Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Introduction

Data Analysis

- Data Analysis is a process of inspecting, cleaning, transforming and modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision-making.
- Intelligent data analysis (IDA) uses the concept from artificial intelligence (AI), information retrieval (IR), machine learning (ML), pattern reorganization, visualization, distributed programming and a host of other computer science concepts to automate the task of extracting unknown, valuable information /knowledge from the large amount of data.
- IDA process demands a combination of processes like extraction, analysis, conversion, classification, organization, and reasoning.
- The IDA process consists of 3 stages namely:
  - Data preparation
  - Data mining and rule finding
  - Result validation and interpretation.

Data Analytics

- Data analytics refers to the process of examining datasets to draw conclusions about the information they contain.
- Data analytic techniques enable to take raw data and uncover patterns to extract valuable insights from it.

Data Analysis vs. Data Analytics

Basis for Comparison	Data Analytics	Data Analysis
Form	Data analytics is 'general' form of analytics which is used in businesses to make decisions from data which are data-driven.	Data analysis is a specialized form of data analytics used in businesses to analyze data and take some insights of it.
Structure	Data analytics consist of data collection and inspect in general and has one or more users.	Data analysis consisted of defining a data, investigation, cleaning, transforming the data to give a meaningful outcome.
Tools	R, Tableau Public, Python, SAS, Apache Spark, Excel are used.	OpenRefine, KNIME, RapidMiner, Google Fusion Tables, Tableau Public, NodeXL, WolframAlpha are used.
Sequence	The life cycle consist of Business Case Evaluation, Data Identification, Data Acquisition & Filtering, Data Extraction, Data Validation & Cleansing, Data Aggregation & Representation, Data Analysis, Data Visualization, Utilization of Analysis Results.	The sequence followed are data gathering, data scrubbing, analysis of data and interpret the data precisely so that you can understand what data want to convey.

Data Analysis vs. Data Analytics cont...

Basis for Comparison	Data Analytics	Data Analysis
Usage	Find masked patterns, anonymous correlations, customer preferences, market trends and other necessary information that can help to make more notify decisions for business purpose.	Descriptive analysis, exploratory analysis, inferential analysis, predictive analysis and take useful insights from the data.
Example	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find that what the customers next possible purchases.	Suppose, 1gb customer purchase related data of past 1 year is available, now one has to find what happened so far.

Summary

- Both data analytics and data analysis are used to uncover patterns, trends, and anomalies lying within data, and thereby deliver the insights businesses need to enable evidence-based decision making.
- Where they differ, data analysis looks at the past, while data analytics tries to predict the future.
- Analysis is the detailed examination of the elements or structure of something. Analytics is the systematic computational analysis of data.

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Types of Data Analytics

- The main goal of big data analytics is to help organizations make **smarter decisions** for better business outcomes.
- With data in hand, you can begin doing analytics.
- **But where do you begin?**
- **And which type of analytics is most appropriate for your big data environment?**
- Looking at all the analytic options can be a daunting task. However, luckily these analytic options can be categorized at a high level into three distinct types.
  - **Descriptive Analytics,**
  - **Predictive Analytics,**
  - **Diagnostic Analytics**
  - **Prescriptive Analytics**

Descriptive Analytics –
(Insight into the past)

- Descriptive Analytics, which use data aggregation and data mining to provide insight into the **past** and answer:
  - **"What has happened in the business?"**

- Descriptive analysis or statistics does exactly what the name implies they "Describe", or summarize raw data and make it something that is interpretable by humans.

- The past refers to any point of time that an event has occurred, whether it is one minute ago, or one year ago.

- Descriptive analytics are useful because they allow us to **learn from past** behaviors, and understand how they might influence future outcomes.
- **Example** : Data Queries, Reports, Descriptive Statistics, Data dashboard, Monthly revenue reports, Sales leads overview

Descriptive Analytics

- **The main objective of descriptive analytics is to find out the reasons behind precious success or failure in the past.**
- The vast majority of the statistics we use fall into this category.
- Common examples of descriptive analytics are **reports that provide historical insights regarding the company's production, financials, operations, sales, finance, inventory and customers**.
- It performs an in-depth analysis of data to reveal details such as frequency of events, operation costs, and the underlying reason for failure. **It helps in identifying the root cause of the problem.**
- **Common examples of Descriptive analytics are company reports that provide historic reviews like:**
  - **Data Queries, Reports, Descriptive Statistics, Data dashboard**

Diagnostic Analysis

- "**Why did it happen?**"
- Diagnostic analysis takes the insights found from descriptive analytics and drills down to find the causes of those outcomes.
- Organizations make use of this type of analytics as it creates more connections between data and identifies patterns of behavior.
- **Example** : A freight company investigating the cause of slow shipments in a certain region
- **Common examples of diagnostic analysis: Data discovery, Data mining, Correlations**

Predictive Analytics – *(Understanding the future)*

- Predictive Analytics, which use **statistical models** and forecasts techniques to understand the future and answer:
  - **"What could happen?"**
- These analytics are about understanding the future.
- **Predictive analytics provide estimates** about the likelihood of a future outcome. It is important to remember that no statistical algorithm can "predict" the future with 100% certainty.
- Companies use these statistics to forecast what might happen in the future. This is because the foundation of **predictive analytics is based on probabilities**.
- These statistics try to take the data that you have, and fill in the missing data with best guesses.
- **Example** : Linear Regression, Time series analysis and forecasting, Data Mining

Predictive Analysis

- Predictive analytics can be further categorized as –
  - Predictive Modelling –What will happen next, if ?
  - Data Mining- Identifying correlated data.
  - Forecasting- What if the existing trends continue?
  - Pattern Identification and Alerts –When should an action be invoked to correct a process.

- **Sentiment analysis** is the most common kind of predictive analytics. The learning model takes **input in the form of plain text** and **the output of the model is a sentiment score that helps determine whether the sentiment is positive, negative or neutral.**

Prescriptive Analytics –
(Advise on possible outcomes)

- Prescriptive Analytics, which use optimization and simulation algorithms to advice on possible outcomes and answer:
  - **"What should we do?"**
- The relatively new field of prescriptive analytics allows users to "prescribe" a number of different possible actions to and guide them towards a solution. In a nut-shell, these analytics are all about providing advice.
- Prescriptive analytics is the next step of predictive analytics that adds the spice of manipulating the future.
- **For example, Prescriptive Analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demography, etc.**

Prescriptive Analytics

- Prescriptive analytics is an advanced analytics concept based on,
  - **Optimization that helps achieve the best outcomes**.
  - Stochastic optimization that helps understand how to achieve the best outcome and identify data uncertainties to make better decisions.
- Prescriptive analytics is a **combination of data, mathematical models and various business rules**. The data for prescriptive analytics can be both internal (within the organization) and external (like social media data).
- Prescriptive analytics can be **used in healthcare to enhance drug development, finding the right patients for clinical trials, etc.**

Types of Analytics

Approach	Explanation
Descriptive	What's happening in my business? • Comprehensive, accurate and historical data • Effective Visualisation
Diagnostic	Why is it happening? • Ability to drill-down to the root-cause • Ability to isolate all confounding information
Predictive	What's likely to happen? • Decisions are automated using algorithms and technology • Historical patterns are being used to predict specific outcomes using algorithms
Prescriptive	What do I need to do? • Recommended actions and strategies based on champion/challenger strategy outcomes • Applying advanced analytical algorithm to make specific recommendations

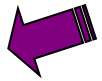
Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Importance of Data Analysis

- Data analysis is important to businesses will be an understatement. In fact, no business can survive without analyzing available data. Visualize the following situations:
  - A pharmacy company is performing trials on number of patients to test its new drug to fight cancer. The number of patients under the trial is well over 500.
  - A company wants to launch new variant of its existing line of fruit juice. It wants to carry out the survey analysis and arrive at some meaningful conclusion.
  - Sales director of a company knows that there is something wrong with one of its successful products, however hasn't yet carried out any market research data analysis. How and what does he conclude?
- These situations are indicative enough to conclude that data analysis is the lifeline of any business. Whether one wants to arrive at some marketing decisions or fine-tune new product launch strategy, data analysis is the key to all the problems.
- Merely analyzing data isn't sufficient from the point of view of making a decision. How does one interpret from the analyzed data is more important. Thus, data analysis is not a decision making system, but decision supporting system.
- Data analysis can offer the following benefits:
  - Structuring the findings from survey research or other means of data collection.
  - Break a macro picture into a micro one.
  - Acquiring meaningful insights from the dataset.
  - Basing critical decisions from the findings.
  - Ruling out human bias through proper statistical treatment.

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications 
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Data Analytics Applications

1. **Understanding and targeting customers:** Data analytics is extremely useful to understand and predict customer behaviour. The trend is towards getting a 360 degree view of each customer which includes data from traditional customer purchase data as well as the non-traditional unstructured data sets like social media, web logs, customer clicks on e-retail sites, etc. This picture helps businesses predict which customers may move to a rival (called customer churn), predict what products will sell, predict living patterns that can help insurance companies charge a differential premium and so on. The list is endless with potential business benefits.
2. **Understanding and optimizing business processes:** Information is collected from social media resulting in sentiment analysis. This along with company profiles data are analyzed using data analytics tools to effectively predict demand for products and thus help to retain stock in warehouses to an optimal level. For example, Apple excessively uses sentiment analysis information from social media to gauge the potential sales of their new iPhone 6s offering. By providing a holistic view of assets and business processes, enterprises are now able to gain unparalleled insight into optimizing those assets and processes. For example, a Fortune 50 company was able to optimise accounts receivable collections without increasing collector headcount. This was possible because data analytics tools crawled and mined historical data, identified the factors that affect late payments, provided insights in the collection system, and provided ongoing recommendations that helped improved Accounts Receivable collections by 65% over the prior year.

Data Analytics Applications cont…

3. **Personal quantification and performance optimization:** Personal quantification is an emerging trend in big data science. Self-quantification of personal health and wellness data are contributing heavily towards more self-managed health care. Advances in network and wearable sensor technologies easily help to capture and share significant health-related information on a daily basis. New functionalities in wearable devices and the associated apps enable individuals to measure vital signs, access analytical tools and quantify data about themselves faster and more ubiquitously than ever before. For example, keeping diaries of food intake, converting these collected data into numbers, analysing them and using them to make better decisions regarding personal health are part of personal quantification.

4. **Improving healthcare and public health:** The healthcare industry historically has massive amounts of data in its archives. This may be due to record keeping, compliance and regulatory requirements, continual patient care, etc. This voluminous data totally renders itself to data analytics applications. This data includes clinical data from hospitals and clinical decision support systems (physician's written notes and prescriptions, medical imaging, laboratory test data, pharmacy prescriptions and sales, insurance data, patient data in electronic patient records (EPRs)) machine generated/sensor data, such as from monitoring vital signs; social media data, news feeds, and articles in medical journals, etc. This data can be analyzed to effectively provide customised medical care to patients, detect epidemics, curtail infections and several such applications. Interesting current application include holistic cancer treatment, genomics, identifying and stopping hospital fraud and the like.

Data Analytics Applications cont…

5. **Improving sports performance:** Any sports, be it football, car racing or sailing, gets affected by advances in the capture, storage and analysis of data. Data analytics allows athletes to train better and more effectively and it allows teams to alter their in-game decision-making based on what they are seeing. Like other businesses, sports teams strive to make better decisions faster. Coaching staff, scouts and players are leveraging analytics to better understand the performance of their own teams as well as that of the opposition.
6. **Improving science and research:** The emergent field of data analytics is rapidly changing the direction and speed of scientific research by letting people fine-tune their inquiries by tapping into giant data sets. In the past, certain fields of science relied heavily on big data sets, such as high-energy particle physics or research on nuclear fusion. But as information becomes available from more sources, collecting and analysing large amounts of data is becoming common in other fields of research too. One such recent example is the research being conducted at CERN, the Swiss nuclear physics laboratory with its Large Hadron Collider. The CERN data centre has 65,000 processors to analyze its 30 petabytes of data. It uses the computing powers of thousands of computers distributed across 150 data centres worldwide to analyze the data.
7. **Optimizing machine and device performance:** Big data analytics helps machines and devices become smarter and more autonomous. Data analytics can be utilised to collect energy usage data from smart meters, analyze usage patterns and effectively provide smart grids that can optimise energy usage. Looking into usage patterns of machines at large manufacturing plants, analytics can effectively predict machine down time and that helps to perform preventive maintenance resulting in saving of huge amount of resources.

Data Analytics Applications cont…

8. **Improving security and law enforcement:** By taking advantage of big data, crime analysts identify trends and make recommendations based on their observations. Through analysis and computer mapping, crime analysts play a crucial role in helping law enforcement agencies quantify, evaluate, and respond to the changing landscape of criminal activity in their jurisdictions. Typical applications use big data techniques to detect and prevent cyberattacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data to detect fraudulent transactions.
9. **Improving and optimizing cities and countries:** Today's towns and cities generate about 5 terabytes per day per square kilometre of urbanized land area. This includes location data collected by smart phones to data generated by GPS instruments, payment cards, smart ID cards, loyalty and store cards, bank cards, toll payments, etc. Further sources of data include data created by traffic management systems, from traffic lights to the sensors on our roads; from the provision of utilities such as gas, electricity and drinking water, etc. All this data can be analyzed to improve many aspects of our cities and citizen's daily life. For example, we can have intelligent route planning systems that are based on real-time traffic information as well as social media and weather data. Smart cities are planned by integrating and analysing all subsystems in a city like energy, traffic, police, etc.
10. **Financial trading:** High-frequency trading is an area where big data finds a lot of use today. Data analytics technologies have advanced sufficiently to provide millisecond latency on large data sets. Here, big data analysis algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that take into account signals from social media networks and news websites, weather predictions, etc. to make, buy and sell decisions in split seconds.

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::** 
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Regression Modelling Techniques

- 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.
- Lets take a simple example: Suppose your manager asked you to predict annual sales. There can be factors (drivers) that affects sales such as competitive pricing, product quality, shipping time & cost, online reviews, easy return policy, loyalty rewards, word of mouth recommendations, ease of checkout etc. In this case, sales is your dependent variable. Factors affecting sales are independent variables.
- Regression analysis would help to solve this problem. In simple words, **regression analysis is used to model the relationship between a dependent variable and one or more independent (predictors) variables and then use the relationships to make predictions about the future.**
- Regression analysis helps to answer the following questions:
  - Which of the drivers have a significant impact on sales?
  - Which is the most important driver of sales?
  - How do the drivers interact with each other?

Regression Modelling Techniques cont…

Two main objectives:

- *Establish if there is a **relationship** between two variables*
  - *Specifically, establish if there is a statistically significant relationship between the two.*
  - *Example: Income and expenditure, wage and gender, etc.*

  ***Forecast** new observations.*
  - *Can we use what we kow about the relationship to forecast unobserved values?*
  - *Example: What will are sales be over the next quarter?*

- The regression analysis allows to model the dependent variable as a function of its predictors

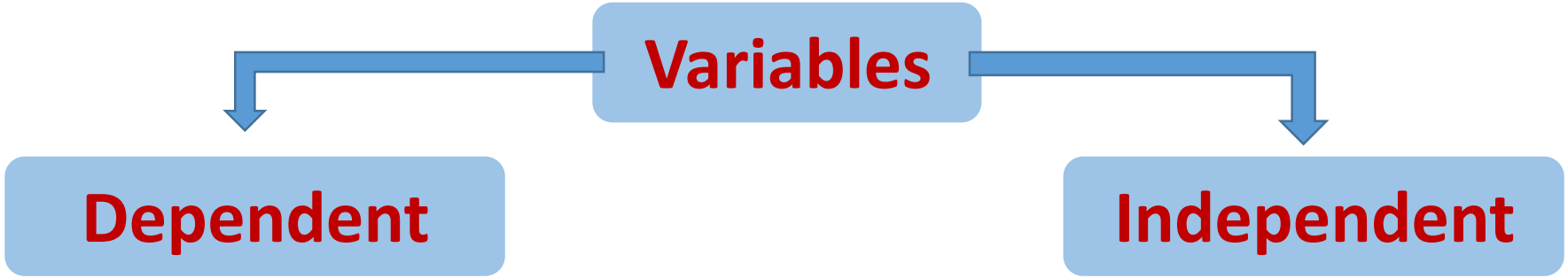
$$\mathbf{Y = f(X_i, \beta) + e_i}$$

where Y is dependent variable,

f is the function,
X_i is the independent variable,
β is the unknown parameters,
e_i is the error term, and i varies from 1 to n.

Regression Modelling Techniques cont…

Variable's Roles in => Y=mX+c



- The variable whos values we want to predict/forecast.
- Its values depends on something else.
- We denote it as Y.

- The variable that explains the other one.
- Its values are independent.
- We denote it as X.

Regression Modelling Techniques Terminology

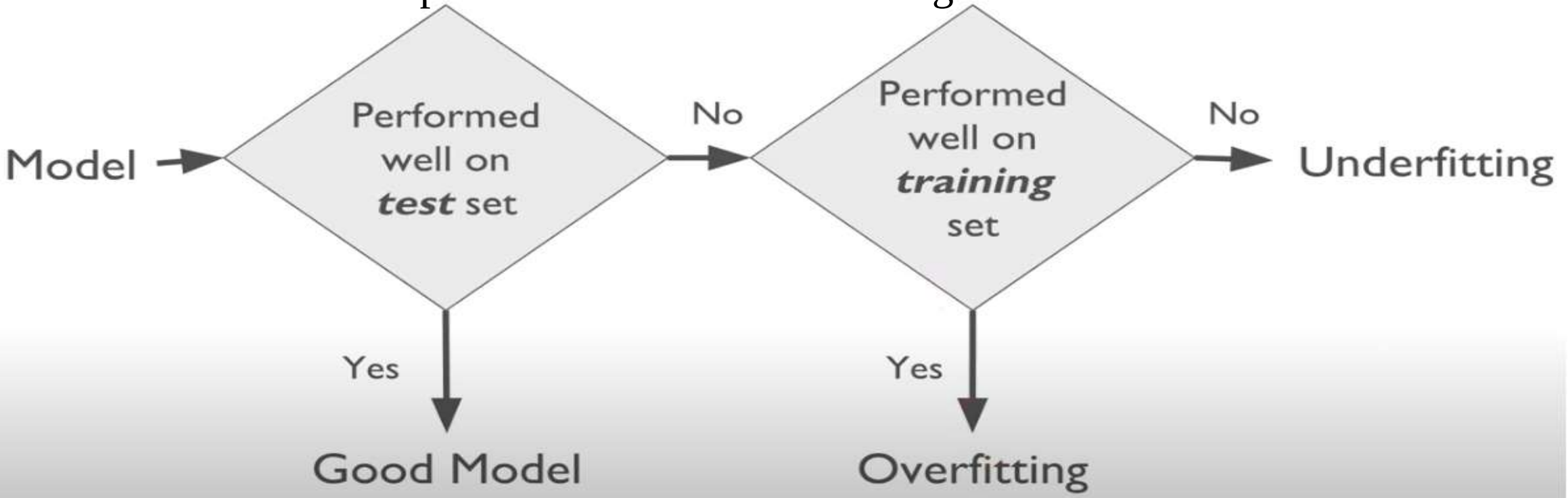
Terminologies

- **Multicollinearity**: When the predictors are highly correlated to each other then the variables are said to be multicollinear. Many types of regression techniques assumes multicollinearity should not be present in the dataset. It is because it causes problems in ranking variables based on its importance or it makes job difficult in selecting the most important independent variable (factor).
- **Heteroscedasticity:** When dependent variable's variability is not equal across values of an independent variable, it is called heteroscedasticity. Example -As one's income increases, the variability of food consumption will increase. A poorer person will spend a rather constant amount by always eating inexpensive food; a wealthier person may occasionally buy inexpensive food and at other times eat expensive meals. Those with higher incomes display a greater variability of food consumption.
- **Training and Test dataset:** In a dataset, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. So, training data is used to fit the model and testing data to test it.

Regression Modelling Techniques Terminology...

Terminologies

- **Overfitting:** It means that model works well on the training dataset but is unable to perform better on the test datasets. It is also known as problem of high variance. Variance indicates how much the estimate of the model will alter if different training data were used.
- **Underfitting:** When the model works so poorly that it is unable to fit even training set well then it is said to be underfitting the data. It is also known as problem of high bias. A bias is the amount that a model's prediction differs from the target value.



Regression Modelling Techniques Terminology...

