AQ061-3-M-ODL-TSF Time Series Analysis and Forecasting

Topic 4 – Box Jenkins Methodology

TOPIC LEARNING OUTCOMES

At the end of this topic, you should be able to:

1. Use Box Jenkins methodology to produce accurate forecasts based on a description of historical patterns in the data.
2. Solve the model using computer software and interpret the results.

Contents & Structure

- Autoregressive (AR)
- Moving Average (MA)
- Autoregressive Moving Average (ARMA)
- Autoregressive Integrated Moving Average (ARIMA)
- Building ARIMA Models
- Seasonal Auto Regressive Integrated Moving Average (SARIMA)
- Building SARIMA Models

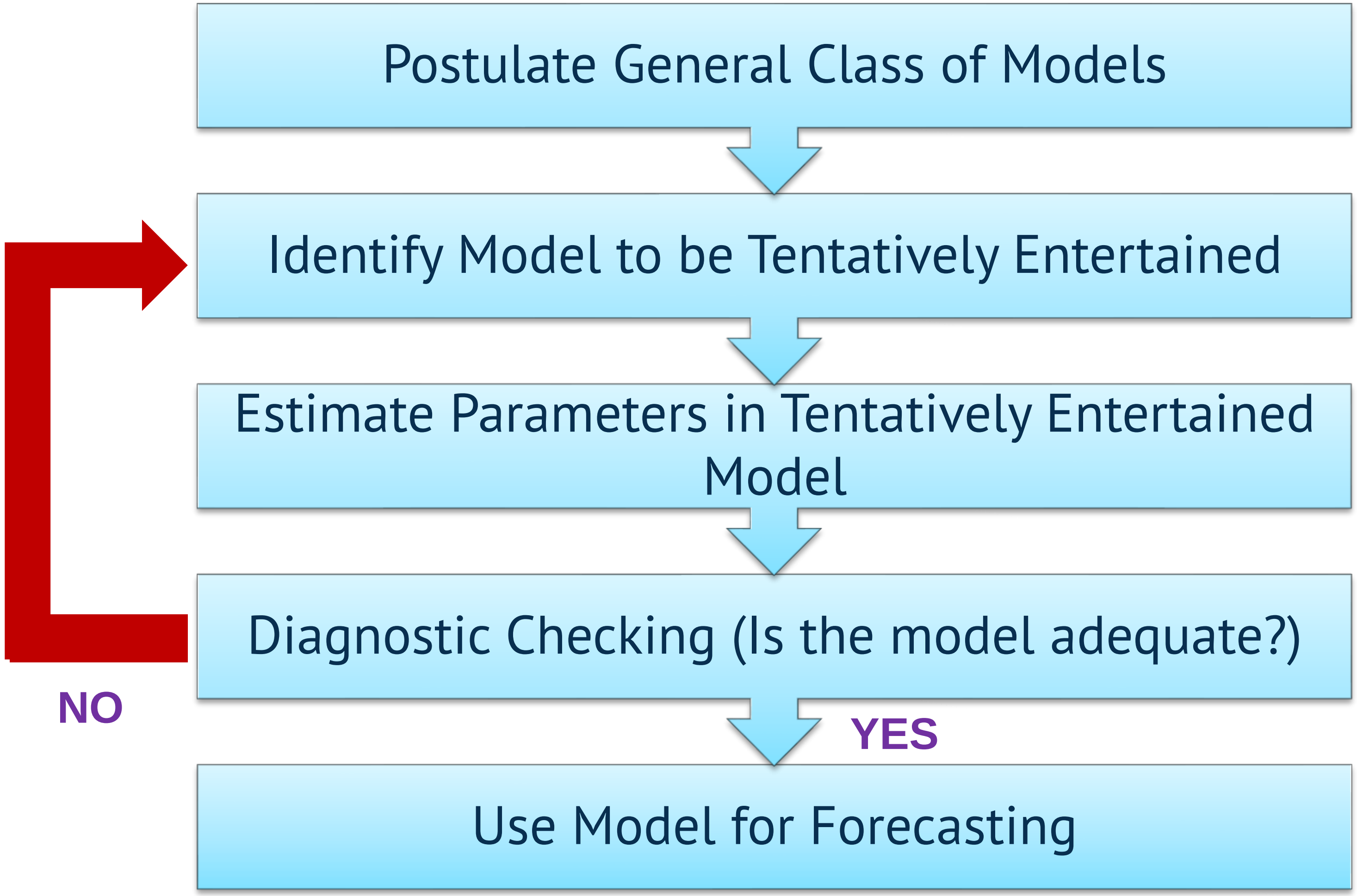
Recap From Last Lesson

- Questions to ask to trigger last week’s key learning points

Introduction

- The Box-Jenkins methodology refers to a set of procedures for identifying and estimating time series models within the class of AutoRegressive Integrated Moving Average (ARIMA) models.
- This models rely heavily on the autocorrelation pattern in the data.

Building ARIMA Models

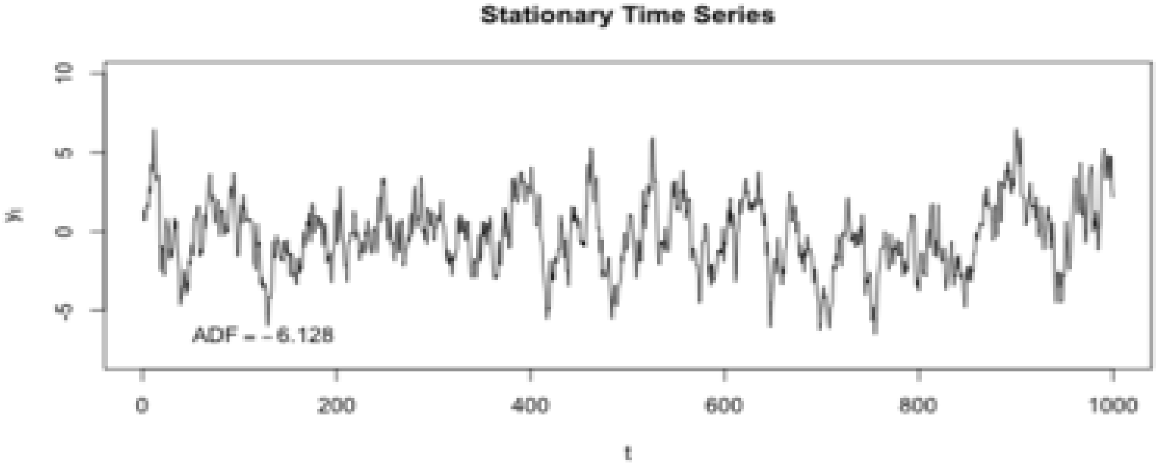


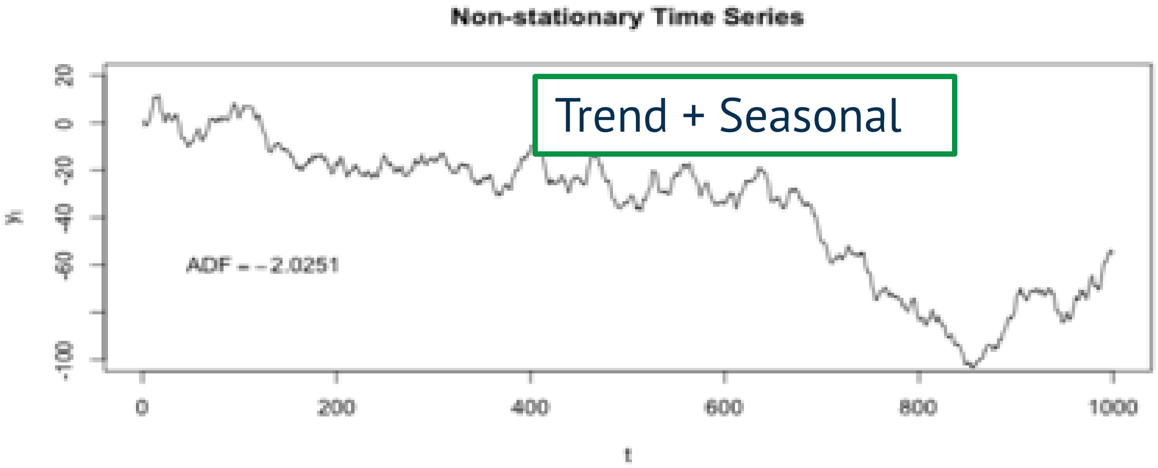
Properties of Stationary Series

Time series are stationary if they do not have trend or seasonal effects

1. $\mathrm{E}(Y_t) = \mu$
2. $\mathrm{Var}(Y_t) = \sigma^2$
3. $\mathrm{Cov}(Y_t, Y_{t-k}) = \gamma_k$
4. $\rho_k = \frac{\gamma_k}{\sigma^2}$

