ARIMA (Auto-Regressive Integrated Moving Average)

An **autoregressive integrated moving average**, or **ARIMA,** is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends.

A statistical model is autoregressive if it predicts future values based on past values.

It's a way of modelling time series data for forecasting (i.e., for predicting future points in the series), in such a way that:

- a pattern of growth/decline in the data is accounted for (hence the "auto- regressive" part)
- the rate of change of the growth/decline in the data is accounted for (hence the "integrated" part)
- noise between consecutive time points is accounted for (hence the "moving average" part)

"time series data" = data that is made up of a sequence of data points taken at successive equally spaced points in time

ARIMA (Auto-Regressive Integrated Moving Average)

- Autoregressive Integrated Moving Average models (ARIMA models) were popularized by George Box and Gwilym Jenkins in the early 1970s.

- ARIMA models are a class of linear models that is capable of representing *stationary as well as non-stationary time series*.

- ARIMA models *do not involve independent variables* in their construction. They make use of the information in the series itself to generate forecasts.

ARIMA (Auto-Regressive Integrated Moving Average)

- ARIMA models rely heavily on autocorrelation patterns in the data.
- ARIMA methodology of forecasting is different because it does not assume any particular pattern in the historical data of the series to be forecast.
- It uses an interactive approach of identifying a possible model from a general class of models.
- The chosen model is then checked against the historical data to see if it accurately describe the series.
- a time series data is a sequence of numerical observations naturally ordered in time,

Box-Jenkins Methodology

- Box and Jenkins suggested a way of converting a non-stationary time series into a stationary time series with the help of some transformations.

- Stationarity is a property of a time series. A stationary series is one where the values of the series is not a function of time.

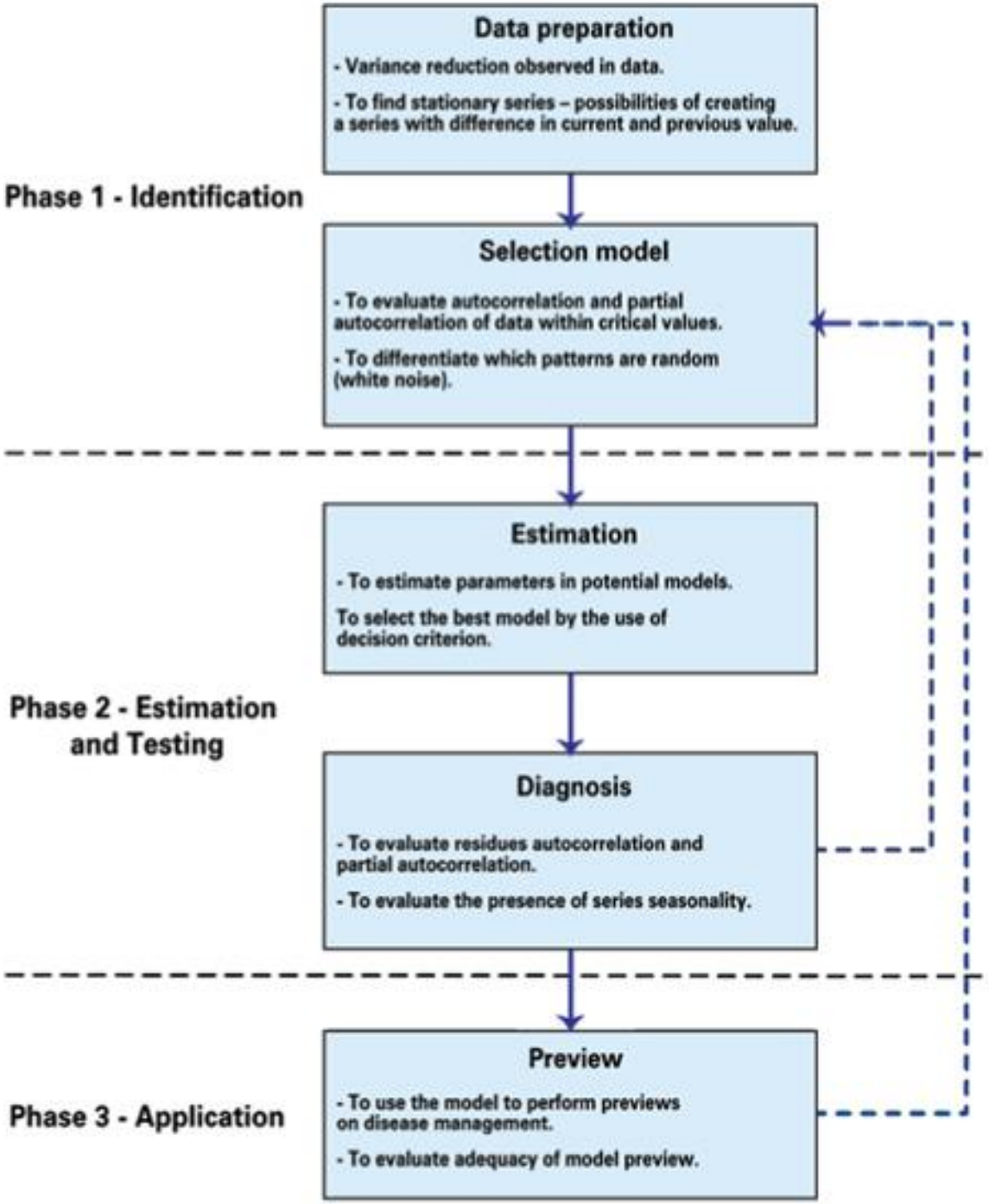
- That is, the statistical properties of the series like mean, variance and autocorrelation are constant over time.

