NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

Forecasting and Time Series
Advanced Time Series
Day 3

Raymond Freth A. Lagria
Instructor
Department Industrial Engineering and Operations Research
University of the Philippines Diliman

Module 5 of the Business Intelligence and Analytics Track of UP NEC and the UP Center of Business Intelligence

Outline for This Training

1. Introduction to Forecasting in Business Intelligence
2. Demand Forecasting Techniques
   - Qualitative
   - Quantitative
3. Accuracy of Forecasts
4. Monitoring of Forecasts
5. Forecasting with R
6. Introduction to Time Series Data Mining
7. **Advanced Time Series**

Advanced Times Series

- Introduction
- ACF and PACF
- Model Identification
- Parameter Estimation
- Forecasting

ARIMA: An Introduction

- 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: An Introduction

- AR – auto-regressive or autoregression part involves regression the variable (Y) on its own lagged values (e.g., past) values.
- I – differencing
- MA – moving average or modelling the error term as a linear combination of error terms simultaneously occurring in the past

ARIMA: An Introduction

- ARIMA models rely heavily on autocorrelation patterns in the data.
- ARIMA methodology of forecasting is different from most methods 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.

ARIMA: An Introduction

- 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:
  - Identification
  - Estimation and testing
  - Application

Methodology

- **Identification**
  - Data preparation
    - Transform data to stabilize variance
    - Differencing data to obtain stationary series
  - Model selection
    - Examine data, use ACF and PACF to identify potential models
- **Estimation and testing**
  - Estimation
    - Estimate parameters in potential models
    - Select best model using suitable criterion
  - Diagnostics
    - Check ACF/PACF of residuals
    - Are the residuals white noise?
- **Application**
  - Forecasting: use model to forecast

Examining Correlation in Time Series Data

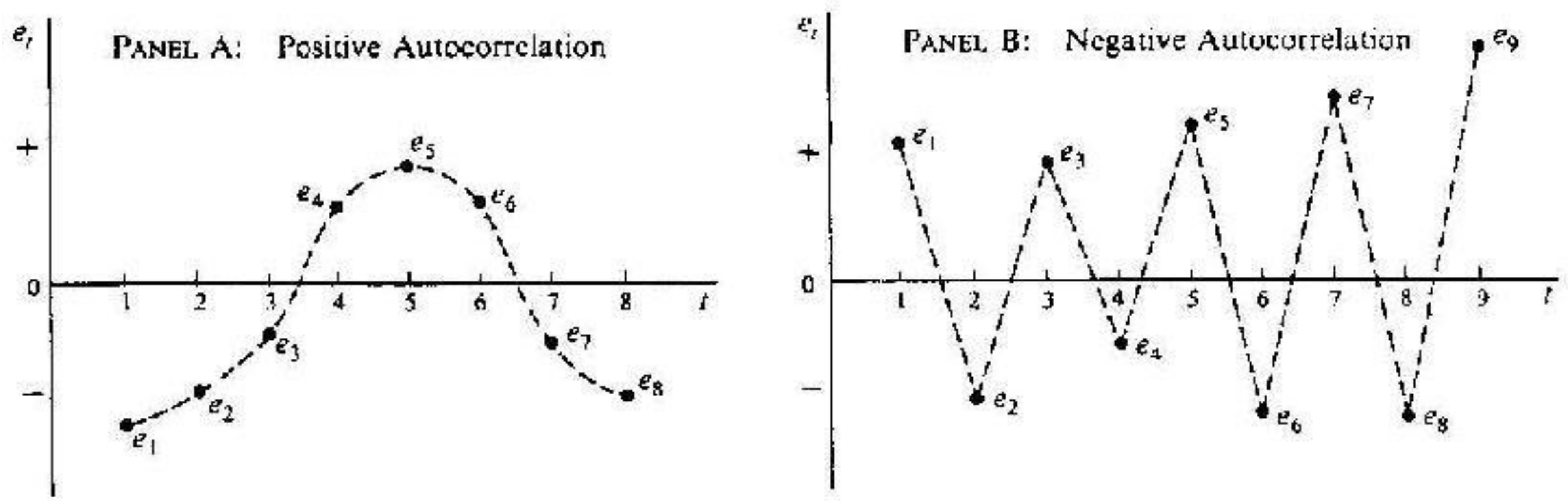
- The key statistic in time series analysis is the autocorrelation coefficient
- Autocorrelation is the correlation of the time series with itself, lagged k =1, 2, or more periods.
- We define the autocorrelation formula:

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

Examining Correlation in Time Series Data

- Recall 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.

Positive and Negative Autocorrelation



- http://www.oocities.org/qecon2002/founda10.html

The White Noise Model

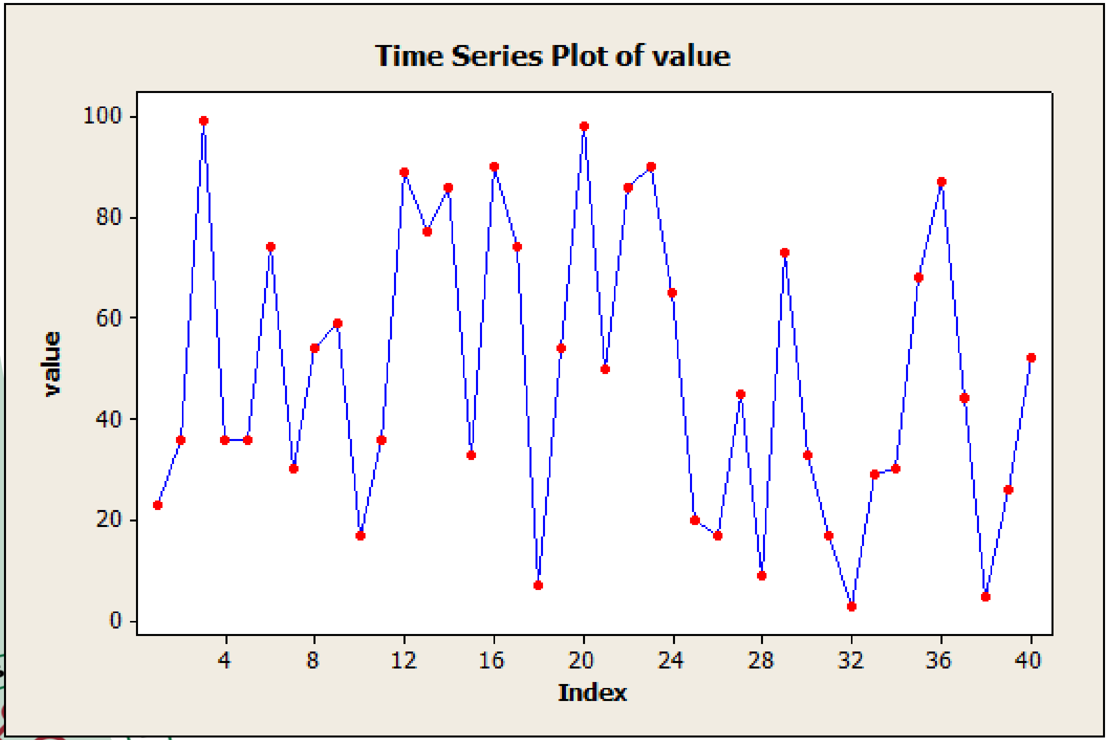
- A white noise model is a model where observations Y_t is made of two parts: a fixed value and an uncorrelated random error component.

$$y_t = C + e_t$$

- For uncorrelated data (a time series which is white noise) we expect each autocorrelation to be close to zero.

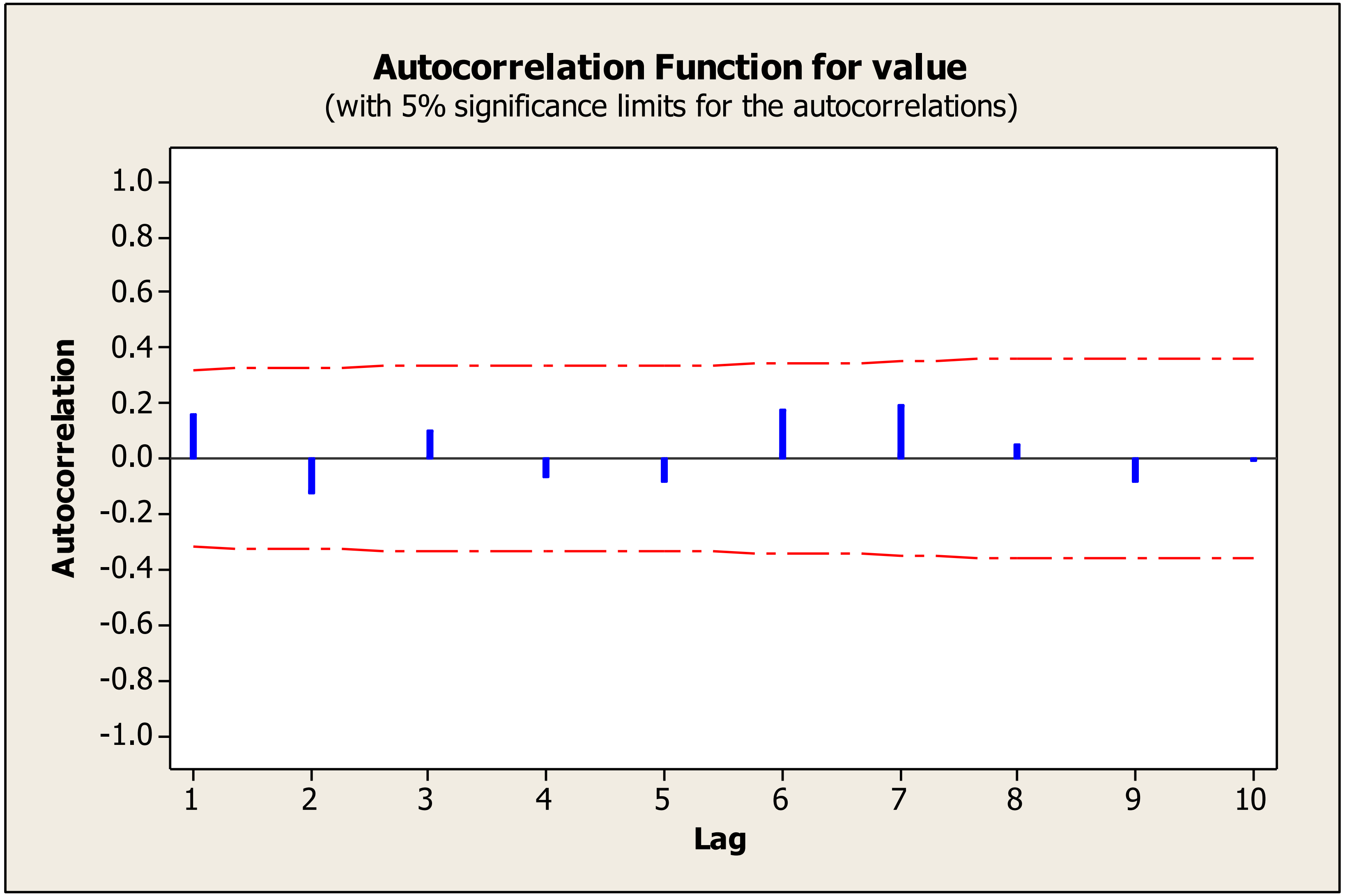
Example: White Noise Series

Consider the following white noise series.



period	value	period	value
1	23	21	50
2	36	22	86
3	99	23	90
4	36	24	65
5	36	25	20
6	74	26	17
7	30	27	45
8	54	28	9
9	59	29	73
10	17	30	33
11	36	31	17
12	89	32	3
13	77	33	29
14	86	34	30
15	33	35	68
16	90	36	87
17	74	37	44
18	7	38	5
19	54	39	26
20	98	40	52

ACF For The White Noise Series



Sampling Distribution of Autocorrelation

- The autocorrelation coefficients of white noise data have a sampling distribution that can be approximated by a normal distribution with mean zero and standard error $\frac{1}{\sqrt{n}}$. where n is the number of observations in the series.
- This information can be used to develop tests of hypotheses and confidence intervals for ACF.

Sampling Distribution of Autocorrelation

- For example
  - For our white noise series example, we expect 95% of **all sample ACF** to be within

$$\pm 1.96 \frac{1}{\sqrt{n}} = \pm 1.96 \frac{1}{\sqrt{40}} = \pm .3099$$

  - If this is not the case then the series is not white noise.
  - The sampling distribution and standard error allow us to distinguish what is randomness or white noise from what is pattern.

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 $y_{t-1}, \ldots y_{y-k}$:

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

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

The Partial Autocorrelation Coefficient

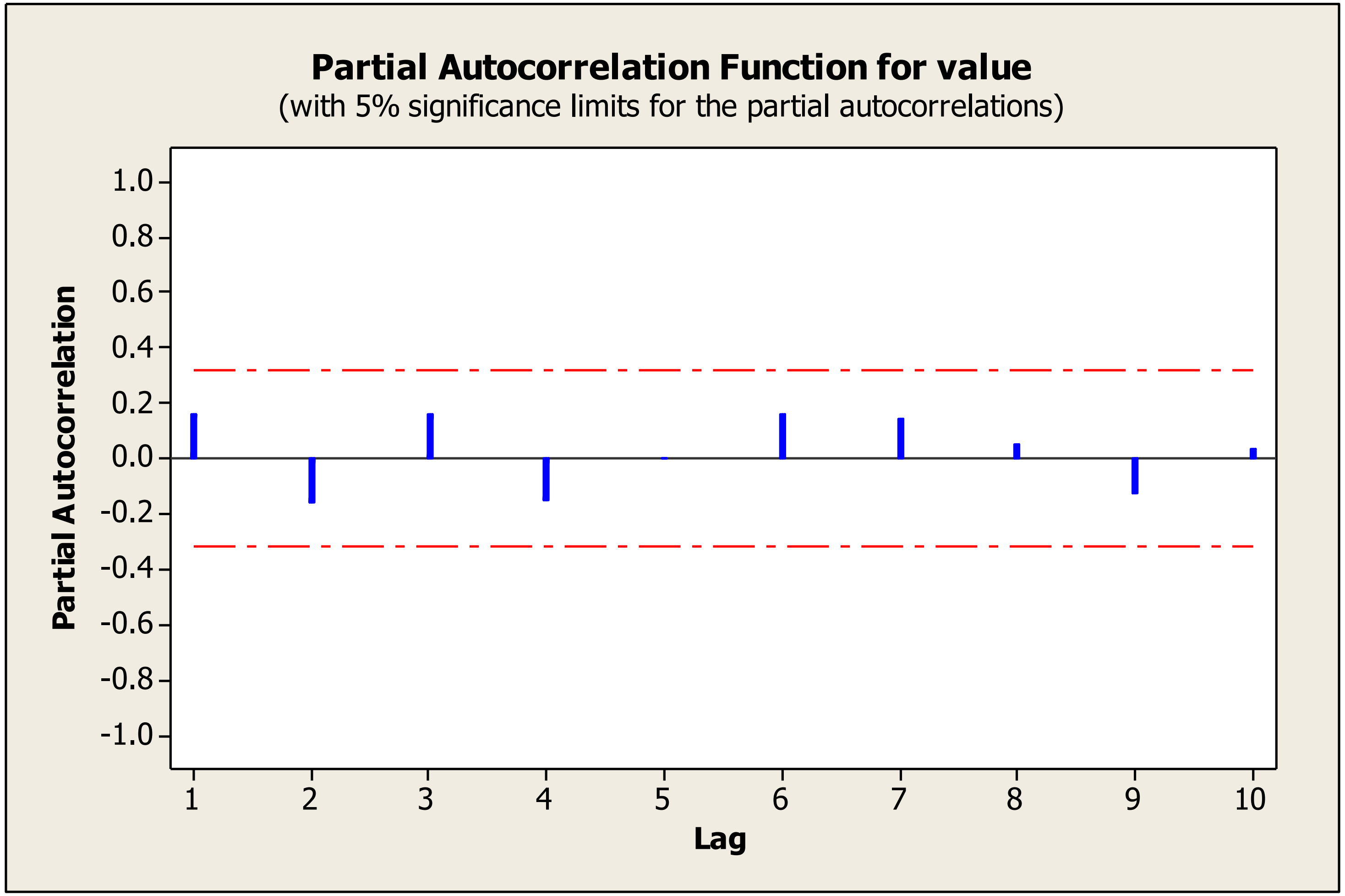
- The partial autocorrelation functions (PACF) should all be close to zero for a white noise series.
- If the time series is white noise, the estimated PACF are approximately independent and normally distributed with a standard error $\frac{1}{\sqrt{n}}$.
- Therefore the same critical values of

$$\pm 1.96 \frac{1}{\sqrt{n}}$$

Can be used with PACF to asses if the data are white noise.

The Partial Autocorrelation Coefficient

- It is usual to plot the partial autocorrelation function or PACF.



Examining Stationarity of Time Series Data

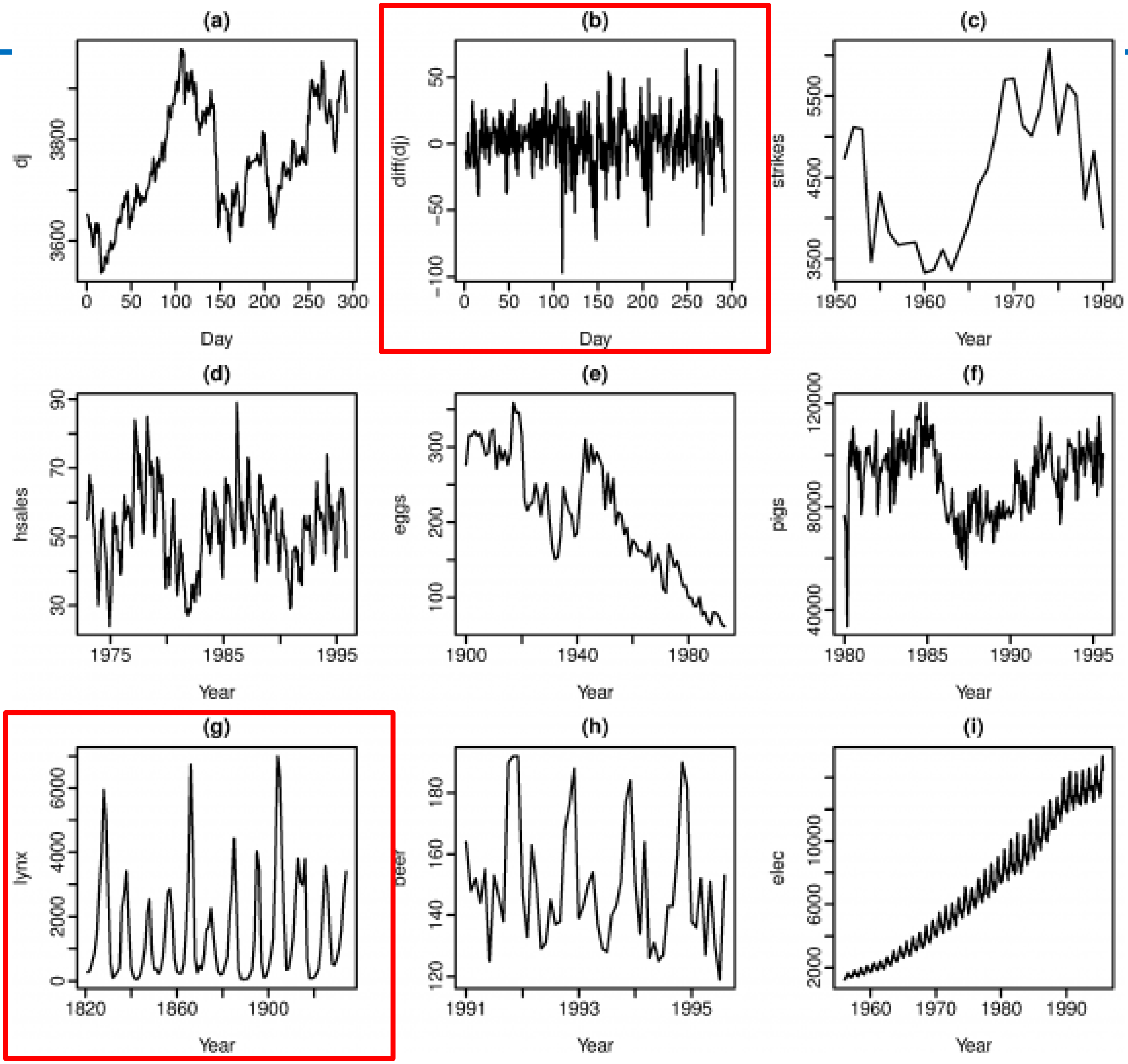
- Stationary time series is one whose properties do not depend on the time at which the series is observed
- Stationarity means no growth or decline
- Time series with trends, with seasonality are not stationary
  - Trend or seasonality will affect the value of the time series at different times
- White noise is stationary
- Sometimes confusing e.g., cyclic behavior since cycles do not have fixed period thus more random-like behavior

Examining Stationarity of Time Series Data

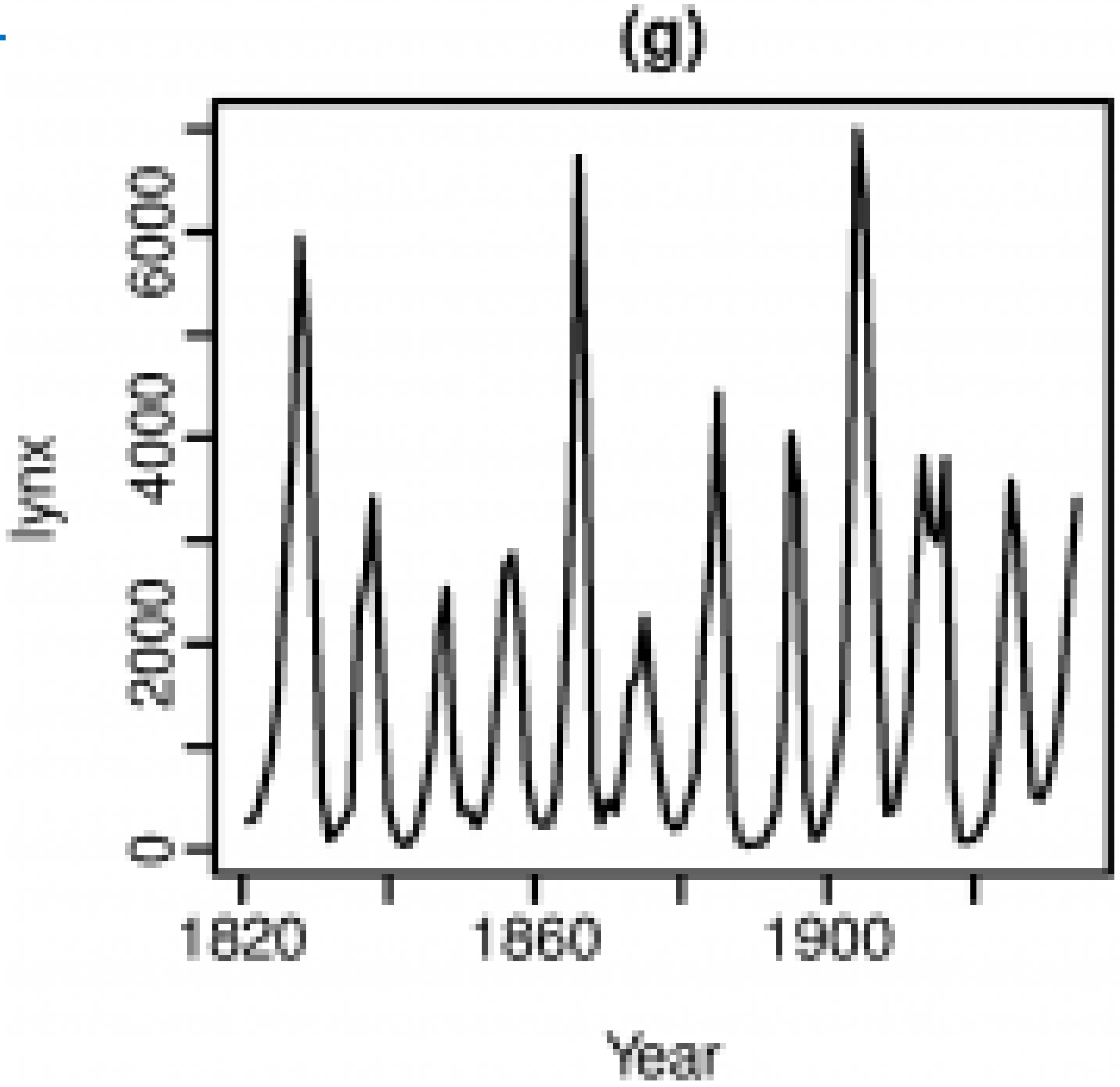
- No predictable patterns
- Data fluctuates around a constant mean independent of time and variance and fluctuation remains constant over time.
- Stationarity can be assessed using a time series plot.
  - Plot shows no change in the mean over time
  - No obvious change in the variance over time.
- Non-stationary data have averages and variances that change over time and must be transformed

Examining Stationarity of Time Series Data

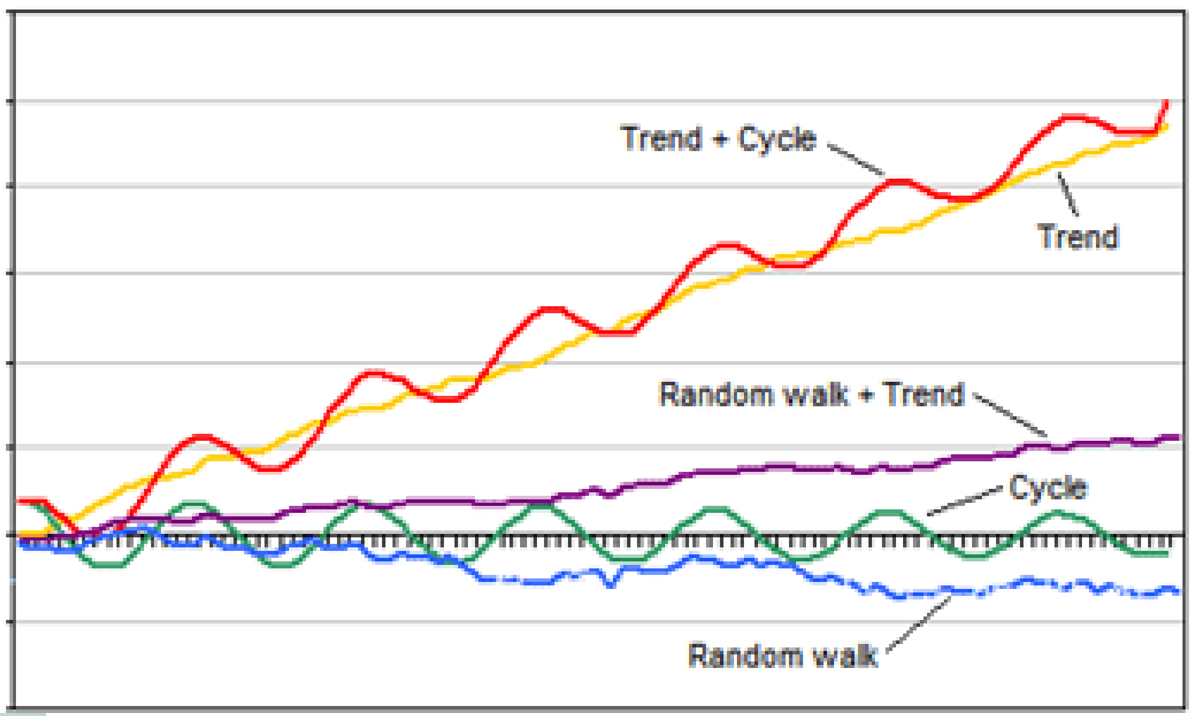
Which of these are stationary?