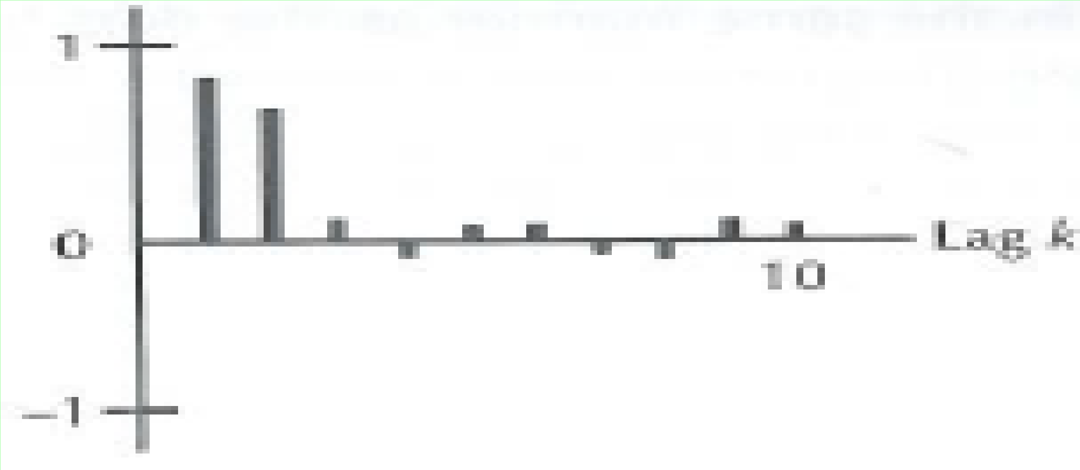
In other words, it has **constant mean and variance**, and covariance (and also correlation) between Y_t and Y_{t-1} is the same for all t.





Behaviors of ACF

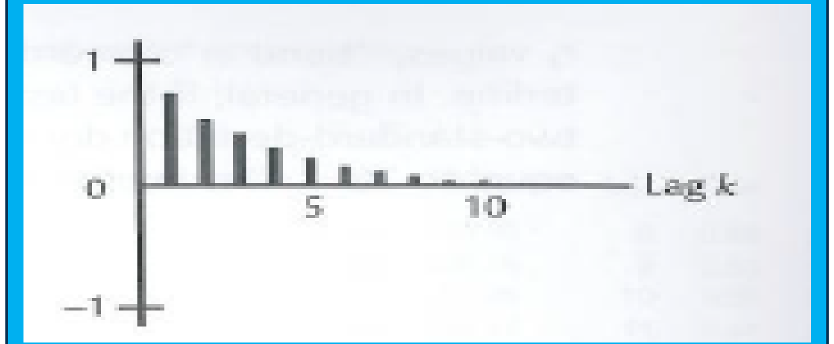
1. The ACF can cut off. A spike at lag ***k*** exists in the ACF if $\boldsymbol{r_k}$ is statistically large. The ACF cuts off after lag ***k*** if there are no spikes at lags greater than ***k*** in the ACF.



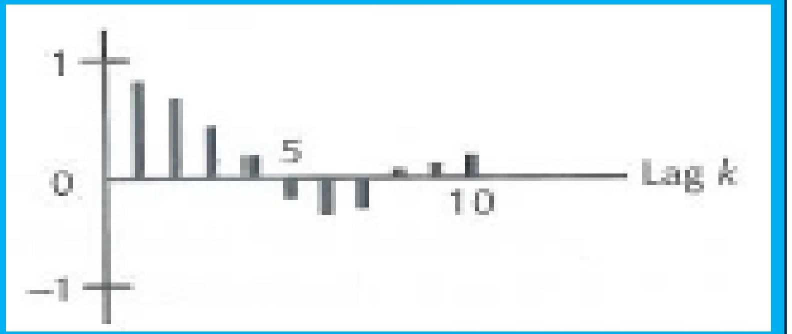
Cut off after lag 2

Behaviors of ACF

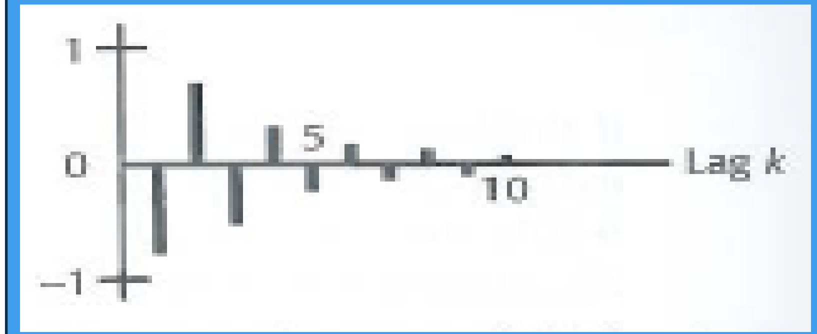
2. The ACF is said to die down if this function does not cut off but rather decreases in a ‘steady fashion’.



Damped exponential dying down



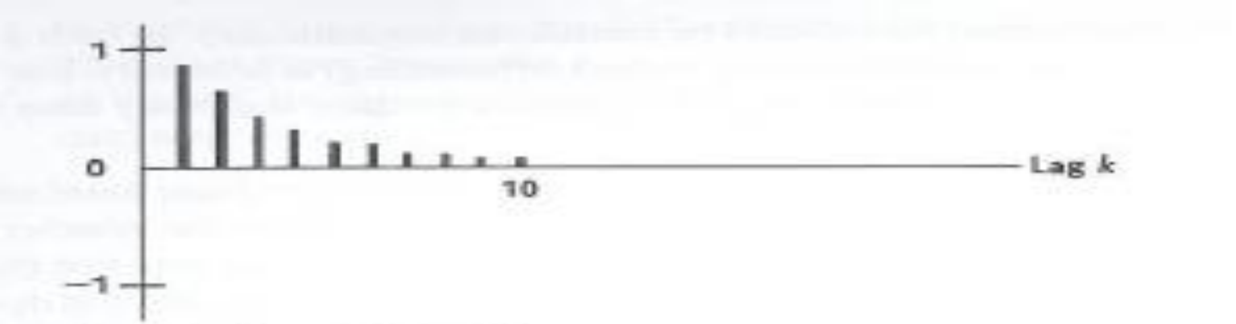
Damped sine-wave dying down



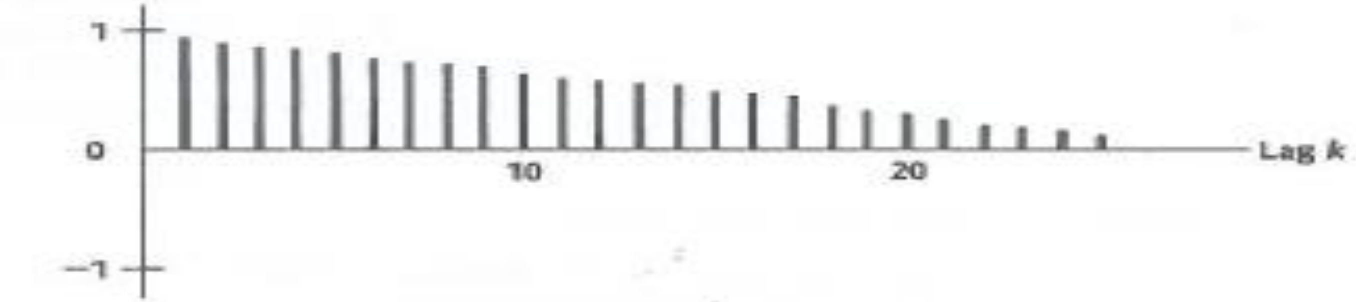
Damped exponential dying down with oscillation

Behaviors of ACF

3. The ACF can die down fairly quickly or extremely slowly.



(a) Dying down fairly quickly (stationary)



(b) Dying down extremely slowly (non-stationary)

Backshift Operator

- Backshift operator is defined as

$$BY_t = Y_{t-1}$$

- In other words, B operating on Y_t has the effect of shifting the data back one period.
- It can be extended,

$$B^k Y_t = Y_{t-k}$$

- The operator is convenient for describing the process of differencing, i.e.

$$(1 - B)^d Y_t$$

Building ARIMA Models

ARIMA(p,d,q)

$$\underbrace{\phi_p(B)}_{\text{Regular AR(p)}}\nabla^d Y_t = \delta + \underbrace{\theta_q(B)}_{\text{Regular MA(q)}}\varepsilon_t$$

$\nabla^d = (1-B)^d$

δ = constant

Y_t = time series data

ε_t = white noise/random error

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$

$\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$

Moving Average (MA)

Moving Average (MA) Model

- The model

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

  is called non-seasonal moving average model of order q.

