Time Series Forecasting using RStudio

Course Taught at SUAD

Dr. Tanujit Chakraborty
Ms. Madhurima Panja
@ Sorbonne
tanujitisi@gmail.com

Introduction

- Time series is a set of observations, each one being recorded at a specific time. (e.g., Annual GDP of a country, Sales figure, etc.)
- Discrete time series is one in which the set of time points at which observations are made is a discrete set. (e.g., All above including irregularly spaced data)
- Continuous time series are obtained when observations are made continuously over some time intervals. (e.g., ECG graph)
- Forecasting is estimating how the sequence of observations will continue in to the future. (e.g., Forecasting of major economic variables like GDP, Unemployment, Inflation, Exchange rates, Production and Consumption)
- Forecasting is very difficult, since it’s about the future! (e.g., forecasts of daily cases of COVID-19)

Time Series Data

- A time series is a sequence of observations over time. What makes it distinguishable from other statistical analyses is the explicit recognition of the importance of the order in which the observations are made. Also, unlike many other problems where observations are independent, in time series observations are most often dependent.

- Why do we need special models for time series data?

  - Prediction of the future based on knowledge of the past (most important).
  - To control the process producing the series.
  - To have a description of the salient features of the series.

- Applications of time series forecasting
  - Economic planning
  - Sales forecasting
  - Inventory (stock) control
  - Exchange rate forecasting
  - Etc…

Use of Time-series Data

Use of time-series data

- To develop forecast model
  - What will be the rate of inflation in next year?

- To estimate dynamic causal effects
  - If the rate of interest increases, what will be the effect on the rates of inflation and unemployment in 3 months? in 12 months?
  - What is the effect over time on electronics good consumption due to a hike in the excise duty?

- Time dependent analysis
  - Rates of inflation and unemployment in the country can be observed over a time period.

Time-series Data

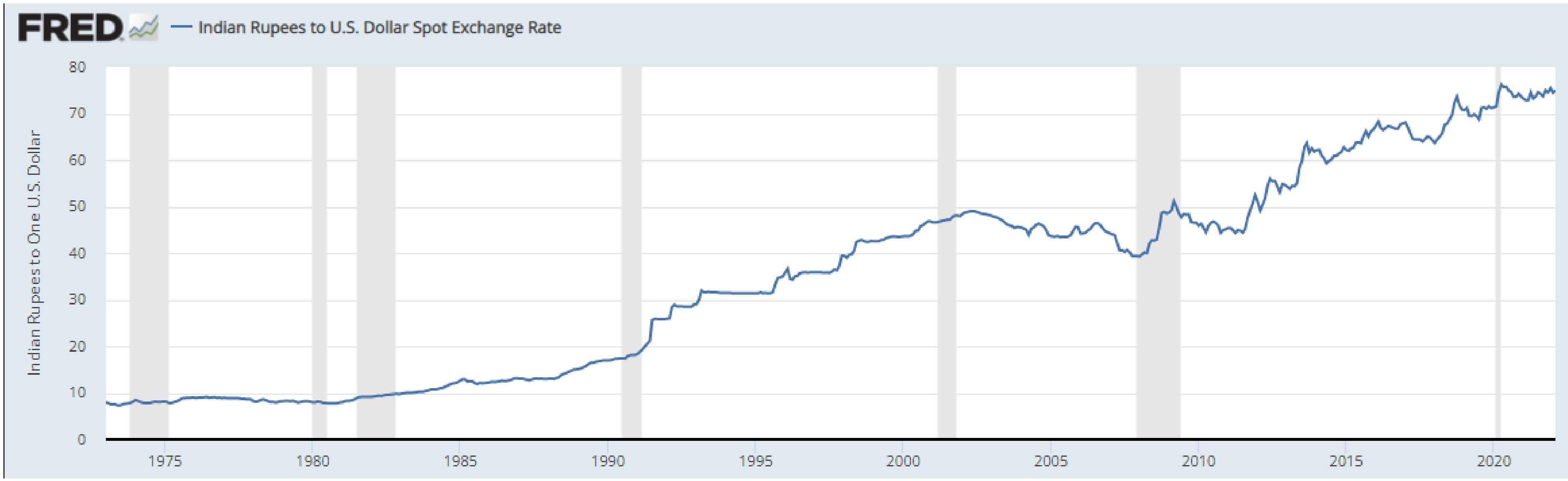
Time-series Data

Time-series data: The data collected on the same observational unit at multiple time periods

Example: Rate of price inflation

A Forecasting Problem: India / U.S. Foreign Exchange Rate (EXINUS)

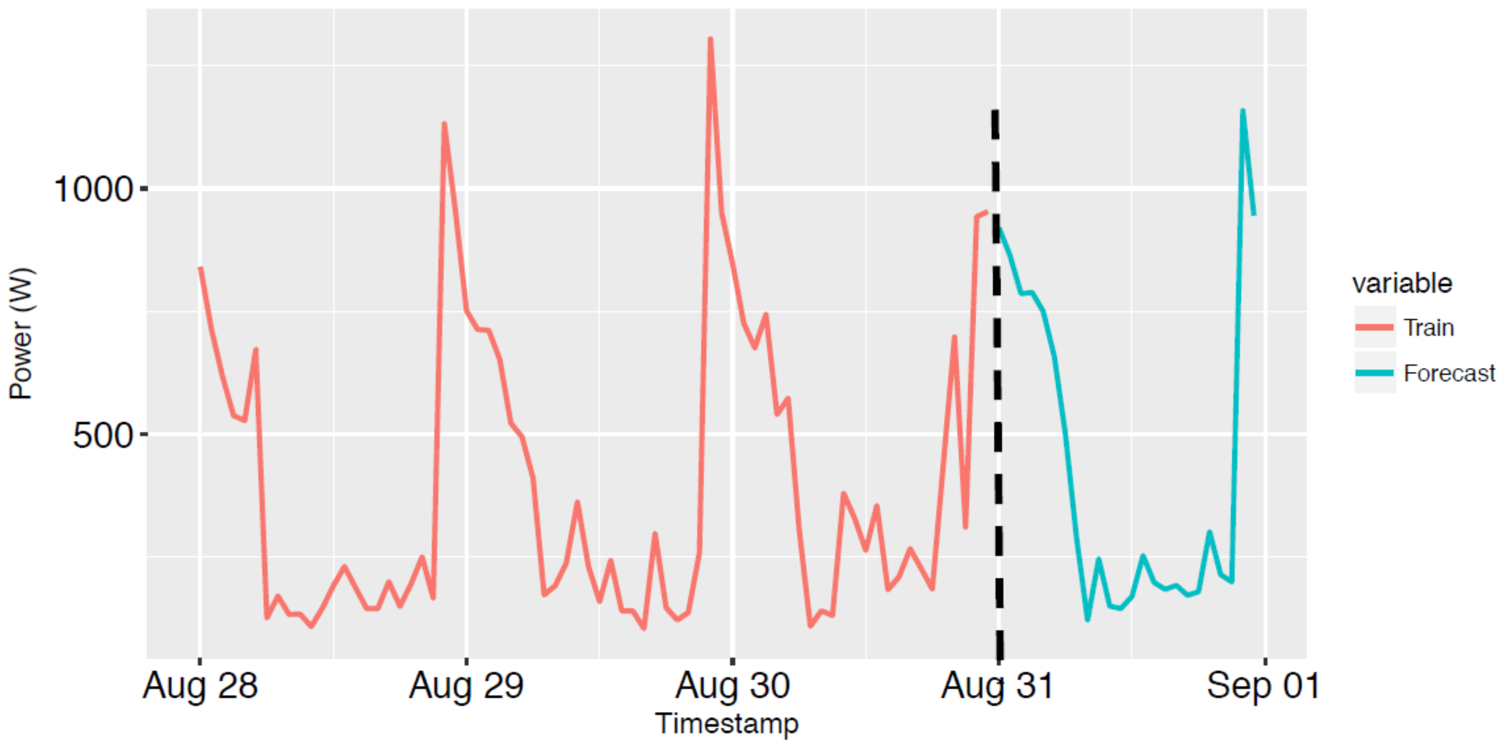
- Source: FRED ECONOMICS DATA (Shaded areas indicate US recessions)
- Units: Indian Rupees to One U.S. Dollar, Not Seasonally Adjusted
- Frequency: Monthly (Averages of daily figures)



Forecasting: Assumptions

- **Time series Forecasting:** Data collected at regular intervals of time (e.g., Weather and Electricity Forecasting).
- **Assumptions:** (a) Historical information is available;

  (b) Past patterns will continue in the future.



Time Series Components

- Trend (T_t) : pattern exists when there is a long-term increase or decrease in the data.
- Seasonal (S_t) : pattern exists when a series is influenced by seasonal factors (e.g., the quarter of the year, the month, or day of the week).
- Cyclic (C_t) : pattern exists when data exhibit rises and falls that are not of fixed period (duration usually of at least 2 years).
- Decomposition : $Y_t = f(T_t; S_t; C_t; I_t)$, where Y_t is data at period t and I_t is irregular component at period t.
- Additive decomposition: : $Y_t = T_t + S_t + C_t + I_t$
- Multiplicative decomposition: $Y_t = T_t * S_t * C_t * I_t$
- A stationary series is : roughly horizontal, constant variance and no patterns predictable in the long-term.

Auto-regression Analysis

Auto Regression Analysis

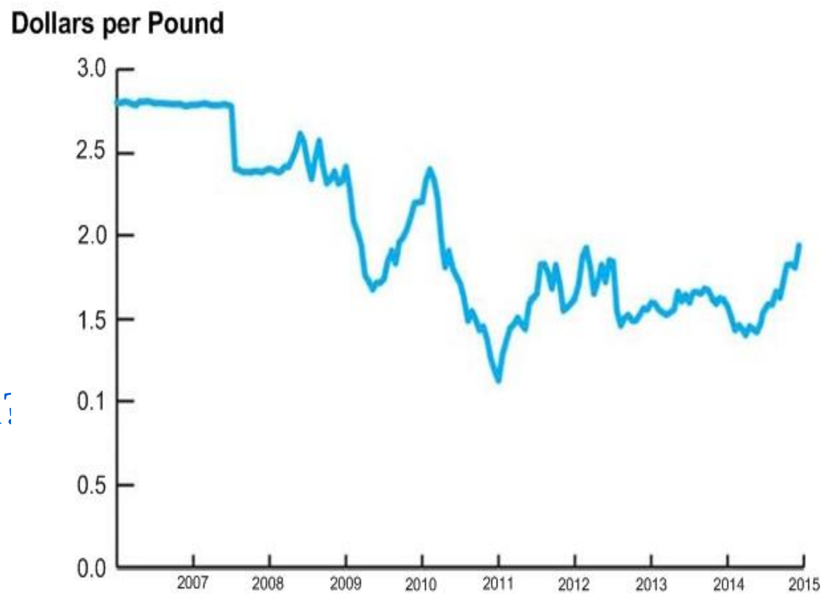
- Regression analysis for time-ordered data is known as Auto-Regression Analysis

- Time series data are data collected on the same observational unit at multiple time periods

Example: Indian rate of price inflation

Modeling with Time Series Data

- Correlation over time
  - Serial correlation, also called autocorrelation
  - Calculating standard error
- To estimate dynamic causal effects
  - Under which dynamic effects can be estimated?
  - How to estimate?
- Forecasting model
  - Forecasting model build on regression model



Can we predict the trend at a time say 2017?