Terminologies

- **Bias(Error in training data):** While making predictions, a difference occurs between prediction values made by the model and actual values/expected values, and this difference is known as bias errors or Errors due to bias
- **Low Bias:** A low bias model will make **fewer assumptions** about the form of the target function.
- **High Bias:** A model with a high bias makes **more assumptions**, and the model becomes unable to capture the important features of our dataset. **A high bias model also cannot perform well on new data.**

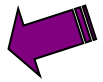
Types of Regression

- Every regression technique has some assumptions attached to it which need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution. The types of regression algorithms are:

- Linear Regression
- Multiple Linear Regression
- Non Linear Regression
- Logistic Regression
- Polynomial Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Principal Components Regression (PCR)
- Partial Least Squares (PLS) Regression
- Support Vector Regression

- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi Poisson Regression
- Cox Regression

Chapter 3: **Data Analytics**

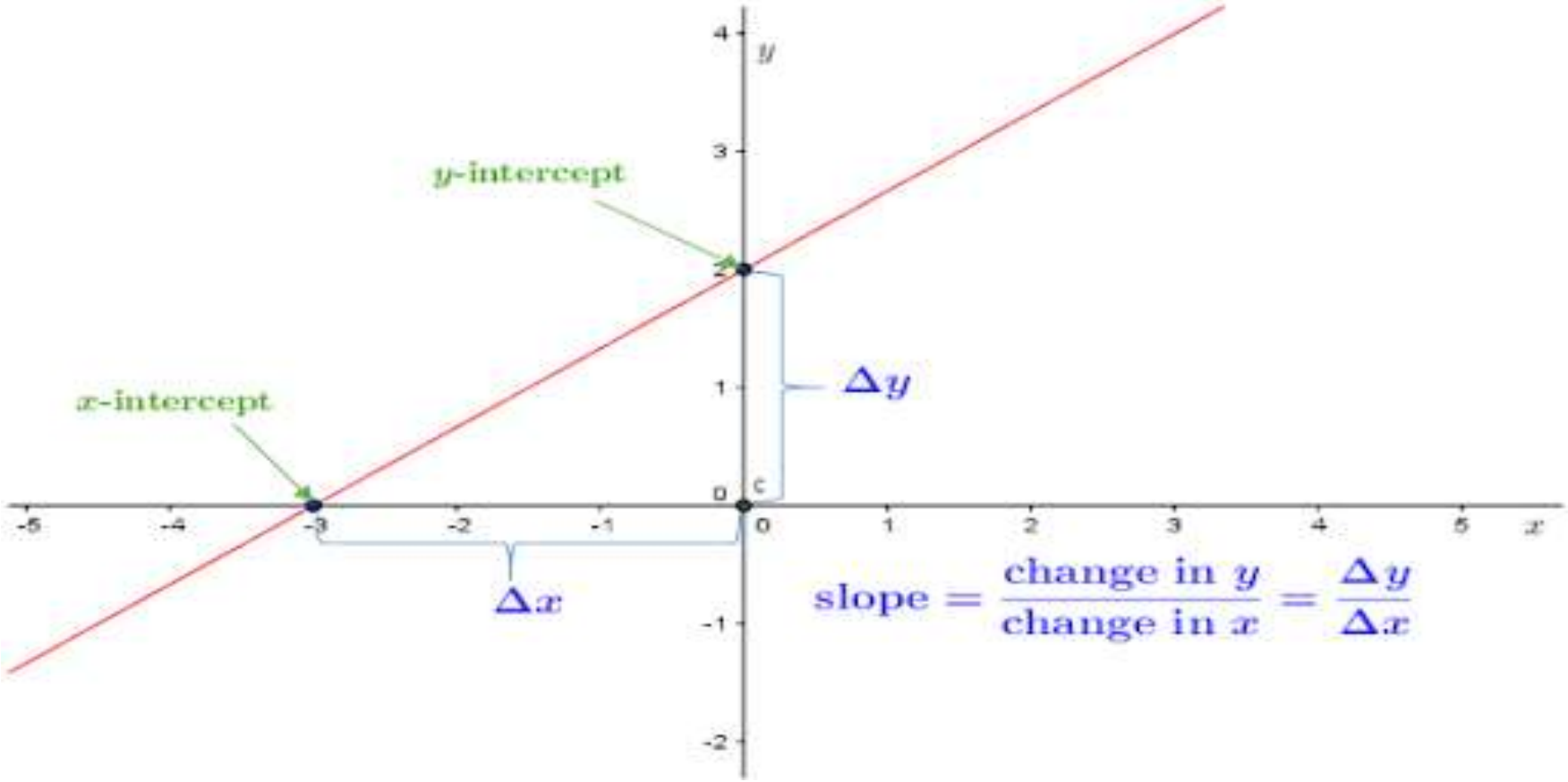
- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression 
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

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. e.g., 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 (higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables.
- A scatterplot can be a helpful tool in determining the strength of the relationship between 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…

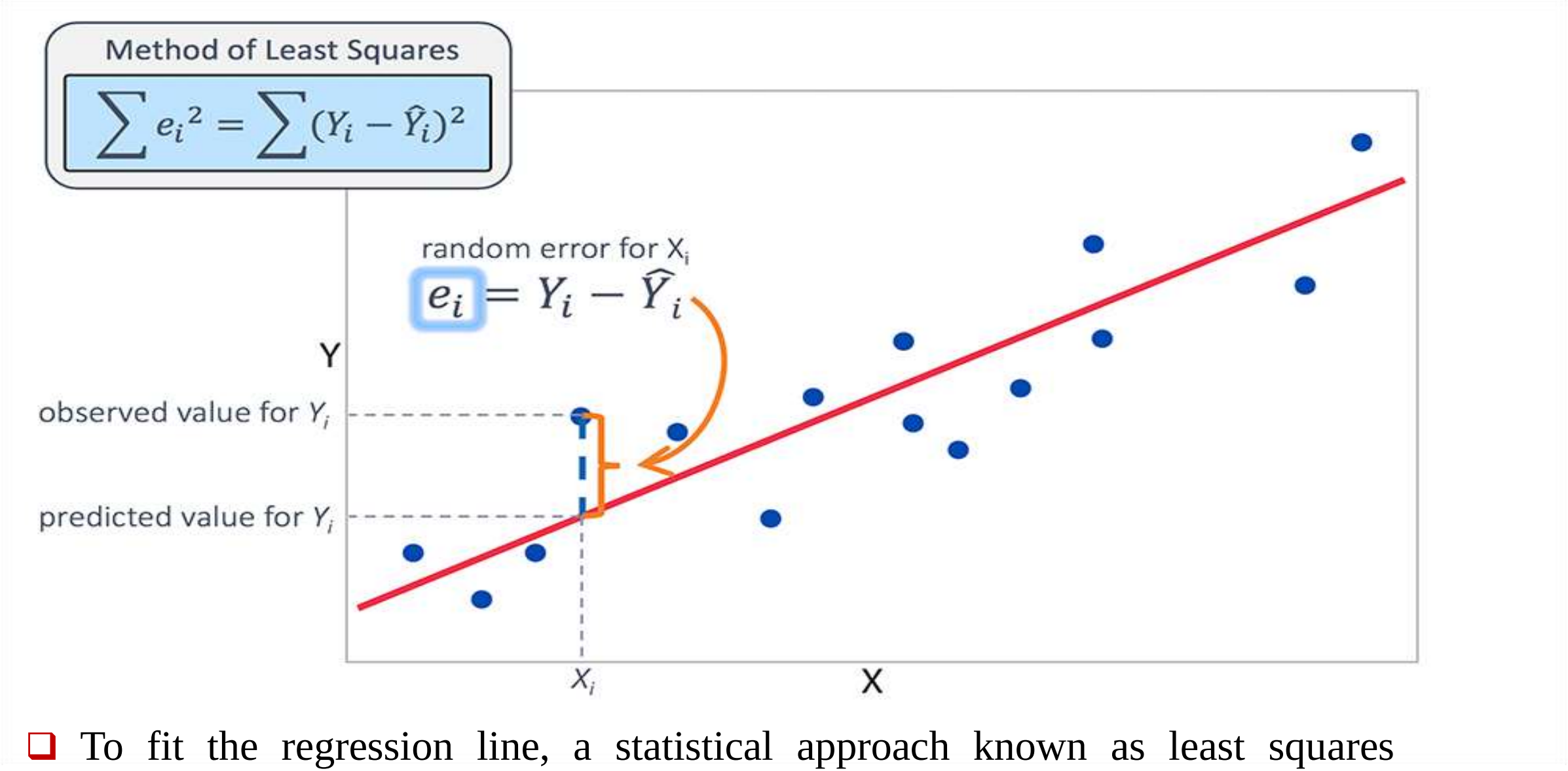
- A linear regression line has an equation of the form **Y = a + bX + e**, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, a is the intercept (the value of y when x = 0), and e is the random error.
- The slope and intercept is as follows in the following linear equation of line:



In the above snap, a is considered as the y-intercept.

Linear Regression cont…

- The random error in the following linear equation of line:



- To fit the regression line, a statistical approach known as least squares

Linear Regression cont...

Least Squares Principle

- Dots are actual values of Y
- Asterisks are the predicted values of Y for a given value of X

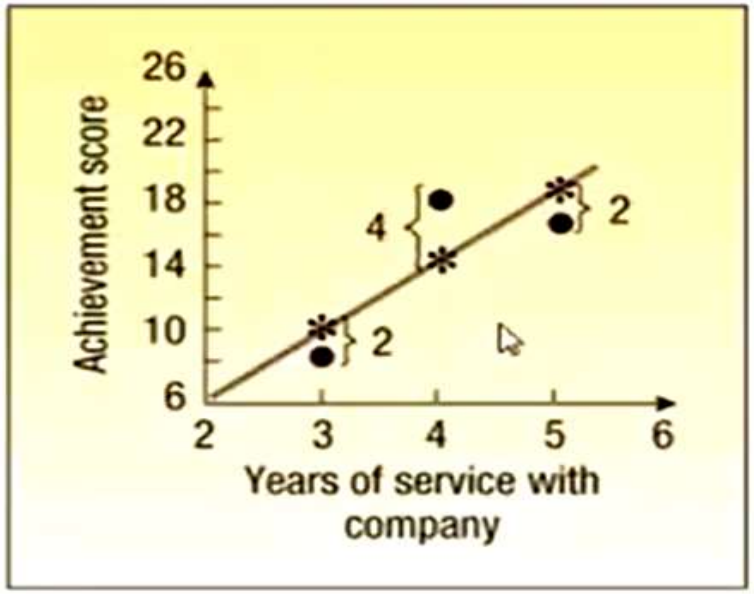


Chart 1

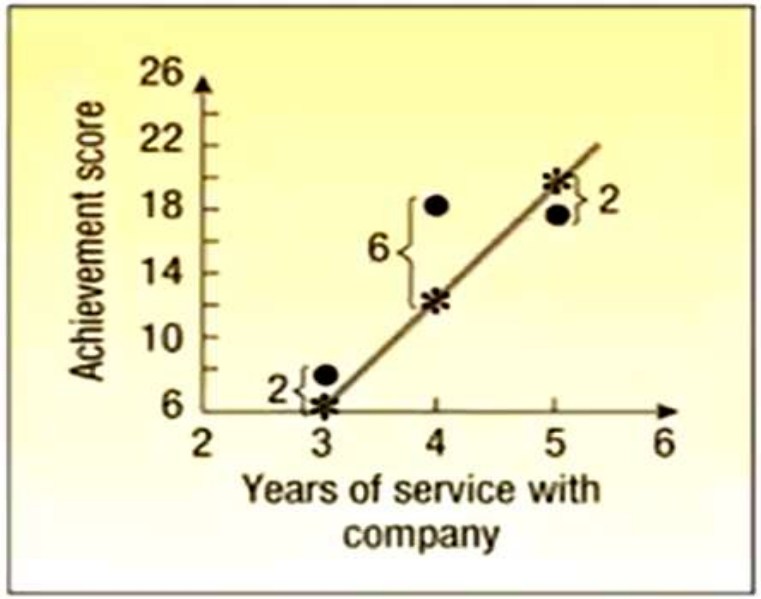


Chart 2

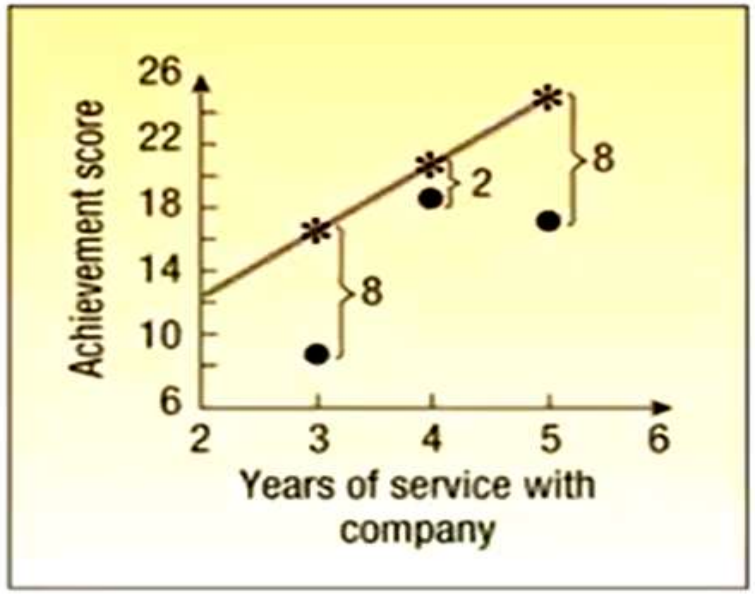


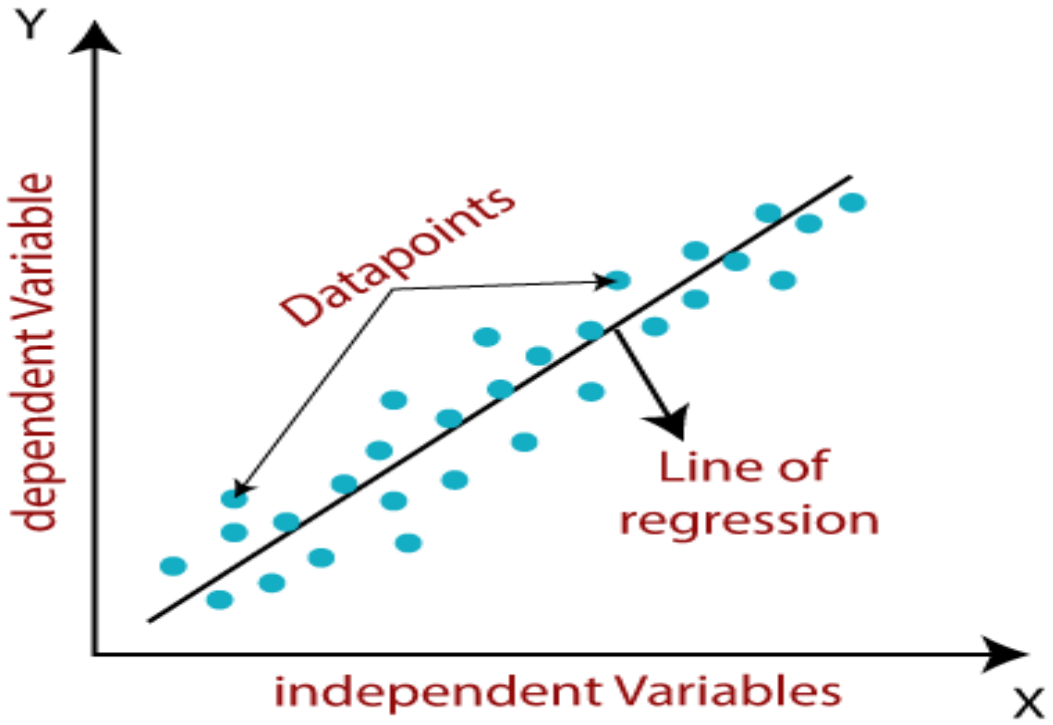
Chart 3

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

According to the Least Square Principle chart-1 error value is less compared to others and thus it is the best fit line

Linear Regression cont…

- The linear regression model provides a sloped straight line (that is most closer to true data point) representing the relationship between the variables. Consider the below image:



When working with linear regression, the main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

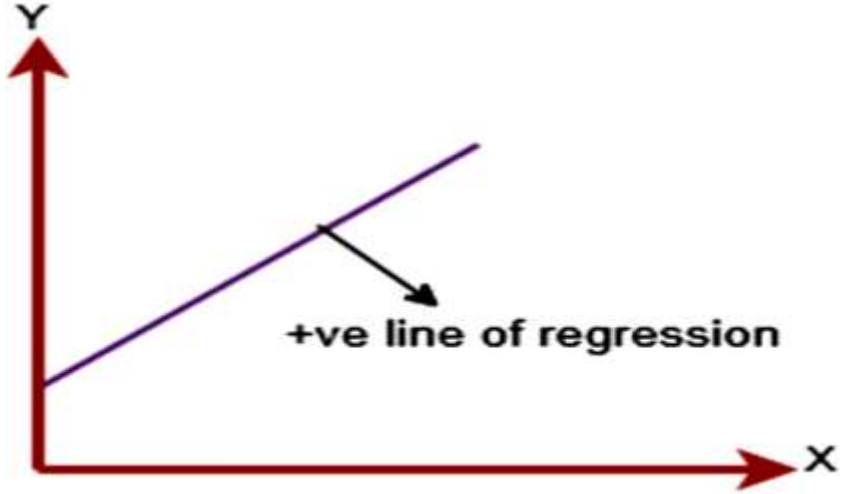
Linear Regression cont...

Simple Linear Regression Model

A linear line showing the relationship between the dependent and independent variables can be shown two types of relationship:

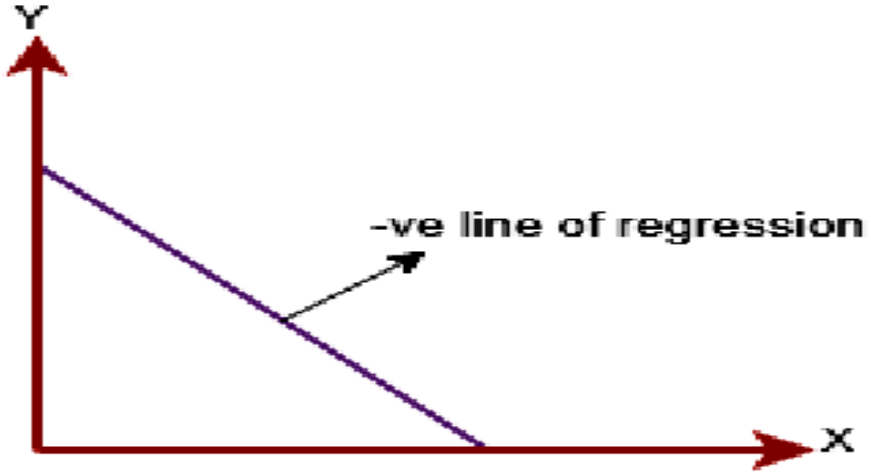
Positive Linear Relationship:

If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



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.

- If sum of squared error is taken as a metric to evaluate the model, then goal to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n} (actual\_output - predicted\_output) ** 2$$

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

$$b = \frac{10 \times 2289 - (80 \times 255)}{[10 \times 756 - (80)^2]} = 2.1466; \qquad a = \frac{255}{10} - 2.1466\frac{80}{10} = 8.3272$$

Linear Regression cont…

- The linear regression will thus be Predicted (Y) = 8.3272 + 2.1466 X
- 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 and a.
  - 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.

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Multiple Linear Regression

- Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression.
- Example:
  - Do age and intelligence quotient (IQ) scores predict grade point average (GPA)?
  - Do weight, height, and age explain the variance in cholesterol levels?
  - Do height, weight, age, and hours of exercise per week predict blood pressure?
- The formula for a multiple linear regression is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots\ldots\ldots\ldots\ldots\ldots + \beta_n x_n + e$$

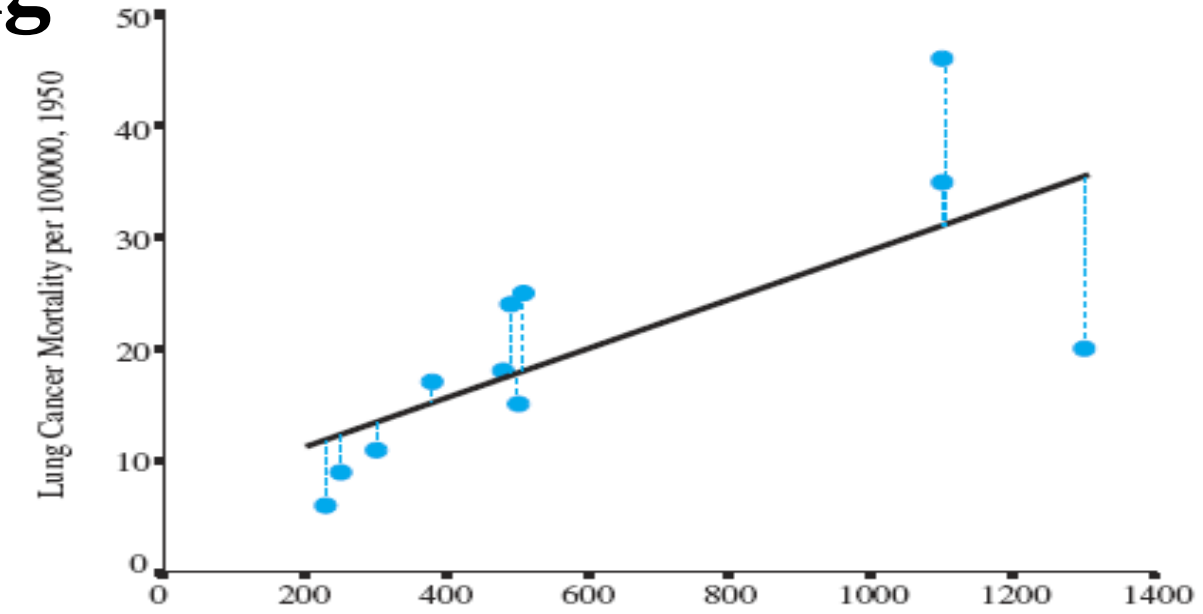
where, y = the predicted value of the dependent variable.