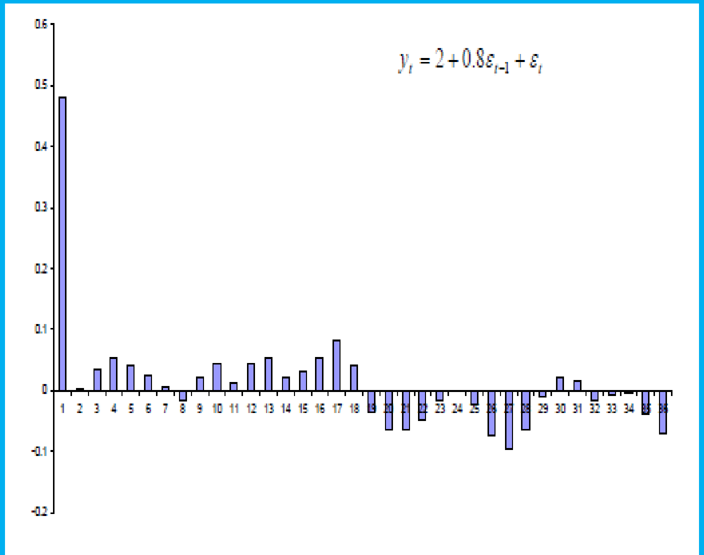
- Denote this process as MA(q).
- The process is described completely by a weighted sum of current and lagged random disturbances.
- $\theta_1, \theta_2, \ldots \theta_p$ are unknown parameter.

Moving Average (MA)

MA(1) Model

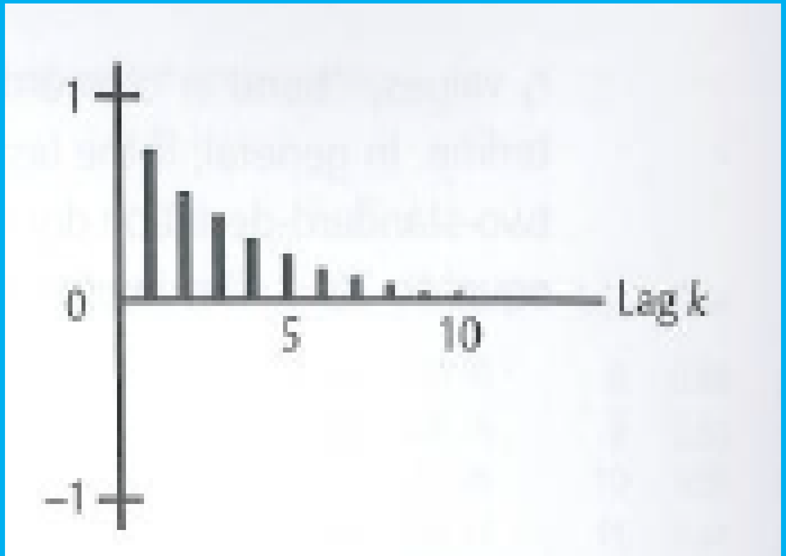
$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

ACF



Cuts off after lag 1

PACF



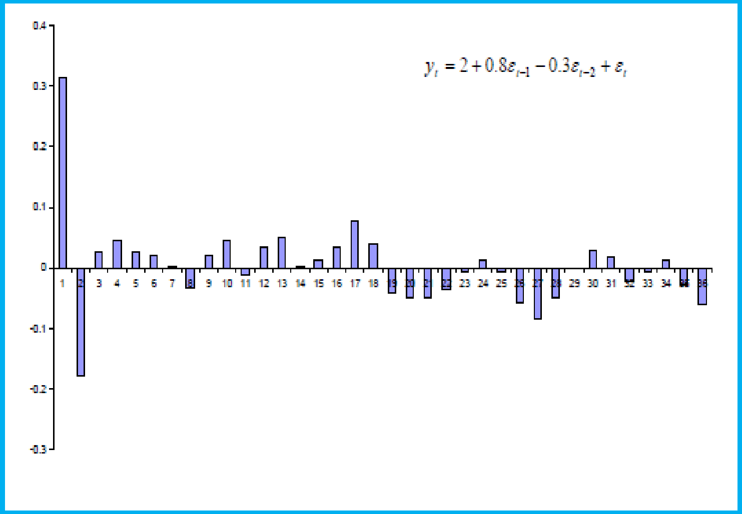
Damped exponential dying down

Moving Average (MA)

MA(2) Model

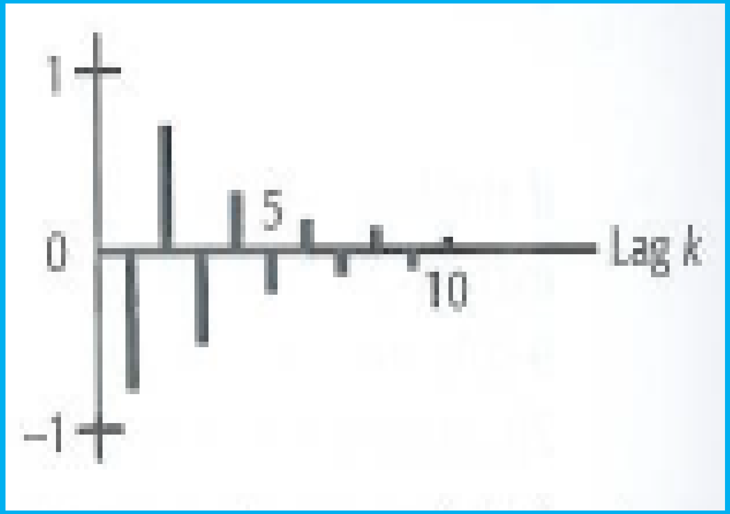
$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$

ACF



Cuts off after lag 2

PACF



Damped exponential dying down with oscillation

Example

Table below shows the result of ARIMA modeling

	Estimates
Constant (Mean)	6.957
MA Lag 1, θ_1	0.765
MA Lag 2, θ_2	0.997
Difference	1

Based on the observation below, <u>forecast the value at period 5</u> if period 4 is the forecast origin assuming $F_1 = 6.957$

Time	**1**	**2**	**3**	**4**
Observed	6	15	10	4

Autoregressive (AR)

Non-seasonal Autoregressive (AR) Model

- The model

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

  is called non-seasonal autoregressive model of order p.

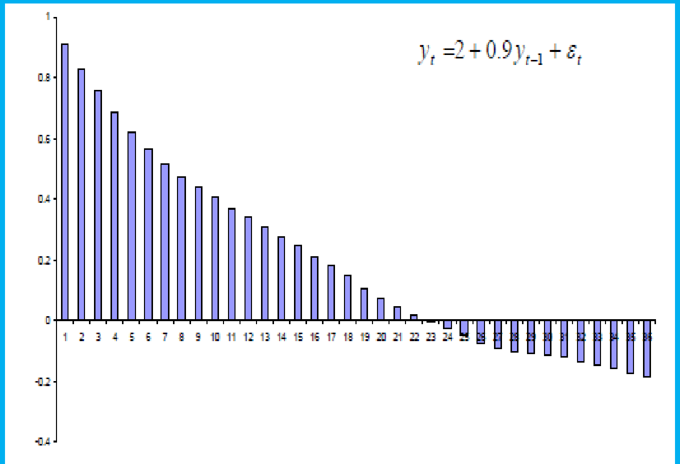
- Denote this process as AR(p)
- The process depends upon a weighted sum of its past values and a random disturbance in the current period .
- $\phi_1, \phi_2, \ldots \phi_p$ are unknown parameter

Autoregressive (AR)