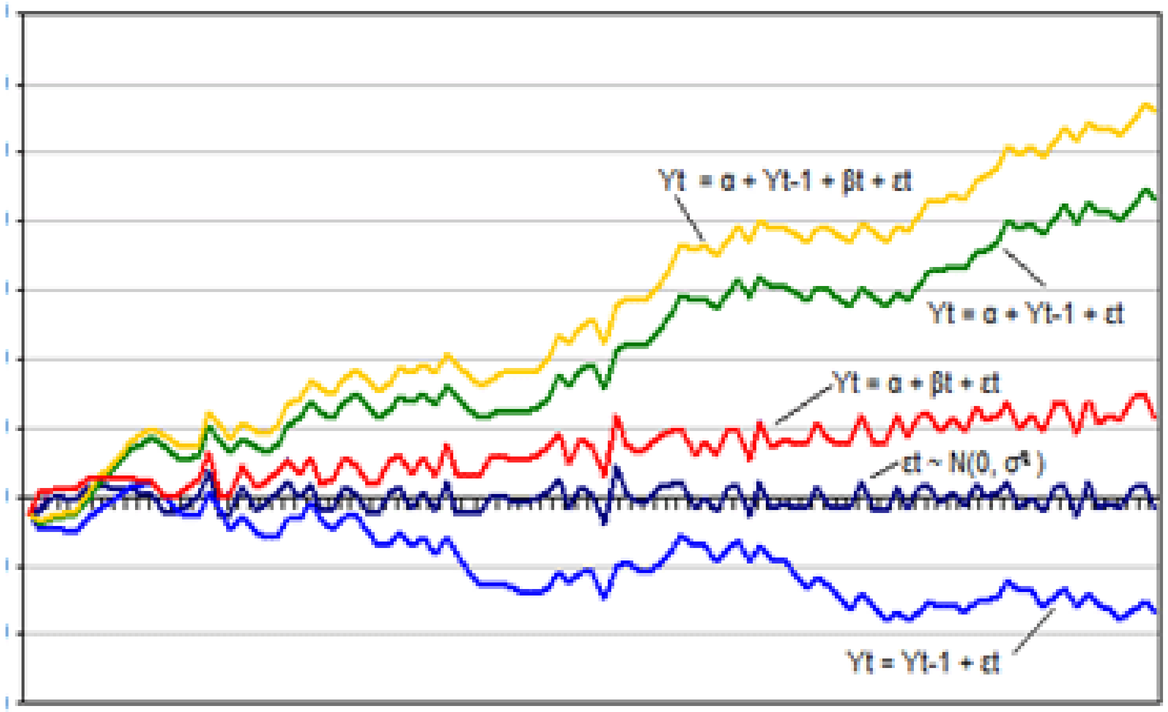


Examining Stationarity of Time Series Data



Stationary versus Non Stationary Plots

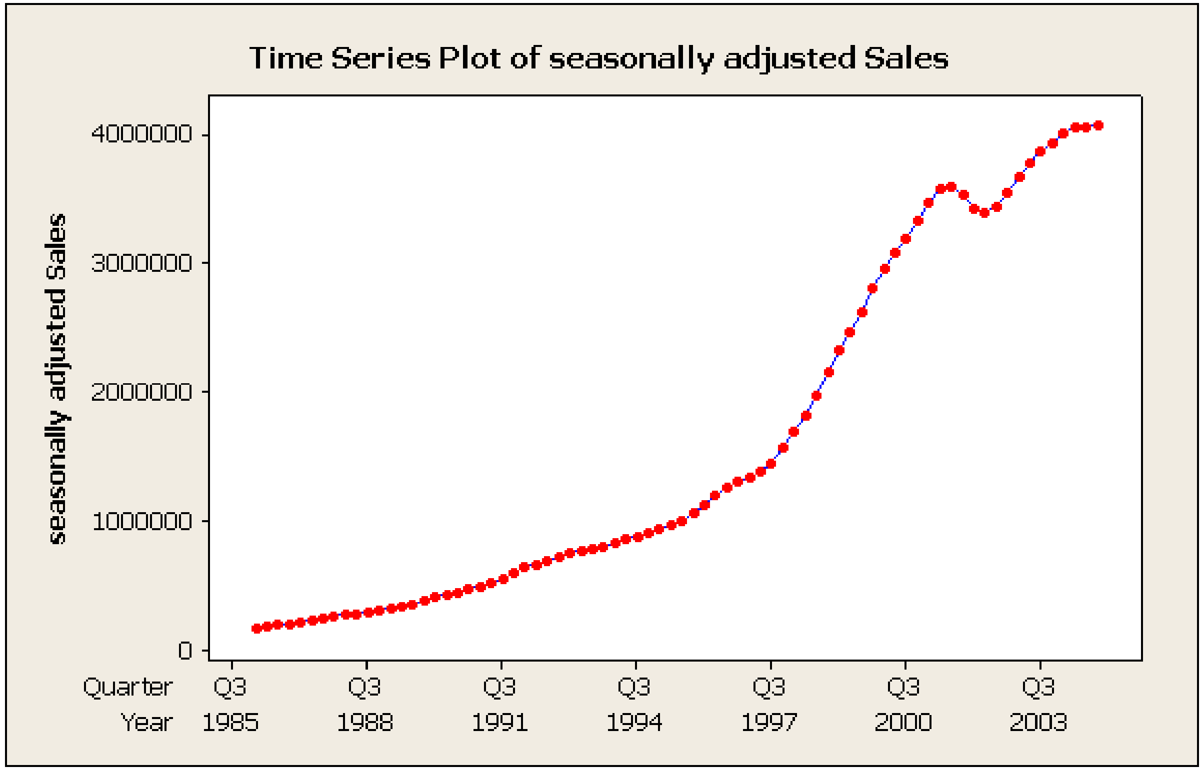




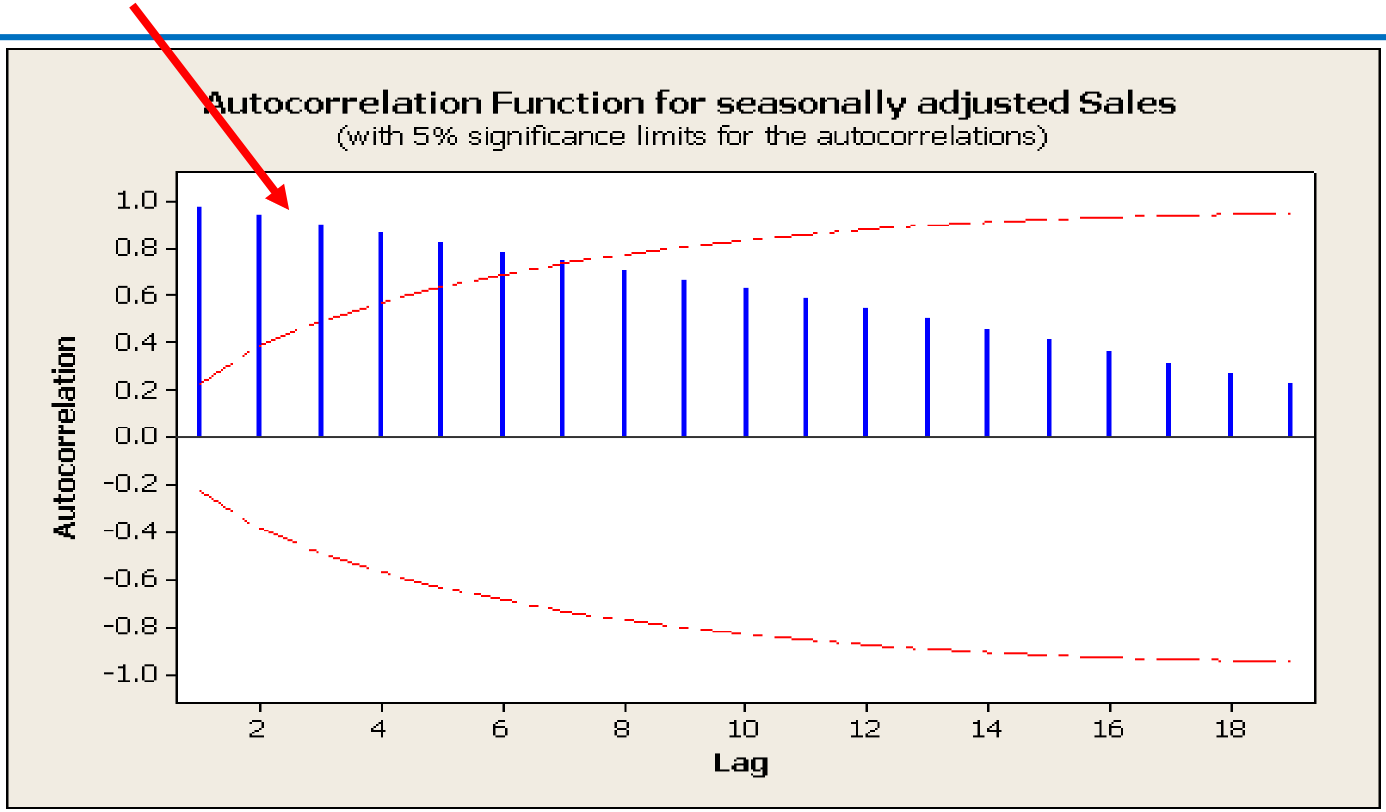
- Source: http://www.investopedia.com/articles/trading/07/stationary.asp

Examining Stationarity of Time Series Data

- The autocorrelation plot can also show non-stationarity.
  – Significant autocorrelation for several time lags and slow decline in r_k indicate non-stationarity.
- The following graph shows the seasonally adjusted sales for Gap stores from 1985 to 2003.

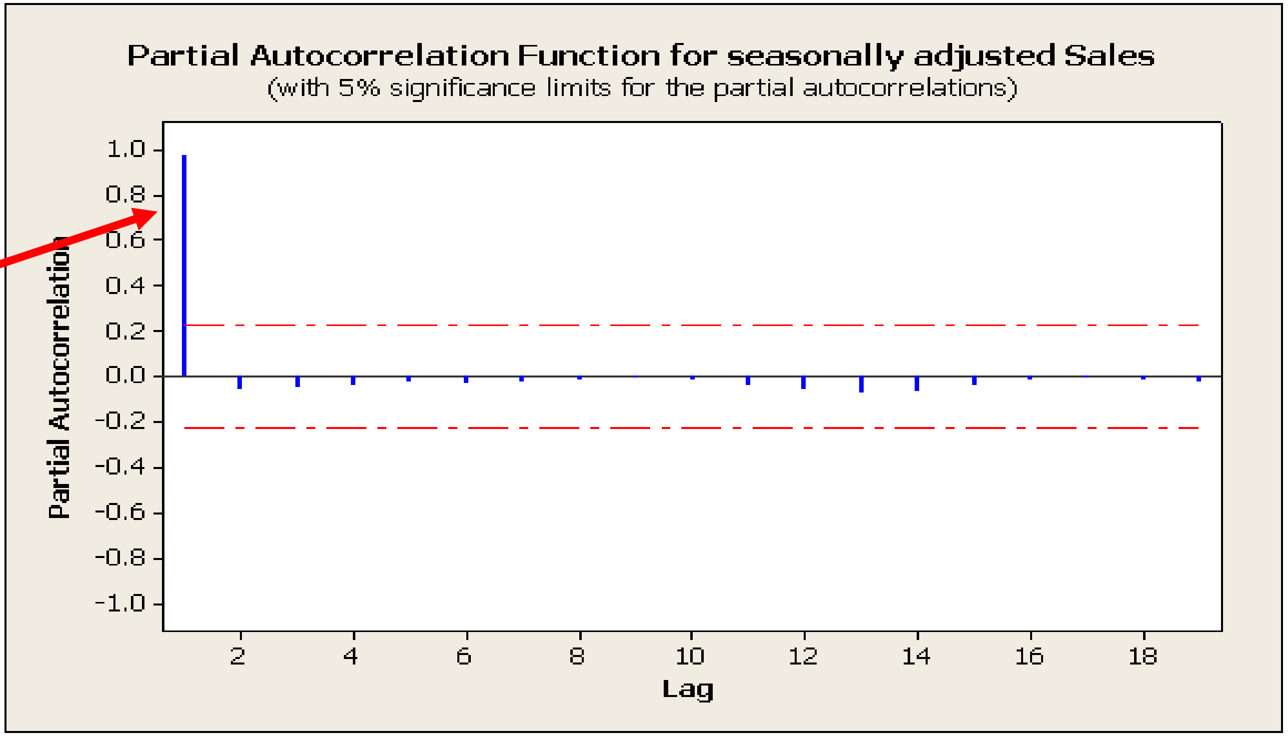


Examining Stationarity of Time Series Data



- 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.

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.

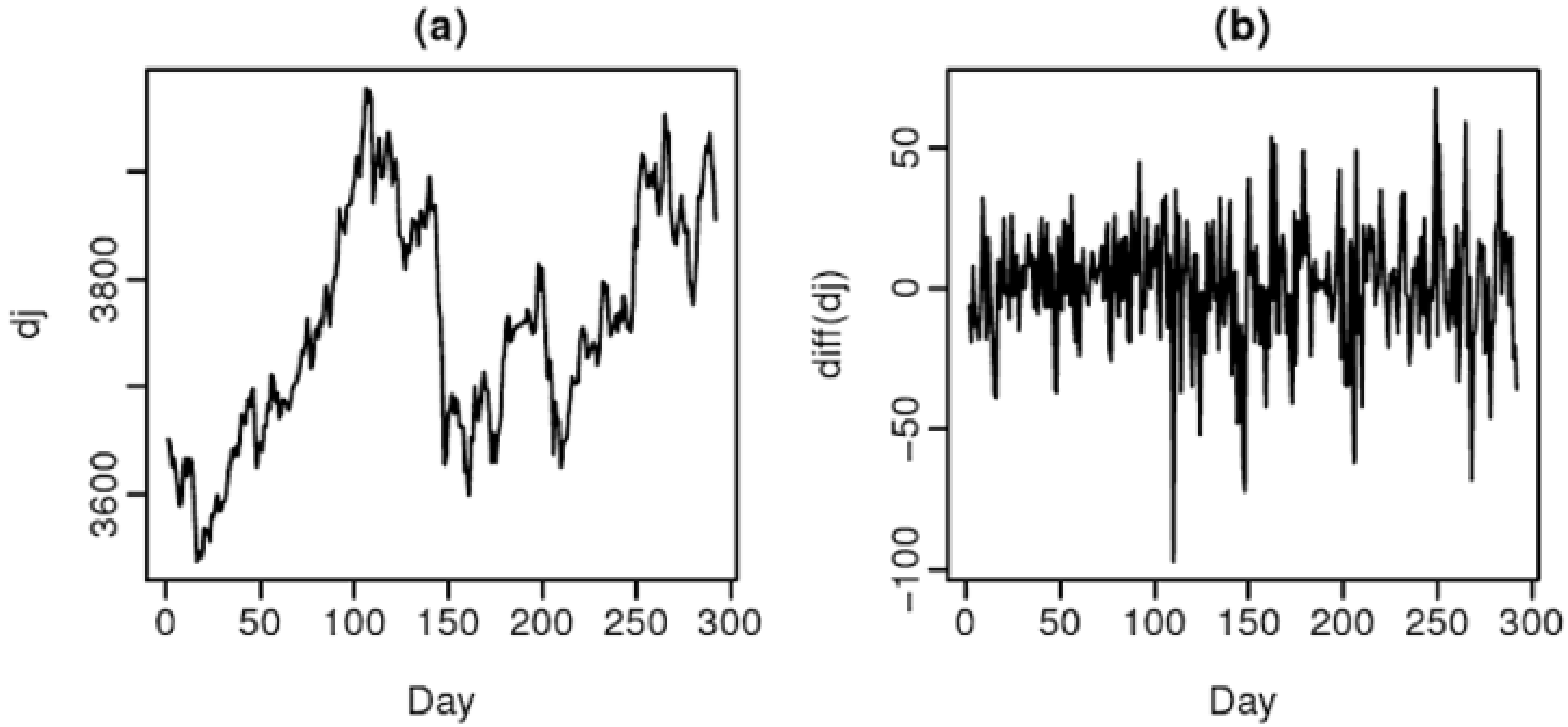
Removing Non-Stationarity in Time Series

- 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.
- Differencing tries to stabilize the mean of a time series by removing changes in the level of a time series and so eliminating trend and seasonality
- The differenced series is defined as:

$$y'_t = y_t - y_{t-1}$$

Differencing

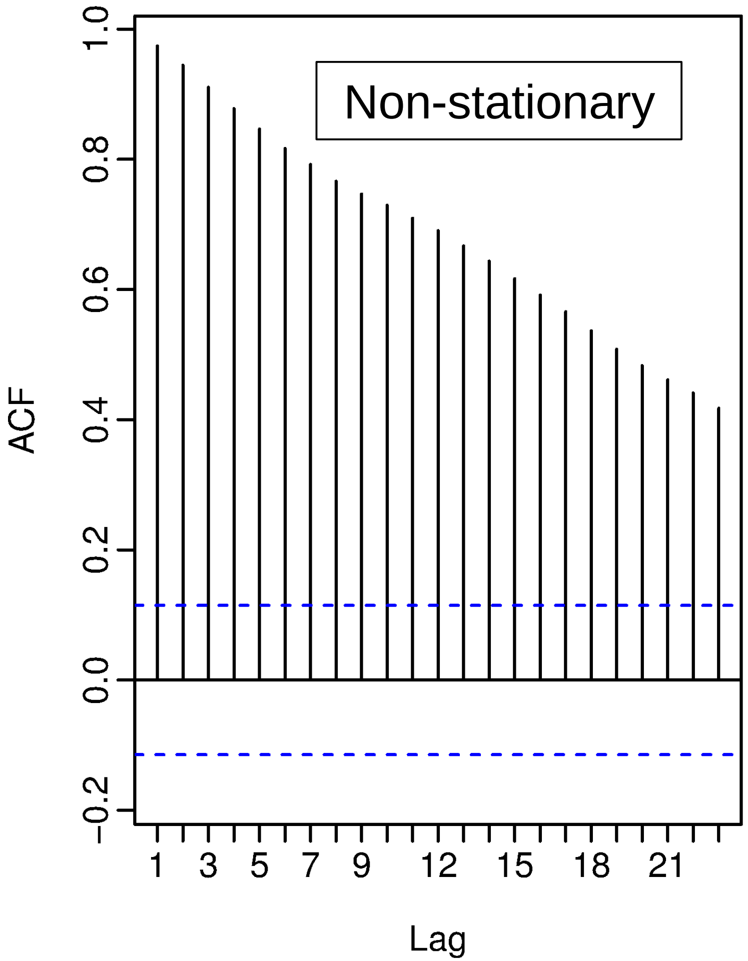
- Observe Dow Jones index in plot (a) and the daily changes in plot (b)

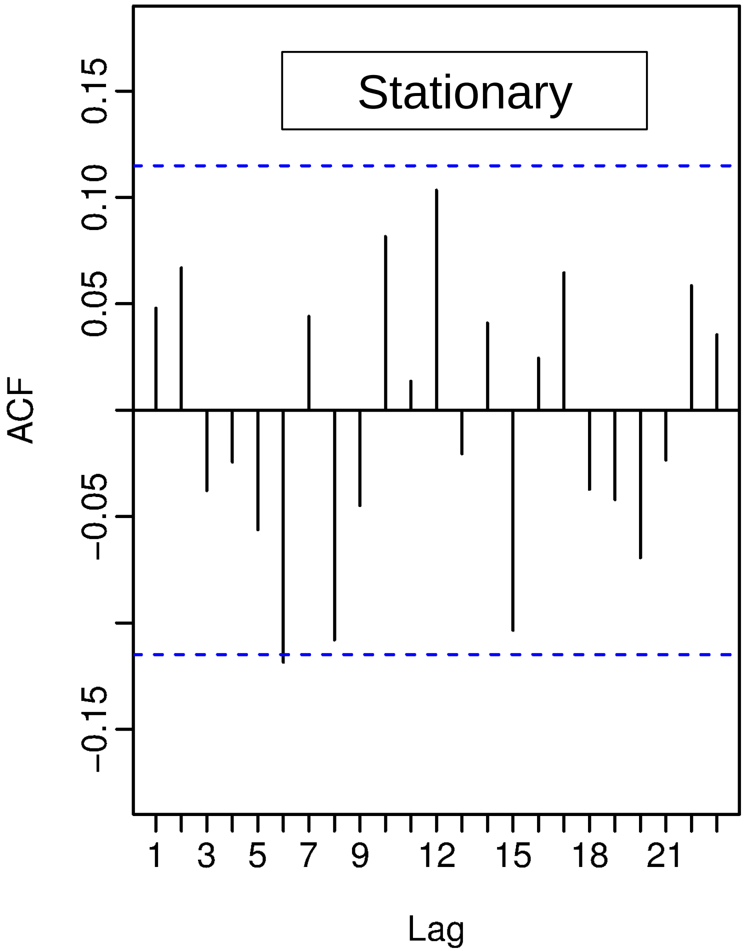


Differencing

- The equivalent ACF of plots (a) and (b)

The ACF can also determine stationarity. For non-stationary time series, ACF decreases slowly.

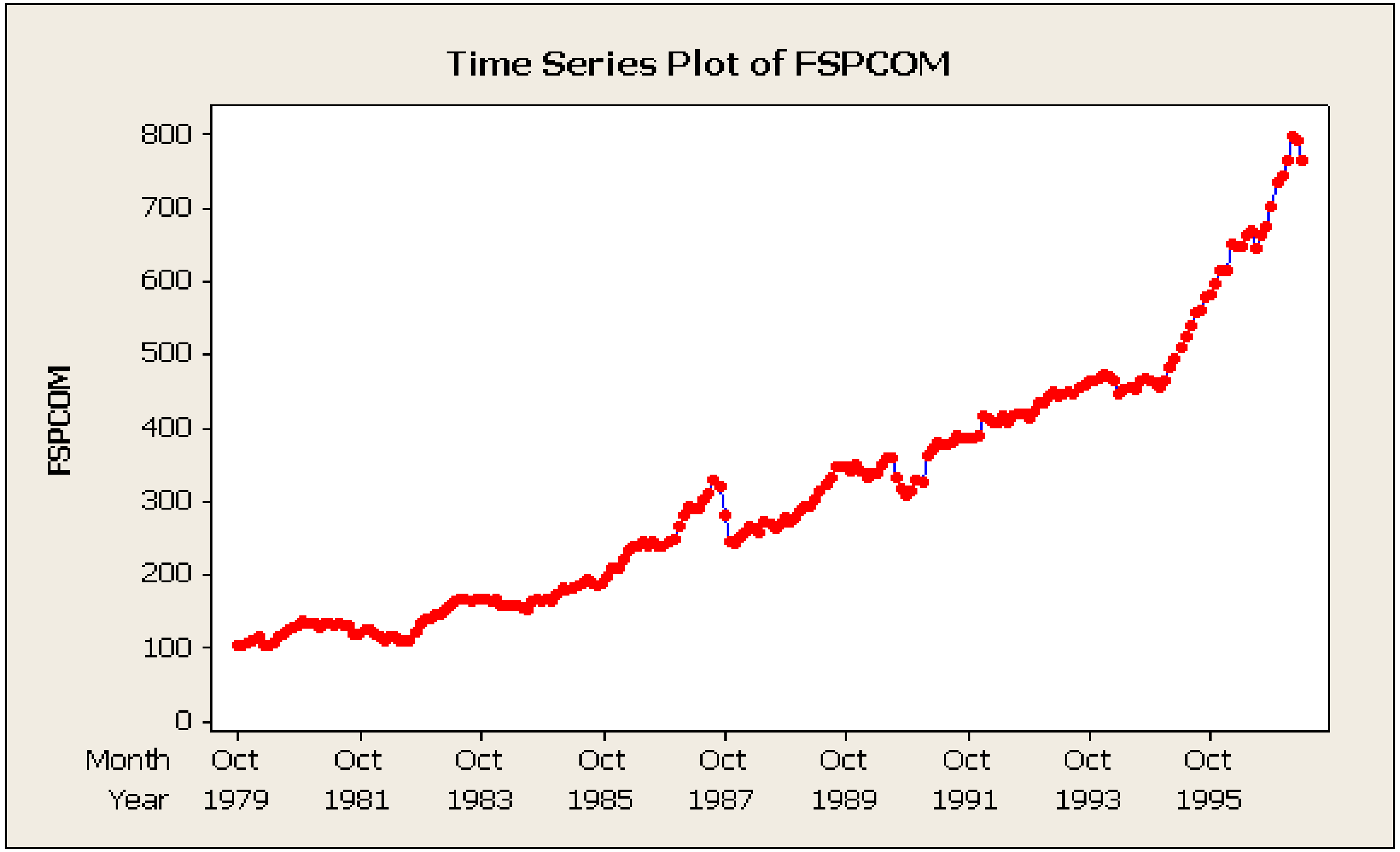




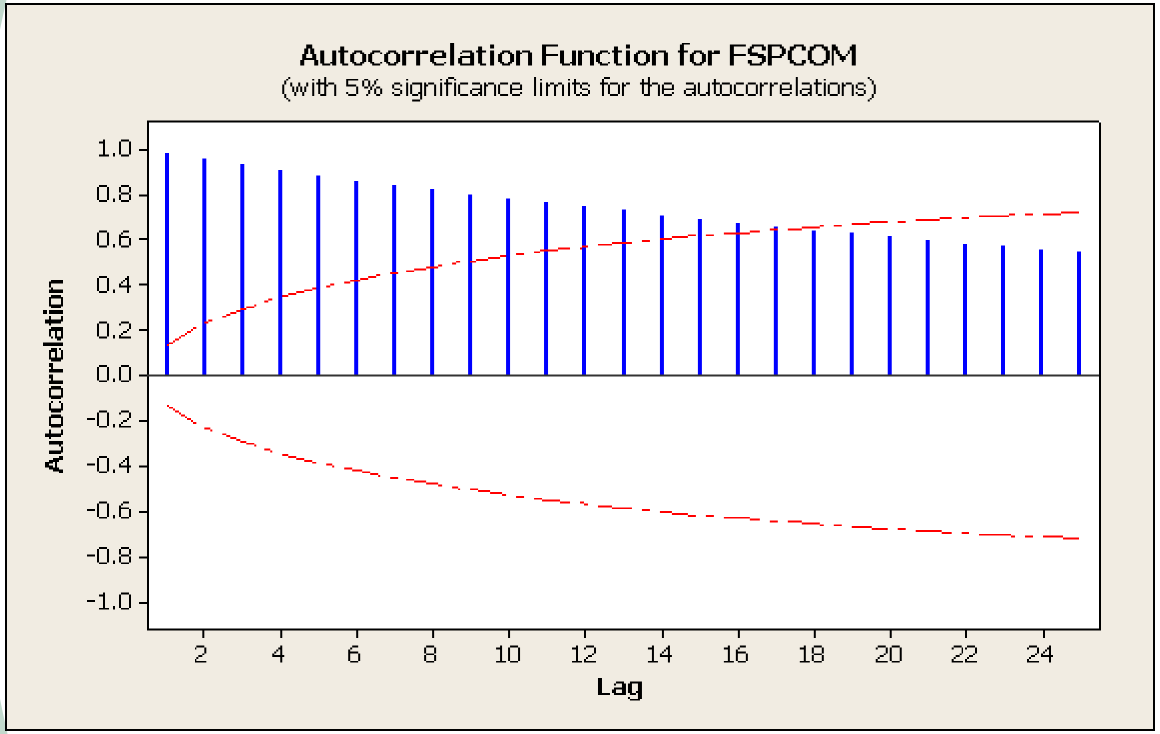
For stationary time series, ACF will drop to zero quickly.

Removing Non-Stationarity in Time Series

The time series plot of the monthly S&P 500 composite index from 1979 to 1997.