β_0 = the y-intercept (value of y when all other parameters are set to 0)
$\beta_1 x_1$= the regression coefficient (β_1) of the first independent variable (x_1)
$\beta_n x_n$= the regression coefficient (β_n) of the last independent variable (x_n)
e = model error

Regression Modeling

- A **simple regression model** (one independent variable) fits a regression *line* in 2-dimensional space

$$X \rightarrow Y$$



- A **multiple regression model with two explanatory** variables fits a regression plane in *3*-dimensional space
- However, it is difficult to visualize/graph 4+- relationship.

$$\begin{matrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{matrix} \rightarrow Y$$

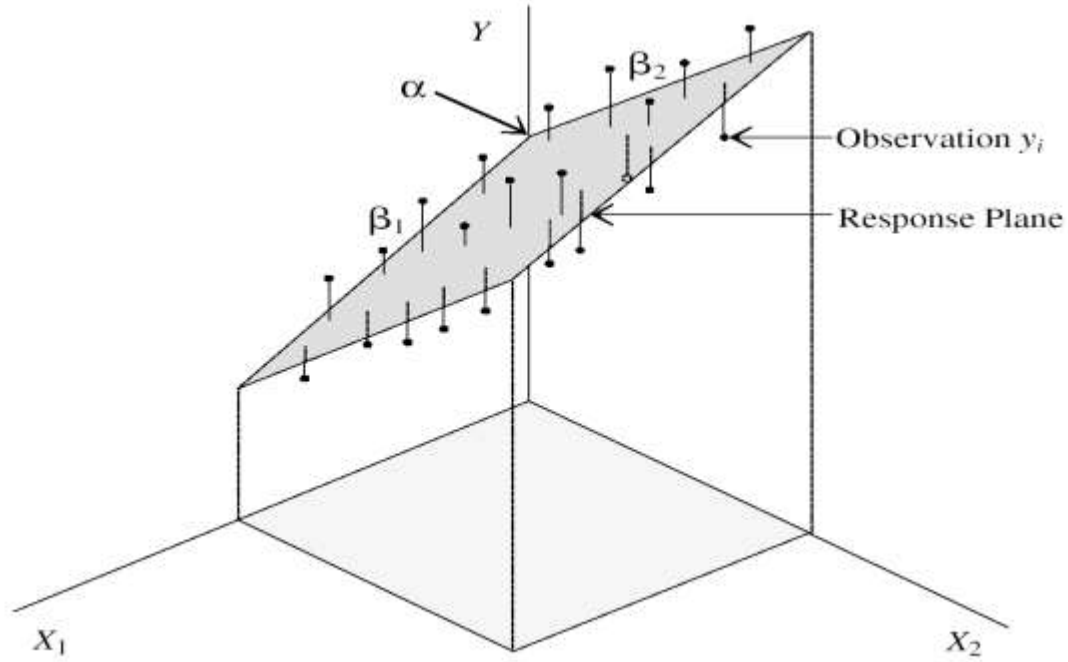


FIGURE 15.1 Three-dimensional response plane.

MLR Model: Basic Assumptions

- **Independence**: The data of any particular subject are independent of the data of all other subjects
- **Normality**: in the population, the data on the dependent variable are normally distributed for each of the possible combinations of the level of the X variables; each of the variables is normally distributed
- **Homoscedasticity**: In the population, the variances of the dependent variable for each of the possible combinations of the levels of the X variables are equal.
- **Linearity**: In the population, the relation between the dependent variable and the independent variable is linear when all the other independent variables are held constant.

Simple vs. Multiple Regression

- One dependent variable Y, predicted from one independent variable X
- One regression coefficient
- **r^2**: proportion of variation in dependent variable Y predictable from X

- One dependent variable Y predicted from a set of independent variables (X1, X2 ….Xk)
- One regression coefficient for each independent variable
- **R^2**: proportion of variation in dependent variable Y predictable by set of independent variables (X's)

Multiple Linear Regression [with Two Independent Variables]

- The formula for a multiple linear regression is:

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

  where, y = the predicted value of the dependent variable.

  β_0 = the y-intercept (value of y when all other parameters are set to 0)

  $\beta_1 x_1$= the regression coefficient (β_1) of the first independent variable (x_1)

  $\beta_2 x_2$= the regression coefficient (β_n) of the second independent variable (x_2)

  e = model error

- β_1 and β_2 is calculated as follows:

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

- β_0 is calculated as follows:

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

where $\bar{y} = \frac{\Sigma Y}{n}$ and $\bar{x}_i = \frac{\Sigma X_i}{n}$

Here, x_1, x_2 (small letter) are variance of regression and can be calculated from given table data X_1, X_2 and Y

Multiple Linear Regression [with Two Independent Variables]...

For β_1 and β_2 the value of variance x_1 x_2 and y can be calculated from the given table data X_1,X_2 and Y as follows. Here N is the number of observations

$$\beta_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\beta_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\sum x_1^2 = \sum X_1 X_1 - \frac{(\sum X_1)(\sum X_1)}{N}$$

$$\sum x_2^2 = \sum X_2 X_2 - \frac{(\sum X_2)(\sum X_2)}{N}$$

$$\sum x_1 y = \sum X_1 Y - \frac{(\sum X_1)(\sum Y)}{N}$$

$$\sum x_2 y = \sum X_2 Y - \frac{(\sum X_2)(\sum Y)}{N}$$

$$\sum x_1 x_2 = \sum X_1 X_2 - \frac{(\sum X_1)(\sum X_2)}{N}$$

Multiple Linear Regression Example [with Two Independent Variables]

Predict the value of Y given X_1 and X_2

SUBJECT	Y	X_1	X_2
1	-3.7	3	8
2	3.5	4	5
3	2.5	5	7
4	11.5	6	3
5	5.7	2	1
6	?	3	2

Multiple Linear Regression Example [with Two Independent Variables]...

$$y = a + b_1x_1 + b_2x_2$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

$$b_1 = \frac{(\sum x_2^2)(\sum x_1y) - (\sum x_1x_2)(\sum x_2y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2y) - (\sum x_1x_2)(\sum x_1y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
1	-3.7	3	8	9	64	24	-11.1	-29.6
2	3.5	4	5	16	25	20	14	17.5
3	2.5	5	7	25	49	35	12.5	17.5
4	11.5	6	3	36	9	18	69	34.5
5	5.7	2	1	4	1	2	11.4	5.7
∑	19.5	20	24	90	148	99	95.8	45.6

Multiple Linear Regression Example [with Two Independent Variables]...

From the table:

SUBJECT	Y	X_1	X_2	X_1X_1	X_2X_2	X_1X_2	X_1Y	X_2Y
∑	19.5	20	24	90	148	99	95.8	45.6

$$\sum x_1^2 = \sum X_1X_1 - \frac{(\sum X_1)(\sum X_1)}{N} = 90 - \frac{20*20}{5} = 10$$

$$\sum x_2^2 = \sum X_2X_2 - \frac{(\sum X_2)(\sum X_2)}{N} = 148 - \frac{24*24}{5} = 32.8$$

$$\sum x_1y = \sum X_1Y - \frac{(\sum X_1)(\sum Y)}{N} = 95.8 - \frac{20*19.5}{5} = 17.8$$

$$\sum x_2y = \sum X_2Y - \frac{(\sum X_2)(\sum Y)}{N} = 45.6 - \frac{24*19.5}{5} = -48$$

$$\sum x_1x_2 = \sum X_1X_2 - \frac{(\sum X_1)(\sum X_2)}{N} = 99 - \frac{20*24}{5} = 3$$

Multiple Linear Regression Example [with Two Independent Variables]...

The value of the regression coefficient b_1 and b_2, y-intercept a is as follows:

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{32.8 * 17.8 - 3 * (-48)}{10 * 32.8 - 3 * 3} = 2.28$$

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2} = \frac{10 * (-48) - 3 * 17.8}{10 * 32.8 - 3 * 3} = -1.67$$

$$a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2 = \frac{19.5}{5} - \frac{2.28 * 20}{5} - \frac{-1.67 * 24}{5} = 2.796$$

Finally, the regression model is: $\boldsymbol{Y = 2.796 + 2.28\, x_1 - 1.67 x_2}$

Now, given is $x_1 = 3$ and $x_2 = 2$, Y=? Y = 2.796 + 2.28 *3 -1.67 *2 = 6.296

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Non-Linear Regression

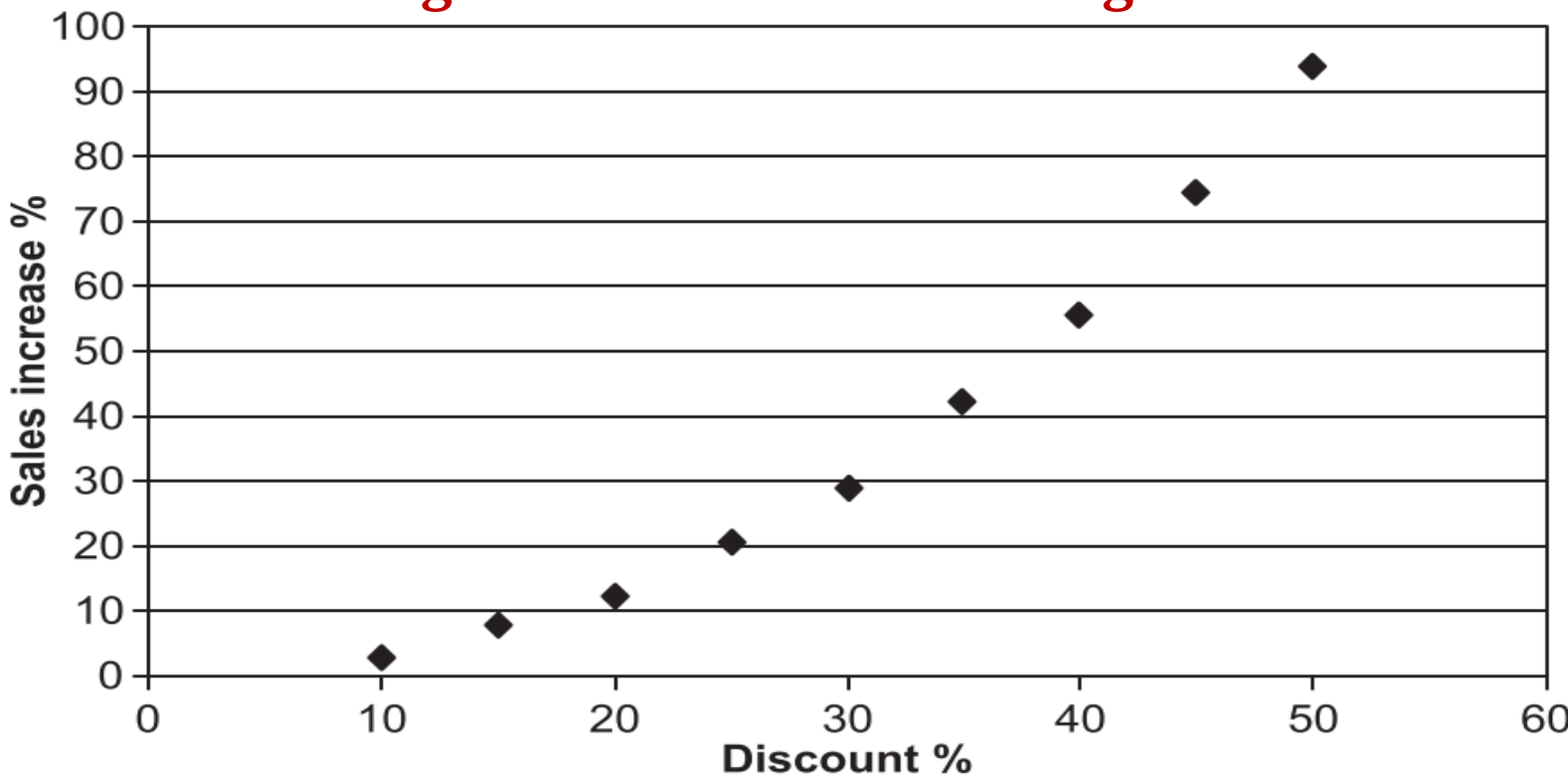
- In the case of linear and multiple linear regression, the dependent variable is linearly dependent on the independent variable(s). But, in several situations, the situation is no simple where the two variables might be related in a non-linear way.
- This may be the case where the results from the correlation analysis show no linear relationship but these variables might still be closely related.
- If the result of the data analysis show that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model.
- The non-linear data can beand handled in 2 ways:
  - Use of polynomial rather than linear regression model
  - Transform the data then use linear regression model.

Non-Linear Regression

- In the case of linear and multiple linear regression, the dependent variable is linearly dependent on the independent variable(s). But, in several situations, the situation is no simple where the two variables might be related in a non-linear way.

Product	Incr in sale % (Y)	Discount in %(X)
A	3.05	10
B	7.62	15
C	12.19	20
D	20.42	25
E	28.65	30
F	42.06	35
G	55.47	40
H	74.68	45
I	93.88	50

For the given data the scatter diagram looks like:



The value of r is 0.97 which is a very strong positive correlation, however the data value appears to form a slight curve.

Non-Linear Regression

- This is the case where the results from the correlation analysis shows no linear relationship but the variables are still be closely related.
- If the result of the data analysis show that there is a non-linear (also known as curvilinear) association between the two variables, then the need is to develop a non-linear regression model instead of optimizing linear model.
- The non-linear data can be handled in 2 ways:

A. **Intrinsically non-linear:** The models that can not be transformed into linear models => **Use of polynomial rather than linear regression model**

B. **Intrinsically Linear:** The model that can be transformed into linear models after applying some suitable transformation=> **Transform the data and then use linear regression model.**

Non-Linear Regression...

A. Intrinsically Non-Linear: [UNIVARIANT]

Use of polynomial rather than linear regression model

- If the data points are non-linear, not fit a linear regression (a straight line), it might be ideal for polynomial regression. Polynomial regression like linear regression uses the relationship between the variables x and y to find the best way to draw a curve (instead of line) through the data points.
- It is a form of linear regression in which the relationship between the independent variable x and the dependent variable y is modelled as as n^{th} degree polynomial.
- Polynomials are the equations that involve powers of the independent variables. A second degree (quadratic), third degree (cubic), and n degree polynomial functions (one variable x):

Second degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$

Third degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$

k degree: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots \dots + \beta_k x^k + e$

Where: β_0 is the intercept of the regression model

$\beta_1, \beta_2, \beta_3$ are the coefficient of the predictors.

Non-Linear Regression...

A. Intrinsically Non-Linear...

k degree polynomial regression model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots \dots + \beta_k x^k + e$

For n given data points: $y = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots \dots + \beta_k x_i^k + e \ (i=1, 2, ..., n)$

Matrix:

$$\begin{bmatrix} y1 \\ y2 \\ y3 \\ . \\ yn \end{bmatrix} = \begin{bmatrix} 1 & x1 & x1^2 & . & x1^k \\ 1 & x2 & x2^2 & . & x2^k \\ 1 & x3 & x3^2 & . & x3^k \\ . & . & . & . & . \\ 1 & xn & xn^2 & . & xn^k \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ . \\ e_n \end{bmatrix}$$

Order of the Model (k): The order k should be as low as possible (compared to n). The higher order polynomials (k>2) shoud be avoided. Hence, from above matrix with n equations and k (n>>k) unkowns, easily the value of unkowns can be derived.

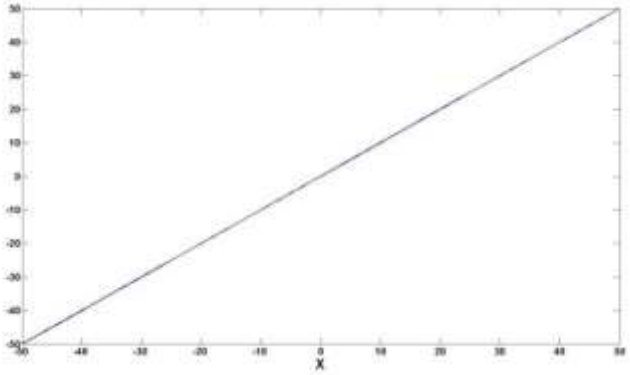
Note: In an extreme case, it is possible to pass a polynomial of order k=n-1 through n point so that a polynomial of sufficient high degree can found to provide a **good-fit** to the data

Non-Linear Regression...

A. Intrinsically Non-Linear...

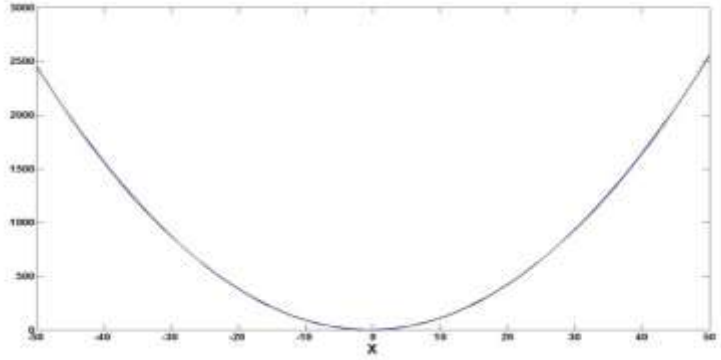
Linear: $Y' = A + BX$

Zero end



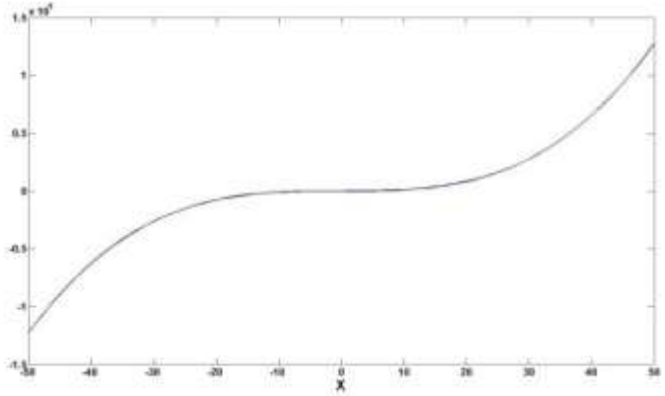
Quadratic: $Y' = A + BX + CX^2$

One Bend



Cubic: $Y' = A + BX + CX^2 + DX^3$

Two Bend



There is one less bend than the highest order in the polynomial model

How to find the right degree of the equation?

As we increase the degree in the model, it tends to increase the performance of the model. However, increasing the degrees of the model also increases the risk of over-fitting and under-fitting the data. So, one of the approach can be adopted:

- **Forward Selection:** This method increases the degree until it is significant enough to define the best possible model.
- **Backward Elimination:** This method decreases the degree until it is significant enough to define the best possible model.

Non-Linear Regression...