AR(1) Model

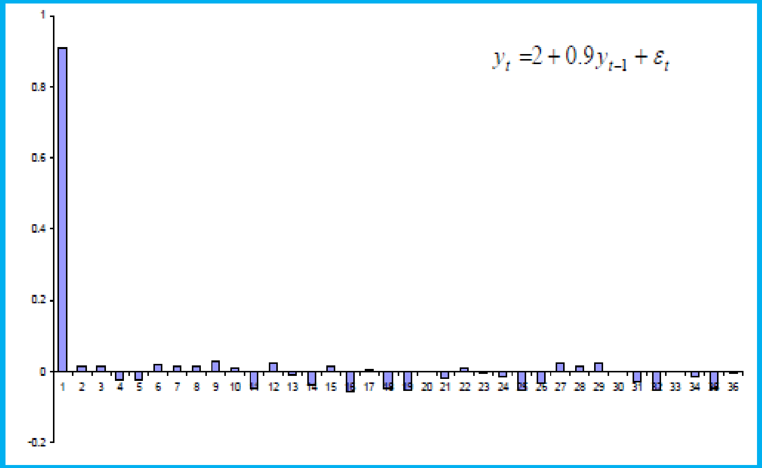
$$y_t = \phi_1 y_{t-1} + \delta + \varepsilon_t$$

ACF



Dies down in a damped exponential fashion

PACF



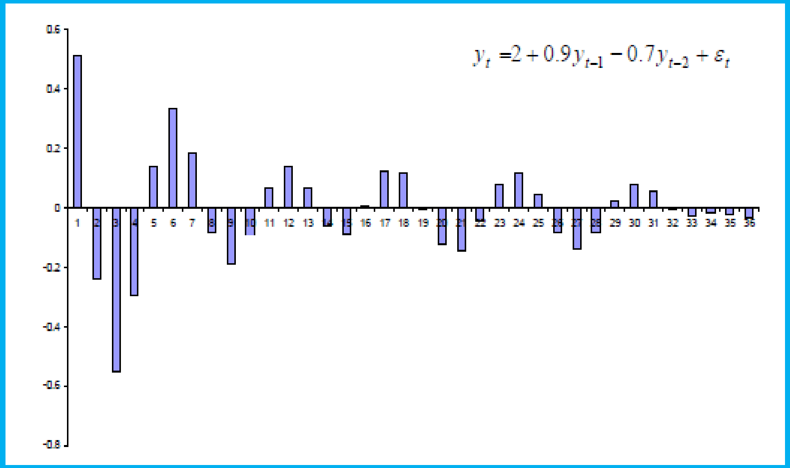
Cuts off after lag 1

Autoregressive (AR)

AR(2) Model

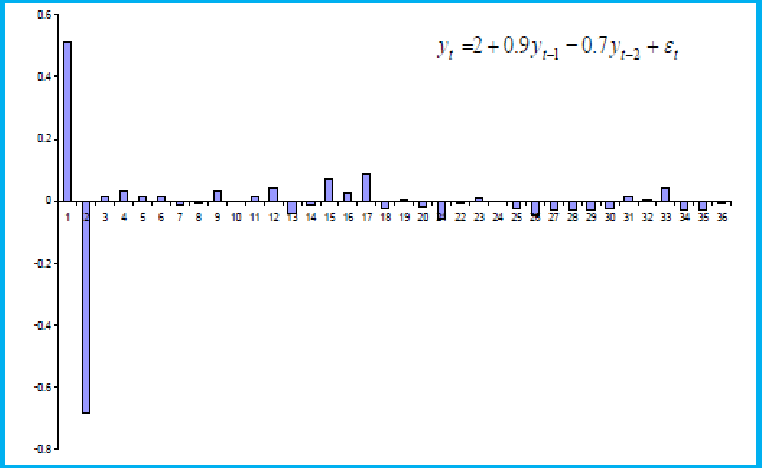
$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \delta + \varepsilon_t$$

ACF



Sine waves dying down.

PACF



Cuts off after lag 2

Practical Exercise

Analyse the following data and formulate the model equation for the ARIMA model you chosen:

- quakes.dat
- population.csv – average growth of population from 1970 to 2017

Autoregressive Moving Average (ARMA)

Non-seasonal Mixed Autoregressive Moving Average (ARMA) Model

- The model

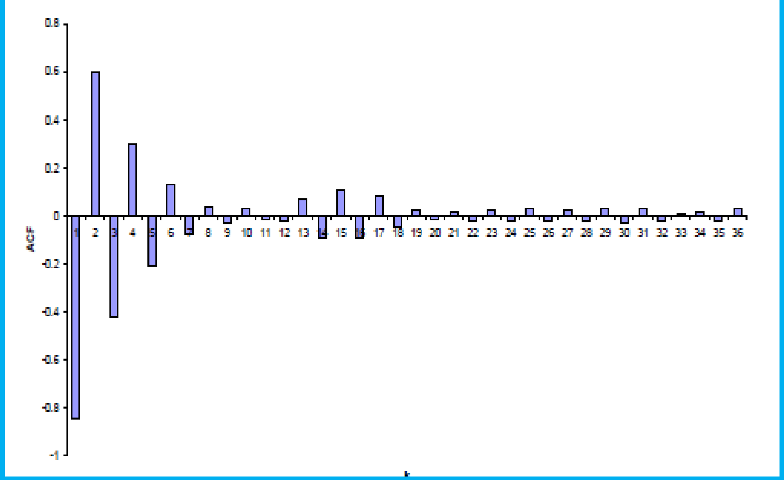
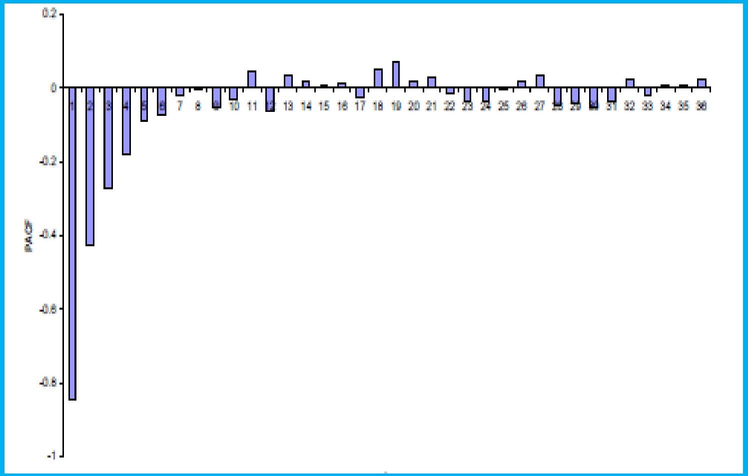
$$\begin{aligned} y_t = \delta + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \end{aligned}$$

  is called non-seasonal mixed autoregressive – moving average model of order (p,q).
- Denote this process as ARMA(p,q)
- Combine features of both MA and AR processes

Autoregressive Moving Average (ARMA)

- **ARMA(1,1) Process**

$$y_t = \delta + \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

ACF	PACF
	
Dies down in a damped exponential fashion with oscillation	Dies down in a fashion dominated by damped exponential decay

Example

Formulate the model equation based on the output below:

