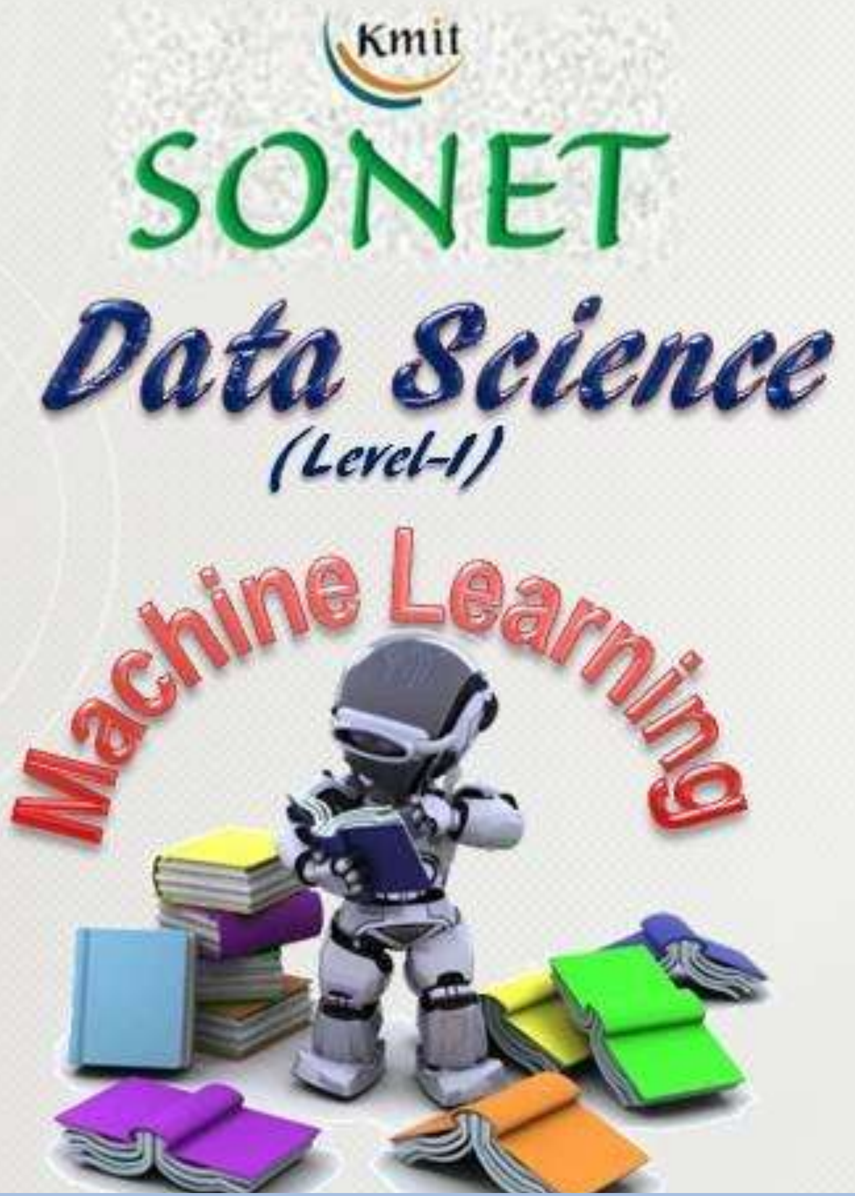


Ajeet K. Jain, M. Narsimlu

(ML TEAM)- SONET, KMIT, Hyderabad

Session – 11

This session deals with

Statistics

Why Statistics?

Inferential Statistics

Statistics Learning

What is Linear Regression

Linear Regression functionality

Statistics

- Statistics plays very important role in Data Science
- All the required algorithms for analysis are available in statistics
- Data representation in graphical and numeric is handled in statistics

Why Statistics?

- Everyone is using statistics unknowingly.
- Ex: to buy a mobile… we may compare…

  cost   Performance   Warranty   Etc…
- We can find statistics in (almost everywhere):

