CHAPTER-III

RANDOM WALK MODELS, GENERAL LINEAR MODELS, TIME SERIES MODELS AND NEURAL NETWORK MODELS FOR FORECASTING

3.1 BRIEF ABOUT RANDOM WALK MODELS:

S.O.N.Agwuegbo, A.P.Adewole, A.N.Maduegbuna(2010) proposed the stock market forecasting is marked more by its failure than by its successes since stock prices reflect the judgments and expectations of investors, based on the information available. If things look good, the price moves upward so quickly that recipient of cheerful information has little or no time to act upon it. Remarkably, efforts have been made to apply econometric techniques of model building in the prediction of stock prices in an effort to demonstrate that the market fluctuations are essentially unpredictable (Bernstein and Bostain, 1974; Black, 1971; Brealey and Myers, 1996; Buhlmann, 2005).

Fama and French (1988) have argued that there are long term pattern in stock prices with several years of upswing followed by more sluggish periods. According to Fama (1965; 1995), a stock market where successive, price changes in individual securities are independent is, by their definition a random walk market. According to Kendal (1953), stock prices following a random walk implies that the price changes are as independent of one another as the gains and losses. The independence assumption of the random walk model is valid as long as knowledge of the past behavior of the series of price changes cannot be used to increase expected gains. More specifically, if successive price changes for a given security are independent, there is no problem in timing purchases and sales of that security. A simple policy of buying and holding the security will be as good as any more complicated mechanical procedure for timing purchases and sales (Fama, 1965,1995). In this study, the random walk approach is presented with the specific aim of giving a definite description of the Nigerian stock market prices. In a world without interest rates, idealized stock prices should be martingales. This is one way of formulating the so called efficient market hypothesis (Buhlmann, 2005; Fama, 1965; 1995).

The random walk model $X_t$ is the price of stock on any particular day $t$. At each subsequent time unit, $X_t$ varies in an unpredictable fashion such that during each time unit the
stock price moves its location either one unit left ward, or one unit right ward or it remains the same. We suppose that the system can either remain in i, or move to state i-1 or to state i+1 and the random direction of each jump is independent of all earlier jumps. More specifically, if \( S_n \) denotes the location of the stock prices at time \( n \) the overall behaviour of the sequence \( (S_n) \) is of the form

\[
S_n = X_1 + X_2 + \cdots + X_n
\]  

(3.1.1)

Where the \( X_n \) are independent random variables Therefore

\[
S_n = S_0 + \sum_{j=1}^{n} X_j
\]  

(3.1.2)

Where

\[
S_0 = \text{The starting position}
\]

\[
X_j = \text{Independent and identically distributed random variables each taking either the value -1 with probability } q, \text{ the value +1 with probability } P, \text{ or remain the same at value 0 with probability } r.
\]

A different perspective on forecasting comes from the random walk theory. Random walk theory gained popularity in 1973 when Burton Malkiel wrote “A random walk Down Wall Street“, a book that is now regarded as an investment classic. Random walk is a stock market theory that states that the past movement or direction of the price of stock overall market cannot be used to forecast its future movement. Originally examined by Maurice Kendall 1953, the theory states the stock price fluctuations are independent of each other and have the same probability distribution, but that over a period of time, prices maintained an upward trend. Random walk theory declares that even with such information, future forecast is ineffective.

3.1.1 Generating Function:

The main interest is to determine the distribution of \( S_n \) for finite \( n \) given assumptions about the distribution of the \( x_j \). Intuitively given

\[
S_n = \sum_{j=1}^{n} X_j
\]  

(3.1.3)
We suppose that the system can be in one of the enumerable sequence of states denoted by $S_1, S_2, \ldots, S_n$. If $S_n$ takes different values for different states then the conditional distribution of $S_{n+1}$ given $S_n, S_{n-1}, \ldots, S_0$ can be written as:

$$p \left\{ S_{n+1} = i_{n+1} / S_n = i_n, S_{n-1} = i_{n-1}, \ldots, S_0 = i_0 \right\}$$  \hspace{1cm} (3.1.4)$$

And if the Markovian property is satisfies, this always equal to

$$p \left\{ S_{n+1} = i_{n+1} / S_n = i_n \right\}$$  \hspace{1cm} (3.1.5)$$

Equation (3.1.5) is also known as the nearest neighbour property. Since the idea of a nearest neighbour system does not involve any prefers direction of time, we then consider only what happens when Markov chain is observed as $n$ decreases in value instead of increasing. These provide a complete description Markov Process.

3.1.2 Martingale Techniques:

It is clear that the investors in stock may make certain choices at each investment on how to invest, of which the choice may depend on the past history of the stock prices.

The expected fortune of the investors after trial $n+1$ is:

$$E \left\{ S_{n+1} = i_{n+1} / S_n = i_n, S_{n-1} = i_{n-1}, \ldots, S_0 = i_0 \right\}$$  \hspace{1cm} (3.1.6)$$

If (3.1.6) is equal $S_n$, the investment is fair.

The investment by the investors is fair for all $n$ when:

$$E \left\{ S_{n+1} = i_{n+1} / S_n = i_n \right\} = S_n$$  \hspace{1cm} (3.1.7)$$

We suppose that the expectation in (3.1.7) is calculated only on the value of $S_n$ but also on the whole previous history of the sequence. Such a system is paid to be a martingale (Karr, 1993; Williams, 1991; Gerber, 1994). Martingale (Ito, 1986; Rolsji et al., 1999) represents a fair game.
and sums of independent random variables. Naturally, the investors unfair investment is said to be a super martingale and the theory of such a system is due to Doob (1953).

We then define an unfair investment by the inequality is:

$$E\left\{ S_{n+1} = i_{n+1} / S_n = i_n \right\} \leq S_n \text{ for all } n$$

(3.1.8)

No investor can alter the fairness or unfairness of a stock price as defined by expectation. All that can be done is to transform the sequence so that the probability of fairness is increased or decreased while at the same time dividend is increased or decreased.

The Defining Conditions for a martingale can be given terms of the difference equation:

$$Z_n = S_n - S_{n-1}$$

(3.1.9)

Where

$$Z_n = \text{Identically distributed}$$

$$S_0 = 0 \text{ with probability one}$$

Equation (3.1.9) has a Markov Property can be seen as solution to recursive stochastic equation:

$$S_n = Z_n + S_{n-1}$$

(3.1.10)

3.1.3 Chain Dependent Model:

Let \{S_n\} be a martingale with increment as defined in (3.1.10)

Then

$$S_n = Z_1 + Z_2 + \cdots + Z_n$$

(3.1.11)

and are still martingale.

We can think of \(Z_{k+1}\) as a score associated with the \(K^{th}\)

Let

$$S_n = \sum_{k=1}^{n} Z_k$$

(3.1.12)
Sn is called an additive functional of the Markov chain the strength of dependence in the Markov chain can be computed from the transition matrix.

By using chapman–Kolmogorov equation, we specify the n-step transition probabilities of lower order by an argument that

\[ p_{ij}^{(n)} = P\{S_n = j | S_o = i\} \]  \hspace{1cm} (3.1.13)

\[ = \sum_k p\{S_n = j | S_{n-1} = k \ldots \ldots S_0 = i\} \]

\[ = \sum_k P \{S_n = j | S_{n-1} = k\} P \{S_{n-1} = k | S_o = i\} \]

\[ = \sum_k P_{kj} p_{ik}^{(n-1)} \]

Then

\[ p_{ij}^{(n)} = \sum_k P_{kj} p_{ik}^{(n-1)} \]  \hspace{1cm} (3.1.14)

The state Probabilities for time n are defined as

\[ p_j^{(n)} = P\{S_n = j\} \]  \hspace{1cm} (3.1.15)

Equation (3.1.14) is given the vector of initial probabilities

\[ p_i^0 = P\{S_o = i\} \]  \hspace{1cm} (3.1.16)

And then-step transition matrix such that

\[ p_j^{(n)} = P\{S_n = j\} \]

\[ = \sum_k P \{S_n = j | S_0 = i\} P \{S_0 = i\} \]

\[ = \sum_{i} p_{ij}^{(n)} p_i^{(0)} \]
Thus therefore

\[ p_j^{(n)} = \sum p_{ij}^{(n)} p_i^{(0)} \]  \hspace{1cm} (3.1.17)

This is completely specified when its transition matrix P and the initial conditions \( P^{(0)} \) are known. Equation (3.1.16) is an ideal procedure in studying the behaviour of the system when it is at equilibrium. It shows that we can study the behaviour of the system over period of time by repeated application of P to any initial distribution vector. After a sufficient long period of time the system settles down to a condition of statistical equilibrium in which the state occupation probabilities are independent of the initial conditions.

We limit our study to the vase where the chain is in stationary state so that the initial state \( P_i^{(0)} \) will become less relevant to the n-step transition probability as n increases.

The limiting values are defined as:

\[ \lim P[S_n = j\, /\, S_0 = j] = P[S_n = j] = \pi_j \]  \hspace{1cm} (3.1.18)

The limiting value is stationery or time invariant. This invariance property of \( \pi_j \) is connected with the notion of ergodicity of markov chain. The ergodicity of the markov chain then implies that \( \pi_j \) is the unique solution to the system of linear equations it turns out that

\[ \pi_j = \lim P_{ij}^n \]  \hspace{1cm} (3.1.19)

The \( \pi_j \) are strictly positive and independent of I and is probability function since.

\[ \sum_{j=1}^{n} \pi_{ij} = 1 \]  \hspace{1cm} (3.1.20)

The limiting distribution in (3.1.18) satisfies the balance equation

\[ \pi_{ij} = \pi_{ij} p_{ij} \]  \hspace{1cm} (3.1.21)

Because of the invariance property, each probability solution \( \pi_{ij} \) in
(3.1.21) is called a stationary initial distribution of \([S_n]\). In matrix notation (3.1.21) can be written as

\[ \pi = \pi P \]  \hspace{1cm} (3.1.22)

Or equivalently

\[ \pi(1 - P) = 0 \]  \hspace{1cm} (3.1.23)

Thus \( \pi \) is left Eigen vector of \( P \) corresponding to Eigen value 1.

### 3.2 General Linear Models or Regression Models:

#### 3.2.1 Simple Applied Linear Regression Model:

The Simple applied Linear Regression Model that is a model with single regressor \( x \) that has a relationship with response variable \( y \) that is a straight line. The Simple Linear Regression Model is

\[ y = \beta_0 + \beta_1 x + \epsilon \]  \hspace{1cm} (3.2.1)

Where the intercept \( \beta_0 \) and the slope are unknown constants and \( \epsilon \) is an error component. The errors are assumed to have mean zero and unknown variance \( \sigma^2 \).

The regressor ‘\( x \)’ is controlled by the data analyst and measured with negligible error, while response variable \( y \) is a random variable. The probability distribution for \( y \) at each possible value for ‘\( x \)’. The mean of the distribution is

\[ E(y/x) = \beta_0 + \beta_1 x \]  \hspace{1cm} (3.2.2)
And the variance is

$$Var(y/ x) = Var(\beta_0 + \beta_1 x + \epsilon) = \sigma^2$$

The mean of y is a linear function of ‘x’ of the variance ‘y’ does not depend on the value of ‘x’.

The parameters $\beta_0, \beta_1$ are called regression coefficients.

The slope $\beta_1$ is the change in the mean, the distribution ‘y’ produced by a unit change ‘x’.

When $x = 0$ the intercept $\beta_0$ is the mean of the distribution of the response variable y. If
the range of ‘x’ does not include zero, then $\beta_0$ has no practical interpretation.

### 3.2.2 Least Squares Estimation of the parameters:

The parameter $\beta_0$ and $\beta_1$ are unknown and must be estimated using sample data.

### 3.2.3 Estimation of $\beta_0$ and $\beta_1$:

The method of least squares is used to estimate $\beta_0$ and $\beta_1$ i.e., estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ so that the sum of squares of the differences between the observation $y_i$ and the straight line is a minimum. The simple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i=1,2,\ldots,n$$

(3.2.3)

Thus the least square criticism is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

(3.2.4)

The least squares estimator of $\beta_0$ and $\beta_1$, say $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy
\[
\left( \frac{\partial s}{\partial \hat{\beta}_0}, \frac{\partial s}{\partial \hat{\beta}_1} \right)_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0
\]

And
\[
\left( \frac{\partial s}{\partial \hat{\beta}_1} \right)_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) x_i = 0
\]

Simplify the above two equations, then
\[
\begin{align*}
n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i
\end{align*}
\]

(3.2.5)

Equations (3.2.5) are called the least squares normal equations. By solving the least squares normal equations,
\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

(3.2.6)

And
\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \sum_{i=1}^{n} y_i \right) \left( \sum_{i=1}^{n} x_i \right)}{n \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2}
\]

(3.2.7)

Where \( \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \) and \( \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \) are the averages of \( Y_i \) and \( X_i \) respectively.

\[\therefore\] \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) in equations (3.2.6) and (3.2.7) are the least square estimators of the intercept and slope respectively.
The fitted simple linear regression model is then
\[ \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \]  
(3.2.8)

Equation (3.2.8) gives a point estimate of the mean of Y for a particular ‘x’.

\[ \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 \] is the corrected sum of squares of \( X_i \).

\[ \sum_{i=1}^{n} y_i x_i - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) \left( \frac{\sum_{i=1}^{n} x_i}{n} \right) \] is the corrected sum of cross products of \( X_i \) and \( Y_i \). We may write these quantities as

\[ S_{xx} = \sum_{i=1}^{n} x_i^2 - \left( \frac{\sum_{i=1}^{n} x_i}{n} \right)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 \]  
(3.2.9)

And

\[ S_{xy} = \sum_{i=1}^{n} y_i x_i - \left( \frac{\sum_{i=1}^{n} y_i}{n} \right) \left( \frac{\sum_{i=1}^{n} x_i}{n} \right) \]

\[ = \sum_{i=1}^{n} y_i (x_i - \bar{x}) \]  
(3.2.10)

Equation (3.2.7) can be written as

\[ \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \]  
(3.2.11)

The difference between the observed value \( Y_i \) and corresponding fitted values \( \hat{y}_i \) is a residual. The \( i^{th} \) residual is

\[ \epsilon_i = y_i - \hat{y}_i = y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_i \right) \]  
(3.2.12)

Residual play an important role in investigating the adequacy of the fitted regression model.
3.2.4 Properties of the Leas Squares Estimators and Fitted Regression model:

The least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have several important properties. In equations (3.2.6) and (3.2.7) that $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of the observations $y$.

$$\hat{\beta}_i = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} C_i y_i, \text{ Where, } C_i = \frac{x_i - \bar{x}}{S_{xx}} \text{ for } i = 1, 2, ..., n$$

The least square estimator $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimator of the model parameter $\beta_0$ and $\beta_1$. To show this for $\hat{\beta}_1$, consider

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} C_i Y_i\right) = \sum_{i=1}^{n} C_i E(Y_i)$$

$$= \sum_{i=1}^{n} C_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i x_i$$

Since, $E(\epsilon_i) = 0$ by the assumption and $\sum_{i=1}^{n} C_i = 0$ and $\sum_{i=1}^{n} C_i x_i = 1$

$$E(\hat{\beta}_1) = \beta_1$$

i.e. if we assume that the model is correct, then $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$. $\hat{\beta}_0$ is an unbiased estimator of $\beta_0$. Or $E(\hat{\beta}_0) = \beta_0$

The variance of $\hat{\beta}_1$ is found as

$$Var(\hat{\beta}_1) = Var\left[\sum_{i=1}^{n} C_i Y_i\right]$$
\[ n \sum_{i=1}^{n} C_i^2 \text{Var}(Y_i) \]  

(3.2.13)

The observations \( Y_i \) are uncorrelated. The variance of each term in the sum is \( C_i^2 \text{Var}(Y_i) \) and assume that \( \text{Var}(Y_i) = \sigma^2 \).

