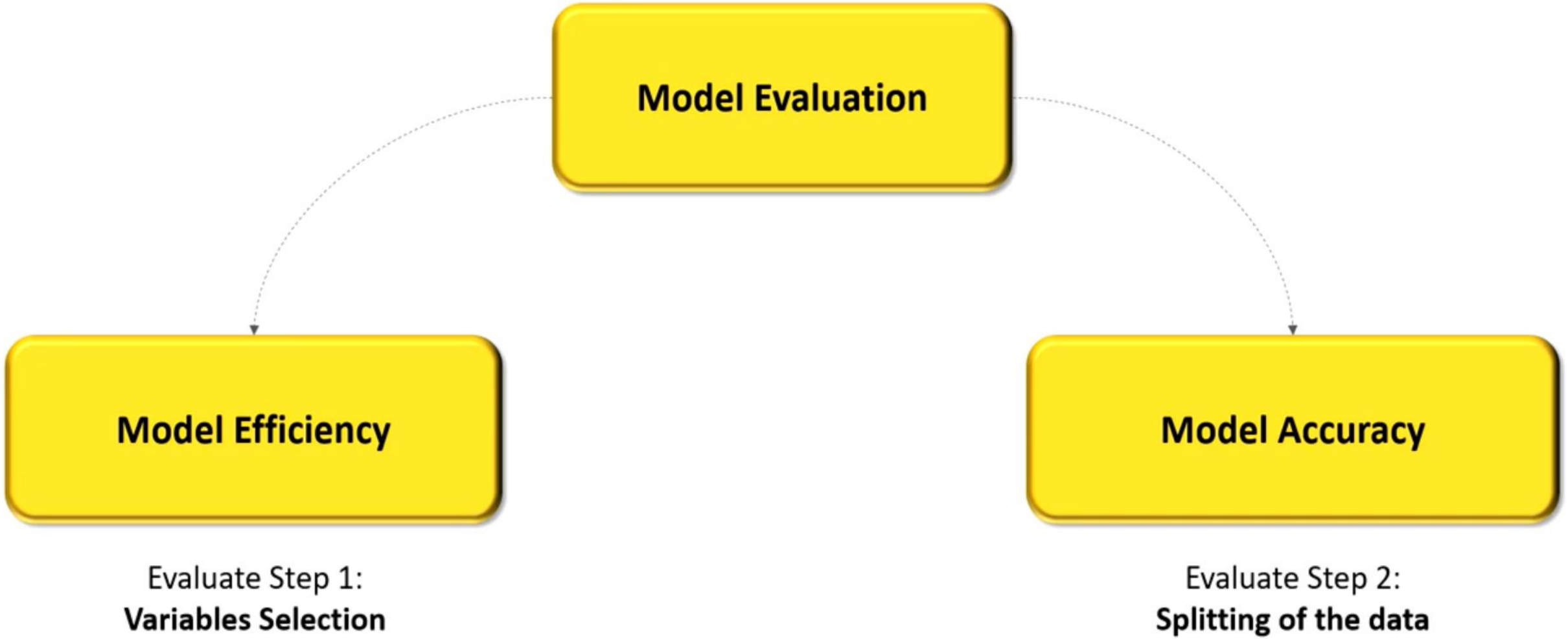
- **Estimate**
- **Significance Codes**
- **R-Square**

$$R^2 = \frac{\sum(y_p - y_m)^2}{\sum(y - y_m)^2}$$

- ★★★★★ Excellent
- ★★★★☆ Above Average
- ★★★☆☆ Average
- ★★☆☆☆ Below Average
- ★☆☆☆☆ Poor

This value helps assess how well the predictors explain the variation in the predicted variable It is a **value between 0 to 100.**

Model Evaluation



Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data			
Weight	Actual Height		
54	150		
61	170		
37	140		
32	120		
35	130		
38	140		

Explanation

Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data

Weight	Actual Height	Predicted Height	
54	150	160	
61	170	180	
37	140	170	
32	120	140	
35	130	180	
38	140	170	

Explanation

Using the model built from the **training** data to predict height for the **testing** data and comparing it to the actual height values that are already available in the testing data (unexposed to the training data).

Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data			
Weight	Actual Height	Predicted Height	Difference (Error)
54	150	160	10
61	170	180	10
37	140	170	30
32	120	140	20
35	130	180	50
38	140	170	30

Explanation

Using the model built from the **training** data to predict height for the **testing** data and comparing it to the actual height values that are already available in the testing data (unexposed to the training data).

Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data

Weight	Actual Height	Predicted Height	Difference (Error)
54	150	160	10
61	170	180	10
37	140	170	30
32	120	140	20
35	130	180	50
38	140	170	30

MAD
25

$$\frac{\sum |Error|}{n}$$

Mean absolute deviation

Mean absolute deviation is a way to describe variation in a data set. Mean absolute deviation helps us get a sense of how "spread out" the values in a data set are.

Explanation

Using the model built from the **training** data to predict height for the **testing** data and comparing it to the actual height values that are already available in the testing data (unexposed to the training data).

Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data

Weight	Actual Height	Predicted Height	Difference (Error)
54	150	160	10
61	170	180	10
37	140	170	30
32	120	140	20
35	130	180	50
38	140	170	30

MAD	RMSE
25	28
$\frac{\sum \lvert Error \rvert}{n}$	$\sqrt{\frac{\sum (Error)^2}{n}}$

Root Mean Square Error
the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are

Explanation

Using the model built from the **training** data to predict height for the **testing** data and comparing it to the actual height values that are already available in the testing data (unexposed to the training data).

Evaluating Model Accuracy

Regression Output

Predicted Variable	Predictor
Height	Weight

Testing Data

Weight	Actual Height	Predicted Height	Difference (Error)
54	150	160	10
61	170	180	10
37	140	170	30
32	120	140	20
35	130	180	50
38	140	170	30

MAD	RMSE	MAPE
25	28	18%
$\frac{\sum \lvert Error \rvert}{n}$	$\sqrt{\frac{\sum (Error)^2}{n}}$	$\frac{\sum \lvert Error / Actual\ Values \rvert}{n}$

Explanation

Using the model built from the **training** data to predict height for the **testing** data and comparing it to the actual height values that are already available in the testing data (unexposed to the training data).

Mean Absolute Percentage Error
most widely used **measure for checking forecast accuracy**. It comes under percentage errors which are scale independent and can be used for comparing series on different scales.

Qualitative Predictors

Types of Variables

Qualitative

Nominal: Lowest level of data measurement

E.g. Location, Mobile Number

Ordinal: Ranks an order or scaling

E.g.: Rank or grade of students

Binary: Has Boolean values

E.g.: 0/1, True/False

Ratio: Measured on a continuous scale

E.g.: Weight, Height, Age

Quantitative

Discrete: Data that can take only integer values, such as counts

E.g.: Number of cars passing by a building

Continuous: Data that can take any value in an interval

E.g.: Price, time, distance

Qualitative Predictors

- ✔ Genders : {Male, Female}
- ✔ Nationality : {India, Japan, Chaina, UK, US, Italy…….}
- ✔ Grades : {A, B, C, D, E, F}

	India	Japan	China
Assign Some Random Value	0.5	0.1	0.2

$$\beta_0+\beta_1 X = \begin{cases} 0.5 & \text{if “India”} \\ 0.1 & \text{if “Japan”} \\ 0.2 & \text{if “China”} \end{cases}$$

Qualitative Predictors

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

India

$$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

Japan

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

China

$$\beta_0 + \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{Bmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

Linear Regression

Linear Regression