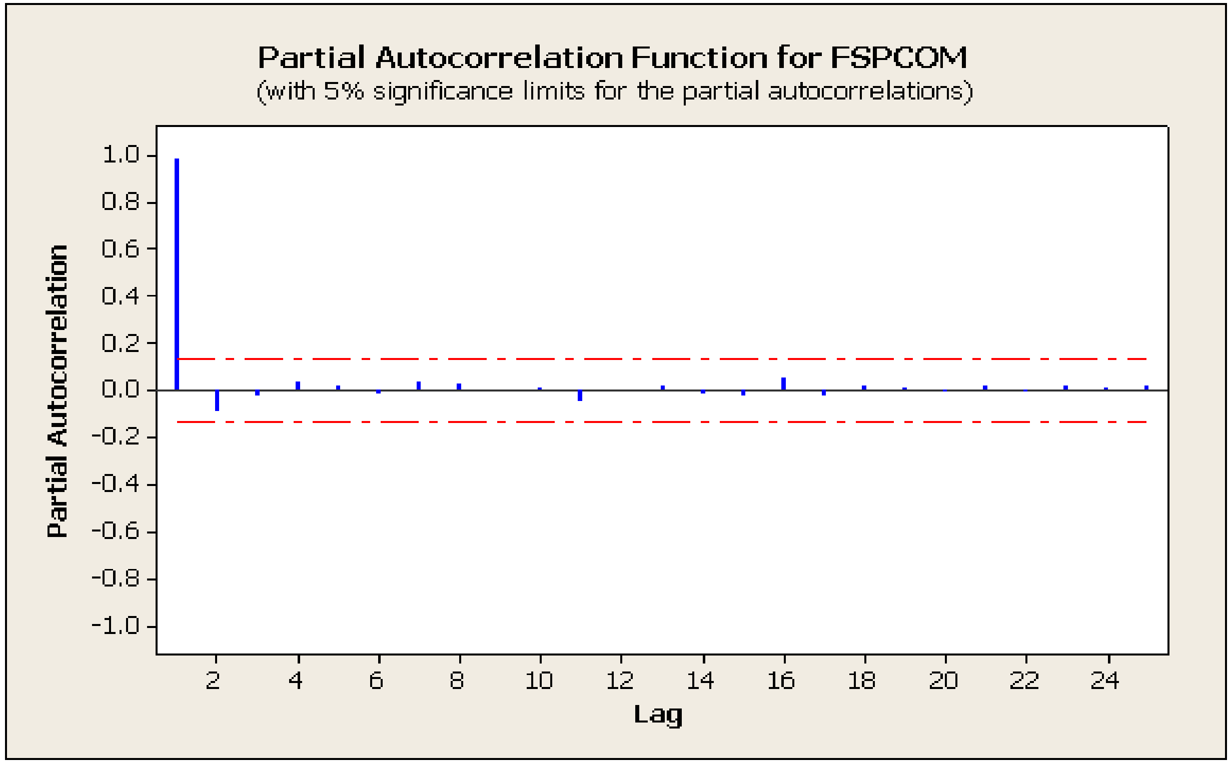


Removing Non-Stationarity in Time Series



Non-stationary

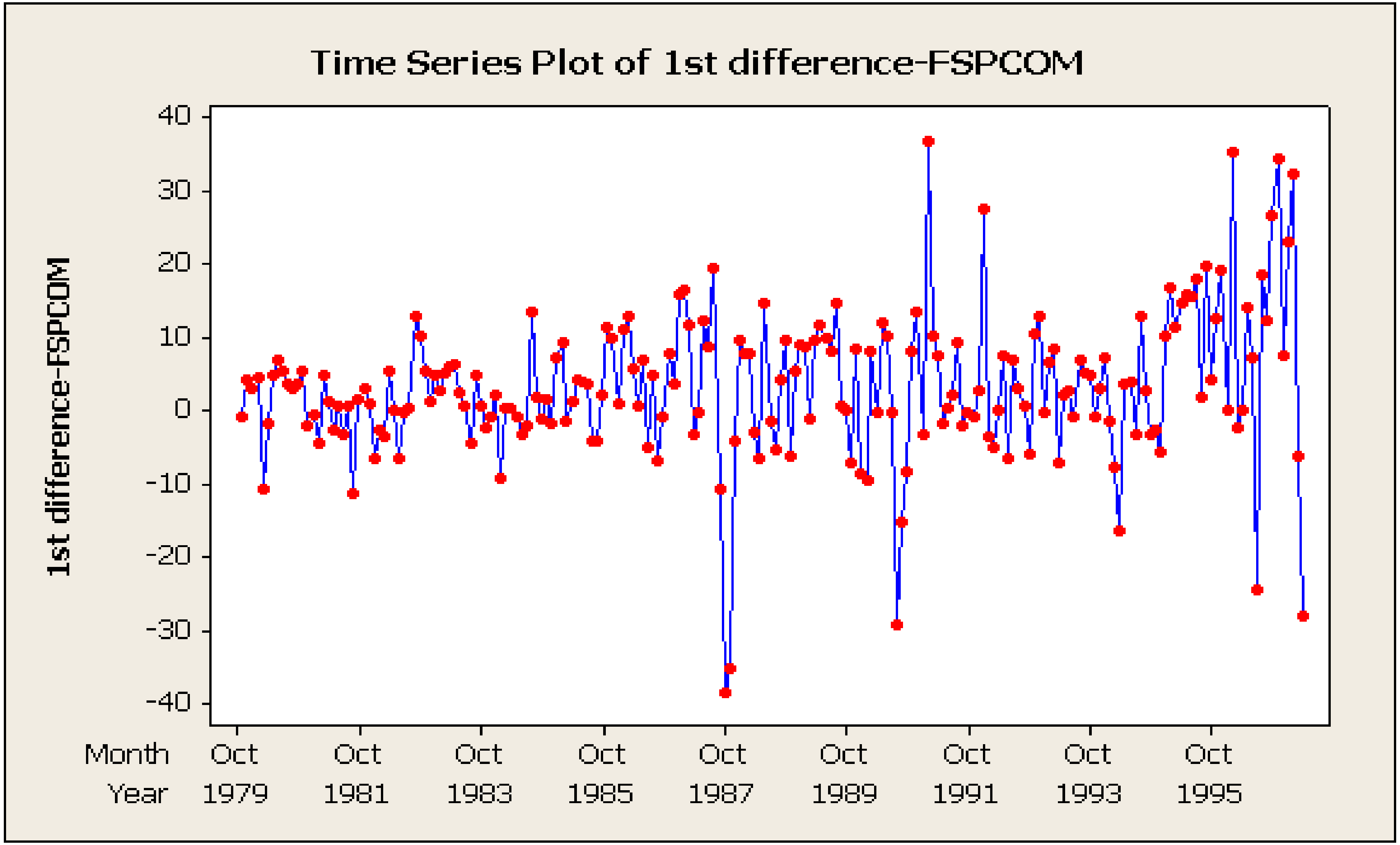
Non-Stationary



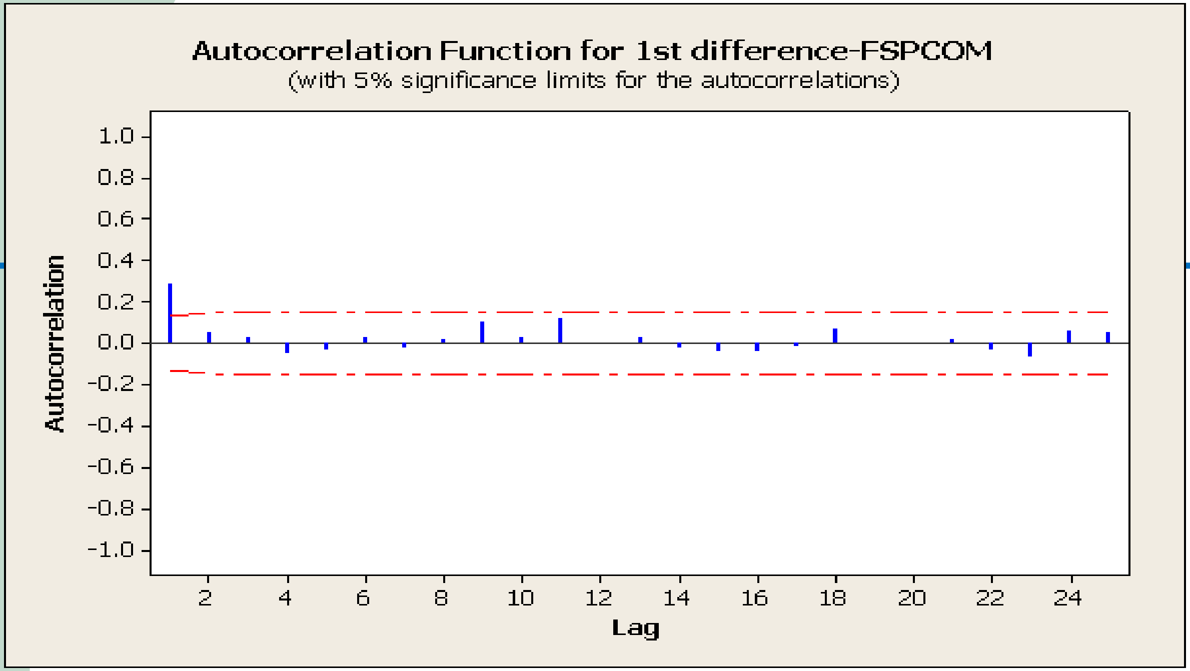
Removing Non-Stationarity in Time Series

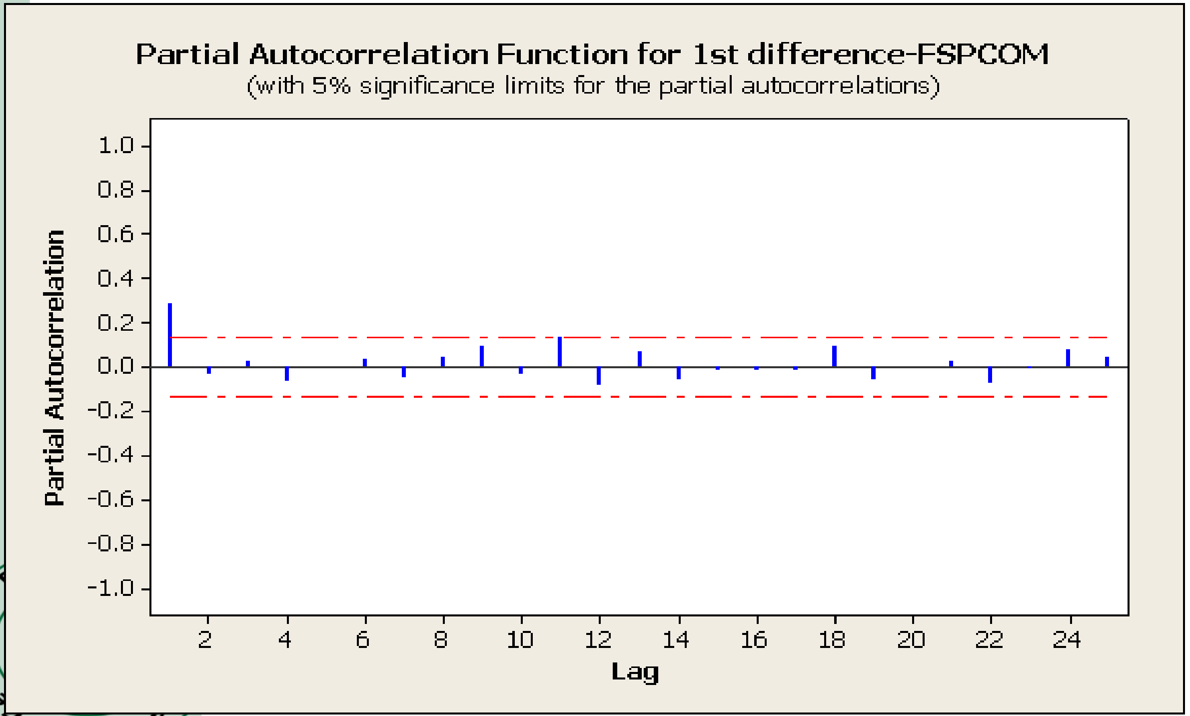
- The time plot shows that it is not stationary in the mean.
- The ACF and PACF plot also display a pattern typical for non-stationary pattern.
- Taking the first difference of the S&P 500 composite index data represents the monthly changes in the S&P 500 composite index.

Removing Non-Stationarity in Time Series



Removing Non-Stationarity in Time Series





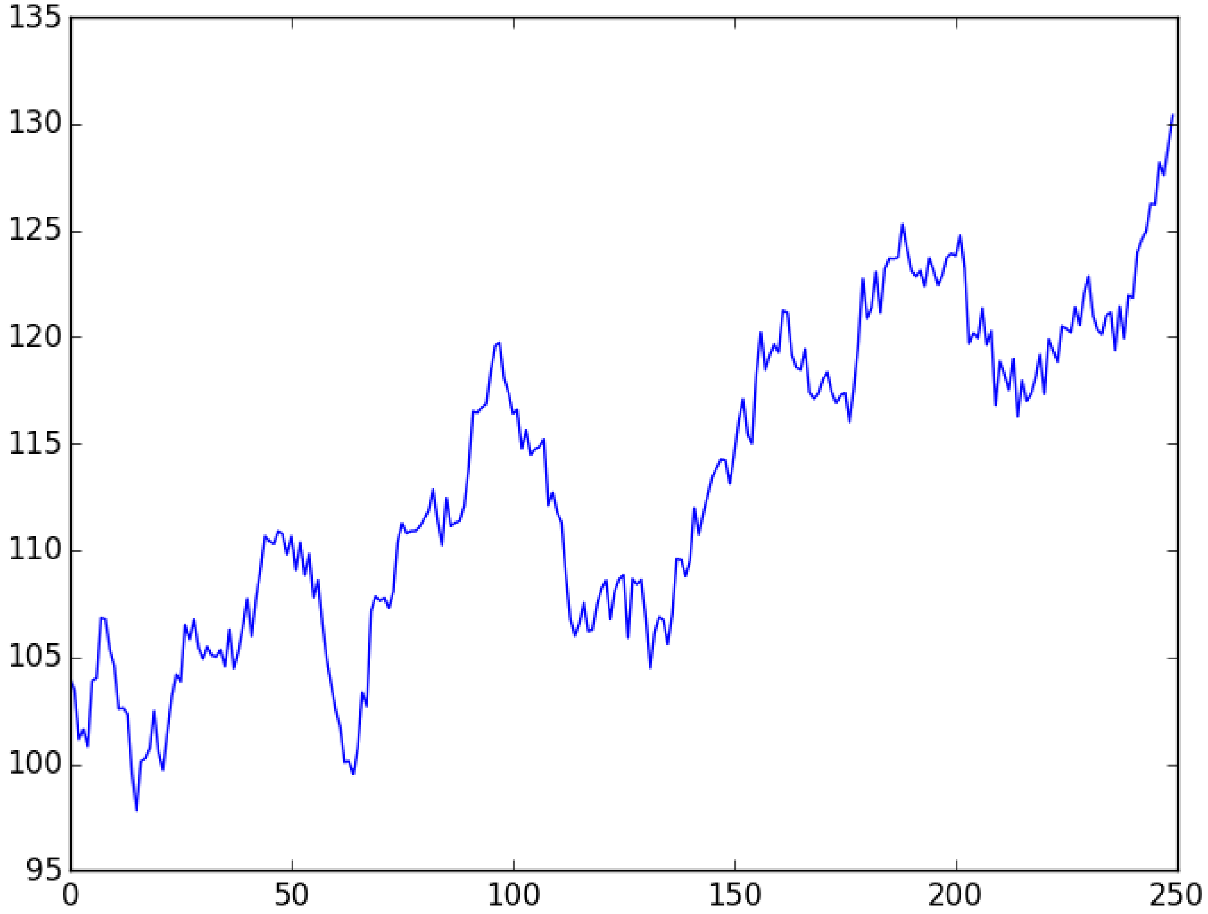
- The time series plot and the ACF and PACF plots indicate that the first difference has removed the growth in the time series data.
- The series looks just like a white noise with almost no autocorrelation or partial autocorrelation outside the 95% limits.

Random Walk

- Let y_t denote the S&P 500 composite index, then the time series plot of differenced S&P 500 composite index suggests that a suitable model for the data might be $y_t - y_{t-1} = e_t$
  - Where e_t is white noise.
- This can be rewritten as: $y_t = y_{t-1} + e_t$
- This model is known as "random walk" model and it is widely used for non-stationary data.

Random Walk

What can you observe?



Random Walk

- Random walks typically have long periods of apparent trends up or down which can suddenly change direction unpredictably
- They are commonly used in analyzing financial, economic and stock price series
- Taking first differencing is a very useful tool for removing non-stationarity

Random Walk

- But sometimes the differenced data will not appear stationary and it may be necessary to difference the data a second time.
- Also, the Random Walk model will only have T-1 values
- Random walks typically have:
  - long periods of apparent trends up or down
  - sudden and unpredictable changes in direction.
- The forecasts from a random walk model are equal to the last observation, as future movements are unpredictable, and are equally likely to be up or down. Thus, the random walk model underpins naïve forecasts.

Second Order Differencing

- Occasionally the differenced data will not appear stationary and it may be necessary to difference the data a second time to obtain a stationary series
- The series of second order difference is defined:

$$y''_t = y'_t - y'_{t-1} = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2}$$

- In practice, it is almost never necessary to go beyond second order differences.
- “change the changes”

Seasonal Differencing

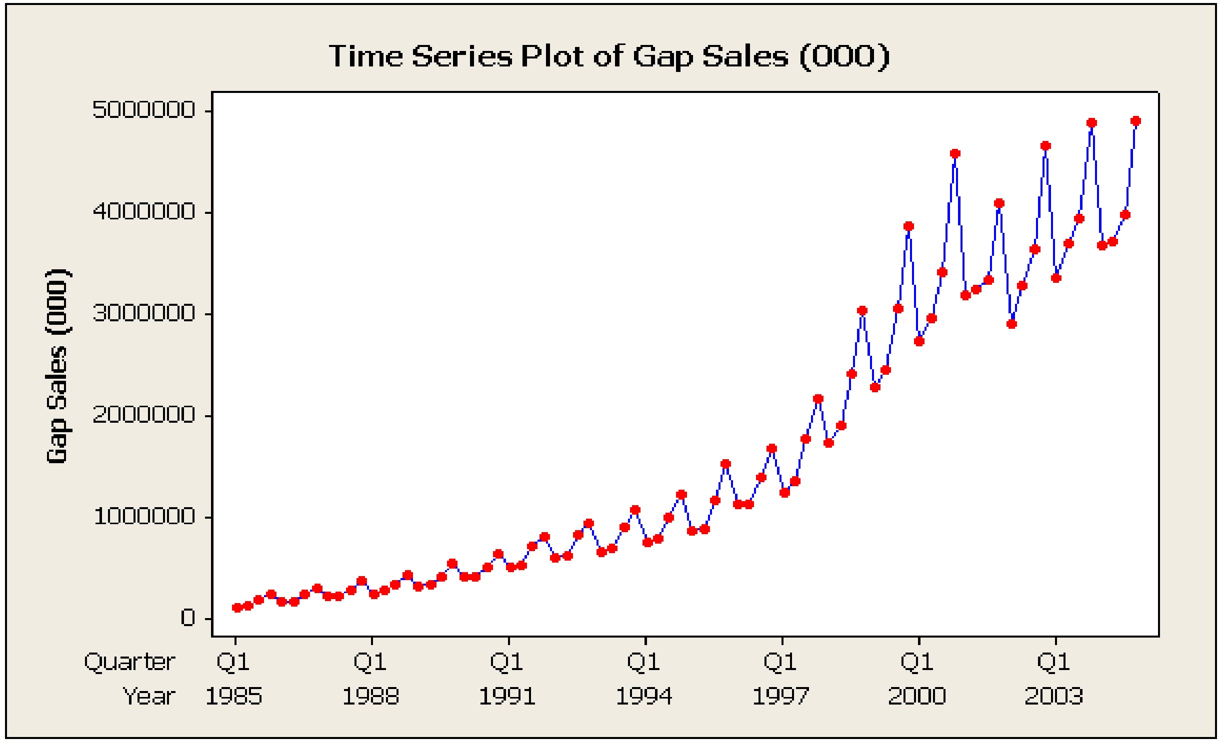
- With seasonal data which is not stationary, it is appropriate to take seasonal differences.
- A seasonal difference is the difference between an observation and the corresponding observation from the previous season (e.g., year).
- We define seasonal differencing of lag s as follows:

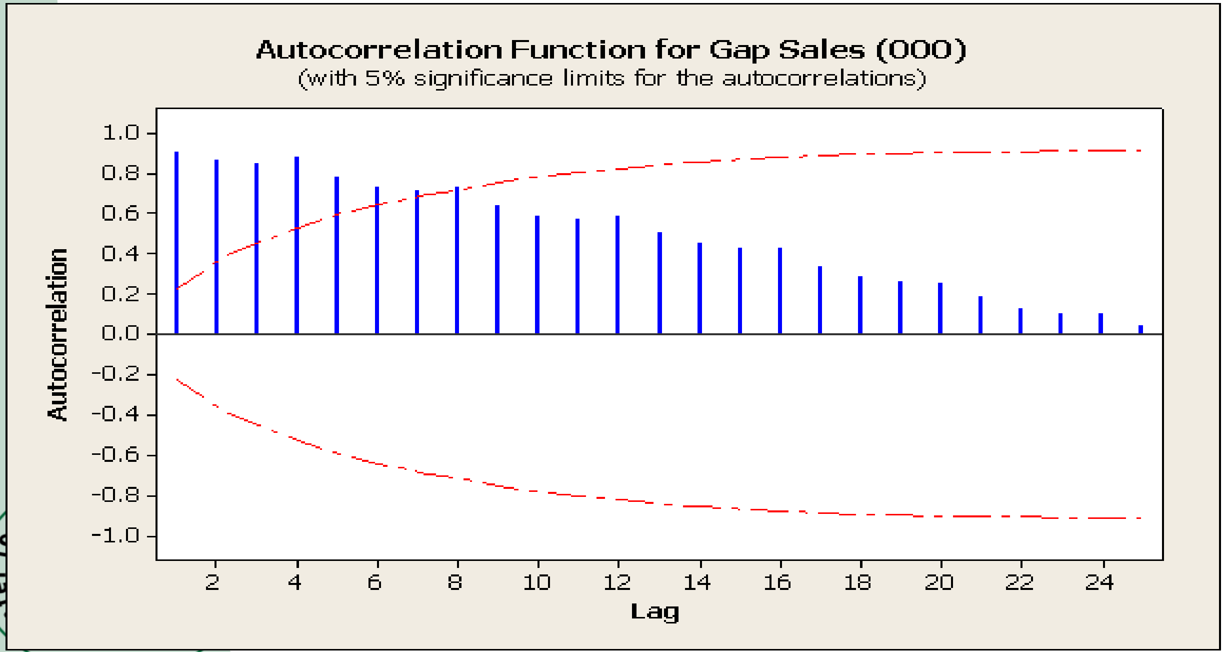
$$y'_t = y_t - y_{t-s}$$

  - Where s is the length of the season
  - Sometimes called "lag-s differences" or "lag-m differences", as we subtract the observation after a lag of s or m periods

Seasonal Differencing Example

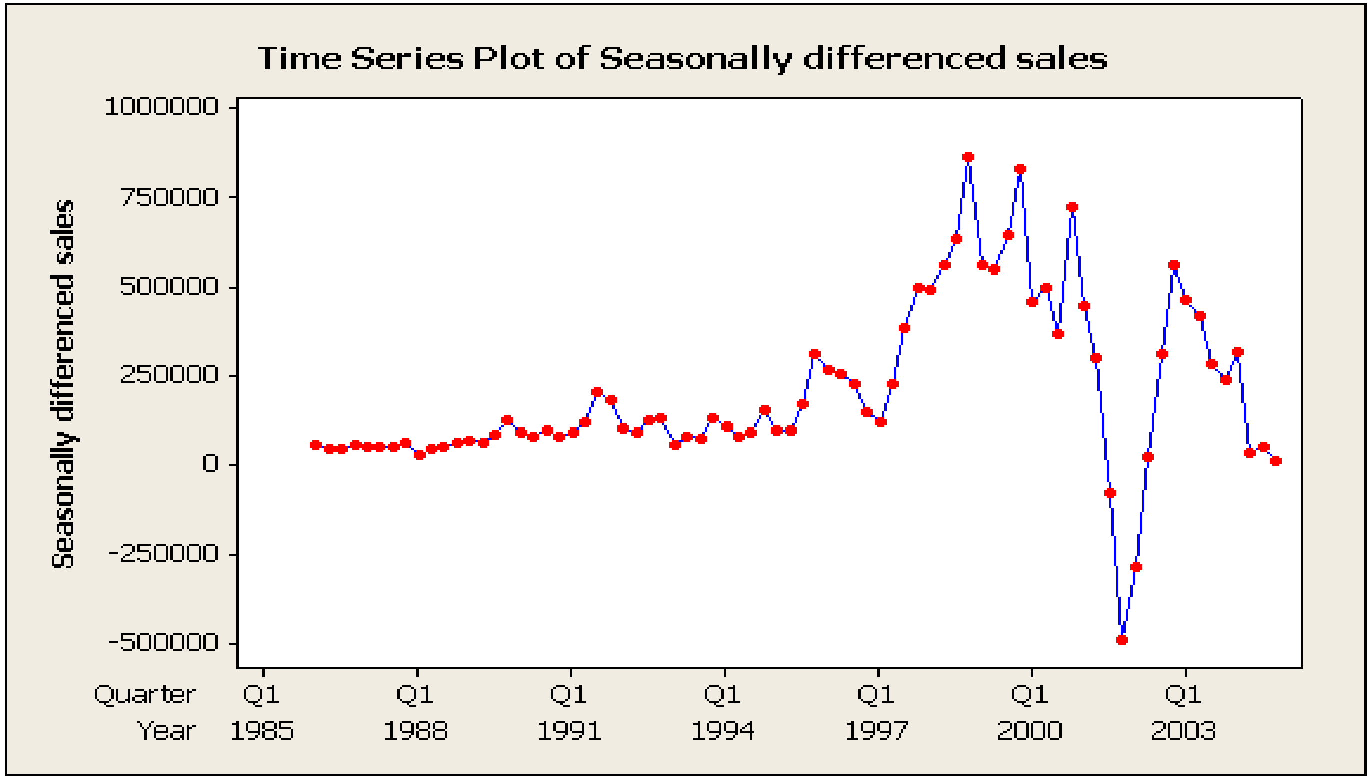
- The Gap quarterly sales is an example of a non-stationary seasonal data.
- The following plot show a trend with a pronounced seasonal component
- The auto correlations show that
  - The series is non-stationary.
  - The series is seasonal.