Consequently,

\[
\text{Var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} C_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}} = \frac{\sigma^2}{S_{xx}} 
\]

(3.2.14)

The variance of \( \hat{\beta}_0 \) is

\[
\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) 
\]

\[
= \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) - 2 \bar{x} \text{Cov}(\bar{y}, \hat{\beta}_1) 
\]

Now the variance of \( \bar{y} \) is just \( \text{Var}(\bar{y}) = \frac{\sigma^2}{n} \) and the covariance between \( \bar{y} \) and \( \hat{\beta}_1 \) can be zero.

\[
\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y}) + \bar{x}^2 \text{Var}(\hat{\beta}_1) 
\]

\[
= \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) 
\]

(3.2.15)

There are several useful properties of the least square fit:

1. The sum of residuals in any regression model that contains an intercept \( \hat{\beta}_0 \) is always zero.
2. The sum of the observed values $Y_i$ equals to the sum of the fitted values $\hat{Y}_i$

\[ \text{i.e., } \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \]

3. The Least Squares Regression line always passes through the centred \( \text{point } (\bar{y}, \bar{x}) \) of the data

4. The sum of residuals weighted by the corresponding value of the regression variable always equals to zero.

\[ \text{i.e., } \sum_{i=1}^{n} x_i e_i = 0 \]

5. The sum of residuals weighted by the corresponding fitted value equals to zero.

\[ \text{i.e., } \sum_{i=1}^{n} y_i \hat{e}_i = 0 \]

### 3.2.5 Estimation of $\sigma^2$:

To estimating $\hat{\beta}_0$ and $\hat{\beta}_1$, an estimate of $\sigma^2$ is required to test hypothesis and construct interval estimates pertinent to the regression model. The estimates of $\sigma^2$ is obtained from the residual (or) error sum of squares

\[ SS_{\text{Res}} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3.2.16) \]

$SS_{\text{Res}}$ may be found by substituting $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ in equation (3.2.16) and simplifying, then

\[ SS_{\text{Res}} = \sum_{i=1}^{n} y_i^2 - n \bar{y}^2 - \hat{\beta}_1 S_{xy} \quad (3.2.17) \]

But
\[ \sum_{i=1}^{n} Y_i^2 - n \bar{Y}^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = SS_T \]

is the corrected sum of squares of the response observations.

\[ SS_{Res} = SS_T - \hat{\beta}_1 S_{xy} \quad \text{(3.2.18)} \]

The residual sum of squares has \( n-2 \) degree of freedom, because two degrees of freedom are associated with the estimates \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) involved in obtaining \( \hat{y}_i \). The expected value of \( SS_{Res} \) is \( E(SS_{Res}) = (n-2) \sigma^2 \), so an unbiased estimator of \( \sigma^2 \) is

\[ \hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res} \quad \text{(3.2.19)} \]

The quantity \( MS_{Res} \) is called the residual mean square. The square root of \( \hat{\sigma}^2 \) is sometimes called the standard error of regression.

### 3.2.6 Important Results on \( SS_R \) AND \( SS_{Res} \):

(a) \( SS_R \)

By definition \( SS_R = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \)

\text{(3.2.20)}

We note that \( \hat{y} = X \left( X'X \right)^{-1} X'y \) and that

\[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} V'y \]

\text{(3.2.21)}

Where ‘\( V \)’ is an \( n \times 1 \) vector all of whose elements are ones. Further, \( n = V'V \) Thus

\[ \bar{y} = \left(V'V \right)^{-1} V'y \]

\text{(3.2.22)}

We can write \( SS_R \) as
$SS_R = \sum (\hat{y}_i - \bar{y})^2$

$= [\hat{y} - V\bar{y}]^t [\hat{Y} - V\bar{y}]$

$= [X(X^tX)^{-1}X^ty - V(V^tV)^{-1}V^ty]^t[X(X^tX)^{-1}X^ty - V(V^tV)^{-1}V^ty]$

$= y^t[X(X^tX)^{-1}X^t - V(V^tV)^{-1}V]^t[X(X^tX)^{-1}X^t - V(V^tV)^{-1}V]y$

(3.2.23)

Note that $X = [V \ X_r]$, where $X_r$ is the matrix formed by the actual values for the regressors. Consequently, $SS_R$ involves a special case of a partitioned matrix. The special identity for partitioned matrices to show that $X(X^tX)^{-1}X^tV = V$ and $V^tX(X^tX)^{-1}X^t = V^t$.

Consequently, we can show that $X(X^tX)^{-1}X^t - V(V^tV)^{-1}V$ is Idempotent. Under the assumption that $Var(\varepsilon) = \sigma^2 I$, then

$$SS_R = \frac{y^t[X(X^tX)^{-1}X^t - V(V^tV)^{-1}V]y}{\sigma^2} = \frac{SS_R}{\sigma^2} = \frac{1}{\sigma^2} y^t[X(X^tX)^{-1}X^t - V(V^tV)^{-1}V]y$$

(3.2.24)

Follow a non central $\chi^2$ distribution with noncentrality parameter $\lambda$ and degrees of freedom equal to the rank of $X(X^tX)^{-1}X^t - V(V^tV)^{-1}V$. Since this matrix is idempotent. Its rank is its trace.

We note that,

$$trace\left[X(X^tX)^{-1}X^t - V(V^tV)^{-1}V\right] = trace\left[X(X^tX)^{-1}X^t\right] - trace\left[V(V^tV)^{-1}V\right]$$

$$= trace\left[X(X^tX)^{-1}\right] - trace\left[V^t(V^tV)^{-1}V\right]$$
\[ = \text{trace}[I_P] - \text{trace}[V] \quad (3.2.25) \]
\[ = P - 1 = K \]

under the assumption that the model is correct.

\[ E(y) = X\beta = [V_{\beta_R}][\begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix}] = \beta_0 + X_R\beta_R \]

Thus the noncentrality parameter is

\[ \lambda = \frac{1}{\sigma^2} E(y)' [X(X'X)^{-1}X' - V(V'V)^{-1}V'] E(y) \]
\[ = \frac{1}{\sigma^2} \beta^T X^T [X(X'X)^{-1}X' - V(V'V)^{-1}V'] X \beta \]

\[ = \frac{1}{\sigma^2} \begin{bmatrix} \beta_0 & \beta_R \end{bmatrix} \begin{bmatrix} V' \\ X_R^T \end{bmatrix} \begin{bmatrix} X(X'X)^{-1}X' - V(V'V)^{-1}V' \\ V'X_R \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix} \quad (3.2.26) \]

\[ = \frac{1}{\sigma^2} \begin{bmatrix} \beta_0 & \beta_R \end{bmatrix} \begin{bmatrix} V'X(X'X)^{-1}X' - V'V(V'V)^{-1}V' \\ X_R^T X(X'X)^{-1}X' - X_R^T V(V'V)^{-1}V' \end{bmatrix} \begin{bmatrix} V'X_R \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix} \quad (3.2.28) \]

\[ = \frac{1}{\sigma^2} \begin{bmatrix} \beta_0 & \beta_R \end{bmatrix} \begin{bmatrix} 0' \\ X_R^T - X_R^T V(V'V)^{-1}V' \end{bmatrix} \begin{bmatrix} V'X_R \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix} \]

\[ = \frac{1}{\sigma^2} \begin{bmatrix} \beta_0 & \beta_R \end{bmatrix} \begin{bmatrix} 0 & 0' \\ 0 & X_R^T X_R - X_R^T V(V'V)^{-1}V'X_R \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix} \quad (3.2.29) \]

\[ = \frac{1}{\sigma^2} \beta_R \begin{bmatrix} X_R^T X_R - X_R^T V(V'V)^{-1}V'X_R \end{bmatrix} \beta_R \]
If we define the matrix of centered regressor values, \( X_c \) by

\[
X_c = \begin{bmatrix}
  x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1k} - \bar{x}_k \\
  x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2k} - \bar{x}_k \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{nk} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{nk} - \bar{x}_k 
\end{bmatrix}
\]

(3.2.30)

where \( \bar{x}_1 \) is the average value for the first regressor.

\( \bar{x}_2 \) is the average value for the second regressor.

The noncentrality parameter as

\[
\lambda = \frac{1}{\sigma^2} \beta_k^T [X_c^T X_c] \beta_k
\]

(3.2.31)

The expected value for \( SS_R \) is

\[
E(SS_R) = E\left[ y^T \left( X^T X \right)^{-1} X^T - V (V^T V)^T \right] y 
\]

(3.2.32)

\[
= \text{trace} \left( X \left( X^T X \right)^{-1} X^T - V (V^T V)^T \right) \sigma^2 I + E(y)^T [X \left( X^T X \right)^{-1} X^T - V (V^T V)^T] E(y) 
\]

(3.2.33)

\[
= K \sigma^2 + \beta_k^T X_c^T X_c \beta_k
\]

(3.2.34)

As a result,

\[
E(MS_R) = E \left( \frac{SS_R}{K} \right) = \sigma^2 + \frac{\beta_k^T X_c^T X_c \beta_k}{K}
\]

(3.2.35)
(a) $SS_{Res}$:

By definition,

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$  \hspace{1cm} (3.2.36)

We note that we can rewrite $SS_{Res}$ as

$$SS_{Res} = (y - \hat{y})' (y - \hat{y})$$

$$= [y - X (X'X)^{-1} X' y]' [y - X (X'X)^{-1} X' y]$$

$$= y' (I - X (X'X)^{-1} X') y$$  \hspace{1cm} (3.2.37)

It is trivial to show that $[I - X (X'X)^{-1} X']$ is symmetric, idempotent consequently.

$$\frac{SS_{Res}}{\sigma^2} = \frac{1}{\sigma^2} y' [I - X (X'X)^{-1} X'] y$$  \hspace{1cm} (3.2.38)

Follow $\chi^2$ distribution. The degrees of freedom come from the rank of $[I - X (X'X)^{-1} X']$, which is turn in the trace. It is straightforward to show that the trace is $n - p$. Under the assumption that the model is correct.

$$E(y) = X\beta$$  \hspace{1cm} (3.2.39)

Thus, the Noncentrality parameter is

$$\frac{1}{\sigma^2} E(y)' [I - X (X'X)^{-1} X'] E(y)$$

$$= \frac{1}{\sigma^2} \beta' X' [I - X (X'X)^{-1} X'] X \beta$$

$$= \frac{1}{\sigma^2} \beta' [X'X - X'X (X'X)^{-1} X'X] \beta = 0$$  \hspace{1cm} (3.2.40)
\[ \therefore \frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p} \]

The expected value for \( SS_{Res} \) is

\[
E(\ SS_{Res} ) = E \left( y^\top \left( I - X \left( X^\top X \right)^{-1} X^\top \right) y \right) \tag{3.2.41}
\]

\[
= \text{trace} \left[ \left( I - X \left( X^\top X \right)^{-1} X^\top \right) \sigma^2 I \right] + E \left[ y^\top \left[ I - X \left( X^\top X \right)^{-1} X^\top \right] \right] E(y) \\
= (n-p)\sigma^2 \tag{3.2.42}
\]

\[
\therefore E(\ MS_{Res} ) = E \left( \frac{SS_{Res}}{n-p} \right) = \sigma^2 \tag{3.2.43} \textbf{3.2.7}
\]

**Hypothesis testing on the slope and intercept:**

Suppose we have to test the hypothesis that the slope equals a constant say \( \beta_0 \). The appropriate hypotheses are

\[
H_0 = \beta_1 = \beta_{10} \]
\[
H_1 = \beta_1 \neq \beta_{10} \tag{3.2.44}
\]

Where alternative hypothesis is two sided. Since the errors \( \varepsilon_i \) are NID \((0, \sigma^2)\). The observation \( y_i \) are NID \((\beta_0 + \beta_1 x_i, \sigma^2)\). Now, \( \hat{\beta}_1 \) is a linear combination of observations, so \( \hat{\beta}_1 \) is normally distributed with mean \( \beta_1 \) and variance \( \frac{\sigma^2}{S_{xx}} \) using the mean and variance of \( \hat{\beta}_1 \).

Therefore, the test statistic is

\[
Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}} \tag{3.2.45}
\]
is distributed N(0,1) if the null hypothesis $H_0: \beta_1 = \beta_{10}$ \hspace{1cm} (3.2.46)

is true. If $\sigma^2$ were known, we could use $z_0$ to test the hypothesis

$$MS_{res}$$ is an unbiased estimator of $\sigma^2$ and $\frac{(n-2)MS_{res}}{\sigma^2}$ follow a $\chi^2_{n-2}$ distribution and that $MS_{res}$ and $\hat{\beta}_1$ are independent.

By the definition of ‘t’ statistic $t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{res}/S_{xx}}} - t_{n-2}$ \hspace{1cm} (3.2.47)

If the Null hypothesis $H_0: \beta_1 = \beta_{10}$ is true. The degrees of freedom associated with to are the number of degrees of freedom associated with $MS_{res}$. Thus, the ratio $t_0$ is the test statistics used to test $H_0: \beta_1 = \beta_{10}$. The test procedure computes $t_0$ and compares the observed value of $t_0$ from equation (3.2.47) with upper $\left(\frac{\alpha}{2}\right)$ percentage point of the $t_{n-2}$ distribution($t_{\alpha/2,n-2}$). This procedure rejects the null hypothesis. If

$$|t_0| > t_{\alpha/2,n-2}$$ \hspace{1cm} (3.2.48)

$\sqrt{MS_{res}/S_{xx}}$ is called the estimated standard error, the standard error of the slop. is called the estimated standard error, the standard error of the slope.

i.e., $S.E(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{S_{xx}}}$ \hspace{1cm} (3.2.49)

t can be written as

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{S.E(\hat{\beta}_1)}$$ \hspace{1cm} (3.2.50)

A similar procedure can be used to test hypothesis about the intercept to test
\[ H_0 : \beta_0 = \beta_{00} \] 
\[ H_i : \beta_0 \neq \beta_{00} \quad (3.2.51) \]

The test statistics is 
\[ t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res} \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{S.E(\hat{\beta}_0)} \quad (3.2.52) \]

Where \( S.E(\hat{\beta}_0) = \sqrt{MS_{Res} \left( \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} \) is the standard error of the intercept. We reject the null hypothesis \( H_0 : \beta_0 = \beta_{00} \) if \( |t_0| > t_{\alpha/2,n-2} \) \( (3.2.53) \)

3.2.8 INTERVAL ESTIMATION IN SIMPLE LINEAR REGRESSION:

CONFIDENCE INTERVALS ON \( \beta_0, \beta_1 \) AND \( \sigma^2 \):

The point estimates of \( \beta_0, \beta_1 \) and \( \sigma^2 \), can also obtain confidence interval estimates of these parameters. The width of these confidence intervals is a measure of the overall quality of the regression line. If the errors are normally and independently distributed, then the sampling distribution of both 
\[ \frac{\hat{\beta}_1 - \beta_1}{S.E(\hat{\beta}_1)} \text{ and } \frac{\hat{\beta}_0 - \beta_0}{S.E(\hat{\beta}_0)} \quad (3.2.54) \]

is with \( n - 2 \) degree of freedom. Therefore, a \( 100(1 - \alpha) \) percent confidence interval on the slope \( \beta_1 \) is given by
\begin{align*}
\hat{\beta}_1 - t_{a/2, n-2}S.E(\hat{\beta}_1) &\leq \beta_1 \leq \hat{\beta}_1 + t_{a/2, n-2}S.E(\hat{\beta}_1) \\
\text{(3.2.55)}
\end{align*}

and a \(100(1-\alpha)\) percent confidence interval on the intercept \(\hat{\beta}_0\) is

\begin{align*}
\hat{\beta}_0 - t_{a/2, n-2}S.E(\hat{\beta}_0) &\leq \beta_0 \leq \hat{\beta}_0 + t_{a/2, n-2}S.E(\hat{\beta}_0) \\
\text{(3.2.56)}
\end{align*}

These confidence intervals have the usual frequency interpretation.

If the errors are normally and identically distributed then sampling distribution of \((n-2)MS_{Res}/\sigma^2\) is Chi-Square with \(n-2\) degrees of freedom. Thus,

\begin{align*}
P \left\{ \chi^2_{a/2, n-2} \leq \frac{(n-2)MS_{Res}}{\sigma^2} \leq \chi^2_{a/2, n-2} \right\} = 1 - \alpha
\end{align*}

And consequently a \(100(1-\alpha)\) percent confidence interval on \(\sigma^2\) is

\begin{align*}
\frac{(n-2)MS_{Res}}{\chi^2_{a/2, n-2}} \leq \sigma^2 \leq \frac{(n-2)MS_{Res}}{\chi^2_{1-a/2, n-2}} \\
\text{(3.2.57)}
\end{align*}

3.2.9 INTERVAL ESTIMATION OF THE MEAN RESPONSE:

Let \(x_0\) be the level of the regressor variable within the range of the original data on ‘x’ used to fit the model. An unbiased point estimation of \(E(Y/x_0)\) is found from the fitted model.

\[ E(Y/x_0) = \hat{\mu}_{Y/x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \]  

To obtain a \(100(1-\alpha)\) percent confidence interval on \(E(Y/x_0)\). First note that \(\hat{\mu}_{Y/x_0}\) is a normally distributed random variable because it is a linear combination of the observations \(y_i\).