- *Autocorrelation of the series is the correlation of the series with its previous values.*

- A stationary time series is devoid of seasonal effects as

Box-Jenkins Methodology

- The Box-Jenkins methodology refers to a set of procedures for *identifying*, *fitting*, and *checking* ARIMA models with time series data.
- Forecasts follow directly from the form of fitted model.
- The basis of BOX-Jenkins approach to modeling time series consists of three phases:


Data preparation
- Variance reduction observed in data.
- To find stationary series – possibilities of creating a series with difference in current and previous value.
Phase 1 - Identification
Selection model
- To evaluate autocorrelation and partial autocorrelation of data within critical values.
- To differentiate which patterns are random (white noise).
Estimation
- To estimate parameters in potential models.
To select the best model by the use of decision criterion.
Phase 2 - Estimation and Testing
Diagnosis
- To evaluate residues autocorrelation and partial autocorrelation.
- To evaluate the presence of series seasonality.
Preview
- To use the model to perform previews on disease management.
- To evaluate adequacy of model preview.
Phase 3 - Application

autocorrelation coefficient

The key statistic in time series analysis is the autocorrelation coefficient (the correlation of the time series with itself, lagged 1, 2, or more periods).

The autocorrelation formula:

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

autocorrelation coefficient

- r_1 indicates how successive values of Y relate to each other, r_2 indicates how Y values two periods apart relate to each other, and so on.

- The auto correlations at lag 1, 2, …, make up the autocorrelation function or ACF.

- Autocorrelation function is a valuable tool for investigating properties of an empirical time series.

The Partial autocorrelation coefficient

Partial autocorrelations measures the degree of association between y_t and y_{t-k}, when the effects of other time lags 1, 2, 3, …, k-1 are removed.

The partial autocorrelation coefficient of order k is evaluated by regressing y_t against y_{t-1},…y_{t-k}:

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \square \quad + b_k y_{t-k}$$

α_k (partial autocorrelation coefficient of order k) is the estimated coefficient b_k.

The Partial autocorrelation coefficient

$$PACF(Y_t, Y_{t-2}) = \frac{Cov(Y_t, Y_{t-2} \mid Y_{t-1})}{\sqrt{Var(Y_t \mid Y_{t-1})\, Var(Y_{t-2} \mid Y_{t-1})}}$$