Time-series Data

Examples of time-series data:

- Aggregate consumption and GDP for a country (for example, 20 years of quarterly observations = 80 observations)
- Yen/$, pound/$ and Euro/$ exchange rates (daily data for 1 year = 365 observations)
- Cigarette consumption per capita in a state, by years
- Rainfall data over a year or a period of years
- Sales of tea from a tea shop in a season

Concept and Notations

Some Notations and Concepts

- Y_t = Value of Y in a period t
- Data set $[Y_1, Y_2, \dots Y_{T-1}, Y_T]$: T observations on the time series random variable Y
- **Assumptions**
  - We consider only consecutive, evenly spaced observations
    - For example, monthly, 2000-2015, no missing months
  - A time series Y_t is **stationary** if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{i+1}, Y_{i+2}, \dots, Y_{i+T})$ does not depend on i.
    - Stationary property implies that history is relevant. In other words, Stationary requires the future to be like the past (in a probabilistic sense).
    - Auto Regression analysis assumes that Y_t is stationary.

Some Notations and Concepts

- There are four ways to have the time series data for AutoRegression analysis
  - **Lag:** The first lag of Y_t is Y_{t-1}, its j-th lag is Y_{t-j}
  - **Difference:** The fist difference of a series, Y_t is its change between period t and $t-1$, that is, $y_t = Y_t - Y_{t-1}$
  - **Log difference:** $y_t = \log(Y_t) - \log(Y_{t-1})$
  - **Percentage:** $y_t = \frac{Y_{t-1}}{Y_t} \times 100$

Related Concepts and Notations

Assumptions

1. **Uniform:** We consider only consecutive, evenly spaced observations
   - For example, say monthly data in 2010-2021 for each year, and without any missing month(s); no other data, for example, on daily basis for a year is admissible.

2. **Stationarity:** A time series Y_t is stationary if its probability distribution does not change over time, that is, if the joint distribution of $(Y_{i+1}, Y_{i+2}, Y_{i+3}, \ldots\ldots Y_{i+T})$ does not depend on i.

   - Stationary property implies that history is relevant. In other words, stationary requires the future to be like the past (in a probabilistic sense).

- Auto-regression analysis assumes that Y_t is **both** uniform and stationary.

Autocorrelation coefficient

Autocorrelation

The correlation of a series with its own lagged values is called autocorrelation (also called serial correlation)

Formula: j^{th} Autocorrelation

The j^{th} autocorrelation, denoted by ρ_j is defined as

$$\rho_j = \frac{Cov(Y_t, Y_{t-j})}{\sigma_{Y_t}\sigma_{Y_{t-j}}}$$

where, $COV(Y_t, Y_{t-j})$ is the j^{th} **auto-covariance**

Covariance

Formula: $COV(Yt, Yt\_{j})$

The covariance between the variables Y_t and $Y_{t\text{-}j}$

$$COV(Y_t, Y_{t-j}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

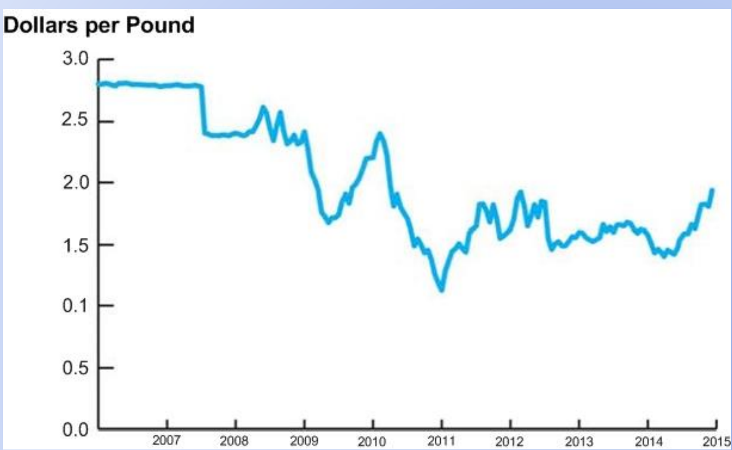
$Y_{t\text{-}j}$	. . .	Y_t
x_1		y_1
x_2		y_2
. . .		. . .
x_j		y_j
.		.
.		.
.		.
x_n		y_n

σ_X is the variance for the variable X

n is the number of observations

Example: Autocorrelation

Example



- For the given data, say $\rho_1 = 0.84$ between two given consecutive years
  - This implies that the Dollars per Pound is highly serially correlated
- Similarly, we can determine $\rho_2, \rho_3 ....$ etc.

Auto-Regression Model

Auto-Regression Model

Definition

An autoregressive model (also called AR model) is used to model a future behavior for a time-ordered data, using data from past behaviors.

- Essentially, it is a linear regression analysis of a dependent variable using one or more variables(s) in a given time-series data.

$$Y_t = f(Y_{t-1}, Y_{t-2}, \ldots\ldots, Y_{t-p})$$

Auto-Regression Model for Forecasting

Definition

- A natural starting point for forecasting model is to use past values of Y, that is, $Y_{t-1}, Y_{t-2}, \ldots \ldots$ to predict Y_t
- An auto-regression is a regression model in which Y_t is regressed against its own lagged values.
- The number of lags used as regressors is called the **order** of auto-regression
  - In first order auto-regression (denoted as AR(1)), Y_t is regressed against Y_{t-1}
  - In p^{th} order auto-regression (denoted as AR(p)), Y_t is regressed against, $Y_{t-1}, Y_{t-2}, \ldots \ldots, Y_{t-p}$

p^{th} order Auto-regression Model

Formula: p^{th} Order Auto-regression Model

In general, the p^{th} order auto-regression model is defined as

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \varepsilon_t$$

where, $\beta_0, \beta_1, \dots, \beta_p$ is called auto-regression coefficients and ε_t is the noise term or residue and in practice it is assumed to Gaussian white noise

For example, *AR(1)* is $Y_t = \beta_0 + \beta_1 Y_{t-1} + \varepsilon_t$

The task in AR analysis is to derive the "best" values for $\beta_i, i = 0, 1, \dots, p$ given a time series$[Y_1, Y_2, \dots, YT_1, Y_T]$

Computing AR Coefficients

Computing AR(p) model

- A number of techniques known for computing the AR coefficients
- The most common method is called Least Squares Method (LSM)
- The LSM is based upon the **Yule-Walker equations**

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \cdots & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \cdots & \rho_{p-3} & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \cdots & \rho_{p-4} & \rho_{p-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \cdots & \rho_{p-5} & \rho_{p-4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & \rho_{p-5} & \cdots & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{bmatrix}$$

Computing AR Coefficients

Computing AR (p): Yule-Walker Equations

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \cdots & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \cdots & \rho_{p-3} & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \cdots & \rho_{p-4} & \rho_{p-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \cdots & \rho_{p-5} & \rho_{p-4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \rho_{p-4} & \rho_{p-5} & \cdots & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \vdots \\ \beta_{p-1} \\ \beta_p \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \vdots \\ \rho_{p-1} \\ \rho_p \end{bmatrix}$$

- Here, $\rho_i, i = 1, 2, \ldots, p$ denotes the i^{th} **auto correlation coefficient**.
- β_0 can be chosen empirically, usually taken as zero.

ARIMA Model

AutoRegressive Integrated Moving Average (ARIMA) Model

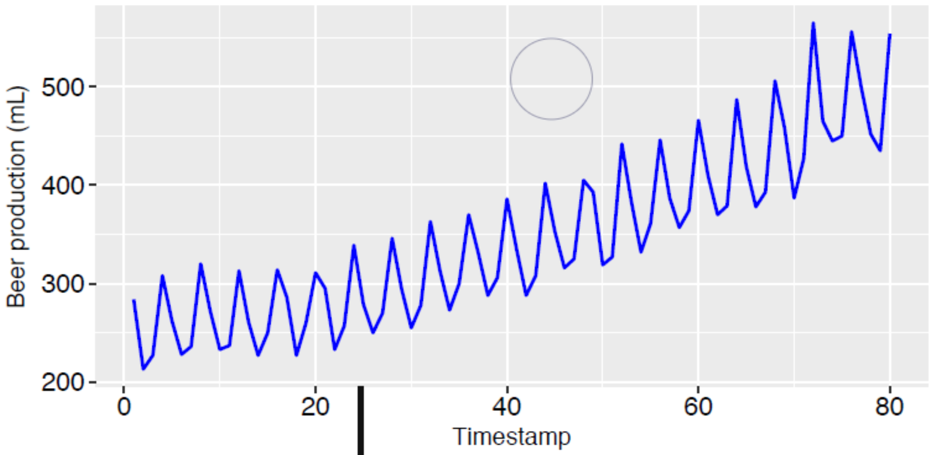
- The ARIMA model, introduced by Box and Jenkins (1976), is a linear regression model indulged in tracking linear tendencies in stationary time series data.
- AR: autoregressive (lagged observations as inputs) I: integrated (differencing to make series stationary) MA: moving average (lagged errors as inputs).
- The model is expressed as ARIMA (p, d, q) where $p, d \ and \ q$ are integer parameter values that decide the structure of the model.
- More precisely, $p \ and \ q$ are the order of the AR model and the MA model respectively, and parameter d is the level of differencing applied to the data.
- The mathematical expression of the ARIMA model is as follows:

$$y_t = \theta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

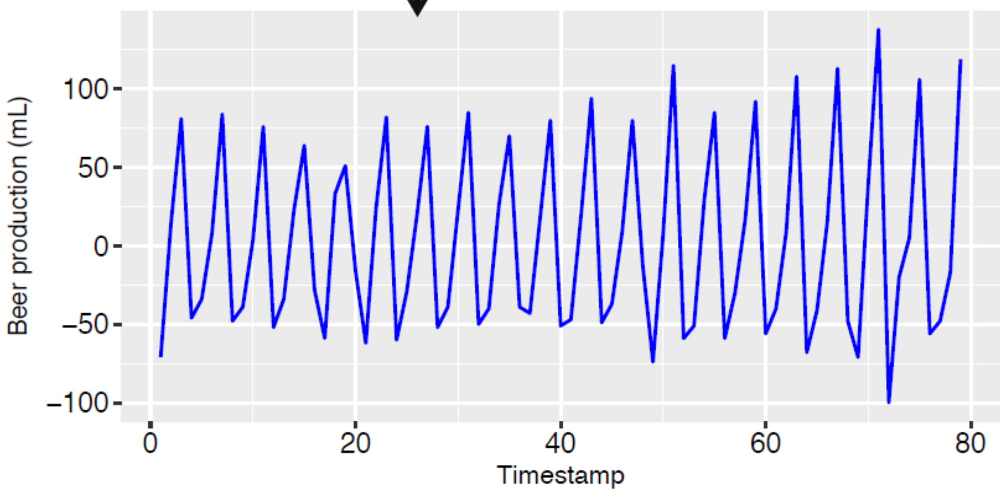
where y_t is the actual value, ε_t is the random error at time t, ϕ_i and θ_j are the coefficients of the model.

- It is assumed that $\varepsilon_{t-1}(\varepsilon_{t-1} = y_{t-1} - \hat{y}_{t-1})$ has zero mean with constant variance, and satisfies the i.i.d. condition.
- Three basic Steps: Model identification, Parameter Estimation, and Diagnostic Checking.