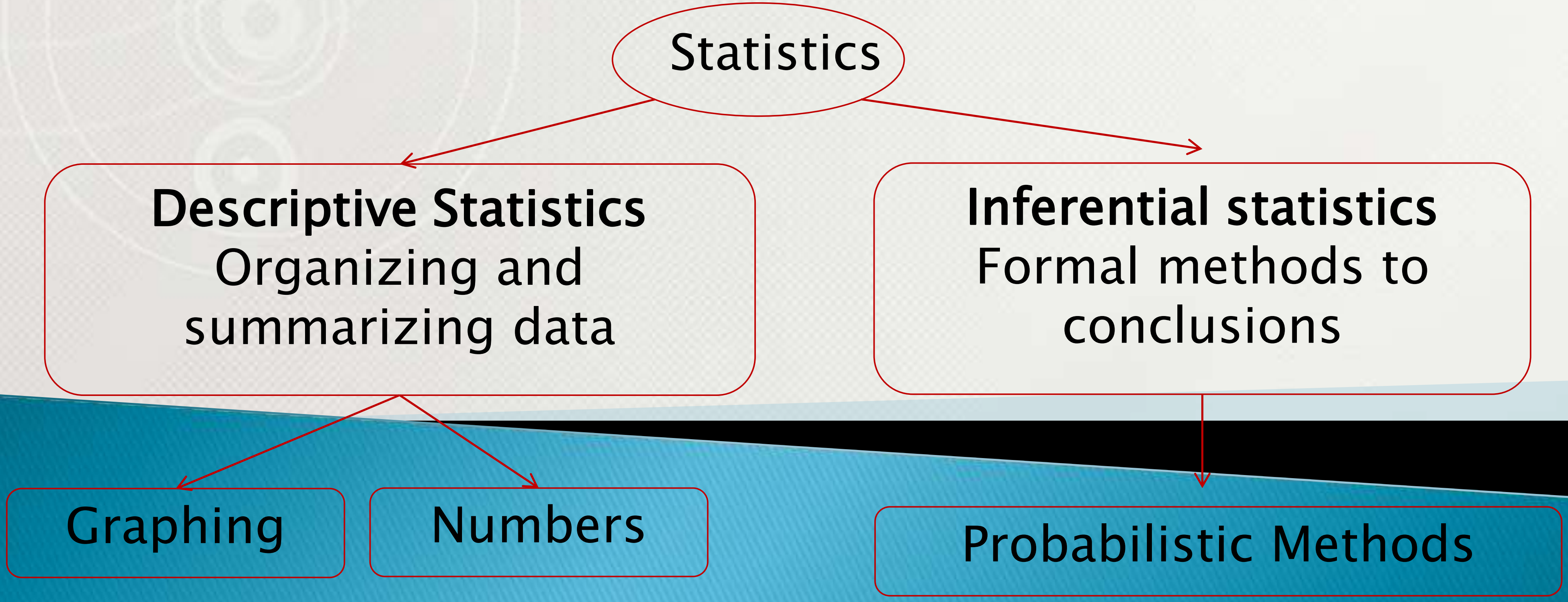
  News paper   Television   Sports   Stock Market

- Sources provide sample information
  - We can understand and take decision

- It is the body of methods for making **wise decisions** in the face of uncertainty on the basis of numerical data and calculated risks.

Statistics

- The science of Numbers that deals with the collection, analysis, interpretation and presentation of data.



Statistics

Descriptive Statistics
Organizing and summarizing data

Inferential statistics
Formal methods to conclusions

Graphing

Numbers

Probabilistic Methods

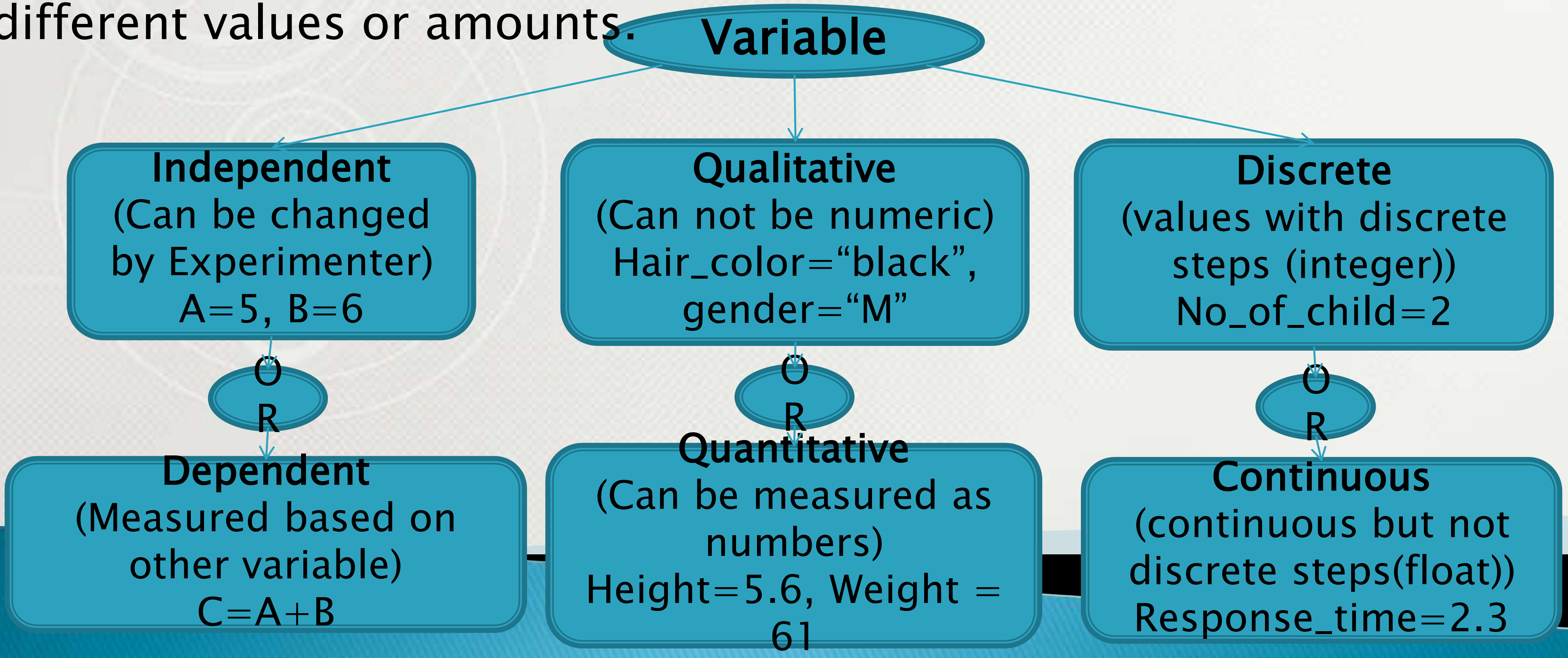
Descriptive Statistics

- Deals with Organizing and summarizing data.
- used when we know all of the values in order to describe a phenomenon

- There are 3 steps
    - Describing Data
    - Summarize / Analyze
    - Visualize

Descriptive statistics

- Variables : properties of some event, object, or person that can take on different values or amounts.