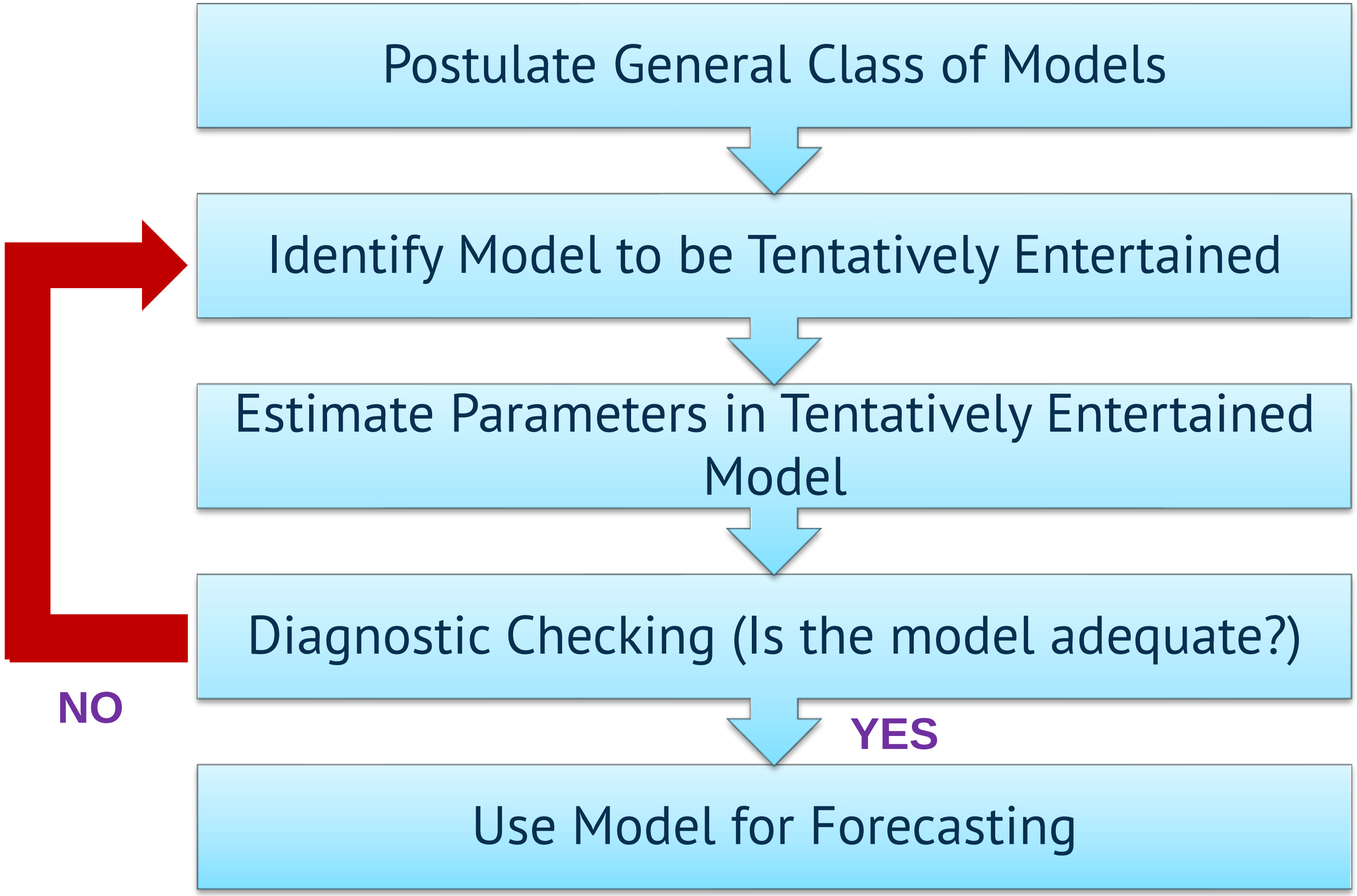
```
ARIMA(1,1,1)
Coefficients:
         ar1      ma1
      0.7713  -0.4422
s.e.  0.1887   0.2639

sigma^2 estimated as 22874248:
log likelihood=-276.07
AIC=558.15   AICc=559.15   BIC=562.14
```

Autoregressive Integrated Moving Average (ARIMA)

ARIMA (p,d,q)

- Models for non-stationary series are called ***autoregressive integrated moving average*** models and denoted by **ARIMA (p,d,q)**
  - **p** indicate the order of AR part
  - **d** indicate the amount of differencing
  - **q** indicate the order of MA part
- If the original series is stationary, then d=0 and the ARIMA models reduce to ARMA models

Building ARIMA Models



Building ARIMA Models

Parameter Estimation

- Once a tentative model has been selected, the parameter for that model must be estimated.
- The parameter in models are estimated by **minimizing the sum of squares of the fitting errors.**

Building ARIMA Models

Parameter Estimation

- Once the least squares estimates and their standard errors are determined, *t* values can be constructed and interpreted in the usual way such as

$$t = \frac{\text{Point estimate of each parameter}}{\text{standard error of the point estimate}}$$

$$t = \frac{\hat{\theta}}{S_{\hat{\theta}}}$$

Building ARIMA Models

Parameter Estimation

- Parameters that are judged significantly different from zero are retained in the fitted model (If p-value < 0.05, Reject H_0).
- Parameters that are not significant are dropped from the model.

$$\text{Null hypothesis, } H_0:\ \theta = 0$$

$$\text{Alternative hypothesis, } H_1:\ \theta \neq 0$$

Building ARIMA Models

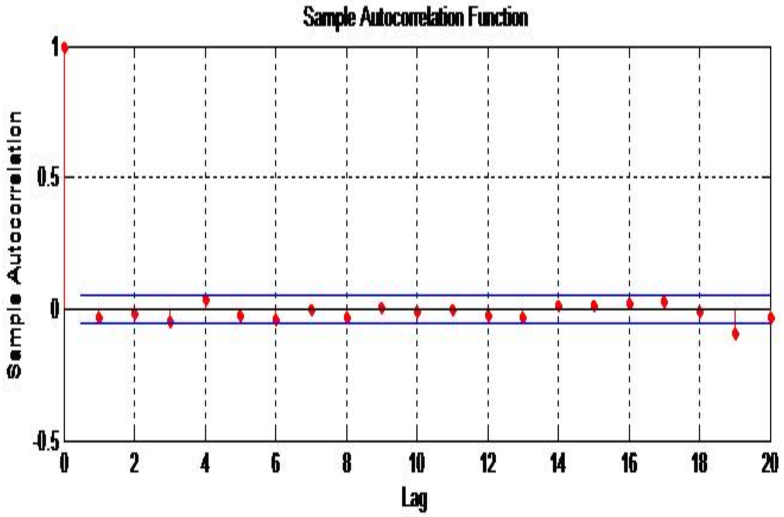
Diagnostic Checking

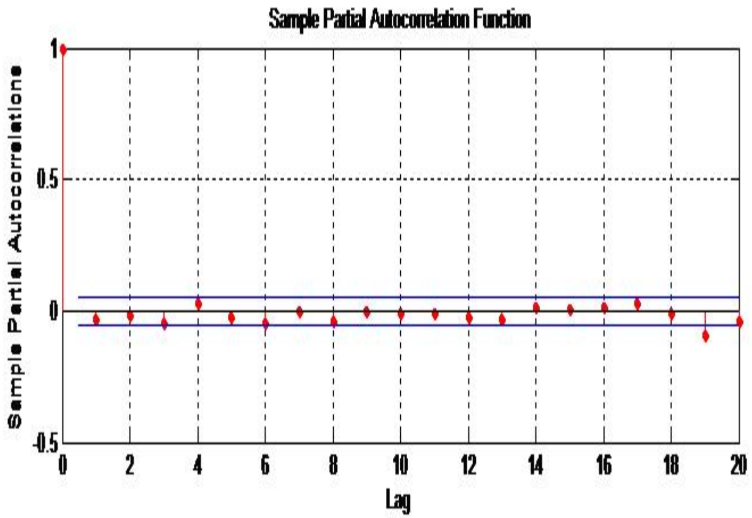
- Check for adequacy of the model.
- Often it is not straightforward to determine a single model that most adequately represents the data generating process, and it is common to estimate several models at the initial stage.
- The model that is finally chosen is the one considered best based on a set of diagnostic checking criteria. These criteria include
    1. t-tests for coefficient significance
    2. residual analysis
    3. model selection criteria

Building ARIMA Models

White Noise Process

- In general, we assume the error term, ε_t is uncorrelated with anything, with **mean 0** and **constant variance, σ^2**. We called this process as White Noise process.





Building ARIMA Models

Diagnostic Checking

- An overall check of model adequacy is provided by a chi-square test based on the Ljung-Box Q statistic.

$$Q = n(n+2)\sum_{k=1}^{m}\frac{r_k^2(e)}{n-k}$$

$r_k(e)$ = residual autocorrelation at lag k

n = number of residuals

k = time lag

m = number of time lags to be tested

Building ARIMA Models

Diagnostic Checking

- If **p-value is small (< 0.05),** the model is considered **inadequate**.
- Then, the analyst should consider a new or modified model and continue the analysis until a satisfactory model has been determined.

Building ARIMA Models

- Once an adequate model has been found, forecasts for one period or several periods into the future can be made.
- Computer programs that fit ARIMA models generate forecasts and prediction intervals at the analyst's request.
- As more data become available, the same ARIMA model can be used to generate revised forecast from another time origin.
- Good to monitor forecast errors. If the forecast error tend to be consistently positive (under predicting) or negative (over predicting).

Seasonal Autoregressive Integrated Moving Average (SARIMA)

- Often time series possess a seasonal component that repeats every observations.
- In order to deal with seasonality, ARIMA processes have been generalized and SARIMA models have then been formulated.
- SARIMA is known as is Seasonal AutoRegressive Integrated Moving Average.