Differencing in ARIMA Model

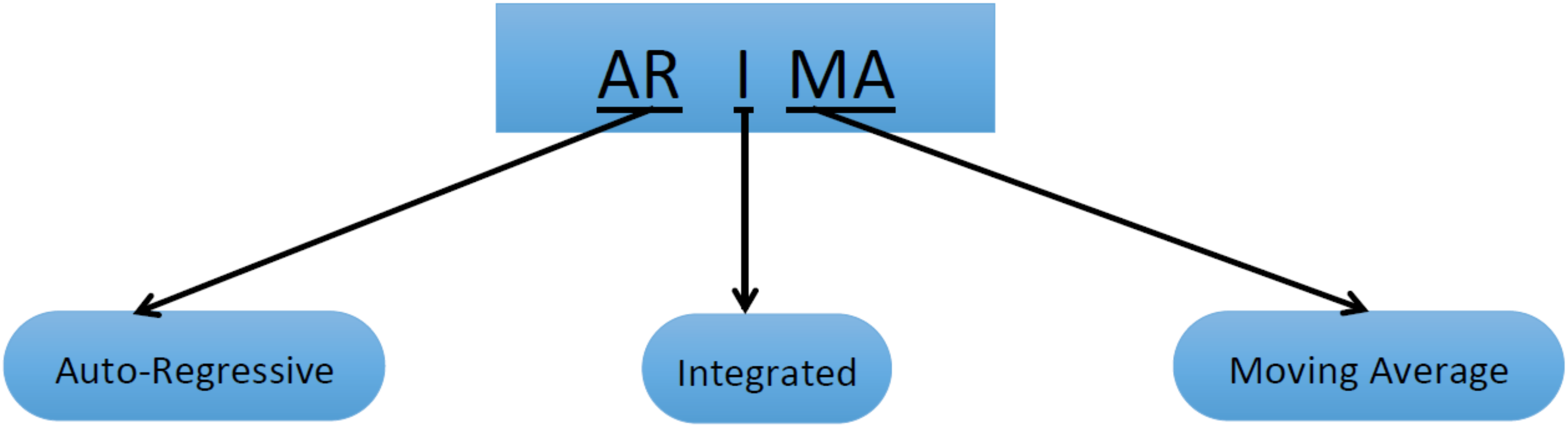


$$Y_t = Y_t - Y_{t-1}$$



Differencing order (*d*): Number of times differencing is done

ARIMA model



$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \ldots + \beta_p Y_{t-p} + \epsilon_t \quad \text{[Order p]}$$

$$Y_t = Y_t - Y_{t-1} \quad \text{[Order d]}$$

$$Y_t = \beta_0 + \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \ldots + \phi_q \epsilon_{t-q} \quad \text{[Order q]}$$

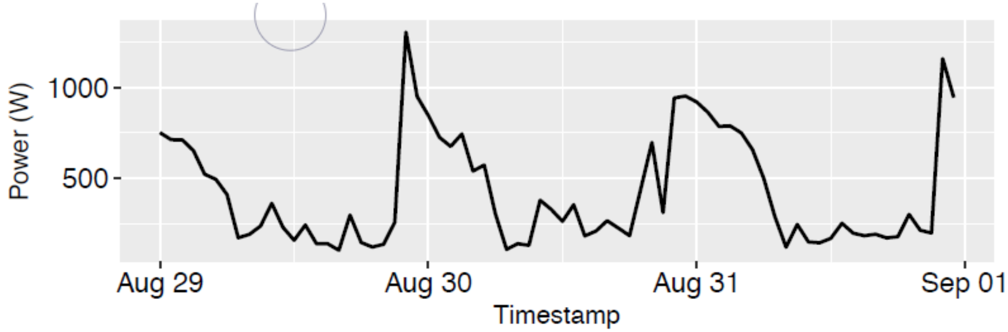
ARIMA is defined by a tuple (*p, d, q*)

ACF / PACF Plots

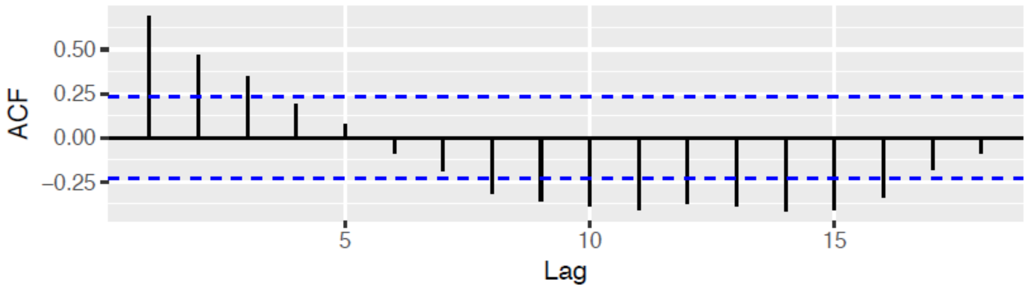
1. Auto-Correlation Function (ACF) Plot:
   - Correlation coefficients of time-series at different lags
   - Defines q order of MA model

2. Partial Auto-correlation Function (PACF) Plot:
   - Partial correlation coefficients of time series at different lags
   - Defines p order of AR model

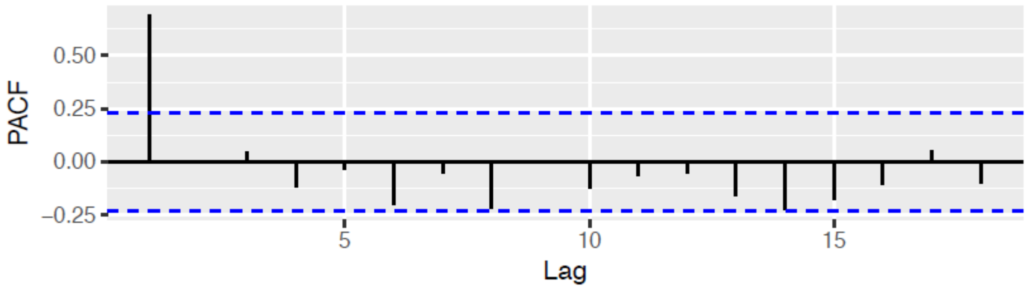
ACF / PACF Plots : Example



Data



ACF plot



PACF plot

Forecast Evaluation

Performance metrics such as mean absolute error (MAE), root mean square error (RMSE), and mean absolute percent error (MAPE) are used to evaluate the performances of different forecasting models for the unemployment rate data sets:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \,;}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \,;$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|,$$

Where y_i is the actual output, $\hat{y}_i$ is the predicted output, and n denotes the number of data points.

By definition, the lower the value of these performance metrics, the better is the performance of the concerned forecasting model.

Time Series Analysis using R

Time Series Plot:

The graphical representation of time series data by taking time on x axis & data on y axis.

A plot of data over time

Example

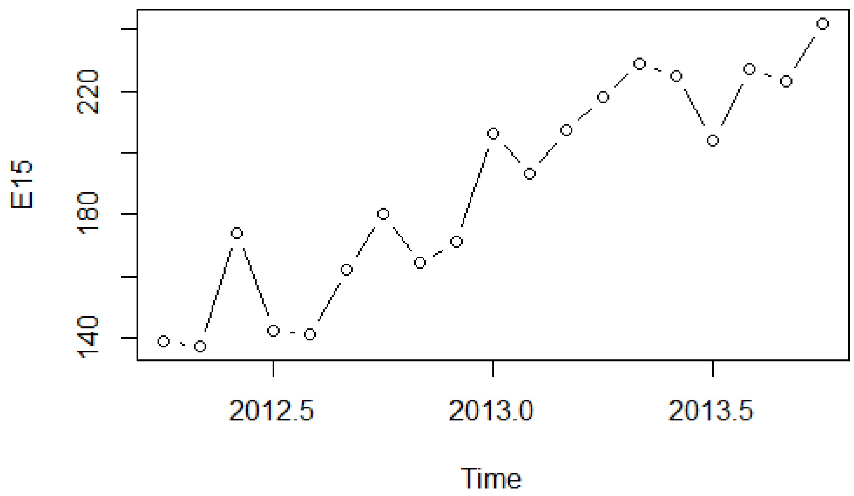
The demand for a commodity E15 for last 20 months from April 2012 to October 2013 is given in E15demand.csv file. Draw the time series plot

Month	Demand	Month	Demand
1	139	11	193
2	137	12	207
3	174	13	218
4	142	14	229
5	141	15	225
6	162	16	204
7	180	17	227
8	164	18	223
9	171	19	242
10	206	20	239

Reading data to R

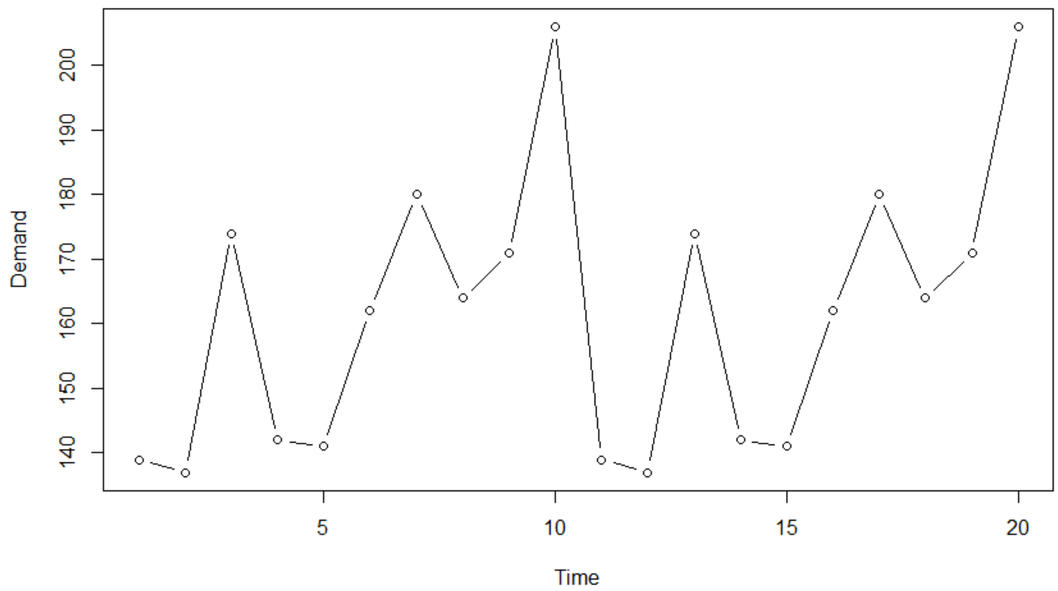
```
mydata <- read.csv("E15demand.csv")

E15 = ts(mydata$Demand, start = c(2012,4), end = c(2013,10), frequency = 12)

E15

plot(E15, type = "b")
```

