# Data Science
# UNIT-V Forecasting Techniques

**Dr. N. Subhash Chandra**
**Professor CSE**

# UNIT-1: Contents

- **UNIT V – Prescriptive Analysis**
- **Forecasting Techniques**: Time series data, Techniques, and accuracy, Moving average method, Single, double, triple exponential smoothing, Regression model for forecasting, Auto-Regression models, ARIMA Process.
- **Graph Analytics:** Path analysis, Connectivity analysis, Community analysis, Centrality analysis, Social-Network Graphs, Communities and Clusters, Betweenness, The Girvan-Newman Algorithm.

# What is Forecasting?

- Forecasting is the process of making predictions about future events based on historical data and analysis.
- It's widely used in various fields such as finance, economics, supply chain management, and meteorology. There are several techniques for forecasting, broadly
- Forecasting techniques categorized into

1. **Qualitative Tools:** These are based on a judgment we make based on experience and analysis of future trends. Due to the dependency of this tool on individual judgment, the forecast is affected by human biases.
2. **Quantitative Tools:** These tools forecast data by analyzing past data. Further, it relies on statistical methods to make future predictions. These methods can be:

1. Time Series Analysis
2. Casual Methods

# TIME-SERIES DATA AND COMPONENTS OF TIME-SERIES DATA

- Time-series data is a data on a response variable, *Yt*, such as demand for a spare parts of a capital equipment or a product or a service or market share of a brand observed at different time points *t*.
- Whenever we have a forecasting problem, we will be using a time-series data. The variable *Yt* is a random variable.
- The data points or measurements are usually collected at regular intervals and are arranged in chronological order.
- If the time-series data contains observations of just a single variable (such as demand of a product at time *t*), then it is termed as **univariate time series.**
- If the data consists of more than one variable, for example, demand for a product at time *t*, price at time *t*, amount of money spent by the company on promotion at time *t*, competitors price at time *t*, etc. then it is called **multivariate time-series data**.

# COMPONENTS OF TIME-SERIES DATA

- **Trend Component (*Tt*):** Trend is the consistent long-term upward or downward movement of the data over a period of time.
- **Seasonal Component (*St*):** Seasonal component (measured using seasonality index) is the repetitive upward or downward movement (or fluctuations) from the trend that occurs within a calendar year such as seasons, quarters, months, days of the week,
- **Cyclical Component (*Ct*):** Cyclical component is fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc.
- **Irregular Component (*It*):** Irregular component is the white noise or random uncorrelated changes that follow a normal distribution with mean value of 0 and constant variance.
- The time-series data can be modelled as an addition of the above components or product of the above components. The additive time-series model is given by