Variable

Independent
(Can be changed by Experimenter)
A=5, B=6

OR

Dependent
(Measured based on other variable)
C=A+B

Qualitative
(Can not be numeric)
Hair_color="black", gender="M"

OR

Quantitative
(Can be measured as numbers)
Height=5.6, Weight = 61

Discrete
(values with discrete steps (integer))
No_of_child=2

OR

Continuous
(continuous but not discrete steps(float))
Response_time=2.3

- Used for conducting analysis on one variable at a time or univariate Analysis

Inferential Statistics

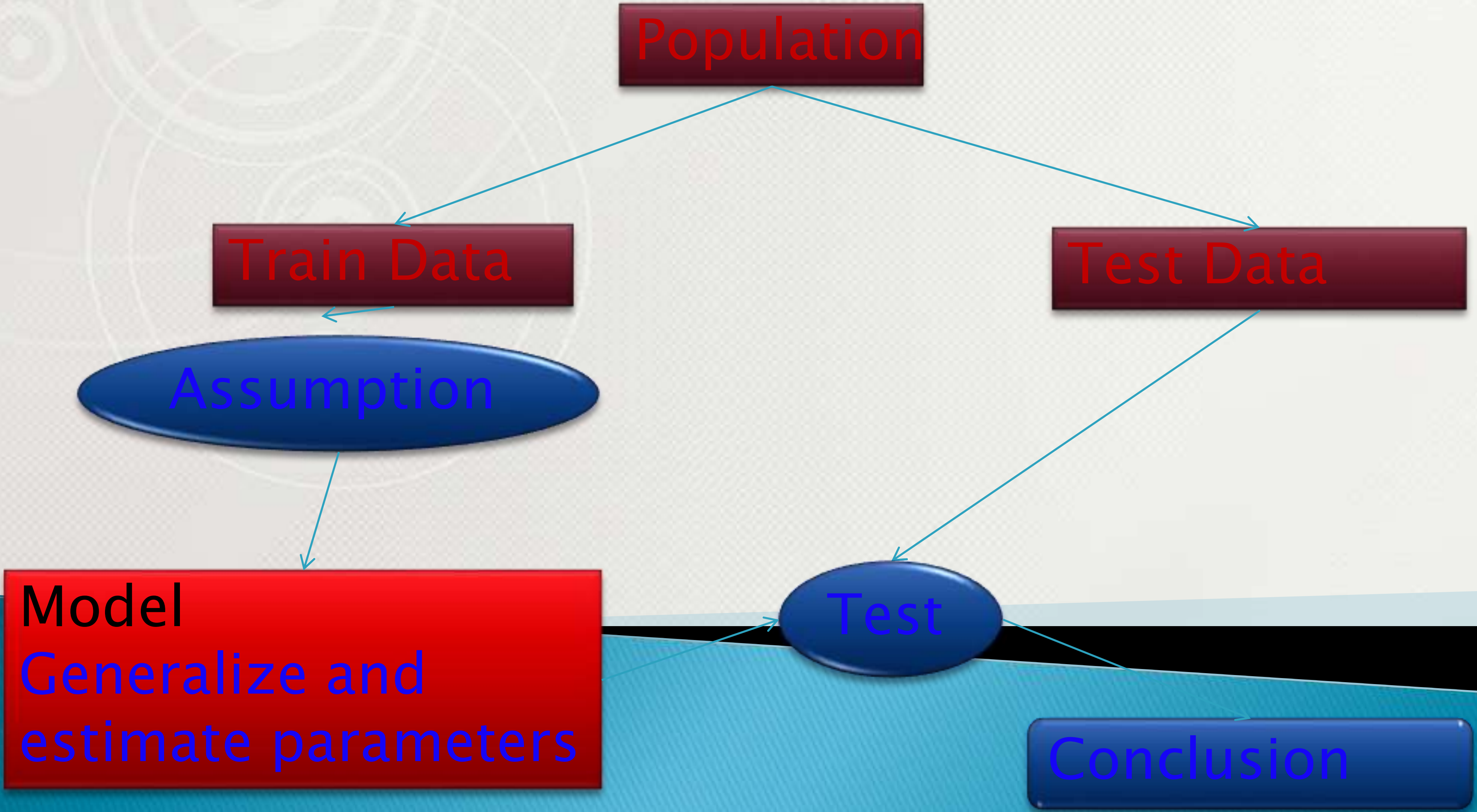
Inferential statistics use a random sample of data taken from a population to describe and make inferences about the population.