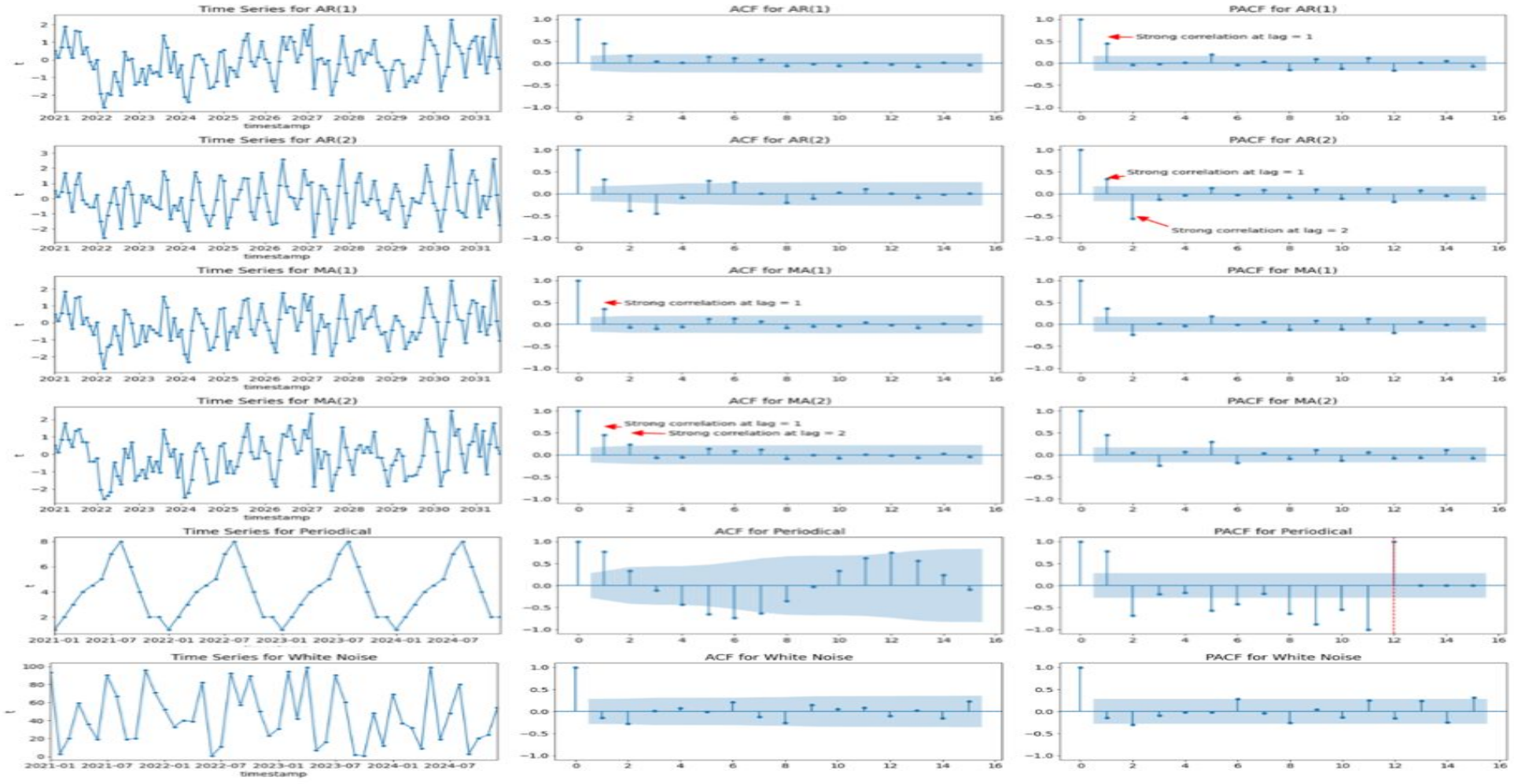
$$PACF(Y_t, Y_{t-3}) = \frac{Cov(Y_t, Y_{t-3} \mid Y_{t-1}, Y_{t-2})}{\sqrt{Var(Y_{t-3} \mid Y_{t-1}, Y_{t-2})} \times \sqrt{Var(Y_t \mid Y_{t-1}, Y_{t-2})}}$$

(ACF) **Autocorrelation Function**	(PACF) **Partial Autocorrelation Function**
Finds the correlation between two values by taking all the past values in time series under consideration (Does not remove the effect of shorter lags autocorrelation and considers them all while estimating longer lags).	Does not take all the past values and instead considers only one past value for finding current one (Removes the effect of shorter lags autocorrelation for estimating longer lags).
There is more than one time lag in values of times series while finding out the correlation between two values.	There is only one time lag between current and one past value.
Uses indirect impacts to the observed value.	Uses direct impact of one past value on the current value.
Does not use coefficient since this type compares all values from the past for finding out the current value.	Uses coefficient since that gives the multiplier effect of one past value to the current value for finding the latter aptly.

Interpreting PACF and ACF

Precondition: Time series must be stationary

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off (Geometric decay)	Significant at lag q / Cuts off after lag q	Tails off (Geometric decay)
PACF	Significant at each lag p / Cuts off after lag p	Tails off (Geometric decay)	Tails off (Geometric decay)



Examining stationarity of time series data

Stationarity means no growth or decline.