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

The multiplicative time-series model is given by

$$Y_t = T_t \times S_t \times C_t \times I_t$$

# Trends in Timeseries data

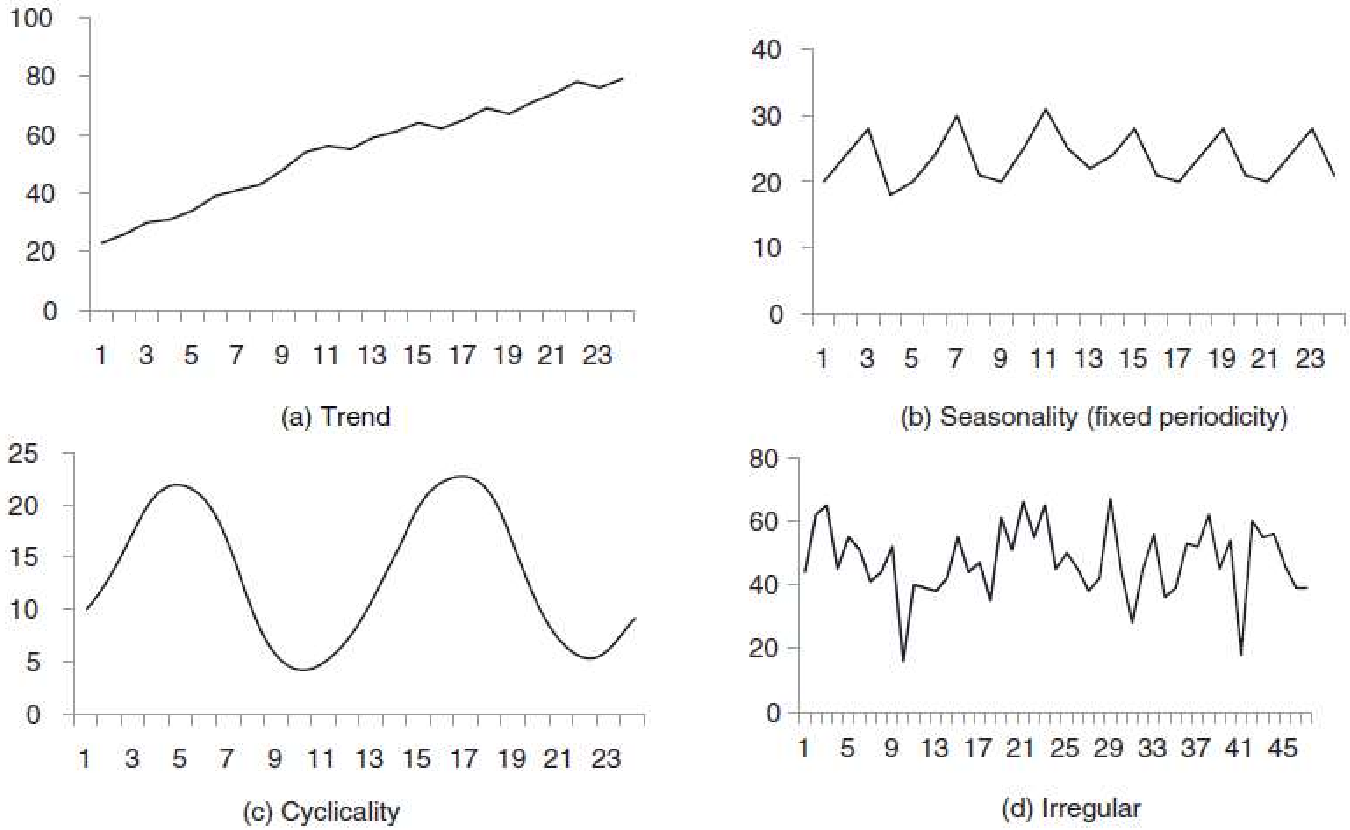



FIGURE 13.1 Trend in time-series data.

# FORECASTING TECHNIQUES AND FORECASTING ACCURACY

### 13.3.1 | Mean Absolute Error (MAE)

Mean absolute error (MAE) is the average absolute error and should be calculated on the validation data set. Assume that the validation data has $n$ observations and forecasting is carried out on these $n$ observations using the model developed. The mean absolute error is given by

$$MAE = \sum_{t=1}^{n} \frac{|Y_t - F_t|}{n} \tag{13.3}$$

In Eq. (13.3), $Y_t$ is the actual value of $Y$ at time $t$ and $F_t$ is the corresponding forecasted value.

### 13.3.2 | Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error (MAPE) is the average of absolute percentage error. Assume that the validation data has $n$ observations and the forecasting is carried out on these $n$ observations. The mean absolute percentage error is given by

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|Y_t - F_t|}{Y_t} \times 100\% \tag{13.4}$$

MAPE defined in Eq. (13.4) is one of the popular forecasting accuracy measures used by practitioners since it expresses the average error in percentage terms and is easy to interpret. Since MAPE is dimensionless it can be used for comparing different models with varying scales.

### 13.3.3 | Mean Square Error (MSE)

Mean square error is the average of squared error calculated over the validation data set. MSE is given by

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_t\right)^2 \tag{13.5}$$

Lower MSE implies better prediction. However, it depends on the range of the time-series data.

### 13.3.4 | Root Mean Square Error (RMSE)

Root mean square error (RMSE) is the square root of mean square error and is given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_t\right)^2} \tag{13.6}$$

RMSE along with MAPE are two most popular accuracy measures of forecasting. RMSE is the standard deviation of errors or residuals. In 2006, Netflix, the movie portal, announced a competition with a prize money worth one million dollars to predict the rating on a 5-point scale likely to be given a customer for a movie[2] (*source*: Wikipedia). The participants were given a target RMSE of 0.8572 to qualify for the prize.

## MOVING AVERAGE METHOD

Moving average is one of the simplest forecasting techniques which forecasts the future value of a time-series data using average (or weighted average) of the past '$N$' observations. Mathematically, a simple moving average is calculated using the formula

$$F_{t+1} = \frac{1}{N} \sum_{k=t+1-N}^{t} Y_k \tag{13.7}$$

The above formula is called simple moving average (SMA) since '$N$' past observations are given equal weights ($1/N$). In a weighted moving average, past observations are given differential weights (usually the weight decrease as the data becomes older). Weighted moving average is given by

$$F_{t+1} = \sum_{k=t+1-N}^{t} W_k \times Y_k \tag{13.8}$$

where $W_k$ is the weight given to value of $Y$ at time $k$ ($Y_k$) and $\sum_{k=t+1-N}^{t} W_k = 1$.

## SINGLE EXPONENTIAL SMOOTHING (ES)

One of the drawbacks of simple moving average technique is that it gives equal weight to all the previous observations used in forecasting the future value. This can be overcome by assigning differential weights to the past observations [Eq. (13.8)]. One easier way to assign differential weight is achieved by using single exponential smoothing (SES) technique (also known as simple exponential smoothing). Just like the moving average, SES assumes a fairly steady time-series data with no significant trend, seasonal or cyclical component. Here, the weights assigned to past data decline exponentially with the most recent observations assigned higher weights.

In single ES, the forecast at time $(t + 1)$ is given by (Winters, 1960)

$$F_{t+1} = \alpha Y_t + (1-\alpha)F_t \tag{13.9}$$

Parameter $\alpha$ in Eq. (13.9) is called the **smoothing constant** and its value lies between 0 and 1. Since the model uses one smoothing constant, it is called **single exponential smoothing**. Substituting for $F_t$ recursively in Eq. (13.9), we get

$$F_{t+1} = \alpha Y_t + \alpha(1-\alpha)Y_{t-1} + \alpha(1-\alpha)^2 Y_{t-2} + \cdots + \alpha(1-\alpha)^{t-1} Y_1 + (1-\alpha)^t F_1 \tag{13.10}$$

From Eq. (13.10), it is evident that the weights assigned to older observations decrease exponentially. Figure 13.3 shows the rate at which the weight decreases for older observations when $\alpha = 0.4$ and 0.8; the plot resembles the exponential decay curve.

## SINGLE EXPONENTIAL SMOOTHING (ES)

In summary, single exponential smoothing technique has the following advantages:

1. It uses all the historic data unlike the moving average where only the past few observations are considered to predict the future value.
2. It assigns progressively decreasing weights to older data.