Population : Any group of data, which includes all the data interested in,

Sample: A smaller set of data, which are used to represent the larger population

The methods of inferential statistics
(1) the estimation of parameter(s)
(2) testing of statistical hypotheses

Inferential Statistics



Population

Train Data

Test Data

Assumption

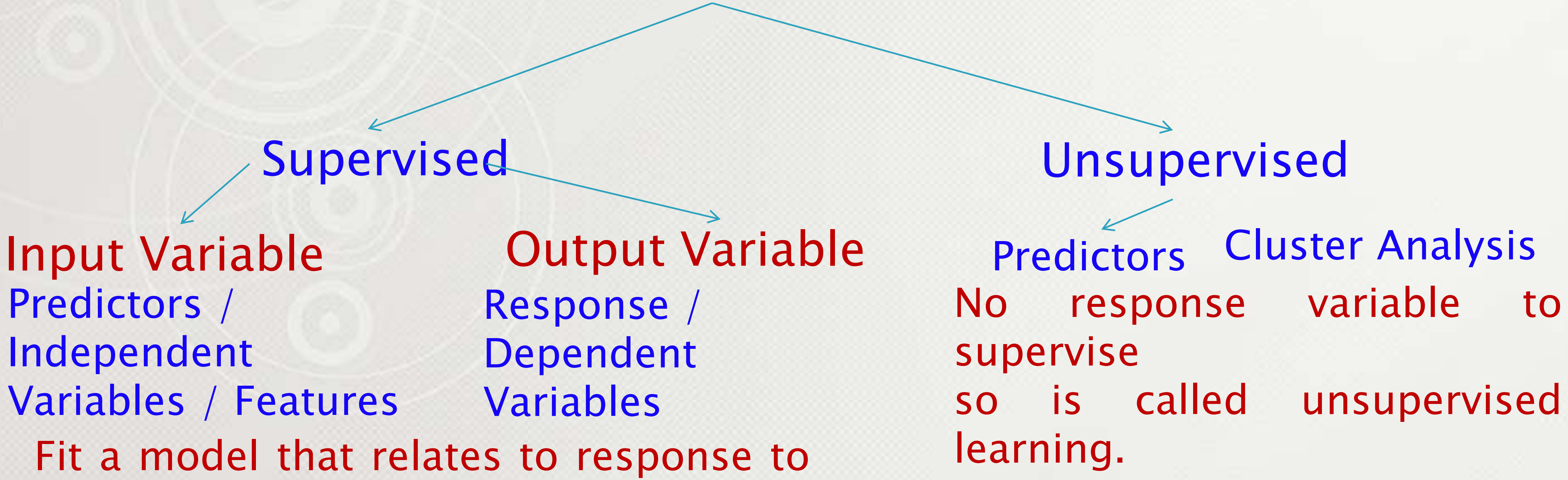
Model

Generalize and estimate parameters

Test

Conclusion

Statistical Learning



Supervised

Unsupervised

Input Variable

Predictors / Independent Variables / Features

Output Variable

Response / Dependent Variables

Predictors

Cluster Analysis

No response variable to supervise so is called unsupervised learning.

Fit a model that relates to response to the predictors, for predicting the response for future observations.

Linear Regression

Logistic Regression

Etc…

Statistical Learning

In the context of Statistical learning, there are two types of data:

Independent variables: Data that can be controlled directly.

Dependent variables: Data that cannot be controlled directly. Dependent variables need to predicted or estimated. To predict output will use model.

A model is a transformation engine that helps us to express dependent variables as a function of independent variables.

Parameters are ingredients added to the model for estimating the output.

Linear Regression

Linear: arranged in or extending along a straight or nearly straight line.

Linear suggests that the relationship between dependent and independent variable can be **expressed in a straight line.**

y = mx + c

y is the dependent variable i.e. the variable that needs to be estimated and predicted.

x is the independent variable i.e. the variable that is controllable. It is the input.

m is the slope. It determines what will be the angle of the line. It is the parameter denoted as β.

c is the intercept. A constant that determines the value of y when x is 0.

Linear Regression

Linear regression models are not perfect.

It tries to approximate the relationship between dependent and independent variables in a straight line.

Approximation leads to errors. Some errors can be reduced. Some errors are inherent in the nature of the problem.

These errors cannot be eliminated. They are called as an **irreducible error**

The noise term in the true relationship that cannot fundamentally be reduced by any model.

The equation can be re-written as

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

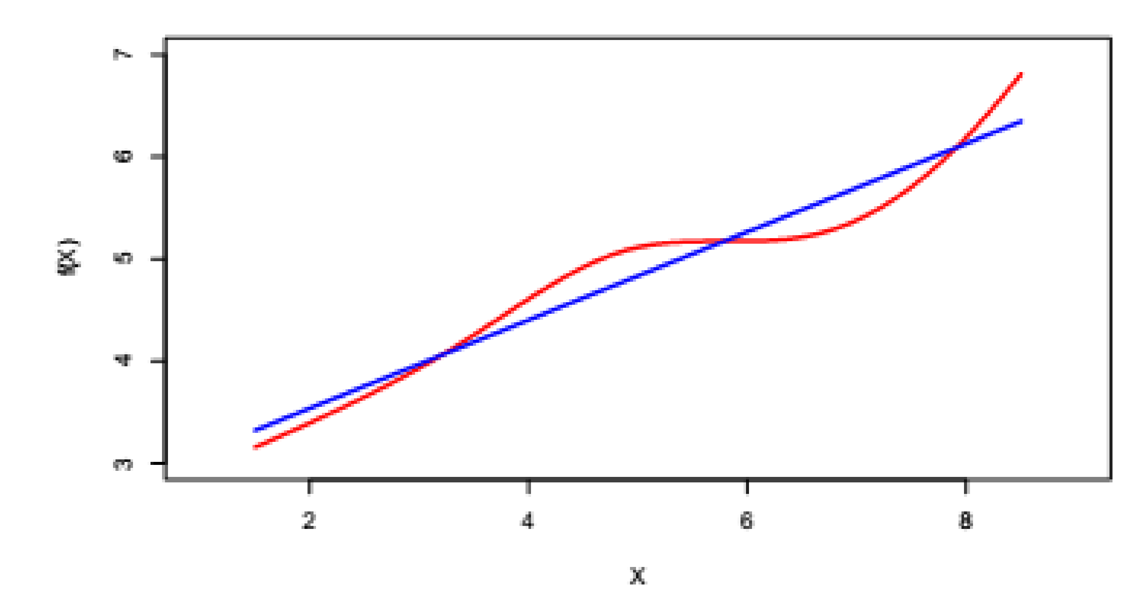
β0 and β1 are two unknown constants that represent the intercept and slope. **ε** is the error term