Data fluctuates around a constant mean independent of time and variance of the fluctuation remains constant over time.

Stationarity can be assessed using a time series plot:

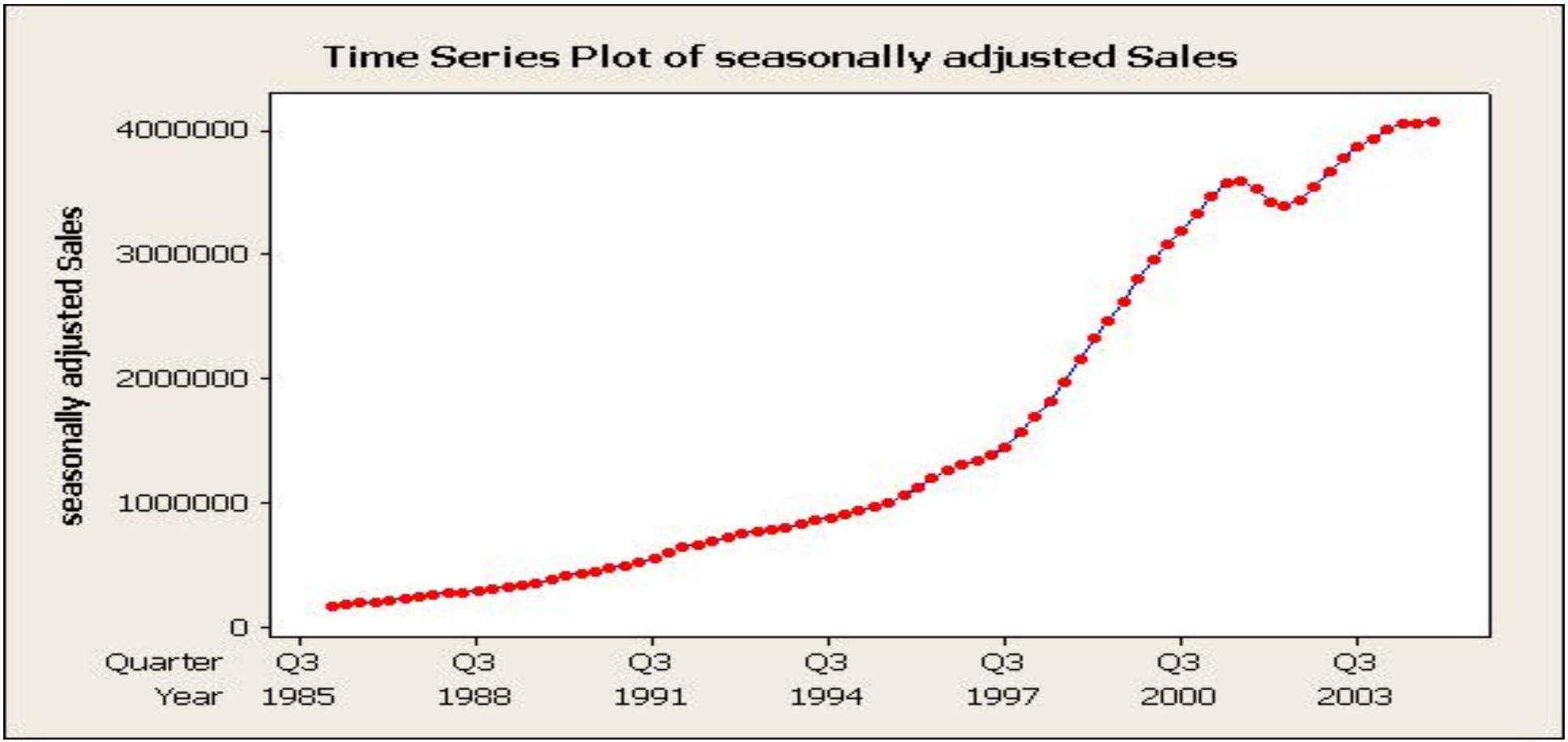
Plot shows no change in the mean over time

No obvious change in the variance over time.

The autocorrelation plot can also show non-stationarity.

