UNIT 13

CORRELATION ANALYSIS IN TIME SERIES

Structure

13.1 INTRODUCTION

With the help of the time series data, we try to fit a time series model so that we can forecast the observations. But one of the essential elements of time series modelling is stationarity. In the previous unit, you have studied what is stationary and nonstationary time series and how to detect and transform nonstationary time series to stationary time series. As you know, a time series is a collection of observations with respect to time, therefore, there is a chance that a value at the present time may relate/depend on the past value. In most of the time series, we observe such relationships. To study the degree of relationship between previous/past values with the current value, we have to study the covariance and correlation between them before modelling the time series. Therefore, in this unit, you will study correlation analysis in time series.

We begin with a simple introduction of autocovariance and autocorrelation functions in time series in Sec. 13.2. In Sec. 13.3, we discuss how to estimate the autocovariance and autocorrelation functions using time series data. When we study the autocorrelation between observations in the presence of the intermediate variables, then it does not give the true picture of the relation. Therefore, to remove the effect of the same, we use partial autocorrelation which is discussed in Sec. 13.4. To present the autocorrelation/ partial autocorrelation in the form of graphs/diagrams, we use a correlogram. In Sec. 13.5, we describe what is correlogram and how to plot it. The

* Dr. Prabhat Kumar Sangal, School of Sciences, IGNOU, New Delhi

interpretation of the correlogram is also explained in Sec. 13.6. In the next unit, you will study different models for time series.

Expected Learning Outcomes

After studying this unit, you would be able to:

- describe the concept of covariance and correlation in time series;
- explain autocovariance and autocorrelation functions;
- describe partial autocorrelation function; and
- plot and interpret the correlogram.

13.3 AUTOCOVARIANCE AND AUTOCORRELATION FUNCTIONS

As you know, a time series is a collection of observations with respect to time. Since time series data are continuous and chronologically arranged, therefore, there is a chance that a value at the present time may depend on the past value. For example, the temperature in the next hour is not a random event since, in most cases, it depends on the current temperature or that has occurred during the past 24 hours. Therefore, the past temperature has an impact more on the feature temperature. In other words, we can say that there exists a strong relationship between the current temperature and the next hour's temperature. Similarly, in cases of current sales of a company depending on past sales, the stock is up today, then it is more likely to be up tomorrow, etc. For measuring such a linear relationship, we use covariance or correlation. The correlation between a series and its lags is called autocorrelation. Since we calculate covariance/correlation between two values of the same time series, therefore, it is called autocovariance/autocorrelation. The information provided by the autocovariance/ autocorrelation is used to understand the properties of time series data, fit the appropriate models, and forecast future events of the series.

You have some idea about the covariance and correlation. You now try to understand the basic concepts of both and what covariance and correlation are.

Covariance

Covariance is defined as a measure of the relationship between two variables. It measures how much two variables change together. If X and Y are two variables, then covariance is defined mathematically as

$$\text{Cov}(x,y)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)$$

It takes values from $-\infty$ to $+\infty$. The covariance tells whether both variables vary in the same direction (positive covariance) or in the opposite direction (negative covariance). If it is positive then it indicates a direct dependency, i.e, increasing the value of one variable will result in an increase in the value of the other variable and vice versa. On the other hand, a negative value signifies negative covariance, which indicates that the two variables have an inverse dependency, i.e., increasing the value of one variable will result in decreasing

the value of another variable and vice versa. A zero value indicates no relationship between the variables.

The main problem with the covariance is that it is hard to interpret due to its wide range ($-\infty$ to $+\infty$). For example, our data set could return a value say 5, or 500. It may take a large value if the variables X and Y are large. Therefore, a large value of covariance does not indicate that there exists a strong relationship between the variables. It means that it does not tell us that there exists a strong relationship between the variables when it is large. A value of 500 tells us that the variables are correlated, but unlike the correlation coefficient, that number doesn't tell us exactly how strong that relationship is. There is no meaning of the numerical value of covariance only the sign is useful. To overcome this problem the covariance is divided by the standard deviation to get the correlation coefficient.

Correlation

Correlation is a measure of identifying and quantifying the linear relationship between two variables. This relationship could vary from having a full dependency or linear relationship between the two, to complete independenc. Correlation explains the change in one variable leads to how much proportion changes in the second variable. One of the most popular methods for measuring the level of correlation between two variables is the Pearson correlation coefficient. It measures the intensity or degree of the linear relationship between two variables. If X and Y are two variables, then the Pearson correlation coefficient (r) is defined mathematically as

$$r_{XY} = r_{YX} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

When two variables are related in such a way that change in the value of one variable affects the value of another variable, then variables are said to be correlated.

The value of the coefficient of correlation can range from – 1 to +1, with a negative value indicating an inverse relationship and a positive value indicating a direct relationship. It reveals not only the nature of the relationship but also its strength. If it is near to $\pm$ 1 then the variables are highly correlated on the other hand if it is near zero then it indicates a poor relationship.

To understand autocovariance/autocorrelation, you have to understand what lag is.

Lag

The number of intervals between the two observations is the **lag**. For example, the lag between the current and previous observations is one. If you go back one more interval, the lag is two, and so on. In mathematical terms, the observations Y_t and Y_{t+k} are separated by k time units, then the lag is k. This lag can be days, quarters, or years depending on the nature of the data. When k = 1, you are assessing adjacent observations.

We now come to our main topic autocovariance, and autocorrelation and we now define autocovariance/autocorrelation formally.

Autocovariance

If we are interested in finding a linear relationship between two consecutive observations of a time series, say, Y_t and Y_{t+1} and also interested in the relationship between observations at k lag apart i.e. Y_t and Y_{t+k}, then we use autocovariance /autocorrelation. Let us start with autocovariance and we will introduce the autocorrelation function after that.

Autocovariance can be defined as

The covariance between a given time series and a lagged version of itself over successive time intervals is called autocovariance.

If Y_t and Y_{t+k} (t = 1, 2, …k = 0,1, 2, …) denote the time series which start from time t and t+k, respectively, then covariance between time series Y_t and Y_{t+k} is called autocovariance at lag k. Mathematically, we can define the autocovariance function as

You also noticed that the summation on the left-hand side of the formula of autocovariance is divided by N instead of N − k as you may expect. This is done because the former ensures that the estimate of the covariance matrix is a nonnegative definite matrix.

$$\gamma_k = \gamma_{-k} = \text{Cov}(Y_t, Y_{t+k}) = \frac{1}{N}\sum_{t=1}^{N-k}\{Y_t - \text{mean}(Y_t)\}\{Y_{t+k} - \text{mean}(Y_{t+k})\}$$

where N is the size of time series.

The autocovariance function is denoted by γ_k and read as gamma. Here, k represents the lag. Since for stationary time series mean remains constant, therefore,

$$\text{mean}(Y_t) = \text{mean}(Y_{t+k}) = \mu$$

Thus,

$$\gamma_k = \gamma_{-k} = \text{Cov}(Y_t, Y_{t+k}) = \frac{1}{N}\sum_{t=1}^{N-k}(Y_t - \mu)(Y_{t+k} - \mu)$$

When the lag is zero, that is, k = 0, then

$$\gamma_0 = \text{Cov}(Y_t, Y_t) = \frac{1}{N}\sum_{t=1}^{N}(Y_t - \mu)(Y_t - \mu) = \frac{1}{N}\sum_{t=1}^{N}(Y_t - \mu)^2$$

The autocovariance is the same as the covariance. The only difference is that the autocovariance is applied to the same time series data, i.e., you compute the covariance of the data say temperature Y with the same data temperature Y, but from a previous period.