The variance of \(\hat{\mu}_{Y/x_0}\) is

\[ \text{Var} \left( \hat{\mu}_{Y/x_0} \right) = \text{Var} \left( \hat{\beta}_0 + \hat{\beta}_1 x_0 \right) = \text{Var} \left[ \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) \right] \]

\[ = \frac{\sigma^2}{n} + \frac{\sigma^2 (x_0 - \bar{x})^2}{s_{xx}} = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right] \]  

\[ \text{(3.2.59)} \]

Since \(Cov(\bar{y}, \hat{\beta}_1) = 0\). The sampling distribution of
Consequently, a $100(1-\alpha)$ percent confidence interval on the mean response at the point $x=x_0$ is

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2} \sqrt{\frac{MS_{Res}}{n} \left( \frac{1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)} \leq E(Y / x_0) \leq \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2} \sqrt{\frac{MS_{Res}}{n} \left( \frac{1 + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)}$$

(3.2.61)

The width of the confidence interval for $E(Y / x_0)$ is a function of $x_0$; the interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.

**3.2.10 PREDICTION OF NEW OBSERVATIONS:**

An important application of the regression model is prediction of new observations $y$ corresponding to a specified level of the regressor variable $x$. If $x_0$ is the values of regressors variable of interest, then

$$\hat{y}_0 = \hat{b}_0 + \hat{b}_1 x_0$$

(3.2.63)

is the point estimate of the new value of the response $y_0$. Now consider interval estimator of the future observation $y_0$. The confidence interval on the mean response at $x = x_0$ is
\[
\hat{\mu}_{y_0} - t_{\alpha/2, n-2} \sqrt{\text{MS}_{\text{Res}} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \leq E(y | x_0) \leq \hat{\mu}_{y_0} + t_{\alpha/2, n-2} \sqrt{\text{MS}_{\text{Res}} \left( \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}
\]

(3.2.64)

Is inappropriate for this problem because it is an interval estimate of the mean \( y \), not a probability statement about future observations from that distribution.

Now we will develop a prediction interval for the future observation \( y_0 \).

Note that the random variable \( \psi = y_0 - \hat{y}_0 \) is normally distributed with mean 0 and variance \( \text{var}(\psi) = \text{var}(y_0 - \hat{y}_0) = \sigma^2 \left[ 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right] \)

because the future observation \( y_0 \) is independent of \( \hat{y}_0 \). If we use \( \hat{y}_0 \) to predict \( y_0 \), then the standard error of \( \psi = y_0 - \hat{y}_0 \), is the appropriate statistic on which to base a prediction interval. Thus, the \( 100(1-\alpha) \) percent prediction interval on a future observation at \( x_0 \) is

\[
\hat{y}_0 - t_{\alpha/2, n-2} \sqrt{\text{MS}_{\text{Res}} \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)} \leq y_0 \leq \hat{y}_0 + t_{\alpha/2, n-2} \sqrt{\text{MS}_{\text{Res}} \left( 1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} \right)}
\]

(3.2.66)

The Prediction interval (3.2.63) is of minimum width at \( x_0 = \bar{x} \) and widens as \( |x_0 - \bar{x}| \) increases. By comparing (3.2.67) with (3.2.58) the prediction interval at \( x_0 \) is always wider than the confidence interval at \( x_0 \) because the prediction interval depends on both the error from the fitted model and the error associated with future observation.

### 3.2.11 Multiple Applied Linear Regression Model:

The response \( y \) may be related to \( k \) regressor (or) predictor variables. The model is

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \ldots \beta_k x_k + \varepsilon
\]

(3.2.67)
is called Multiple Applied Linear Regression Model with k regressor. The parameter $\beta_j, j = 0, 1, \ldots, K$ are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of regressor variables $x_j$. The parameter $\beta_j$ represents the expected change in the response $Y$ per unit change in $x_j$ when all of the remaining regressor variables $x_i (i \neq j)$ are held constant. The parameters $\beta_j, j = 1, 2, \ldots, k$ are often called partial regression coefficients.

Multiple Applied Linear Regression Models are often used as empirical models (or) approximating functions. The true functional relationship between $y$ and $x_1, x_2, \ldots, x_k$ is unknown, but over certain ranges of the regression variables the Applied Linear Regression Model is an adequate approximation to the true unknown function.

### 3.2.12 Estimation of the model parameters:

(i) **Least – Squares estimation of the regression coefficients:**

The method of least squares can be used to estimate the regression coefficients in (3.2.67).

Let $y_i$ denote the $i^{th}$ observed response and $x_{ij}$ denote the $i^{th}$ observation (or) level of regressor $x_j$ and the error term $\epsilon$ in the model has $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$ and that the errors are uncorrelated.

The multiple linear regression models is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$$  \hspace{1cm} (3.2.68)

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, i = 1, 2, \ldots, n$$  \hspace{1cm} (3.2.69)

The least square function is

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \epsilon_i^2$$
\[
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2
\]  
(3.2.70)

The function \( S \) must be minimized with respect to \( \beta_0, \beta_1, \ldots, \beta_k \). The least-squares estimators of \( \beta_0, \beta_1, \ldots, \beta_k \) must satisfy:

\[
\left[ \frac{\partial S}{\partial \beta_0} \right]_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0
\]  
(3.2.71)

And

\[
\left[ \frac{\partial S}{\partial \beta_j} \right]_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0, \quad j = 1, 2, \ldots, k
\]  
(3.2.72)

We obtain the least-square normal equations:

\[
\begin{align*}
n \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\
M M M M M &
\end{align*}
\]  
(3.2.73)

There are \( \rho = k + 1 \) normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the least-squares estimators \( \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k \).

In matrix notation, the model given by is \( Y = X \beta + \varepsilon \)
Here $Y$ is an $n \times 1$ vector of the observations

$X$ is a $n \times k$ matrix of levels of the Regressor variables

$\beta$ is a $k \times 1$ vector of the regression coefficients.

$\varepsilon$ is an $n \times 1$ vector of random errors.

Now to find the vector of least – squares estimators, $\hat{\beta}$ that minimizes

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon^T \varepsilon = (y - X\beta)^T (y - X\beta)$$

(3.2.75)

And $S(\beta)$ may be expressed as

$$S(\beta) = y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X \beta$$

$$= y^T y - 2\beta^T X^T y + \beta^T X^T X \beta$$

Since $\beta^T X^T y$ is a $1 \times 1$ matrix (or) a scalar .The least squares estimators must satisfy

$$\left[ \frac{\partial}{\partial \beta} \right] = -2X^T y + 2X^T X \hat{\beta} = 0$$

Which simplifies to

$$X^T X \hat{\beta} = X^T y$$

(3.2.76)

These equations are the least – squares normal equations.

To solve the normal equations, multiply both side of (3.2.75) by the inverse of $X^T X$.Thus, the least – squares estimator of $\beta$ is
\[ \hat{\beta} = (X'X)^{-1} X'Y \quad (3.2.77) \]

Provided that the inverse matrix \((X'X)^{-1}\) exists. The \((X'X)^{-1}\) matrix will always exist if the regressors are linearly independent.

The matrix form of the normal equations (3.2.75) is identical to the scalar form (3.2.64) writing out (3.2.76) in detail, we obtain

\[
\begin{bmatrix}
\sum_{i=1}^{n} x_{1i} & \sum_{i=1}^{n} x_{1i} & \cdots & \sum_{i=1}^{n} x_{ki} \\
\sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i} x_{12} & \cdots & \sum_{i=1}^{n} x_{1i} x_{1k} \\
\vdots & \vdots & \ddots & \vdots \\
\sum_{i=1}^{n} x_{ki} & \sum_{i=1}^{n} x_{ki} x_{11} & \cdots & \sum_{i=1}^{n} x_{ki}^2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta}_0 \\
\hat{\beta}_1 \\
\vdots \\
\hat{\beta}_k
\end{bmatrix} =
\begin{bmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} x_{1i} y_i \\
\vdots \\
\sum_{i=1}^{n} x_{ki} y_i
\end{bmatrix}
\quad (3.2.78)
\]

Here \(X'X\) is a \(P \times P\) symmetric matrix

\(X'Y\) is a \(p \times 1\) column vector.

The diagonal elements of \(X'X\) are the sums of square of the elements in the columns of \(X\) and the off-diagonal elements are the sums of cross products of the elements in the columns of \(X\).

The fitted regression model corresponding to the levels of regressor variables \(X_i = [1, x_1, x_2, \ldots, x_k]\) is

\[ \hat{\gamma} = X \hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j \quad (3.2.79) \]

The vector of fitted values \(\hat{\gamma}_i\), corresponding to the observed values is

\[ \hat{\gamma} = X \hat{\beta} = X (X'X)^{-1} X'Y = Hy \quad (3.2.80) \]
The $n \times n$ matrix $H = X \left( X' X \right)^{-1} X'$ is usually called the hat matrix.

The difference between observed value $y_i$ and the corresponding fitted value $\hat{y}_i$ is the residual $e_i = y_i - \hat{y}_i$. The $n$ residuals may be conveniently written in matrix notation as

$$e = y - \hat{y} = y - X\hat{\beta} = I - Hy = (1 - H)y$$  \hspace{1cm} (3.281)

### 3.2.13 Properties of the Least – Square estimator:

The statistical properties of the least – squares estimator $\hat{\beta}$ may be easily demonstrated.

1. $E(\hat{\beta}) = E \left[ (X'X)^{-1} X'Y \right]$

   $$= E \left[ (X'X)^{-1} X'(X\beta + \varepsilon) \right]$$

   $$= E \left[ (X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon \right] = \beta$$  \hspace{1cm} (3.2.82)

Since $E(\varepsilon) = 0$ and $(X'X)^{-1} X'X = I$. Thus $\hat{\beta}$ is an unbiased estimator of $\beta$

2. The variance property of $\hat{\beta}$ is expressed by the covariance matrix

   $$Cov(\hat{\beta}) = E \left\{ (\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))' \right\}$$  \hspace{1cm} (3.2.83)
Which is \( P \times P \) Symmetric matrix whose \( j^{th} \) diagonal element is the variance of \( \hat{\beta}_j \) and who’s \((ij)^{th}\) off – diagonal element is the covariance between \( \hat{\beta}_i \) and \( \hat{\beta}_j \). The covariance of \( \hat{\beta} \) is 
\[
Cov(\hat{\beta}) = \sigma^2 \left( X^T X \right)^{-1}.
\]

Let \( C = \left( X^T X \right)^{-1} \), the variance of \( \hat{\beta}_j \) is \( \sigma^2 c_{ji} \) and the covariance between \( \hat{\beta}_i \) and \( \hat{\beta}_j \) is \( \sigma^2 c_{ij} \).

### 3.2.14 Estimation of \( \sigma^2 \):

In Simple Applied Linear Regression an estimator of \( \sigma^2 \) from the residual sum of squares.

\[
SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \varepsilon^T \varepsilon
\]  

(3.2.84)

Substituting \( \varepsilon = y - X \hat{\beta} \) we have

\[
SS_{Res} = (y - X \hat{\beta})^T (y - X \hat{\beta}) = y^T y - \hat{\beta} X^T y - y^T X \hat{\beta} + \hat{\beta} X^T X \hat{\beta} = y^T y - 2 \hat{\beta} X^T y + \hat{\beta} X^T X \hat{\beta}
\]

(3.2.86)

Since \( X^T X \hat{\beta} = X^T y \)

(3.2.87)
the above equation becomes

\[ SS_{Res} = y^\top y - \hat{\beta}^\top X^\top y \]  

(3.2.88)

The residual sum of squares has \( n - p \) degrees of freedom associated with \( P \) parameters are estimated in the regression model.

The residual mean square is

\[ MS_{Res} = \frac{SS_{Res}}{n - p} \]  

(3.2.89)

And the unbiased estimator of \( \sigma^2 \) is given by

\[ \hat{\sigma}^2 = MS_{Res} \]  

(3.2.90)

3.3 IMPORTANCE OF TIME SERIES FORECASTING:

Time series forecasting is very importance even in the financial as well as other domains which needs some kind of prediction over time. One of the reasons for its importance is preventing undesirable events by forecasting the event, identifying the circumstances preceding the event, and taking corrective action so the event can be avoided. It means that we can avoid low sales even if it is predicted correctly by implementing several sales development programs.

Forecasting also helps to reduce the impact of some unavoidable events by predicting them well in advance. We can save the lives of thousands in case if predictions of weather forecasts done properly. Many public and private organizations are concentrating more on improving their predictive power so that they can reduce the bad effects of us avoidable events.

Finally, many people, primarily in the financial markets, would like to profit from time series forecasting. Whether this is viable (or) not is most likely a never-to-be-resolved question, nevertheless many products are available for financial forecasting. Because of this
importance, a new branch of time series called financial time series has been come in to the picture and several models like ARCH and GARCH are developed as a result of this.

3.3.1 Character of Time Series data:

Many of the time series data exhibits the following characteristics depending on the nature and domain from which they adopted for forecasting. It is noted that all time series data may not contain all the following characteristics together every time.

(i) Limited availability of the data:

In many situations, it is observed that the historical data used for forecasting is very limited and the forecaster needs utilize this with appropriate assumptions. Some of the time series data like number of breaks or failures involves lot of amount spend on them to repeat the experiment and hence user will get only limited data.

(ii) Noise in the data:

Noise otherwise disturbances in the data are mainly because of two reasons. The first reason is erroneous data due to improper entries or improper metrics of measurement. For example, in one shift the weight of a particular commodity is measured in kilograms and in other shift it is measured in pounds.

The second reason for the noise is due to components of time series. For example, we may get very high or low values due to cyclical or seasonal fluctuations prevailing in that time period. The researcher needs to analyze the cause of the noise and then remove it from the data by appropriate method.

(iii) Non-stationary in the data:

A time series data is said to be stationary if it exhibits same statistical properties like mean and variance at each time point. Forecasting with non stationary data produce misleading or
erroneous predictions. After some data validation checks, one should always check the stationary in the data to proceed further.

(iv) **Forecasting technique selection:**

From statistics to artificial intelligence, there is a huge list of techniques. One of the simplest techniques is to search a data series for similar past events and use the matches to make a forecast. One of the most complex techniques is to train a model on the series and use the model to make a forecast. K-nearest-neighbor and neural networks are examples of the first and second techniques, respectively.

3.3.2 Assumptions of Time – Series models:

The random disturbances that are prevailing in the system are considered to be independent and normally distributed with mean zero and constant variance over time. Contingencies among scores over time are part of the model that is developed during identification and estimation. It is assumed that there are correlations in the sequence of observations over time that has been adequately modeled. This assumption is tested during the diagnostic phase of the time series modeling. Outliers among scores are sought before modeling and among the residuals once the model is developed.

(i) **Normality of Distributions of Residuals:**

A initial time series model is constructed for the given data and then normality of residuals is evaluated in time-series analysis. We will test the normalized plot of residuals for the model before evaluating an intervention. Transform the respective dependent variable if residuals are nonnormal. The normalized plot of residuals is examined as part of the diagnostic phase of modeling. The usual square root, logarithmic, or inverse transformations are appropriate in the event of nonnormally distributed residuals.

(ii) **Zero mean of residuals:**
The common understanding in any time series model is that sum of all residual terms should always approaches to zero. In other words, the sum of all positive residuals should be almost equal to the sum of all negative residuals. We can check this assumption graphically also by plotting the residuals along with zero line in the middle. The proper behavior should be such that we should find equal distribution of residual terms on the both sides of the zero line.

(iii) **Homogeneity of Variance of Residuals:**

After the appropriate times series model construction, we will construct a plot of standardized residuals versus predicted values to assess homogeneity of variance over time. If clear patterns are observed in the plot like upward trend or downward patterns, then we can decide that the homogeneity of variance is not clearly observed in the data. We can transform the dependent variable in this case to avoid the problem of heterogeneous variances. McCleary and Hay (1980) recommend a logarithmic transformation to remedy heterogeneity of variance.

(iv) **Independence of Residuals:**

During the diagnostic phase, once the model is developed and residuals are computed, we can test the independence of residuals. In general, ACF and PACF are the proper graphical tools to test the correlations between the residual terms. If the correlation patterns are observed at various lags in the ACFs and PACFs. Then the data that have not been properly modeled. Examine the ACFs and PACFs for other patterns and adjust the model accordingly.

(v) **Absence of Outliers:**

Outliers are defined as the set of observations that are highly inconsistent with the remainder of the time-series data and are found at irregular intervals of data. They can greatly affect the results of the analysis and must be treated properly in the beginning of the time series analysis. They sometimes show up in the original plot of the dependent variable against time,
but are often more noticeable after initial modeling is complete. We can find the outliers by examine the time-series plot before and after adjusting for autocorrelation and seasonality to identify obvious outliers. Unfortunately, there are no concrete guidelines to determine how discrepant a case must be to be labeled an outlier in a time-series analysis. An outlier is treated in the usual manner by checking the original data for errors, deleting the observation, replacing the observation with an imputed value.
3.4 BASIC MODELS OF TIME SERIES:

In this section, we discuss some of the basic models that are getting utilized in the analysis of time series data. These models are just to explain the overall significance of the model and the forecaster needs to apply further complex models to fulfill all his forecasting requirements.