For quarterly data, frequency = 4
For monthly data, frequency = 12



Reading data to R

```
E15 = ts(mydata$Demand)

E15

plot(E15, type = "b")
```

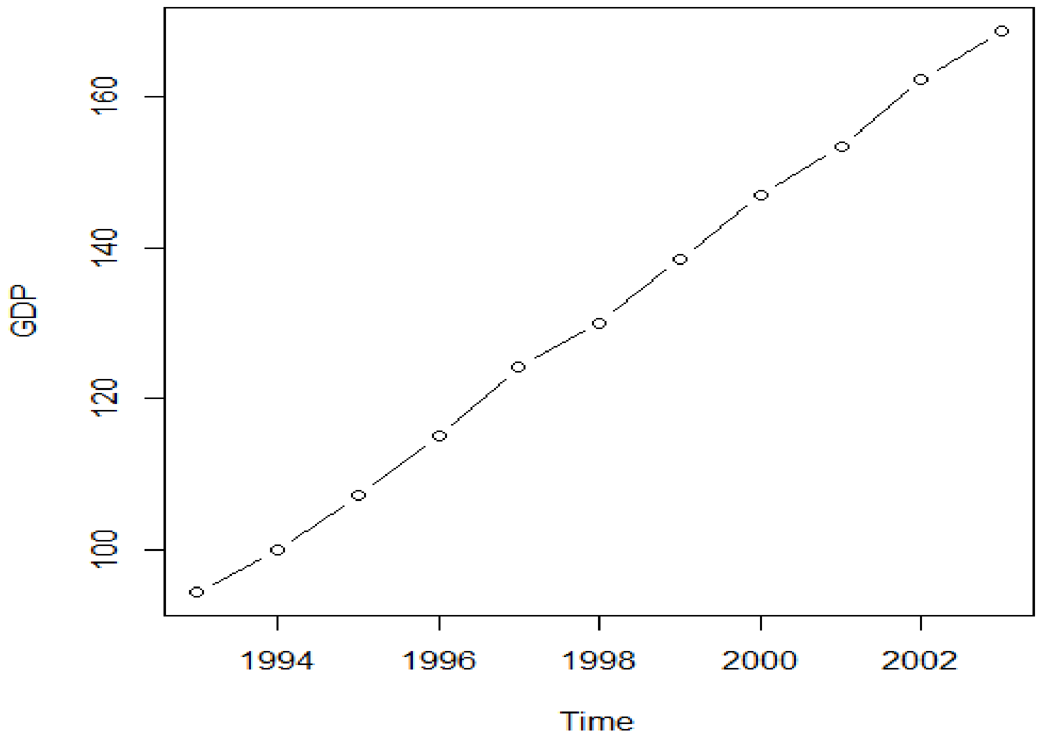


Trend:

A long term increase or decrease in the data

Example: The data on Yearly average of Indian GDP during 1993 to 2005.

Year	GDP
1993	94.43
1994	100.00
1995	107.25
1996	115.13
1997	124.16
1998	130.11
1999	138.57
2000	146.97
2001	153.40
2002	162.28
2003	168.73



Seasonal Pattern:

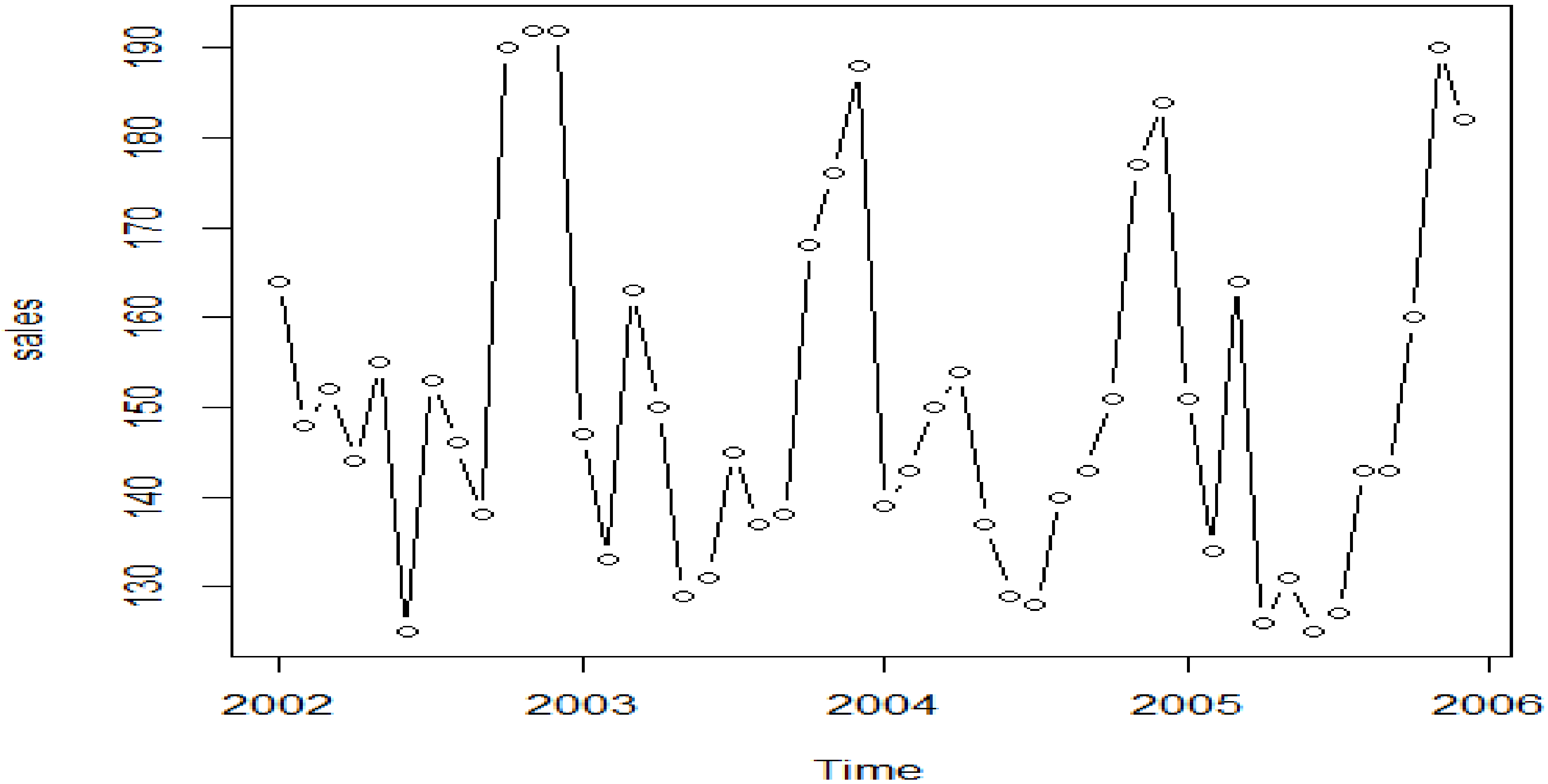
The time series data exhibiting rises and falls influenced by seasonal factors

Example: The data on monthly sales of a branded jackets

Month	Sales	Month	Sales	Month	Sales	Month	Sales
Jan-02	164	Jan-03	147	Jan-04	139	Jan-05	151
Feb-02	148	Feb-03	133	Feb-04	143	Feb-05	134
Mar-02	152	Mar-03	163	Mar-04	150	Mar-05	164
Apr-02	144	Apr-03	150	Apr-04	154	Apr-05	126
May-02	155	May-03	129	May-04	137	May-05	131
Jun-02	125	Jun-03	131	Jun-04	129	Jun-05	125
Jul-02	153	Jul-03	145	Jul-04	128	Jul-05	127
Aug-02	146	Aug-03	137	Aug-04	140	Aug-05	143
Sep-02	138	Sep-03	138	Sep-04	143	Sep-05	143
Oct-02	190	Oct-03	168	Oct-04	151	Oct-05	160
Nov-02	192	Nov-03	176	Nov-04	177	Nov-05	190
Dec-02	192	Dec-03	188	Dec-04	184	Dec-05	182

Seasonal Pattern:

The time series data exhibiting rises and falls influenced by seasonal factors

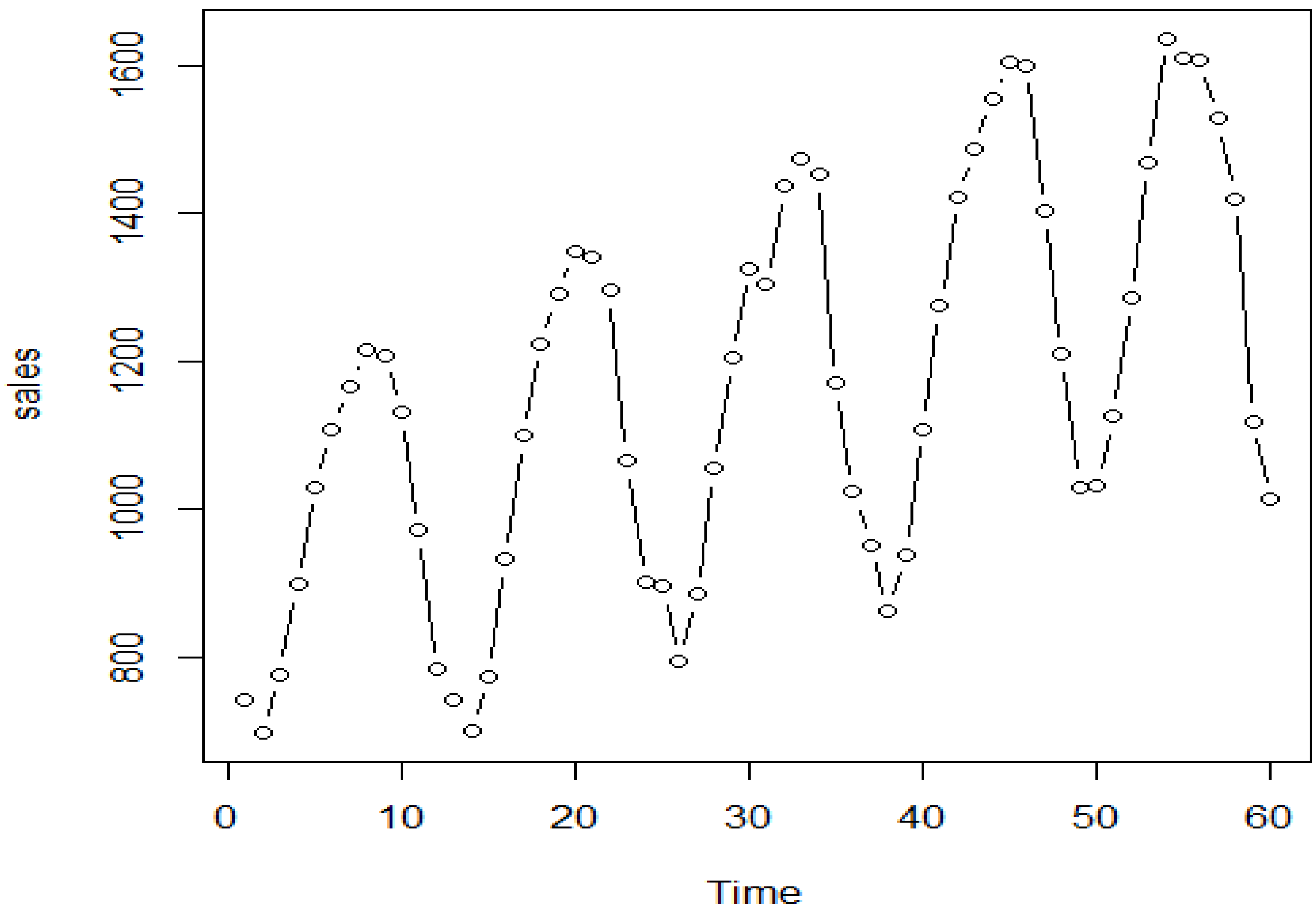


Trend and Seasonal Patterns Combined

The time series data may include a combination of trend and seasonal patterns

Example: The data on monthly sales of an aircraft component is given below:

Month	Sales	Month	Sales	Month	Sales
1	742	21	1341	41	1274
2	697	22	1296	42	1422
3	776	23	1066	43	1486
4	898	24	901	44	1555
5	1030	25	896	45	1604
6	1107	26	793	46	1600
7	1165	27	885	47	1403
8	1216	28	1055	48	1209
9	1208	29	1204	49	1030
10	1131	30	1326	50	1032
11	971	31	1303	51	1126
12	783	32	1436	52	1285
13	741	33	1473	53	1468
14	700	34	1453	54	1637
15	774	35	1170	55	1611
16	932	36	1023	56	1608
17	1099	37	951	57	1528
18	1223	38	861	58	1420
19	1290	39	938	59	1119
20	1349	40	1109	60	1013



sales
800
1000
1200
1400
1600
0
10
20
30
40
50
60
Time

Stationary Series:

A series free from trend and seasonal patterns

A series exhibits only random fluctuations around mean

Test for Stationary: Unit root test

Augmented Dickey Fuller Test (ADF) :

Checks whether any specific patterns exists in the series

H0: data is non stationary

H1: data is stationary

A small p-value suggest data is stationary.