Autocorrelation

In time series analysis, the autocorrelation is the fundamental technique for calculating the degree of correlation between a series and its lags. This method is fairly similar to the Pearson correlation coefficient but autocorrelation uses the same time series twice: one in its original form and the second lagged one or more time periods as in autocovariance. We now define autocorrelation as

Autocorrelation is a measure of the degree of relationship between a given time series and a lagged version of itself over successive time intervals.

If Y_t and Y_{t+k} denote the value of a stationary time series which start from time t and t+k, respectively,then the autocorrelation function/ coefficient between time series Y_t and its lag value Y_{t+k} is defined as

$$\rho_k = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{Var(Y_t)\,Var(Y_{t+k})}}$$

Since for stationary time series variance of the series remains constant, therefore,

$$Var(Y_t) = Var(Y_{t+k})$$

Thus, the autocorrelation function at lag k becomes as

$$\rho_k = \frac{Cov(Y_t, Y_{t+k})}{Var(Y_t)} = \frac{\sum_{t=1}^{N-k}(Y_t - \mu)(Y_{t+k} - \mu)}{\sum_{t=1}^{N}(Y_t - \mu)^2}$$

The autocorrelation function (ρ_k), can also be defined in terms of autocovariance as

$$\rho_k = \frac{\sum_{t=1}^{N-k}(Y_t - \mu)(Y_{t+k} - \mu)}{\sum_{t=1}^{N}(Y_t - \mu)^2} = \frac{\gamma_k}{\gamma_0}$$

When lag is zero, that is, k = 0 then

$$\rho_0 = \frac{\sum_{t=1}^{N}(Y_t - \mu)(Y_t - \mu)}{\sum_{t=1}^{N}(Y_t - \mu)^2} = \frac{\gamma_0}{\gamma_0} = 1$$

The degree of correlation between a series and its lags indicates the pattern/characteristics of the series. For example, if a time series has a seasonality component say monthly then we will observe a strong correlation with its seasonal lags, say, 12, 24, and 36 months.

Some important properties of time series can be studied with the help of autocovariance and autocorrelation functions. They measure the linear relationship between observations at different time lags apart. They provide useful descriptive properties of the time series under study. This is also an important tool for guessing a suitable model for the time series data.

After understanding the concept of autocovariance and autocorrelation functions, we now study how to estimate them using sample data.

13.3 ESTIMATION OF AUTOCOVARIANCE AND AUTOCORRELATION FUNCTIONS

In the previous section, we consider the theoretical aspect of autocovariance and autocorrelation functions for time series. In practice, we have a finite time series and based on the observations of the given time series we estimate the mean, autocovariance and autocorrelation function. Suppose $y_t = y_1, y_2, ..., y_n$ represent the observations of a finite time series and it is assumed a sample of theoretical time series Y_t. We can estimate the mean of the time series as (μ) by the sample mean as

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$$

and estimate autocovariance function as

$$\hat{\gamma}_k = c_k = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y}); \quad k = 1,2,...,n-1$$

It is known as the sample autocovariance function.

You also noticed that the summation on the left-hand side of the formula of sample autocovariance is divided by n instead of n − k as you may expect. This is done because the former ensures that the estimate of the covariance matrix is a nonnegative definite matrix.

Similarly, we estimate the autocorrelation function at lag k as

$$\hat{\rho}_k = r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{c_k}{c_0}; \quad k = 1,2,...,n-1$$

It is known as the sample autocorrelation function.

As you know that the correlation coefficient is calculated between variables with multiple values of the same length. Therefore, to compute the sample autocorrelation, first of all, we make two series of the same length as discussed in Example 1.

You will also notice that as we increase lag k, that is, if we calculate autocorrelation between observations further and further apart, then we create two variables, y_{t+k} and y_t and they will each have n – k observations, therefore, as we increase k, the number of observations decrease. Therefore, after a while, the estimates of autocovariance and autocorrelation will become more and more unreliable. Hence, to find a reliable estimate of the autocorrelation function, we should require at least 50 observations and the sample autocorrelation function should be calculated up to lag k = n/4, where n is the number of observations in the time series. For illustration purposes, we just consider small time series data (less than 50 observations).

Let's look at an example which helps you to understand how to calculate the sample autocovariance and autocorrelation functions.

Example 1: The meteorological department collected the following data of temperature (in °C) in a particular area on different days:

Day	Temperature	Day	Temperature
1	22	9	28
2	23	10	30
3	23	11	31
4	24	12	30
5	23	13	30
6	25	14	31
7	26	15	30
8	28		

Calculate mean, variance and autocorrelation functions for the given data.

Solution: As you know that the autocovariance/autocorrelation function is calculated between variables with multiple values of the same length. Therefore, to compute the sample autocorrelation, first of all, we make two

series of the same length. If y_t denotes the value of the temperature/series at any particular time t then y_{t+1} denotes the value of the temperature/series one time after time t. That is, y_{t+1} is the lag 1 value of y_t as shown in the following table:

Day	Temperature (y_t)	y_{t+1}	Day	Temperature (y_t)	y_{t+1}
1	22	--	9	28	28
2	23	22	10	30	28
3	23	23	11	32	30
4	24	23	12	32	31
5	23	24	13	34	30
6	25	23	14	33	31
7	26	25	15	34	31
8	28	26			

Since y_t and y_{t+1} have different dimensions (the first one has 15 observations, while the second one has 14), therefore, we use data from day 2 onwards to day 15 to make equal length for k = 1. Consequently, our data is as follows:

Day	Temperature (y_t)	y_{t+1}
1	22	--
2	23	22
3	23	23
4	24	23
5	23	24
6	25	23
7	26	25
8	28	26
9	28	28
10	30	28
11	32	30
12	32	31
13	34	30
14	33	31
15	34	31

Since there are 15 observations, therefore, we prepare the data up to n/4 = 15/4 ~ 4 lags in a similar way as shown below:

Day	Temperature(y_t)	y_{t+1}	y_{t+2}	y_{t+3}	y_{t+4}
1	22				
2	23	22			
3	23	23	22		
4	24	23	23	22	
5	23	24	23	23	22
6	25	23	24	23	23
7	26	25	23	24	23
8	28	26	25	23	24

For k = 2, we consider data from day 3 and for k = 3, we start from day 4.