(i) **Time series decomposition model:**

An important step in analyzing the time series data is to identify the data patterns overs certain period of time. Generally the following four components of data being identified in any time series model.

i) General trend: It is the pattern in the time series data which is the result of usual patterns without attributing to any specific reason. For example, in population data, we can find an increasing trend over a long period of time which is the general tendency of the population.

ii) Seasonal trends: These are the patterns which are due to particular seasons prevailing in that time period. These patterns repeat every year and they can be eliminated by some seasonal adjustments. Sales data of any product over a long period of time have tendency to show the seasonal fluctuations in the data.

iii) Cyclical trends: These are mainly attributed to business cycles which may distributed over more that one or two years. These fluctuations are having the stages of recovery, prosperity, recession and depression which are represented as follows.

iv) Random or irregular variations: These fluctuations area due some un expected factors in the real time like wars, earthquakes, strikes etc.
We can note that any time series can consists of all or some of the above components in its data. Time plot (data plotted over time) and seasonal plot (data plotted against the seasons in which they occur) are usually good tools to identify the trend patterns in the data.

A practical way of decomposing the time series data is either simple additive or multiplicative model with these four components as follows.

\[ Y_t = T_t + S_t + C_t + I_t \quad \text{or} \quad Y_t = T_t \times S_t \times C_t \times I_t \quad (3.4.1) \]

Where \( Y_t \) = Time series value at time period \( t \)

\( T_t \) = General trend component

\( S_t \) = Seasonal trend component

\( C_t \) = Cyclical trend component

\( I_t \) = Irregular trend component

### 3.5 AVERAGING METHODS:

The mean was discussed as an estimator that minimizes the mean squared error (MSE) of actual – minus – fitted– values, and it could have been shown the mean with all forecasting
methods. The data can be smoothed in many ways. In this section, several straightforward averaging methods, including the mean, simple moving averages, double moving averages, have been explained. The objective is to make use of past data to develop a forecasting system for future periods.

$$M_T = \frac{X_{T-N+1} + \ldots + X_{T-1} + X_T}{N} \quad (3.5.1)$$

$M_T$ is called the moving average of time $\tau$. It is noted that the moving averages at $t = T$ and $t = T-1$ are related as follows.

$$M_T = M_{T-1} + \frac{X_T - X_{T-N}}{N} \quad (3.5.2)$$

The cumulative forecast can be easy to calculate as $\hat{X}_T(\tau) = \tau M_T$.

### 3.5.1 Single Moving Averages Method:

Most common forecasting model in use today is the “single smoothing model”. One way to modify the influence of past data on the mean as a forecast is to specify at the outset. The term moving average is used to describe this procedure because as each new observation becomes available.

‘$Z$’ is data points and decision to use ‘$K$’ observations for each average as $X_1, X_2, \ldots, X_k, \ldots X_{k+1}, \ldots X_Z$ (Test Set)

<table>
<thead>
<tr>
<th>Time</th>
<th>Moving Average</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>$\bar{X} = \frac{X_1 + X_2 + \ldots + X_k}{K}$</td>
<td>$F_{k+1} = \bar{X} = \frac{\sum_{i=1}^{k} X_i}{K}$</td>
</tr>
<tr>
<td>$K+1$</td>
<td>$\bar{X} = \frac{X_2 + \ldots + X_{k+1}}{K}$</td>
<td>$F_{k+2} = \bar{X} = \frac{\sum_{i=2}^{k+1} X_i}{K}$</td>
</tr>
<tr>
<td>$K+2$</td>
<td>$\bar{X} = \frac{X_3 + \ldots + X_{k+2}}{K}$</td>
<td></td>
</tr>
</tbody>
</table>
To apply the single smoothing model, the forecaster selects a smoothing parameter $\alpha$. The forecast at time $t$ is found using $\alpha$, the current demand $x_t$ and one period a head forecast $\hat{x}_{t-1}(1)$. The forecasts become

$$\hat{x}_t(\tau) = \alpha x_t + (1-\alpha) \hat{x}_{t-1}(1) \quad \tau = 1, 2, \ldots$$

(3.5.3)

$$y_t = \hat{x}_t(1), \quad y_{t-1} = \hat{x}_{t-1}(1)$$

$y_t$ is called the smoothed average of $X_t (t = 1, 2, \ldots T)$. Algebraically, the moving average can be written as follows.

$$y_{t+1} = \frac{1}{t} \sum_{i=1}^{t} x_i$$

(3.5.4)

$$y_{t+2} = \frac{1}{t} \sum_{i=2}^{t+1} x_i$$

(3.5.5)

Comparing $y_{t+1}$ and $y_{t+2}$, it can be seen that $y_{t+2}$ requires dropping the value $X_1$ and the value $X_{t+1}$ as it becomes available, so that another way to write $y_{t+2}$ is

$$y_{t+2} = y_{t+1} + \frac{1}{t} \left(X_{t+1} - X_1\right)$$

(3.5.6)

It can be seen that new forecast $y_{t+2}$ is simply an adjustment of the immediately preceding forecast ($F_{t+1}$). This is the adjustment of the difference between $y_{t+1}$ and $y_t$. Clearly if $y$ is a big number, then the technique of moving averages as a forecasting procedure is not used often because the methods of exponential smoothing.

### 3.5.2. Double Moving Averages:
Two moving averages \((M_T \text{ and } M_T^{(2)})\) are measured at time period \(T\) and are used to generate the forecasts. The forecaster selects a parameter \(N\), which gives the number of time periods upon which the moving averages are calculated.

The moving averages are calculated by using,

\[ M_T = \frac{S_T + S_{T-1} + \ldots + S_{T-N+1}}{N} \]  

(3.5.7)

and \(M_T^{(2)} = \frac{M_T + M_{T-1} + \ldots + M_{T-N+1}}{N} \)  

(3.5.8)

\[ \Rightarrow M_T = M_{T-1} + \frac{X_T - X_{T-N}}{N} \]  

(3.5.9)

and \(M_T^{(2)} = M_{T-1}^{(2)} + \frac{M_T - M_{T-N}}{N} \)  

(3.5.10)

Forecast relation for the \(\tau^{th}\) future time period in given by

\[ \hat{x}_T(\tau) = \hat{a}_\tau + \hat{b}_\tau \]  

(3.5.11)

Where

\[ \hat{b}_\tau = \frac{2}{N-1}(M_T - M_T^{(2)}) \]  

(3.5.12)

\[ \hat{a}_\tau = 2M_T - M_T^{(2)} \]  

(3.5.13)

3.6 SMOOTHING METHODS:

Two or more observations taken from periods during with the same causal factor combing provides a smoothed value, or estimate. The term smoothed is used because such combinations tend to reduce randomness by allowing positive and negative random effects to partially offset each other.

3.6.1. Single Exponential Smoothing Technique:
It is one of the forecasting techniques with wide number of applications. In particular, it is used in sales and inventory forecasting. Exponential smoothing overcomes almost all the problems associated with the Moving averages forecasting technique. A moving average gives equal weightage to all values (or Items) in a time series, whether these are more recent ones or old ones. Under exponential smoothing the weights vary in Geometric progression with a constant ratio \( \alpha (0 < \alpha < 1) \) in such a way that the weights get lower and lower as we go back in the distant past. The ratio of the Geometric progression of the time series of weights is \( (1-\alpha) \) where \( \alpha \) is called the ‘smoothing coefficient’.

In Single Exponential Smoothing technique, the Forecaster selects a smoothing Parameter \( \alpha \), which must lie between 0 and 1.

The single exponential smoothing, forecasts can be obtained by using the equation.

\[
\hat{x}_t(\tau) = \alpha x_{t-1} + (1-\alpha) \hat{x}_{t-1}(1) \quad \tau = 1, 2, \ldots
\]  

(3.6.1)

By considering, \( S_T = \hat{x}_T(1) \) and \( S_{T-1} = \hat{x}_{T-1}(1) \) in the equation (3.6.1), one may write equation (2.6.1) as

\[
S_T = \alpha x_T + (1-\alpha) S_{T-1}
\]  

(3.6.2)

Here, \( S_T \) is called the smoothed average of \( x_t (t = 1, 2, \ldots, T) \) as of time \( T \);
\( x_T \) is the current demand; and \( \hat{x}_{T-1}(1) \) is the prior one period ahead forecast.

The forecast for the \( T^{th} \) time period is given by

\[
\hat{x}_T(T) = S_T
\]  

(3.6.3)

The cumulative forecast for the t-time periods is given by

\[
\hat{x}_T(T) = \tau S_T.
\]  

(3.6.4)

Another problem in the single exponential smoothing technique concerns the choice of the smoothing parameter at the early time periods with a higher \( \alpha \), more weight is given to the current demand entry, which reduces the bias caused from \( S_0 \) should it be initially set too low or too high.

Selection of \( \alpha \)
There are several techniques and among them, the following three alternative schemes are explained:

1. In the first technique, the forecaster selects a desired value for the smoothing parameter $\alpha$. At the early time periods, the forecaster chooses a smoothing parameter which is greater than $\alpha$. Suppose the value chosen at the $t^{th}$ time period is defined by $\alpha_t$. Then the values of $\alpha_t$ are obtained from $\alpha_t = \frac{1}{t}$, as long as $\frac{1}{t}$ is greater than $\alpha$.

2. Under the second technique, the forecaster selects specific values for $\alpha_t$ as

   $\alpha_t = 1.0$ for $t = 1$
   $0.3$ for $t = 2, 3, 4$
   $0.2$ for $t = 5, 6$
   $0.1$ for $t \geq 7$  \hspace{1cm} (3.6.5)

3. A third method uses one value of $\alpha$ throughout and begins each time period ($T$) by measuring the average ($\overline{x}$) of all prior demands. This is given by

   $$\overline{x} = \frac{x_1 + x_2 + \ldots + x_T}{T}$$  \hspace{1cm} (3.6.6)

   The smoothing parameter $\alpha$ is used to assign weights to the past demand entries.

   $$S_T = \alpha x_T + (1 - \alpha) S_{T-1}$$  \hspace{1cm} (3.6.7)

   $$= \alpha x_T + (1 - \alpha) \left[ \alpha x_{T-1} + (1 - \alpha) S_{T-2} \right]$$
   $$= \alpha x_T + \alpha (1 - \alpha) x_{T-1} + (1 - \alpha)^2 S_{T-2}$$
   $$= \alpha x_T + \alpha (1 - \alpha) x_{T-1} + (1 - \alpha)^2 \left[ \alpha x_{T-2} + (1 - \alpha) S_{T-3} \right]$$

   $$\Rightarrow S_T = \alpha x_T + \alpha (1 - \alpha) x_{T-1} + \alpha (1 - \alpha)^2 x_{T-2} + \ldots + \alpha (1 - \alpha)^T x_1 + (1 - \alpha)^T S_0$$
   $$\text{or } S_T = \alpha \sum_{j=0}^{T-1} (1 - \alpha)^j x_{T-j} + (1 - \alpha)^T S_0$$  \hspace{1cm} (3.6.8)

   Since, $0 < \alpha < 1$, the weights assigned get smaller as the demands become older.

   The sum of all the weights is given by
\[
\alpha \sum_{j=0}^{T-1} (1-\alpha)^j + (1-\alpha)^T = 1 \tag{3.6.9}
\]

Using \( T = \alpha \), we have \( \alpha \sum_{j=0}^{\infty} (1-\alpha)^j = 1 \)

The expansion of \( S_T \) becomes
\[
S_T = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j x_{T-j} \tag{3.6.10}
\]

Assume that the demands are independent with mean \( \mu \) and standard deviation \( \sigma \).

Since, \( \hat{x}_T(\tau) = S_T = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j x_{T-j} \),

We have, \( E(\hat{x}_T(\tau)) = \alpha \sum_{j=0}^{\infty} (1-\alpha)^j \mu = \mu \)

(3.6.11)

The forecast error is given by,
\[
e_{(\tau)} = \left[ \hat{x}_T(\tau) - x_{T+\tau} \right], \text{ Also} \quad E[e_{(\tau)}] = E[\hat{x}_T(\tau) x_{T+\tau}] = \mu - \mu = 0 \tag{3.6.12}
\]
and \( \text{Var} \left( \hat{x}_T(\tau) \right) = \alpha^2 \sum_{j=0}^{\infty} (1-\alpha)^{2j} \sigma^2 = \frac{\alpha}{2-\alpha} \sigma^2 \)

\[
\text{Var} \left( e_{(\tau)} \right) = \text{Var} \left( \hat{x}_T(\tau) - x_{T+\tau} \right) = \frac{\alpha}{2-\alpha} \sigma^2 + \sigma^2
\]

\[
\Rightarrow \text{Var} \left( e_{(\tau)} \right) = \frac{2}{2-\alpha} \sigma^2 \tag{3.6.13}
\]

(Warnings: For the \( \tau \) period cumulative forecast,

We have \( \hat{x}_T(\tau) = \tau S_T \)

\[
E \left( \hat{x}_T(\tau) \right) = \tau S_T \quad \text{and} \quad \text{Var} \left( \hat{x}_T(\tau) \right) = \tau^2 \text{Var} \left( S_T \right) = \frac{\tau^2 \alpha}{2-\alpha} \sigma^2 \tag{3.6.15}
\]

For the \( \tau \) - period cumulative forecast error,
\[ E_{\tau} = \hat{x}_{\tau}(\tau) - x_{\tau+1} \cdots - x_{\tau+\tau} \]
\[ = \tau \mu - \tau \mu = 0 \]

and  \[ \text{Var} \left( E_{\tau} \right) = \text{Var}\left[ \hat{x}_{\tau}(\tau) - x_{\tau+\tau} \cdots - x_{\tau+\tau} \right] \]
\[ = \frac{\tau^2 \alpha}{2 - \alpha} \sigma^2 + \tau \sigma^2 \]

\[ \Rightarrow \text{Var}(E_{\tau}) = \left[ \frac{\tau^2 \alpha + \tau(2 - \alpha)}{2 - \alpha} \right] \sigma^2 \quad (3.6.17) \]

3.6.2 Double Exponential Smoothing Technique:

It is an extension of single exponential smoothing technique and is applied with values that follow a trend demand pattern. A smoothing parameter \( \alpha \) is chosen and is used to assign weights to the past values of demand entries. As in the single smoothing, higher weightage is assigned to the more current demand entries. In the double smoothing for fewer data are required to maintain the system. Under double smoothing, two smoothing averages are defined by

\[ S_{\tau} = \alpha x_{\tau} + (1 - \alpha) s_{\tau-1} \quad (3.6.18) \]
\[ S_{\tau}^{(2)} = \alpha S_{\tau} + (1 - \alpha) S_{\tau-1}^{(2)} \quad (3.6.19) \]

where \( T \) is current time period.

The equations (3.6.18) and (3.6.19) can be rewritten as

\[ S_{\tau} = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j x_{\tau-j} \quad (3.6.20) \]
\[ S_{\tau}^{(2)} = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j S_{\tau-j} \quad (3.6.21) \]

We have, \[ E(S_{\tau}) = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j \left( a_{\tau} - j b \right) \]
\[ = a_{\tau} - \left( \frac{1 - \alpha}{\alpha} \right) b \]

And \[ E(S_{\tau}^{(2)}) = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j \left( a_{\tau} - \left( \frac{1 - \alpha}{\alpha} \right) b - j b \right) \]
\[ = a_r - \frac{2(1-\alpha)}{\alpha} b \]

By the moments method of estimation, \( S_r, S_r^{(2)}, \hat{a}_r \) and \( \hat{b} \) are substituted for \( E(S_r), E(S_r^{(2)}), a_r \) and \( b \) respectively. This gives

\[
S_r = \hat{a}_r - \frac{1-\alpha}{\alpha} \hat{b} \tag{3.6.22}
\]

\[
S_r^{(2)} = \hat{a}_r - \frac{2(1-\alpha)}{\alpha} \hat{b} \tag{3.6.23}
\]

\[
\Rightarrow \hat{a}_r = 2S_r - S_r^{(2)} \tag{3.6.24}
\]

\[
\hat{b} = \frac{\alpha}{1-\alpha} \left( S_r - S_r^{(2)} \right) \tag{3.6.25}
\]

Consider, \( \hat{a}_r = 2S_r - S_r^{(2)} \)

\[ = 2\left[ \alpha x_r + (1-\alpha)S_{r-1} \right] - \alpha^2 - \alpha(1-\alpha)S_{r-1} - (1-\alpha)S_{r-1}^{(2)} \]

Again, by substituting, \( S_{r-1} = \hat{a}_{r-1} - \left( \frac{1-\alpha}{\alpha} \right) \hat{b}_{r-1} \)

\[
S_{r-1}^{(2)} = \hat{a}_{r-1} - \frac{2(1-\alpha)}{\alpha} \hat{b}_{r-1} \]

We get, \( \hat{a}_r = x_r + (1-\alpha)^2 \left( \hat{a}_{r-1} + \hat{b}_{r-1} - x_r \right) \)

\[ \Rightarrow \hat{a}_r = x_r + (1-\alpha)^2 e_r \tag{3.6.26} \]

Similarly, we get

\[
\hat{b}_r = \frac{\alpha}{1-\alpha} \left( S_r - S_r^{(2)} \right)
\]

\[ = \frac{\alpha}{1-\alpha} \left[ \alpha x_r + (1-\alpha)S_{r-1} - \alpha^2 x_r - \alpha(1-\alpha)S_{r-1} - (1-\alpha)S_{r-1}^{(2)} \right] \]

\[ = \alpha^2 x_r + \alpha(1-\alpha)S_{r-1} - \alpha S_{r-1}^{(2)} \]

\[ = \hat{b}_{r-1} - \alpha^2 \left( \hat{a}_{r-1} + \hat{b}_{r-1} - x_r \right) \]
\[ \hat{b}_T = \hat{b}_{T-1} - \alpha^2 e_T \]  \hspace{1cm} (3.6.27)

Where \( e_T = (\hat{x}_{T-1}(1) - x_T) \) = one period ahead forecast error at time T.