Linear Regression

Tool for predicting an unknown value based on existing data.

Linear Regression is Supervised Learning

Predictors X1,X2,X3,… Response Y=f(x)

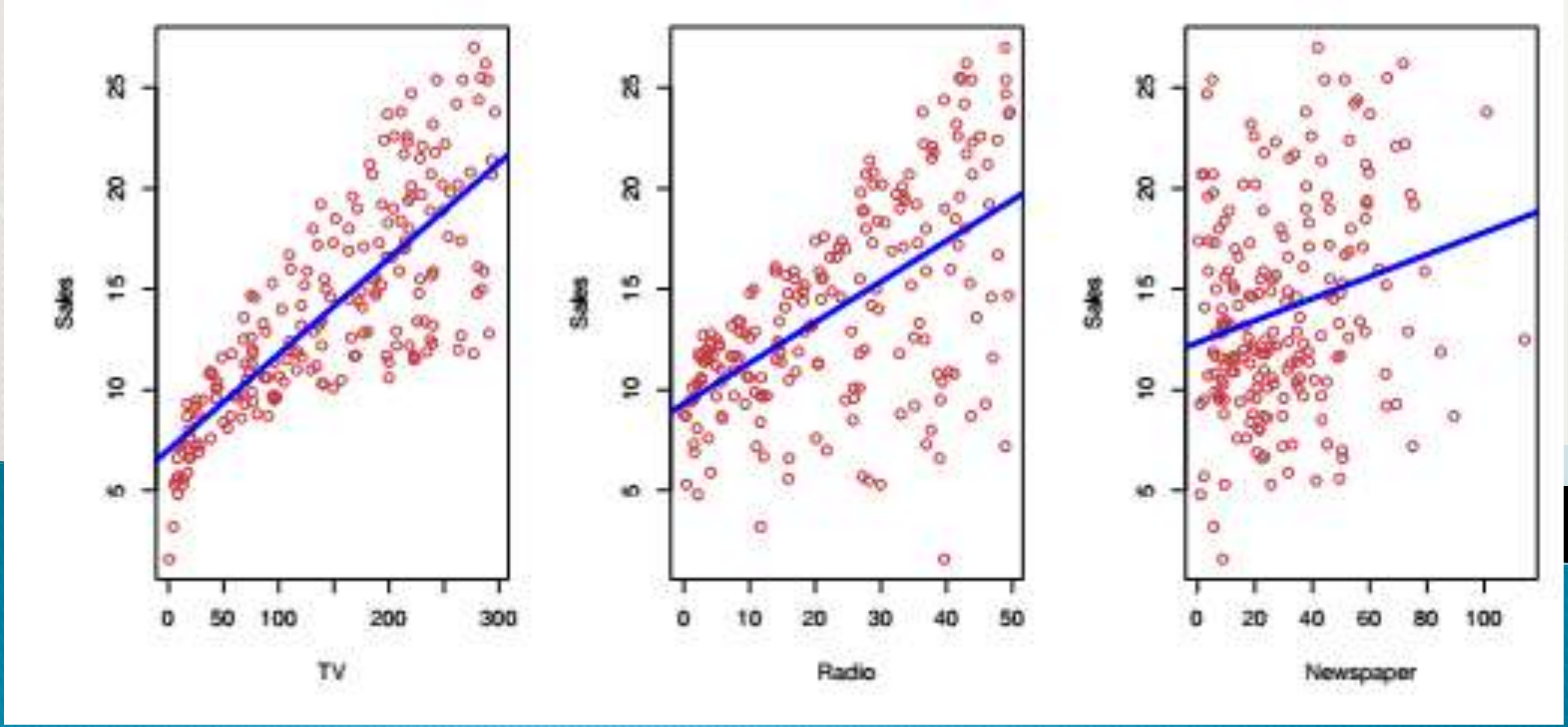


—— Linear Regression

—— Actual Regression

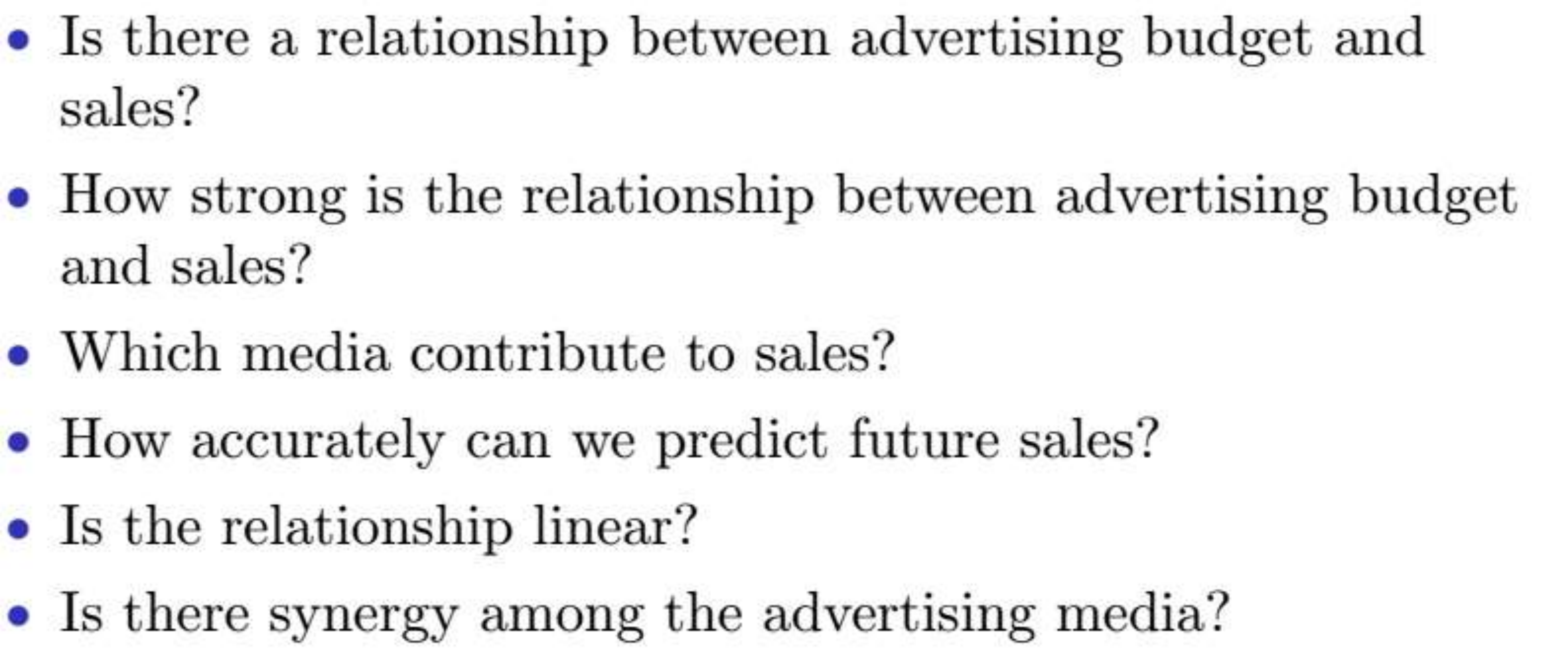
Linear Regression

In Advertising data,
Predictors: Budget for TV, Radio, News Paper