Data points on a Scatter gram

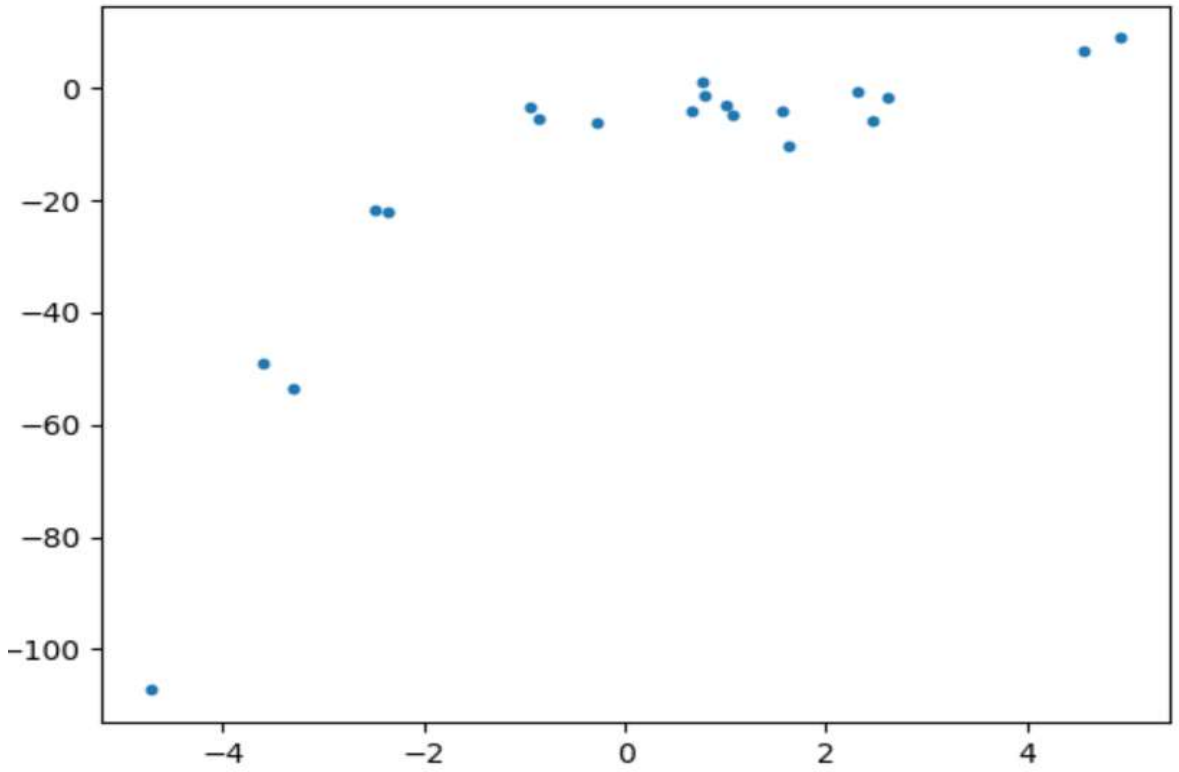


Chart-1

- Linear Regression Applied
- Under-fitting/ unable to capture data pattern
- R^2 : 0.638

$$Y = \theta_0 + \theta_1 x$$

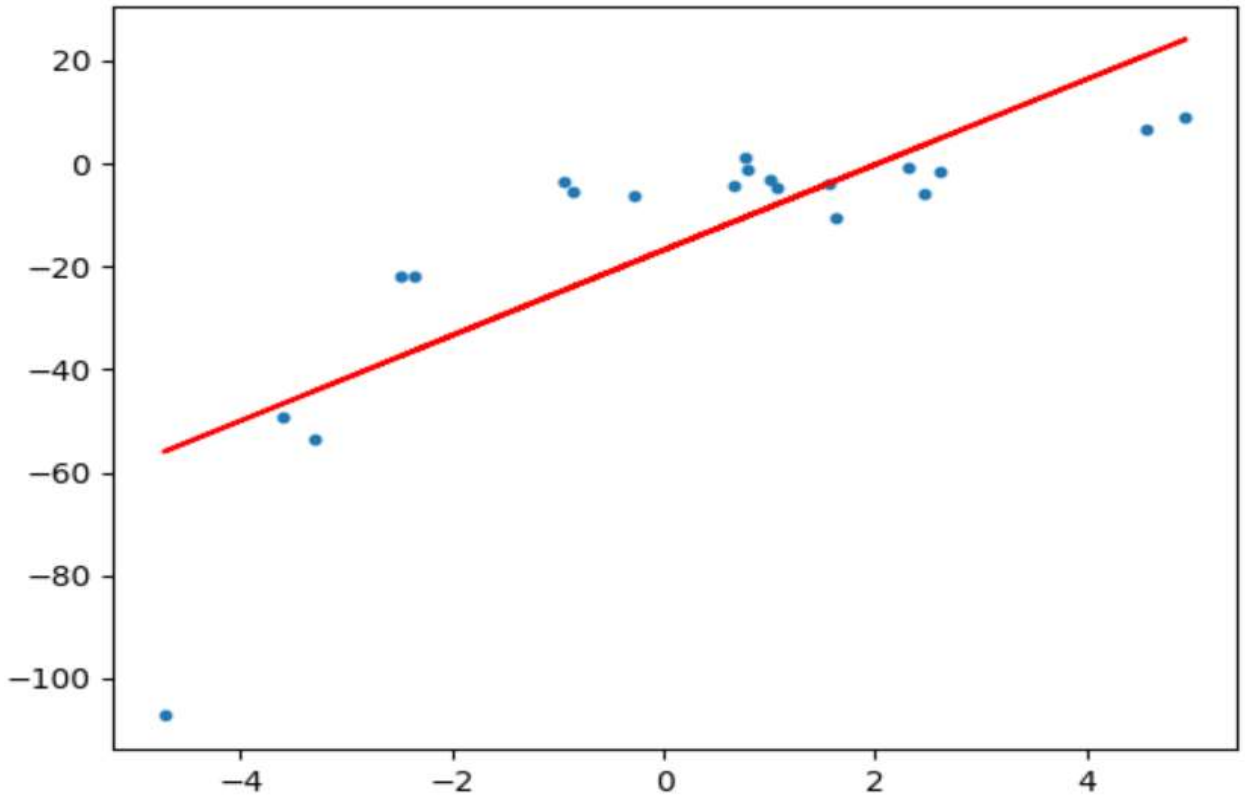


Chart-2

Non-Linear Regression...

- To overcome under-fitting, we need to increase the complexity of the model.

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2$$

- Model is non-linear (degree=2), but curve is quadratic.
- R^2 score = 0.85

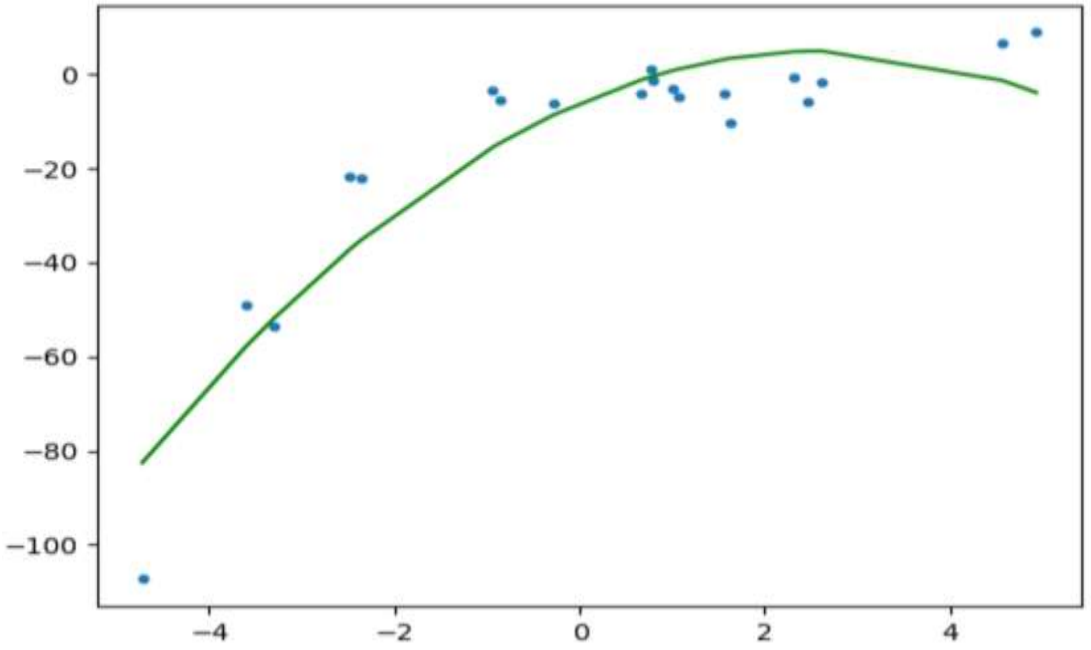


Chart-3

- If we try to fit a cubic curve (degree=3) to the dataset, we can see that it passes through more data points than the quadratic and the linear plots.
- R^2 score =0.98

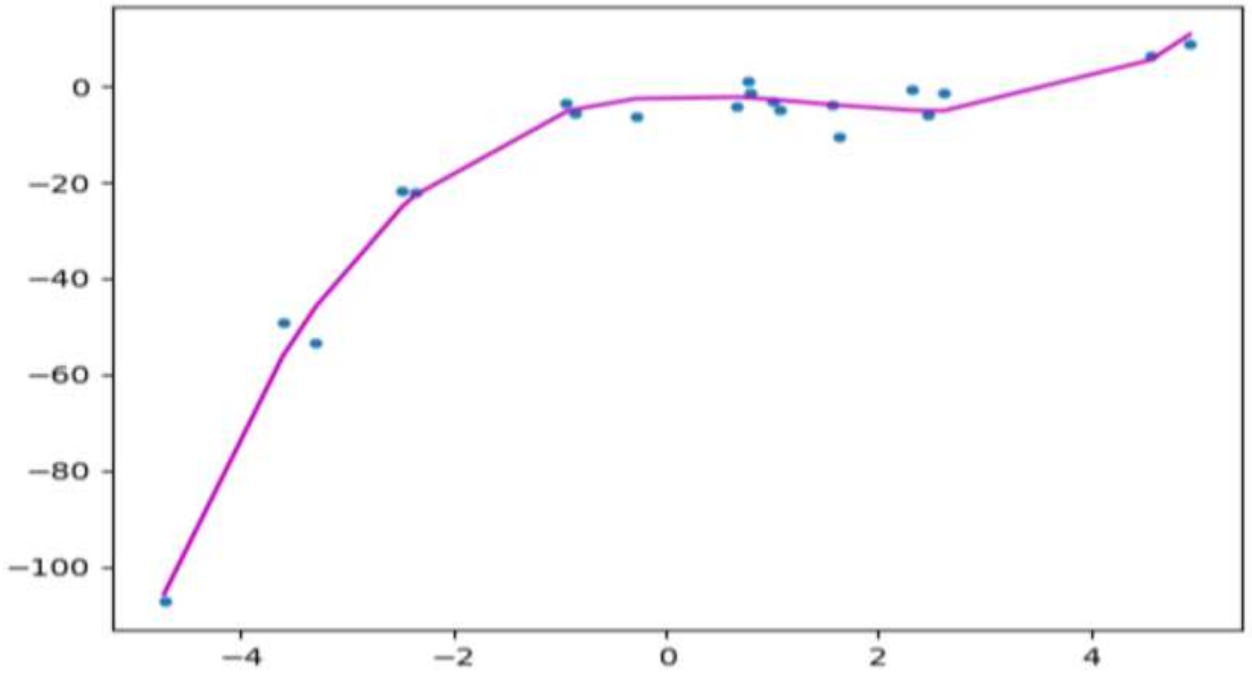


Chart-4

Non-Linear Regression...

➢A comparison of fitting linear, quadratic and cubic curves on the dataset

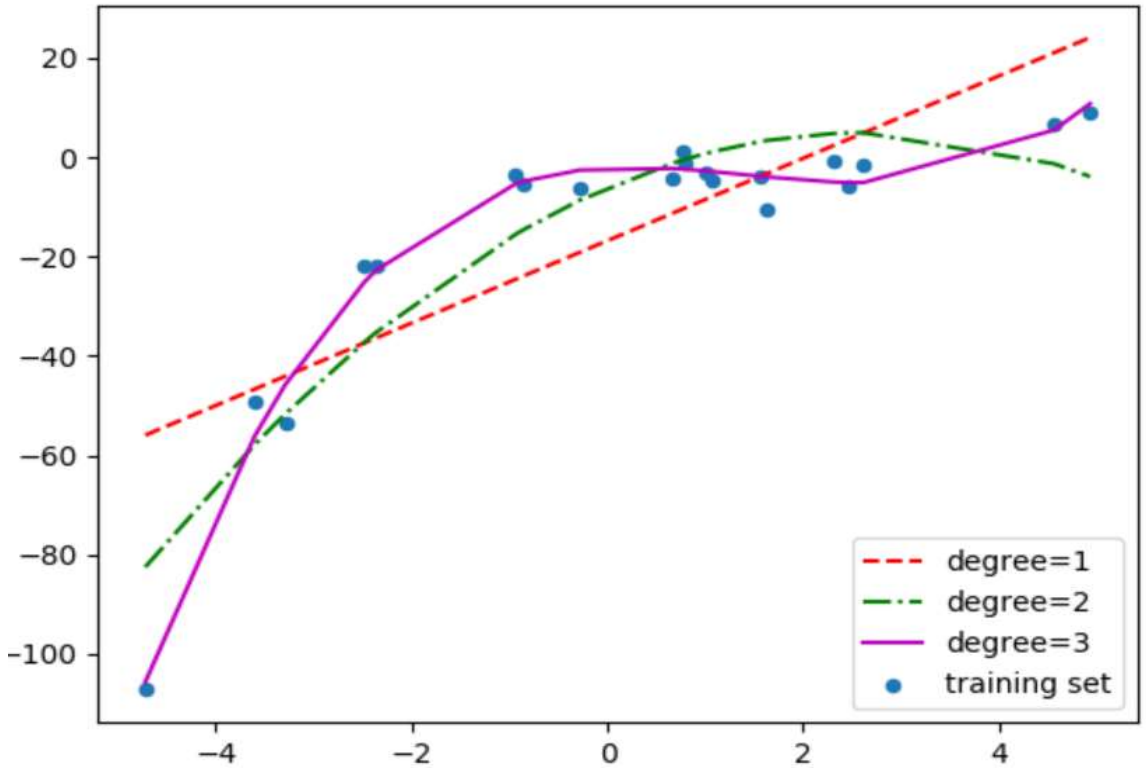


Chart-3

➢For degree to 20, we can see that the curve passes through more data points. A comparison of curves for degree 3 and 20 on the data set.

➢For degree=20, the model is also capturing the noise in the data.

➢This is an example of over-fitting.

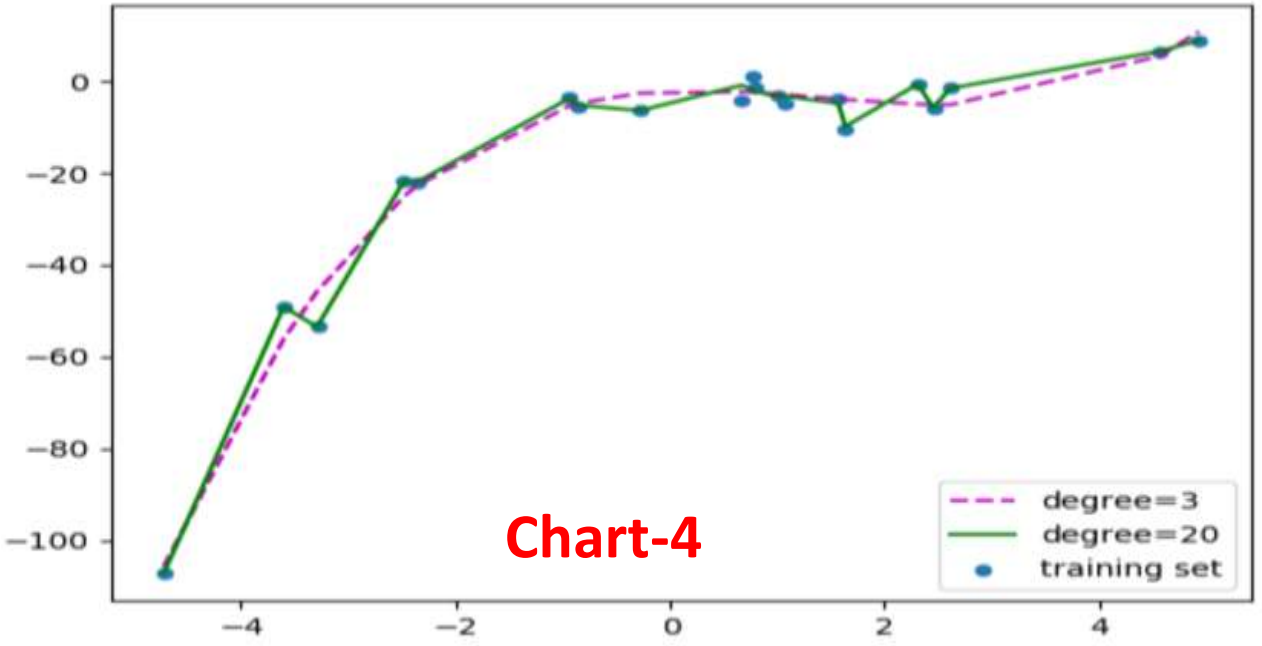
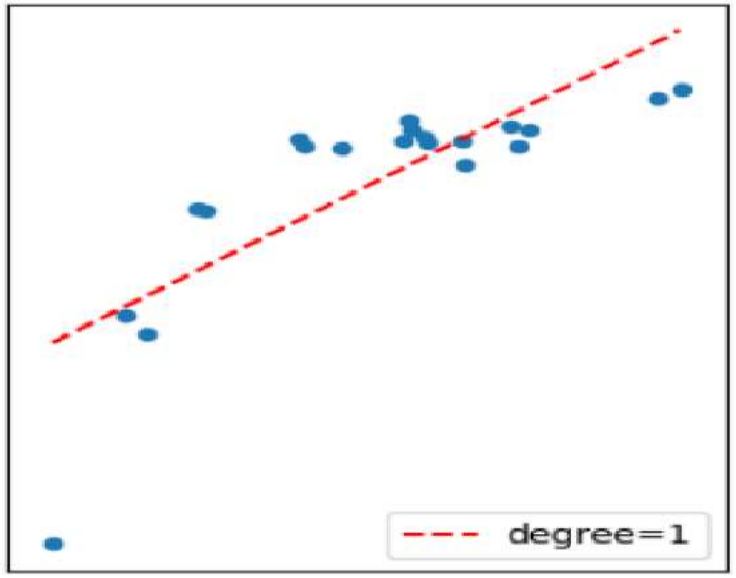
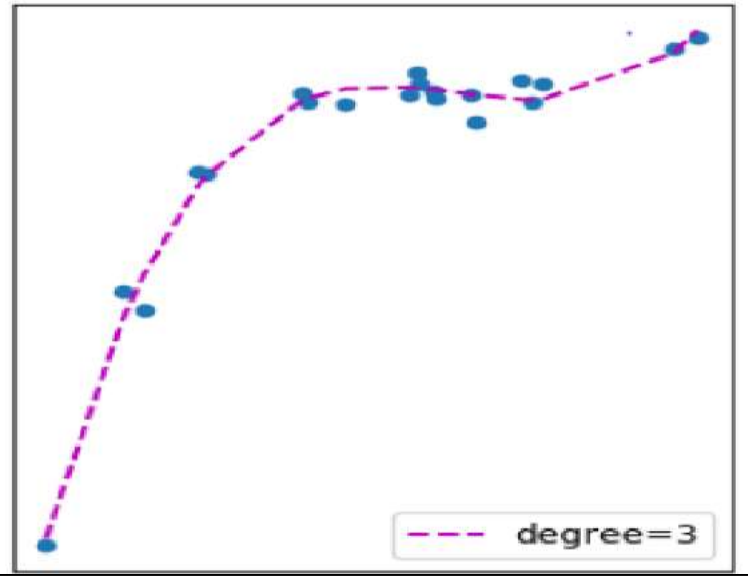
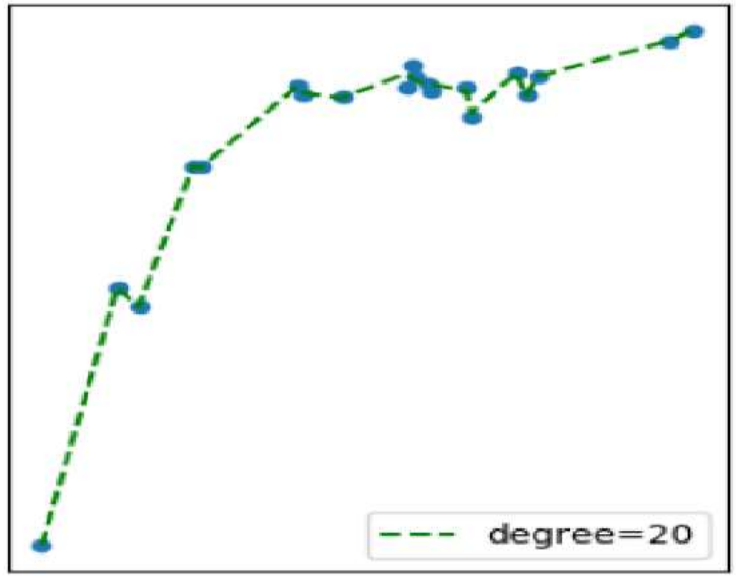


Chart-4

Non-Linear Regression...







Correct Fit, Low Bias, Low Variance

Under Fit, High Bias, Low Variance

Over Fit, Low Bias, High Variance

The Bias vs Variance trade-off

- **Bias** refers to the error due to the model's simplistic assumptions in fitting the data. A high bias means that the model is unable to capture the patterns in the data and this results in **under-fitting**.
- **Variance** refers to the error due to the complex model trying to fit the data. High variance means the model passes through most of the data points and it results in **over-fitting** the

Non-Linear Regression...

A. Intrinsically Non-Linear... [MULTIVARIANT]

- The techniques of fitting of the polynomial model in one variable can be extended to the fitting of polynomial models in two or more independent variables.
- A second-order polynomial is more used in practice, and its model with two independent variables is specified by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + e$$

- This is also termed as **response surface**. The methodology of response surface is used to fit such models and helps in designing an experiment.

Class work

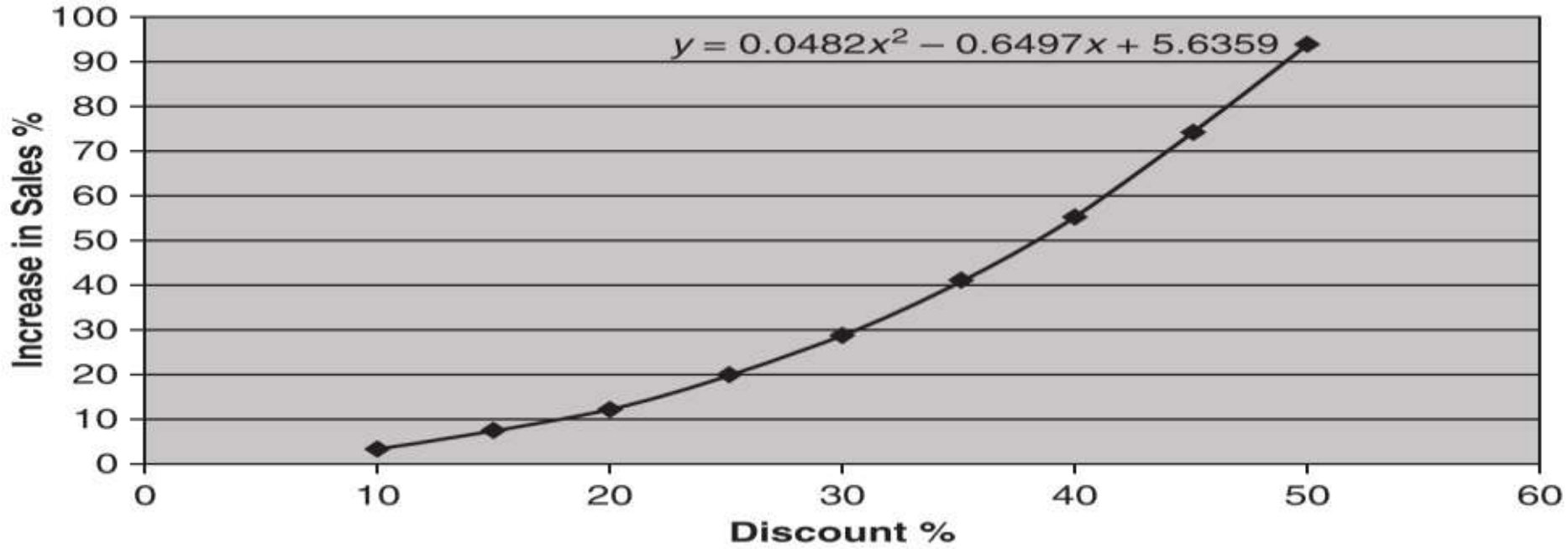
- Define the second-order polynomial model with two independent variables.
- Define the second-order polynomial model with three independent variables.
- Define the third-order polynomial model with two independent variables.