And \( \hat{x}_{T-1}(1) = \hat{a}_{T-1} + \hat{b}_{T-1} \)  \hspace{1cm} (3.6.28)

Thus, it should be noted that the current one period ahead forecast error is used as the principal basis from which the coefficients \( a \) and \( b \) are updated.

The current forecast for the \( \tau \)th future time period is given by

\[ \hat{x}_T(\tau) = \hat{a}_T + \hat{b}_T \tau \]  \hspace{1cm} (3.6.29)

To obtain initial forecast, there is a need of the value of \( \hat{a}_0 \) and \( \hat{b}_0 \), but these are not available to the forecaster. There are three methods by which the system can be initialized.

(i) Initialize using \( x_1 \).
(ii) Estimate the level and slope at \( t=0 \).
(iii) Fit a straight line through the past demand entries.

In the first two methods, no demand entries prior to the current time period (\( T=1 \)) are used to initialize the system. The third method uses past demand entries in order to set started.

If at \( t=1 \), the current demand entry \( x_1 \) is used to yield the initial estimates. Here, \( \hat{a}_1 = x_1 \) and \( \hat{b}_1 = 0 \). This gives \( \hat{x}_1(\tau) = x_1 \) as the initial forecast. There after, as each new demand entry becomes available, new estimates of \( \hat{a}_T \) and \( \hat{b}_T \) are found in the standard manner.

### 3.6.3 Choosing the Smoothing Parameter:

In the single smoothing technique, the level at time T can be estimated by

\[ \hat{a}_T = x_T + (1-\alpha) \left[ \hat{x}_{T-1}(1) - x_T \right] \]  \hspace{1cm} (3.6.30)

In Double smoothing technique, the corresponding level is given by
\[ \hat{a}_t = x_t + (1 - \alpha)^2 \left[ \hat{T}_{t-1} - (1 - \alpha) \right] \]  

(3.6.31)

It should be noted that the above two relations for \( \hat{a}_t \) are quite similar. These are used to choose a value for \( \alpha \) in double smoothing which corresponds to an equivalent smoothing parameter (say \( \alpha_0 \)) in single smoothing.

For this purpose, the forecaster seeks an \( \alpha \) where

\[ (1 - \alpha)^2 = 1 - \alpha_0 \]

or \( \alpha = 1 - \sqrt{1 - \alpha_0} \)  

(3.6.32)

### 3.6.4 Single Smoothing Technique with Linear Trend:

Under this method, \( \hat{a} \) and \( \hat{b} \) are given by

\[ \hat{a}_t = \alpha x_t + (1 - \alpha) \left( \hat{a}_{t-1} + \hat{b}_{t-1} \right) \]  

(3.6.33)

\[ \hat{b}_t = \alpha \left( \hat{a}_t - \hat{a}_{t-1} \right) + (1 - \alpha) \hat{b}_{t-1} \]  

(3.6.34)

The forecast for future time period \( T + \tau \) is given by

\[ \hat{T}_t(\tau) = \hat{a}_t + \hat{b}_t \tau \]  

(3.6.35)

Equation (3.6.33) can be written as

\[ \hat{a}_t = \alpha x_t + (1 - \alpha) \left( \hat{a}_{t-1} + \hat{b}_{t-1} \right) \]

\[ = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j x_{t-j} + \sum_{j=1}^{\infty} (1 - \alpha)^j b_{t-1} \]  

(3.6.36)

And \[ E \left( \hat{a}_t \right) = \alpha \sum_{j=0}^{\infty} (1 - \alpha)^j (a_t - j b) + b \sum_{j=1}^{\infty} (1 - \alpha)^j \]

\[ = a_t \]  

(3.6.37)

Equation (3.6.34) can be written as

\[ \hat{b}_t = \alpha \left( \hat{a}_t - \hat{a}_{t-1} \right) + (1 - \alpha) \hat{b}_{t-1} \]

\[ = \alpha \hat{a}_t - \alpha^2 \sum_{j=0}^{\infty} (1 - \alpha)^j \hat{a}_{t-j-1} \]  

(3.6.38)

and \[ E \left( \hat{b}_t \right) = \alpha a_t - \alpha^2 \sum_{j=0}^{\infty} (1 - \alpha)^j \left[ a_t - (j+1) b \right] \]
It should be noted that these estimate $\hat{a}_t, \hat{b}_t$ and the forecasts are unbiased. The forecast for future time period $(T + \tau)$ is given by

$$\hat{x}_T(\tau) = \hat{a}_t + \hat{b}_t \tau$$  \hspace{1cm} (3.6.40)

Under linear regression method at the subsequent time period $t = T + 1$, the estimate of the coefficients are given by

$$\hat{a}_{t+1} = \alpha x_{t+1} + (1 - \alpha) (\hat{a}_t + \hat{b}_t)$$  \hspace{1cm} (3.6.41)

$$\hat{b}_{t+1} = \alpha (\hat{a}_{t+1} - \hat{a}_t) + (1 - \alpha) \hat{b}_t.$$  \hspace{1cm} (3.6.42)

### 3.6.5 Triple Smoothing Method:

It is an extension to double smoothing and is used as forecast model for items whose demand pattern in quadratic. A smoothing parameter $\alpha$ is selected by the forecaster and is used to assign weights to the past demands. In Triple Smoothing, the current time period $(T)$ is always assumed as the origin of time. In this way the level at time $t = T + K$ is

$$\mu_{T+m} = a + bm + \frac{1}{2} cm^2$$

Not that $\mu_t = a$.

$$\mu_{T-j} = a - bj + \frac{1}{2} cj^2$$  \hspace{1cm} (3.6.43)

and $\mu_{T+\tau} = a + \tau b + \frac{1}{2} c \tau^2$  \hspace{1cm} (3.6.44)  \hspace{1cm} The

co-efficient c is divided by 2 be consistent with the Taylor series expansion.

Suppose $X_t$ be current demand;

$\hat{a}_{t-1}, \hat{b}_{t-1}, \hat{c}_{t-1}$ be the estimates of a, b, c from the prior time period.

The one period ahead forecast at $T - 1$ is given by
\[
\hat{x}_{T-1}(1) = \hat{a}_{T-1} + \hat{b}_{T-1} + \frac{1}{2} \hat{c}_{T-1}
\]  
(3.6.45)

Also, the current one period ahead forecast error is given by
\[
e_T = \hat{x}_{T-1}(1) - x_T
\]  
(3.6.46)

Now, the updated coefficient estimates can be written as
\[
\hat{\alpha}_T = x_T + (1 - \alpha)^3 e_T
\]  
(3.6.47)
\[
\hat{b}_T = \hat{b}_{T-1} + \hat{c}_{T-1} - 1.5\alpha^3 (2 - \alpha) e_T
\]  
(3.6.48)
\[
\hat{c}_T = \hat{c}_{T-1} - \alpha^3 e_T
\]  
(3.6.49)

With \(\hat{a}_T, \hat{b}_T\) and \(\hat{c}_T\) now available, the forecast for the \(\tau^{th}\) future time period can be generated. As before, the forecast for the sum of the next \(\tau\) time periods is given by
\[
\hat{x}_T(\tau) = \hat{x}_T(1) + \hat{x}_T(2) + \ldots + \hat{x}_T(\tau)
\]  
(3.6.50)

### 3.7 AUTOREGRESSIVE MOVING AVERAGE (ARMA) MODELING IN FORECASTING:

There are mainly three main steps in ARMA modeling which are given by:

(i) Check the series for stationary and if necessary transform the series to induce stationary;

(ii) From the autocorrelation properties of the transformed series, choose a few ARMA specifications for estimation and testing in order to arrive at a preferred specification with white noise residuals;

(iii) Calculate forecasts over a relevant time horizon from the preferred specification.

For the purpose of inference in ARMA models, Box - Jenkins methodology has to be applied and this methodology based on the basic concepts of stationary and non-stationary time series, stationary models, Autocorrelation and the lag operator.
A stationary time series is one where all the demands are in equilibrium with a common mean $\mu$ and a common variance $\sigma^2$. A Non-stationary time series is a series where the above conditions do not apply. The various types of forecasting models that are a part of the Box-Jenkins family of models. Two types of stationary models are given by (i) Autoregressive models and (ii) Moving Average Models. Combination of these stationary models gives another stationary model namely the mixed Autoregressive Moving Average Models. The various types of stationary models have been denoted by the notation $(p,q)$ where $p$ gives the number of coefficients in the AR model and $q$ denotes the corresponding number in the MA model.

3.7.1 Autoregressive Models $(P,O)$:

In the Autoregressive model, the current entry $y_t$ is related in a linear manner to its $p$ most current entries and to an unknown white noise by the relation.

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \epsilon_t$$  \hspace{1cm} (3.7.1)

The coefficient $\alpha_i$ $(i=1,2,\ldots,p)$ gives the assigned weight to the $i^{th}$ prior entry. The white noise disturbance $\epsilon_t$ is a random with mean zero and variance $\sigma^2$.

The commonly applied autoregressive models are Autoregressive Model $(1,0)$ or AR(1) model and Autoregressive model $(2,0)$ or AR (2) models.

(a) AR(1) Process and properties:

Consider a specification for AR (1) process as

$$y_t = \delta + \alpha y_{t-1} + \epsilon_t$$  \hspace{1cm} (3.7.2)

Where $\epsilon_t$ is a white noise process.

Define the lag operator as $L (y_t) = y_{t-1}$  \hspace{1cm} (3.7.3)
\[ L^2 (y_t) = L \left[ L(y_t) \right] = L(y_{t-1}) = y_{t-2} \]
and \((1 - L)y_t = y_t - y_{t-1} = \Delta y_t, \]
\[ L \left( 1 - L \right) y_t = y_{t-1} - y_{t-2} = \Delta y_{t-1} \quad (3.7.4) \]

Where \( \Delta \) is the first difference operation. By using the Lag operator \( L \), one can write the equation (3.7.2) as \((1 - \alpha L)y_t = \delta + \epsilon_t, \]
\[ \Rightarrow y_t = \left( 1 + \alpha L + \alpha^2 L^2 + \ldots \right) \delta + \left( \epsilon_t + \alpha \epsilon_{t-1} + \alpha^2 \epsilon_{t-2} + \ldots \right) \quad (3.7.5) \]

Since, a constant '\( \delta \)' has the same value at all periods; application of the lag operator, any number of times simply reproduces the constant. Thus, one can write

\[ y_t = \left( 1 + \alpha + \alpha^2 + \ldots \right) \delta + \left( \epsilon_t + \alpha \epsilon_{t-1} + \alpha^2 \epsilon_{t-2} + \ldots \right) T \ldots \quad (3.7.6) \]

Provided \(|\alpha| < 1.\]
\[ \Rightarrow E(y_t) = \left( 1 + \alpha + \alpha^2 + \ldots \right) \delta \quad (3.7.7) \]

This expectation only exists if the infinite geometric series on the R.H.S. has a limit. The necessary and sufficient condition is \(|\alpha| < 1.\]

Now, the expectation is given by
\[ E(y_t) = \frac{\delta}{1 - \alpha} = \mu \quad (3.7.8) \]
It shows that \( Y_t \) series has a constant unconditional mean \( \mu \) at all points, independent of time.

To determine the variance of \( y_t \) one may define
\[ (y_t - \mu) = \epsilon_t + \alpha \epsilon_{t-1} + \alpha^2 \epsilon_{t-2} + \ldots \quad (3.7.9) \]
\[ \Rightarrow Var(y_t) = E[y_t - \mu]^2 = E \left[ \epsilon_t^2 + \alpha^2 \epsilon_{t-1}^2 + \alpha^4 \epsilon_{t-2}^2 + \ldots + 2\alpha \epsilon_t \epsilon_{t-1} + 2\alpha^2 \epsilon_t \epsilon_{t-2} + \ldots \right] \quad (3.7.10) \]

By considering the assumptions,
\[ E(\epsilon_t) = 0, \]
\[ E(\epsilon_t^2) = \sigma^2, \forall t \quad (3.7.11) \]
Equation (3.7.10) yields
\[ \sigma^2_y = \text{Var}(y_t) = E\left((y_t - \mu)^2\right) = \frac{\sigma^2_e}{1 - \alpha^2} \] (3.7.12)

Using the equation (3.7.12), rewrite the equation (3.7.2) as
\[ (y_t - \mu) = \alpha [y_{t-1} - \mu] + \epsilon_t \]
or
\[ x_t = \alpha x_{t-1} + \epsilon_t \] (3.7.13)

Where \( x_t = y_t - \mu \)

Now, one may have,
\[ E(x_t^2) = \alpha^2 E(x_{t-1}^2) + E(\epsilon_t^2) + 2 \alpha E(x_{t-1} \epsilon_t) \] (3.7.14)

Since, \( x_{t-1} \) depends only \( \epsilon_{t-1}, \epsilon_{t-2}, \ldots, \epsilon_t \), and \( x_{t-1} \) is uncorrelated with all the previous values by the white noise disturbance assumption, we have
\[ E(x_{t-1} \epsilon_t) = 0 \]. Also, if \( \alpha \) satisfies the stationary condition \( |\alpha| < 1 \), then we have,
\[ \sigma^2_y = E(x_t^2) = E(x_{t-1}^2) = \ldots. \] (3.7.15)

Thus the equation (3.7.14) gives
\[ \sigma^2_y = \alpha^2 \sigma^2_y + \sigma^2_e \] (3.7.16)
\[ (1 - \alpha^2) \sigma^2_y = \sigma^2_e \] (or)
\[ \sigma^2_y = \frac{\sigma^2_e}{1 - \alpha^2} \] (or) \( \sigma^2_y \), which gives equation (3.7.12).

From equation (3.7.13), we have
\[ E(x_t x_{t-1}) = \alpha E(x_{t-1}^2) + E(x_{t-1} \epsilon_t) \] (3.7.17)

Denote the “Autocovariance coefficients” by
\[ \gamma_o = E(x_t x_{t-1}) \], the equation (3.7.17) gives
\[ \gamma_1 = \alpha \gamma_o \] (3.7.18)
Where \( r_0 \) is another symbol for \( \sigma^2_y \).

In the similar fashion, multiplying equation (3.7.13) by \( x_{t-2} \), followed by taking expectations, gives,
\[ \gamma_2 = \alpha \gamma_1 \] (3.7.19)
In general, we have \( \gamma_k = \alpha \gamma_{k-1} = \alpha^k \gamma_0, \ k=1, 2 \ldots \)

The Autocorrelation coefficients for a stationary series are given by

\[
\rho_k = \frac{E(x_t x_{t-k})}{\sqrt{Var(x_t) \sqrt{Var(x_{t-k})}}} = \frac{\gamma_k}{\gamma_0}
\]  

(3.7.20)

Hence, the autocorrelation coefficients for AR (1) process are given by

\[
\rho_k = \alpha, \ \rho_{k-1} = \alpha^k, \ k=1, 2, \ldots
\]  

(3.7.21)

The formula given is known as the Autocorrelation Function (ACF) of the AR (1) process.

### 3.7.3 Second Order Autoregressive [AR(2)] Process:

The second order autoregressive AR (2) process can be specified as

\[
y_t = \delta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \epsilon_t\]  

(3.7.22)

By assuming stationarity, the unconditional mean is given by \( \mu = \frac{\delta}{1-\alpha_1-\alpha_2} \).