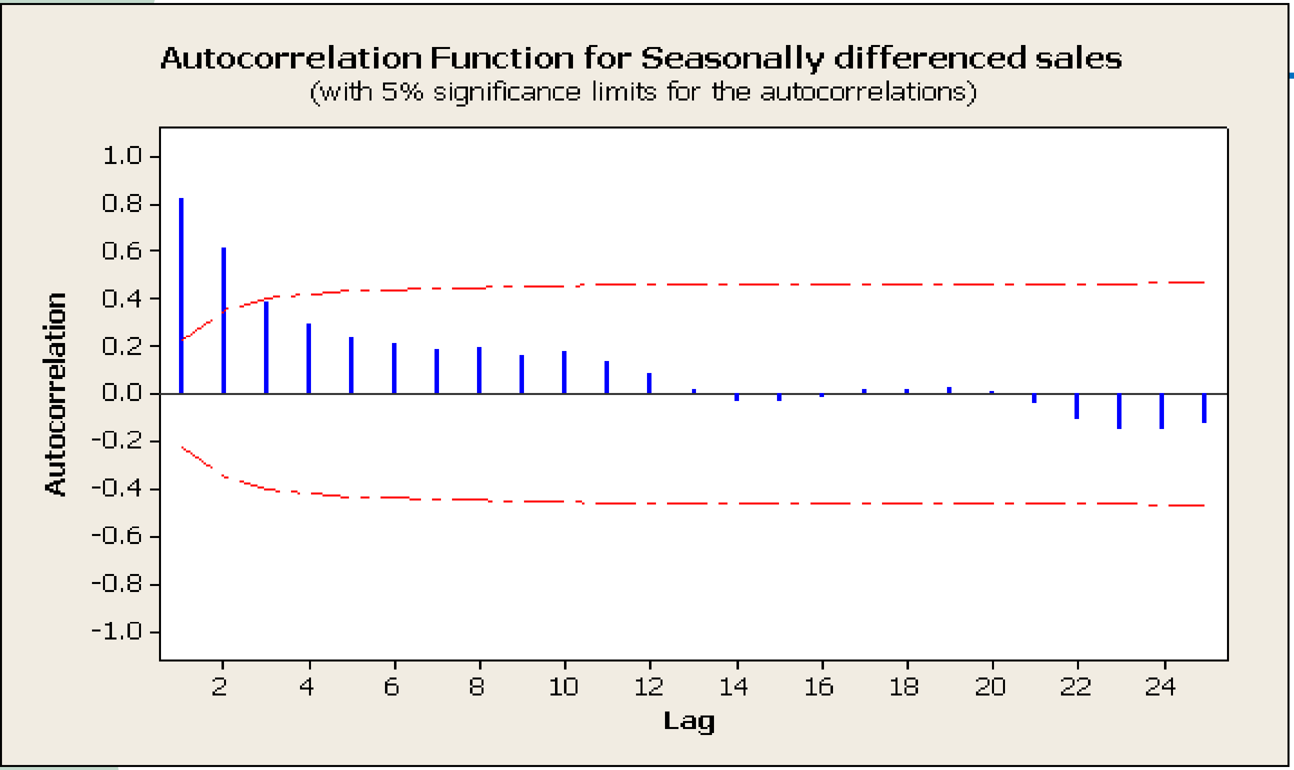


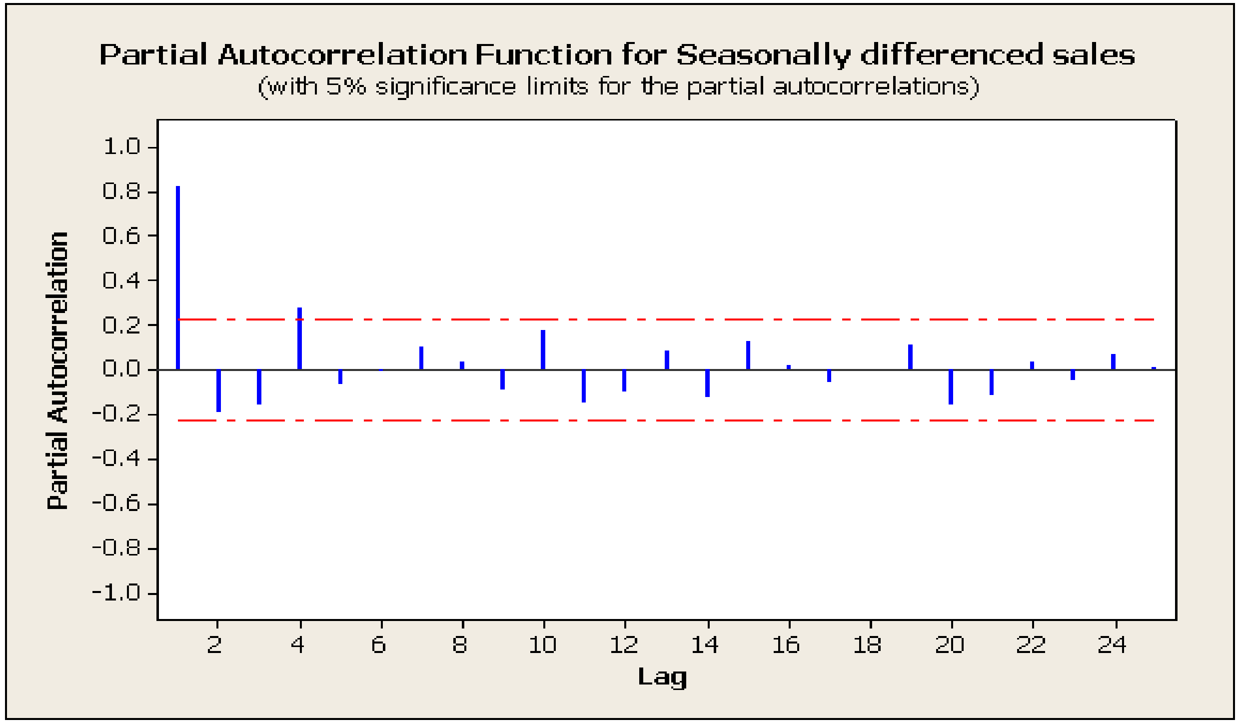
Seasonal Differencing

- The seasonally differenced series represents the change in sales between quarters of consecutive years.
- The time series plot, ACF and PACF of the seasonally differenced Gap's quarterly sales are as follows:



Seasonal Differencing





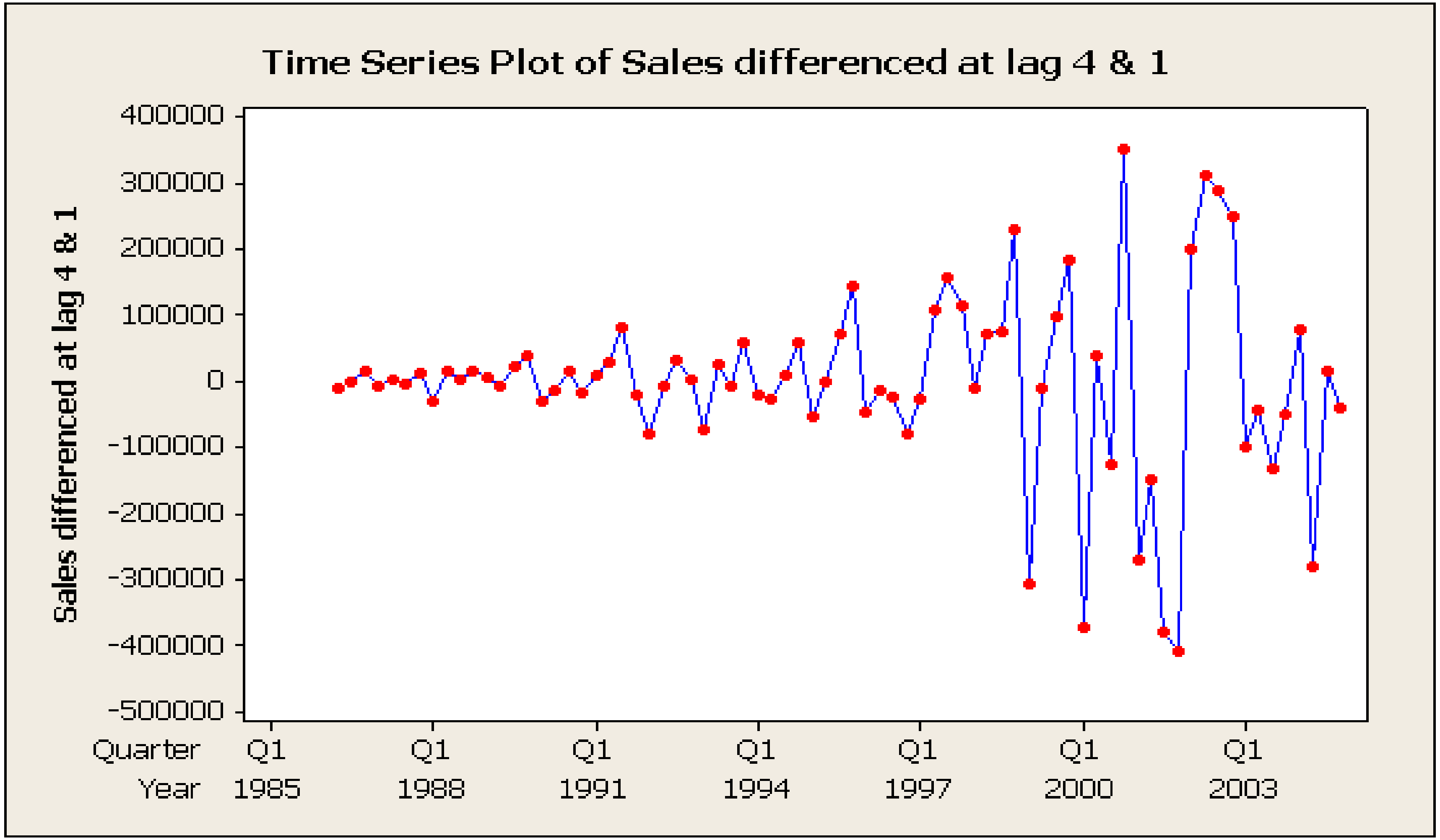
Seasonal Differencing

- The series is now much closer to being stationary, but more than 5% of the spikes are beyond 95% critical limits and autocorrelation show gradual decline in values.
- The seasonality is still present as shown by spikes at time lag 1 and 4 in the PACF.
- The remaining non-stationarity in the mean can be removed with a further first difference.

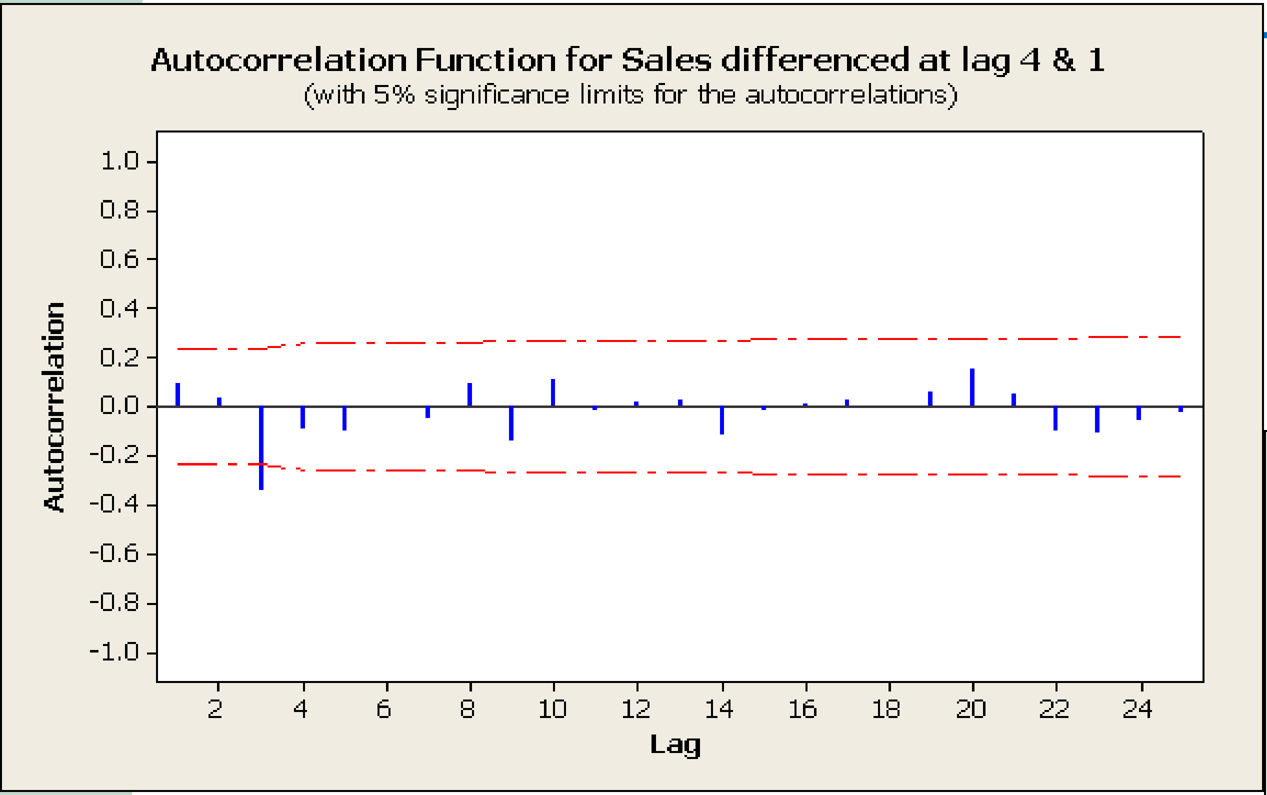
Seasonal Differencing

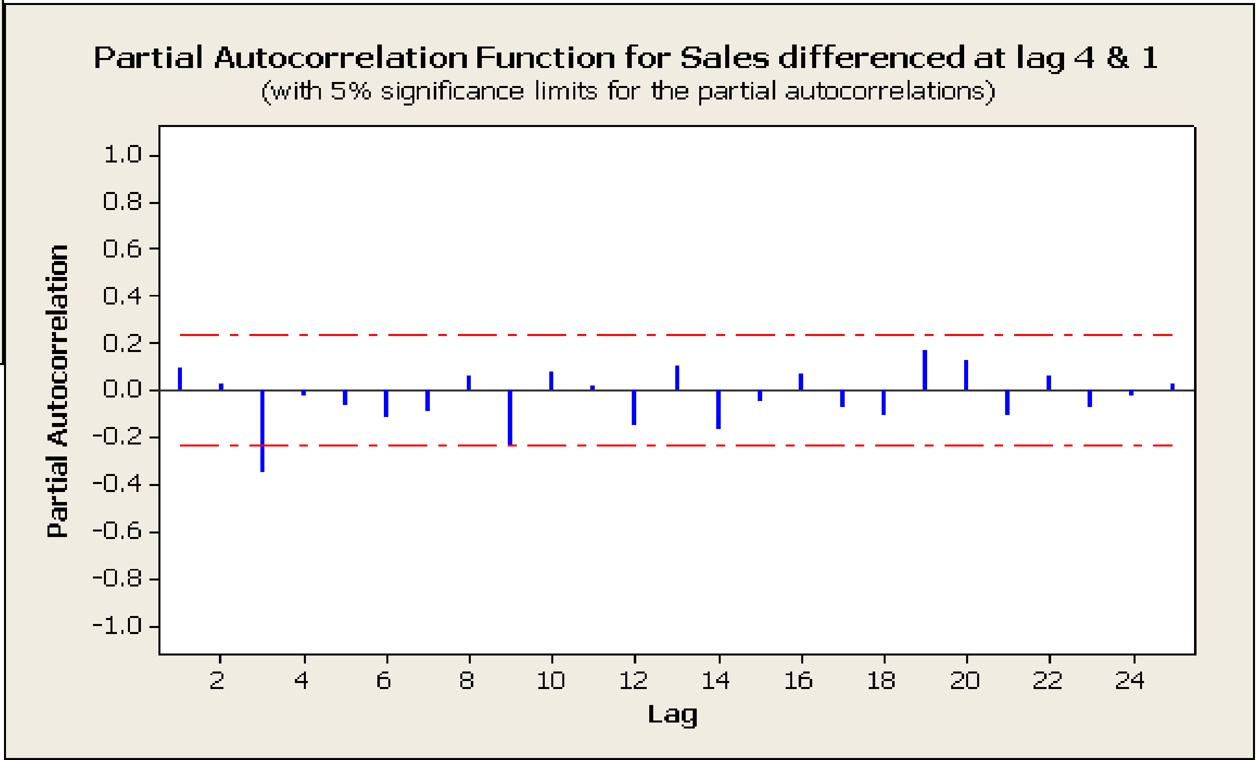
- It is recommended to do the seasonal differencing first since sometimes the resulting series will be stationary and hence no need for a further first difference.
- When differencing is used, it is important that the differences be interpretable.
- The series resulted from first difference of seasonally differenced Gap’s quarterly sales data is reported in the following two slides.
- Is the resulting series white noise?

Seasonal Differencing



Seasonal Differencing





ARIMA Models For Time Series Data

- Autoregression
  - Consider regression models of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

  - Define

$$x_1 = y_{t-1}$$
$$x_2 = y_{t-2}$$
$$\vdots$$
$$x_p = y_{t-p}$$

ARIMA Models For Time Series Data

- Then the previous equation becomes

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + \varepsilon_t$$

- The explanatory variables in this equations are time-lagged values of the variable y.
- Autoregression (AR) is used to describe models of this form.

ARIMA Models For Time Series Data

- Autoregression models should be treated differently from ordinary regression models since:
  - The explanatory variables in the autoregression models have a built-in dependence relationship.
  - Determining the number of past values of y_t to include in the model is not always straight forward.

ARIMA Models For Time Series Data

- Moving average model
  - A time series model which uses past errors as explanatory variable:

$$y_t = \beta_0 + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \cdots \beta_p e_{t-q} + \varepsilon_t$$

    is called moving average(MA) model
  - Note that this model is defined as a moving average of the error series, while the moving average models we discussed previously are the moving average of the observations.

ARIMA Models For Time Series Data

- Autoregressive (AR) models can be coupled with moving average (MA) models to form a general and useful class of time series models called *Autoregressive Moving Average (ARMA)* models.
- These can be used when the data are stationary.

ARIMA Models For Time Series Data

- This class of models can be extended to non-stationary series by allowing the differencing of the data series.
- These are called *Autoregressive Integrated Moving Average(ARIMA)* models.
- There are a large variety of ARIMA models.

ARIMA Models For Time Series Data

- The general non-seasonal model is known as ARIMA(p, d, q):
  - p is the number of autoregressive terms.
  - d is the number of differences.
  - q is the number of moving average terms.

ARIMA Models For Time Series Data

- A white noise model is classified as ARIMA(0, 0, 0)
  - No AR part since y_t does not depend on y_{t-1}.
  - There is no differencing involved.
  - No MA part since y_t does not depend on e_{t-1}.

ARIMA Models For Time Series Data

- A random walk model is classified as ARIMA (0, 1, 0)
  - There is no AR part.
  - There is one difference.
  - There is no MA part.

ARIMA Models For Time Series Data

- Note that if any of p, d, or q are equal to zero, the model can be written in a shorthand notation by dropping the unused part.
- Example
  - ARIMA(2, 0, 0) = AR(2)
  - ARIMA (1, 0, 1) = ARMA(1, 1)

An Autoregressive Model Of Order One AR(1)

- The basic form of an ARIMA (1, 0, 0) or AR(1) is:

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

  - Observation y_t depends on y_{t-1}.
  - The value of autoregressive coefficient ϕ_1 is between –1 and 1.

ARIMA Models For Time Series Data

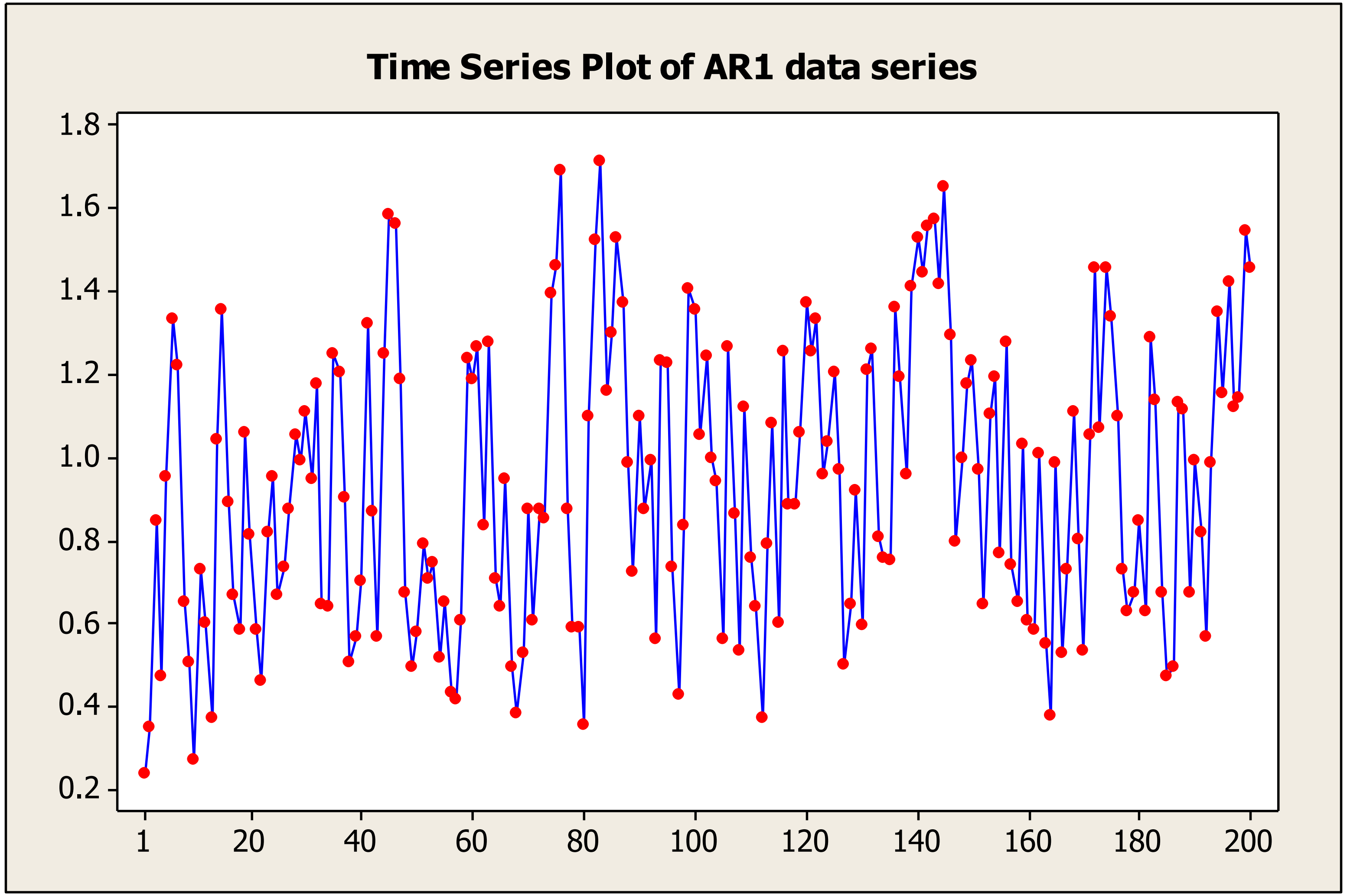
- The AR(p) component is referring to the use of past values in the regression equation for the series Y. The parameter specifies the number of lags used in the model
- For example, AR(2), ARIMA(2,0,0) is represented

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$$

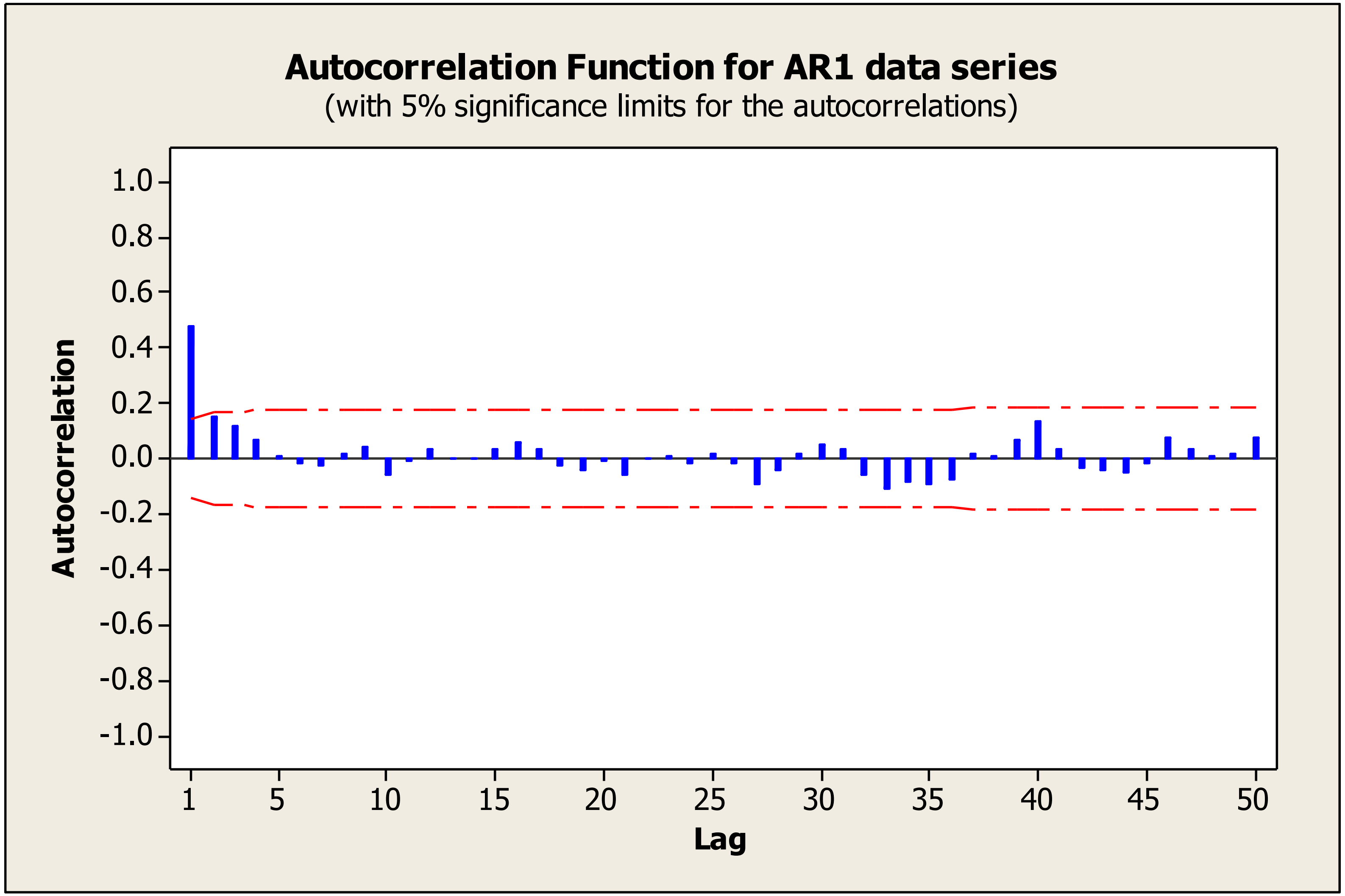
An Autoregressive Model of Order One

- The time plot of an AR(1) model varies with the parameter ϕ_1.
  - When ϕ_1 = 0, y_t is equivalent to a white noise series.
  - When ϕ_1 = 1, y_t is equivalent to a random walk series
  - For negative values of ϕ_1, the series tends to oscillate between positive and negative values.
- The following slides show the time series, ACF and PACF plot for an ARIMA(1, 0, 0) time series data.

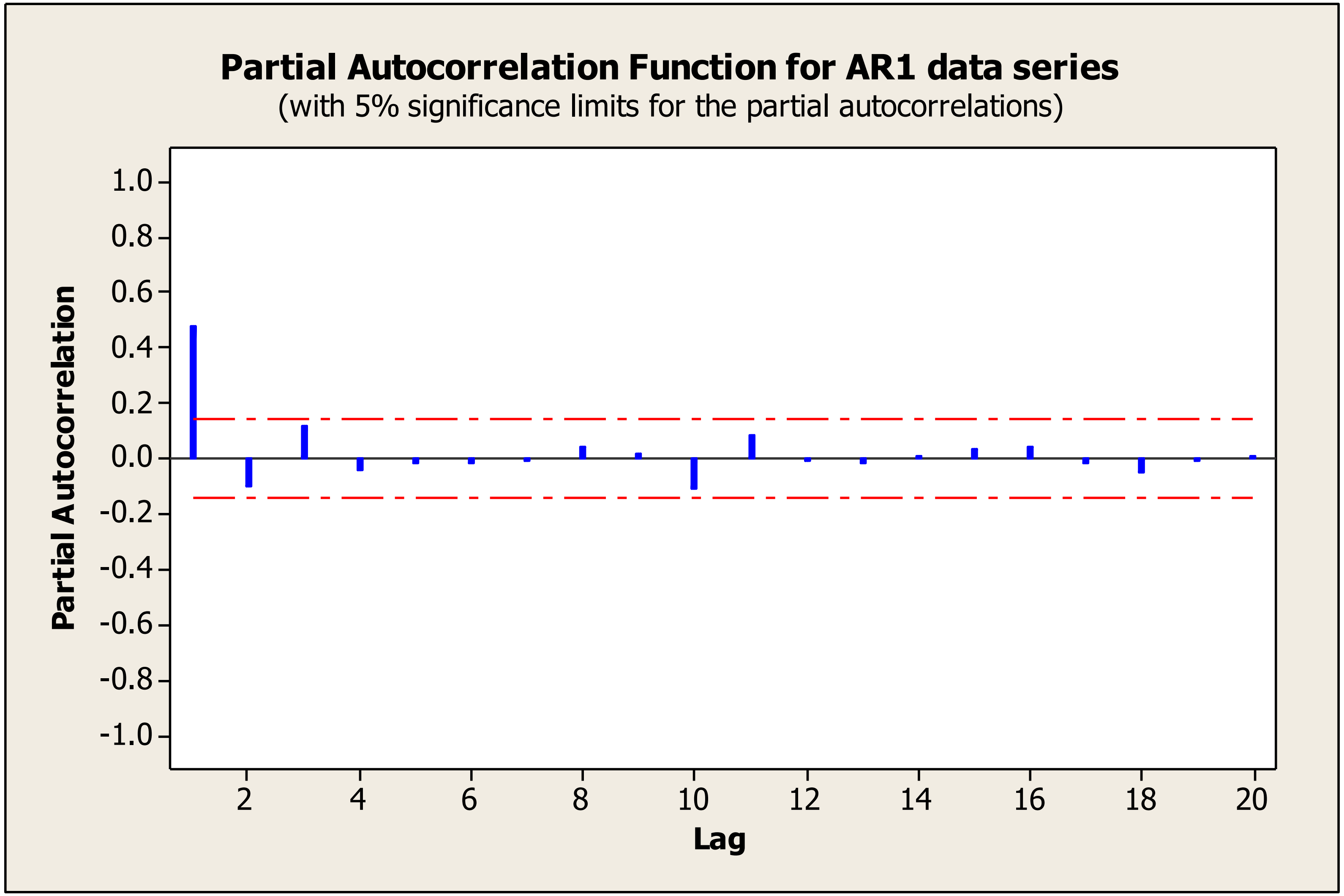
An Autoregressive Model of Order One



An Autoregressive Model of Order One



An Autoregressive Model of Order One



An Autoregressive Model of Order One

- The ACF and PACF can be used to identify an AR(1) model.
  - The autocorrelations decay exponentially.
  - There is a single significant partial autocorrelation.