Some disadvantages of smoothing methods are:

1. Increasing $n$ makes forecast less sensitive to changes in data.
2. It always lags behind trend as it is based on past observations. The longer the time period $n$, the greater the lag as it is slow to recognize the shifts in the level of the data points.
3. Forecast bias and systematic errors occur when the observations exhibit strong trend or seasonal patterns.

# Optimal Smoothing Constant in a Single Exponential Smoothing (SES)

Choosing optimal smoothing constant $\alpha$ is important for accurate forecast. Whenever the data is smooth (without much fluctuations), we may choose higher value of $\alpha$. However, when the data is highly fluctuating, then it is better to choose lower value of $\alpha$. We can find the optimal value of the smoothing constant by solving a non-linear optimization problem. For example, assume that we have to find the optimal $\alpha$ that will give the minimum RMSE. This can be achieved by solving the following optimization problem:

$$\underset{\alpha}{Min}\left[\sqrt{\frac{1}{n}\sum_{t}\left(Y_t - F_t\right)^2}\right] \tag{13.11}$$

subject to the constraint: $0 < \alpha < 1$. For the data in Table 13.1, the optimal value of $\alpha$ that minimizes the RMSE is 0.1574 and the corresponding RMSE is 739399.76. Table 13.4 shows the forecasted value, RMSE, and MAPE calculations for $\alpha = 0.1574$ (rounded to 4 decimals).

# DOUBLE EXPONENTIAL SMOOTHING – HOLT'S METHOD

- One of the drawbacks of single exponential smoothing is that the model does not do well in the presence of trend. This can be improved by introducing an additional equation for capturing the trend in the time-series data. Double exponential smoothing uses two equations to forecast the future values of the time series, one for forecasting the level (short term average value) and another for capturing the trend. The two equations are provided in Eqs. (13.12) and (13.13).

Level (or Intercept) equation ($L_t$):

$$L_t = \alpha \times Y_t + (1 - \alpha) \times F_t \tag{13.12}$$

The trend equation is given by ($T_t$)

$$T_t = \beta \times (L_t - L_{t-1}) + (1 - \beta) \times T_{t-1} \tag{13.13}$$

$\alpha$ and $\beta$ are the smoothing constants for level and trend, respectively, and $0 < \alpha < 1$ and $0 < \beta < 1$.

The forecast at time $t + 1$ is given by

$$F_{t+1} = L_t + T_t \tag{13.14}$$

$$F_{t+n} = L_t + nT_t \tag{13.15}$$

where $L_t$ is the level which represents the smoothed value up to and including the last data, $T_t$ is the slope of the line or the rate of increase or decrease at period $t$, $n$ is the number of time periods into the future.

Initial value of $L_t$ is usually taken same as $Y_t$ (that is, $L_t = Y_t$). The starting value of $T_t$ can be taken as $(Y_t - Y_{t-1})$ or the difference between two previous actual values of observations prior to the period for which forecasting is carried out. Another option for $T_t$ is $(Y_t - Y_1)/(t - 1)$.

The value of

$$L_1 = Y_1 = 3002666$$

and $$T_1 = (Y_{36} - Y_1)/35 = (4732677 - 3002666)/35 = 49428.8857$$

The value of $$F_2 = L_1 + T_1 = 3002666 + 49428.8857 = 3052095$$

The forecasted values for periods 37 to 48 are shown in Table 13.5 ($\alpha = 0.0328$ and $\beta = 0.9486$).

The RMSE and MAPE of the forecast using double exponential smoothing is given by 659888.9554 and 0.1135 (11.35%). The values of $\alpha$ and $\beta$ used in Table 13.5 are optimized values of $\alpha$ and $\beta$ that minimize the root mean square error.

# TRIPLE EXPONENTIAL SMOOTHING (HOLT-WINTER MODEL)

Moving averaging and single and double exponential smoothing techniques discussed so far can handle data as long as the data do not have any seasonal component associated with it.

However, when there is seasonality in the time-series data, techniques such as moving average, exponential smoothing, and double exponential smoothing are no longer appropriate.

- In most cases, the fitted error values (actual demand minus forecast) associated with simple exponential smoothing and Holt’s method will indicate systematic error patterns that reflect the existence of seasonality.
- Triple exponential smoothing is used when the data has trend as well as seasonality. The following three equations which account for level, trend, and seasonality are used for forecasting (for a multiplicative model, Winters 1960):

**Level (or Intercept) equation:**

$$L_t = \alpha \frac{Y_t}{S_{t-c}} + (1-\alpha)\left[L_{t-1} + T_{t-1}\right] \tag{13.16}$$

**Trend equation:**

$$T_t = \beta \times (L_t - L_{t-1}) + (1-\beta)\, T_{t-1} \tag{13.17}$$

**Seasonal equation:**

$$S_t = \gamma \frac{Y_t}{L_t} + (1-\gamma) S_{t-c} \tag{13.18}$$

The forecast $F_{t+1}$ using triple exponential smoothing is given by

$$F_{t+1} = \left[L_t + T_t\right] \times S_{t+1-c} \tag{13.19}$$

where $c$ is the number of seasons (if it is monthly seasonality, then $c = 12$; in case of quarterly seasonality $c = 4$; and in case of daily data $c = 7$). The initial values of $L_t$ and $T_t$ are calculated using the following equations:

$$L_t = Y_t \tag{13.20}$$

Alternatively

$$L_t = \frac{1}{c}\left(Y_1 + Y_2 + \cdots + Y_c\right) \tag{13.21}$$

$$T_t = \frac{1}{c}\left[\frac{Y_t - Y_{t-c}}{12} + \frac{Y_{t-1} - Y_{t-c-1}}{12} + \frac{Y_{t-2} - Y_{t-c-2}}{12} + \ldots + \frac{Y_{t-c+1} - Y_{t-2c+1}}{12}\right] \tag{13.22}$$

# Predicting Seasonality Index Using Method of Averages

The following steps are used for predicting the seasonality index using method of averages:

**STEP 1**

Calculate the average of value of $Y$ for each season (that is, if the data is monthly data, then we need to calculate the average for each month based on the training data). Let these averages be $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_t$.