Kwiatkowski-Phillips-Schmidt-Shin Test (KPSS) :

Another test for stationary.

Checks especially the existence of trend in the data set

H0: data is stationary

H1: data is non stationary

A large p-value suggest data is stationary.

Check stationary of data

Example : The data on daily shipments is given in shipment.csv. Check whether the data is stationary

Day	Shipments	Day	Shipments
1	99	13	101
2	103	14	111
3	92	15	94
4	100	16	101
5	99	17	104
6	99	18	99
7	103	19	94
8	101	20	110
9	100	21	108
10	100	22	102
11	102	23	100
12	101	24	98

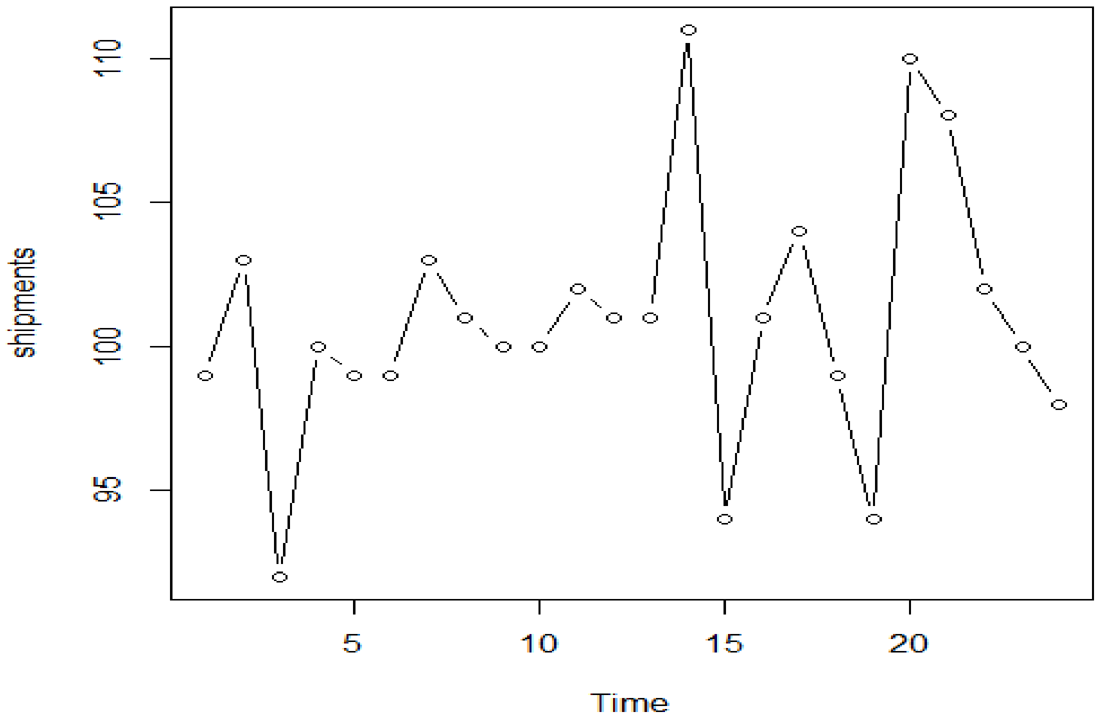
Stationary Series: A series free from trend and seasonal patterns.

A series exhibits only random fluctuations around mean

Example : The data on daily shipments is given in shipment.csv. Check whether the data is stationary

R code

```
mydata <- read.csv("shipment.csv")
shipments = ts(mydata$Shipments)
plot(shipments, type = "b")
```



Test for checking series is Stationary: Unit root test in R

ADF Test

R Code

```
install.packages("tseries")
library("tseries")
adf.test(shipments)
```

Statistic	Value
Dickey-Fuller	-3.2471
P value	0.09901

Since p value = 0.099 < 0.1, the data is stationary at 10% significant level

Test for checking series is Stationary : Unit root test in R

KPSS test

R Code

```
kpss.test(shipments)
```

Statistic	Value
KPSS Level	0.24322
P value	> 0.1

Since p value > 0.1 >= 0.1, the data is stationary at 10% level of significance

Differencing: A method for making series stationary

A differenced series is the series of difference between each observation Y_t and the previous observation Y_{t-1}

$$Y_t' = Y_t - Y_{t-1}$$

A series with trend can be made stationary with 1st differencing

A series with seasonality can be made stationary with seasonal differencing

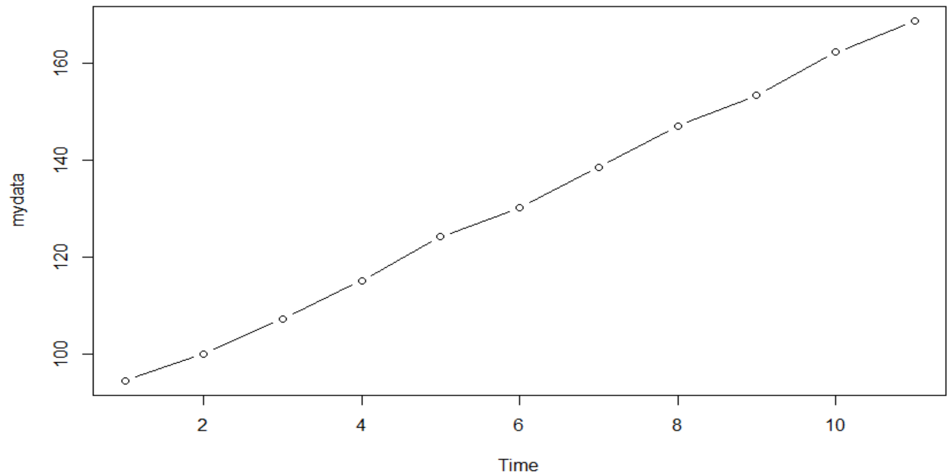
Example: Is it possible to make the GDP data given in GDP.csv stationary?

Differencing: A method for making series stationary

Example: Is it possible to make the GDP data given in GDP.csv stationary?

R Code

```
>mydata = ts(GDP$GDP)
> plot(mydata, type = "b")
```



KPSS Statistic	0.48402
P value	0.04527

Conclusion

Series has a linear trend

KPSS test (p value < 0.05) shows data is not stationary

Differencing: A method for making data stationary

Example: Is it possible to make the GDP data given in GDP.csv stationary?

Identify the number of differencing required

R Code

```
install.packages("forecast")
library(forecast)
ndiffs(GDP)
```

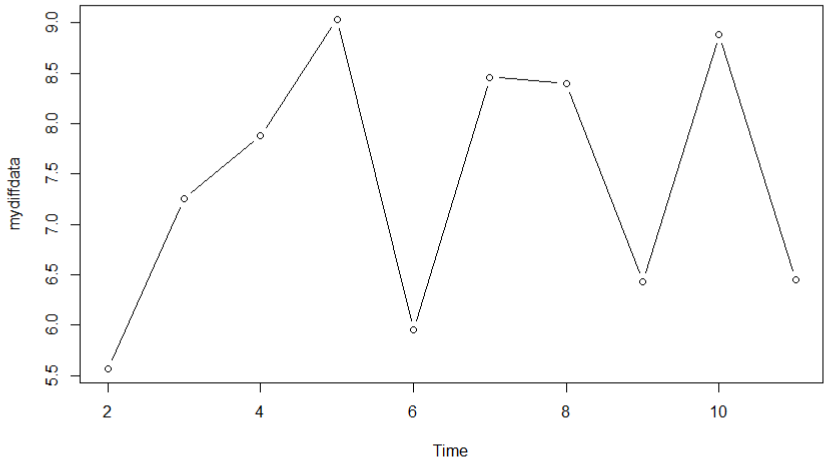
Differencing required is 1

Yt’ = Yt – Yt-1

```
mydiffdata = diff(GDP, difference = 1)
plot(mydiffdata, type = "b")
adf.test(mydiffdata)
kpss.test(mydiffdata)
```

Differencing: A method for making series stationary

Example: Is it possible to make the GDP data given in GDP.csv stationary?



Test	Statistic	P value
ADF	-5.0229	< 0.01
KPSS	0.20905	>0.1

Conclusion: Series became stationary after 1^{st} order differencing

Single Exponential Smoothing:

Give more weight to recent values compared to the old values

More efficient for stationary data without any seasonality and trend

Single Exponential Smoothing: Methodology

Let $y_1, y_2, - - - y_t$ be the values, then

y_{t+1} estimate = $S_{t+1} = \alpha\, y_t + (1- \alpha)\, S_t$

where $0 \leq \alpha \leq 1$ and $S_1 = y_1$

Example: The data on ad revenue from an advertising agency for the last 12 months is given in Amount.csv. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

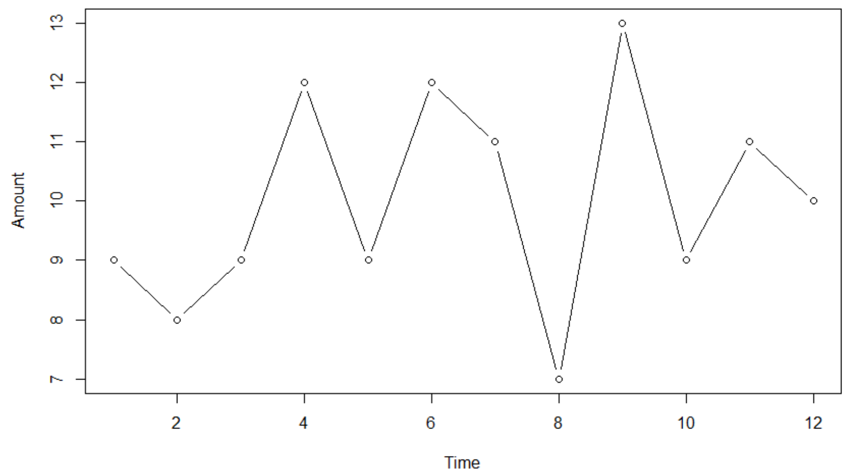
Month	Amount	Month	Amount
1	9	7	11
2	8	8	7
3	9	9	13
4	12	10	9
5	9	11	11
6	12	12	10

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

R code

Reading and plotting the data

```
mydata <- read.csv("Amount.csv")
amount = ts(mydata$Amount)
plot(amount, type ="b")
```



Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

R code

Checking whether series is stationary

```
library(forecast)
adf.test(amount)
kpss.test(amount)
```

Test	Statistic	P value
ADF	-2.3285	0.4472
KPSS	0.24038	>0.1

ADF and KPSS tests show that the series is stationary

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

R code

Fitting the model

```
mymodel = HoltWinters(amount, beta = FALSE, gamma = FALSE)
mymodel
```