9	28	28	26	25	23
10	30	28	28	26	25
11	31	30	28	28	26
12	30	31	30	28	28
13	31	30	31	30	28
14	31	31	30	31	30
15	30	31	31	30	31
Total	**405**				

Since for the calculation of the autocorrelation function, we assume that the time series is stationary, therefore, mean and variance of the series will be constant. Thus, we calculate the sample mean and variance of the given original time series and make the necessary calculations for calculating the autocovariance and autocorrelation function in the following table:

$y_t - \bar{y}$	$(y_t - \bar{y})^2$	$y_{t+1} - \bar{y}$	$y_{t+2} - \bar{y}$	$y_{t+3} - \bar{y}$	$y_{t+4} - \bar{y}$	$(y_t - \bar{y})(y_{t+1} - \bar{y})$	$(y_t - \bar{y})(y_{t+2} - \bar{y})$	$(y_t - \bar{y})(y_{t+3} - \bar{y})$	$(y_t - \bar{y})(y_{t+4} - \bar{y})$
–5	25								
–4	16	–5				20			
–4	16	–4	–5			16	20		
–3	9	–4	–4	–5		12	12	15	
-4	16	–3	–4	–4	–5	12	16	16	20
–2	4	–4	–3	–4	–4	8	6	8	8
–1	1	–2	–4	–3	–4	2	4	3	4
1	1	–1	–2	–4	–3	–1	–2	–4	–3
1	1	1	–1	–2	–4	1	–1	–2	–4
3	9	1	1	–1	–2	3	3	–3	–6
4	16	3	1	1	-1	12	4	4	–4
3	9	4	3	1	1	12	9	3	3
4	16	3	4	3	1	12	16	12	4
4	16	4	3	4	3	16	12	16	12
3	9	4	4	3	4	12	12	9	12
Total	**164**	**–3**	**–7**	**–11**	**–14**	**137**	**111**	**77**	**46**

Therefore,

$$\text{Mean} = \frac{1}{n}\sum_{t=1}^{n} y_t = \frac{405}{15} = 27$$

$$\text{Variance} = c_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2 = \frac{164}{15} = 10.933$$

Autocovariance function

$$c_1 = \frac{1}{n}\sum_{t=1}^{n-1}(y_t - \bar{y})(y_{t+1} - \bar{y}) = \frac{1}{15} \times 137 = 9.133$$

$$c_2 = \frac{1}{n}\sum_{t=1}^{n-2}(y_t - \bar{y})(y_{t+2} - \bar{y}) = \frac{1}{15} \times 111 = 7.4$$

$$c_3 = \frac{1}{n}\sum_{t=1}^{n-3}(y_t - \bar{y})(y_{t+3} - \bar{y}) = \frac{1}{15} \times 77 = 5.133$$

$$c_4 = \frac{1}{n}\sum_{t=1}^{n-4}(y_t - \bar{y})(y_{t+4} - \bar{y}) = \frac{1}{15} \times 46 = 3.067$$

After calculating the autocovariance function, we now calculate the sample autocorrelation function as

$$r_1 = \frac{c_1}{c_0} = \frac{9.133}{10.933} = 0.835$$

$$r_2 = \frac{c_2}{c_0} = \frac{7.4}{10.933} = 0.677$$

$$r_3 = \frac{c_3}{c_0} = \frac{5.133}{10.933} = 0.470$$

$$r_4 = \frac{c_4}{c_0} = \frac{3.067}{10.933} = 0.280$$

You may like to try the following Self Assessment Question before studying further.

SAQ 1

A researcher wants to study the pattern of the unemployment rate in his country. He collected quarterly unemployment rate data and given in the following table:

Quarter	Unemployment rate	Quarter	Unemployment rate
1	91	7	64
2	45	8	99
3	89	9	64
4	36	10	89
5	72	11	68
6	51	12	108

Compute:

(i) mean and variance, and

(ii) Autocovariance and autocorrelation functions.

13.4 PARTIAL AUTOCORRELATION FUNCTION

In the previous section, you studied autocorrelation function which measures the linear dependency between a time series Y_t with its own lagged values Y_{t+k}. However, a time series tend to carry information and dependency structures in steps and therefore autocorrelation at lag k is also influenced by the intermediate variables Y_{t+1}, Y_{t+2},…, Y_{t+k-1}. Therefore, autocorrelation is not the correct measure of the mutual correlation between Y_t and Y_{t+k} in the presence of the intermediate variables. Partial autocorrelation solves this problem by measuring the correlation between Y_t and Y_{t+k} when the influence of the intermediate variables has been removed. Hence partial autocorrelation in time series analysis defines the correlation between Y_t and Y_{t+k} which is not accounted for by lags t +1 to t + k –1. The partial autocorrelation function is similar to the atocorrelation function except that it displays only the correlation between two observations after removing the effect of intermediate variables. For example, if we are interested in the direct relationship between today's

consumption of patrol and that of a year ago then we don't blame what happens in between. The consumption of the previous 12 months has an effect on the consumption of the previous 11 months, and the cycle continues until the most current period. In partial autocorrelation estimates, these indirect effects are ignored. Therefore, we can define the partial autocorrelation function as

The partial autocorrelation function calculates the degree of relationship between a time series Y_t with its own lagged values Y_{t+k} after their mutual linear dependency on the intervening variables Y_{t+1}, Y_{t+2},…, Y_{t+k-1} has been removed.

You can understand the same using the diagram given in Fig. 13.1.

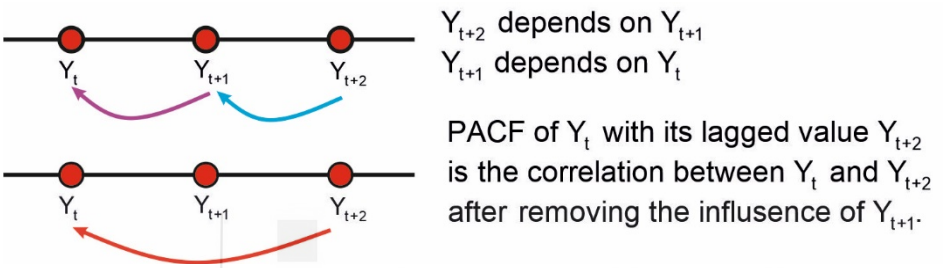

Fig. 13.1: PACF of order 2.

In other words, we can define the partial autocorrelation function between Y_t and Y_{t+k} as

The conditional correlation between Y_t and Y_{t+k}, conditional on $Y_{t+1}, Y_{t+2},..., Y_{t+k-1}$ (the set of observations that come between the time points Y_t and Y_{t+k}), is known as the kth order PACF.

Therefore, we can define the kth order (lag) partial autocorrelation function mathematically as

$$\phi_{kk} = \frac{\mathrm{Cov}\left(Y_t, Y_{t+k} \mid Y_{t+1}, Y_{t+2},..., Y_{t+k-1}\right)}{\sqrt{\mathrm{Var}\left(Y_t \mid Y_{t+1}, Y_{t+2},..., Y_{t+k-1}\right)\mathrm{Var}\left(Y_{t+k} \mid Y_{t+1}, Y_{t+2},..., Y_{t+k-1}\right)}}$$

This is the correlation between values two time periods apart conditional on knowledge of the value in between. (By the way, the two variances in the denominator will equal each other in a stationary series.), therefore,

$$\phi_{kk} = \frac{\mathrm{Cov}\left(Y_t, Y_{t+k} \mid Y_{t+1}, Y_{t+2},..., Y_{t+k-1}\right)}{\mathrm{Var}\left(Y_t \mid Y_{t+1}, Y_{t+2},..., Y_{t+k-1}\right)}$$

The formula for calculating the partial autocorrelation function looks scary, therefore, we calculate it using the autocorrelation function instead of it.