Response: Sales



Linear Regression

Linear regression answers…



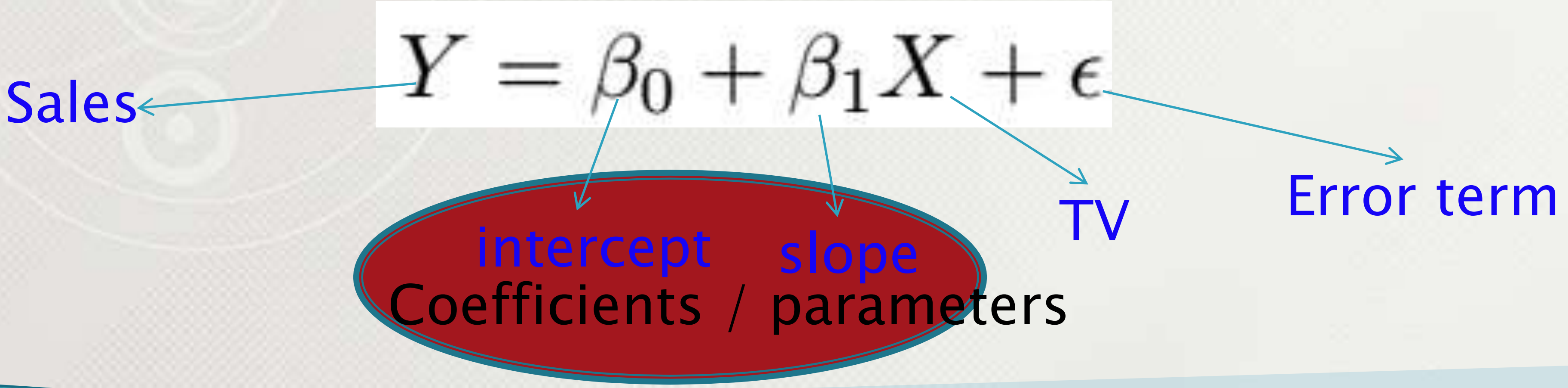
- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Simple Linear Regression

Predicting a quantitative response Y on the basis of single predictor variable X

Assumption: there is a linear relationship between x and y

Model:



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

Sales

intercept

slope

Coefficients / parameters

TV

Error term

Estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

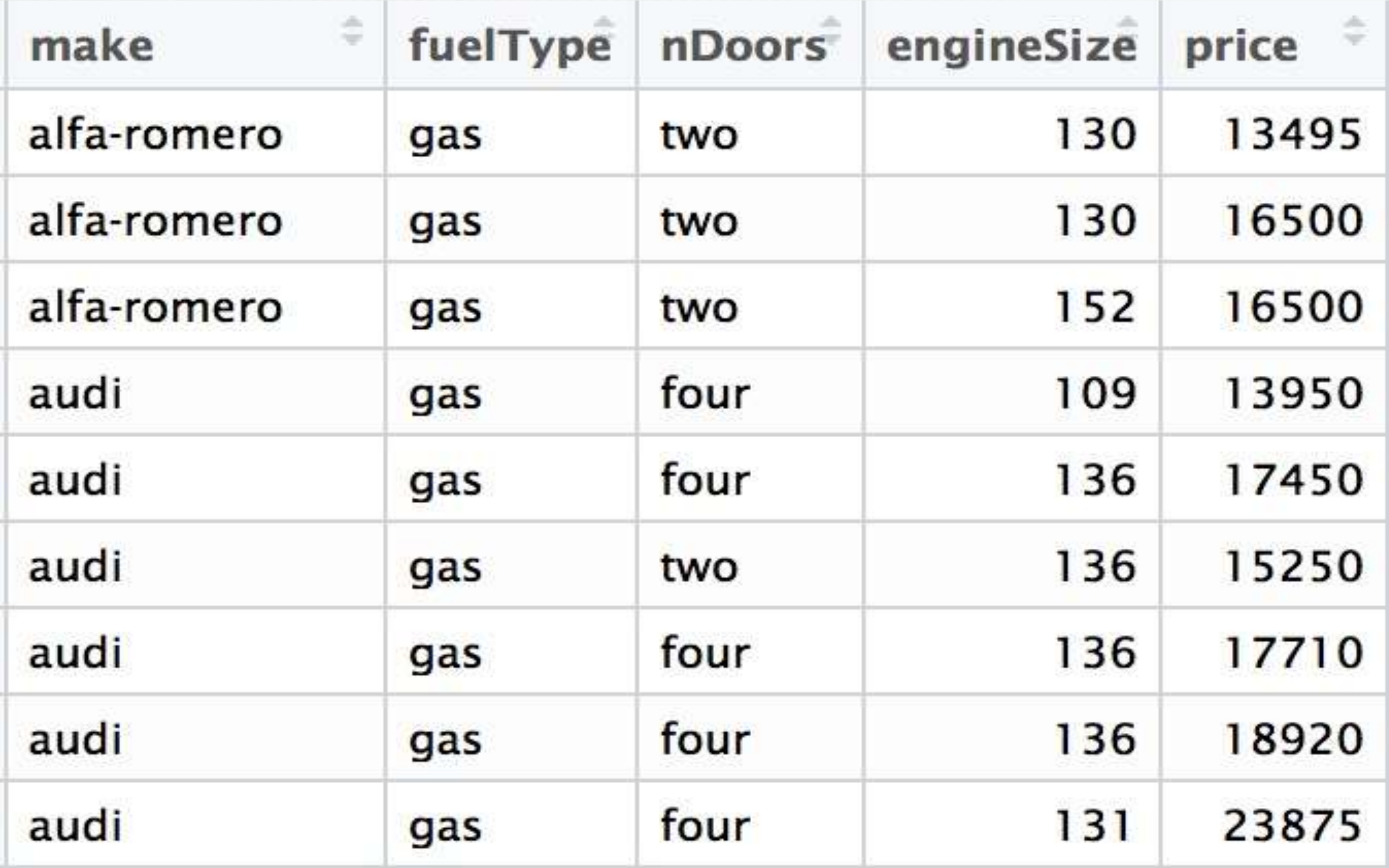
^ represents estimated term

Linear Regression

To buy a car

Predictors/features: make, fuel Type, nDoor, engine Size

Response: Price



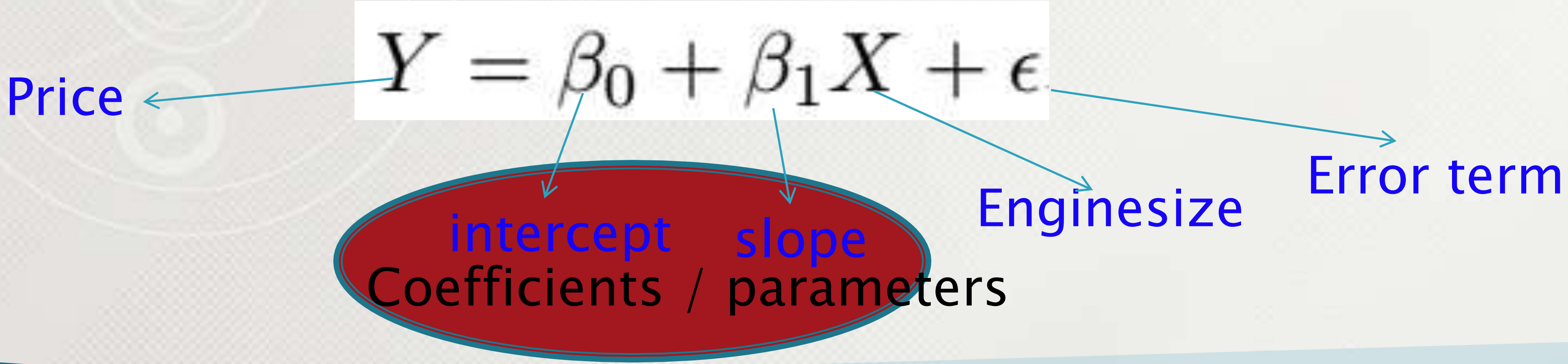
make	fuelType	nDoors	engineSize	price
alfa-romero	gas	two	130	13495
alfa-romero	gas	two	130	16500
alfa-romero	gas	two	152	16500
audi	gas	four	109	13950
audi	gas	four	136	17450
audi	gas	two	136	15250
audi	gas	four	136	17710
audi	gas	four	136	18920
audi	gas	four	131	23875

Simple Linear Regression

Predicting a quantitative response Y on the basis of single predictor variable X.