**STEP 2**

Calculate the average of the seasons' averages calculated in step 1 (say $\overline{\overline{Y}}$ ).

**STEP 3**

The seasonality index for season $k$ is given by the ratio $\overline{Y}_k / \overline{\overline{Y}}$. Variation to the procedure explained above is first divide the value of $Y_t$ with its yearly average and calculate the seasonal average. We will use first 3 years of data in Table 13.1 to calculate the seasonality index for various months. The seasonality index based on first 3 years of data using method of averages is shown in Table 13.6.

# Forecasting Time-Series Data with Seasonal Variation

**STEP 1**

Estimate the seasonality index (using techniques such as method of averages or ratio to moving average).

**STEP 2**

De-seasonalize the data using either additive or multiplicative model. For example, in multiplicative model, the de-seasonalized data $Y_{d,t} = Y_t / S_t$, where $Y_{d,t}$ is the de-seasonalized data and $S_t$ is the seasonality index for period $t$.

**STEP 3**

Develop a forecasting model on the de-seasonalized data ($F_{d,t}$).

**STEP 4**

The forecast for period $t + 1$ is $F_{t+1} = F_{d,t+1} \times S_{t+1}$.

# REGRESSION MODEL FOR FORECASTING

The forecasted value at time $t$, $F_t$, can be written as a regression equation as follows:

$$F_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} + \varepsilon_t \qquad (13.27)$$

Here $F_t$ is the forecasted value of $Y_t$, and $X_{1t}$, $X_{2t}$, etc. are the predictor variables measured at time $t$. The regression equation for Example 13.1 is

$$F_t = \beta_0 + \beta_1 \text{ promotion expenses at time } t + \beta_2 \text{ competition promotion at time } t$$

# AUTO-REGRESSIVE (AR) MODELS

### 1. Introduction to Autoregression Models

- **Autoregression (AR)** is a type of time series model where the current value of a variable is regressed on its own previous values.
- It is based on the assumption that past data points have a direct influence on future data points.
- **Time series** data refers to data points that are collected or recorded at specific intervals over time (e.g., daily stock prices, monthly sales, etc.).

### :. Basic Idea of Autoregression (AR)

- An autoregressive model assumes that the future value of a time series is a **linear combination** of its past values.

  For example, in an AR(1) model:

  $$y_t = \phi_1 y_{t-1} + \epsilon_t$$

  - $y_t$: the value of the series at time $t$
  - $\phi_1$: coefficient that describes the influence of $y_{t-1}$ on $y_t$
  - $\epsilon_t$: white noise error term

- In general, for an AR(p) model:

  $$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

  where $p$ is the order of the autoregressive process, i.e., how many previous time steps are considered.

### 3. AR Model Components

- **Lag (p)**: This refers to how many prior time points (lags) influence the current value.
  - An AR(1) model uses one lag (the immediate prior value).
  - An AR(2) model uses two lags.
- **Coefficients ($\phi$)**: These coefficients describe the relationship between the lagged values and the current value.
- **Noise term ($\epsilon_t$)**: This is an error term that represents randomness and accounts for unexplained variance in the model.

**4. Stationarity Assumption**

•**Stationarity**: A key assumption of AR models is that the time series is **stationary**. This means that its properties (mean, variance, autocorrelation) do not change over time.

•Stationarity ensures that the relationships between lags and the series are stable and meaningful over time.
•**If the series is non-stationary (e.g., has a trend), differencing may be applied to make it stationary before fitting an AR model.**

# Types of AR Models

1. **AR(1) Model**: Only the immediately previous value influences the current value.
   - Example: $y_t = \phi_1 y_{t-1} + \epsilon_t$
2. **AR(2) Model**: Two prior time steps influence the current value.
   - Example: $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$
3. **Higher-Order AR Models:** Use more lags, like AR(p), where $p$ can be any positive integer.

**Forecasting with AR Models**

- AR models are widely used for **forecasting** future values of time series.
- The future value is predicted based on past observations.
- Forecast intervals can also be generated to represent the uncertainty in the predictions.

**Limitations of AR Models**

- **Linearity**: AR models assume linear relationships between the current and lagged values.
- **Stationarity**: They require the series to be stationary, which may not be the case in many real-world time series.
- **Overfitting**: Including too many lags can result in overfitting, where the model captures noise rather than the underlying pattern.

**Applications of AR Models**

- **Economics and Finance**: Predicting stock prices, interest rates, and inflation rates.
- **Weather Forecasting**: Estimating future weather conditions based on historical data.
- **Sales Forecasting**: Predicting future sales based on past trends.

## Extensions of AR Models

- **ARMA (Autoregressive Moving Average) Models**: Combine AR with a moving average (MA) component.
- **ARIMA (Autoregressive Integrated Moving Average) Models**: Extend ARMA to handle non-stationary series using differencing.

# Moving Average (MA) Models

- **Moving Average (MA)** models are a type of time series model where the current value of the series is expressed as a linear combination of **past error terms** (also known as residuals or shocks).