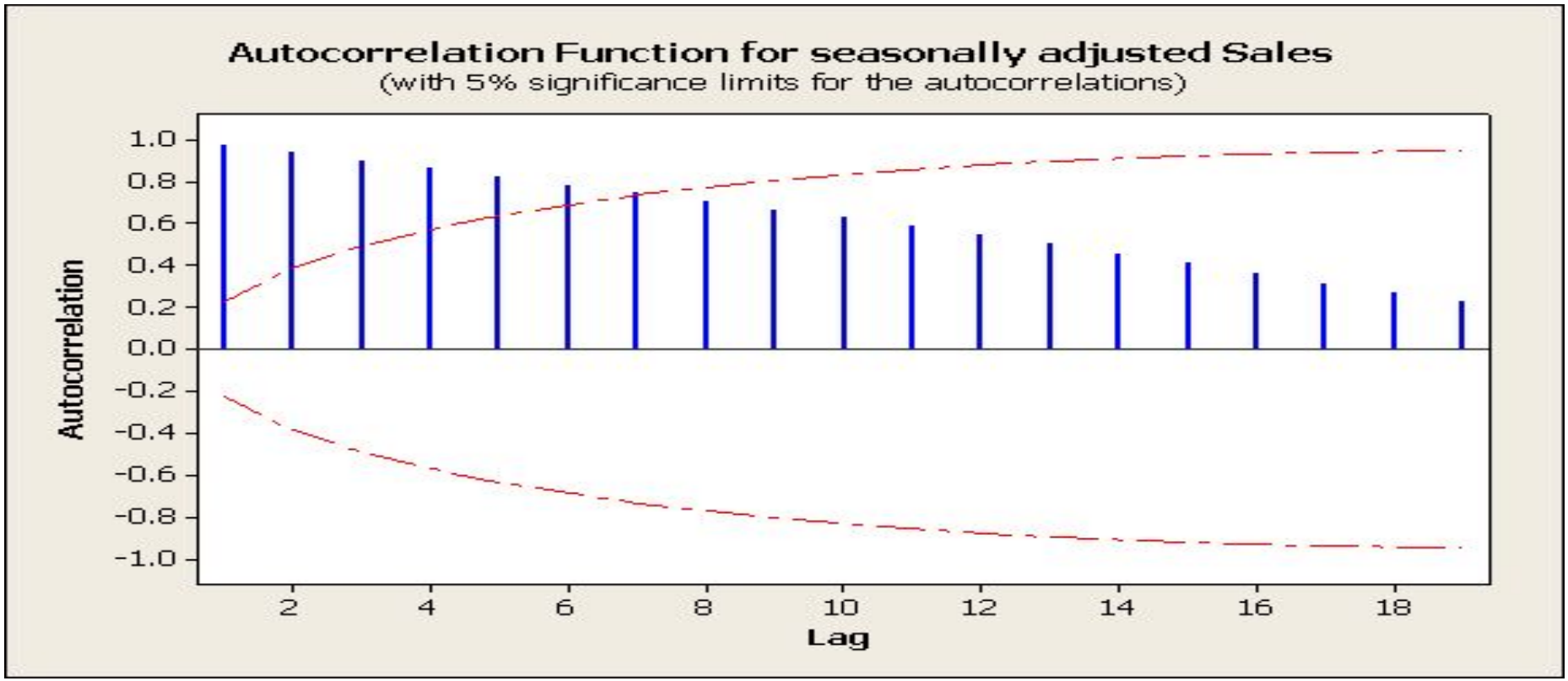
- Significant autocorrelation for several time lags and slow decline in r_k indicate non-stationarity.