The 1st order partial autocorrelation function equals to the 1st order autocorrelation function, that is,

$$\phi_{11} = \rho_1$$

Similarly, we can define the 2nd order (lag) partial autocorrelation function in terms of autocorrelation function as

$$\phi_{22} = \frac{\left(\rho_2 - \rho_1^2\right)}{\left(1 - \rho_1^2\right)}$$

The general form for calculating partial autocorrelation function in terms of ACF is given in the matrix form as shown:

$$\begin{bmatrix}\phi_{1k}\\ \phi_{2k}\\ \phi_{3k}\\ \vdots\\ \phi_{kk}\end{bmatrix}=\begin{bmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1}\\ \rho_1 & 1 & \rho_3 & \cdots & \rho_{k-2}\\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1\end{bmatrix}^{-1}\begin{bmatrix}\rho_1\\ \rho_2\\ \rho_2\\ \vdots\\ \rho_k\end{bmatrix}$$

Or

$$\phi_k = P_k^{-1}\Psi_k$$

where

$$\phi_k=\begin{bmatrix}\phi_{1k}\\ \phi_{2k}\\ \phi_{3k}\\ \vdots\\ \phi_{kk}\end{bmatrix}, P_k=\begin{bmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1}\\ \rho_1 & 1 & \rho_3 & \cdots & \rho_{k-2}\\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-3}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1\end{bmatrix} \text{ and } \Psi_k=\begin{bmatrix}\rho_1\\ \rho_2\\ \rho_2\\ \vdots\\ \rho_k\end{bmatrix}$$

In the above expression, the last coefficient, ϕ_{kk}, is the partial autocorrelation function of order k. Since we are interested only in this coefficient, therefore, we can solve the above expression for ϕ_{kk} using the Cramer-Rule. We get

$$\phi_{kk}=\frac{\left|P_k^*\right|}{\left|P_k\right|}$$

where $|.|$ indicates the determinant and $\left|P_k^*\right|$ is given as

$$\left|P_k^*\right|=\begin{vmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_1\\ \rho_1 & 1 & \rho_3 & \cdots & \rho_2\\ \rho_2 & \rho_1 & 1 & \cdots & \rho_3\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_k\end{vmatrix}$$

It is equal to the matrix P_k in which the kth column is replaced with Ψ_k.

Therefore, the 3rd order partial autocorrelation function is

$$\phi_{33}=\frac{\begin{vmatrix}1 & \rho_1 & \rho_1\\ \rho_1 & 1 & \rho_2\\ \rho_2 & \rho_1 & \rho_3\end{vmatrix}}{\begin{vmatrix}1 & \rho_1 & \rho_2\\ \rho_1 & 1 & \rho_3\\ \rho_2 & \rho_1 & 1\end{vmatrix}}$$

As you saw, the autocorrelation function helps assess the properties of a time series. In contrast, the partial autocorrelation function (PACF) is more useful for finding the order of an autoregressive, autoregressive integrated moving average (ARIMA) model. You will study these models in the next unit.

Sample Partial Autocorrelation Function

In practice, we have a finite time series and on the basis of the observations of the given time series, we estimate the partial autocorrelation function. The

Cramer's rule can be used in any system of n linear equations in n variables. If we have following equations

$$a_{11}x+a_{12}y+a_{13}z=b_1$$
$$a_{21}x+a_{22}y+a_{23}z=b_2$$
$$a_{31}x+a_{32}y+a_{33}z=b_3$$

and if $\Delta \neq 0$, then according to Cramer-Rule the system has unique solution and is given by

$$x=\frac{\Delta_1}{\Delta},\ y=\frac{\Delta_2}{\Delta},\ z=\frac{\Delta_3}{\Delta}$$

Where

$$\Delta=\begin{vmatrix}a_{11} & a_{12} & a_{13}\\ a_{21} & a_{22} & a_{23}\\ a_{31} & a_{32} & a_{33}\end{vmatrix}$$

$$\Delta_1=\begin{vmatrix}b_1 & a_{12} & a_{13}\\ b_2 & a_{22} & a_{23}\\ b_3 & a_{32} & a_{33}\end{vmatrix}$$

$$\Delta_2=\begin{vmatrix}a_{11} & b_1 & a_{13}\\ a_{21} & b_2 & a_{23}\\ a_{31} & b_3 & a_{33}\end{vmatrix}$$

$$\Delta_3=\begin{vmatrix}a_{11} & a_{12} & b_1\\ a_{21} & a_{22} & b_2\\ a_{31} & a_{32} & b_3\end{vmatrix}$$

estimate of the partial autocorrelation function is known as sample partial autocorrelation and the formulae for the same are obtained by replacing autocorrelation function (ρ) with sample autocorrelation function (r) which are given as follows:

$$\hat{\phi}_{11} = r_1$$

We define the 2nd order (lag) sample partial autocorrelation function as

$$\hat{\phi}_{22} = \frac{\left(r_2 - r_1^2\right)}{\left(1 - r_1^2\right)}$$

The general form for calculating the sample partial autocorrelation function of order k is given in the matrix form as shown below:

$$\hat{\phi}_{kk} = \frac{\left|\hat{P}_k^*\right|}{\left|\hat{P}_k\right|} = \frac{\begin{vmatrix} 1 & r_1 & r_2 & \cdots & r_1 \\ r_1 & 1 & r_3 & \cdots & r_2 \\ r_2 & r_1 & 1 & \cdots & r_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & r_{k-3} & \cdots & r_k \end{vmatrix}}{\begin{vmatrix} 1 & r_1 & r_2 & \cdots & r_{k-1} \\ r_1 & 1 & r_3 & \cdots & r_{k-2} \\ r_2 & r_1 & 1 & \cdots & r_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{k-1} & r_{k-2} & r_{k-3} & \cdots & 1 \end{vmatrix}}$$

Let's consider an example which helps you to understand how to calculate the sample partial autocorrelation function.

Example 2: For the data given in Example 2 of Unit 12, calculate the sample partial autocorrelation up to order 3.

Solution: For calculating sample partial autocorrelation, first of all, we have to compute the sample autocorrelation function. We have already calculated these in Example 2. Therefore, for the sake of time, we just write them here

$$r_1 = 0.835,\ r_2 = 0.677,\ r_3 = 0.470,\ r_4 = 0.280$$

Since, the 1st-order partial autocorrelation function equals the 1st-order autocorrelation function, therefore,

$$\phi_{11} = r_1 = 0.835$$

We can calculate the 2nd order (lag) sample partial autocorrelation function as

$$\phi_{22} = \frac{\left(r_2 - r_1^2\right)}{\left(1 - r_1^2\right)} = \frac{0.677 - (0.835)^2}{1 - (0.835)^2}$$

$$= \frac{-0.020}{0.303} = -0.067$$

Similarly, We now compute the 3rd order sample partial autocorrelation function as

$$\phi_{33} = \frac{\begin{vmatrix} 1 & r_1 & r_1 \\ r_1 & 1 & r_2 \\ r_2 & r_1 & r_3 \end{vmatrix}}{\begin{vmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_3 \\ r_2 & r_1 & 1 \end{vmatrix}}$$

$$\begin{vmatrix} 1 & r_1 & r_1 \\ r_1 & 1 & r_2 \\ r_2 & r_1 & r_3 \end{vmatrix} = \begin{vmatrix} 1 & 0.835 & 0.835 \\ 0.835 & 1 & 0.677 \\ 0.677 & 0.835 & 0.470 \end{vmatrix}$$

$$= 1\times(0.470 - 0.835\times 0.677) - 0.835\times(0.835\times 0.470 - 0.677\times 0.677) + 0.835\times(0.835\times 0.835 - 0.676\times 1)$$

$$= -0.095 + 0.055 + 0.017 = -0.023$$

Similarly,

$$\begin{vmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_3 \\ r_2 & r_1 & 1 \end{vmatrix} = \begin{vmatrix} 1 & 0.835 & 0.677 \\ 0.835 & 1 & 0.470 \\ 0.677 & 0.835 & 1 \end{vmatrix}$$

$$= 1\times(1 - 0.835\times 0.470) - 0.835\times(0.835\times 1 - 0.677\times 0.470) + 0.677\times(0.835\times 0.835 - 0.677)$$

$$= 0.608 - 0.432 + 0.014 = 0.191$$

Therefore,

$$\hat{\phi}_{33} = \frac{-0.023}{0.191} = -0.120$$

Before going to the next session, you may like to compute the sample partial autocorrelation function yourself. Let us try a Self Assessment Question.