- In contrast to **Autoregressive (AR)** models, where the past values of the series itself are used, MA models use past forecast errors to explain the current observation.

- **Time series data** involves observations taken over regular time intervals (e.g., monthly sales data, daily temperatures).

**Moving Regression vs. Traditional Regression:**

- **Traditional Regression**: Assumes relationships between variables are static over time.
- **Moving Regression**: Assumes relationships can change as time progresses, fitting models to **rolling windows** of data.
- Better for **time series forecasting** in dynamic environments.

### 2. Basic Idea of Moving Average (MA)

- The **Moving Average model** assumes that the current value of the time series is influenced by a finite number of past error terms (white noise shocks).

  In an MA(1) model:

  $$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}$$

  - $y_t$: the value of the series at time $t$
  - $\mu$: the mean of the series (assumed to be constant)
  - $\epsilon_t$: the current white noise error term
  - $\theta_1$: the coefficient for the lagged error term
  - $\epsilon_{t-1}$: the error term from the previous time step

- In general, for an MA(q) model:

  $$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

  - **q** is the number of lagged error term[illegible]rder of the MA process).

### 3. Components of an MA Model

1. **White Noise (Error Terms):**
   - $\epsilon_t$ represents the **white noise** in the series, which is normally distributed with a mean of zero and constant variance.
   - These error terms are uncorrelated with each other.
2. **Lagged Errors:**
   - Past error terms are used to predict the current value.
   - The coefficients $\theta_1, \theta_2, \ldots, \theta_q$ represent the influence of these past errors on the current observation.
3. **Order (q):**
   - The order of the model determines how many past error terms are considered.
   - For example, an MA(1) model only includes the immediate past error term, while an MA(2) includes two past error terms.

**8. Types of MA Models**

1. **MA(1) Model**: Involves only one past error term.

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}$$

   - The current value is influenced by the present error and the error from the previous time period.

2. **MA(2) Model**: Involves two past error terms.

$$y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2}$$

   - This extends the model by considering errors from both the last and second-to-last time periods.

3. **Higher-Order MA Models (MA(q))**: Involve more past errors, with **q** representing the number of lagged error terms.

### 11. Forecasting with MA Models

- **Short-term forecasting**: MA models are especially good at capturing short-term dependencies in the data.
- **Forecasting future values**: The model predicts the next value based on past errors, assuming that these errors follow the same pattern.

  Example forecast for an MA(1) model:

  $$\hat{y}_{t+1} = \mu + \theta_1 \epsilon_t$$

- **Uncertainty of predictions**: Because predictions rely on past errors, the uncertainty grows as you forecast further into the future.

### 12. Limitations of MA Models

- **Limited Scope**: MA models work well for series that exhibit short-term dependencies but are not as effective for long-term trends or seasonality.
- **Need for Invertibility**: To ensure model stability, MA models must satisfy the **invertibility condition**—i.e., the MA process should have an AR representation.
- **Complex Parameter Estimation**: As $q$ increases, estimating the parameters becomes more complex, especially for large datasets.

# ARMA Models

Auto-regressive moving average (ARMA) is a combination auto-regressive and moving average process. ARMA(*p, q*) process combines AR(*p*) and MA(*q*) processes. ARMA(*p, q*) model is given by

The equation for an ARMA(p, q) model is:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

Where:

- $y_t$ is the value of the time series at time $t$.
- $\phi_1, \phi_2, \ldots, \phi_p$ are the **autoregressive (AR)** coefficients.
- $y_{t-1}, y_{t-2}, \ldots, y_{t-p}$ are the past values of the series (lags).
- $\epsilon_t$ is the error term (random shocks, also called "white noise") at time $t$.
- $\theta_1, \theta_2, \ldots, \theta_q$ are the **moving average (MA)** coefficients.
- $\epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_{t-q}$ are the past forecast errors (lags of the error term).

### Example of ARMA(1,1):

For an ARMA(1,1) model, where $p = 1$ and $q = 1$:

$$y_t = \phi_1 y_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

This indicates:

- The current value $y_t$ is influenced by the previous value $y_{t-1}$ (AR term), the current error $\epsilon_t$, and the previous error $\epsilon_{t-1}$ (MA term).

### Components:

- **Autoregressive (AR)**: Describes how the current value is related to its own previous values.
- **Moving Average (MA)**: Describes how the current value is influenced by past errors or random shocks.

An ARMA model assumes stationarity, meaning the mean, variance, and autocorrelation of the series are constant over time. If the series is not stationary, differencing might be applied, leading to an ARIMA model.

# ARIMA Models

- **ARIMA** stands for **Autoregressive Integrated Moving Average**.
- ARIMA models are a popular class of statistical models used for **time series forecasting**.
- They combine **Autoregression (AR)**, **Moving Average (MA)**, and **Integration (I)** to handle non-stationary time series data.
- Widely used in fields like finance, economics, and environmental science.

The ARIMA model has three key components:

1. **Autoregressive (AR) Part:**
   - This part uses past values to model the current value.
   - The AR term is denoted by **p**, representing the number of lagged values used.
   - Example: In an AR(1) model, $y_t = \phi_1 y_{t-1} + \epsilon_t$.
2. **Integrated (I) Part:**
   - Integration refers to **differencing** the data to achieve stationarity.
   - The I term is denoted by **d**, representing how many times the data has been differenced.
   - Differencing means subtracting the previous value from the current value: $y_t - y_{t-1}$.
   - Non-stationary data is transformed into stationary data through this process.