Substituting \( x_t = y_t - \mu \), the equation (3.7.22) can be written as

\[
x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \epsilon_t\]  

(3.7.23)

\[
\Rightarrow E\left[x_t^2\right] = \alpha_1 E\left(x_{t-1}\right) + \alpha_2 E\left(x_{t-2}\right) + E\left(x_t \epsilon_t\right)
\]

or \( \gamma_0 = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + E\left(x_t \epsilon_t\right) \)  

(3.7.24)

From equation (3.7.24) one may obtain \( E\left(x_t \epsilon_t\right) = \sigma_e^2 \)

\[
\therefore \gamma_0 = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \sigma_e^2
\]  

(3.7.25)

Again from equation (3.7.23) we get,

(i) \( E\left[x_t x_{t-1}\right] = \alpha_1 E\left(x_{t-1}^2\right) + \alpha_2 E\left[x_{t-1} x_{t-2}\right] + E\left[x_{t-1} \epsilon_t\right] \)

(or) \( \gamma_1 = \alpha_1 \gamma_0 + \alpha_2 \gamma_1 \)  

(3.7.26)

(ii) \( E\left[x_t x_{t-2}\right] = \alpha_1 E\left[x_{t-1} x_{t-2}\right] + \alpha_2 E\left[x_{t-2}^2\right] + E\left[x_{t-2} \epsilon_t\right] \)

(or) \( \gamma_2 = \alpha_1 \gamma_1 + \alpha_2 \gamma_0 \)  

(3.7.27)

\[
\left[Q \gamma_o = \sigma_e^2\right]
\]
Substituting expressions for $\gamma_1$ and $\gamma_2$ in (3.7.25) gives

$$r_o = \frac{(1-\alpha_2)^2 \sigma_e^2}{(1+\alpha_2)(1-\alpha_1-\alpha_2)(1+\alpha_1-\alpha_2)}$$  \hspace{1cm} (3.7.28)

Under stationary, this variance $\gamma_o$ must be a constant and positive number. This requires, each term for the parantheses to be positive. Thus, the stationarity conditions for the AR (2) process are given by

$$\begin{align*}
\alpha_2 + \alpha_1 &< 1 \\
\alpha_2 - \alpha_1 &< 1 \\
|\alpha_2| &< 1
\end{align*}$$  \hspace{1cm} (3.7.29)

The relations between autocovariances in (3.7.26) and (3.7.27) may be restated in terms of autocorrelation coefficients as

$$\begin{align*}
\rho_1 &= \alpha_1 + \alpha_2 \rho_1 \\
\rho_2 &= \alpha_1 \rho + \alpha_2
\end{align*}$$  \hspace{1cm} (3.7.30)

Equations (3.7.30) are called the yule-walker equations for AR (2) process.

From (3.7.30), the first two autocorrelation coefficients are given by

$$\rho_1 = \frac{\alpha_1}{1-\alpha_2} \text{ and } \rho_2 = \frac{\alpha_1^2}{1-\alpha_2} + \alpha_2$$  \hspace{1cm} (3.7.31)

Now, the ACF for the AR (2) process is given by

$$\rho_k = \alpha_1 \rho_{k-1} + \alpha_2 \rho_{k-2}, \hspace{1cm} k = 3, 4, \ldots.$$

Consider the equation (3.7.23) as

$$x_i = \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + \varepsilon_i$$  \hspace{1cm} (3.7.33)

Using the lag operator $L(x_i) = x_{i-1}$, $L^2(x_i) = x_{i-2}$, $L^k x_i = x_{i-k}$, equation (3.7.33) may be expressed as

$$A(L)x_i = \varepsilon_i$$  \hspace{1cm} (3.7.34)

Where, $A(L) = 1-\alpha_1 L - \alpha_2 L^2$  \hspace{1cm} (3.7.35)

By expressing this quadratic equation as the product of two factors one may write

$$A(L) = 1-\alpha_1 L - \alpha_2 L^2 = (1-\lambda_1 L)(1-\lambda_2 L)$$  \hspace{1cm} (3.7.36)

The relation between the $\alpha$'s and the $\lambda$'s are given by
\[ \lambda_1 + \lambda_2 = \alpha \text{ and } \lambda_1\lambda_2 = -\alpha \]

(3.7.37)

Here, \( \lambda_1 \) and \( \lambda_2 \) are the roots of quadratic equation \( \lambda^2 - \alpha \lambda - \alpha = 0 \) this equation is known as the characteristic equation of the AR (2) process. The two roots are given by

\[ \lambda_1, \lambda_2 = \frac{\alpha \pm \sqrt{\alpha^2 + 4\alpha}}{2} \]

(3.7.38)

Now the \( A^{-1}(L) \) may be written as

\[ A^{-1}(L) = \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} = \frac{c}{1-\lambda_1 L} + \frac{d}{1-\lambda_2 L} \]

(3.7.39)

Where, \( C = -\lambda_1 / (\lambda_2 - \lambda_1) \) and \( d = \frac{\lambda_2}{(\lambda_2 - \lambda_1)} \)

Thus equation (3.7.34) can be written as

\[ x_t = A^{-1}(L)\varepsilon_t \]

(3.7.40)

Or \[ x_t = \frac{c}{1-\lambda_1 L}\varepsilon_t + \frac{d}{1-\lambda_2 L}\varepsilon_t \]

(3.7.41)

The stationarity of the AR (2) process requires

\[ |\lambda_1| < 1 \text{ And } |\lambda_2| < 1. \]

Restating these conditions in terms of \( \alpha \)'s gives the stationarity conditions for AR (2) process already given in (3.7.29).

AR (1), AR (2)… AR (p) processes may be often denoted by AR (1, 0), AR (2, 0)….AR (p, 0).

The \( p^{th} \) order Autoregressive model AR (p,0) can be written as

\[ y_t = \delta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \varepsilon_t \]

(3.7.42)

### 3.7.4 Moving Average (MA) Processes:

In the moving average model, the current entry \((y_t)\) is related to the q most current one-period-ahead forecast errors \((\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-q})\) and the current noise \( \varepsilon_t \).
Consider the AR (1) process specification as

\[ x_t = \alpha x_{t-1} + \epsilon_t \]  
(3.7.43)

Where \( x_t = y_t - \mu \) and \( E(y_t) = \mu = \frac{\delta}{1 - \alpha} \).

AR (1) in equation (3.7.41) may be inverted to give

\[ x_t = \epsilon_t + \alpha \epsilon_{t-1} + \alpha^2 \epsilon_{t-2} + \ldots \]  
(3.7.44)

This is a MA process of an infinite order.

Thus, is a pure MA process, the study variable \( y_t \) is expressed solely in terms of the current and past white noise disturbances.

The MA (1) process may be specified as

\[ x_t = \epsilon_t - \beta_1 \epsilon_{t-1} \]  
(3.7.45)

It can be easily shown that the autocovariances are

\[
\begin{align*}
\gamma_0 &= (1 + \beta_1^2) \sigma^2_e \\
\gamma_1 &= -\beta_1 \sigma^2_e \\
\gamma_2 &= \gamma_3 = \ldots = 0
\end{align*}
\]
(3.7.46)

Which gives the Auto Correlation Coefficients as

\[
\rho_1 = \frac{-\beta_1}{1 + \beta_1^2} \\
\rho_2 = \rho_3 = \ldots = 0
\]  
(3.7.47)

The MA (1) process may be inverted to give \( \epsilon_t \) as an infinite series in \( x_t, x_{t-1} \ldots \) namely

\[ \epsilon_t = x_t + \beta_1 x_{t-1} + \beta_1^2 x_{t-2} + \ldots \]  
(3.7.48)

\[ \Rightarrow x_t = -\beta_1 x_{t-1} - \beta_1^2 x_{t-2} + \ldots + \epsilon_t \]  
(3.7.49)

The properties of a pure MA process are the converse of those of a pure AR process.

Equation (3.7.47) exists if \( |\beta_1| < 1 \).
This condition $| \beta | < 1$ is known as Invertibility condition. It is similar to the stationary condition for an AR (1) process, but stationary of the MA (1) process itself does not impose any condition on $\beta$.

Similar results can be obtained for a stationary MA (2) process. In general, for a stationary MA (q) process, the first q autocorrelation coefficients will be non-zero and the rest zero.

$\text{MA}(1), \text{MA}(2), ..., \text{MA}(q)$ processes may be often denoted by MA (0,1), MA (0,2), ..., MA (0,q).

The q\textsuperscript{th} order moving Average model MA (o, q) can be written as

$$x_t = -\beta_1 \epsilon_{t-1} - \beta_2 \epsilon_{t-2} - \ldots - \beta_q \epsilon_{t-q} + \epsilon_t.$$  \hspace{1cm} (3.7.50)

### 3.7.5 Autoregressive Moving Average (ARMA) Processes:

In the ARMA model, the entry $Y_t$ is related to the P most recent entries $(y_{t-1}, y_{t-2}, ..., y_{t-p})$ and the q most recent forecast errors $(\epsilon_{t-1}, \epsilon_{t-2}, ..., \epsilon_{t-q})$ and the current white noise $\epsilon_t$. It gives

$$y_t = \delta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \ldots + \alpha_p y_{t-p} + \epsilon_t - \beta_1 \epsilon_{t-1} - \beta_2 \epsilon_{t-2} - \ldots - \beta_q \epsilon_{t-q}.$$  \hspace{1cm} (3.7.51)

This can be rewritten as

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \ldots + \alpha_p x_{t-p} + \epsilon_t - \beta_1 \epsilon_{t-1} - \beta_2 \epsilon_{t-2} - \ldots - \beta_q \epsilon_{t-q}.$$  \hspace{1cm} (3.7.52)

Where $x_t = y_t - E(y_t)$. It can be represented by ARMA (p,q).

The general ARMA (p, q) process can be written as

$$A(L) x_t = B(L) \epsilon_t.$$  \hspace{1cm} (3.7.53)

Where $A(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \ldots - \alpha_p L^p$

$$B(L) = 1 - \beta_1 L - \beta_2 L^2 - \ldots - \beta_q L^q$$  \hspace{1cm} (3.7.54)

ARMA (p,q) process may alternatively expressed as a pure infinite order AR process or a pure infinite order MA process in the following manner:

$$B^{-1}(L) A(L) x_t = \epsilon_t.$$  \hspace{1cm} (3.7.55)
or \( x_t = A^{-1}(L)B(L)\varepsilon_t \) \hspace{1cm} (3.7.56)

We have ARMA (1, 1) as \( x_t = \alpha x_{t-1} + \varepsilon_t - \beta \varepsilon_{t-1} \) \hspace{1cm} (3.7.57)

By squaring both sides of (3.7.47) and taking expectations, one may obtain by some manipulations,

\[
\text{Var}(x_t) = \sigma^2 = \gamma_o = \left[ \frac{1-2\alpha\beta + \beta^2}{1-\alpha^2} \right] \sigma^2_e
\]

Multiplying through equation (3.7.47) by \( x_{t-1} \) and taking expectations, gives

\[
\gamma_1 = \alpha \gamma_o - \beta \sigma^2_e = \left[ \frac{(\alpha - \beta)(1-\alpha \beta)}{1-\alpha^2} \right] \sigma^2_e
\]

The higher order covariances are given by

\[
\gamma_k = \alpha \gamma_{k-1}, \quad k=2, 3\ldots \]

The ACF of the ARMA (1, 1) process is given by

\[
\rho_1 = \frac{(\alpha - \beta)(1-\alpha \beta)}{1-2\alpha\beta + \beta^2}
\]

And \( \rho_k = \alpha \rho_{k-1}, \quad k=2,3,\ldots \)

In general, in a ARMA (p, q) process, there will be p autoregressive and q moving average terms

### 3.8 AUTOREGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA) PROCESS:

In the case of AR, MA, ARMA processes, the time series involved are weakly stationary. The mean and variance for a weakly stationary time series are constant and covariance is time invariant. But many of the economic time series are integrated.

If we difference a non-stationary time series ‘d’ times to make it stationary and then apply the ARMA (p,q) model to it, yields an Autoregressive integrated moving Average time series namely ARIMA (p,d,q). Here ‘p’ denotes the number of autoregressive terms; ‘d’ denotes the number of times the series has to the differenced before it becomes stationary and ‘q’ denotes the number of moving average terms. Thus, an ARIMA (2, 1, 2) time series has to be differenced once (d=1) before it becomes stationary and the first differenced stationary time series can be
modeled as an ARMA (2,2) process. We have ARMA (p, o, q) = ARMA (p, q). It should be noted that an ARIMA (p, o, o) process means a purely AR (p) stationary process; an ARIMA (o, o, q) means a purely MA (q) stationary process.

For instance,

\[
\text{ARIMA (1, d, o): } x_t = \alpha_1 x_{t-1} + \varepsilon_t
\]

(3.8.1)

\[
\text{ARIMA (2, d, o): } x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \varepsilon_t
\]

(3.8.2)

\[
\text{ARIMA (o, d, 1): } x_t = \varepsilon_t - \beta_1 \varepsilon_{t-1}
\]

(3.8.3)

\[
\text{ARIMA (o, d, 2): } x_t = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2}
\]

(3.8.4)

\[
\text{ARIMA (1, d, 1): } x_t = \alpha_1 x_{t-1} - \beta_1 \varepsilon_{t-1} - \varepsilon_t
\]

(3.8.5)

3.9 THE BOX AND JENKINS METHODOLOGY FOR IDENTIFICATION, ESTIMATION AND TESTING OF ARIMA MODELS:

Box and Jenkins (1968) have suggested forecasting technique that seeks in a systematic manner, the forecasting model that is best suited to each item under investigation. This methodology is highly sophisticated in both mathematical and computational aspects.

The objective of Box-Jenkins methodology is identifies and estimates a statistical model which can be interpreted as having generated the sample data. If this estimated model is then to be used for forecasting, we must assume that the features of this model are constant through time, and particularly over future time periods. By this method, one can identify the time series which follows a purely AR process or a purely MA process or an ARMA process or an ARIMA process with the values of \( p \), \( d \), and \( q \). The Box-Jenkins methodology consists of four steps, namely: (i) Identification (ii) Estimation (iii) Diagnostic checking and (iv) Forecasting.

(i). Identification

The objective of this first stage is to select the forecasting model that seems most appropriate to the time series under study. The data are used to generate a series of sample autocorrelation functions. These are now compared to certain theoretical autocorrelation
functions from known forecasts models to seek the best match between the sample and theoretical results. Under this stage, the forecast model is identified and selected.

In the ARIMA (p, d, q) models, we have three parameters, namely,

\[ p = \text{order of the AR component.} \]
\[ d = \text{number of differences required for stationary.} \]
\[ q = \text{order of the MA component.} \]

Typically d is zero or one; or very occasionally two; and one seeks a parsimonious representation with low values of p and q.

The choice of the order of p and q may be obtained by a numerical procedure suggested by Hannan and Rissanen (1983). This method has three steps. In the first step, some pure AR process of fairy higher order are estimated by OLS, which is not unreasonable since an unknown ARMA process is equivalent to an infinite AR process. Under second step, the regression model with the smallest value of the Akaike Information Criterion (AIC) is selected. Then the OLS residuals \( \{e_t\} \) from this regression model are taken as estimates of the unknown \( \varepsilon \)'s in an ARMA model. In the third step, a number of ARMA models are fitted using these estimated residuals.

For instance, if an ARMA (2,1) is fitted, the regression is

\[
y_t = \delta + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t - \beta_1 e_{t-1} + u_t
\]

(3.9.1)

Where \( u_t \) is error

Regression models of type (3.9.1) are fitted by OLS for various values of \( p \) and q. The residual variance \( \hat{\sigma}^2_{p,q} \) is obtained and the specification has been chosen that has the lowest value of

\[
SC = \ln \hat{\sigma}^2_{p,q} + (p + q) \frac{\ln(n)}{n}
\]

(3.9.2)

Where SC is Schwarz Criterian.
It should be noted that even though the Hannan - Rissanen procedure yields numerical estimates of an ARMA model, they are the by-product of an identification process and are not meant as final estimates of the relevant coefficients. Granger and Newbold (1986) have discussed in detail about Hannan and Rissanen Procedure. To identify the model that best represents the series, one may find estimates of the parameters (p, d, q) from an analysis of the historical demands. The concept of the Autocorrelation function (ACF) may be used as a tool for identification. The ACF for the $d^{th}$ difference is denoted by $r_j(d)$, $j = 0, 1, 2, \ldots k$, and $d = 0, 1, 2, \ldots$.

Here, the value of $k$ depends on the number of demand entries ($n$) available in the data. Mabert (1975) suggested a rule to choose $k \leq \frac{n}{4}$.

