Business Analytics

Data Exploration & Modeling

CONTENTS – Part I

SL No.	Topics
1	Time series analysis
2	Multivariate time series analysis (VAR models)
3	Co integration modelling
4	Panel data analysis
5	Association rule mining
6	Conjoint Analysis

TME SERIES ANALYSIS

INTRODUCTION

Time Series:

A collection of observations or data made sequentially in time.

A dataset consisting of observations arranged in chronological order

A sequence of observations over time

Forecast:

An estimate of the future value of some variable

Example:

The number of 2 wheeler sales in Bangalore during next month

The average volume of an airline passengers in the next quarter

INTRODUCTION

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

INTRODUCTION

Reading data to R

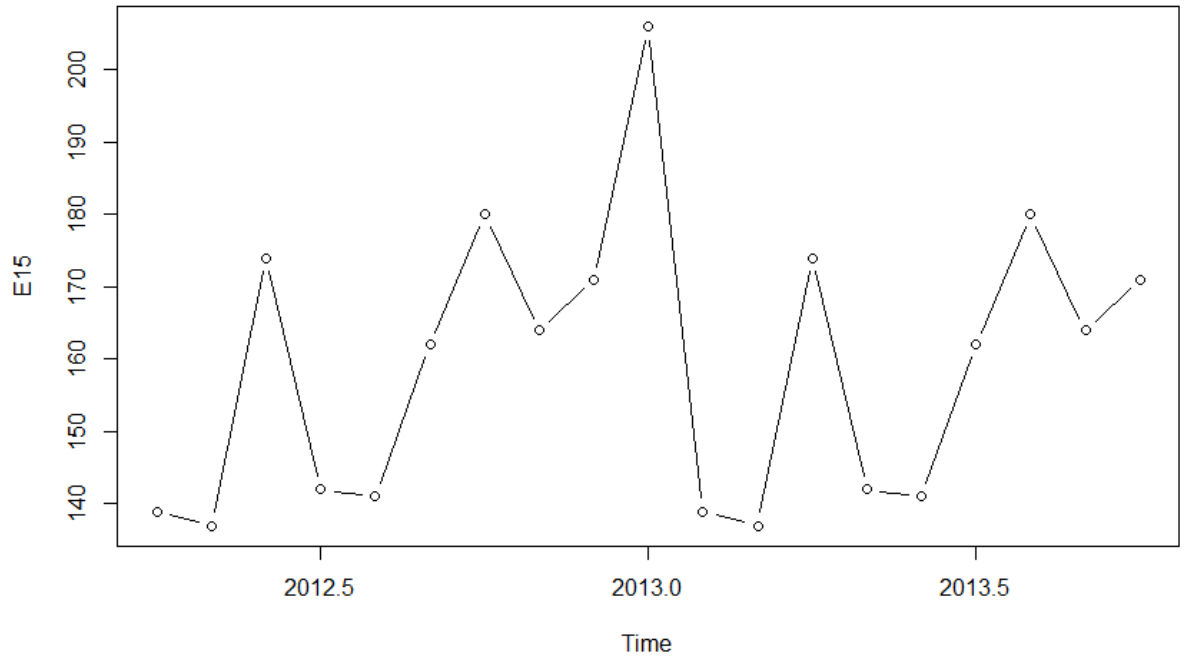
```
> E15 = ts(mydata, start = c(2012,4), end = c(2013,10), frequency = 12)

> E15

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

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

INTRODUCTION

Reading data to R

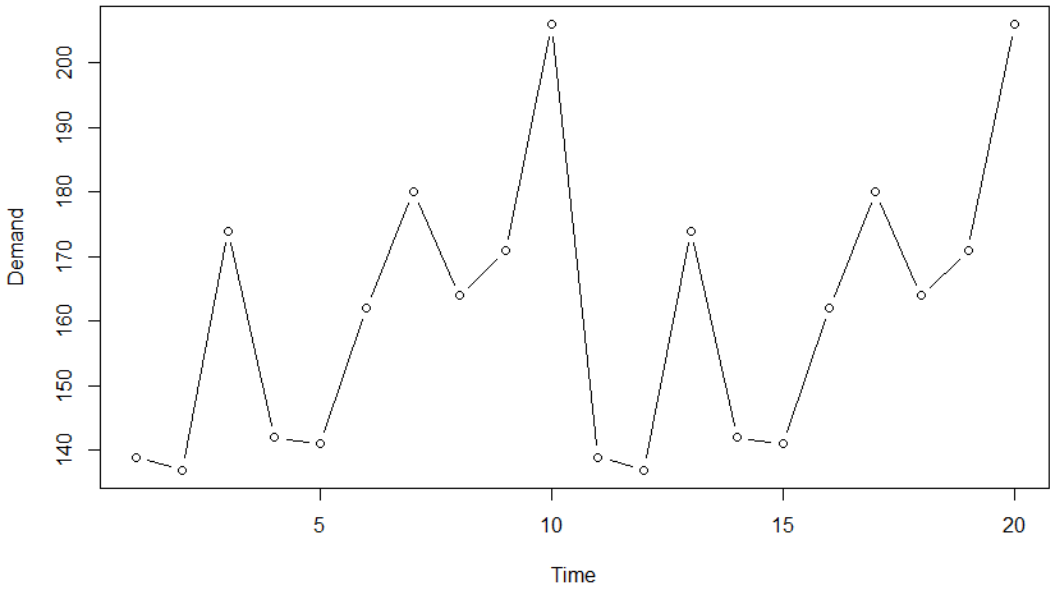
```
> E15 = ts(mydata, start = c(2012,4), end = c(2013,10), freq = 12)
> E15
> plot(E15, type = "b")
```



INTRODUCTION

Reading data to R