# ARIMA Models

3. **Moving Average (MA) Part:**

- The MA part models the error terms (random shocks) from previous time points.
- The MA term is denoted by **q**, representing how many lagged error terms are used.
- Example: In an MA(1) model, $y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1}$.

The general equation for an **ARIMA(p, d, q)** model combines the **autoregressive (AR)**, **integrated (I)**, and **moving average (MA)** components. The equation is:

$$y'_t = \phi_1 y'_{t-1} + \phi_2 y'_{t-2} + \cdots + \phi_p y'_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

Where:

- $y'_t$ is the differenced series at time $t$ (obtained by differencing the original series $y_t$ $d$-times to make it stationary).

$$y'_t = \Delta^d y_t$$

($\Delta^d$ is the differencing operator applied $d$-times.)

# ARIMA Models

- $\phi_1, \phi_2, \ldots, \phi_p$ are the **autoregressive (AR)** coefficients that describe how past values $y'_{t-1}, y'_{t-2}, \ldots$ influence $y'_t$.
- $\epsilon_t$ is the white noise error term at time $t$ (random shocks).
- $\theta_1, \theta_2, \ldots, \theta_q$ are the **moving average (MA)** coefficients that describe how past forecast errors $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$ influence $y'_t$.

### Example of ARIMA(1, 1, 1):

For ARIMA(1,1,1), where $p = 1$, $d = 1$, and $q = 1$:

$$y'_t = \phi_1 y'_{t-1} + \epsilon_t + \theta_1 \epsilon_{t-1}$$

Here:

- $y'_t = y_t - y_{t-1}$ (since $d = 1$, the first difference is taken).
- The next value of the series depends on the lagged value $y'_{t-1}$ and past error $\epsilon_{t-1}$, along with a random shock $\epsilon_t$.

# ARIMA Models

### 3. ARIMA Notation

- ARIMA models are typically denoted as **ARIMA(p, d, q)**, where:
  - **p**: The number of autoregressive terms (AR component).
  - **d**: The number of times the series is differenced (I component).
  - **q**: The number of moving average terms (MA component).

For example, an **ARIMA(1, 1, 1)** model means:

- **p = 1**: The model includes one lag of the series in the AR part.
- **d = 1**: The series has been differenced once to make it stationary.
- **q = 1**: The model includes one lag of the error term in the MA part.

# ARIMA Models

$Y_t$ → AR filter → Integration filter → MA filter → $\varepsilon_t$

(long term) (stochastic trend) (short term) (white noise error)

ARIMA (2,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + b_1 \varepsilon_{t-1}$

ARIMA (3,0,1) $y_t = a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + b_1 \varepsilon_{t-1}$

ARIMA (1,1,0) $\Delta y_t = a_1 \Delta y_{t-1} + \varepsilon_t$, where $\Delta y_t = y_t - y_{t-1}$

ARIMA (2,1,0) $\Delta yt = a_1 \Delta y_{t-1} + a_2 \Delta y_{t-2} + \varepsilon_t$ where $\Delta yt = yt - yt\text{-}1$

To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q

# ARIMA Modelling

To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q

- **Ensuring Stationarity**
  - Determine the appropriate values of d
- **Identification:**
  - Determine the appropriate values of p & q using the ACF, PACF, and unit root tests
  - p is the AR order, d is the integration order, q is the MA order
- **Estimation :**
  - Estimate an ARIMA model using values of p, d, & q you think are appropriate.
- **Diagnostic checking:**
  - Check residuals of estimated ARIMA model(s) to see if they are white noise; pick best model with well behaved residuals.
- **Forecasting:**
  - Produce out of sample forecasts or set aside last few data points for in-sample forecasting.

## 8. ARIMA Model Identification

- **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** are used to identify appropriate values for **p** and **q**.
  - **ACF** helps identify the order of the MA term (**q**).
  - **PACF** helps identify the order of the AR term (**p**).

Typical patterns:

- If the ACF tails off slowly, consider an AR term.
- If the PACF tails off slowly, consider an MA term.