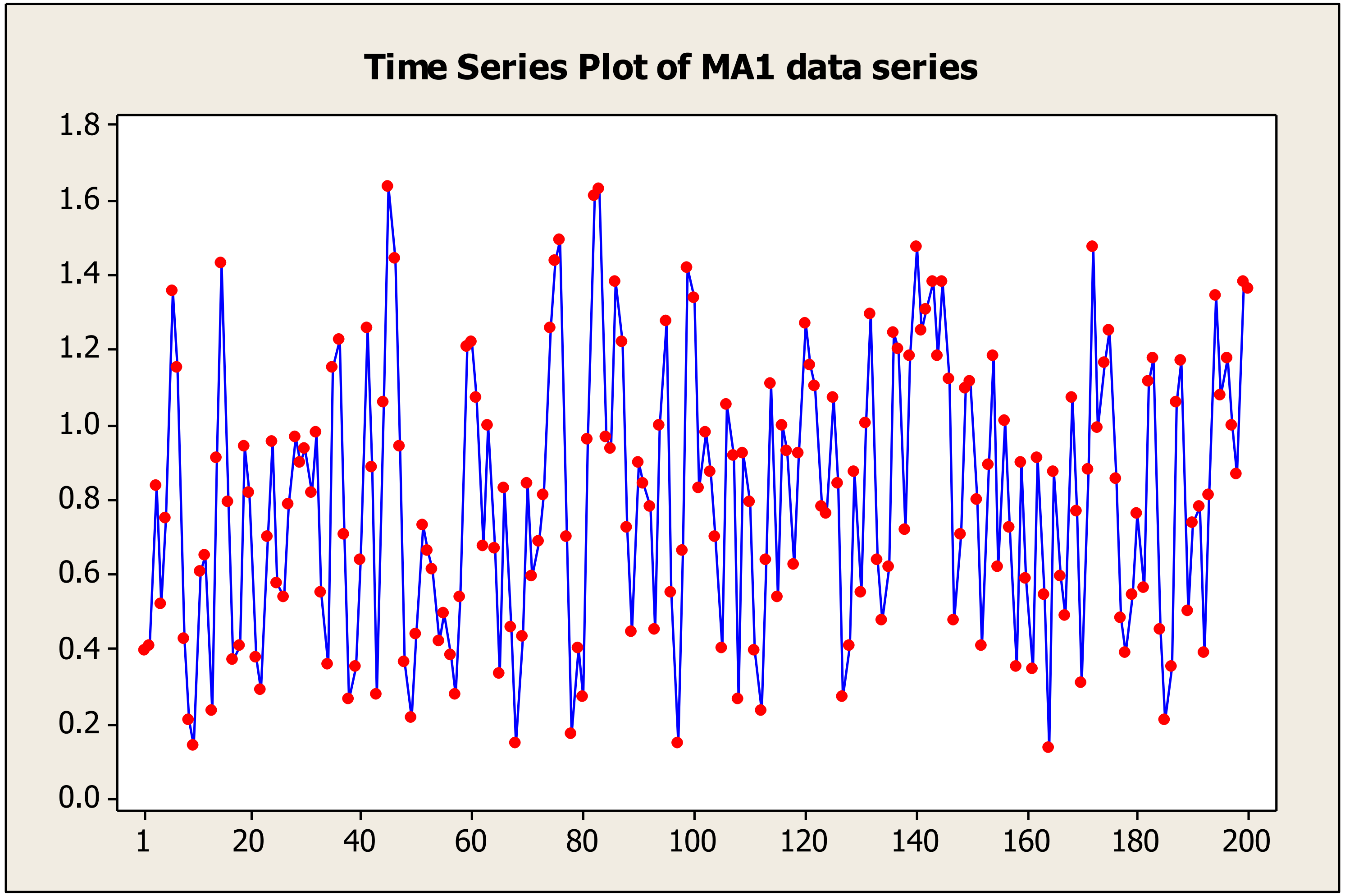
A Moving Average of Order One MA(1)

- The general form of ARIMA (0, 0, 1) or MA(1) model is

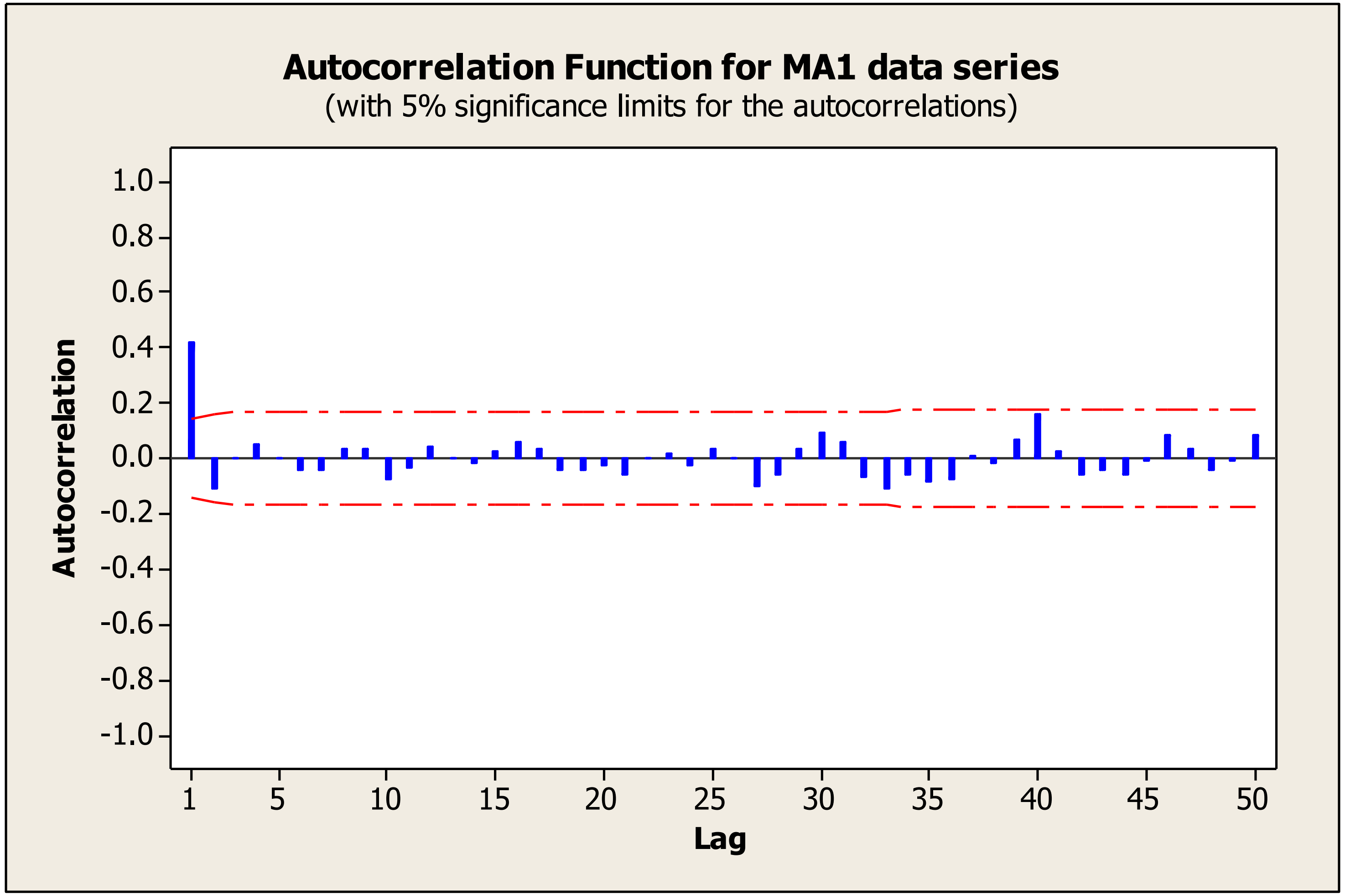
$$y_t = C + e_t - \theta_1 e_{t-1}$$

  - Y_t depends on the error term e_t and on the previous error term e_{t-1} with coefficient $-\theta_1$.
  - The value of θ_1 is between –1 and 1.
  - The following slides show an example of an MA(1) data series.

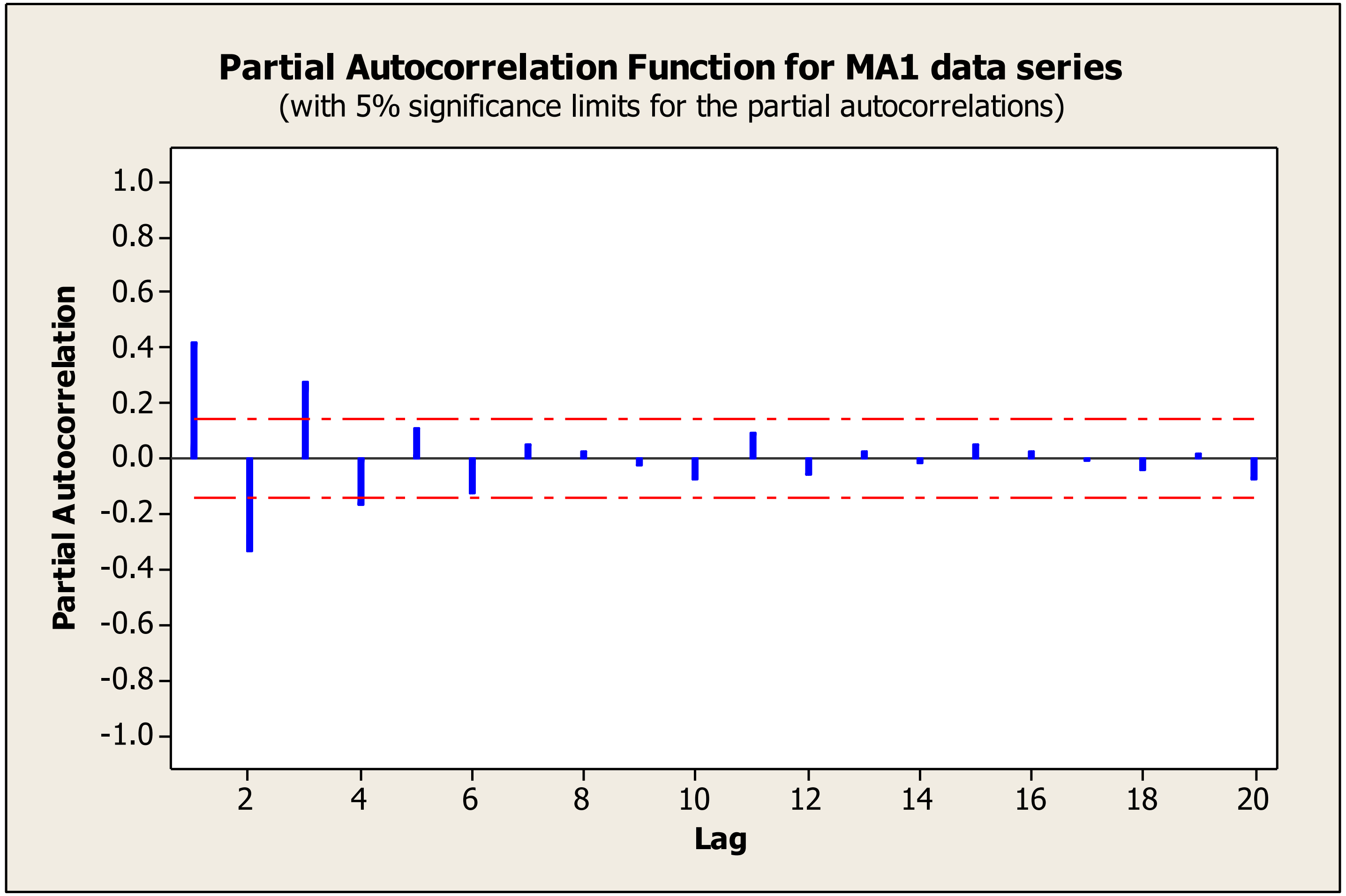
A Moving Average of Order One MA(1)



A Moving Average of Order One MA(1)



A Moving Average of Order One MA(1)



A Moving Average of Order One MA(1)

- Note that there is only one significant autocorrelation at time lag 1.
- The partial autocorrelations decay exponentially, but because of random error components, they do not die out to zero as do the theoretical autocorrelation.

Higher Order Auto Regressive Models

- A pth-order AR model is defined as

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

  - C is the constant term
  - ϕ_j is the jth auto regression parameter
  - e_t is the error term at time t.

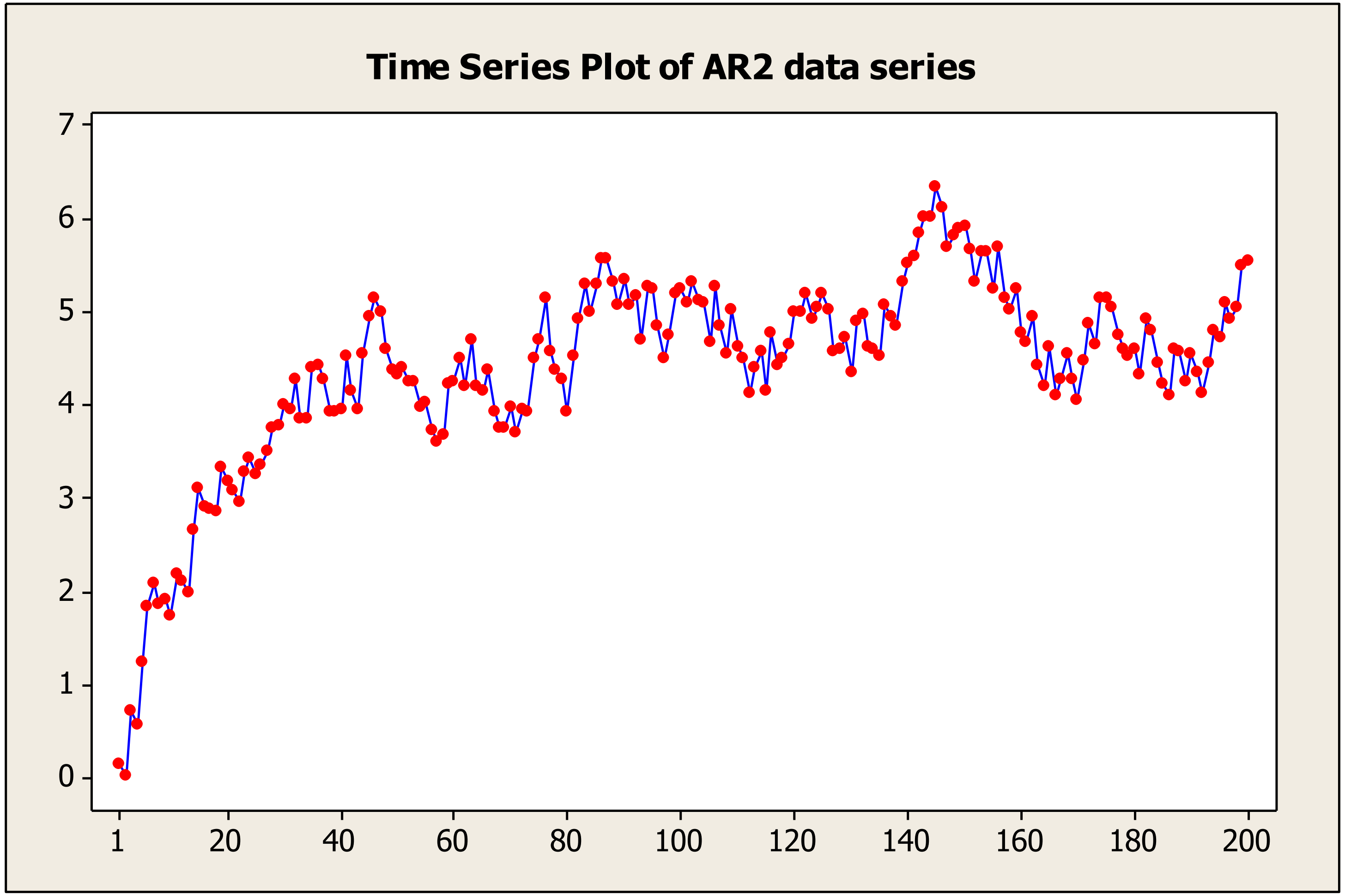
Higher Order Auto Regressive Models

- Restrictions on the allowable values of auto regression parameters
  - For p =1
    - $-1 < \phi_1 < 1$
  - For p = 2
    - $-1 < \phi_2 < 1$
    - $\phi_1 + \phi_2 < 1$
    - $\phi_2 - \phi_1 < 1$

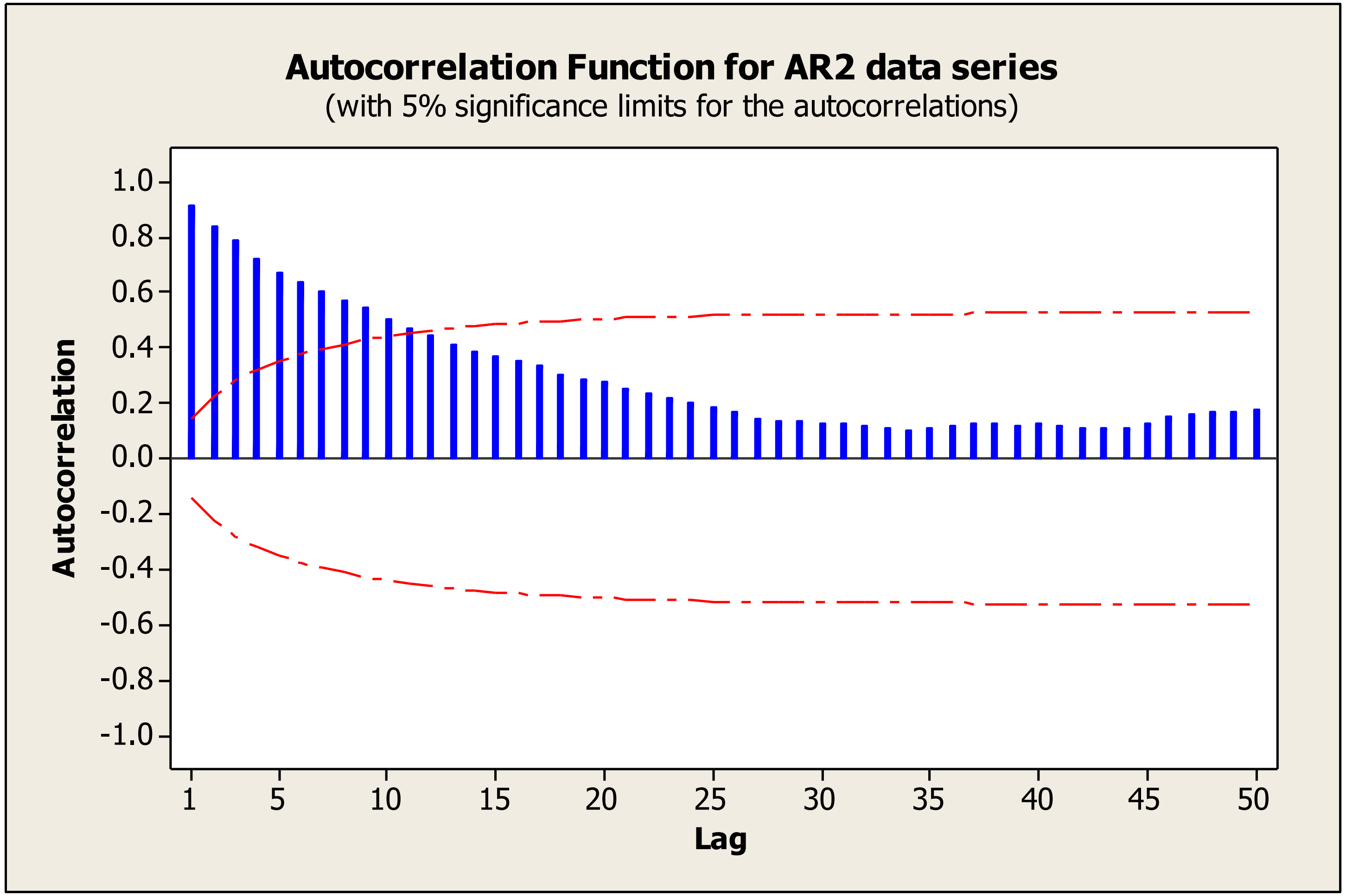
Higher Order Auto Regressive Models

- A great variety of time series are possible with autoregressive models.
- The following slides shows an AR(2) model.
- Note that for AR(2) models the autocorrelations die out in a damped Sine-wave patterns.
- There are exactly two significant partial autocorrelations.

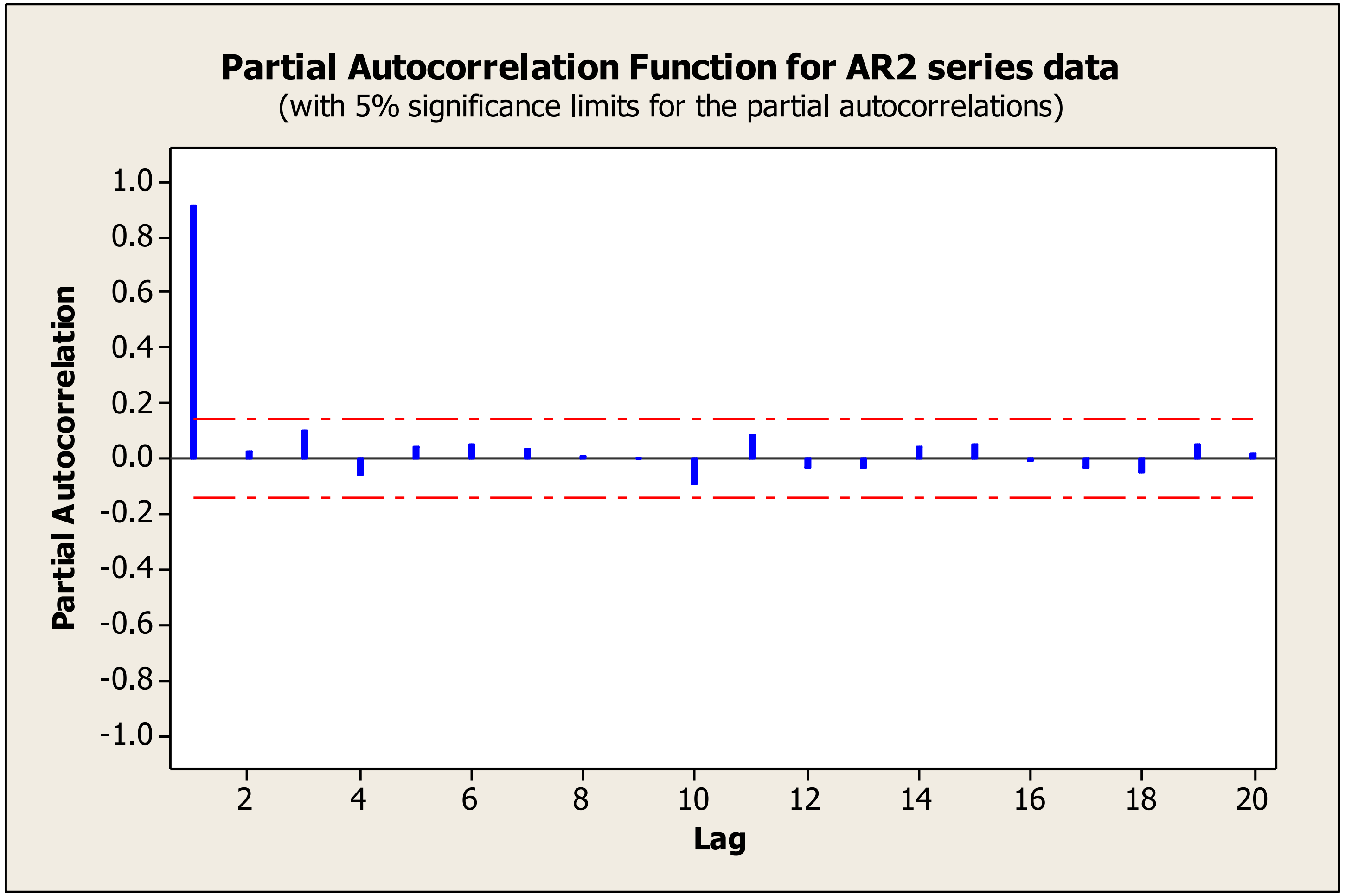
Higher Order Auto Regressive Models



Higher Order Auto Regressive Models



Higher Order Auto Regressive Models



Higher Order Moving Average Models

- The general MA model of order q can be written as

$$y_t = C + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q}$$

  - C is the constant term
  - θ_j is the *j*th moving average parameter.
  - e_{t-k} is the error term at time $t-k$

Higher Order Moving Average Models

- Restrictions on the allowable values of the MA parameters.
  - For q =1
    - $-1 < \theta_1 < 1$
  - For q =2
    - $-1 < \theta_2 < 1$
    - $\theta_1 + \theta_2 < 1$
    - $\theta_2 - \theta_1 < 1$

Higher Order Moving Average Models

- A wide variety of time series can be produced using moving average models.
- In general, the autocorrelations of an MA(q) models are zero beyond lag q
- For $q \geq 2$, the PACF can show exponential decay or damped sine-wave patterns.

Mixtures ARMA Models

- Basic elements of AR and MA models can be combined to produce a great variety of models.
- The following is the combination of MA(1) and AR(1) models

$$y_t = C + \phi_1 y_{t-1} + e_t - \theta_1 e_{t-1}$$

- This is model called ARMA(1, 1) or

  ARIMA (1, 0, 1)
- The series is assumed stationary in the mean and in the variance.

Mixtures ARIMA Models

- If non-stationarity is added to a mixed ARMA model, then the general ARIMA (p, d, q) is obtained.
- The equation for the simplest ARIMA (1, 1, 1) is given below.

$$y_t = C + \phi_1 y_{t-1} - \phi_1 y_{t-2} + e_t - \theta_1 e_{t-1}$$

Mixtures ARIMA Models

- The general ARIMA (p, d, q) model gives a tremendous variety of patterns in the ACF and PACF, so it is not practical to state rules for identifying general ARIMA models.
- In practice, it is seldom necessary to deal with values p, d, or q that are larger than 0, 1, or 2.
- It is remarkable that such a small range of values for p, d, or q can cover such a large range of practical forecasting situations.

Seasonality and ARIMA Models

- The ARIMA models can be extended to handle seasonal components of a data series.
- The general shorthand notation is

  $$\text{ARIMA } (p, d, q)(P, D, Q)_s$$

  - Where s is the number of periods per season.

Seasonality and ARIMA Models

- The general ARIMA(1,1,1)(1,1,1)$_4$ can be written as

$$y_t = (1+\phi_1)y_{t-1} - \phi_1 y_{t-2} + (1+\Phi_1)y_{t-4} - (1+\phi_1+\Phi_1+\phi_1\Phi_1)y_{t-6}$$
$$- \Phi_1 y_{t-8} + (\Phi_1+\phi_1\Phi_1)y_{t-9} - \phi_1\Phi_1 y_{t-10} + e_t - \theta_1 e_{t-1} - \Theta_1 e_{t-4} + \theta_1\Theta_1 e_{t-5}$$

- Once the coefficients ϕ_1, Φ_1, θ_1, and Θ_1 have been estimated from the data, the above equation can be used for forecasting.

Seasonality and ARIMA Models

- The seasonal lags of the ACF and PACF plots show the seasonal parts of an AR or MA model.
- Examples:
  - Seasonal MA model:
    - ARIMA(0,0,0)(0,0,1)$_{12}$
      - will show a spike at lag 12 in the ACF but no other significant spikes.
      - The PACF will show exponential decay in the seasonal lags i.e. at lags 12, 24, 36,…

Center for Business Intelligence

Seasonality and ARIMA Models

Seasonal AR model:

- ARIMA(0,0,0)(1,0,0)$_{12}$
  - will show exponential decay in seasonal lags of the ACF.
  - Single significant spike at lag 12 in the PACF.

Implementing the Model–Building Strategy

- The Box –Jenkins approach uses an iterative model-building strategy that consist of
  - Selecting an initial model (model identification)
  - Estimating the model coefficients (parameter estimation)
  - Analyzing the residuals (model checking)
- If necessary, the initial model is modified and the process is repeated until the residual indicate no further modification is necessary. At this point the fitted model can be used for forecasting.

Model Identification Methodology

- Plot the data
  - Identify any unusual observations
  - If necessary, transform the data to stabilize the variance
- Check the time series plot, ACF, PACF of the data (possibly transformed) for stationarity.
- IF
  - Time plot shows the data scattered horizontally around a constant mean
  - ACF and PACF to or near zero quickly
- Then, the data are stationary.

Model Identification Methodology

- Use differencing to transform the data into a stationary series
  - For no-seasonal data take first differences
  - For seasonal data take seasonal differences
- Check the plots again if they appear non-stationary, take the differences of the differenced data.
- When the stationarity has been achieved, check the ACF and PACF plots for any pattern remaining.
- There are three possibilities
  - AR or MA models
    - No significant ACF after time lag q indicates MA(q) may be appropriate.
    - No significant PACF after time lag p indicates that AR(p) may be appropriate.