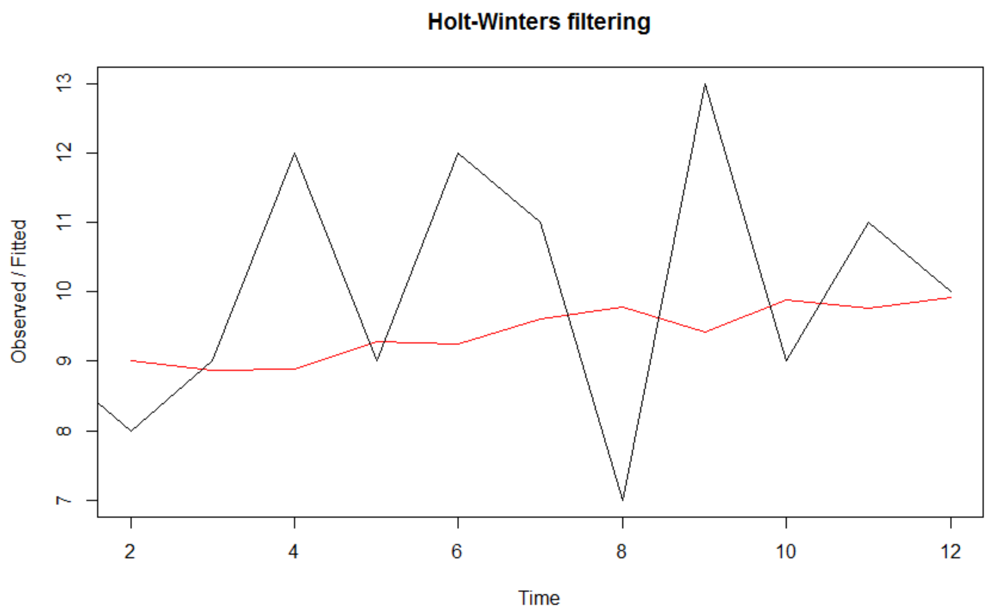
Smoothing parameter	value
alpha	0.1285076

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

R code

Actual Vs Fitted plot

```
plot(mymodel)
```



Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

R code

Computing predicted values and residuals (errors)

```
pred = fitted(mymodel)
res = residuals(mymodel)
outputdata = cbind(amount, pred[,1], res)
write.csv(outputdata, “amount_outputdata.csv")
```

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Month	Actual	Predicted	Error
1	9		
2	8	9	-1
3	9	8.8715	0.12851
4	12	8.8880	3.11199
5	9	9.2879	-0.2879
6	12	9.2509	2.74908
7	11	9.6042	1.3958
8	7	9.7836	-2.7836
9	13	9.4259	3.57414
10	9	9.8852	-0.8852
11	11	9.7714	1.22859
12	10	9.9293	0.0707

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics

Residual = Actual – Predicted

Mean Absolute Error: MAE

Root Mean Square Error: RMSE

Mean Absolute Percentage Error: MAPE

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics – R Code

```
abs_res = abs(res)
res_sq = res^2
pae = abs_res/ amount
```

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics

Month	Absolute Error	Error Squares	Absolute Error / Actual
1.0000	1.0000	1.0000	0.1250
2.0000	0.1285	0.0165	0.0143
3.0000	3.1120	9.6845	0.2593
4.0000	0.2879	0.0829	0.0320
5.0000	2.7491	7.5574	0.2291
6.0000	1.3958	1.9483	0.1269
7.0000	2.7836	7.7483	0.3977
8.0000	3.5741	12.7745	0.2749
9.0000	0.8852	0.7835	0.0984
10.0000	1.2286	1.5094	0.1117
11.0000	0.0707	0.0050	0.0071

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics

Statistic	Description	R Code	Value
ME	Average residuals	mean(res)	0.6638322
MAE	Average of absolute residuals	mean(abs_res)	1.565
MSE	Average of residual squares	mse = mean(res_sq)	3.919
RMSE	Square root of MSE	sqrt(mse)	1.980
MAPE	Average of absolute error / actual	mean(PAE)*100	15.23%

Criteria

MAPE < 10% is reasonably good
MAPE < 5 % is very good

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics - Normality of Errors with zero

R Code

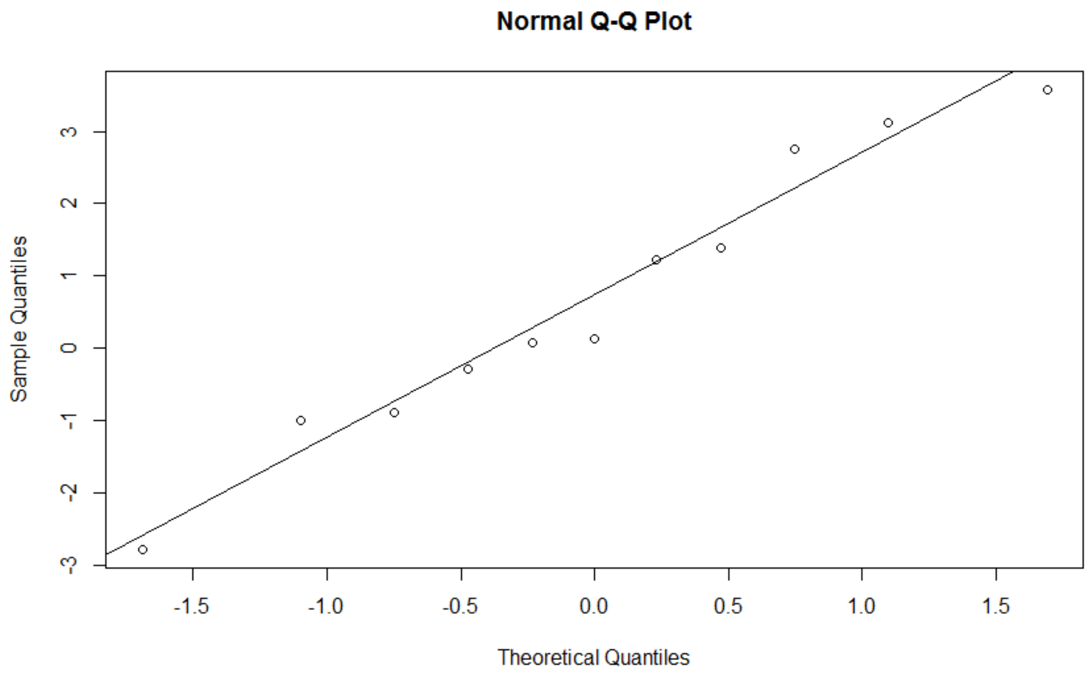
```
qqnorm(res)
qqline(res)
shapiro.test(res)
mean(res)
```

Statistic (w)	P value
0.962	0.7963

Error Mean	0.6638

Example: The data on ad revenue from an advertising agency for the last 12 months is given below. Forecast the ad revenue from the agency in the future month using single exponential smoothing method with best value of α?

Model diagnostics – Normal Q – Q plot



Forecast and Prediction Interval

Prediction interval : Predicted value $\pm$ z $\sqrt{MSE}$

where z = width of prediction interval

Prediction Interval	z
90%	1.645
95%	1.960
99%	2.576

Forecasted value $S_{t+1} = \alpha y_t + (1 - \alpha)S_t$
Forecasted value $S_{13} = \alpha y_{12} + (1 - \alpha)S_{12}$
Forecasted value $S_{13} = 0.1285076 \times 10 + (1 - 0.1285076) \times 9.9293 = 9.9383$

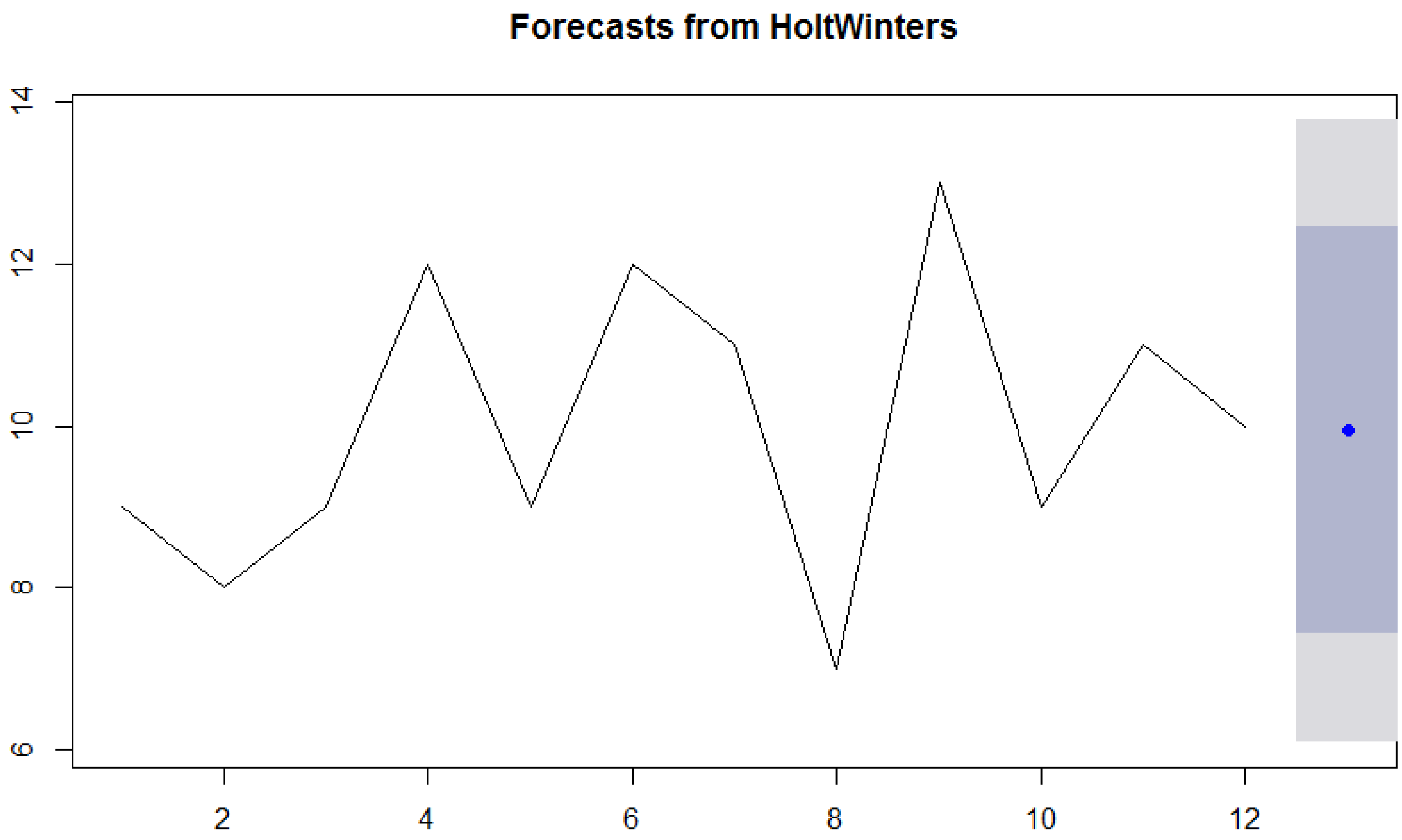
Forecast

R Code

```
library(forecast)

forecast = forecast(mymodel, 1)

forecast

plot(forecast)
```

Month	Forecast	80% Prediction Interval		95% Prediction Interval	
		Lower	Upper	Lower	Upper
13	9.938382	7.431552	12.44521	6.104517	13.77225