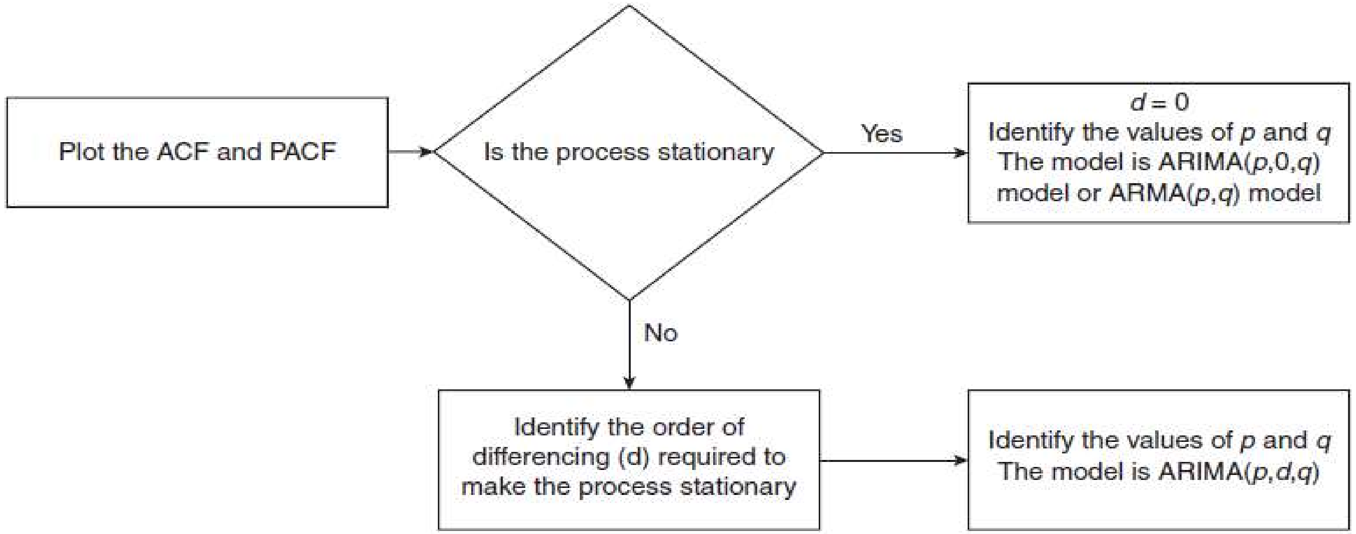



FIGURE 13.14 Model identification in ARIMA model.

# Autocorrelation Function(ACF)

The **Autocorrelation Function (ACF)** measures the correlation between observations of a time series at different time lags. The ACF at lag $k$ is the correlation between the time series values at time $t$ and $t-k$.

The formula for the autocorrelation at lag $k$ is:

$$\rho_k = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

Where:

- $y_t$ is the value of the time series at time $t$.
- $y_{t-k}$ is the value of the time series at time $t-k$.
- $\bar{y}$ is the mean of the time series.
- $T$ is the total number of observations.
- $k$ is the lag at which the autocorrelation is being calculated.

# Partial Autocorrelation Function(PACF)

The **Partial Autocorrelation Function (PACF)** measures the correlation between the time series values at two points, after accounting for the values of the time series at intervening lags.

The PACF at lag $k$ is the correlation between $y_t$ and $y_{t-k}$, controlling for the intermediate lags (i.e., lags 1, 2,..., $k-1$).

The PACF can be computed by solving the Yule-Walker equations, but for simpler interpretation, the PACF at lag $k$ can be described as:

$$\phi_{kk} = \frac{\rho_k - \sum_{i=1}^{k-1} \phi_{k-1,i}\rho_{k-i}}{1 - \sum_{i=1}^{k-1} \phi_{k-1,i}\rho_i}$$

Where:

- $\phi_{kk}$ is the partial autocorrelation at lag $k$.
- $\rho_k$ is the autocorrelation at lag $k$.
- $\phi_{k-1,i}$ are the partial autocorrelations at lower lags (from lag 1 to $k-1$).

## Step-by-Step Guide to Selecting $p$ and $q$ for ARIMA:

1. **Visualize ACF and PACF Plots**

- After differencing the time series (to make it stationary, if necessary), plot the **ACF** and **PACF** of the differenced series.

2. **Identify $p$ from PACF**

- Look at the **PACF** plot to identify the significant lags:
  - If the **PACF cuts off** (i.e., shows significant spikes) after lag $p$, you can select $p$ as the lag where the PACF cuts off.
  - For example, if PACF has significant spikes up to lag 2 and then diminishes, this suggests an **AR(2)** model.

3. **Identify $q$ from ACF**

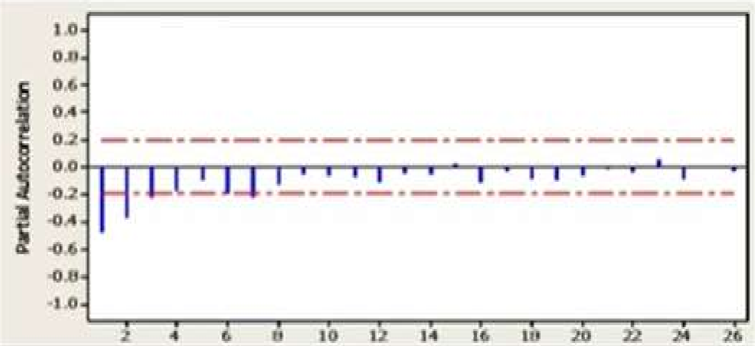
- Look at the **ACF** plot to identify the significant lags:
  - If the **ACF cuts off** after lag $q$, this suggests that a **MA(q)** model is appropriate.
  - For example, if ACF shows significant spikes up to lag 1 and no significant spikes afterward, this suggests an **MA(1)** model.

- Identify using **Partial Autocorrelation (PACF) plot**.
- PACF conveys the pure correlation between a lag and the series.

$$Yt = \alpha 0 + \alpha 1Yt - 1 + \alpha 2Yt - 2 + \alpha 3Yt - 3$$

current series lag 1 of Y lag 2 of Y lag 2 of Y

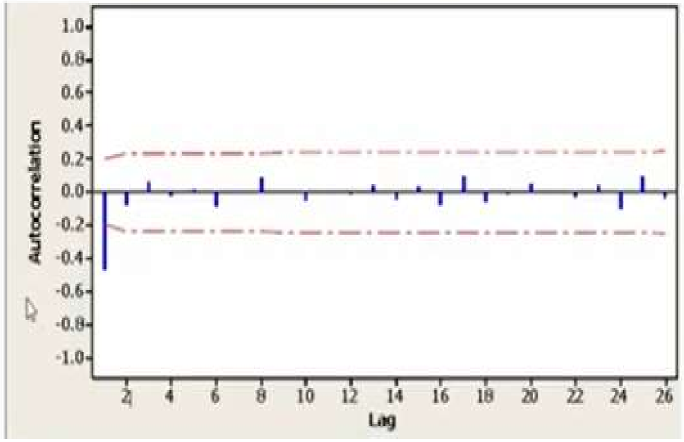
- Take the order of AR term to be equal to as many lags that crosses the significance limit in the PACF plot.