The autocorrelation with a lag of $k$ for $d = 0$ is given by

$$r_j(0) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_{i-j} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$  \hspace{1cm} (3.9.3)

Where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Further, for $d = 1$, we have,

$$r_j(1) = \frac{\sum_{i=2}^{n} (\Delta x_i - \bar{\Delta} x)(\Delta x_{i-j} - \bar{\Delta} x_i)}{\sum_{i=2}^{n} (\Delta x_i - \bar{\Delta} x_i)^2}$$  \hspace{1cm} (3.9.4)

Where $\bar{\Delta} x_i = \frac{\sum_{i=2}^{n} \Delta x_i}{n-1}$  \hspace{1cm} (3.9.5)

The value for $d$ may be chosen corresponds to the smallest value of ACF, which yields a stationary time series. Naturally, a larger value of ACF for many time periods characterizes a non stationary time series.
Having selected $d$, the forecaster investigates $r_j(d) \ j=0,1,2,\ldots,k$ to seek which values of $p$ and $q$ are most appropriate. This is done by comparing correlogram, a plot of the function $r_j(d)$ with known functions ($\rho_j$) for given values of $p$ and $q$. The best match is chosen and the corresponding values of $p$ and $q$ are selected.

(ii) Estimation:

Having identified the appropriate $p$ and $q$ values, the next stage is to estimate the parameters of the autoregressive and moving average terms included in the model. These estimates can be obtained by using either OLS method or non-linear least squares method.

Software statistical packages typically offer least squares, whether linear or nonlinear, or maximum likelihood; whether conditional or full estimation methods for ARMA models.

Consider the AR (1) specification as

$$y_t = \delta + \alpha y_{t-1} + \epsilon_t \quad (3.9.5)$$

Where $\epsilon_t$ is white noise OLS method can be applied to obtain estimators for the parameters. Here, the value of $y_1$ is taken as given and summation run over $t = 2, 3, \ldots, n$. It can be shown that the OLS estimator as a conditional Maximum likelihood estimator.

If we take $y_1$ as given, the conditional likelihood for the remaining $(n-1)$ observations may be written as

$$L = f(y_2, y_3, \ldots, y_n / y_1)$$

$$= f(y_2 / y_1) f(y_3 / y_1) \ldots f(y_n / y_1) \quad (3.9.6)$$

If we assume normal distribution for white noise $\epsilon_t$, we have,

$$f(y_t / y_{t-1}) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} (y_t - \delta - \alpha y_{t-1})^2 \right] \quad (3.9.7)$$

The conditional log–likelihood function is given by
\( \ln L = \text{constant} - \left(\frac{n-1}{2}\right) \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=2}^{n} (y_t - \delta - \alpha y_{t-1})^2 \). Maximizing \( \ln L \) with respect to \( \delta \) and \( \alpha \) gives the OLS estimates for \( \delta \) and \( \alpha \) respectively.

Consider the unconditional likelihood function,

\[
L_u = f(y_1) L
\]

(3.9.8)

Under the assumption of the AR (1) process,

\[
y_1 \sim N\left(\frac{\delta}{1-\alpha}, \frac{\sigma^2}{1-\alpha^2}\right)
\]

(3.9.9)

Thus, \( l_n f(y_1) = \text{constant} + \frac{1}{2} \ln \left(1 - \alpha^2\right) - \frac{1}{2} \ln \sigma^2 - \left(\frac{1 - \alpha^2}{2\sigma^2}\right) \left(y_1 - \frac{\delta}{1-\delta}\right)^2 \)

(3.9.10)

The unconditional log-likelihood is given by

\[
\phi = \ln f(y_1) + \ln L = \text{constant}
\]

\[
-\frac{n}{2} \ln \sigma^2 + \frac{1}{2} \ln \left(1 - \alpha^2\right) - \left(\frac{1 - \alpha^2}{2\sigma^2}\right) \left(y_1 - \frac{\delta}{1-\delta}\right)^2 - \frac{1}{2\sigma^2} \sum_{t=2}^{n} (y_t - \delta - \alpha y_{t-1})^2
\]

(3.9.11)

Taking first order derivatives with respect to \( \delta \) and \( \alpha \), no longer yields linear equations in these parameters. Iterative techniques may be used to maximize equation (3.9.10) to obtain full maximum likelihood estimators. In the similar manner, the higher order AR schemes may be fitted.

However, the fitting of MA schemes is more complicated. Even the lowest – order MA processes involve non linear methods. Consider the MA (1) process as

\[
y_t = \mu + \varepsilon_t - \beta \varepsilon_{t-1}
\]

(3.9.12)

If \( \varepsilon_n \) is set to zero, then \( \varepsilon_1 = y_1 - \mu \) and \( \varepsilon_2 = y_2 - \mu + \beta \varepsilon_1 \). For \( n \) values of \( \varepsilon \), \( \varepsilon_n \) can be expressed in terms of \( \mu \) and \( \beta \). Now, \( \sum_{t=1}^{n} \varepsilon_t^2 \) is a complicated non linear function of the parameters which requires Iterative techniques to obtain even conditional estimators. Hamilton (1994) described full Maximum likelihood procedures for ARMA processes which share all the estimation problems of pure MA processes.
(iii) Diagnostic checking:

The next stage in the Box-Jenkins methodology is to check on the adequacy of the fit. This check is carried out by using the residuals.

If an adequate model has been fitted, the residuals should be approximately white noise. The residuals of the model provide important information for testing the adequacy of the model.

Three crucial aspects of the residuals are their autocorrelations, their partial autocorrelations and the values of the Box- Pierce – Ljung statistic, which tests the joint significance of subsets of autocorrelation coefficients. If the residuals depart significantly from white noise, the model is unsatisfactory and has to be re-specified.

Box and Pierce (1970) suggested Q statistic which is based on the squares of the first $p$ autocorrelation coefficients of the OLS residuals. The statistic is given by

$$Q = n \sum_{j=1}^{p} r_j^2$$  \hspace{1cm} (3.9.13)

Where

$$r_j = \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2}$$  \hspace{1cm} (3.9.14)

Here, $e_t$’s are the OLS residuals.

The limiting distribution of Q-statistic was derived under the assumption that the residuals come from an AR scheme, or from ARMA scheme fitted to some variable $Y$. under the hypothesis of zero autocorrelations for the residuals, $Q$ follows an asymptotic $\chi^2$ distribution with $(p - m)$ degrees of freedom. Here $m$ is the number of parameters estimated in fitting the ARMA model.

L Jung and Box (1978) revised Q statistic to improve small sample performance, which is given by

$$Q' = n(n+2) \sum_{j=1}^{p} r_j^2/(n-j)$$  \hspace{1cm} (3.9.15)
(iv) Forecasting:

One of the reasons for the popularity of the ARIMA modeling is its success in forecasting. In many cases, the forecasts obtained by this method are more reliable than those obtained from the traditional econometric modeling, particularly for short-term forecasts. Once the forecaster is satisfied that the ARIMA model is appropriate and the estimates of the parameters are significant, then the model may be used to forecast the future demand entries $x_t$. Further, the confidence limits of the forecasts may be obtained.

The Box-Jenkins methodology may be extended to the simultaneous study of two or more time series.

3.10. A BRIEF ABOUT NEURAL NETWORKS:-

3.10.1 Introduction:

Neural networks are parameterized nonlinear functions. Their parameters are the weights and biases of the network. Adjustment of these parameters results in different shaped nonlinearities. Typically these adjustments are achieved by a gradient descent approach on an error function that measures the difference between the output of the neural network and output of the actual system. Additionally there is no restriction on the unknown function to be linear. In this way, neural networks provide a logical extension to create nonlinear robust control schemes where there is no need to assume that the plant is a linear parameterization of known nonlinear functions.

3.10.2 Biological Neural Networks:

Muller and Reinhardt (1991), the features of the biological neural networks are attributed to its structure and its function. The Fundamental Unit of the network is called a neuron or a nerve cell. The below figure shows a schematic of the structure of a neuron.
Fig: 3.10.1 Schematic diagram of a typical neuron or nerve cell.

It consists of a cell body or soma where the cell nucleus is located. Tree like nerve fibres called dendrites are associated with the cell body. These dendrites receive signals from other neurons. Extending from the cell body is a signal long fibre called the axon, which eventually branches into strands and substrands connecting to many other neurons at synaptic junctions, or synapses. The receiving ends of these junctions on other cells can be found both on the dendrites and on the cell bodies themselves. The axon of a typical neuron leads to few thousand synapses associated with other neurons.

### 3.10.3 Characteristics of Neural Networks:

Some attractive features of the biological neural network that make it superior to even the most sophisticated AI Computer system for pattern recognition tasks are the following:

a) Robustness and fault tolerance: The decay of nerve cells does not seem to affect the performance significantly.

b) Flexibility: The Network automatically adjusts to a new environment without using any preprogrammed instructions.

c) Ability to deal with a variety of data situations: The network can deal with information that is fuzzy, probabilistic, noisy and inconsistent.
d) Collective computation: the network performs routinely many operations in parallel and also given task in a distributed manner

3.11. ARTIFICIAL NEURAL NETWORKS:

Over the last few recent years, there has been much research directed at predicting the future and making better decisions. This research has led to many developments in forecasting methods. Most of these methodological advances have been based on statistical techniques.

Statistical methods and neural networks are commonly used for time series prediction. Empirical results have shown that Neural Networks outperform linear regression especially in the case of more complex behavior of dependent variables like nonlinear, dynamic and chaotic behaviors. Neural networks are reliable for modeling nonlinear, dynamic market predictions. Neural Network makes very few assumptions as opposed to normality assumptions commonly found in statistical methods. Neural network can perform prediction after learning the underlying relationship between the input variables and outputs. From a statistician’s point of view, neural networks are analogous to nonparametric, nonlinear regression models.

![Artificial Neuron](image)

**Fig:3.11 Artificial Neuron**

3.11.1 Components of neural networks:

i). Nodes or neurons: The nodes can be seen as computational units. They receive inputs, and process them to obtain an output. This processing might be very simple such as summing the...
inputs, or quite complex like node might contain another network or involves huge iterative calculations.

![Fig: 3.11.1Nodes or neurons](image)

**ii). Connectors:** The connections determine the information flow between nodes. They can be unidirectional, when the information flows only in one sense, and bidirectional, when the information flows in either sense.

**iii). Input and output nodes:** Input nodes are the starting points of input data into the system. There are no preceding nodes to this input node. Each ANN may contain more than one input node also. Similarly output nodes are final resultants of the forecasting process.

**3.11.2 Architecture of neural networks:**

The basic architecture consists of three types of neuron layers: input, hidden, and output layers. In the forward networks, the signal flow is from input to output units, which is termed as feed-forward network. There are several other neural network architectures (Elman network, adaptive resonance theory maps, competitive networks, etc.), depending on the properties and requirement of the application. A simple architecture can be represented diagrammatically as follows.
Fig: 3.11.2 Multilayered artificial neural network

Neural network architecture comprises of the following four main elements

a) Processing Elements (Neurons)

b) Connection between the elements

c) Weights of the connection

d) Activation function

A neural network has processing elements called neurons which are used for processing the information. For a simple network, these neurons are arranged in two layers namely input layer and hidden layer. One hidden layer can approximately map any type of non-linear calculation (Weatherford, Lawrence R., 2002). The issue of determining the number of neurons required in the hidden layer is a matter of trial and error process or experimentation.

Each neuron has an activation function which is applied to the inputs provided to the neuron. For propagation of information between the various layers, the weightage connections are used. So ANN also called interconnections network. The weight of these connections depends upon which training algorithm is used to train the network.

The inputs that neurons receive from other neurons are summed up and are provided to the transfer function to get the output from neurons. There are various types of transfer functions used viz. Sigmoid Function, Hyperbolic tangent function, exponential function etc.
In order for ANN to learn, various training algorithms are used. The selection of an optimal learning algorithm is an open problem (Walczak et al., 1999). Some of them are gradient descent with momentum, gradient descent with adaptive learning rate, quasi-Newton, conjugate gradient, Scaled conjugate gradient and Levenberg-Marquardt.

3.11.3 Steps involved in neural network forecasting:

We can not find a constant and regular procedure for neural network models forecasting. We can generalize the fitting process in the following steps.

**Step 1:** Usually the modeling of ANN begins by postulating an initial network based on general modeling practices. One regular practice is to add very important variables in the model based on user knowledge on the data and respective scenario.

**Step 2:** Only one hidden layer is needed for a artificial neural network to be a universal function approximate to a continuous function (Hornik et al., 1989) so often only one hidden layer is used; however, more than one hidden layer maybe used since overall fewer neurons will be required. If the function to be approximated is discontinuous, the model will require at most two hidden layers (Cybenko, 1988).

**Step 3:** Determine the weights which are usually initialized with random values. The observations are then input to the network and parameters adjusted by several available methods. The adjustment process is repeated until the error converges on a minimum point.

**Step 4:** The results calculated in step 3 are now transformed as output nodes. The number of nodes in the output layer corresponds to the number of variables to be predicted.
3.12 Advantages and disadvantages of artificial neural network models:

Artificial neural networks are mathematical models inspired by the organization and functioning of biological neurons. There are numerous artificial neural network variations that are related to the nature of the task assigned to the network. The literature suggests several potential advantages that artificial neural networks have over statistical methods.

3.12.1 Advantages of artificial neural networks:

The primary advantages of neural networks can be outlined as follows.

i) ANN can be universal function approximators for even non-linear functions:

Artificial neural networks can be mathematically shown to be universal function approximators (Hornik et al., 1989). This means that artificial neural networks can automatically approximate whatever functional form best characterizes the data. While this property gives little value if the functional form is simple (e.g., linear), this property allows artificial neural networks to extract more signal from complex underlying functional forms. Also artificial neural networks can at least partially transform the input data automatically (Connor, 1988; Donaldson, Kamstra and Kim, 1993).

ii) ANN can also estimate piece-wise approximations of functions:

With artificial neural networks using one or more hidden layers, the networks can partition the sample space automatically and build different functions in different portions of different sample spaces. This means that artificial neural networks have a modest capability for building piece-wise non-linear models. The artificial neural network model for the exclusive operations research function is a good example of such a model (Wasserman, 1989, pp. 30-33). Collopy and Armstrong (1992) surveyed forecasting experts and found that the experts considered it important to select extrapolation techniques that identified and treated abrupt changes in historical data patterns, suggesting the utility of piece-wise models.

iii) ANN adopts automated estimation procedure:

Some statistical time series methods like ARIMA have inherent limitations due to the process. The estimation of many kinds of statistical time series models require human interaction
and evaluation. The estimation of artificial neural networks, however, can be automated (Hoptroff, 1993). Also, many statistical models must be re-estimated periodically when new data arrive. Many artificial neural network algorithms learn incrementally (Widrow and Sterns, 1985).

3.12.2 Disadvantages of artificial neural networks:

Apart from its various advantages, the following are some of its disadvantages.

- Artificial neural network methodology and modeling techniques are rapidly changing, whereas many statistical modeling techniques are stable and well developed. This is the alarming for the researcher who needs to keep an eye on the latest developments through articles and journals.

- Software is readily available for statistical techniques but commercial artificial neural network software, although of good quality, often lags developments in the field. Many software still utilizes the initial developments in ANN and not using the latest algorithm like conjugate radiant algorithm etc.

- Artificial neural network models are harder to interpret and to give physical meaning than many forecasting models. There has to be a deeper understanding of the concepts like hidden layers, training algorithms etc otherwise it is hard to take business decisions out of neural network output.

- Artificial neural networks contain more parameters to estimate than do most statistical forecasting models; this can result in over fitting problems. In many cases it requires more time and resources to carry out the forecasting using neural network models.

3.12.3 Applications of Neural Networks:

Neural networks are applicable in virtually every situation in which a relationship between the predictor variables (independents, inputs) and predicted variables (dependents, outputs) exists, even when that relationship is very complex and not easy to articulate in the usual terms of "correlations" or "differences between groups." A few representative examples of problems to which neural network analysis has been applied successfully are:

- Detection of Medical Phenomena:
A variety of health-related indices (e.g., a combination of heart rate, levels of various substances in the blood, respiration rate) can be monitored. The onset of a particular medical condition could be associated with a very complex (e.g., nonlinear and interactive) combination of changes on a subset of the variables being monitored. Neural networks have been used to recognize this predictive pattern so that the appropriate treatment can be prescribed.

- **A Stock Market Prediction:**
  Fluctuations of stock prices and stock indices are another example of a complex, multidimensional, but in some circumstances at least partially-deterministic phenomenon. Neural networks are being used by many technical analysts to make predictions about stock prices based upon a large number of factors such as past performance of other stocks and various economic indicators.

- **Credit Assignment:**
  A variety of pieces of information are usually known about an applicant for a loan. For instance, the applicant's age, education, occupation, and many other facts may be available. After training a neural network on historical data, Neural network analysis can identify the most relevant characteristics and use those to classify applicants as good or bad credit risks.

- **Condition Monitoring:**
  Neural networks can be instrumental in cutting costs by bringing additional expertise to scheduling the preventive maintenance of machines. A neural network can be trained to distinguish between the sounds a machine makes when it is running normally ("false alarms") versus when it is on the verge of a problem. After this training period, the expertise of the network can be used to warn a technician of an upcoming breakdown, before it occurs and causes costly unforeseen "downtime."
- **Engine Management:**

  Neural networks have been used to analyze the input of sensors from an engine. The neural network controls the various parameters within which the engine functions, in order to achieve a particular goal, such as minimizing fuel consumption.