Seasonal Autoregressive Integrated Moving Average (SARIMA)

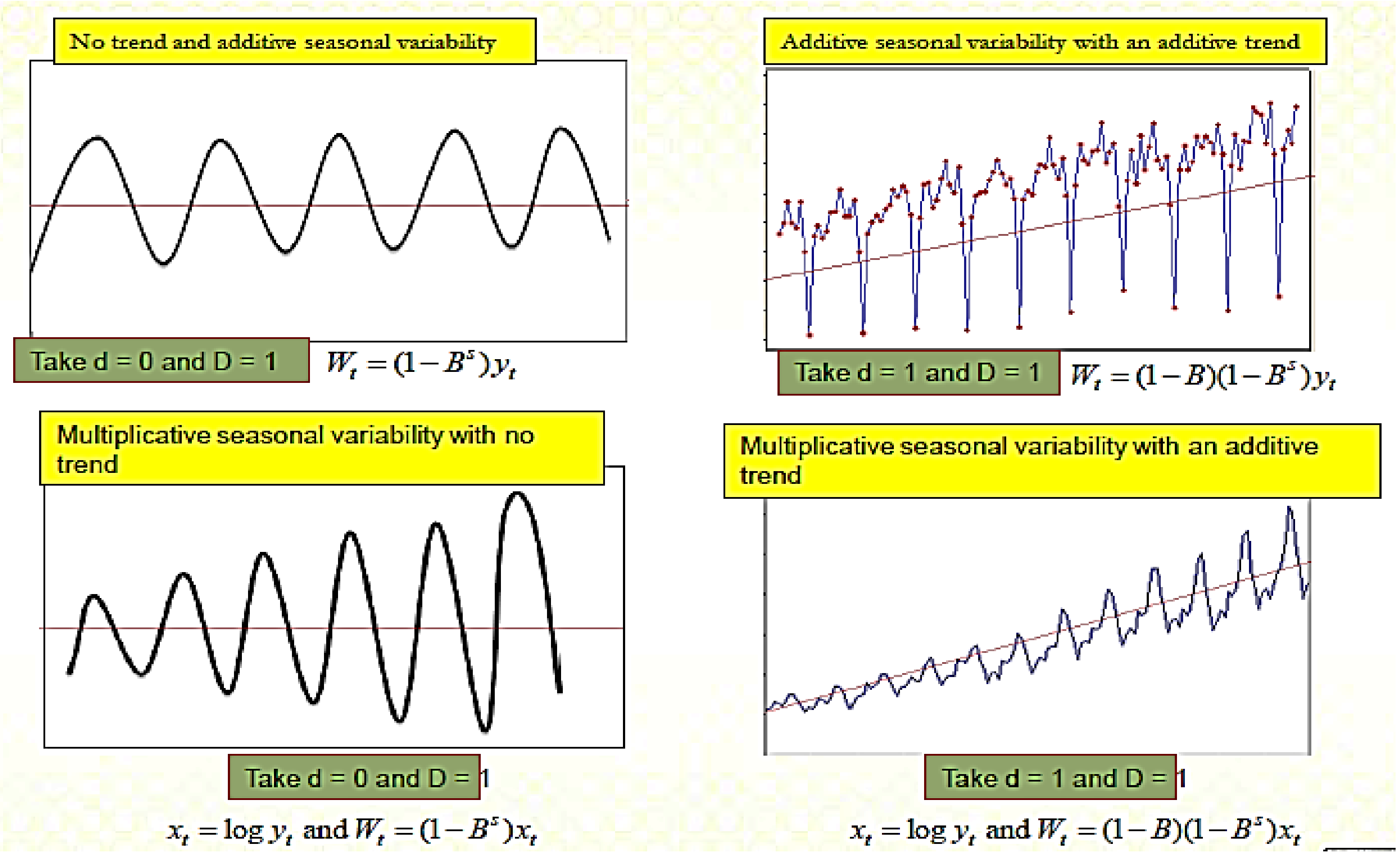
- The Box-Jenkins methodology for modeling seasonal data is no different to that from non-seasonal data. Consists of:
  - Stationary
  - Select an initial model
  - Estimate the model coefficients
  - Analyse the residuals
  - Forecasting
- The slight change introduced by seasonal data of period k is that the seasonal coefficients of the ACF and PACF appear at lags $k, 2k, 3k, \ldots$, rather than at lags $1, 2, 3, \ldots$

Seasonal Autoregressive Integrated Moving Average (SARIMA)

- **Seasonal (periodic) model with S observations per period.**
  - Monthly data has 12 observations per year (S = 12)
  - Quarterly data has 4 observations per year (S = 4)
  - Daily data has 5 or 7 (or some other number) of observations per week (S = 5 or 7)
- **Stationary**
  - General way to transform non-stationary to stationary series is given as:

$$(1 - B)^d (1 - B^S)^D Y_t$$

Seasonal Autoregressive Integrated Moving Average (SARIMA)



Example

Seasonal MA model:

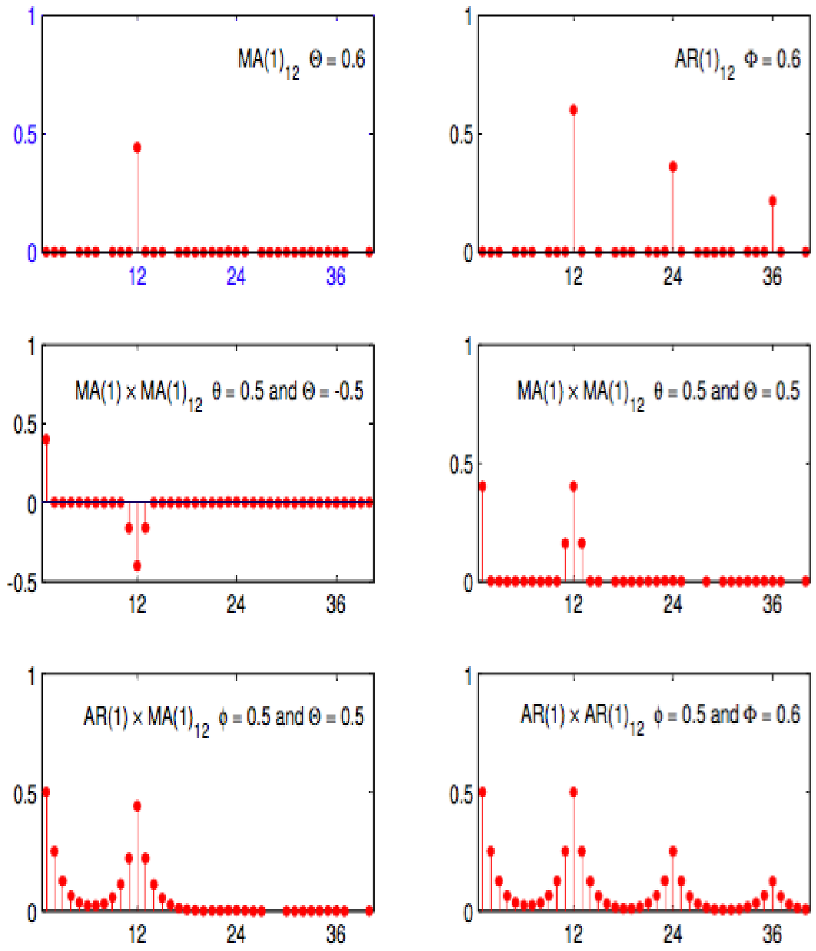
- $ARIMA(0,0,0)(0,0,1)_{12}$
  - will show a spike at lag 12 in the ACF but no other significant spikes.
  - The PACF will show exponential decay in the seasonal lags i.e. at lags 12, 24, 36,...