Summary: To determine *p* and *q*.

Use the following table.

	MA(q)	AR(p)	ARMA(p,q)
ACF	Cuts after q	Tails off	Tails off
PACF	Tails off	Cuts after p	Tails off

Note: Usually $p + q \leq 4$. There is no harm in over identifying the time series. (allowing more parameters in the model than necessary. We can always test to determine if the extra parameters are zero.)

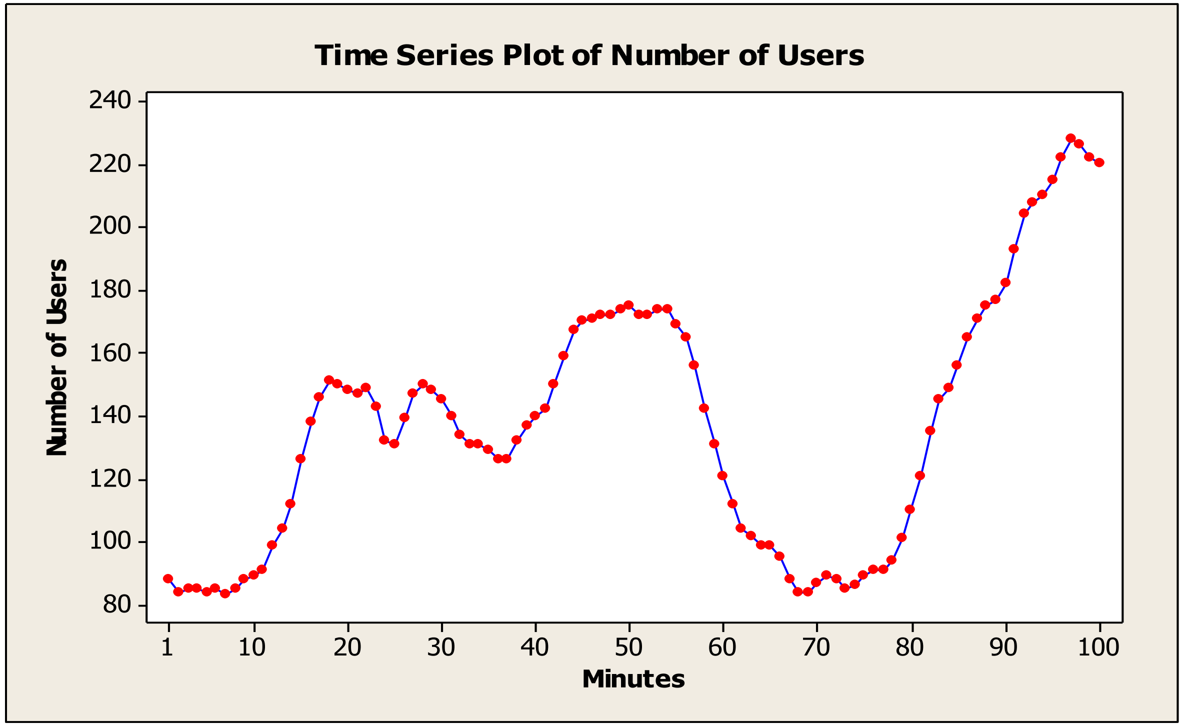
Model Identification Methodology

- Seasonality is present if ACF and/or PACF at the seasonal lags are large and significant.
- If no clear MA or AR model is suggested, a mixture model may be appropriate.

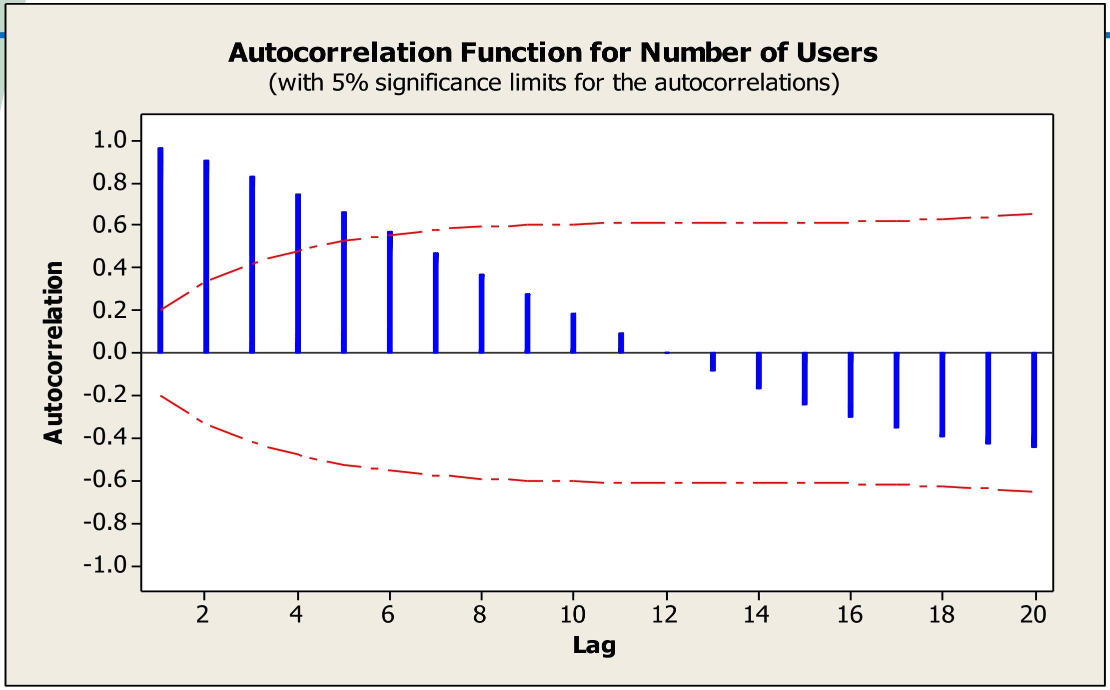
Center for Business Intelligence

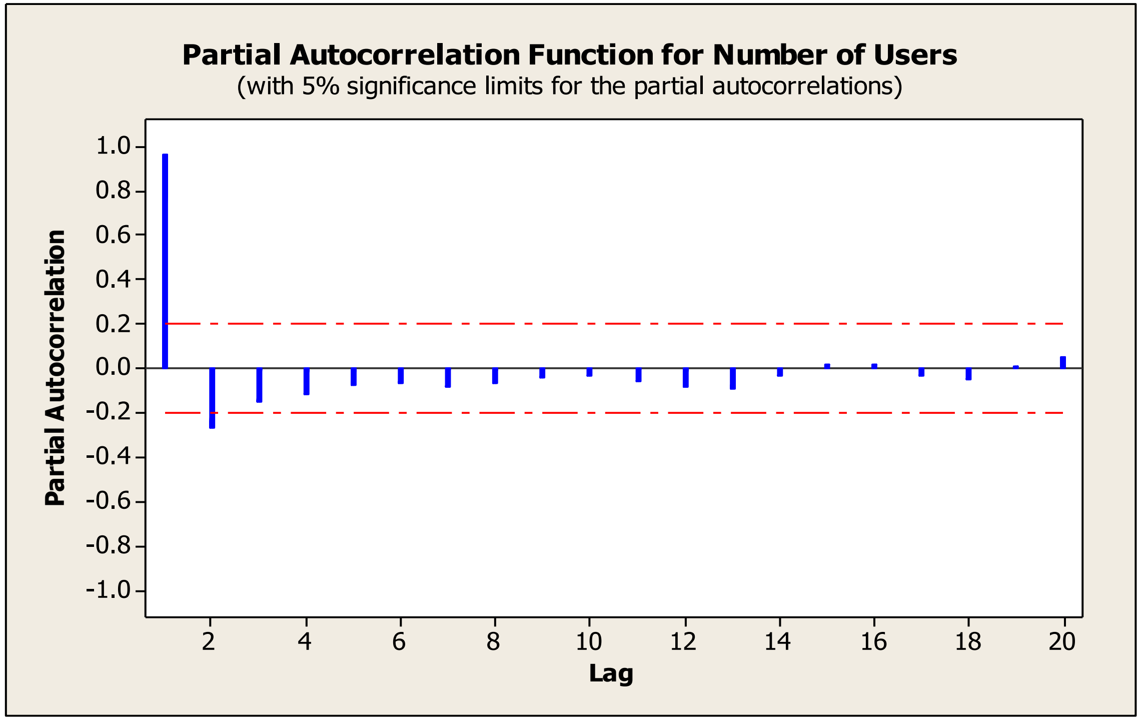
Model Identification Example

- Non seasonal time series data.
  - The following example looks at the number of users logged onto an internet server over a 100 minutes period.
  - The time plot, ACF and PACF is reported as follows:



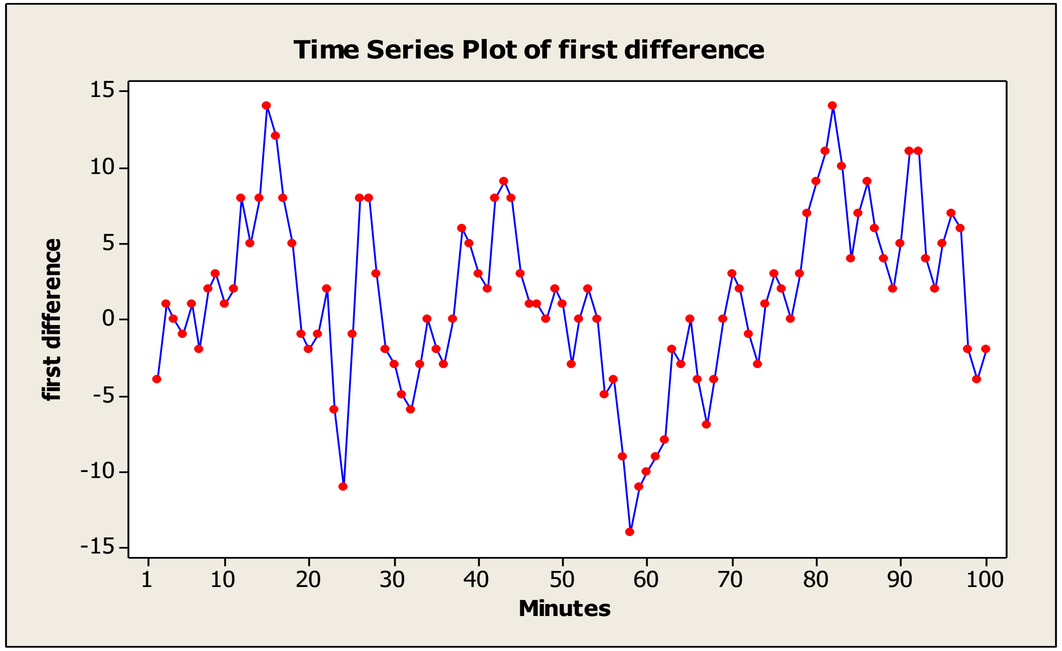
Model Identification Example



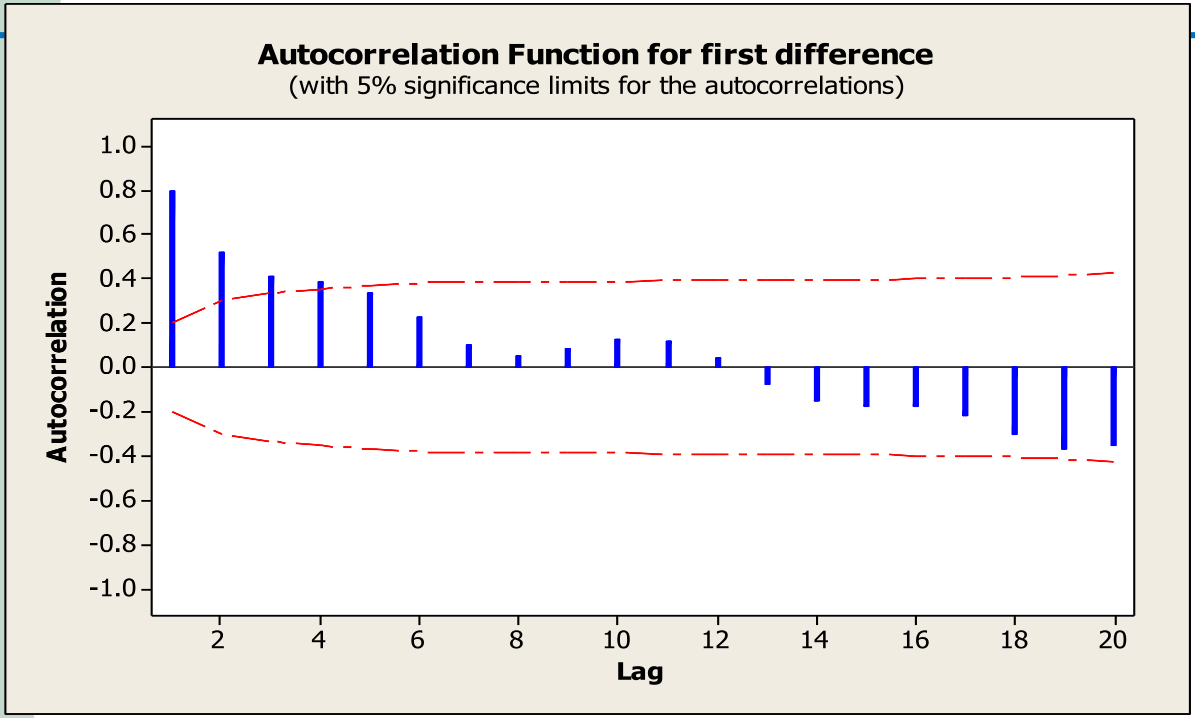


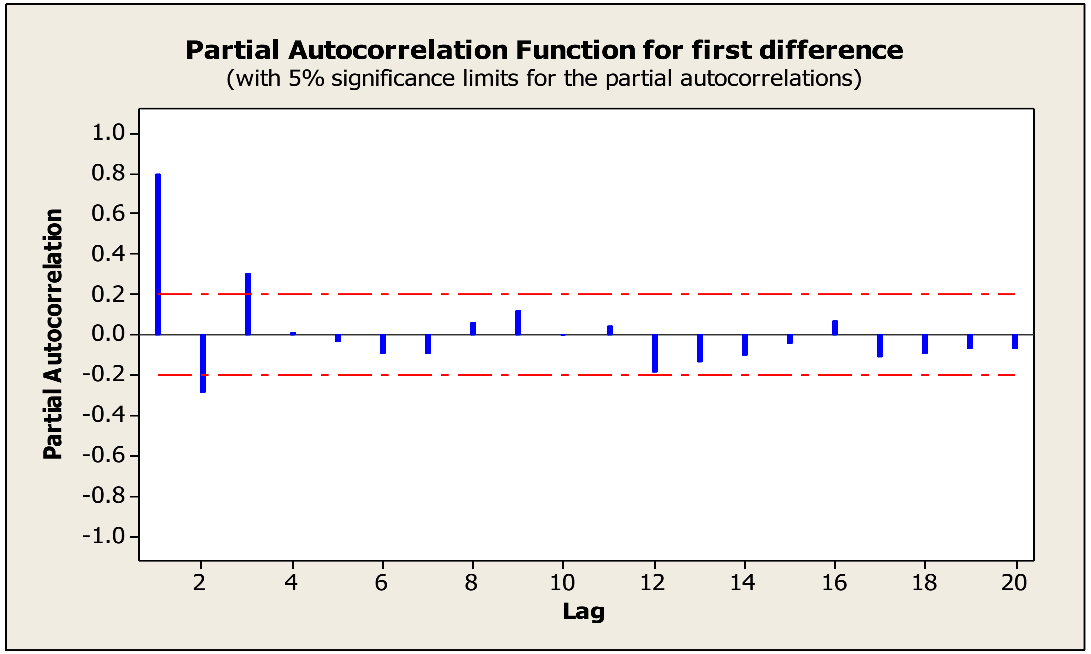
Model Identification Example

- The gradual decline of ACF values indicates non-stationary series.
- The first partial autocorrelation is very dominant and close to 1, indicating non-stationarity.
- The time series plot clearly indicates non-stationarity.
- We take the first differences of the data and reanalyze.



Model Identification Example





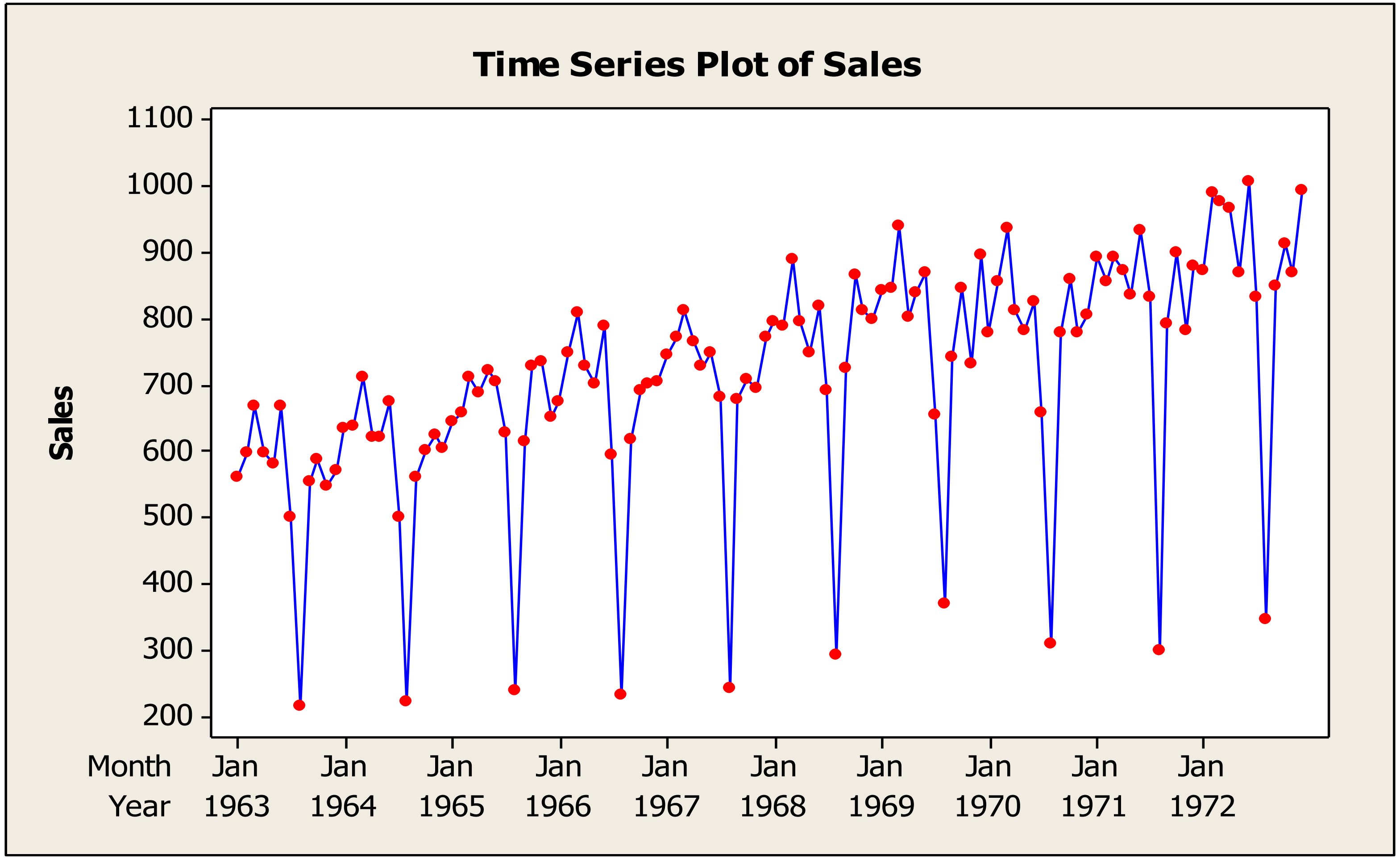
Model Identification Example

- ACF shows a mixture of exponential decay and sine-wave pattern
- PACF shows three significant PACF values.
- This suggests an AR(3) model.
- This identifies an ARIMA(3,1,0).

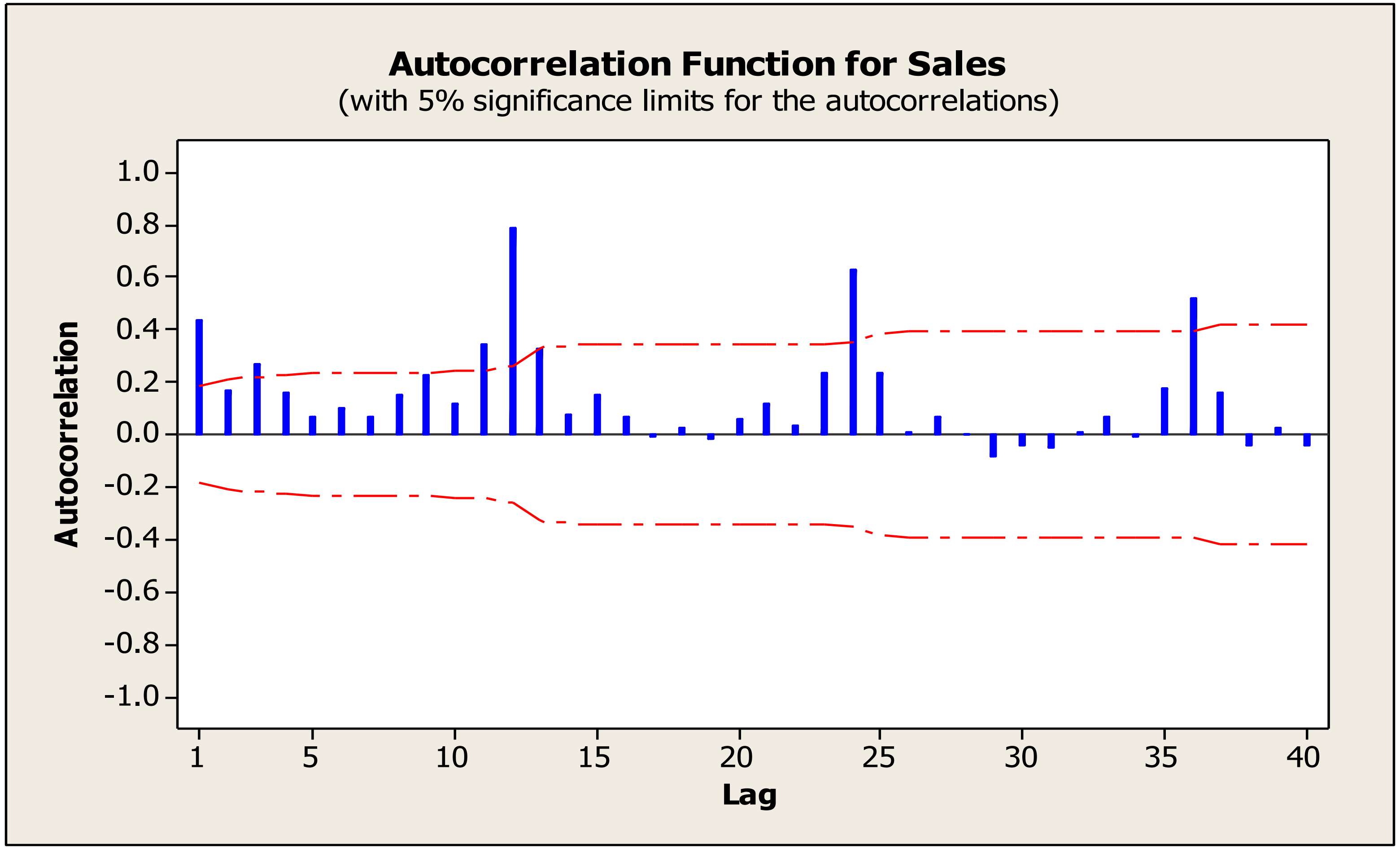
Model Identification Example

- A seasonal time series.
  - The following example looks at the monthly industry sales (in thousands of francs) for printing and writing papers between the years 1963 and 1972.
  - The time plot, ACF and PACF shows a clear seasonal pattern in the data.
  - This is clear in the large values at time lag 12, 24 and 36.