Non-Linear Regression...

A. Intrinsically Non-Linear...

A polynomial regression is regression that involves multiple powers of predictor(s).
So, **regression tools and diagnostics** can be applied to polynomial regression.

- The tools exists in software such as SAS, Excel or the language such as Python, R can estimate the value of coefficients of predictor such as β_0, β_1 etc and to fit a curve in a non-linear fashion for the given data.
- Following figure depicts the graph of increase in sale vs. discount.



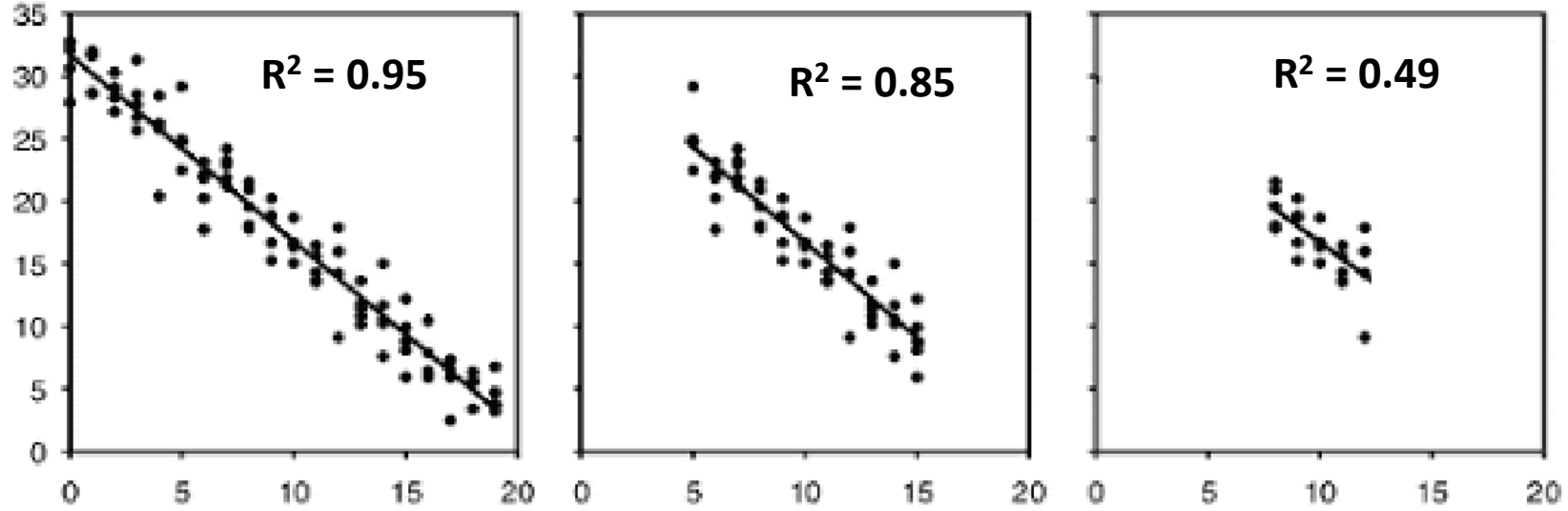
The predicted model is Y = 5.6359 – 0.6497 x + 0.0482 x^2

Non-Linear Regression...

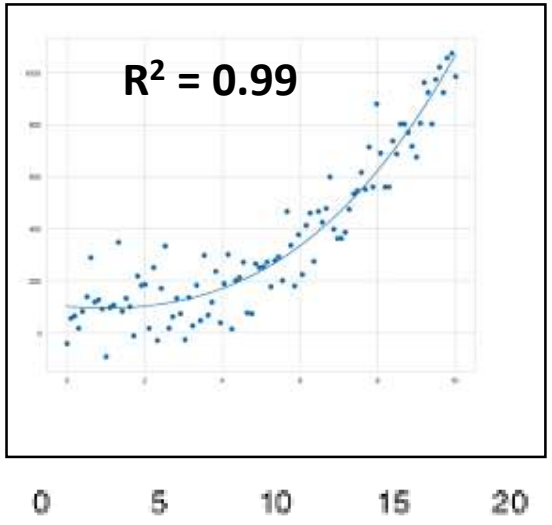
A. Intrinsically Non-Linear...

R^2 is a known coefficient of determination and it's value indicates how well the data fits into the developed model i.e. a line or curve.

Linear



Curve



Here, a value of 0.99 for R^2 indicates that a quadratic model is good fit for the data.

An **R^2 of 1** indicates that the regression model **perfectly fits** the data while an **R^2 of 0** indicate that model **does not fit** the data at all.

Non-Linear Regression...

A. Intrinsically Non-Linear...

- An **R^2 of 1** indicates that the regression model **perfectly fits** the data while an **R^2 of 0** indicate that model **does not fit the data at all**.
- An R^2 is calculated as follows:

Coefficient of Determination →
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Sum of Squares Total →
$$SST = \sum (y - \bar{y})^2$$

Sum of Squares Regression →
$$SSR = \sum (y' - \bar{y'})^2$$

Sum of Squares Error →
$$SSE = \sum (y - y')^2$$

where Y & Y': Actual value of and predicted value of individual Y

$\bar{Y}$ & $\bar{Y'}$: Mean of actual and predicted value

Non-Linear Regression...

A. Intrinsically Linear...

- Another preferable way to perform non-linear regression is to try to transform the data in order to make the relationship between the two variables more linear and then use a regression model rather than a polynomial one. Transformations aim to make a non-linear relationship between two variables more linear so that it can be described by a linear regression model.
- Three most popular transformations are the:
  - Square root ($\sqrt{X}$)
  - Logarithm (log X)
  - Negative reciprocal (- 1/ X)

Non-Linear Regression...

A. Intrinsically Linear...

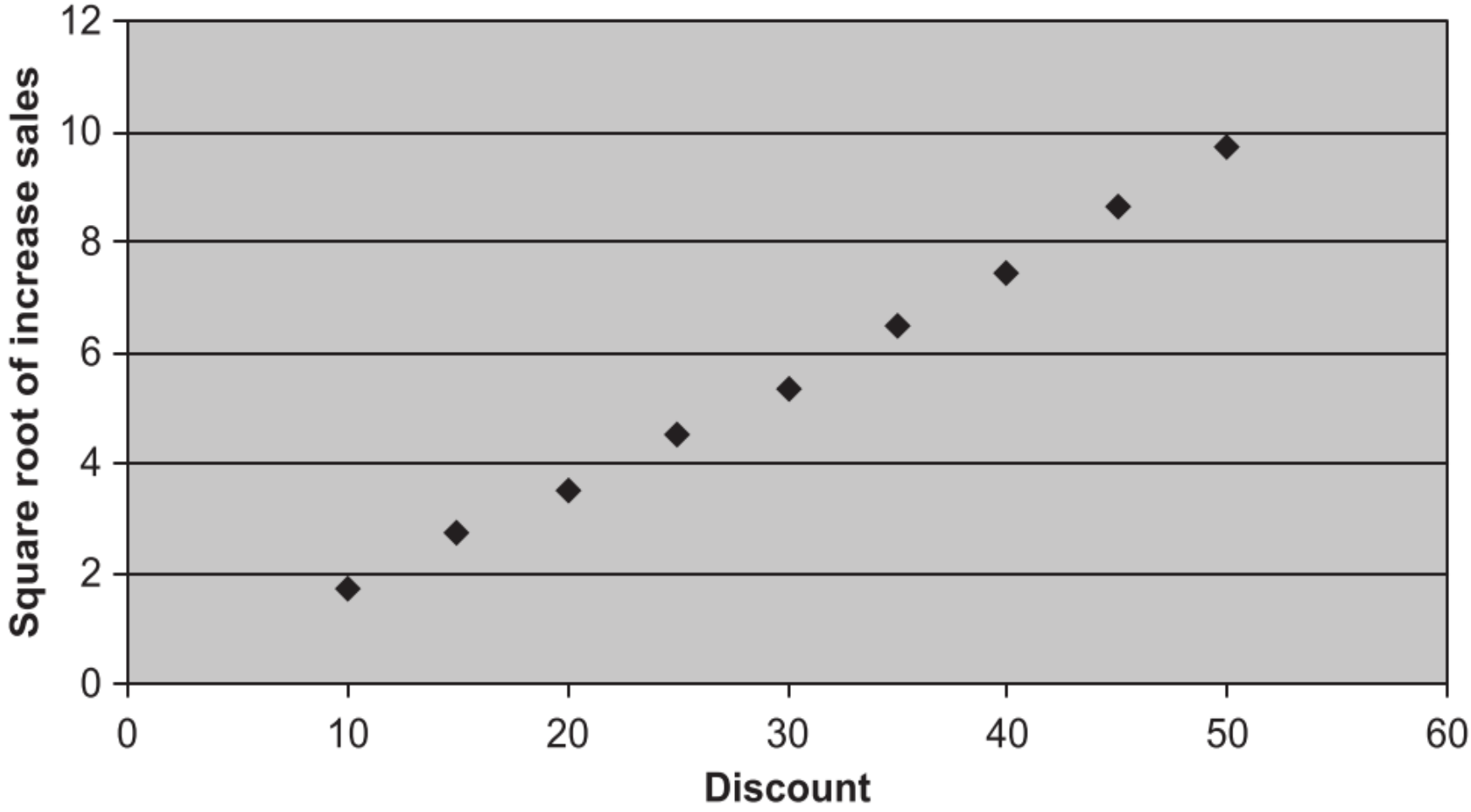
Application of Square root (√Y)

Product	Discount in %(X)	Increase in sale in% (Y)	SQRT (Y)
A	10	3.05	√3.05 = 1.75
B	15	7.62	√7.62 = 2.76
C	20	12.19	√12.19 = 3.49
D	25	20.42	√20.42 = 4.52
E	30	28.65	√28.65 = 5.35
F	35	42.06	√42.06 = 6.49
G	40	55.47	√55.47 = 7.45
H	45	74.68	√74.68 = 8.64
I	50	93.88	√93.88 = 9.69

Non-Linear Regression...

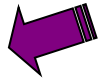
A. Intrinsically Linear...

Square root of transformation



In the similar fashion, Logarithm and negative reciprocal techniques can be applied to the dependent variable followed up by the application of linear regression model.

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression 
- **Time Series Analysis**
- **Performance analysis**
  - RMSE
  - MAPE

Logistic Regression

- In linear regression, the Y variable is always a continuous variable.
- When Y variable is categorical, linear regression model cannot be used.
- **So what one would do when the Y is a categorical variable with 2 classes?**
- Logistic regression can be used to model and solve such problems, also called as **binary classification problems**.
- Logistic Regression is one of the most commonly used Machine Learning algorithms that is used to model a binary variable that takes only 2 values – 0 and 1.
- The objective of Logistic Regression is to develop a mathematical equation that can give us a score in the range of 0 to 1.

Logistic Regression

Example

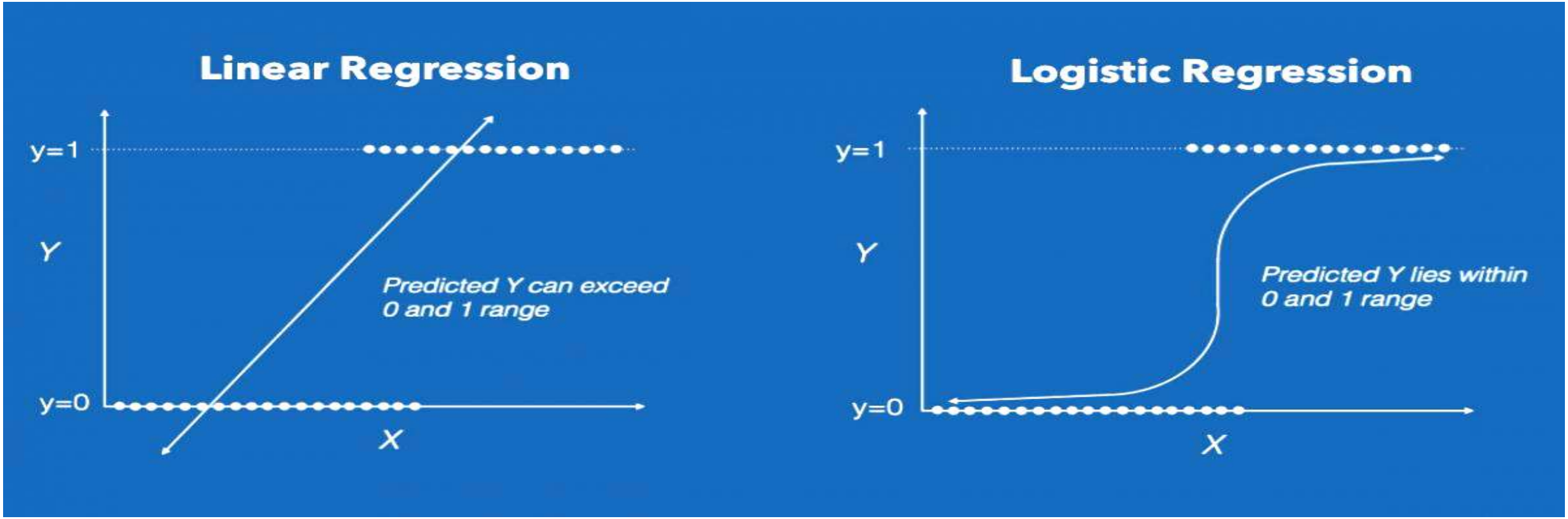
- **Spam Detection:** Spam detection is a binary classification problem where we are given an email and we need to classify whether or not it is spam. If the email is spam, we label it 1; if it is not spam, we label it 0.
- **Tumour Prediction:** A Logistic Regression classifier may be used to identify whether a tumour is malignant or if it is benign. Several medical imaging techniques are used to extract various features of tumours. For instance, the size of the tumour, the affected body area, etc. These features are then fed to a Logistic Regression classifier to identify if the tumour is malignant or if it is benign.
- **Health :** Predicting if a given mass of tissue is benign or malignant
- **Marketing :** Predicting if a given user will buy an insurance product or not
- **Banking :** Predicting if a customer will default on a loan.

- **Dichotomous categorical response variable *Y***
  - **e.g. Success/Failure, Remission/No Remission, Survived/Died, CHD/No CHD, Low Birth Weight/Normal Birth Weight, etc…**

Logistic Regression

- Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications.
- Logistic Regression is used when the dependent variable (target) is categorical. For example:
  - To predict whether an email is spam (1) or not (0). If the model infers a value of 0.932 on a particular email message, it implies a 93.2% probability that the email message is spam. The model predicts the email message is spam 93.2% of the time and the remaining 6.8% will not.
  - Whether the tumor is malignant (1) or not (0)
- There are 3 types of Logistic Regression
  - **Binary Logistic Regression:** The categorical response has only two 2 possible outcomes. Example: Spam or Not.
  - **Multinomial Logistic Regression:** Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
  - **Ordinal Logistic Regression:** Three or more categories with ordering. Example: Movie rating from 1 to 5.

Logistic Regression



- When the response variable has only 2 possible values, it is desirable to have a model that predicts the value either as 0 or 1 or as a probability score that ranges between 0 and 1.
- Linear regression does not have this capability. Because, If you use linear regression to model a binary response variable, the resulting model may not restrict the predicted Y values within 0 and 1.

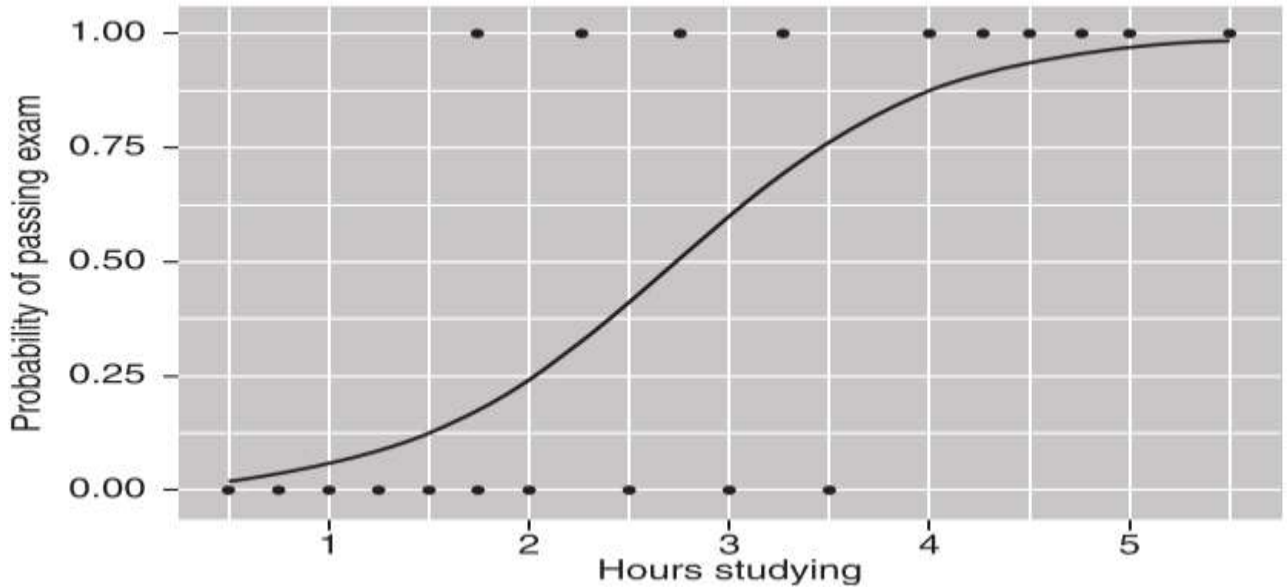
Logistic Regression ...

- It measures the relationship between the categorical dependent variable and one or more independent variables by **estimating probabilities using a logistic function** which is the cumulative logistic distribution.
- Since the predicted values are probabilities and therefore are restricted to (0, 1), a logistic regression model **only predicts the probability of particular outcome** given the values of the existing data.
- **Example:** A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam? The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used. The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1

Logistic Regression

- In logistic regression, we don't directly fit a straight line to the data like in linear regression. Instead, we fit a S shaped curve, called **sigmoid** or **logistic regression curve**. A logistic regression curve showing probability of passing an exam versus hours studying is shown below.
- Y-axis goes from 0 to 1. This is because the sigmoid function always takes as maximum (i.e. 1) and minimum (i.e. 0), and this fits very well to the goal of classifying samples in two different categories (fail or pass).
- The sigmoid function is $\text{sigmoid}(x) = 1 / (1 + e^{-x})$ where x is the weighted sum of independent variable i.e. $x = \beta_0 + \beta_1 x_i$ where i is the individual independent variable instance.



Logistic Regression ...

Consider a model with one predictor X_1, and one binary response variable Y, which we denote p = P(Y = 1 | X_1 = x), where p is the probability of success. p should meet criteria: (i) it must always be positive, (ii) it must always be less than equals to 1.

We assume a linear relationship between the independent variable and the logit of the event i.e. Y = 1. In statistics, the logit is the logarithm of the **odds** i.e. p / (1-p). This linear relationship can be written in the following mathematical form (where ℓ is the logit, b is the base of the logarithm, and β is the parameter of the model.

$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

The odds can be recovered by exponentiation of the logit:

$$\frac{p}{1-p} = b^{\beta_0+\beta_1 x_1} \rightarrow p = \frac{b^{\beta_0+\beta_1 x_1}}{b^{\beta_0+\beta_1 x_1}+1} = \frac{1}{1+b^{-(\beta_0+\beta_1 x_1)}} = S_b(\beta_0+\beta_1 x_1)$$

Where S_b is the sigmoid function with base b. However in some cases it can be easier to communicate results by working in base 2, base 10, or exponential constant e.

In reference to the students example, solving the equation with software tool and considering base as e, the coefficient is β_0 = -4.0777 and β_1= 1.5046

Logistic Regression cont…

- For example, for a student who studies 2 hours, entering the value Hours = 2 in the equation gives the estimated probability of passing the exam of 0.26.

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 2 - 4.0777))} = 0.26$$

- Similarly, for a student who studies 4 hours, the estimated probability of passing the exam is 0.87:

$$\text{Probability of passing exam} = \frac{1}{1 + \exp(-(1.5046 \cdot 4 - 4.0777))} = 0.87$$

- Following table shows the probability of passing the exam for several values of hours studying.