SAQ 2

For the data given in SAQ 1, calculate the sample partial autocorrelation function up to order 2.

13.5 CORRELOGRAM

In the previous sessions, you learnt autocovariance, autocorrelation, and partial autocorrelation functions which are used to understand the properties of time series, fit the appropriate models, and forecast future events of the series. With the help of the autocorrelation/partial autocorrelation function, we can also diagnose whether the time series is stationary or not. But a group of a large number of autocorrelation always makes misperceptions to the reader and he/she may understand it wrongly. If we present the autocorrelation/ partial autocorrelation function in the form of graphs/diagrams, then it attracts the reader and it can be understood better.

A plot in which we take the autocorrelation function on the vertical axis and different lags on the horizontal axis is known as a correlogram. The technique of drawing a correlogram is the same as that of a simple bar diagram. The only difference is that we just take a line instead of a bar of the same width. Each bar in the correlogram represents the level of correlation between the series and its lags in chronological order. A correlogram is also known as an **autocorrelation function (ACF) plot** or **autocorrelation plot**. It gives a **summary of autocorrelation** at different lags. With the help of a correlogram, we can easily examine the nature of the time series and diagnose a suitable model for the time series data.

The correlogram suggests that observations with smaller lag are positively correlated and autocorrelation decreases as lag k increases. In most of the time series, it is noticed that the absolute value of r_k i.e. $|r_k|$ decreases as k increases. This is because observations which are located far away are not much related to each other, whereas observations close may be positively or negatively correlated.

Let us understand how we plot a correlogram with the help of an example.

Example 3: For the data given in Example 2 of Unit 12, plot the correlogram.

Solution: A correlogram is a plot of the autocorrelation function with respect to its lag, therefore, first of all, we have to compute the sample autocorrelation coefficients. In Example 2, we have already calculated these. Therefore, to the sake of time, we just write them here

$$r_1 = 0.835,\ r_2 = 0.676,\ r_3 = 0.469,\ r_4 = 0.280$$

For the correlogram, we take lags on the X-axis and sample autocorrelation function on the Y-axis. At each lag, we draw a line, which represents the level of correlation between the series and a lagged version of itself, as shown in the following Fig. 13.2.

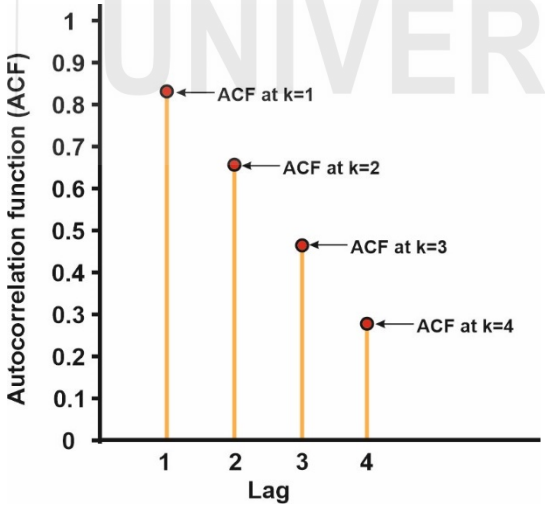

Fig. 13.2: The correlogram for lag k = 1, 2, 3 and 4.

After learning what is correlogram and how we plot it, we now understand how the correlogram helps us to recognise the nature of a time series.

13.6 INTERPRETATION OF CORRELOGRAM

A correlogram is a graph used to interpret a set of autocorrelation functions in which the autocorrelation function is plotted against its lag. It is often very

helpful for visual inspection to recognise the nature of time series, though it is not always easy. We now describe certain types of time series and the nature of their correlograms.

Random Series

A time series is completely random if it contains only independent observations. Therefore, the values of the autocorrelation function for such a series are approximately zero, that is, $r_k \simeq 0$ and the correlogram of such a random time series will be moving around the zero line. The typical correlogram is shown in Fig. 13.3.

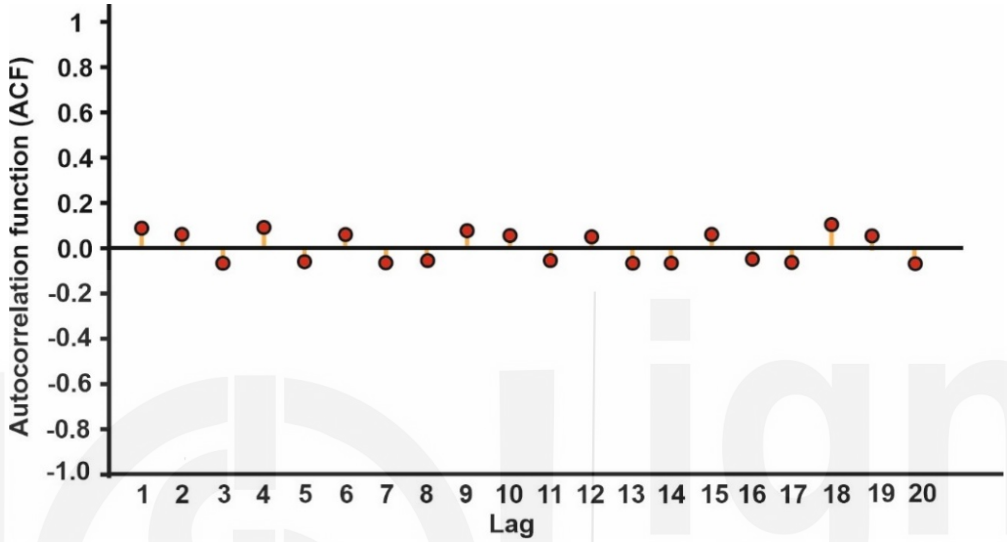

Fig. 13.3: The correlogram of random series.

Alternating Series

If a time series behaves in a very rough and zig-zag manner, alternating between above and below mean, then it indicates by negative r_k and positive r_{k+1} and vice-versa. The correlogram of an alternating time series is shown in Fig. 13.4.