Seasonal AR model:

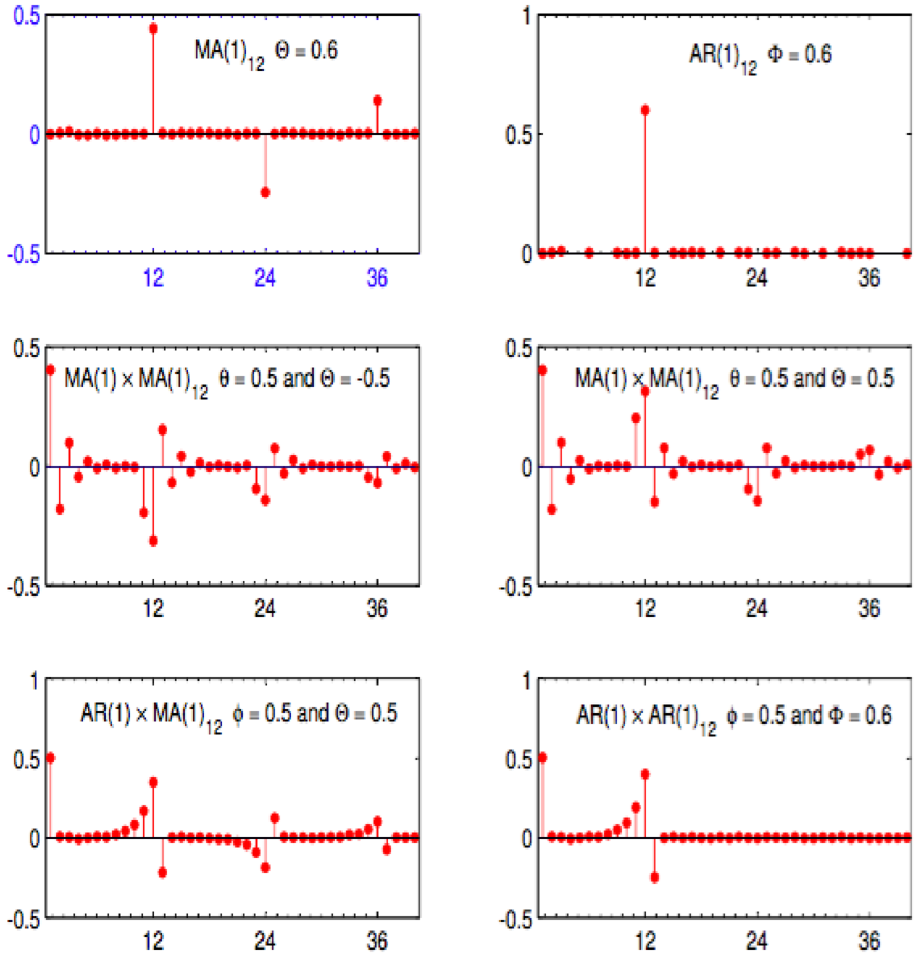
- $ARIMA(0,0,0)(1,0,0)_{12}$
  - will show exponential decay in seasonal lags of the ACF.
  - Single significant spike at lag 12 in the PACF.

Seasonal Autoregressive Integrated Moving Average (SARIMA)

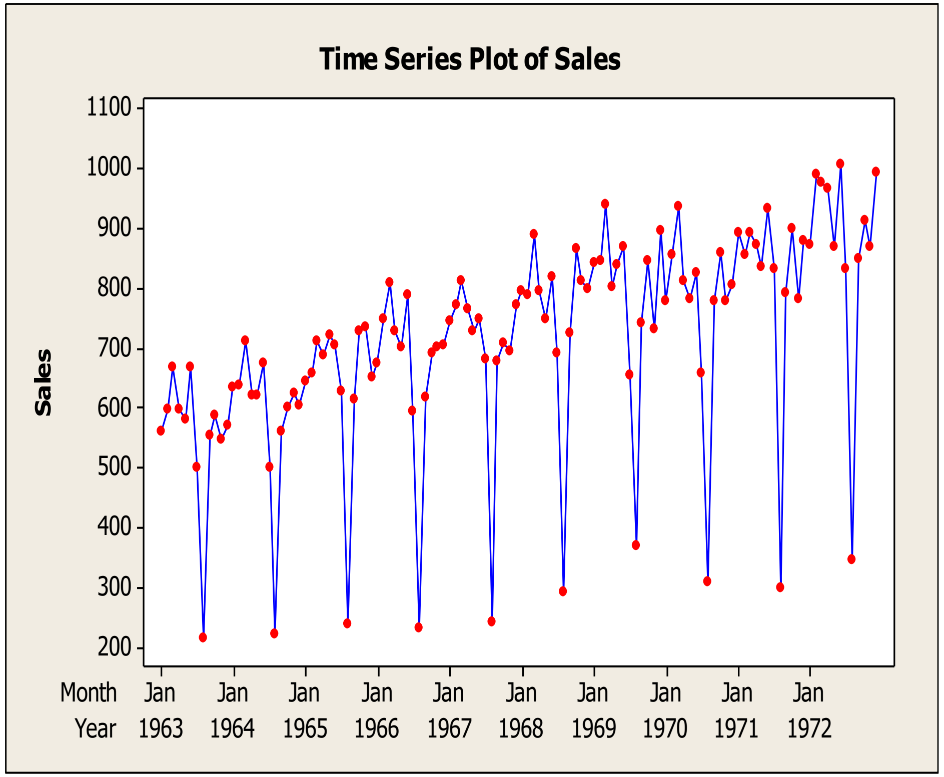
Simple autocorrelation function - Examples

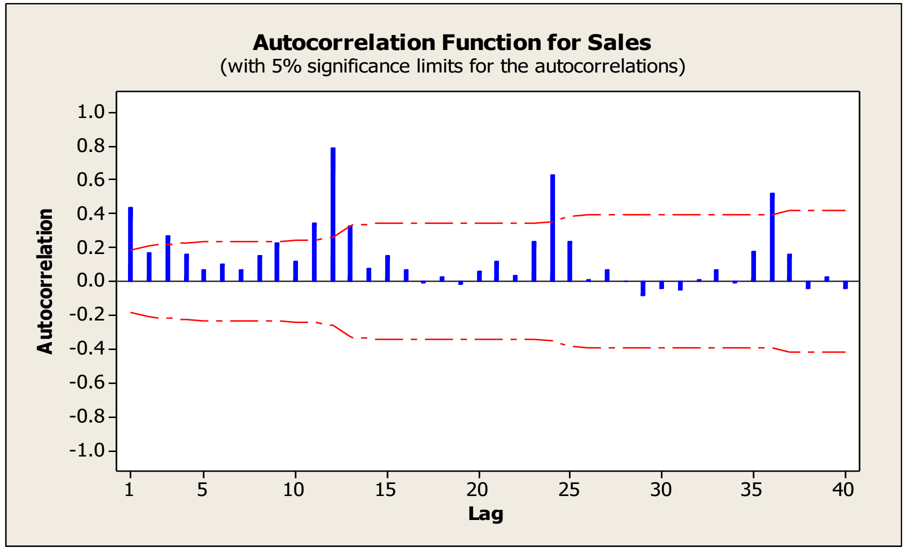


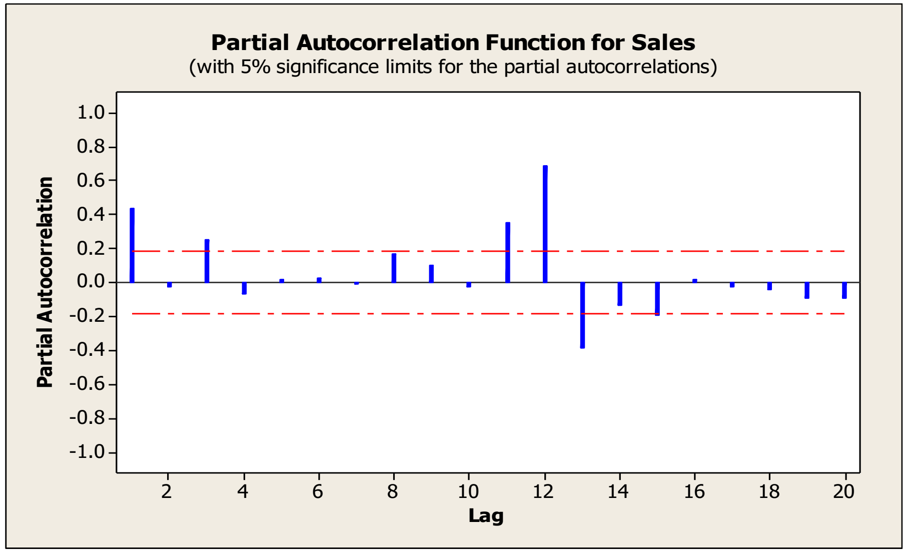
Partial autocorrelation function - Examples