Hours of study	Probability of passing the exam
1	0.07
2	0.26
3	0.61
5	0.97

Chapter 3: **Data Analytics**

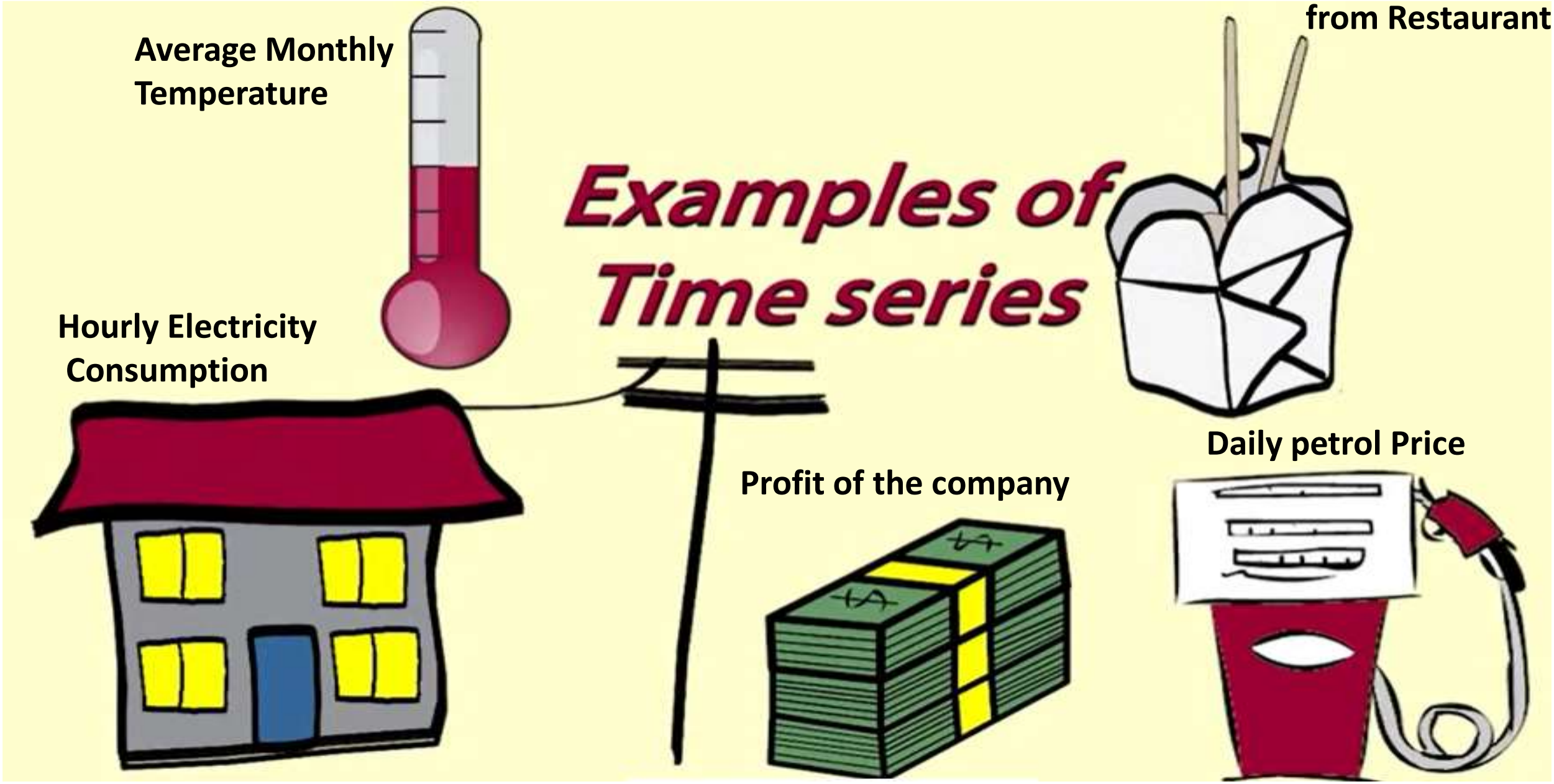
- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis** 
- **Performance analysis**
  - RMSE
  - MAPE

Time Series Analysis

- Whether to predict the trend in financial markets or electricity consumption, time is an important factor that must be considered in the model. For example, it would be interesting to forecast at what hour during the day it is going to be a peak consumption in electricity, such as to adjust the price or the production of electricity.

- **A time series is simply a series of data points ordered in time. In a time series, time is often the independent variable and the goal is usually to make a forecast for the future.** As the name suggests, it involves working on time (years, weeks, days, hours, minutes) based data, to derive hidden insights to make informed decision making.

Time Series Analysis Examples



Time Series Analysis Examples

Field	Example Topics
Economics	Gross Domestic Product (GDP), Consumer Price Index (CPI), and unemployment rates
Medicine	Blood pressure tracking, weight tracking, cholesterol measurements, heart rate monitoring
Physical sciences	Global temperatures, monthly sunspot observations, pollution levels.
Social Sciences	Birth rates, population, migration data, political indicators
Epidemiology	Disease rates, mortality rates, mosquito populations

Time Series Model

- A time series is a sequential set of data points, measured typically at successive times. It is mathematically defined as a set of vectors x(t) where t = 0,1, 2… where t represents the time elapsed. The variable x(t) is treated as random variable.
- A time series model generally reflect the fact that observations close together in time which are closely related than the observations further apart.
- The data shown below represent the weekly demand of some product. The model uses x to indicate an observation and t to represent the index of the time period. The data from 1 to t is: x_1, x_2, … x_t. Time series considering 25 periods is shown below.

Weekly demand

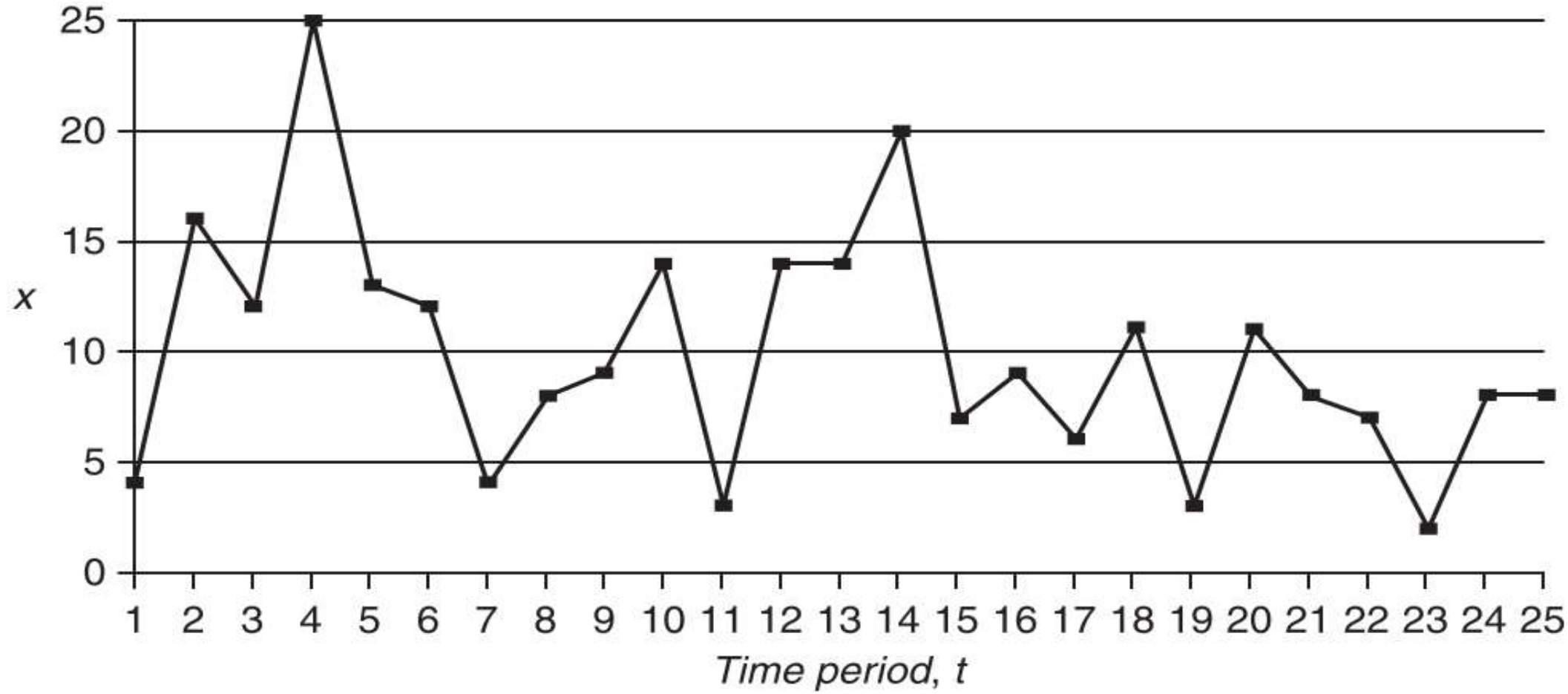
Time	Observations									
1 – 10	4	16	12	25	13	12	4	8	9	14
11 – 20	3	14	14	20	7	9	6	11	3	11
21 – 30	8	7	2	8	8	10	7	16	9	4

Time Series Model

Weekly demand

Time	Observations									
1 – 10	4	16	12	25	13	12	4	8	9	14
11 – 20	3	14	14	20	7	9	6	11	3	11
21 – 30	8	7	2	8	8	10	7	16	9	4

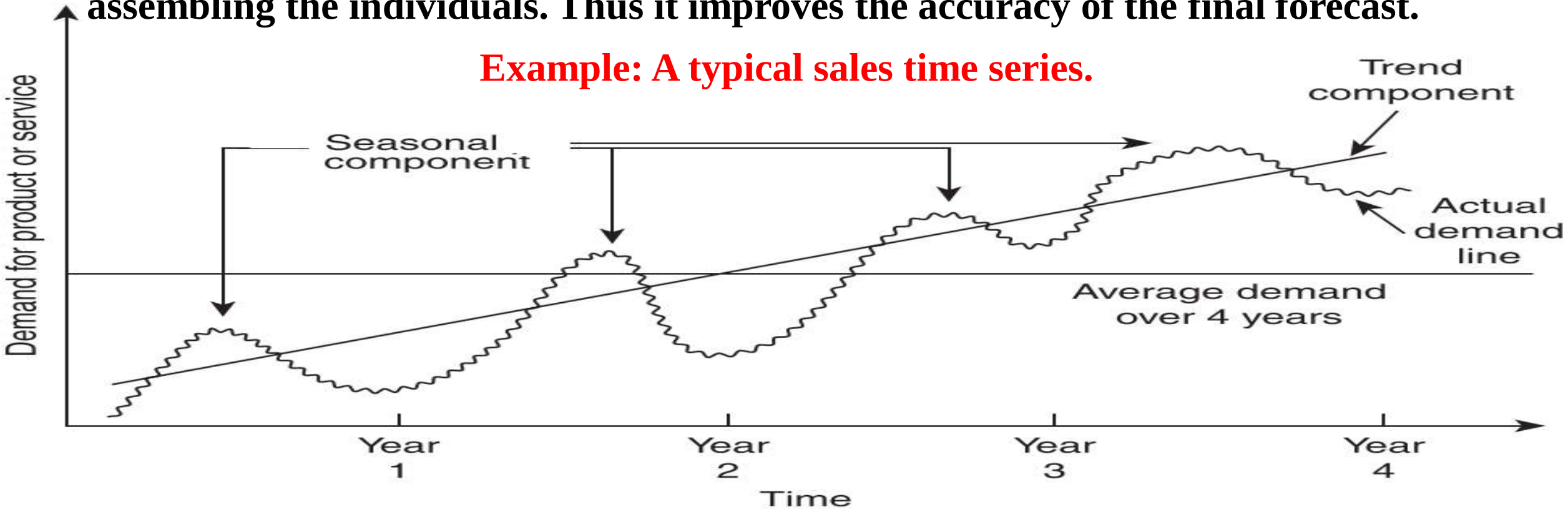
Time series of weekly demand - 25 periods



Time Series Model...

- Any time series is **composition of many individual component** times series. Some of these components are predictable whereas other components may be almost random which can be difficult to predict.
- This calls for **the decomposition methods that will generate individual component** series from the original series. Decomposing a series into such components **enable to analyze the behaviour of each component and then conclude the Forecast by assembling the individuals. Thus it improves the accuracy of the final forecast.**

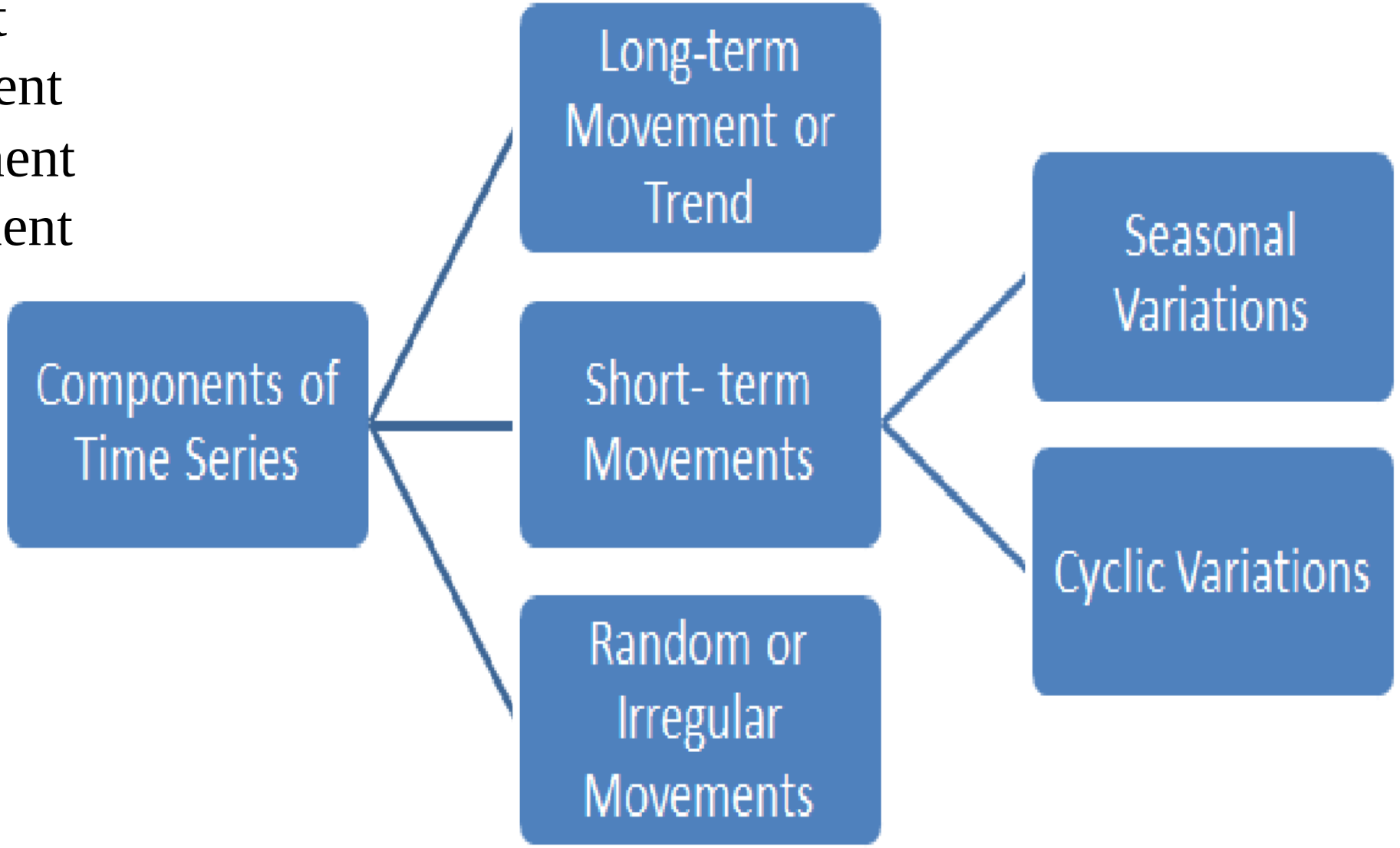
Example: A typical sales time series.



Time Series Model Component

Time series models are characterized of four components:

- Trend component
- Cyclical component
- Seasonal component
- Irregular component



Time Series Model Component contd...

I. Trend component

- The trend shows the general tendency of the data to increase or decrease during a long period of time. A trend is a smooth, general, long-term, average tendency.
- It is not always necessary that the increase or decrease is in the same direction throughout the given period of time.
- It is observable that the tendencies may increase, decrease or are stable in different sections of time. But the overall trend must be **upward, downward or stable.**
- **The population, agricultural production, items manufactured, number of births and deaths, number of industry or any factory, number of schools or colleges are some of its example showing some kind of tendencies of movement.**

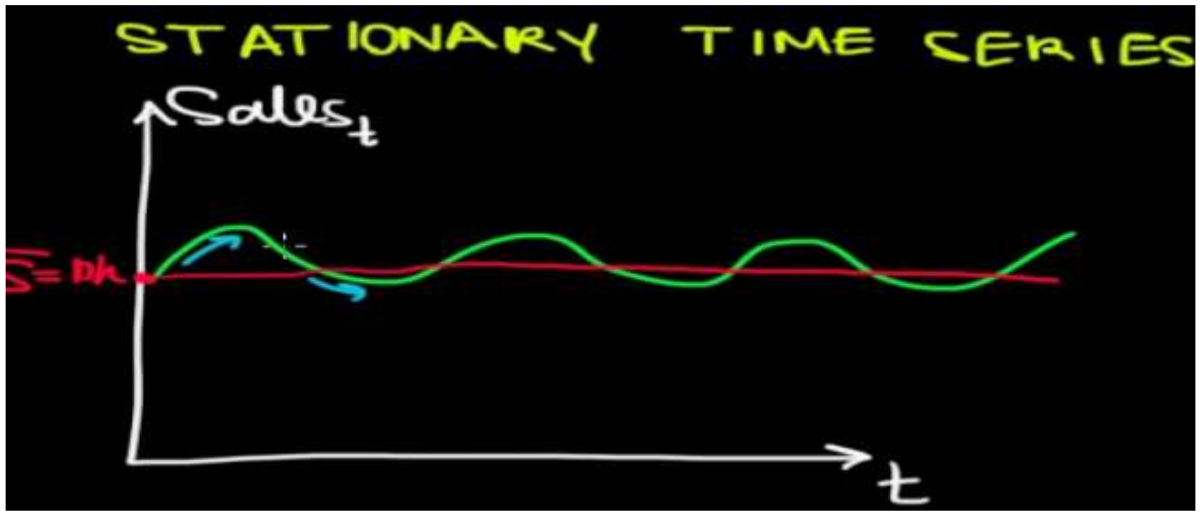
Time Series Model Component contd...

I. Trend component

Downword

Upward

Stable



Time Series Model Component contd...

II. Seasonal component

- These are the rhythmic forces which operate in a regular and periodic manner over a span of less than a year. They have the same or almost the same pattern during a period of 12 months. This variation will be present in a time series if the data are recorded hourly, daily, weekly, quarterly, or monthly.
- These variations come into play either because of the **natural forces** or **person-made** conventions.

Natural forces

Production of crops depends on seasons, the sale of umbrella and raincoats in the rainy season, and the sale of electric fans and A.C. shoots up in summer seasons.

Person-made

Some festivals, customs, habits, fashions, and some occasions like marriage They recur themselves year after year. An upswing in a season should not be taken as an indicator of better business conditions.

Time Series Model Component cont…

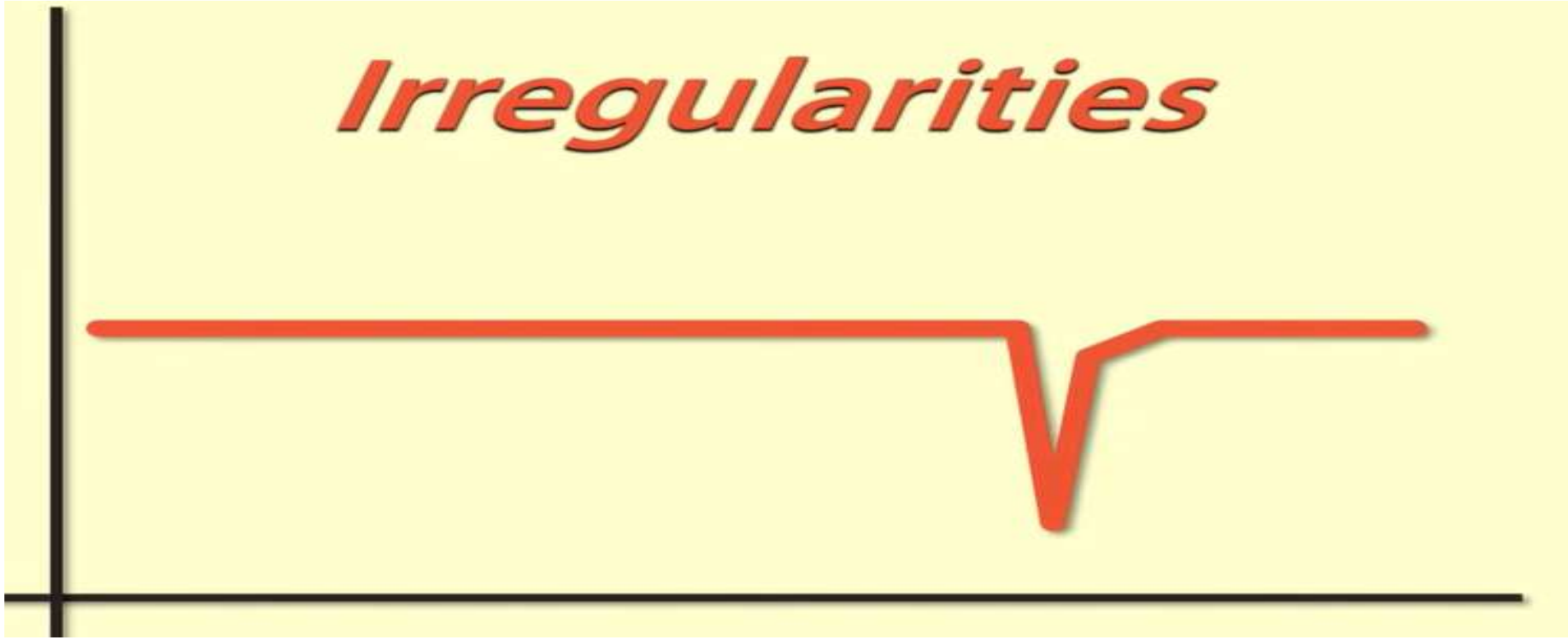
III. Cyclical component

- The variations in a time series which operate themselves over a span of more than one year are the cyclic variations. This oscillatory movement has a period of oscillation of more than a year. One complete period is a cycle. This cyclic movement is sometimes called the 'Business Cycle'.
- '**Business Cycle'**: It is a four-phase cycle comprising of the phases of **prosperity, recession, depression, and recovery**. The cyclic variation may be regular but not periodic. The upswings and the downswings in business depend upon the joint nature of the economic forces and the interaction between them.