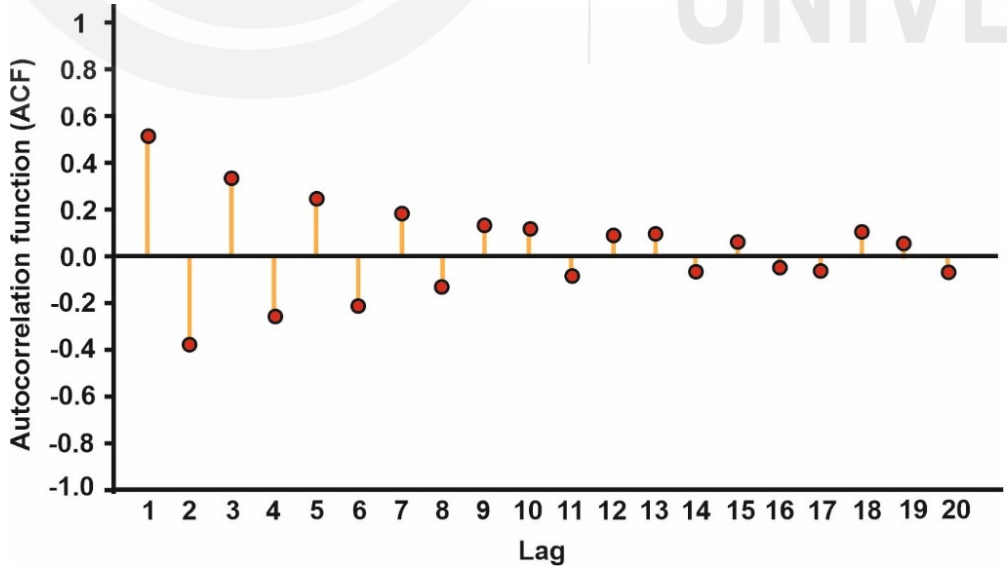

Fig. 13.4: The correlogram of alternating series.

Stationary Time Series

A time series is said to be stationary if its mean, variance and covariance are almost constant and it is free from trend and seasonal effects. The

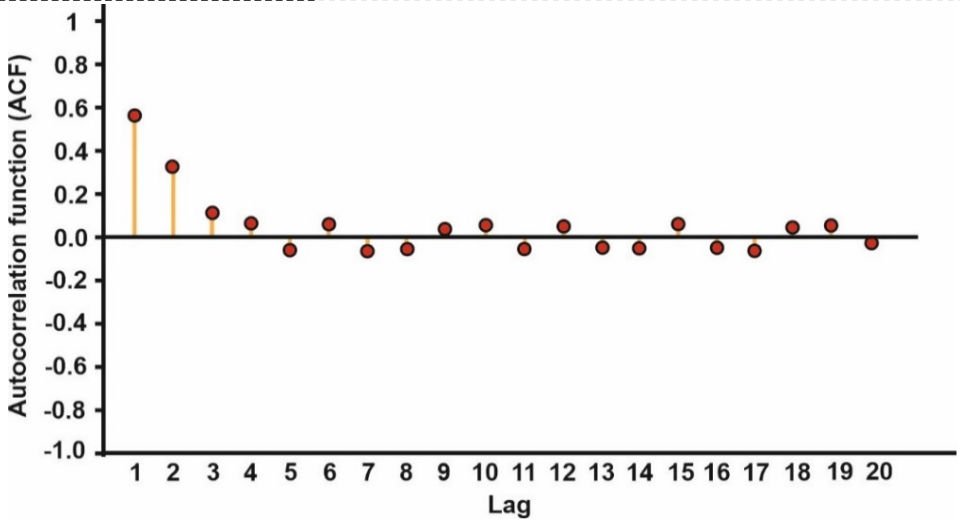

Fig. 13.5: The correlogram of stationary time series.

correlogram of the stationary series has a few large autocorrelations in absolute value for small lag k and they tend to zero very rapidly with an increase in lag k (See Fig. 13.5). A model called an autoregressive model (you will study in the next session), may be appropriate for a series of this type.

Nonstationary Time Series

A time series is said to be nonstationary if its mean, variance, and covariance change over time. Therefore, a time series which contains trend, seasonality cycles, random walks, or combinations of these is nonstationary. Such a series is usually very smooth in nature and its autocorrelations go to zero very slowly as the observations are dominated by trend. We should remove the trend from such a time series before doing any further analysis. The time series with trend and seasonal effects are as follows:

(i) Trend Time Series

If a time series has a trend effect, then a time plot will show an upward or down word pattern as you have seen in Unit 10. For such type of time series, the correlogram decreases in an almost linear fashion as the lags increase as shown in Fig. 13.6. Hence a correlogram of this type is a clear indication of a trend.

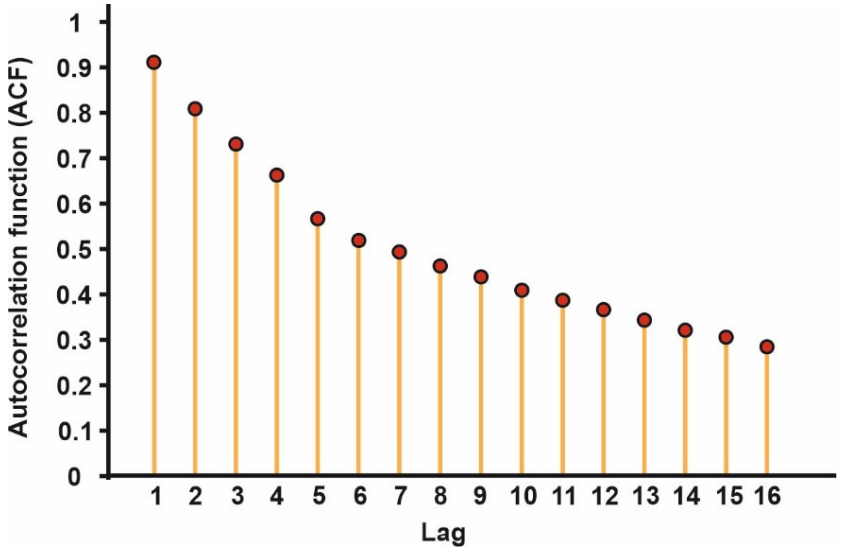

Fig. 13.6: The correlogram of time series having trend effect.

(ii) Seasonal Time Series

If a time series has a dominant seasonal pattern, then a time plot will show cyclical behaviour with a periodicity of the season. The correlogram will also exhibit an oscillation behaviour as shown in Fig. 13.7. If there is seasonality, say, 12 months, then the ACF value will be large and positive at lag 12 and possibly also at lags 24, 36, . . .Similarly, for quarterly seasonal data, a large ACF value will be seen at lag 4 and possibly also at lags 8,12, …. If the seasonal variation is removed from time series data then the correlogram may provide useful information. Therefore, in this case, the correlogram may not contain any more information than what is given by the time plot of the time series.

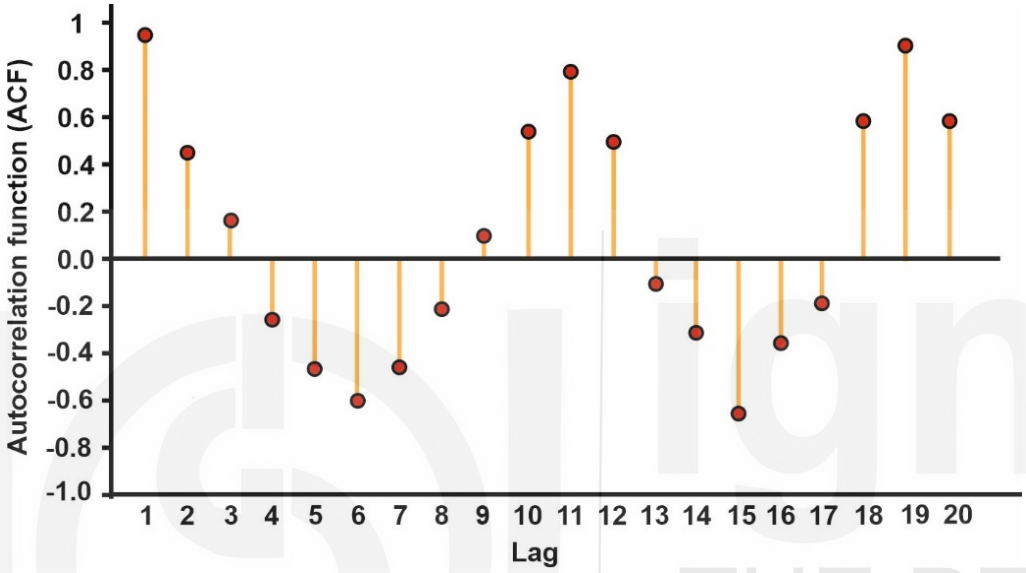

Fig. 13.7: The correlogram of time series having seasonal effect.