- **Signature Analysis:**

  Signature Analysis, as a mechanism for comparing signatures made (e.g. in a bank) with those stored. This is one of the first large-scale applications of neural networks in the USA, and is also one of the first to use a neural network chip.

- **Process Control:**

  Process Control, Most processes cannot be determined as computable algorithms. Neural Networks can be used to adaptively control the process.

- **Non-Linear Identification & Adaptive Control:**

  This is one of the main areas of application of the neural networks. Neural Networks find applications in situations where the plant dynamics are uncertain or un-modeled.

### 3.13 MULTI-LAYER PERCEPTRON:

MLP consists of sensory units that make up the input layer, one or more hidden layers of processing units (perceptrons), and one output layer of processing units (perceptrons). The MLP performs a functional mapping from the input space to the output space. An MLP with a single hidden layer having \( H \) hidden units and a single output, \( y \), implements mappings of the form

\[
y = F(W_0 + \sum_{h=1}^{H} W_h Z_h)
\]

\[
Z_h = F(\beta_{oh} + \sum_{j=1}^{n} \beta_{th} X_j)
\]

Where \( Z_h \) is the output of the \( h \)th hidden unit, \( W_h \) is the weight between the \( h \)th hidden and the output unit, and \( W_0 \) is the output bias. There are \( N \) sensory inputs, \( X_j \). The \( j \)th input is weighted by an amount \( \beta_j \) in the \( h \)th hidden unit. The output of an MLP is compared to a target output and an error is calculated. This error is back-propagated to the neural network and used to adjust the weights. This process aims at minimizing the mean square error between the network’s prediction output and the target output.
MLP was first developed to mimic the functioning of the brain. It consists of interconnected nodes referred to as processing elements that receive, process, and transmit information. MLP consists of three types of layers: the first layer is known as the input layer and corresponds to the problem input variables with one node for each input variable. The second layer is known as the hidden layer and is useful in capturing non-linear relationships among variables. The final layer is known as the output layer and corresponds to the classification being predicted (Baranoff, Sager, & Shively, 2000). The below figure represents the typical structure of MLP.

First of all the network has to be trained to produce the correct output with minimum error. To achieve the minimum error the network first has to be trained until it produces a tolerable error. This is how the training is done. Input is fed to the input nodes, from here the middle layer nodes take the input value and start to process it. These values are processed based on the randomly allocated initial weight of the links. The input travels from one layer to another and every layer process the value based on the weights of its links. When the value finally reaches the output node, the actual output is compared with the expected output. The difference is calculated and it
is propagated backwards, this is when the links adjust their weights. After the error has propagated all the way back to first layer of middle level nodes, the input is again fed to the input nodes. The cycle repeats and the weights are adjusted over and over again until the error is minimized. The key here is the weight of different links. The weights of the links will decide the output value.

Bishop (1999) Sharda (1994) proposed The MLP is the most frequently used neural network technique in pattern recognition and classification problems. However, numerous researchers document the disadvantages of the MLP approach. For example, Calderon and Cheh (2002) argue that the standard MLP network is subject to problems of local minima. Swicegood and Clark (2001) claim that there is no formal method of deriving a MLP network configuration for a given classification task. Thus, there is no direct method of finding the ultimate structure for modeling process. Consequently, the refining process can be lengthy, accomplished by iterative testing of various architectural parameters and keeping only the most successful structures. Wang (1995) argues that standard MLP provides unpredictable solutions in terms of classifying statistical data.

3.13.1 GENERALIZED REGRESSION NEURAL NETWORK:

GRNN was devised by Specht (1991), casting a statistical method of function approximation into a neural network form. The GRNN, like the MLP, is able to approximate any functional relationship between inputs and outputs. Structurally, the GRNN resembles the MLP. However, unlike the MLP, the GRNN does not require an estimate of the number of hidden units to be made before training can take place. Furthermore, the GRNN differs from the classical MLP in that every weight is replaced by a distribution of weight which minimizes the chance of ending up in local minima. Therefore, no test and verification sets are required, and in principle all available data can be used for the training of the network proposed by Parojcic, Ibric, Djuric, Jovanovic, & Corrigan, 2005.

The GRNN is a method of estimating the joint probability density function (pdf) of $x$ and $y$, giving only a training set. The estimated value is the most probable value of $y$ and is defined by
\[
E(\frac{\partial}{\partial x})y = \hat{y}(x) = \frac{\int_{-\alpha}^{\alpha} y f(x, y) dy}{\int_{-\alpha}^{+\alpha} f(x, y) dy} \tag{3.13.3}
\]

Parzen, 1962 derived the density function \( f(x, y) \) can be estimated from the training set using Parzen’s estimator

\[
f(x, y) = \frac{1}{(2\pi)^{(p+1)/2} \sigma^{(p+1)}} \frac{1}{n} \sum_{i=1}^{n} \exp \left[-\frac{(x - x^i)^T(x - x^i)}{2\sigma^2} \right] \exp \left[-\frac{(y - y^i)^2}{2\sigma^2} \right] \tag{3.13.4}
\]

Specht, (1991) defines the probability estimate \( f(x, y) \) assigns a sample probability of width \( \sigma \) for each sample \( x^i \) and \( y^i \), and the probability estimate is the sum of these sample probabilities. Defining the scalar function \( D^2_i \)

\[
D^2_i = (x - x_i)^T(x - x_i) \tag{3.13.5}
\]

and assessing the indicated integration yields the following:

\[
Y(x) = \sum_{i=1}^{n} Y^i \exp \left(-\frac{D^2_i}{2\sigma^2} \right) / - \frac{D^2_i}{2\sigma^2} \tag{3.13.6}
\]

The resulting regression (3.13.6) is directly applicable to problems involving numerical data.

The first hidden layer in the GRNN contains the radial units. A second hidden layer contains units that help to estimate the weighted average. This is a specialized procedure. Each output has a special unit assigned in this layer that forms the weighted sum for the corresponding output. To get the weighted average from the weighted sum, the weighted sum must be divided through by the sum of the weighting factors. A single special unit in the second layer calculates the latter value. The output layer then performs the actual divisions (using special division units). Hence, the second hidden layer always has exactly one more unit than the output layer. In regression problems, typically only a single output is estimated, and so the second hidden layer usually has two units. The below figure shows the general structure of the GRNN. The GRNN
can be modified by assigning radial units that represent clusters rather than each individual training case: this reduces the size of the network and increases execution speed. Centers can be assigned using any appropriate algorithm (i.e., sub-sampling, K-means or Kohonen). The below figure shows the General structure of the GRNN.

![Diagram of GRNN](image)

**Fig: 3.13.2 General structure of the generalized regression neural network (GRNN)**

### 3.14 NEURAL NETWORK ALGORITHMS (BACK PROPAGATION):

The realization of back propagation method comprises of two particular calculation movements, Foreword movement does not change the connection weights and progressively computes the outcome of the network output signal of each is transmitted to the neurons in following layer, until the last layer is reached. The result of network is then compared with expected target and network generates corresponding error signal for all output for all layer neurons.

Backward Movement in contrast to the former one, sends the error signals from output neurons back over the network, applying the weight update according to some specific rule. This
recursive calculation is performed Neuron by neuron, providing the modifications to all connection weights. Subsequent sections are devoted to several optimization algorithms that are surely used to modify connection weights in feed forward networks with back propagation learning.

3.14.1 Hebbian Learning:

Hebb (1949), proposed a learning which was based on the modification of synaptic connections between neurons. The basic idea is that if two neurons are active simultaneously, their interconnection must be strengthened. If we consider a single layer net, one of the interconnected neurons will be an input unit and one an output unit. If the data are represented in bipolar form, it is easy to express the desired weight update as

\[ \omega_i (\text{new}) = \omega_i (\text{old}) + x_i o \]

Where \( o \) is the desired output for \( i = 1 \) to \( n(\text{inputs}) \).

Unfortunately, plain Hebbian learning continually strengthens its weights without bound (unless the input data is properly normalized).

3.14.2 Perceptron Learning Rule:

The perceptron is a single layer neural network whose weights and biases could be trained to produce a correct target vector when presented with the corresponding input vector. The training technique used is called the perceptron learning rule. Perceptron’s are especially suited for simple problems in pattern classification.

Suppose we have a set of learning samples consisting of an input vector \( x \) and a desired output \( d(k) \). For a classification task, the \( d(k) \) is usually +1 or −1. The perceptron-learning rule is very simple and can be stated as follows:

1. Start with random weights for the connections.
2. Select an input vector \( x \) from the set of training samples.
3. If output \( y_k \neq d(k) \) (the perceptron gives an incorrect response), modify all connections \( wi \) according to:
\[ \delta w_i = \eta (d_k - y_k) x_i; \ (\eta = \text{learning rate}). \]

4. Go back to step 2.

Note that the procedure is very similar to the Hebb rule; the only difference is that when the network responds correctly, no connection weights are modified.

### 3.14.3 Conjugate Gradient Learning Algorithm:

Conjugate gradient algorithm is one of the popular search methods to minimize the network output error in conjugate directions. Conjugate gradient method uses orthogonal and linearly independent non zero vector. Two sectors \(d_i\) and \(d_j\) are neutrally \(G\)-Conjugate if

\[ d_i^T G d_j = 0 \quad \text{for} \ i \neq j \quad \ldots \quad (3.14.1) \]

The algorithm firstly developed to minimize the quadratic function of \(n\)-variables

\[ F(w) = C-b^T W + 1/2 W^T G W \quad s \quad (3.14.2) \]

Where \(W\) is vector with \(n\) elements and \(G\) is \(nxn\) symmetric and positive definite matrix. The algorithm was than extended to minimization of general linear functions interpretation (3.14.2) as a secured order Taylor series expansion of the objective function \(G\).

A starting point \(W\) is selected and the first search direction \(d_1\) is said to be negative gradient \(g_1\) i.e., \((d_1 = -g_1)\) conjugate gradient method is to minimise differential function (3.14.2) by generating a sequence of appropriation \(W_{k+1}\) interactively according to

\[ W_{k+1} = W_k + \alpha_k d_k \quad \ldots \quad (3.14.3) \]

\[ d_{k+1} = -g_{k+1} + \beta_k d_k \quad \ldots \quad (3.14.4) \]

\(\alpha\) and \(\beta\) are moment term to avoid oscillations

Let \(\mu = \frac{1}{1 + \beta_s}\) then equation (3.14.4) can be written as

\[ d_{k+1} = \frac{1}{\mu_s} [\mu(-g_{k+1}) + (1-\mu)d_k] \quad \ldots \quad (3.14.5) \]
The value of $\alpha_k$ can be determined by line search techniques. Such as golden search and brent algorithm in the way that $f(W_k+\alpha_k d_k)$ is minimized along the direction $d_k$. $\beta_k$ can be calculated by any of the following formula.

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{d_k^T (g_{k+1} - g_k)}$$

(3.14.6)

Polka and Reeves formula is given by

$$\beta_k = \frac{g_{k+1}^T (g_{k+1} - g_k)}{g_k^T g_k}$$

(3.14.7)

Shanno derives the formula for $d_{k+1}$ by considering conjugate method as memory less quasi-nefon method.

$$d_{k+1} = g_{k+1} + \left[1 + \frac{y_k^T y_k}{p_k^T y_k}\right] \left(\frac{p_k^T g_k}{p_k^T y_k} \left(y_k^T g_k - \frac{y_k^T g_k}{p_k^T y_k}\right) + \frac{p_k^T g_k}{p_k^T y_k} y_k\right) + \frac{p_k^T g_k}{p_k^T y_k} y_k$$

(3.14.8)

Where $p_k = \alpha_k d_k$ and $y_k = g_{k+1} - g_k$

The method performs an appropriation line minimization in a descent direction in order to increase numerical stability.

The summary of conjugate gradient algorithm is described below

i) Set K=1, Initialize W

ii) Compute $g_1 = \nabla f(W_1)$

iii) Set $d_1 = -g_1$

iv) Compute $\alpha_k = \text{avg min} \ [f(W_k+\alpha_k d_k)]$

v) Update neighbour vector by $W_{k+1} = W_k + \alpha_k d_k$

vi) If network error is less than a pre set minimum value (or) the maximum number of iterations one reached, then stop. Else go to step vii
vii) If \( k+1 > n \) then \( W_1 = W_{k+1}, K=1 \) and go to step 2.

### 3.14.4 Multiple Linear Regression Weight Initializations:

Back propagation technique has risk of being stopped at local minimum. Multiple linear regression Weight method, neither between hidden layer and output layer are obtained by multiple linear regression.

The weight \( W_{ij} \) between input mode \( i \) and output mode \( j \) is initialized by uniform randomization. Once input \( x_i^s \) of sample \( s \) has been fed into input mode and \( w_{ij} \)'s have been assigned values, output value \( R_j^s \) of the hidden layer can be calculated as

\[
Y_s = \left( \sum_j V_j R_j^s \right) \quad \text{... (3.14.9)}
\]

Where \( V_j \) is weight between the hidden layer and output layer.

the Assume that sigmoid function \( f(x) = \frac{1}{1 + e^{-x}} \) is used as the transfer function of network. By Taylor expansion \( f(x) \equiv \frac{1}{2} + \frac{x}{4} \quad \text{... (3.14.10)} \)

Applying linear approximation is (3.14.7) in (3.4.6) we have the following linear approximate relationship between output \( y \) and \( v_j \)'s.

\[
y^s = \frac{1}{2} + \frac{1}{4} \left( \sum_j V_j R_j^s \right) \quad \text{(or) } 4y^s - 2 = V_i R_i^s + \ldots + V_m R_m^s \quad \text{... (3.14.11)}
\]

Where \( m \) is number of hidden layers. The set of equation in (3.14.8) is multiple linear regression model. \( R_j^s \) are considered as repressors, \( V_j^s \) can be estimated by standard regression methods.

### 3.14.5 Levenberg-Marquardt Method:

A condition that can cause erratic behavior of Guass-Newton iteration is singularity of the derivative matrix \( V \) caused by collinearity of the columns. When \( V \) is the nearly singular \( \delta \) can be very large causing the parameters to go onto undesirable regions of the parameter space.
One solution to the problem of near-singularity is to perform the calculations for the increment in numerically stable way, which is why we recommend. Using the QR decomposition rather than the Normal equation we also recommended using double precision (or) extended precision arithmetic for the calculations.

Another general method for dealing with near – singularity is to modify the Guass – Newton increment to

\[
\delta(k) = (V^TV + KI)^{-1} V^T (Y - \eta) \tag{3.14.12}
\]

as suggested in Levenberg (1944)

\[
\delta(k) = (V^TV + KD)^{-1} V^T (Y - \eta) \tag{3.14.13}
\]

As suggested on Marquardt (1963), where K is a conditioning factor D is a diagonal matrix with entries equal to the diagonal element of \( V^TV \). This is called the Levenberg – Marquardt method the direction of \( \delta(k) \) is intermediate between the direction of the Guass – Newton increment \((k \rightarrow 0)\) And the direction of steepest descent.

\[
V^T (Y - \eta) / |V^T (y - \eta)| (K \rightarrow \infty) \tag{3.14.14}
\]

Note that Levenberg recommends inflating the diagonal of \( V^TV \) by an additive factor while Marquardt recommends inflating the diagonal by a multiplicative factor; \( 1+K \). Marquardts method produces an increment which is invariant under scaling transformations of the parameters. So that if the scale for one component of the parameter vector is doubled, the increment calculated, and the corresponding component of the increment halved, the result will be the same as the calculating the increment in original scale. Box and Kanemasu (1984) suggested, however, that if one requires in-varaince of the increment under linear transformations of the parameter space, the resulting increment is the Guass-Newton increment with a step factor.

The Levenberg – Marqardt method is more difficult to implement than the Guass – Newton method. Since one must decide how to manipulate both the conditioning factor k and the step factor \( \lambda \). Nevertheless it is implemented in many non – linear least squares programs.
Although we presented the increment in terms of the inverse of an augmented $V^T V$ matrix, the actual calculations for the increment should be done using a $QR$ decomposition of $V$ and applying updates from a diagonal matrix (Dogarra et al., 1979), Golub and Pereyra, 1973). Since the Levenberg increment is the least squares solution of the system with derivative matrix.

$$\begin{bmatrix} V \\ \sqrt{K} I \end{bmatrix} \text{ and response vector } \begin{bmatrix} Y - \eta \\ 0 \end{bmatrix}$$

(3.14.15)

For the Marquardt increment (derivative matrix is change to

$$\begin{bmatrix} V \\ \sqrt{K} D^{1/2} \end{bmatrix}$$

(3.14.16)