Time Series Model Component contd...

IV. Irregular component

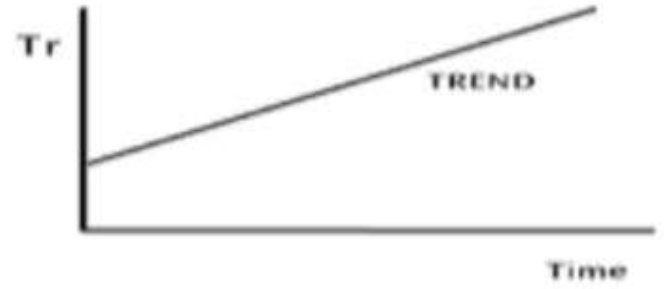
- They are not regular variations and are purely random or irregular. These fluctuations are unforeseen, uncontrollable, unpredictable, and are erratic. These forces are earthquakes, wars, flood, famines, and any other disasters.

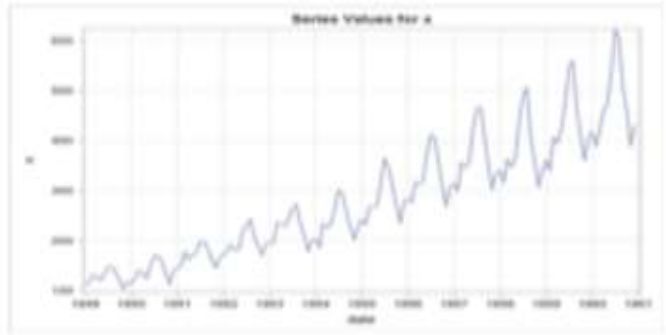


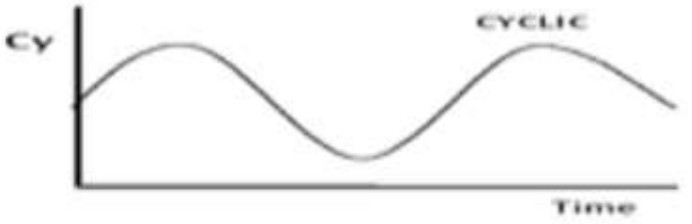
Time Series Model Component contd...

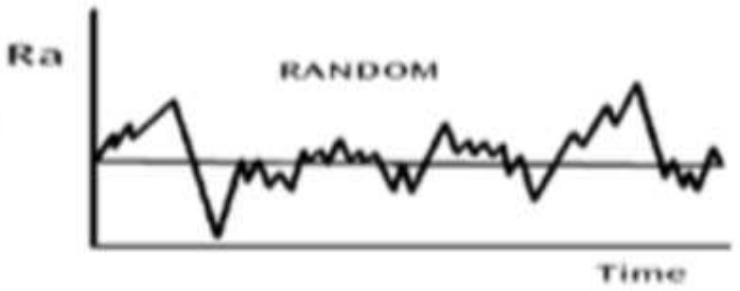
Time series can be decomposed into four components:

- **Trend (T):** *The trend is the long term pattern of a time series. A trend can be positive or negative depending on whether the time series exhibits an increasing long term pattern or a decreasing long term pattern.*
- **Seasonal (S):** *Seasonality occurs when the time series exhibits regular fluctuations during the same month (or months) every year, or during the same quarter every year. For instance, retail sales peak during the month of December.*
- **Cyclical (C):** *Any pattern showing an up and down movement around a given trend is identified as a cyclical pattern. The duration of a cycle depends on the type of business or industry being analyzed.*
- **Irregular(I):** *This component is unpredictable. Every time series has some unpredictable component that makes it a random variable. In prediction, the objective is to "model" all the components to the point that the only component that remains unexplained is the random component.*



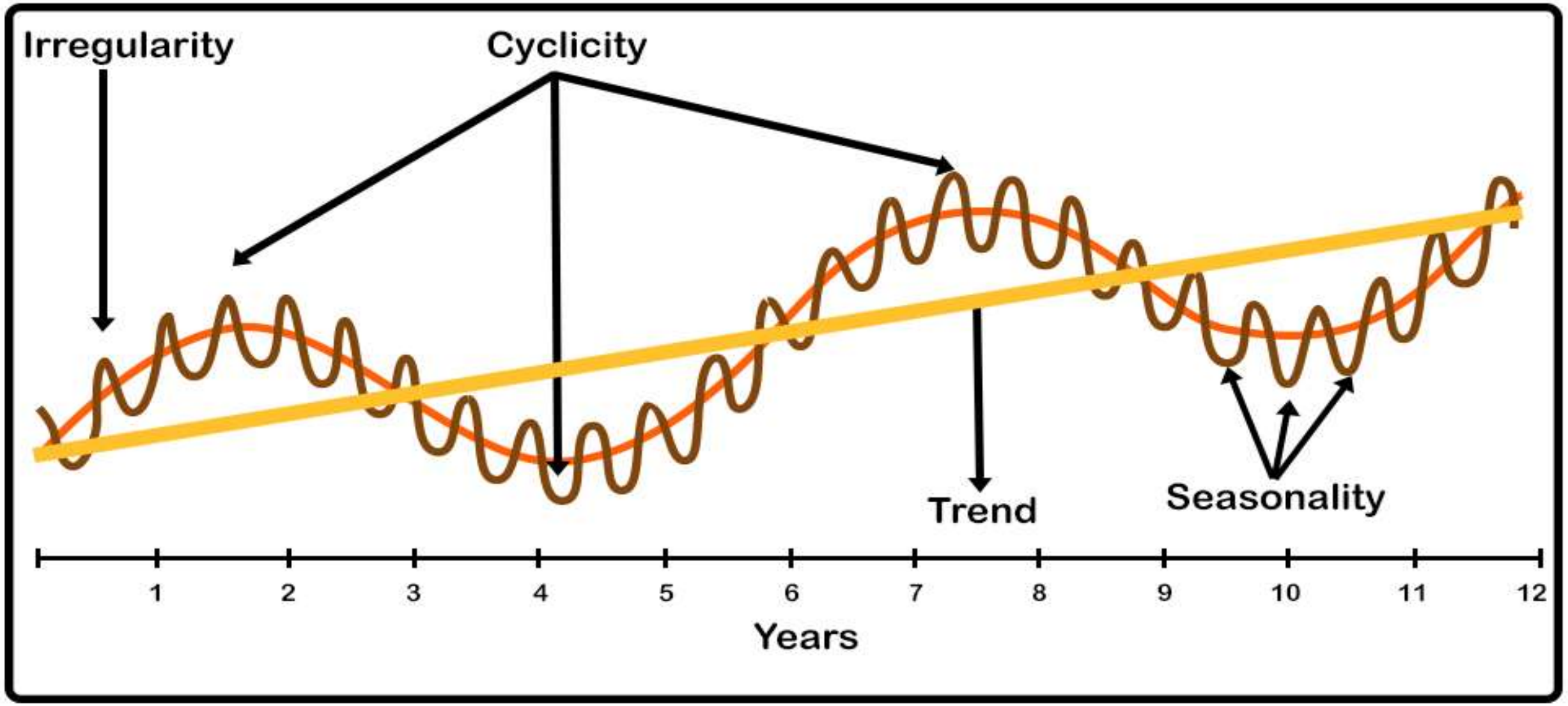






Time Series Model Component cont…

Pictorial depiction of different component



Decomposition Model

- The decomposition model assumes that sales are affected by four factors:
  - the general trend in the data, general economic cycles, seasonality, and irregular or random occurrences.
- Owing to this, the decomposition of time series is a statistical task that deconstructs a time series into several components, each representing one of the categories of four patterns.
- To forecast a time series using a decomposition model, the future values are calculated for each separate component and then added back together to obtain a prediction.
- Mathematical representation of the decomposition approach is $Y_t = f(T_t, S_t, C_t, I_t)$ where Y_t is the time series value at time t. T_t, S_t, C_t, and I_t are the trend, seasonal, cyclic and irregular component value at time t respectively.

Decomposition Model

- There are 3 types of decomposition model:
  - Additive model
  - Multiplicative model
  - Mixed Model

Additive model

- According to this model, a time series is expressed as $Y_t = T_t + S_t + C_t + I_t$
- The model is appropriate when the amplitude of both the **seasonal and irregular variations do not change as the level of trend rises or falls**.
- Assumption: **all four components** of the time series act **independently.**

Multiplicative model

- According to this model, a time series is expressed as $Y_t = T_t * S_t * C_t * I_t$
- The model is appropriate when the **amplitude of both the seasonal and irregular variations increase as the level of trend rises**.
- Assumption: **various components operate proportionately** to each other.

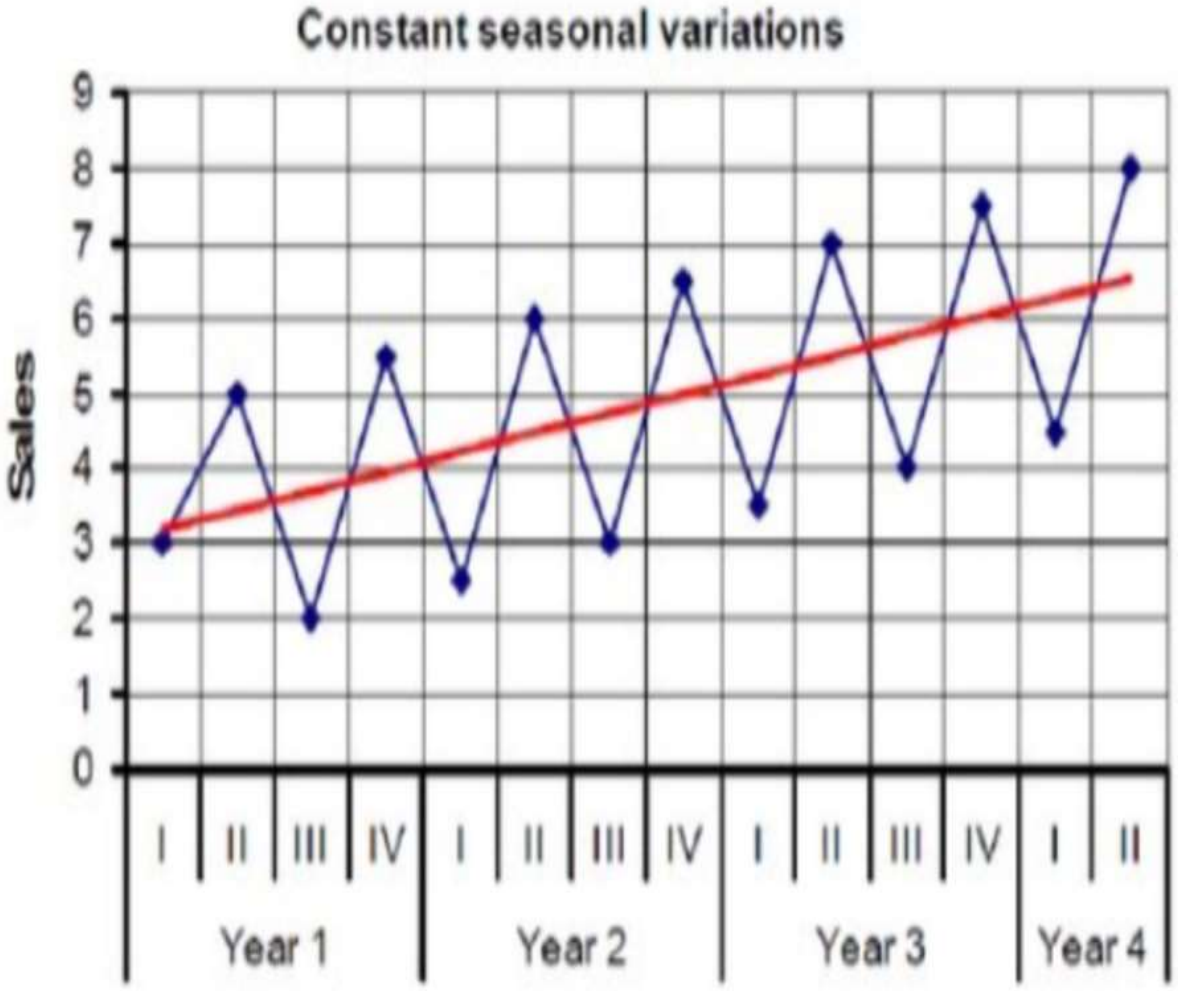
Decomposition Model cont…

Additive Model

$Y_t = T_t + S_t$ Where, Y_tActual figure in period t

T_t Trend in period t

S_t Seosonal variation in period t

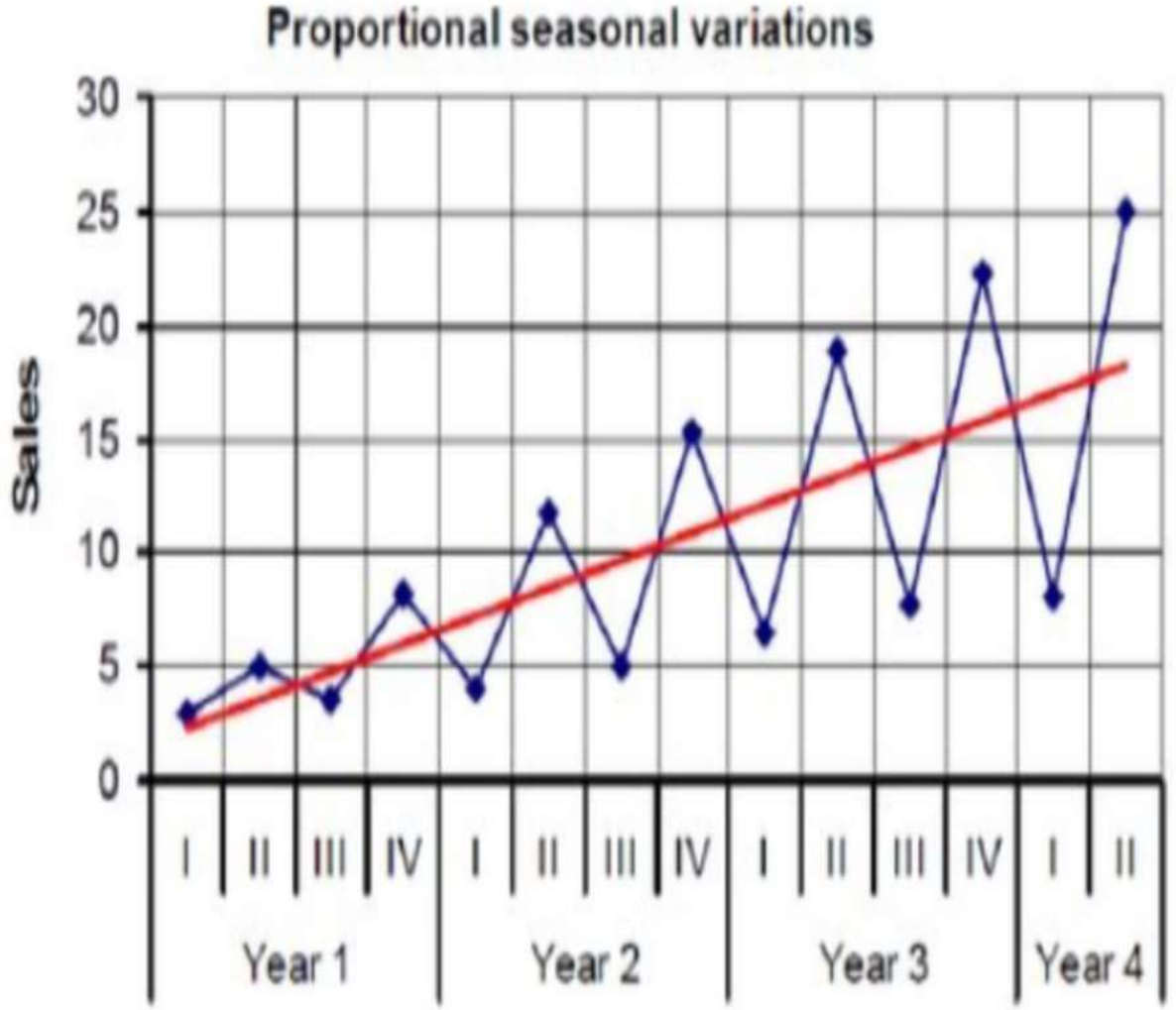


Multiplicative Model

$Y_t = T_t * S_t$ Where, Y_tActual figure in period t

T_t Trend in period t

S_t Seosonal variation in period



Decomposition Model cont...

III. Mixed model

- Different assumptions lead to different combinations of additive and multiplicative models as

$$Y_t = T_t + S_t + C_t * I_t$$

- The time series analysis can also be done using the model as:
  - $Y_t = T_t + S_t * C_t * I_t$
  - $Y_t = T_t * S_t + C_t * I_t$

NOTE: STATIONARITY TIME SERIES

- A stationary time series is one whose properties (trend, seasonality,..) do not depend on the time at which the series is observed. Thus, the statistical property (mean and variance) are constant.

Home Work

- How to determine if a time series has a trend component?
- How to determine if a time series has a seasonal component?
- How to determine if a time series has both a trend and seasonal component?

Time Series Forecasting Model

Time series forecasting is required to make scientific predictions based on historical time stamped data. It involves building models [called Time Series Forecasting Model] through historical analysis and using them to make observations and drive future strategic decision-making.

Time series forecasting models can be classified into 2 categories.

1. **Averaging methods** in which all observations (time series values) are **equally weighted**. The variations are:
    - 1.1. Averaging Model
    - 1.2 Moving Averages Model

2. **Exponential smoothing methods** that applies **unequal weights** to past data, typically decaying in an exponential manner as one goes from recent to distinct past. The variations are:
    - 2.1 Simple Exponential Smoothing/ Weighted Moving Averages Model
    - 2.2 Holt's Method
    - 2.3 Holt Winter's Method

Time Series Forecasting Model (Averaging Model)

1.1 Averaging Model

- The simple average method uses the **mean of all the past values** to forecast the next value. This method is seen to be no use in a practical scenario.
- This method is used when the time series has attained some level of **stability** and no longer dependent on any external parameters.
- This would happen in sales forecasting, only when the product for which the forecast is needed is at a **mature stage in its life cycle**.
- The Averaging Model is represented as follows where F is the forecasted value at instance of time t+1, t is the current time and Y_i is the value of series at time instant i.

$$F_{t+1} = \frac{1}{t}\sum_{i=1}^{t} Y_i$$

Averaging Model

Supplier	Amount
1	9
2	8
3	9
4	12
5	9
6	12
7	11
8	7
9	13
10	9
11	11
12	10

A manager of a warehouse wants to know how much a typical supplier delivers in 10 dollar units. He/she has taken a sample of 12 suppliers at random, obtaining the result as shown in the table.

The **computed mean of the amount is 10** and hence the manager decides to use this as the estimate for the expenditure of a typical supplier.

It is more reasonable to assume that the recent points in past are better predictors than the whole history. This is particularly true for sales forecasting.

Every product has a life cycle: Initial stage, Middle volatile period and a more or less stable Mature stage and an End stage. Hence, a better method of forecasting would be to use **Moving Averages (MAs).**

e.g. Keypad Phone as the product

Time Series Forecasting Model (Moving Average Model)

1.2 Moving Average Model or Simple Moving Average Model

- The MA approach calculates an average of a **finite number of past observations** and then employs that average as the forecast for the next period.
- The number of sample observations to be included in the calculation of the average is specified at the start of the process. The term MA refer to the fact that as a new observation becomes available, a new average is calculated by dropping the oldest observation in order to include the newest one.
- An MA of order k, represented with MA(k) is calculated as:

$$F_{t+1} = \frac{1}{k} \sum_{i=t-k+1}^{t} Y_i$$

- MA(3), MA(5) and MA(12) are commonly used for monthly data and MA(4) is normally used for quarterly data.
- MA(4), and MA(12) would average out the seasonality factors in quarterly and monthly data respectively.
- The advantage of MA method is that the data requirement is very small.
- The major disadvantage is that it assumes the data to be stationary.