In general, the interpretation of a correlogram is not easy and requires a lot of experience and insight.

You may like to try the following Self Assessment Question.

SAQ 3

A share market expert wants to study the pattern of a particular share price. For that, he calculates the autocorrelation for different lags which are given as follows:

$r_0 = 1, r_1 = 0.482$, $r_2 = 0.050$, $r_3 = -0.159$, $r_4 = 0.253$, $r_5 = -0.024$, $r_6 = 0.053$,

$r_7 = 0.025$, $r_8 = -0.252$, $r_9 = -0.177$, $r_{10} = 0.006$, $r_{11} = 0.390$, $r_{12} = -0.838$,

$r_{13} = 0.407$, $r_{14} = 0.010$, $r_{15} = -0.181$, $r_{16} = -0.257$, $r_7 = -0.057$, $r_{18} = 0.016$

$r_{19} = -0.051$

For the above information:

(i) Plot the correlogram.

(ii) Interpret the correlogram. Is the seasonality apparent in the correlogram?

We end this unit by giving a summary of its contents.

13.7 SUMMARY

In this unit, we have discussed:

- Role of correlation analysis in time series.
- The covariance between a given time series and a lagged version of itself over successive time intervals is called autocovariance. The formula for calculating the autocovariance function is given as

$$\gamma_k = \gamma_{-k} = \text{Cov}(Y_t, Y_{t+k}) = \frac{1}{N}\sum_{t=1}^{N-k}(Y_t - \mu)(Y_{t+k} - \mu)$$

  and its estimate using sample data is as follows:

$$\hat{\gamma}_k = c_k = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y}); \quad k = 1,2,...,n-1$$

- Autocorrelation is a measure of the degree of relationship between a given time series and a lagged version of itself over successive time intervals. The formula for calculating the autocorrelation function is given as

$$\rho_k = \frac{\sum_{t=1}^{N-k}(Y_t - \mu)(Y_{t+k} - \mu)}{\sum_{t=1}^{N}(Y_t - \mu)^2} = \frac{\gamma_k}{\gamma_0}$$

  and its estimate using sample data is as follows:

$$\hat{\rho}_k = r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} = \frac{c_k}{c_0}; \quad k = 1,2,...,n-1$$

- The partial autocorrelation function calculates the degree of relationship between a time series Y_t with its own lagged values Y_{t+k} after their mutual linear dependency on the intervening variables Y_{t+1}, Y_{t+2},…, Y_{t+k-1} has been removed. We can calculate it using the autocorrelation function as

$$\phi_{11} = \rho_1$$

$$\phi_{22} = \frac{(\rho_2 - \rho_1^2)}{(1 - \rho_1^2)}$$

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|}$$

- A plot in which we take the autocorrelation function on the vertical axis and different lags on the horizontal axis is known as a correlogram.

13.8 TERMINAL QUESTIONS

For the data which is obtained after the first difference of the time series data (sales of a new single house in a region) of TQ 1 of Unit 12

(i) Calculate ACF.

(ii) Interpret the correlogram. Is the trend apparent in the correlogram?

13.9 SOLUTION/ANSWERS

Self Assessment Questions (SAQs)

1. Since there are 12 observations, therefore, we prepare the data up to n/4 = 12/4 = 3 lags as follows:

Quarter	Unemployment (y_t)	y_{t+1}	y_{t+2}	y_{t+3}
1	91			
2	45	91		
3	89	45	91	
4	36	89	45	91
5	72	36	89	45
6	51	72	36	89
7	64	51	72	36
8	99	64	51	72
9	64	99	64	51
10	89	64	99	64
11	68	89	64	99
12	108	68	89	64
Total	**876**			

For the calculation of the autocorrelation function, we assume that the time series is stationary, therefore, mean and variance of the series will be constant. Thus, we calculate the sample mean and variance of the given original time series and make the necessary calculations for calculating the autocovariance and autocorrelation function in the following table:

$y_t - \bar{y}$	$(y_t - \bar{y})^2$	$y_{t+1} - \bar{y}$	$y_{t+2} - \bar{y}$	$y_{t+3} - \bar{y}$	$(y_t - \bar{y})(y_{t+1} - \bar{y})$	$(y_t - \bar{y})(y_{t+2} - \bar{y})$	$(y_t - \bar{y})(y_{t+3} - \bar{y})$
18	324						
–28	784	18			–504		
16	256	–28	18		–448	288	
–37	1369	16	–28	18	–592	1036	–666
–1	1	–37	16	–28	37	–16	28
–22	484	–1	–37	16	22	814	–352
–9	81	–22	–1	–37	198	9	333
26	676	–9	–22	–1	–234	–572	–26
–9	81	26	–9	–22	–234	81	198
16	256	–9	26	–9	–144	416	–144
–5	25	16	–9	26	–80	45	–130
35	1225	–5	16	–9	–175	560	–315
0	**5562**				**–2154**	**2661**	**–1074**

Therefore,

$$\text{Mean} = \frac{1}{n}\sum_{t=1}^{n} y_t = \frac{876}{12} = 73,$$

$$\text{Variance} = c_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2 = \frac{5562}{12} = 463.5$$

Autocovariance

$$c_1 = \frac{1}{n}\sum_{t=1}^{n-1}(y_t - \bar{y})(y_{t+1} - \bar{y}) = \frac{1}{15} \times -2154 = -179.5$$

$$c_2 = \frac{1}{n}\sum_{t=1}^{n-2}(y_t - \bar{y})(y_{t+2} - \bar{y}) = \frac{1}{12} \times 2661 = 221.75$$

$$c_3 = \frac{1}{n}\sum_{t=1}^{n-3}(y_t - \bar{y})(y_{t+3} - \bar{y}) = \frac{1}{12} \times -1074 = -89.5$$

After calculating the autocovariance function, we now calculate the sample autocorrelation function as

$r_1 = -0.387$, $r_2 = 0.478$, $r_3 = -0.193$

2. In SAQ 1, we have already calculated the sample autocorrelation coefficients which are as follows:

$$r_1 = \frac{c_1}{c_0} = \frac{-179.5}{463.9} = -0.387, \; r_2 = \frac{c_2}{c_0} = \frac{221.75}{463.9} = 0.478$$

$$r_3 = \frac{c_3}{c_0} = \frac{-89.5}{463.9} = -0.193$$

Since the 1st-order partial autocorrelation equals to the 1st-order autocorrelation, therefore,

$\phi_{11} = r_1 = -0.387$

We can compute the 2nd order (lag) sample partial autocorrelation function as

$$\hat{\phi}_{22} = \frac{(r_2 - r_1^2)}{(1 - r_1^2)} = \frac{0.478 - (-0.387)^2}{1 - (-0.193)^2} = 0.386$$

3. For plotting the correlogram, we take lags on the X-axis and sample autocorrelation coefficients on the Y-axis. At each lag, we draw a line, which represents the level of correlation between the series and its lags, as shown in the following Fig. 13.8.