Forecast Plot



TIME SERIES MODELING

General form of linear model

y is modeled in terms of x's

$$Y = a + b_1 x_1 + b_2 x_2 + - - - + b_k x_k$$

Step 1: Check Correlation between y and x's

y should be correlated with some of the x's

Time series model

Generally there will not be any x's

Hence patterns in y series is explored

y will be modeled in terms of previous values of y

$$y_t = a + b_1 y_{t-1} + b_2 y_{t-2} + - -$$

Step 1: Check correlation between y_t and y_{t-1}, etc

correlation between y and previous values of y are called autocorrelation

Example: Check the auto correlation up to 3 lags in GDP data

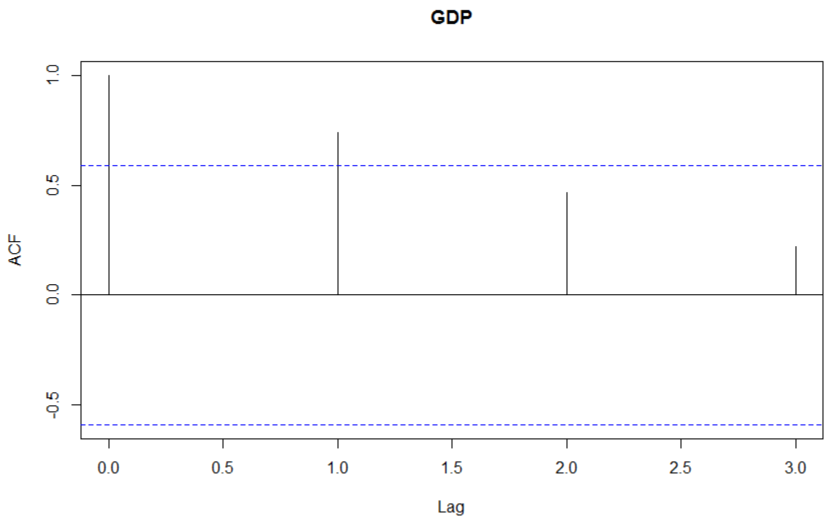
Year	GDP(y_t)	y_{t-1}	y_{t-2}	y_{t-3}
1993	94.43			
1994	100	94.43		
1995	107.3	100	94.43	
1996	115.1	107.3	100	94.43
1997	124.2	115.1	107.3	100
1998	130.1	124.2	115.1	107.3
1999	138.6	130.1	124.2	115.1
2000	147	138.6	130.1	124.2
2001	153.4	147	138.6	130.1
2002	162.3	153.4	147	138.6
2003	168.7	162.3	153.4	147

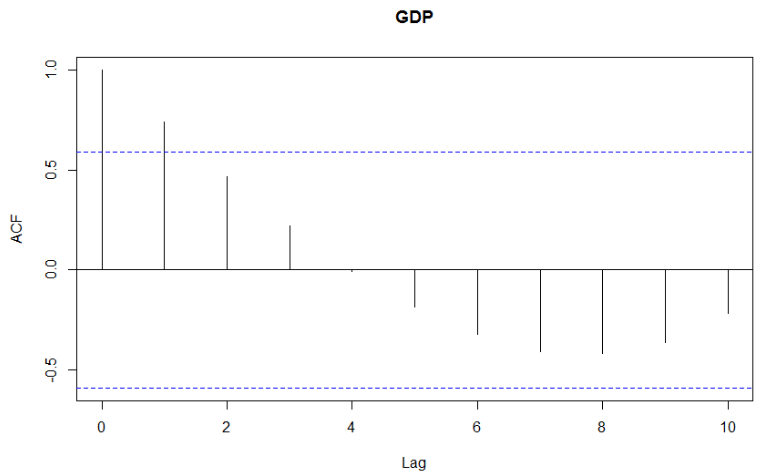
Lag	variables	Auto Correlation
1	y_t vs y_{t-1}	0.9985
2	y_t vs y_{t-2}	0.9984
3	y_t vs y_{t-3}	0.9981

Example: Check the auto correlation up to 3 lags in GDP data

R Code

```
mydata <- read.csv("Trens_GDP.csv")
GDP <- ts(mydata$GDP, start = 1993, end = 2003)
acf(GDP, 3)
acf(GDP)
```





Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Widely used and very effective modeling approach

Proposed by George Box and Gwilym Jenkins

Also known as Box – Jenkins model or ARIMA(p,d,q)

where

p: number of auto regressive (AR) terms

q: number of moving average (MA) terms

d: level of differencing

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

General Form

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + - - - + \theta_1 e_{t-1} + + \theta_2 e_{t-2} - - - -$$

Where

c: constant

$\phi_{1,}\ \phi_{2,}\ \theta_{1,}\ \theta_2$, - - - are model parameters

$e_{t-1} = y_{t-1} - s_{t-1}$, e_t are called errors or residuals

s_{t-1} : predicted value for the t-1th observation (y_{t-1})

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1:

Draw time series plot and check for trend, seasonality, etc

Step 2:

Draw Auto Correlation Function (ACF) and Partially Auto Correlation Function (PACF) graphs to identify auto correlation structure of the series

Step 3:

Check whether the series is stationary using unit root test (ADF test, KPSS test)

If series is non stationary do differencing or transform the series

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4:

Identify the model using ACF and PACF or automatically

The best model is one which minimizes AIC or BIC or both

Step 5:

Estimate the model parameters using maximum likelihood method (MLE)

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 6:

Do model diagnostic checks

The errors or residuals should be white noise and should not be auto correlated

Do Portmanteau and Ljung & Box tests. If p value > 0.05, then there is no autocorrelation in residuals and residuals are purely white noise.

The model is a good fit

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

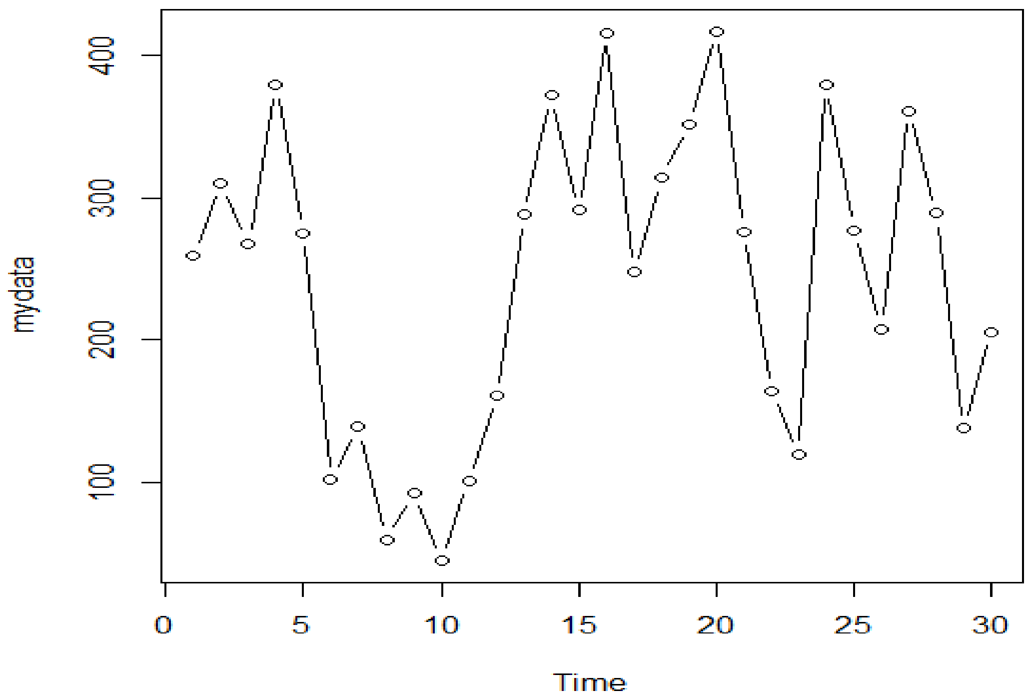
Example: The number of visitors to a web page is given in Visits.csv. Develop a model to predict the daily number of visitors?

SL No.	Data	SL No.	Data
1	259	16	416
2	310	17	248
3	268	18	314
4	379	19	351
5	275	20	417
6	102	21	276
7	139	22	164
8	60	23	120
9	93	24	379
10	45	25	277
11	101	26	208
12	161	27	361
13	288	28	289
14	372	29	138
15	291	30	206

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 1: Read and plot the series

```
mydata <- read.csv("Visits.csv")
mydata <- ts(mydata$Data)
plot(mydata, type = "b")
```



Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 2: Descriptive Statistics

```
summary(mydata)
```

Statistic	Value
Minimum	45
Quartile 1	144.5
Median	271.5
Mean	243.6
Quartile 3	313
Maximum	417

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 3: Check whether the series is stationary

```
library(tseries)
adf.test(mydata)
kpss.test(mydata)
ndiffs(mydata)
```

Test	Statistic	P value
ADF	-2.494	0.3829
KPSS	0.15007	> 0.1

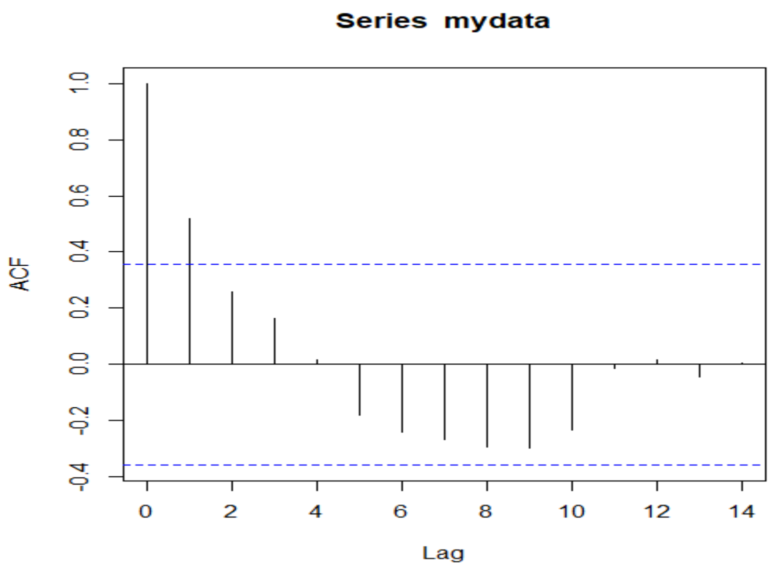
Both tests shows that series is stationary
Number of differences required = 0

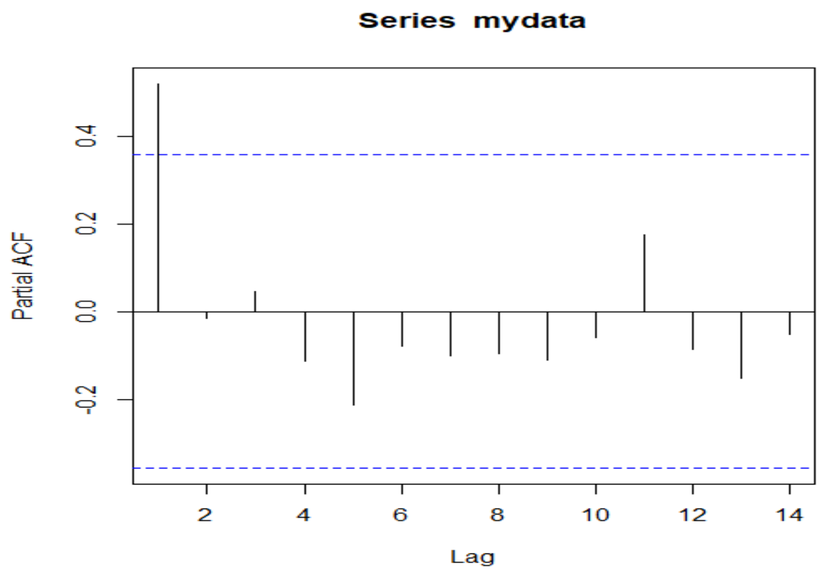
Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 4: Draw ACF & PACF Graphs

acf(mydata)

pacf(mydata)