Examining stationarity of time series data



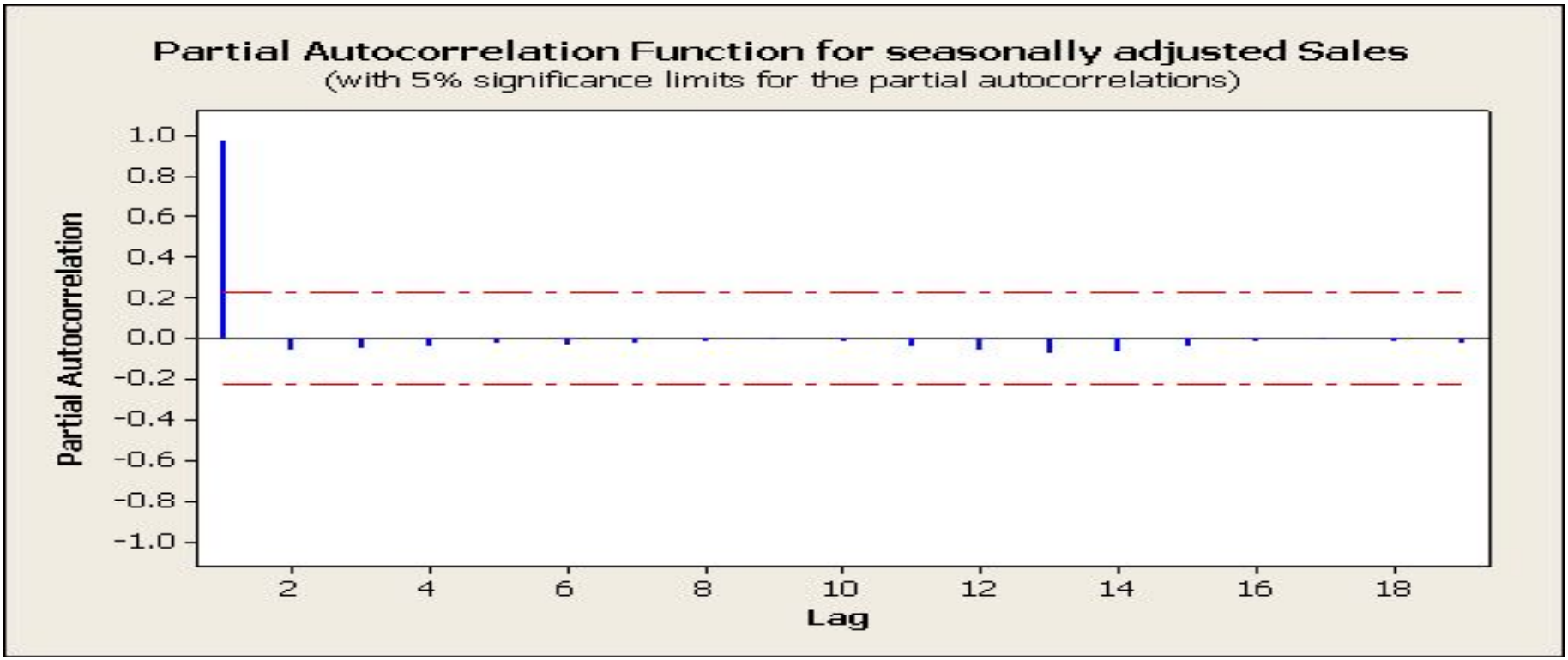
- The time series plot shows that it is **non-stationary** in the mean.
- The next slide shows the ACF plot for this data series.

Examining stationarity of time series data



- The ACF also shows a pattern typical for a non-stationary series:
  - Large significant ACF for the first 7 time lag
  - Slow decrease in the size of the autocorrelations.
- The PACF is shown in the next slide.

Examining stationarity of time series data



This is also typical of a non-stationary series. Partial autocorrelation at time lag 1 is close to one and the partial autocorrelation for the time lag 2 through 18 are close to zero

Removing non-stationarity in time series

The non-stationary pattern in a time series data needs to be removed in order that other correlation structure present in the series can be seen before proceeding with model building.

One way of removing non-stationarity is through the method of differencing.

How to make a time series stationary?

One can make series stationary by:

- ✔ Differencing the Series (once or more)
- ✔ Take the log of the series
- ✔ Take the nth root of the series
- ✔ Combination of the above

The most common and convenient method to stationarize the series is by differencing the series at least once until it becomes approximately stationary.

Differencing of Series

If Y_t is the value at time 't', then the first difference is given by:

$$\mathbf{Y = Y_t - Y_{t-1}}$$

In simpler terms, differencing the series is nothing but subtracting the next value by the current value.

If the first difference doesn't make a series stationary, you can go for the second differencing. And so on.

Differencing of Series

For example, consider the following series: [1, 5, 2, 12, 20]

First differencing gives: [5-1, 2-5, 12-2, 20-12] = [4, -3, 10, 8]

Second differencing gives: [-3-4, 10-(-3), 8-10] = [-7, 13, -2]

Why make a non-stationary series stationary before forecasting?

- Forecasting a stationary series is relatively easy and the forecasts are more reliable.

- An important reason is, autoregressive forecasting models are essentially linear regression models that utilize the lag(s) of the series itself as predictors.