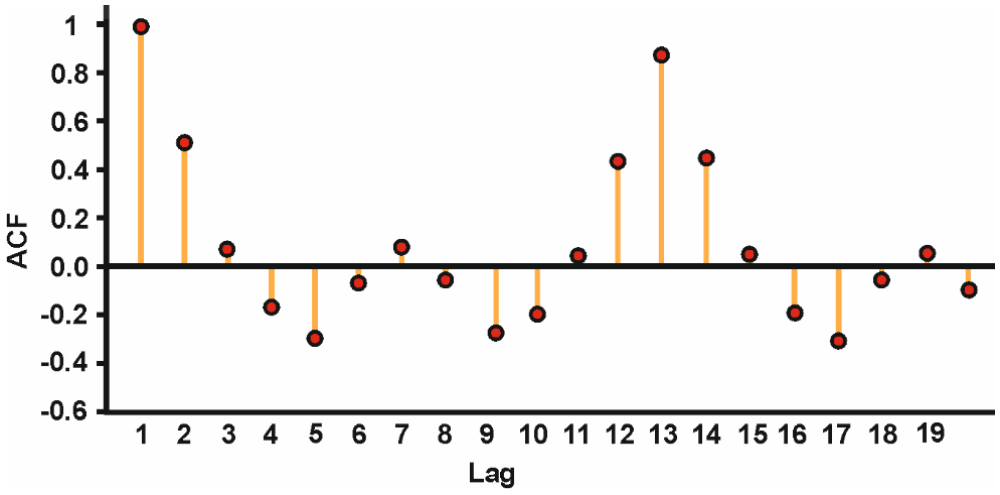

Fig. 13.8: The correlogram of time series of the share price.

Since the correlogram shows an oscillation, therefore, the time series of the share price is not stationary. The frequency of the oscillations is almost the same, therefore, it has a seasonal effect.

Terminal Questions (TQs)

1. For calculating the first sample autocorrelation of the first-order difference, we prepare the data of lags. Since there are 14 observations, therefore, we prepare the data up to n/4 = 14/4 ~ 4 lags as follows:

Month	First Difference (y_t)	y_{t+1}	y_{t+2}	y_{t+3}	y_{t+4}
2	38				
3	21	38			
4	32	21	38		
5	18	32	21	38	
6	5	18	32	21	38
7	15	5	18	32	21
8	25	15	5	18	32
9	20	25	15	5	18
10	10	20	25	15	5
11	15	10	20	25	15
12	30	15	10	20	25
13	8	30	15	10	20
14	32	8	30	15	10
15	25	32	8	30	15

For the calculation of the autocorrelation function, we assume that the time series is stationary, therefore, mean and variance of the series will be constant. Thus, we calculate the sample mean and variance of the time series (first difference data) and make the necessary calculations for calculating the autocovariance and autocorrelation function in the following table:

$y_t-\bar{y}$	$(y_t-\bar{y})^2$	$y_{t+1}-\bar{y}$	$y_{t+2}-\bar{y}$	$y_{t+3}-\bar{y}$	$y_{t+4}-\bar{y}$	$(y_t-\bar{y})(y_{t+1}-\bar{y})$	$(y_t-\bar{y})(y_{t+2}-\bar{y})$	$(y_t-\bar{y})(y_{t+3}-\bar{y})$	$(y_t-\bar{y})(y_{t+4}-\bar{y})$
17	289								
0	0	17				0			
11	121	0	17			0	187		
–3	9	11	0	17		–33	0	–51	
–16	256	–3	11	0	17	48	–176	0	–272
–6	36	–16	–3	11	0	96	18	–66	0
4	16	–6	–16	–3	11	–24	–64	–12	44
–1	1	4	–6	–16	–3	–4	6	16	3
–11	121	–1	4	–6	–16	11	–44	66	176
–6	36	–11	–1	4	–6	66	6	–24	36
9	81	–6	–11	–1	4	–54	–99	–9	36
–13	169	9	–6	–11	–1	–117	78	143	13
11	121	–13	9	–6	–11	–143	99	–66	–121
4	16	11	–13	9	–6	44	–52	36	–24
Total	**1272**					**–10**	**–41**	**33**	**–109**

$$\text{Mean} = \frac{1}{n}\sum_{t=1}^{n} y_t = \frac{294}{14} = 21$$

$$\text{Variance} = c_0 = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \bar{y})^2 = \frac{1272}{14} = 90.86$$

Autocovariance

$$c_1 = \frac{1}{n}\sum_{t=1}^{n-1}(y_t - \bar{y})(y_{t+1} - \bar{y}) = \frac{1}{14} \times -110 = -7.86$$

$$c_2 = \frac{1}{n}\sum_{t=1}^{n-2}(y_t - \bar{y})(y_{t+2} - \bar{y}) = \frac{1}{14} \times -41 = -2.93$$

$$c_3 = \frac{1}{n}\sum_{t=1}^{n-3}(y_t - \bar{y})(y_{t+3} - \bar{y}) = \frac{1}{14} \times 33 = 2.36$$

$$c_4 = \frac{1}{n}\sum_{t=1}^{n-4}(y_t - \bar{y})(y_{t+4} - \bar{y}) = \frac{1}{14} \times -109 = -7.79$$

After calculating the autocovariance, we now calculate the sample autocorrelation function as

$$r_1 = \frac{c_1}{c_0} = \frac{-7.86}{90.86} = -0.086\,,\ r_2 = \frac{c_2}{c_0} = \frac{-2.93}{9086} = -0.032$$

$$r_3 = \frac{c_3}{c_0} = \frac{2.36}{90.86} = 0.026\,,\ r_4 = \frac{c_4}{c_0} = \frac{-7.79}{90.86} = -0.086$$

For the correlogram, we take lags on the X-axis and sample autocorrelation function on the Y-axis. At each lag, we draw a line, which represents the level of correlation between the series and its lags, as shown in the following Fig. 13.9.

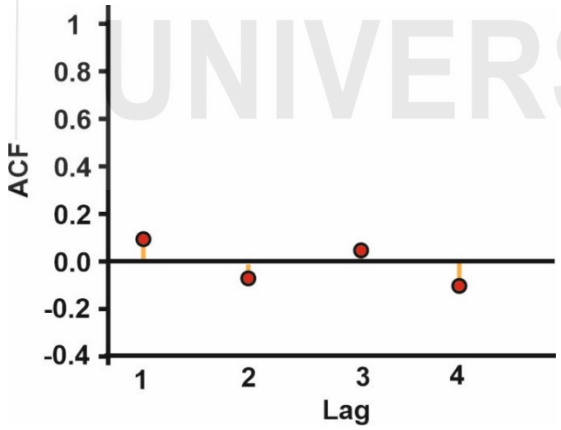

Fig. 13.9: The correlogram of time series data of sales of new single houses.

Since the autocorrelation function is approximately zero and moving around the zero line, therefore, the time series is stationary and no trend appeared in the correlogram.