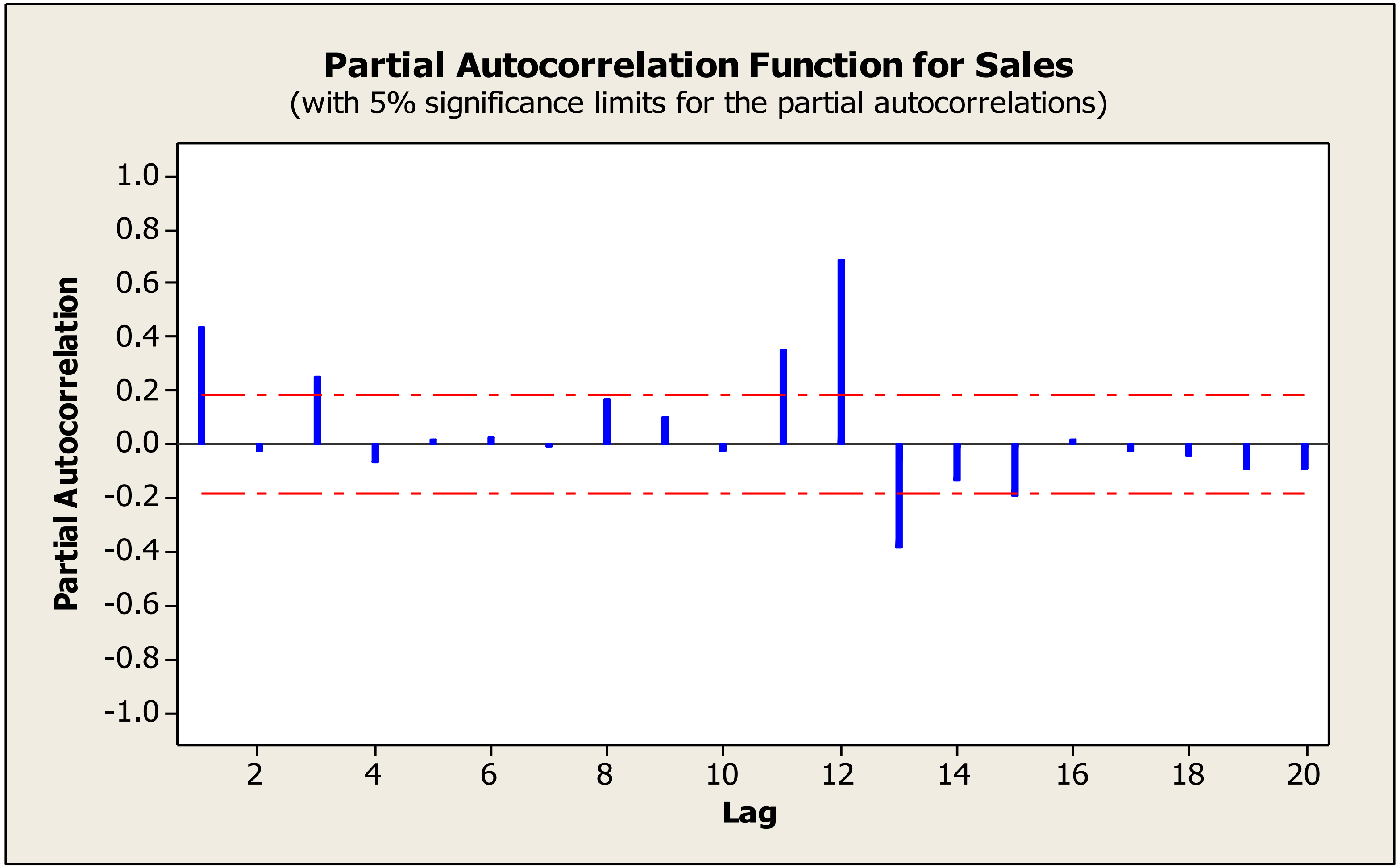
Model Identification Example



Model Identification Example



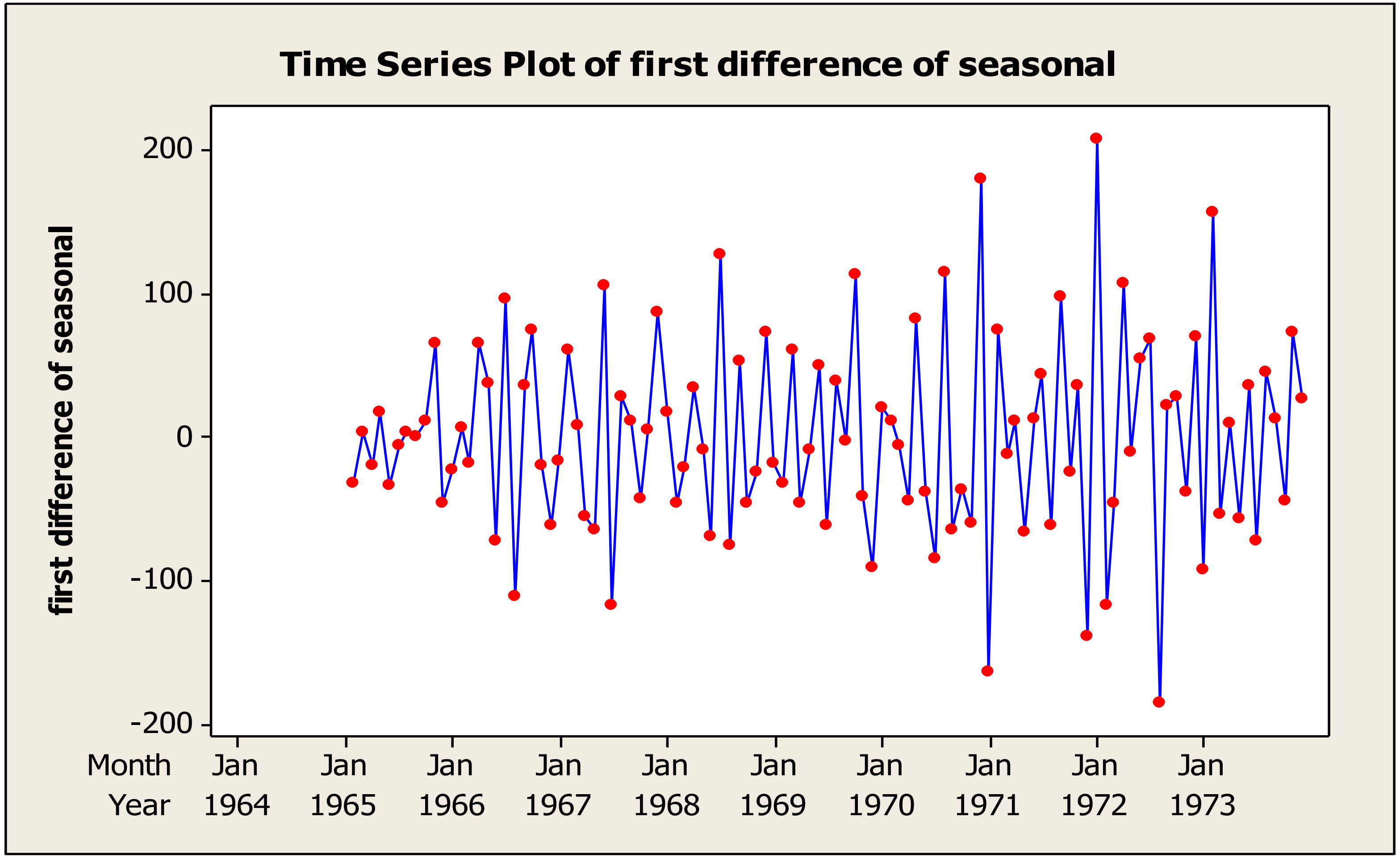
Model Identification Example



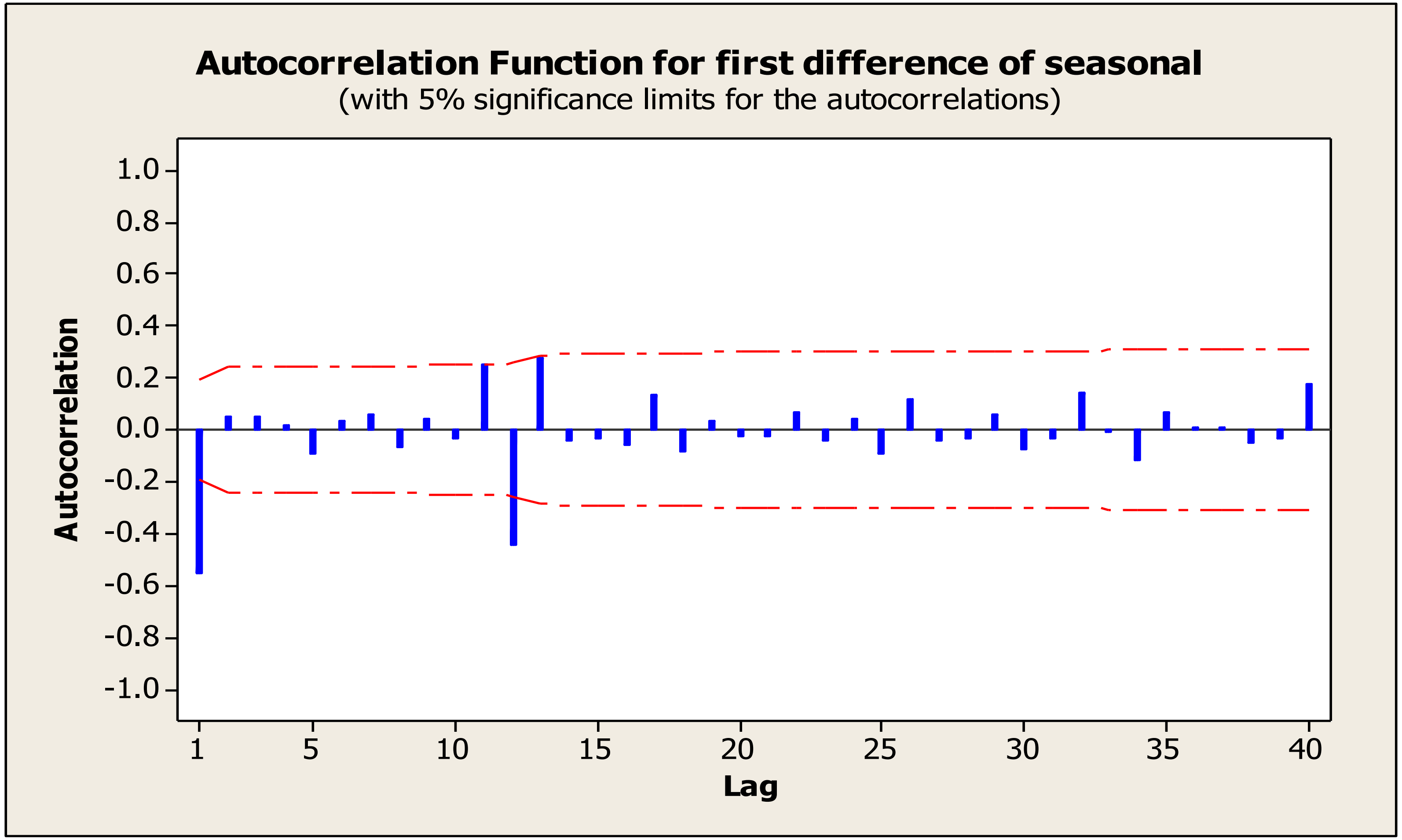
Model Identification Example

- We take a seasonal difference and check the time plot, ACF and PACF.
- The seasonally differenced data appears to be non-stationary (the plots are not shown), so we difference the data again.

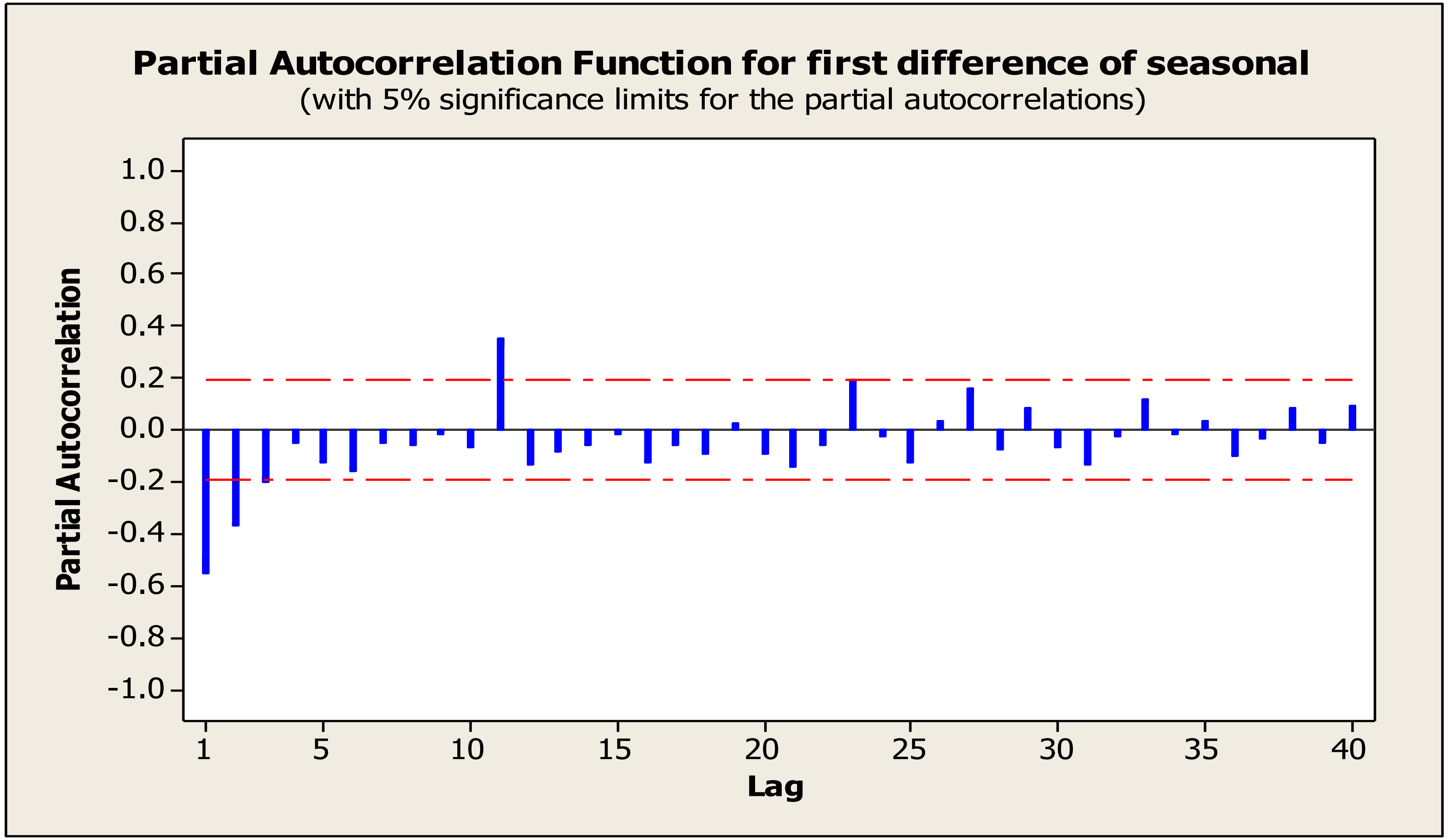
Model Identification Example



Model Identification Example



Model Identification Example



Model Identification Example

- The PACF shows the exponential decay in values.
- The ACF shows a significant value at time lag 1.
  - This suggest a MA(1) model.
- The ACF also shows a significant value at time lag 12
  - This suggest a seasonal MA(1).

Model Identification Example

- Therefore, the identified model is

  ARIMA $(0,1,1)(0,1,1)_{12}$.
- This model is sometimes is called the "airline model" because it was applied to international airline data by Box and Jenkins.
- It is one of the most commonly used seasonal ARIMA model.

Center for Business Intelligence

Summary

- The process of identifying an ARIMA model requires experience and good judgment. The following guidelines can be helpful.
  – Make the series stationary in mean and variance
    - Differencing will take care of non-stationarity in the mean.
    - Logarithmic or power transformation will often take care of non-stationarity in the variance.

Summary

- Consider non-seasonal aspect
  - The ACF and PACF of the stationary data obtained from the previous step can reveal whether MA of AR is feasible
    - Exponential decay or damped sine-wave. For ACF, and spikes at lags 1 to p then cut off to zero, indicate an AR(P) model.
    - Spikes at lag1 to q, then cut off to zero for ACF and exponential decay or damped sine-wave for PACF indicates MA(q) model.

Summary

- Consider seasonal aspect
  - Examination of ACF and PACF at the seasonal lags can help to identify AR and MA models for the seasonal aspect of the data.
    - For example, for quarterly data the pattern of r_4, r_8, r_{12}, r_{16}, and so on.

Estimating the Parameters

- Once a tentative model has been selected, the parameters for the model must be estimated.
- The method of least squares can be used for RIMA model.
- However, for models with an MA components, there is no simple formula that can be used to estimate the parameters.
- Instead, an iterative method is used. This involves starting with a preliminary estimate, and refining the estimate iteratively until the sum of the squared errors is minimized.

Estimating the Parameters

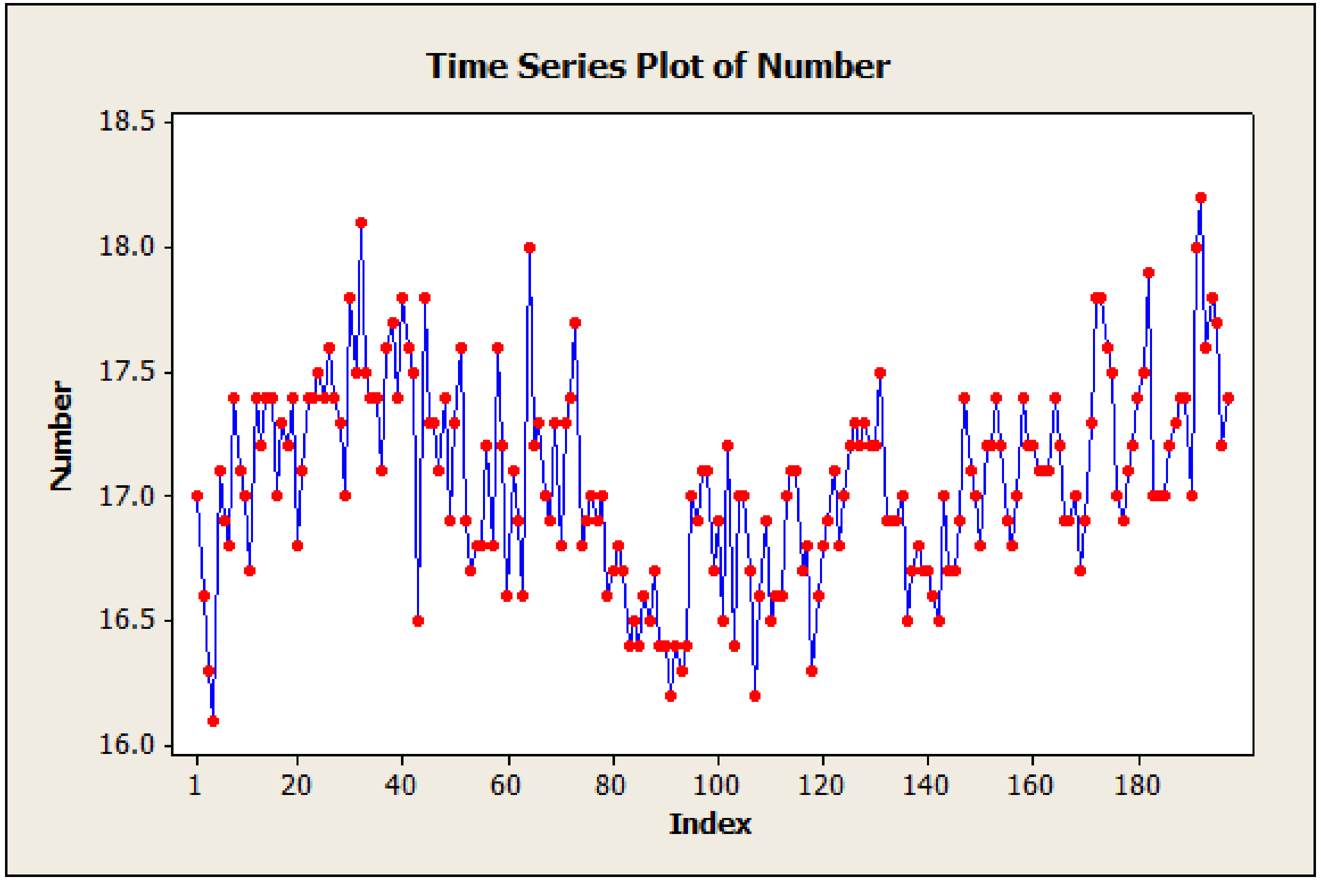
- Another method of estimating the parameters is the maximum likelihood procedure.
- Like least squares methods, these estimates must be found iteratively.
- Maximum likelihood estimation is usually favored because it has some desiable statistical properties.

Estimating the Parameters

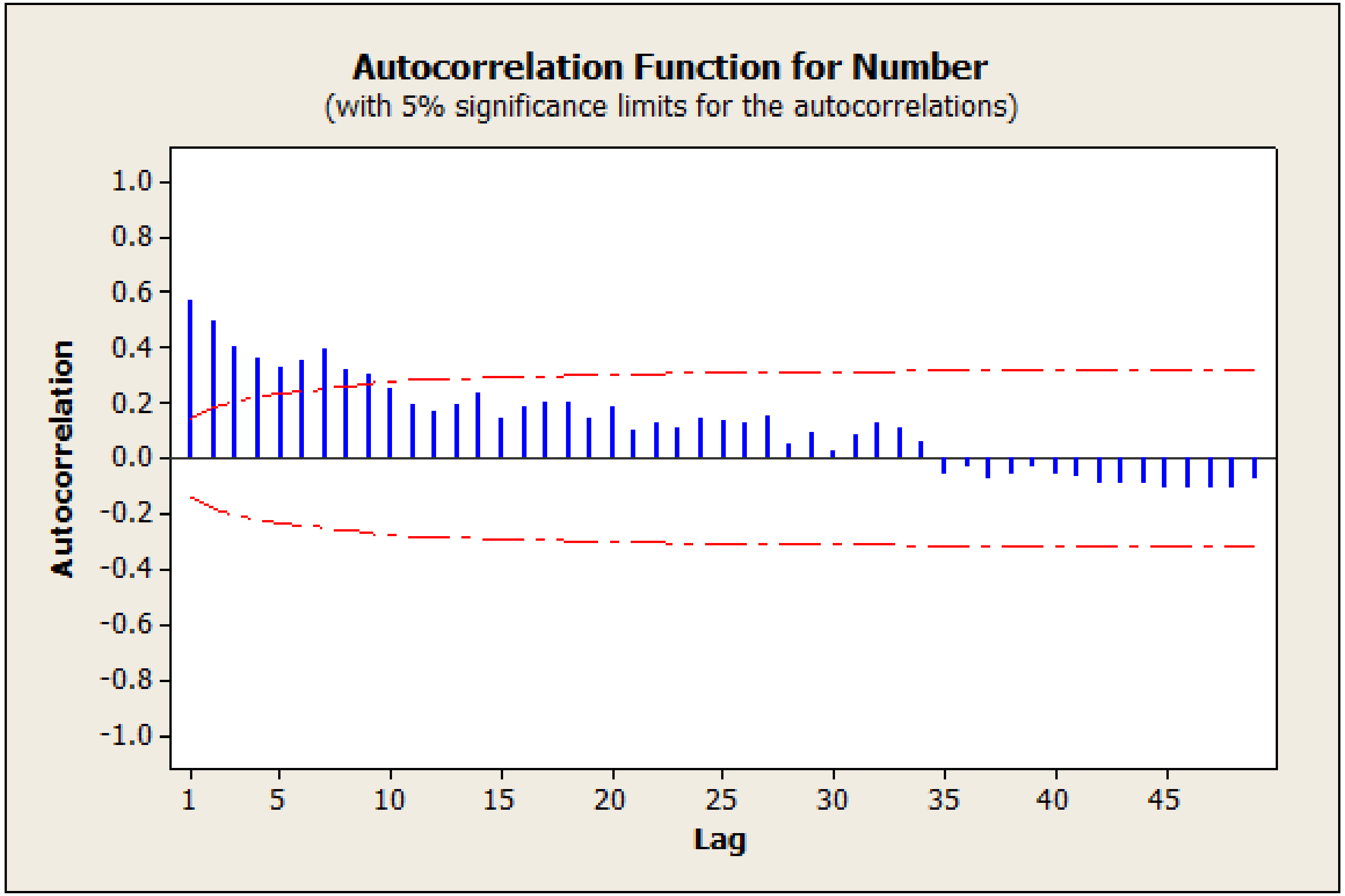
- After the estimates and their standard errors are determined, t values can be constructed and interpreted in the usual way.
- Parameters that are judged significantly different from zero are retained in the fitted model; parameters that are not significantly different from zero are dropped from the model.

Example

- Annual Sunspots Dataset

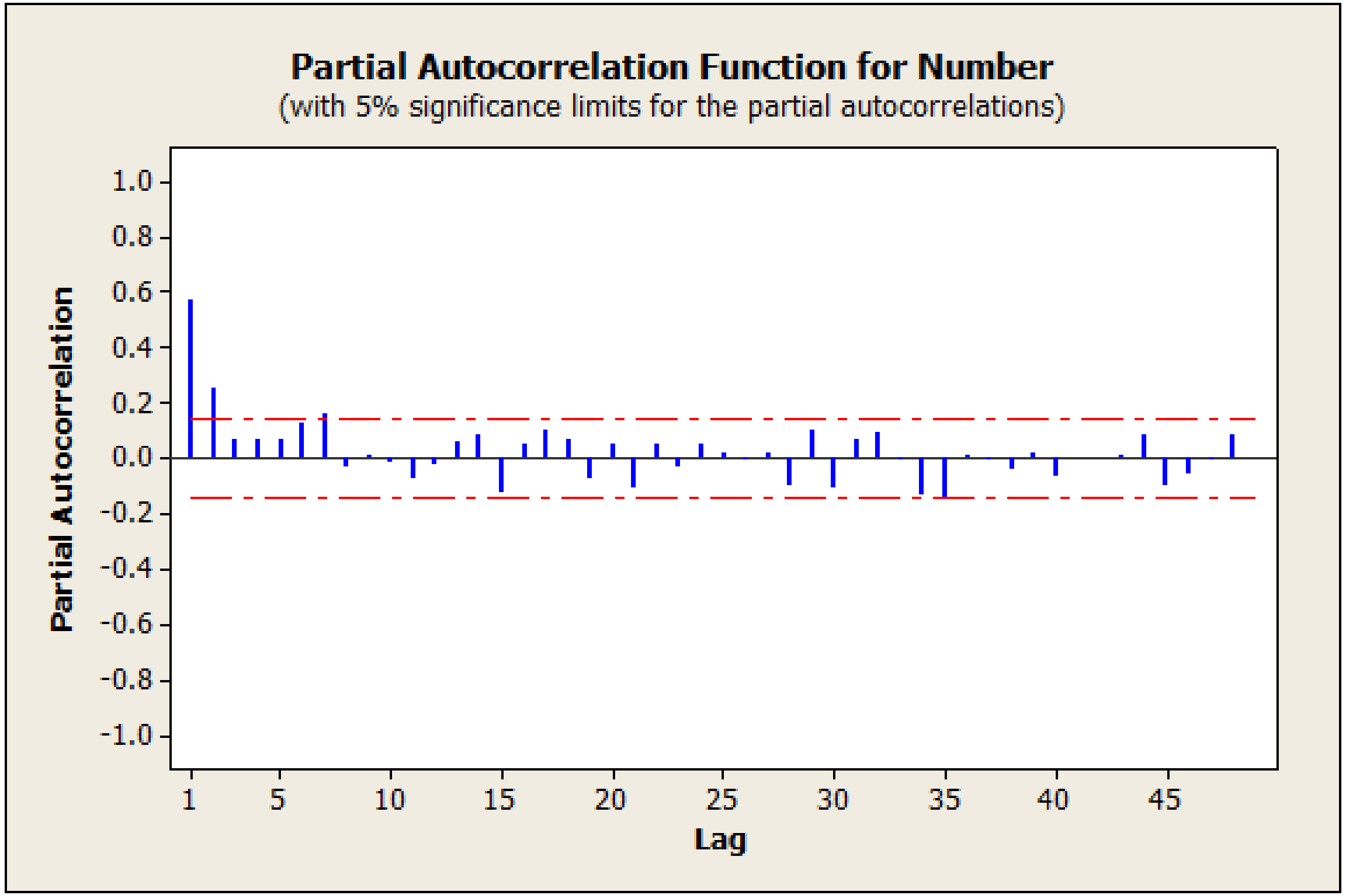


ACF

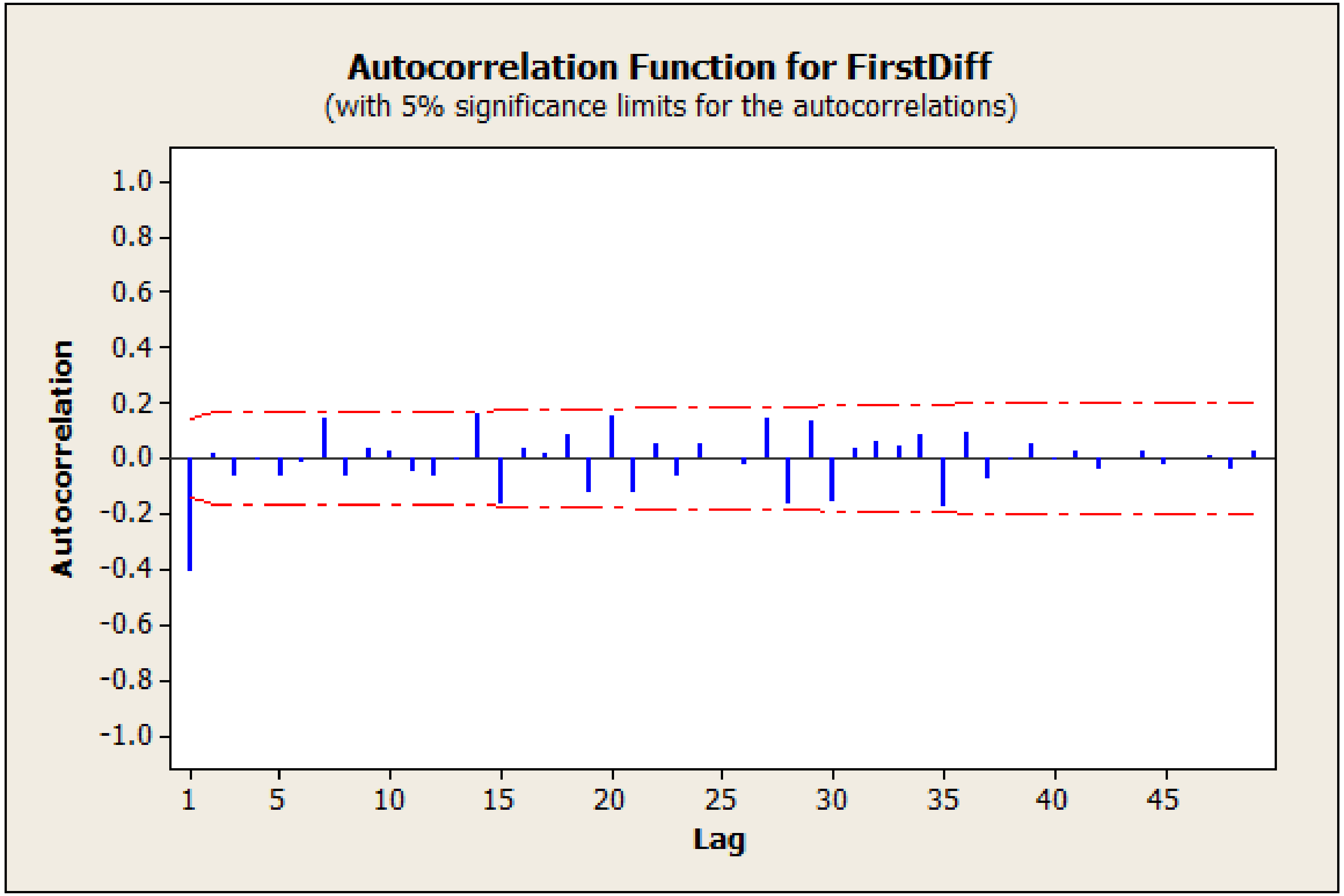


Autocorrelation Function for Number
(with 5% significance limits for the autocorrelations)
Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1 5 10 15 20 25 30 35 40 45
Lag