p=1,2,3

# Determine of MA term (q)

- Identify using **Autocorrelation (ACF) plot**.
- An MA term is technically, the error of the lagged forecast.
- The ACF tells how many MA terms are required to remove any autocorrelation in the stationaries series.




q=1

### Summary of Steps to Determine $d$ (Number of Differences)

1. **Visual Inspection:**
   - Look for trends or seasonality in the time series. If they are present, differencing might be necessary.
2. **First Differencing:**
   - Compute the first difference of the time series and check if the series appears stationary (through visual inspection or by looking at the ACF/PACF plots).
3. **ADF Test:**
   - Use the **Augmented Dickey-Fuller (ADF) test** to confirm if the series is stationary.
   - If the p-value > 0.05 (indicating non-stationarity), apply differencing and test again.
4. **Second Differencing (if needed):**
   - If the first difference does not make the series stationary, apply a second difference and repeat the process.

# Example for differencing

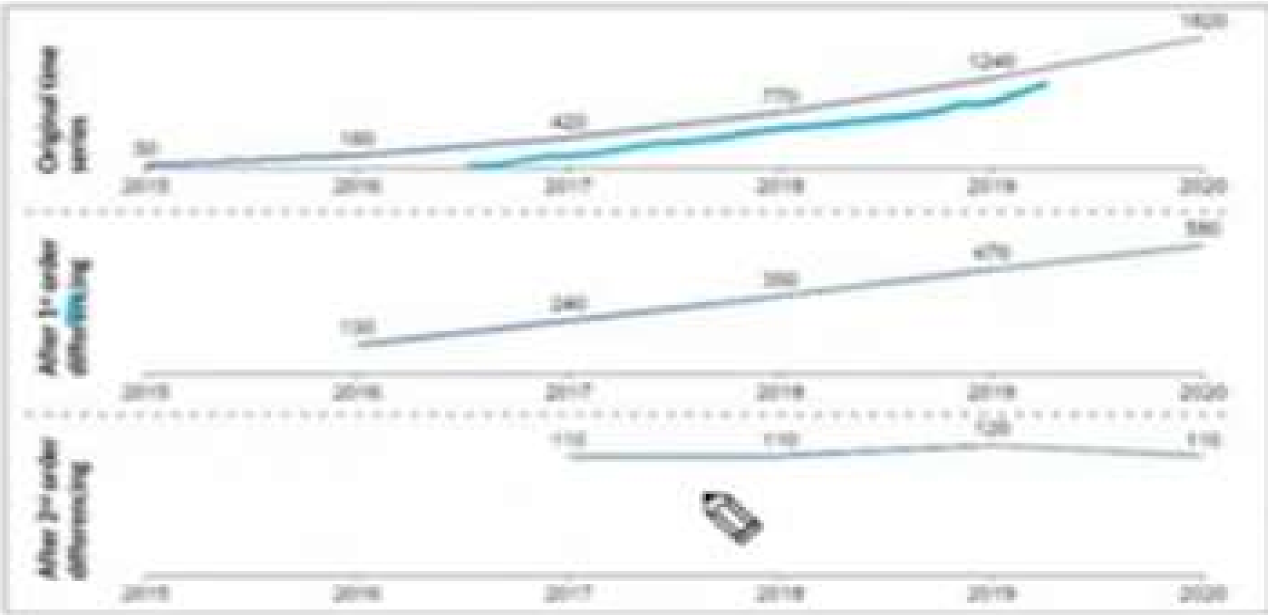
## Differencing:

- Compute the differences between consecutive observations.
- It's just like differencing a sloped line so as to make the slope zero.
- Differencing stabilizes the mean of a time series by removing changes in the level of a time series and therefore eliminating (or reducing) trend and seasonality.

**For Example:**

| Original time series | After 1st order differencing | After 2nd order differencing |
|---|---|---|
| 50 | | |
| 180 | 130 | |
| 420 | 240 | 110 |
| 770 | 350 | 110 |
| 1240 | 470 | 120 |
| 1820 | 580 | 110 |

**DIFFERENCING**



**Seasonal ARIMA (SARIMA)**

•When time series data exhibits **seasonality**, the basic ARIMA model is extended to **SARIMA** (Seasonal ARIMA).

•SARIMA adds seasonal components to the AR, MA, and differencing parts.

•SARIMA notation is **ARIMA(p, d, q)(P, D, Q)m**, where:

- **P, D, Q** are the seasonal orders for AR, differencing, and MA.
- **m** is the number of periods per season (e.g., m = 12 for monthly data with annual seasonality).

# Limitations of ARIMA Models:

1. **Linear Assumption**: ARIMA models assume that the relationships between past values and future values are linear.
2. **Stationarity Requirement**: The model requires stationary data, which may not always be the case.
3. **Short-Term Forecasting**: ARIMA models are typically more effective for short-term forecasts, as long-term forecasts may suffer from increasing uncertainty.

**Applications of ARIMA Models:**

1. **Economics and Finance**: ARIMA is used to predict stock prices, interest rates, GDP growth, and inflation.
2. **Sales Forecasting**: Retailers use ARIMA to forecast future sales based on historical data.
3. **Environmental Science**: ARIMA models are used to predict temperature changes, rainfall, and pollution levels.