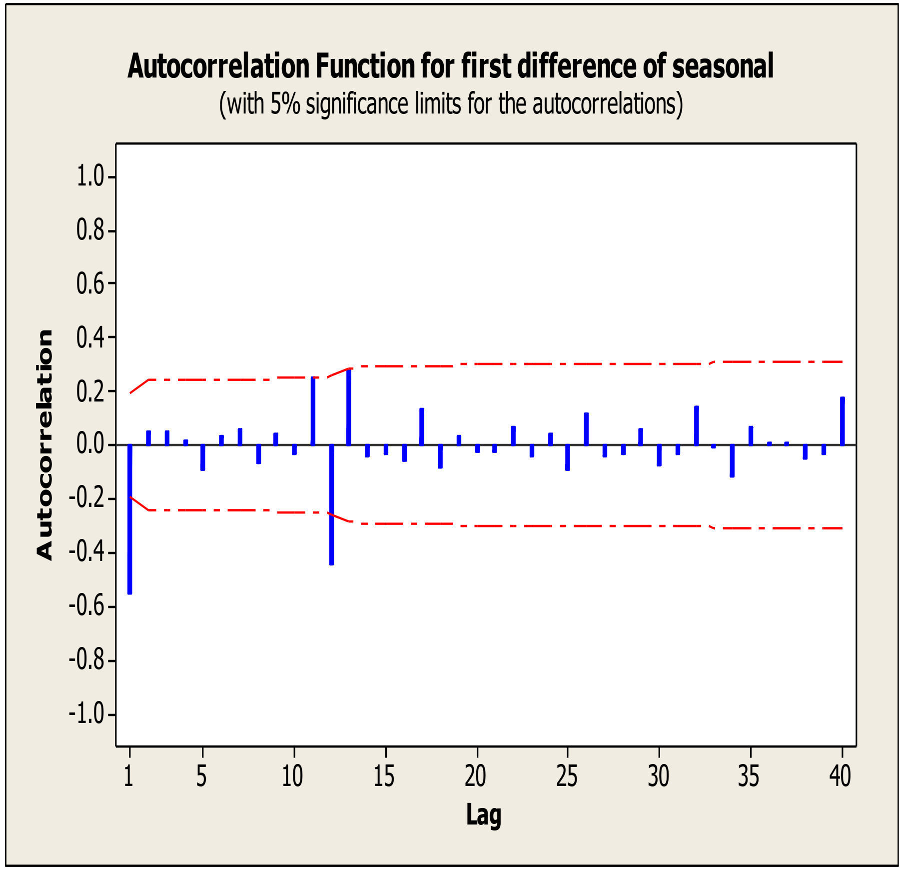
Seasonal Autoregressive Integrated Moving Average (SARIMA)

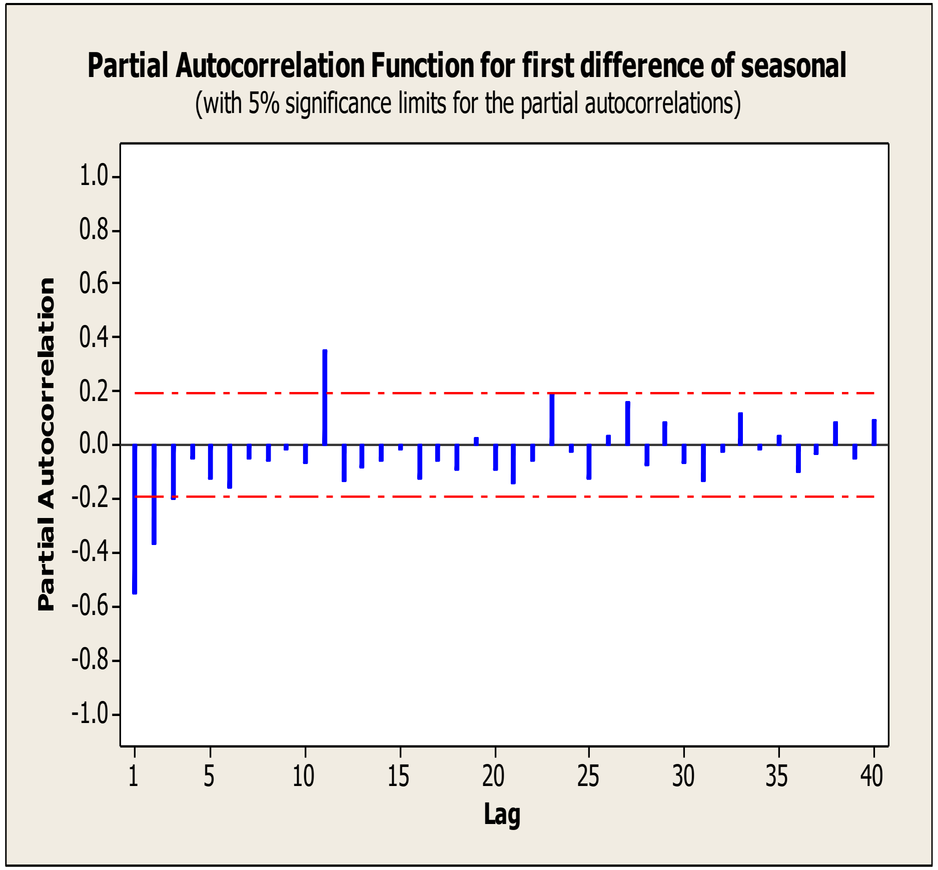






Seasonal Autoregressive Integrated Moving Average (SARIMA)





Seasonal Autoregressive Integrated Moving Average (SARIMA)

- 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).
- **ARIMA $(0,1,1)(0,1,1)_{12}$.**

Seasonal Autoregressive Integrated Moving Average (SARIMA)

- **SARIMA model is denoted by**

$$ARIMA(p,d,q)(P,D,Q)_S$$

  - **p** indicate the order of regular AR part
  - **d** indicate the regular amount of differencing
  - **q** indicate the order of regular MA part
  - **P** indicate seasonal AR part at period *S* (lag *S*)
  - **D** indicate seasonal difference at period *S*
  - **Q** indicate seasonal MA term at period *S*
  - ***S*** indicate seasonal period/lag

Seasonal Autoregressive Integrated Moving Average (SARIMA)

$$\underbrace{\phi_p(B)}_{\text{Regular AR(p)}}\underbrace{\Phi_P(B^S)}_{\text{Seasonal AR(p)}}\underbrace{\nabla_S^D}_{\text{Seasonal Differences}}\underbrace{\nabla^d}_{\text{Regular Differences}}Y_t = \delta + \underbrace{\theta_q(B)}_{\text{Regular MA(q)}}\underbrace{\Theta_Q(B^S)}_{\text{Seasonal MA(q)}}\varepsilon_t$$

Seasonal Autoregressive Integrated Moving Average (SARIMA)

- $\nabla^d = (1-B)^d$
- $\nabla_S^D = (1-B^S)^D$
- δ = constant
- Y_t = time series data
- ε_{t} = white noise process/random error
- $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p$
- $\theta_q(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$
- $\Phi_P(B^S) = 1 - \Phi_1 B^S - \Phi_2 B^{2S} - \ldots - \Phi_P B^{SP}$
- $\Theta_Q(B^S) = 1 + \Theta_1 B^S + \Theta_2 B^{2S} + \cdots + \Theta_Q B^{SQ}$

Example

Formulate the model equation based on the output below:

```
ARIMA(1,0,0)(0,0,1)[4] with non-zero mean

Coefficients:
         ar1     sma1       mean
      0.1051   0.8037  1630.9404
s.e.  0.1753   0.1650    76.6915

sigma^2 estimated as 61818:  log likelihood=-250.15
AIC=508.29   AICc=509.58   BIC=514.63
```

Example

Formulate the model equation based on the output below and test the model:

```
z test of coefficients:
     Estimate Std. Error z value Pr(>|z|)
ar1   0.51225    0.21535  2.3786  0.01738 *
ma1   0.23030    0.19510  1.1804  0.23784
sma1 -0.21569    0.20762 -1.0389  0.29886

Box-Pierce test
data:  fit1$residual
X-squared = 10.699, df = 5, p-value = 0.05768
```

Practical Exercise

Split the below data into training (80%) and testing data (20%). Analyse the training data and formulate the model equation for the ARIMA model you chosen:

- sales.dat – quarterly sales data (in $'000) starting 01-01-2007
- USABeerproduction.csv

Then, compute the accuracy of the model in the testing data. Check the residuals and test whether the model you chosen is satisfactory.

Review Questions

Summary / Recap of Main Points

1. Use Box Jenkins methodology to produce accurate forecasts based on a description of historical patterns in the data.
2. Solve the model using computer software and interpret the results.

What To Expect Next Week

In Class

Preparation for Class

- Volatile Models