Assumption: there is a linear relationship between x and y

Model:



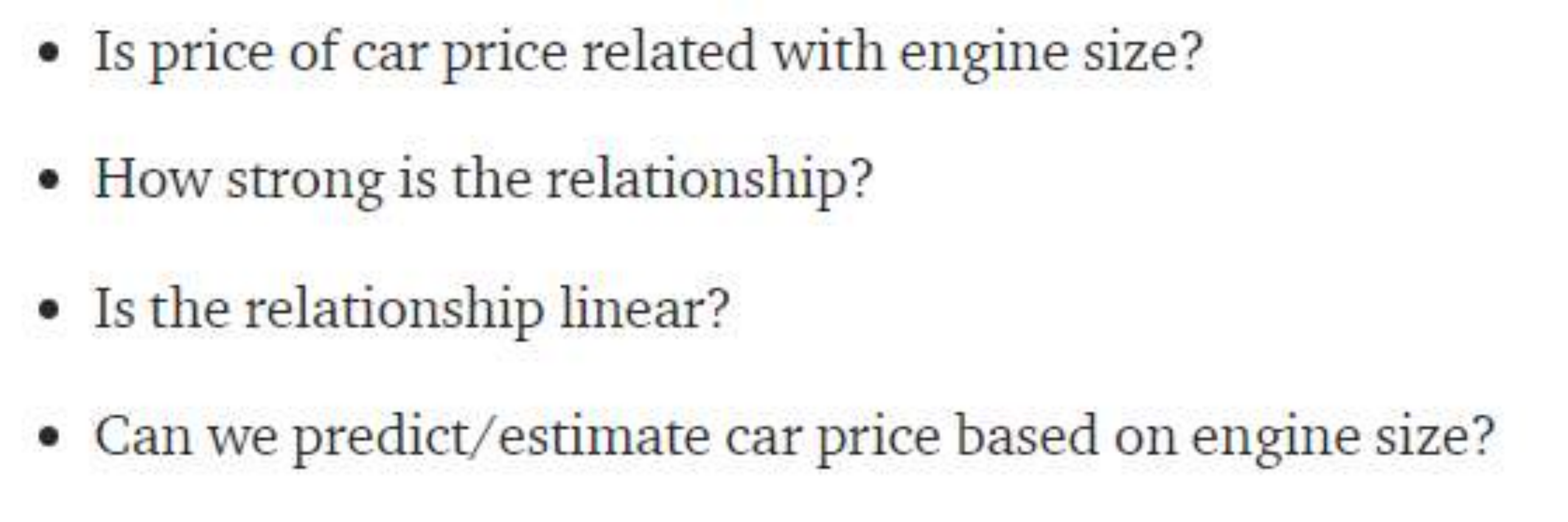
$$Y = \beta_0 + \beta_1 X + \epsilon$$

Estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

^ represents estimated term

Linear Regression

Linear regression answers…



- Is price of car price related with engine size?
- How strong is the relationship?
- Is the relationship linear?
- Can we predict/estimate car price based on engine size?

Linear Regression

Correlation is a measure of how much the two variables are related.

It is measured by a metric called as the **correlation coefficient**. Its value is between 0 and 1.

Correlation

- **Correlation** – show whether and how strongly pairs of variables are related.
  - For Example**:** engine size, price are related; *A decent prediction of price can be made using engine size.*
  - **Correlation can tell you just how much of the variation in products' engine size is related to price.**

Correlation

- **Correlation works for quantifiable data** (Numerical Data).
- It cannot be used for purely categorical data, such as gender, brands purchased, or favourite colour.
- Correlation is used to understand the relationship between variables such as:
  - Whether the relationship is +ve or -ve
  - the strength of the relationship.
- **Positive correlation**: is a relationship between two variables where if one variable increases, the other one also increases and vice-versa
  - Eg: family size , family expenditure will increase or decrease together.

Correlation

- **Negative Correlation**: there is an inverse relationship between two variables – when one variable decreases, the other increases and vice-versa.
  - **Eg: negative correlation**, between **price** and **demand** for goods and services. As the price of goods and services increases, the quantity demanded falls.
- **Coefficient of Correlation (r):** Statistical correlation is measured is known as "**coefficient of correlation (r)**". Its numerical value ranges from +1.0 to -1.0. It gives us **the strength of relationship.**

Correlation

- In general, r > 0 – positive relationship
- r<0 – negative relationship
- r = 0 – no relationship (meaning the variables are independent and not related)
- when r = +1.0 – describes a perfect +ve correlation
- when r = -1.0 – describes a perfect -ve correlation
- closer the coefficients are to +1.0 and -1.0, greater is the strength of the relationship between the variables.

Pre-requisite for Implementing algorithms

IDE

NumPy

Data Analysis Pandas

Python Data Structures

DATA VISUALIZATION

Data Science Case Studies

scikit learn

MATPLOTLIB

Data Visualization in Python

Seaborn

ANACONDA

spyder