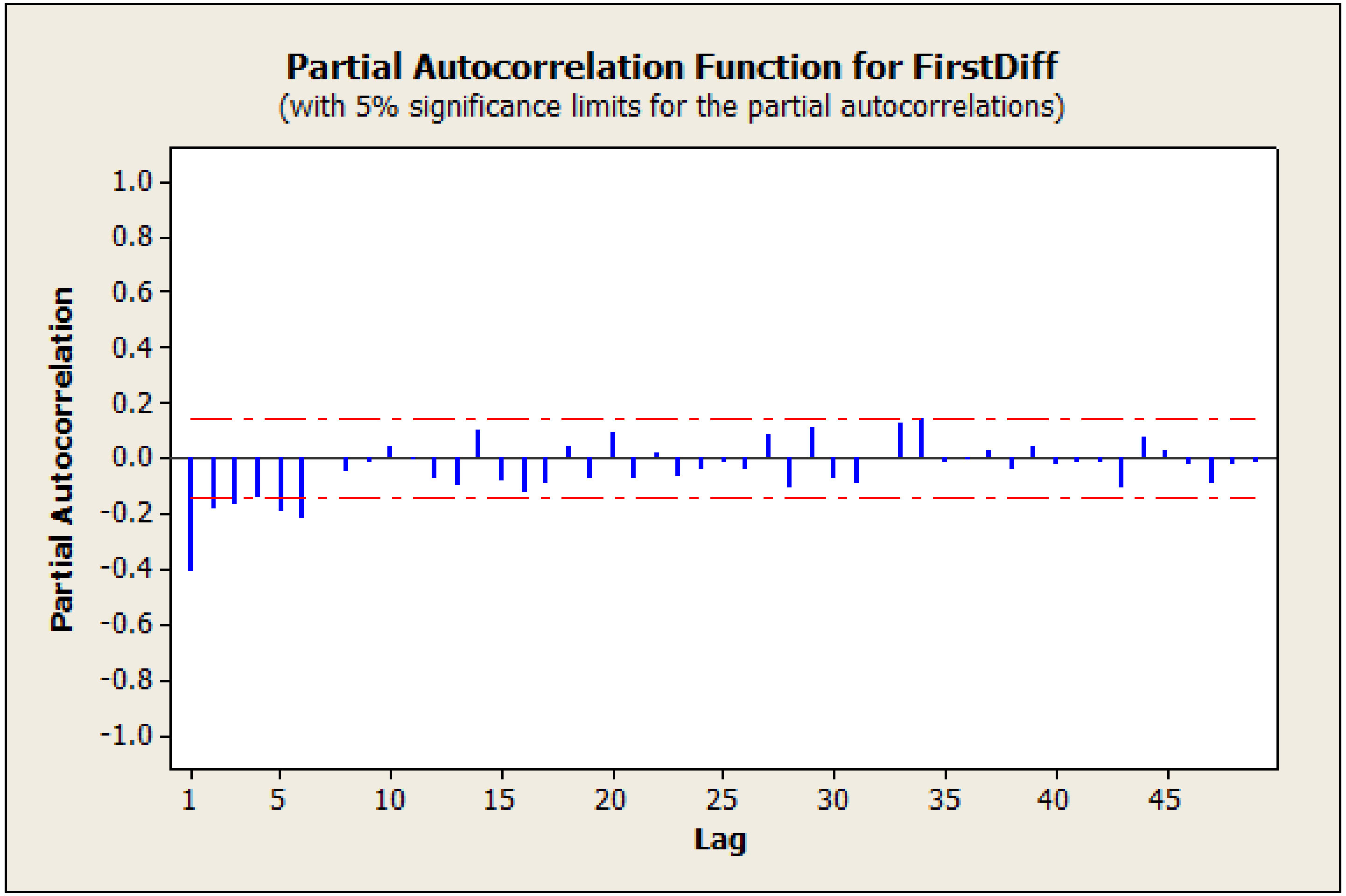
PCF


Partial Autocorrelation Function for Number
(with 5% significance limits for the partial autocorrelations)
Partial Autocorrelation
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
1
5
10
15
20
25
30
35
40
45
Lag

First Difference ACF



First Difference PCF



Minitab Results

- p=1, d=1, q=1,

Final Estimates of Parameters

Type	Coef	SE Coef	T	P
AR 1	0.2202	0.0935	2.36	0.020
MA 1	0.8295	0.0529	15.68	0.000
Constant	0.002803	0.003879	0.72	0.471

Differencing: 1 regular difference
Number of observations: Original series 197, after differencing 196
Residuals: SS = 19.2258 (backforecasts excluded)
MS = 0.0996 DF = 193

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	11.2	24.9	46.5	50.9
DF	9	21	33	45
P-Value	0.262	0.252	0.060	0.253

Forecasts

- Forecast for period 198

$$y_t = C + \phi_1 y_t - \phi_1 y_{t-1} + e_t - \theta_1 e_{t-1}$$

Period	Forecast	Actual
196	17.6379	17.2000
197	17.6271	17.4000

ARIMA Modelling in R

- `auto.arima()` is a function that uses a variation of algorithms that obtains an ARIMA Model
- It follows the steps below:
  - Number of differences d is determined using KPSS tests
  - Values of p and q are chosen by minimizing AICc after differencing the data d times
  - Uses stepwise search to achieve this
- `Arima()` can be used to model your own Arima model

ARIMA Modelling in R

- Modelling procedure:
  - Plot the data and identify unusual observations
  - Transform data, if necessary, to stabilize the variance
  - If data is non-stationary: take the first differences until it becomes stationary
  - Examine ACF and PACF: Is an AR(p) or MA(q) model appropriate?
  - Try chosen model/s and use the AICc to search for a better model
  - Check residuals by plotting its ACF
  - Once residuals are like white noise, calculate the forecasts

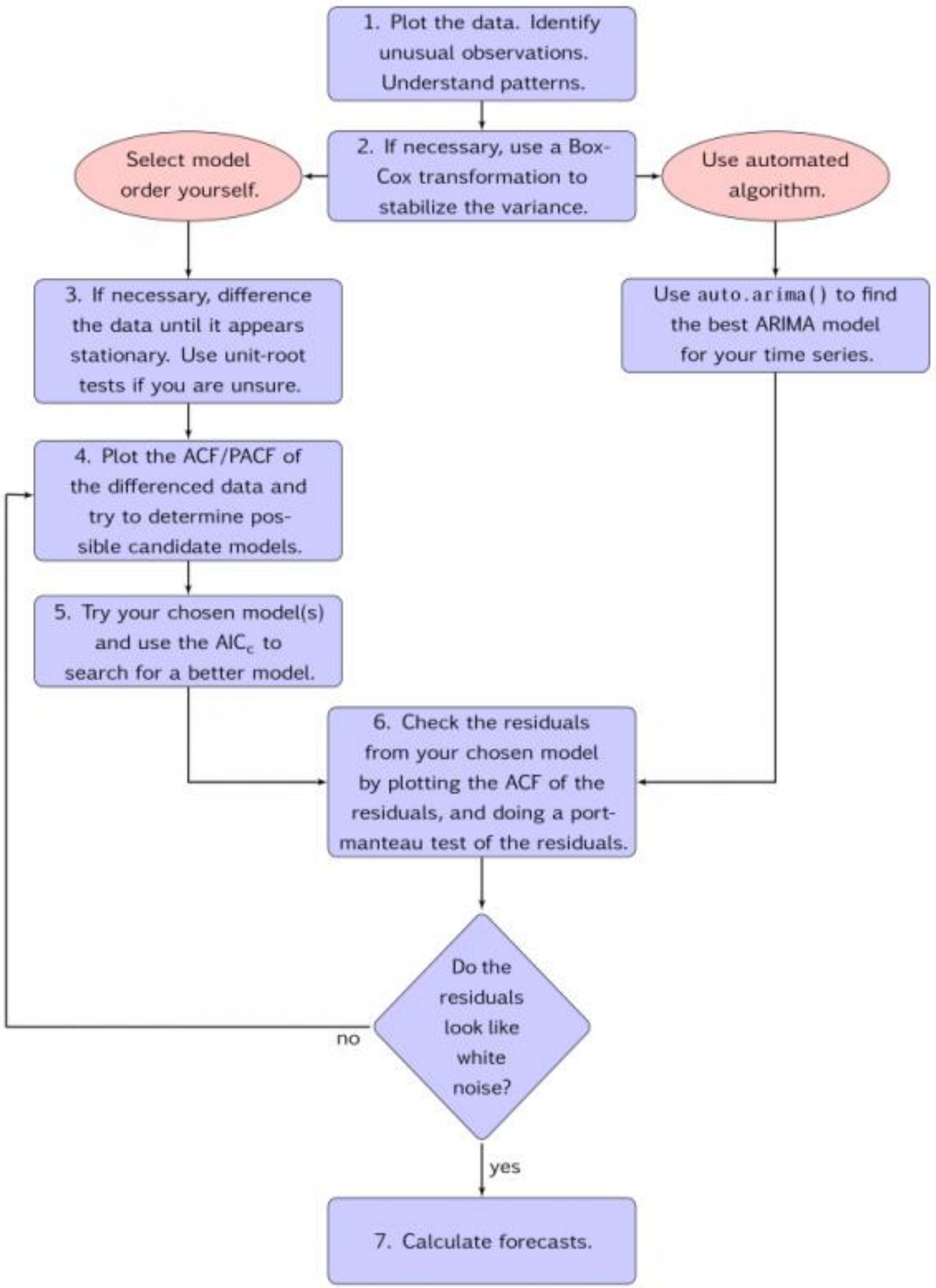
Methodology

1. Examine your data
   - Plot the data and examine its patterns and irregularities
   - Clean any outliers or missing values
   - `tsclean()` to remove outlier
   - Transform (e.g., logarithm) to help stabilize a strong growth trend
2. Decompose your data
   - Does data have trends or seasonality
   - Use `decompose()` or `stl()` to examine and remove components
3. Stationarity
   - Use `adf.test()`, ACF, PACF plots to determine order of differencing

Methodology

4. Autocorrelations and choosing model order
    - Choose order of the ARIMA by examining ACF and PACF plots
5. Fit an ARIMA model
6. Evaluate and iterate
    - Check residuals, which should have no patterns and be normally distributed
    - If there are patterns, plot ACF/PACF
    - Refit model if needed and compare model errors using criteria such as AIC or BIC
    - Calculate forecast using chosen model

ARIMA Modelling in R

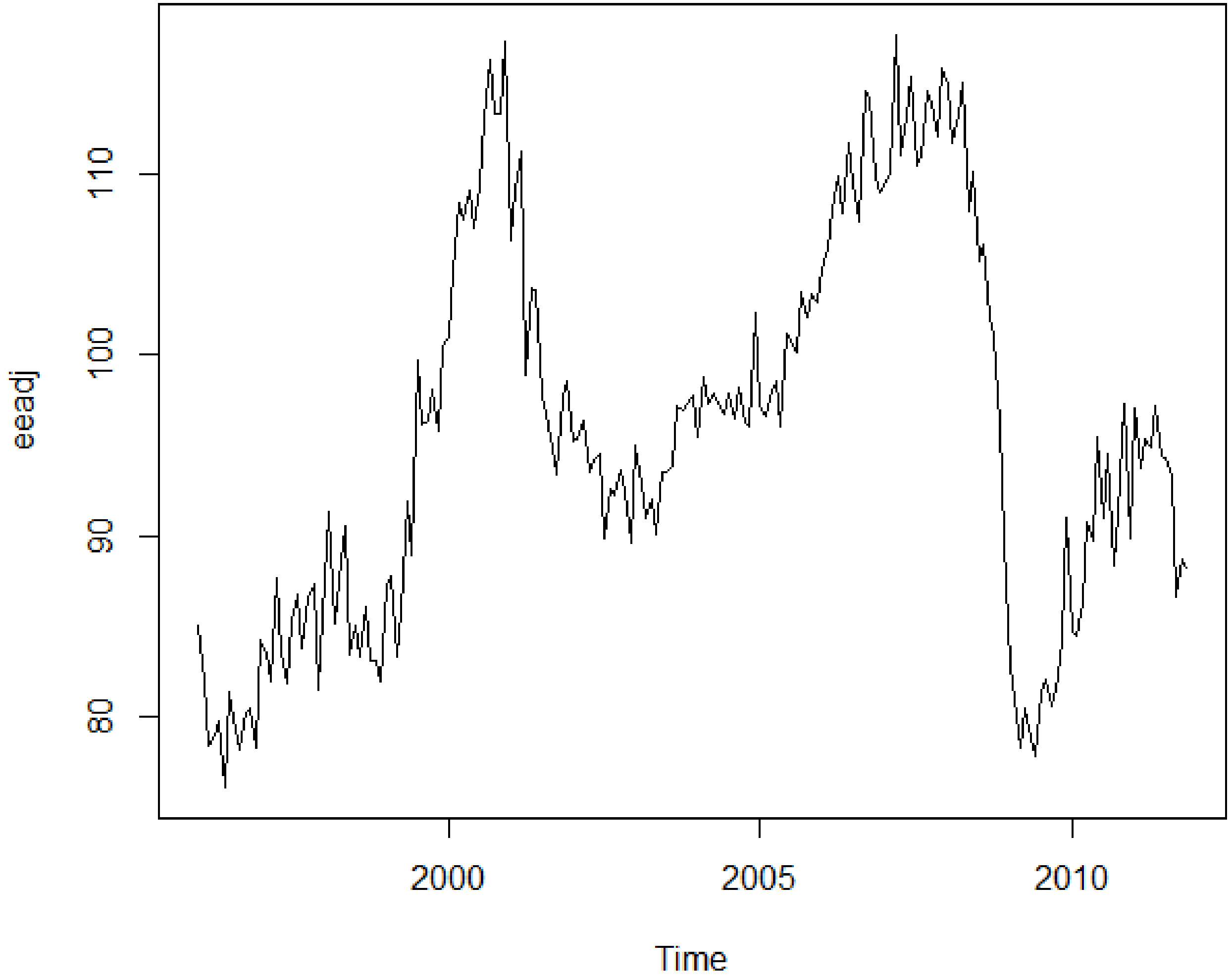


https://www.otexts.org/fpp/8/7

R Code: Electrical Equipment Orders Example

- `eeadj <- seasadj(stl(elecequip, s.window="periodic"))`
- `eeadj`
- `plot(eeadj)`

R Code: Electrical Equipment Orders Example



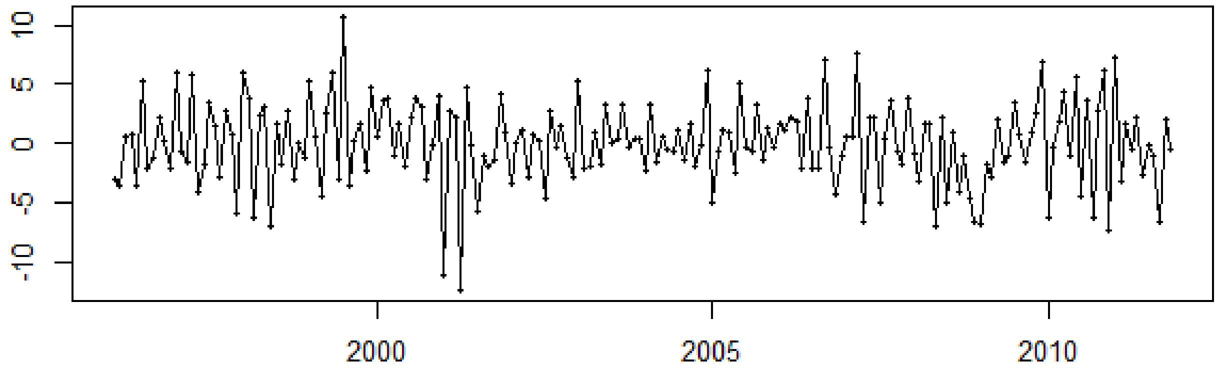
Observations on the Example

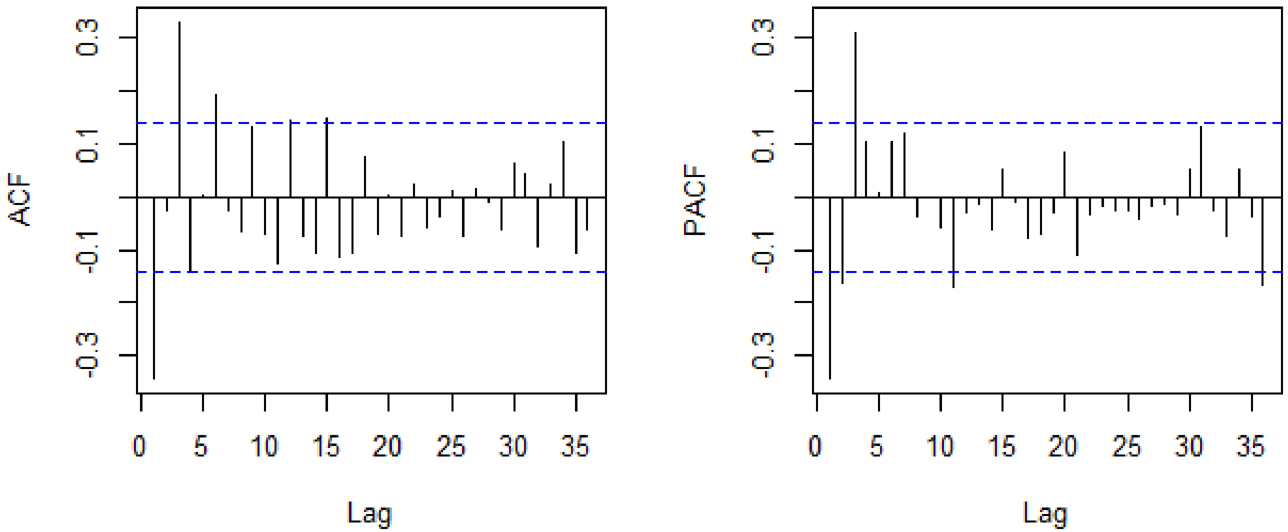
- The time plot shows some sudden changes, particularly the big drop in 2008/2009. These changes are due to the global economic environment. Otherwise, there is nothing unusual about the time plot and there appears to be no need to do any data adjustments.
- There is no evidence of changing variance, so we will not do a Box-Cox transformation.
- The data are clearly non-stationary as the series wanders up and down for long periods.

R Code: Electrical Equipment Orders Example

- `diff(eeadj, lag=1)`
- `tsdisplay(diff(eeadj),main="")`

R Code: Electrical Equipment Orders Example





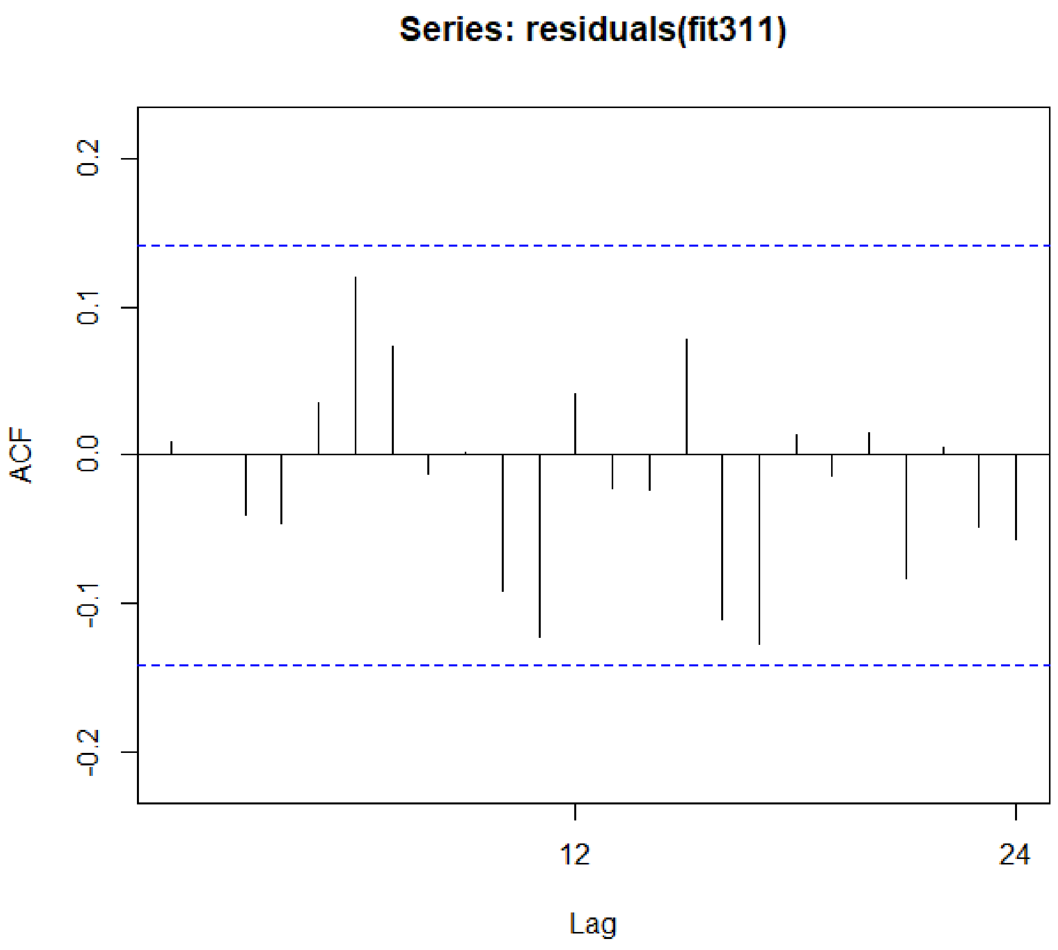
Observations on the Example

- The PACF shown suggests an AR(3) model so an initial candidate model is an ARIMA(3,1,0)
- Fit an ARIMA(3,1,0) along with other variations including ARIMA(4,1,0), ARIMA(2,1,0), ARIMA(3,1,1), etc.
- Check smallest AICc value

R Code: Fitting the ARIMA Model

```
• #Fitting or identifying a model
• fit300 <- Arima(eeadj, order=c(3,0,0))
• fit410 <- Arima(eeadj, order=c(4,1,0))
• fit210 <- Arima(eeadj, order=c(2,1,0))
• fit311 <- Arima(eeadj, order=c(3,1,1))
• summary(fit300)
• summary(fit410)
• summary(fit210)
• summary(fit311)
```

R Code: Fitting the ARIMA Model



R Code: Fitting the ARIMA Model

```
> summary(fit311)
Series: eeadj
ARIMA(3,1,1)

Coefficients:
         ar1     ar2     ar3      ma1
      0.0519  0.1191  0.3730  -0.4542
s.e.  0.1840  0.0888  0.0679   0.1993

sigma^2 estimated as 9.532:  log likelihood=-484.08
AIC=978.17   AICc=978.49   BIC=994.4

Training set error measures:
                      ME     RMSE      MAE         MPE     MAPE      MASE        ACF1
Training set -0.001227744 3.079373 2.389267 -0.04290849 2.517748 0.2913919 0.008928478
```

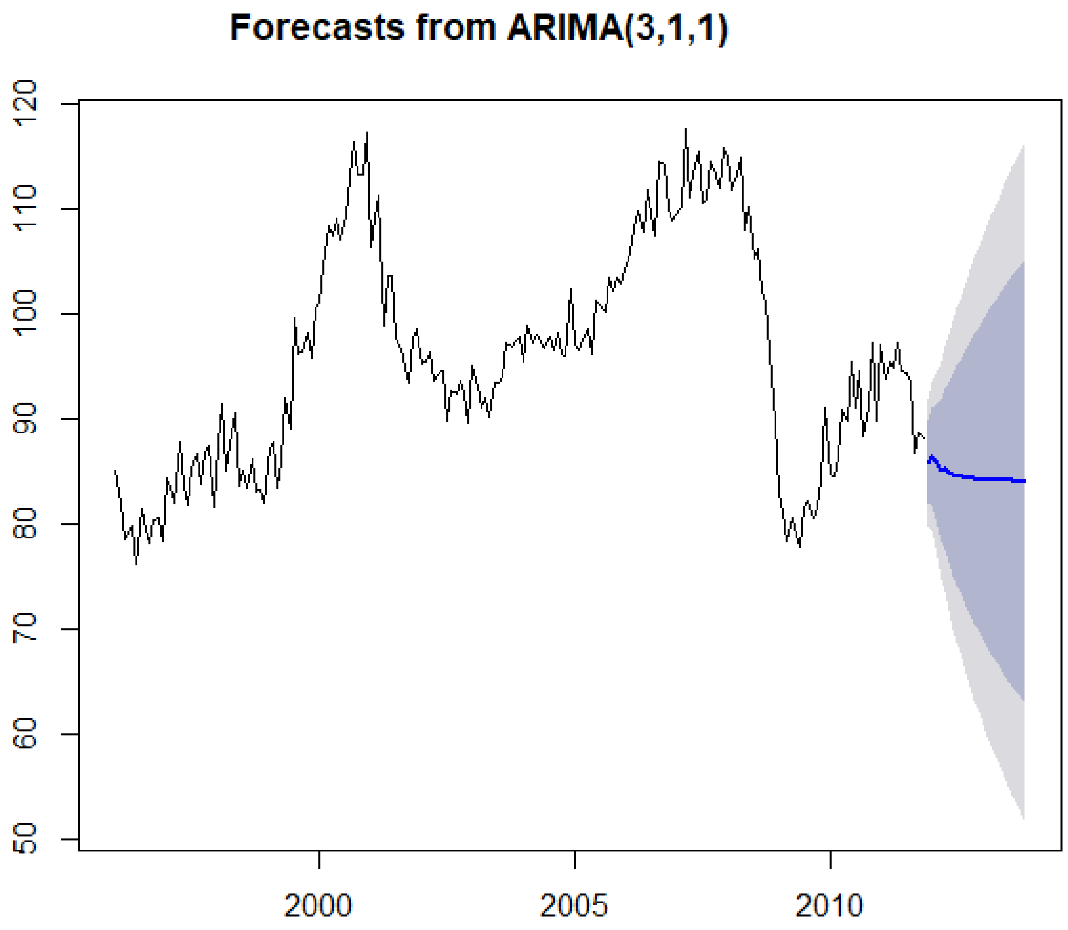
Observations on the Example

- The ACF plot of the residuals from the ARIMA(3,1,1) shows all correlations are within the threshold limits indicating that the residuals are somewhat like white noise
- Then, we can use the model as a representative of the time series as well as for forecasting

R Code: Fitting the ARIMA Model

- `#Fitting or identifying a model`
- `plot(forecast(fit311))`
- `forecastvalues <- forecast(fit311)`
- `forecastvalues`

R Code: Fitting the ARIMA Model



R Code: Fitting the ARIMA Model

- `#Using auto.arima`
- `auto.arima(eeadj)`

R Code: Fitting the ARIMA Model

```
> auto.arima(eeadj)
Series: eeadj
ARIMA(3,1,1)

Coefficients:
         ar1     ar2     ar3      ma1
      0.0519  0.1191  0.3730  -0.4542
s.e.  0.1840  0.0888  0.0679   0.1993

sigma^2 estimated as 9.532:  log likelihood=-484.08
AIC=978.17   AICc=978.49   BIC=994.4
```

Final Note

1. Forecasting makes a statement about the future but it does not tell us the actual future.
2. Forecasting needs regular monitoring and adjustments.
3. Forecasting should help the organization, the business, and the team decide on strategic decisions.
4. Forecasting assists in decision making.
5. Forecasting is just a subset of Business Analytics.

ralagria@up.edu.ph

References

- Notes and Datasets from Montgomery, Peck and Vining, Introduction to Linear Regression Analysis 4th Ed. Wiley
- Notes from G. Runger, ASU IEE 578
- Trevor Hastie, Rob Tibshirani, Friedman: Elements of Statistical Learning (2nd Ed.) 2009
- Time Series Data Library: Australian Bureau of Statistics
- http://datamarket.com/data/list/?q=provider:tsdl
- For R: http://robjhyndman.com/hyndsight/r/
- Time Series Data in R - http://www.rdatamining.com

References

- Duong Tuan Anh, Faculty of Computer Science and Engineering, September 2011
- http://faculty.wiu.edu/F-Dehkordi/DS-533/Lectures/Multiple%20Regression.ppt
- http://faculty.wiu.edu/F-Dehkordi/DS-533/Lectures/The%20Box-Jenkins%20Methodology%20for%20RIMA%20Models.ppt
- www.cse.hcmut.edu.vn/~dtanh/download/TS_PartI_new.ppt

R

- A free software environment for statistical computing and graphics
- http://cran.r-project.org/web/views/TimeSeries.html

R

- R: http://cran.r-project.org/bin/windows/base/R-3.1.3-win.exe
- R Studio: http://download1.rstudio.org/RStudio-0.98.1103.exe
- Make sure to install R first before R Studio which is a GUI of R.
- R is the most widely used Data Mining tool since it is fast, comprehensive and free.