Moving Averages (MAs) cont…

Month	Demand Y_i
1	89
2	57
3	144
4	221
5	177
6	280
7	223
8	286
9	212
10	275
11	188
12	312

- MA(3) = $(Y_{10}+Y_{11}+Y_{12})/3$ [Recent past 3 Y_i]
  =(275 + 188 + 312) / 3
  = 258.33

- MA(6) = (223+286+212+275+188+312)/6
  = 249.33

- MA(12) = (89+57+144+221+177+280+223
  +286+212+ 275+188+312)/ 12
  = 205.33

Home Work

Calculate:

- MA(5)
- MA(4)
- MA(10)

Time Series Forecasting Model - (Exponential Smoothing Model)

- The extension to the MA method is to have a weighted MA.
- In Single Moving Averages the past observations are weighted equally where as in Exponential Smoothing assigns exponentially decreasing weights as the observation get older. In other words, **recent observations are given relatively more weight in forecasting than the older observations.**
- In the case of moving averages, the weights assigned to the last k observations are the same and are equal to 1/k. In exponential smoothing, however, there are one or more smoothing parameters to be determined (or estimated) and these choices determine the weights assigned to the observations.

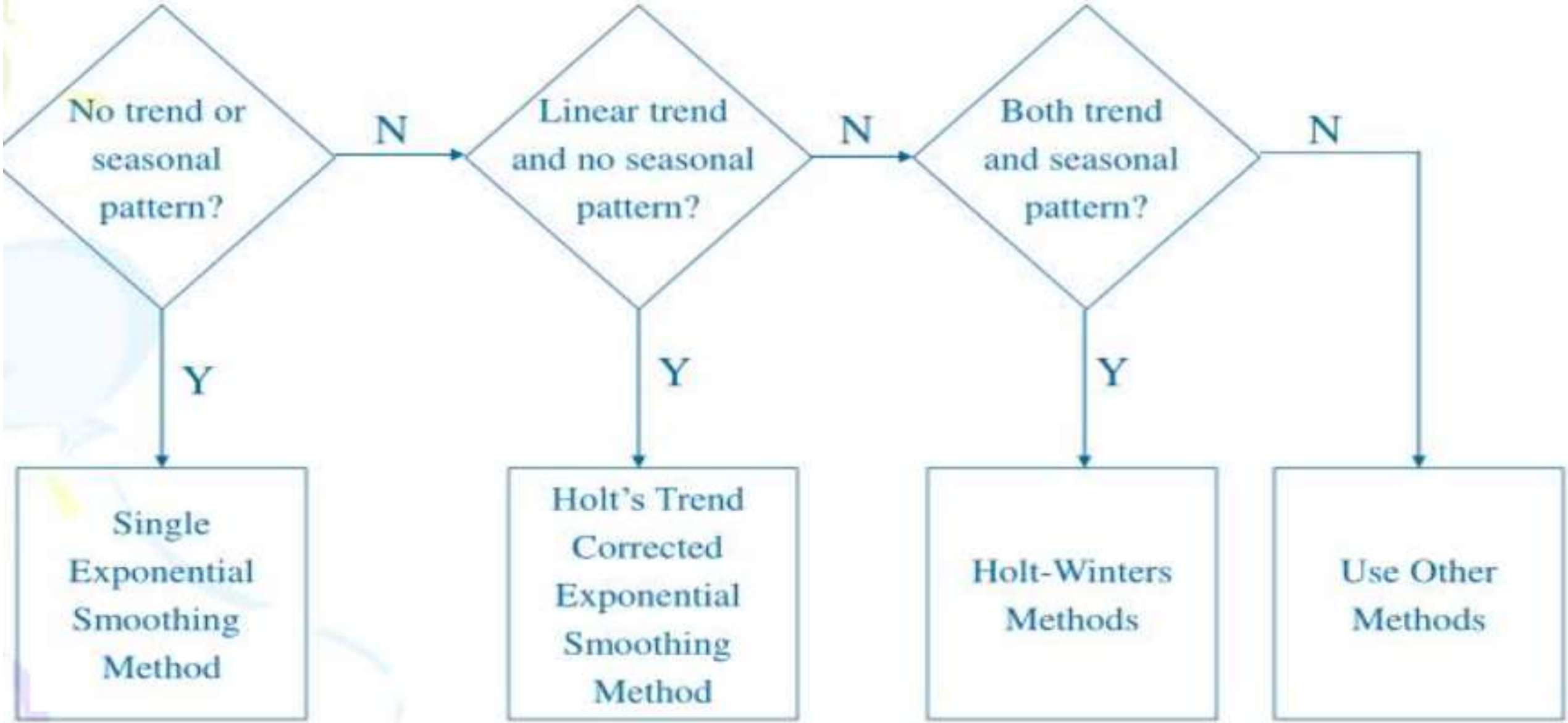
Time Series Forecasting Model - (Exponential Smoothing Model)...

Exponential smoothing consists of a range of methods

1. **Simple Exponential Smoothing (SES)/ Weighted Moving Average:** used to forecast data that have no **no trend and no seasonality**
2. **Double Exponential Smoothing or Holt's Method** used to forecast for data that **exhibit trend but not seasonality.**
3. **Triple Exponential Smoothing or Holt Winter's Method** used to forecast for data that **exhibit trend as well as seasonality.**

In all these methods, the observations are weighted in an exponentially decreasing manner as they become older.

Time Series Forecasting Model - (Exponential Smoothing Model)...



Time Series Forecasting Model - (Exponential Smoothing Model)...

2.1. Simple Exponential Smoothing Model

For any time period t, the smoothed value S_t [forecast vale] is found by computing

$$\mathbf{S_t=\alpha * y_{t-1}+(1-\alpha) * S_{t-1} \text{ and } t\geq 2}$$

where $0<\alpha\leq 1$, α is smoothing constant

y_{t-1} is the actual demand value at time t-1

or just previous demand

S_{t-1} last smoothed observation at time t-1

or just previous forecast

- When $\alpha = 1$, $S_t = Y_{t-1}$ or forecast will be just previous demand
- When $\alpha = 0$, $S_t = S_{t-1}$ or forecast will be just previous forecast

Time Series Forecasting Model - (Exponential Smoothing Model)...

2.1. Simple xponential Smoothing Model...

Why is it called Exponential?

Let us expand the basic recurrence equation by first substituting for S_{t-1} in the basic equation $S_t=\alpha * y_{t-1}+(1-\alpha) * S_{t-1}$ to obtain the exential decaying:

$$S_t=\alpha * y_{t-1}+(1-\alpha) * [\alpha * y_{t-2}+(1-\alpha) * S_{t-2}]$$

$$= \alpha * y_{t-1}+ \alpha * (1-\alpha) * y_{t-2}+ (1-\alpha)^2 * S_{t-2}$$

$$= \alpha *(1-\alpha)^{1-1} y_{t-1}+ \alpha * (1-\alpha)^{2-1} * y_{t-2}+ (1-\alpha)^2 * S_{t-2}$$

S_{t-2}

	Yi	Si
1	10	10 ($Y_1=S_1$)
2	20	S_2
3	30	S_3
…	…	…
t-1	35	S_{t-1}
t		S_t ?

By substituting for S_{t-2}, then for S_{t-3}, and so forth, until we reach S_2 (which is just y_1), it can be shown that the expanding equation can be written as:

$$S_t = \alpha \sum_{i=1}^{t-2} (1-\alpha)^{i-1} y_{t-i} + (1-\alpha)^{t-2} S_2\,, \quad t \geq 2$$

This illustrates the exponential behavior. The weights assigned to ith past Y is, $\alpha * (1-\alpha)^{i-1}$ which decreases geometrically.

Time Series Forecasting Model - (Exponential Smoothing Model)...

2.1. Simple Exponential Smoothing Model...

Month	Demand (Y_i)	Forecast (S_i)
JAN	500	400
FEB	600	
MARCH		

Q: For $\alpha = 0.75$, find the Forecast for March

$$S_t = \alpha * y_{t-1} + (1-\alpha) * S_{t-1}$$

$$S_{March} = \alpha * y_{Feb} + (1-\alpha) * S_{Feb}$$

$$S_{Feb} = \alpha * y_{Jan} + (1-\alpha) * S_{Jan}$$

$$= 0.75 * 500 + 0.25 * 400 = 475$$

$$S_{March} = \alpha * y_{Feb} + (1-\alpha) * S_{Feb}$$

$$= 0.75 * 600 + 0.25 * 475 = 568.75$$

Time Series Forecasting Model - (Exponential Smoothing Model)...

2.1. Simple xponential Smoothing Model...

- **What is the best value for α?**

  The speed at which the older responses are dampened (smoothed) is a function of the value of α. When α is close to 1, dampening is quick and when α is close to 0, dampening is slow. This is illustrated in the table below.

α	(1-α)	$(1-\alpha)^2$	$(1-\alpha)^3$	$(1-\alpha)^4$
0.9	0.1	0.01	0.001	0.0001
0.5	0.5	0.25	0.125	0.0625
0.1	0.9	0.81	0.729	0.6561

Time Series Forecasting Model - (Exponential Smoothing Model)...

2.1. Simple Exponential Smoothing Model...

Month	Demand (Y_i)	Forecast (S_i)
JAN	500	400
FEB	600	
MARCH	550	
APRIL		

For $\alpha = 0.75$ and $\alpha=0.25$ find the Forecast for April and justify which α is best

Solution: Find the error of forecast for both α and conclude the α with lesss MSE(Mean Squared Error)

Error Calculation

- The **error is calculated** as $E_t = y_t - S_t$ (difference of actual and smooth/forecast at time t)
- Then **error square is calculated** i.e. $ES_t = E_t * E_t$
- Then, **sum of the squared errors (SSE)** is calculated i.e. $SSE = \Sigma ES_i$ for i = 2 to n where n is the number of observations.
- Then, the **mean of the squared errors** is calculated i.e. $MSE = SSE/(n-1)$

The best value for α is choosen to result the smallest MSE.

Simple Exponential Smoothing cont…

Let us illustrate this principle with an example. Consider the following data set consisting of 12 observations taken over time with **α** as 0.1:

Time	y_t	S_t	E_t	ES_t
1	71	NA	NA	
	70	71		
2	69	0.1 * 70 + (1-0.1) * 71 = 71	70 - 71= -1.0	$(-1.0)^2 = 1.00$
3	68	0.1 * 69 + (1-0.1) * 71 = 70.9	69 – 70.9 = -1.90	$(-1.90)^2 = 3.61$
4	64	70.71	-2.71	7.34
5	65	70.44	-6.44	41.47
6	72	69.80	-4.80	23.04
7	78	69.32	2.68	7.18
8	75	69.58	8.42	70.90
9	75	70.43	4.57	20.88
10	75	70.88	4.12	16.97
11	70	71.29	3.71	13.76
12	70	71.67	-1.67	2.79

Simple Exponential Smoothing cont…

- Sum of the Squared Errors (SSE) = 208.94.
- Mean of the Squared Errors (MSE) is the SSE /n = 19.0.
- In the similar fashion, the MSE can be calculated for α=0.5 and let turned out to be 16.29, so in this case **we would prefer** an α of 0.5.
- **Can we do better?**
  - We could apply the proven trial-and-error method. This is an iterative procedure beginning with a range of α between 0.1 and 0.9.
  - We determine the best initial choice for α and then search between $\alpha-\Delta$ and $\alpha+\Delta$. We could repeat this perhaps one more time to find the best α to 3 decimal places.

In general, most well designed statistical software programs should be able to find the value of α that minimizes the MSE.

Time Series Forecasting Model (Holt's Model)

2.2 Holt's Method or Double Exponential Smoothing

- Holt (1957) extended simple exponential smoothing to **allow the forecasting of data with a trend but with no seasonality.**
- This method involves a forecast equation with two smoothing equations, one for the level (smoothed series) and one for the trend component.
- The m step ahead Holt forecast function for a given time series F is

$$F_{t+m} = L_t + m * b_t$$

where at time t,

L_t denotes an estimate of the level of the series

b_t denotes an estimate of the trend of the time series

Holt's Model or Double Exponential Smoothing

- **Equation of Level**

$$L_t = \alpha * y_t + (1 - \alpha) * (L_{t-1} + b_{t-1})$$

Note: L_t value depends on Y_t not Y_{t-1}

where,

α is the smoothing parameter for the level, $0 \leq \alpha \leq 1$

- **Equation for trend:**

$$b_t = \beta * (L_t - L_{t-1}) + (1- \beta) * b_{t-1}$$

where,

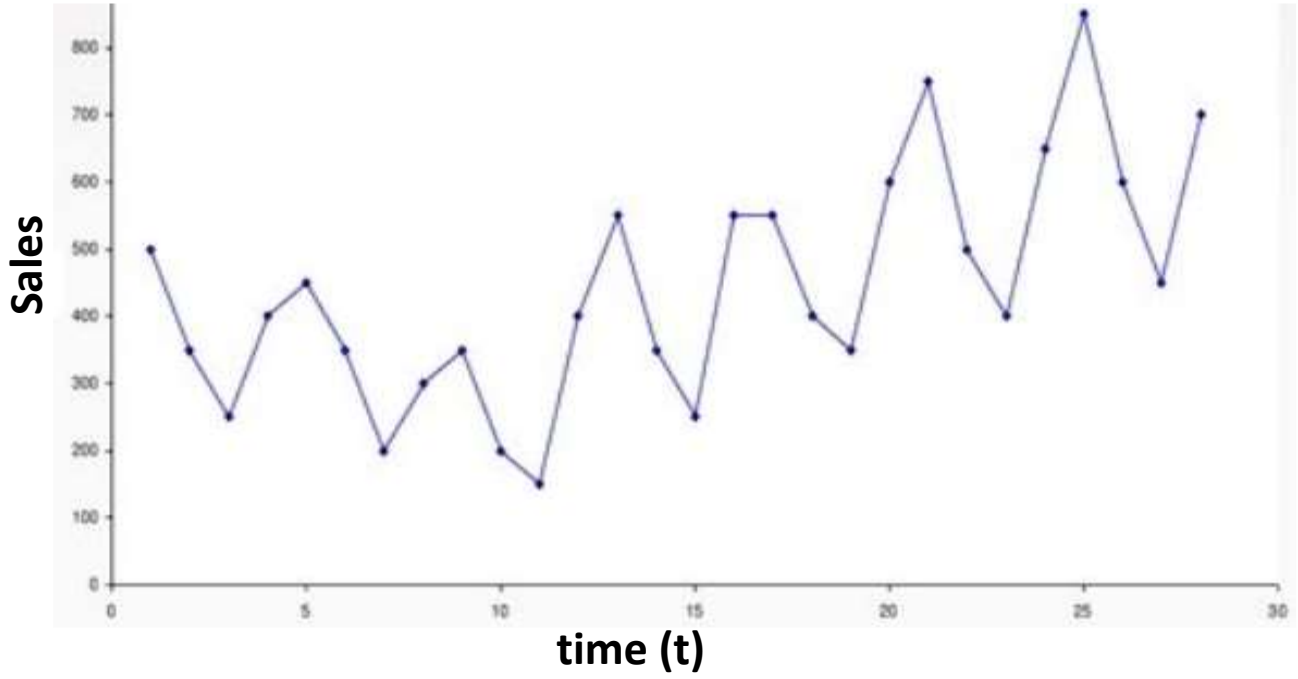
β is the smoothing parameter for the trend, $0 \leq \beta \leq 1$

- The weight of α and β can be selected subjectively or by minimizing a measure of forecast error such as MSE.
- Large value of α and β results in more rapid changes and small weights result in less rapid changes in the component.

Initialization of level: $L_1 = Y_1$ Initialization of trend $b_1 = L_2 - L_1$

Holt's Model

- The table shows quaterly sale of a company from 1994 to 1997.
- The plotting of the data shows a non-stationary time series data with trending effect and thus Holt's method can be applied.



YEAR	QUARTER	T	SALES
1994	1	1	500
	2	2	350
	3	3	250
	4	4	400
1995	1	5	450
	2	6	350
	3	7	200
	4	8	300
1996	1	9	350
	2	10	200
	3	11	150
	4	12	400
1997	1	13	550
	2	14	350
	3	15	250
	4	16	550

Holt's Model

Apply Holt's model on the given data to produce forecasts.

Use α=.3 and β =.1

Holt's Model $F_{t+m} = L_t + m * b_t$

Equation of Level: $L_t = \alpha * y_t + (1 - \alpha) * (L_{t-1} + b_{t-1})$

Equation of Trend: $b_t = \beta * (L_t - L_{t-1}) + (1- \beta) * b_{t-1}$

Initialization of $L_t = L_1$ = First Observation

Initialization of $b_t = b_0 = 0$ [Initial Slope]

YEAR	QUARTER	T	SALES
1994	1	1	500
	2	2	350
	3	3	250
	4	4	400
1995	1	5	450
	2	6	350
	3	7	200
	4	8	300
1996	1	9	350
	2	10	200
	3	11	150
	4	12	400
1997	1	13	550
	2	14	350
	3	15	250
	4	16	550

Holt's Model (α=.3 and β =.1)

YEAR	T	SALES(Y)	$L_t = \alpha * y_t + (1 - \alpha) * (L_{t-1} + b_{t-1})$	$b_t = \beta * (L_t - L_{t-1}) + (1- \beta) * b_{t-1}$	$F_{t+m} = L_t + m * b_t$
1994	1	500	500	0	500
	2	350	0.3*350 + 0.7*(500+0) = 455	0.1*(455-500) + 0.9*0 = -4.5	500 + 0 = 500
	3	250	0.3*250 + 0.7 (455 + -4.5) = 390.35	0.1*(390.35-455)+0.9*-4.5=-10.52	455 + -4.5 = 450.5
	4	400	385.88	-9.91	379.84
1995	5	450	398.18	-7.69	375.97
	6	350	378.34	-8.90	390.49
	7	200	318.61	-13.99	369.44
	8	300	303.23	-14.13	304.62
1996	9	350	307.38	-12.30	289.11
	10	200	266.55	-15.15	295.08
	11	150	220.98	-18.19	251.40
	12	400	261.95	-12.28	202.79
1997	13	550	339.77	-3.27	249.67
	14	350	355	-2.86	336.50
	15	250	311.38	-5.49	337.69
	16	550	379.12	1.83	305.89

Time Series Forecasting Model (Holt Winter's Model)

2.3 Holt Winter's Method or Double Exponential Smoothing [Multiplicative Model]

- Winter's exponential smoothing model is the second extension of the basic exponential smoothing model.
- It is used for data that exhibit both **trend** and **seasonality**.
- It is a **three parameter multiplicative model** that is an extension of Holt's method.
- An additional equation adjusts the model for the seasonal component.
- The **m** step ahead Holt forecast function for a given time series F is

$$F_{t+m} = (L_t + m * b_t) S_{t-m+s}$$

Where L_t: Level Series

b_t: Trend Estimate

S_t: Seasonal Component

s: Length of seasonality

Time Series Forecasting Model (Holt Winter's Model)

- In **three parameter multiplicative Winter's model L_t, b_t and S_t are as follows.**
- The exponential smoothed series at time t (L_t):

$$L_t = \alpha \frac{y_t}{S_{t-s}} + (1-\alpha)(L_{t-1} + b_{t-1})$$

- The trend estimate at time t (b_t):

$$b_t = \beta(L_t - L_{t-1}) + (1-\beta)b_{t-1}$$

Where α: Smoothing constant for data

β: Smoothing constant for trend esimate

ϒ: Smoothing constant for seasonal esimate

- The seasonality estimate at time t (S_t):

$$S_t = \gamma \frac{y_t}{L_t} + (1-\gamma)S_{t-s}$$

Time Series Forecasting Model (Holt Winter's Model)

- The intial value of α smoothes the data to estimate randomness. The smoothing constant β smoothes the trend in the data set. The seasonality constant ϒ smoothes the seasonality in the data.
- The weights of α, β and ϒ can be selected subjectively or by minimizing a measure of forecast error such as MSE.
- As with all exponential smoothing methods we need initial values for the components to start the algorithm, the intial value of L_t, b_t and S_t must be set.
- To determine the initial estimates of the seasonal indices we need to use at least one complete season's data (i.e. s period).

Chapter 3: **Data Analytics**

- **Introduction:**
  - Types of Data Analytics
  - Importance of Data Analytics
  - Data Analytics Applications
- **Regression Modelling Techniques::**
  - Linear Regression
  - Multiple Linear Regression
  - Non-Linear Regression
  - Logistic Regression
- **Time Series Analysis**
- **Performance analysis**
  - MSE
  - RMSE
  - MAPE

Evaluation of Forecasting Accuracy [Performance Analysis]

- What makes a good forecast? Of course, a good forecast is an accurate forecast.
- A forecast "error" is the difference between an observed value and its forecast.
- 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 forecasting model/method and helps the analyst to improve forecasting method.

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

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 Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	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

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 Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	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

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 forecasted 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 Demand	42	45	49	55	57	60	62	58	54	50	44	40
Forecasted Demand	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 forecasted value and the actual value is 11.5%.**
- The lower the value for MAPE, the better a model is able to forecast values e.g. a model with a **MAPE of 2% is more accurate** than a model with a MAPE of 10%.

THANK YOU!