Potential Models

ARMA(1,0) since acf at lag 1 is crossing 95% confidence interval

ARMA(0,1) since pacf at lag 1 is crossing 95% confidence interval

ARMA(1,1) since both acf and pacf at lag 1 is crossing 95% confidence interval

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 5: Identification of model automatically

```
library(forecast)
mymodel = auto.arima(mydata)
mymodel
```

Model	Log likelihood	AIC	BIC
ARIMA(1,0,0)	-178.31	362.62	366.82

Model Parameters	Value
Intercept	242.8594
AR1	0.5064

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 6: Identification of model manually

```
arima(mydata, c(0,0,1))
arima(mydata, c(1,0,0))
arima(mydata, c(1,0,1))
```

Model		Log likelihood	AIC
p=0,q=1	ARIMA(0,0,1)	-179.07	364.15
p=1,q=0	ARIMA(1,0,0)	-178.31	362.62
p=1,q=1	ARIMA(1,0,1)	-178.31	364.62

Conclusion:

The best model which minimizes AIC & BIC is p=1, q=0 or ARIMA(1,0,0)

Identified automatically

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 7: Estimation of parameters

ARIMA(1,0,0) Parameters	Value	Std Error
Intercept	242.8594	32.8552
AR1	0.5064	0.1520

The model is: $Y_t = 242.8594 + 0.5064\, Y_{t-1}$

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

```
summary(mymodel)
```

Statistic	Description	Value
ME	Residual average	-0.3470709
MAE	Average of absolute residuals	76.90398
RMSE	Root mean square of residuals	91.81328
MAPE	Mean absolute percent error	47.78088

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

```
pred = fitted(mymodel)
res = residuals(mymodel)
```

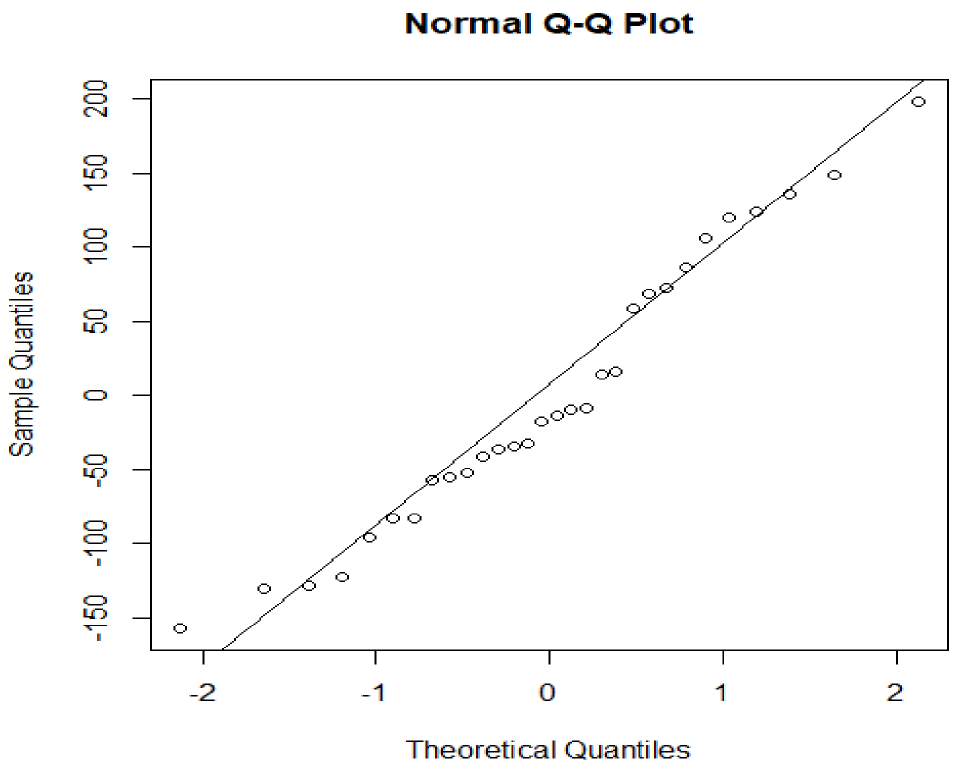
Normality check on Residuals

```
qqnorm(res)
qqline(res)
shapiro.test(res)
hist(res, col = "grey")
```

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

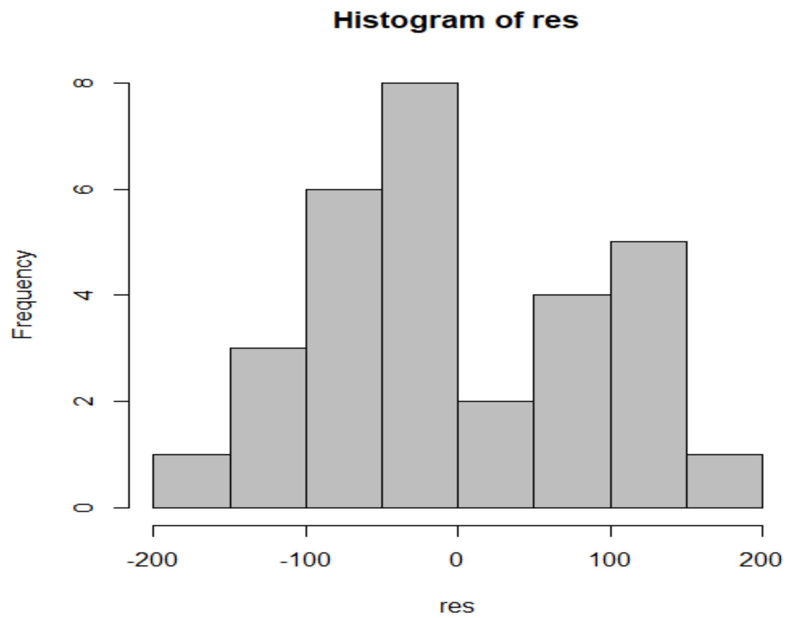
Normality check on Residuals : Normal Q – Q Plot



Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

Normality check on Residuals: Histogram of Residuals



Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

Normality check on Residuals: Shapiro Wilk Normality test

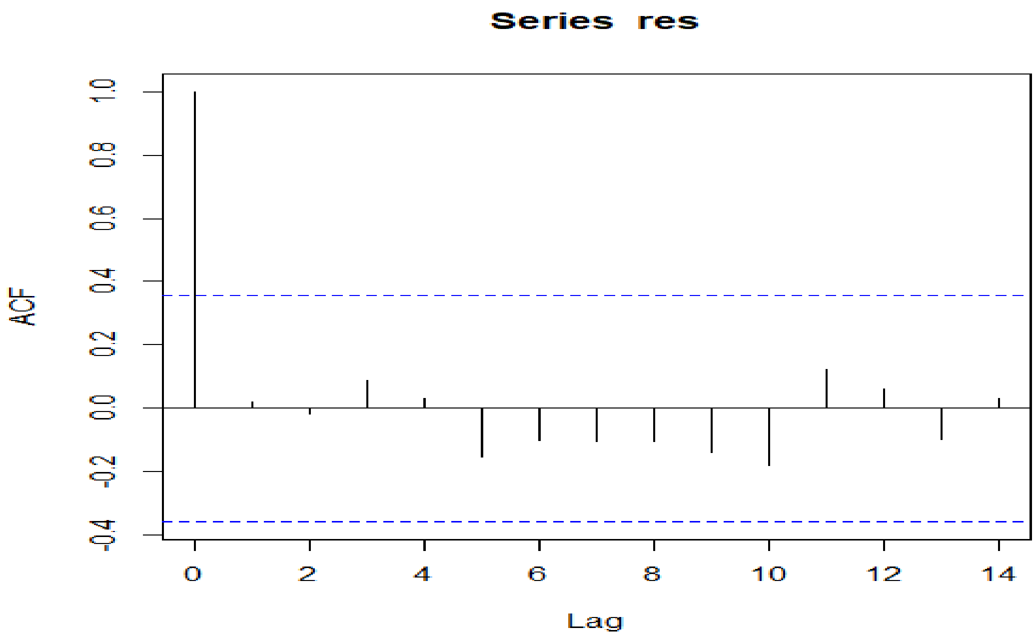
Statistic	p value
0.96445	0.4004

P > 0.05, Residuals are normal

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

Checking auto correlation among residuals: ACF of Residuals



None of the autocorrelation values is exceeding 95% confidence interval
Residuals are not auto correlated

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model Diagnostics

Tests for checking auto correlation among residuals

Ljung-Box Test

Test whether the residuals are independent or not auto correlated
If p value ≥ 0.05, then the residuals are not auto correlated and independent

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 8: Model diagnostics

Ljung & Box Test

```
Box.test(res, lag = 15, type = "Ljung-Box")
```

Test	Lag	Statistic	df	p value
Ljung & Box	15	6.5528	15	0.9689

Since the p value $\geq$ 0.05, The residuals are not auto correlated

The residuals are white noise

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Step 9: Forecasting upcoming values

```
forecast = forecast(mymodel, h = 3)

forecast
```

Point	Forecast	80% Prediction Interval		95% Prediction Interval	
		Lower	Upper	Lower	Upper
31	224.1953	102.40201	345.9885	37.92856	410.4620
32	233.4086	96.89144	369.9258	24.62361	442.1936
33	238.0739	98.03062	378.1172	23.89618	452.2516

Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

Exercise 1: The data on sales of a electro magnetic component is given in Sales.csv. Develop a forecasting methodology?

Period	Data	Period	Data
1	4737	16	4405
2	5117	17	4595
3	5091	18	5045
4	3468	19	5700
5	4320	20	5716
6	3825	21	5138
7	3673	22	5010
8	3694	23	5353
9	3708	24	6074
10	3333	25	5031
11	3367	26	5648
12	3614	27	5506
13	3362	28	4230
14	3655	29	4827
15	3963	30	3885

Cheatsheet

Dependent Variable Type (Ys)	Independent Variable Type (Xs)	Modelling Technique
Numerical	Numerical	1. Linear Regression or Best Subset Regression 2. Non-linear Regression or Regression Splines 3. Regression Trees, Neural Nets, etc.
Numerical	Categorical + Numerical	1. Linear Regression with Dummy Variables 2. Polynomial Regression with Dummy Variables 3. Regression Trees, Neural Nets, etc.
Categorical	Numerical	1. Logistics Regression 2. Classification Trees 3. Support Vector Machines, Neural Nets, etc.
Categorical	Categorical + Numerical	1. Logistic Regression with Dummy Variables 2. Classification Trees 3. Advanced Neural Nets, etc.
Numerical (Time dependent)	Numerical Exogenous Variables	1. ARIMA, ETS, Naïve Model 2. Autoregressive Neural Network 3. RNN, LSTM, etc.

References

Read Online: https://otexts.com/fpp3/

A very updated Survey Paper:

https://arxiv.org/abs/2010.05079