```
> E15 = ts(mydata)

> E15

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



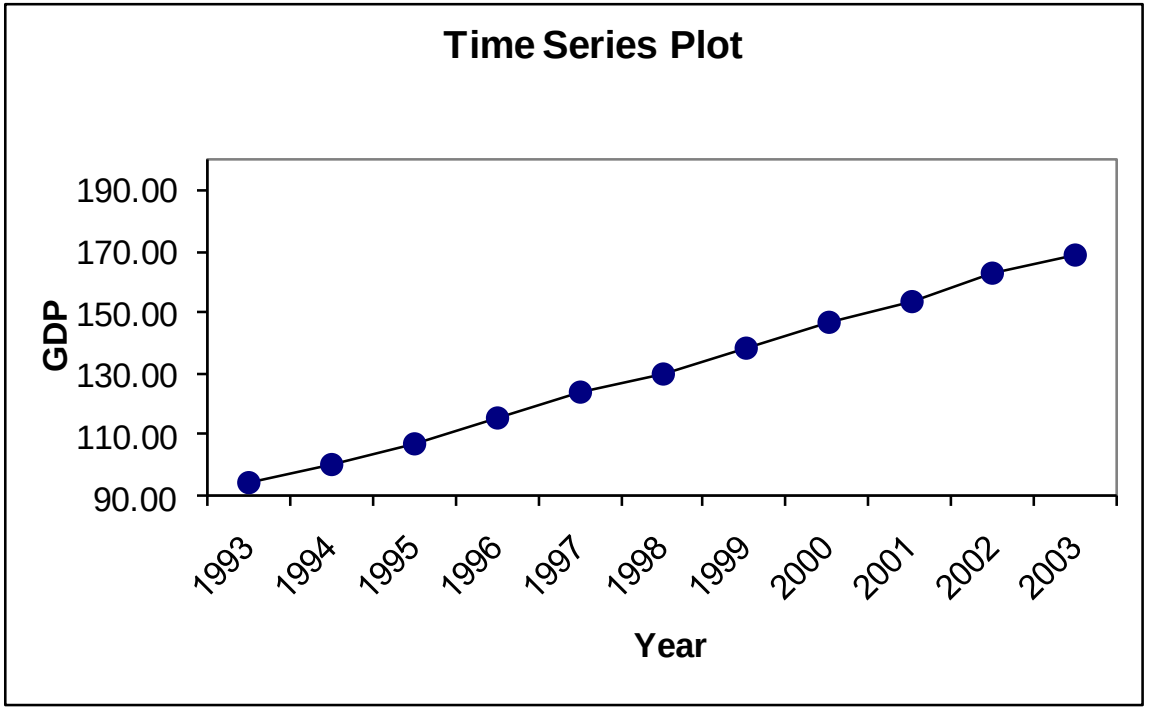
INTRODUCTION

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



INTRODUCTION

Seasonal Pattern:

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

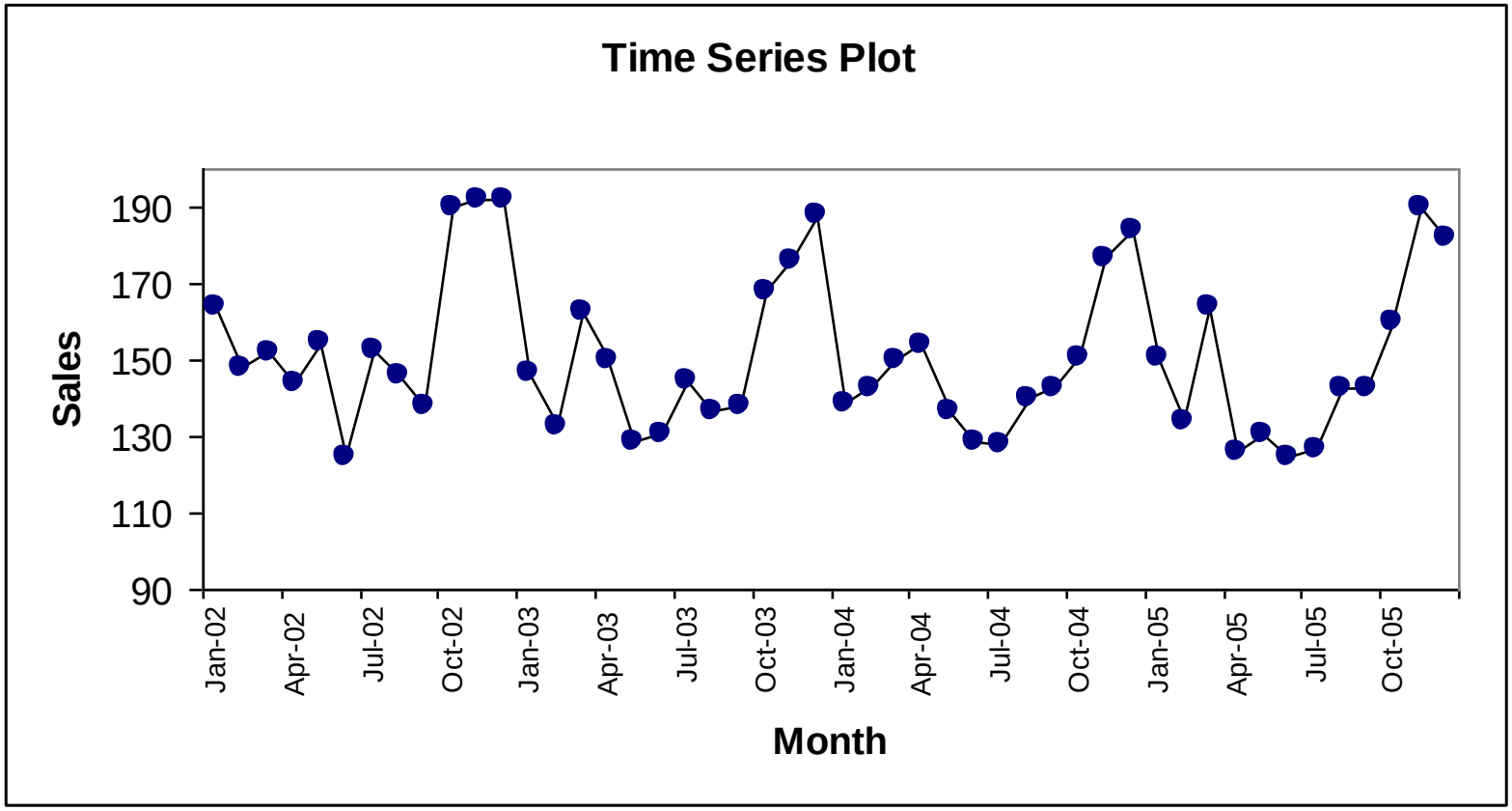
Example: The data on monthly sales of a branded jackets

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

INTRODUCTION

Seasonal Pattern:

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

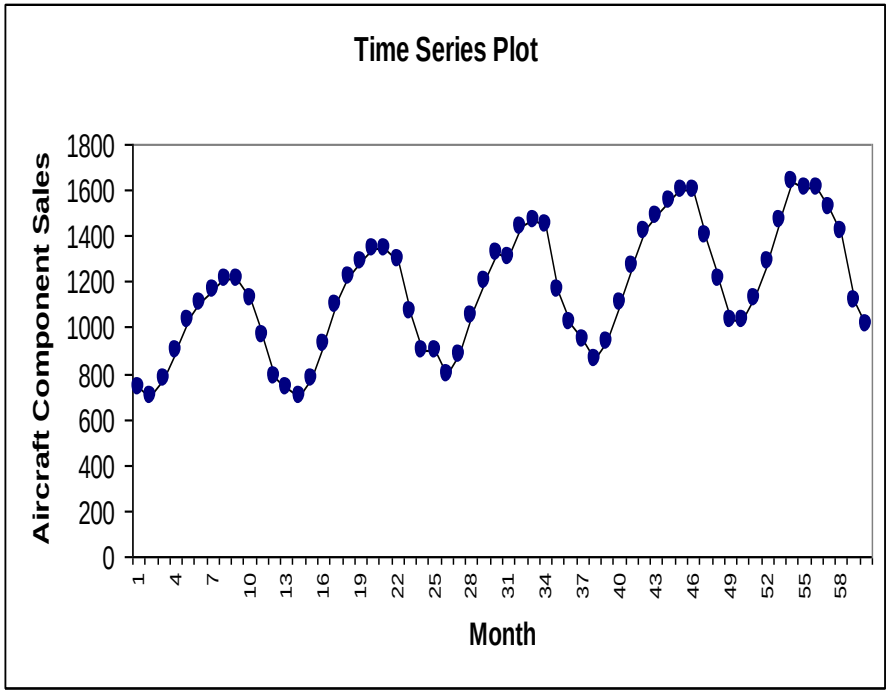


INTRODUCTION

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



INTRODUCTION

Stationary Series:

A series free from trend and seasonal patterns

A series exhibits only random fluctuations around mean

INTRODUCTION

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.

INTRODUCTION

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

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

INTRODUCTION

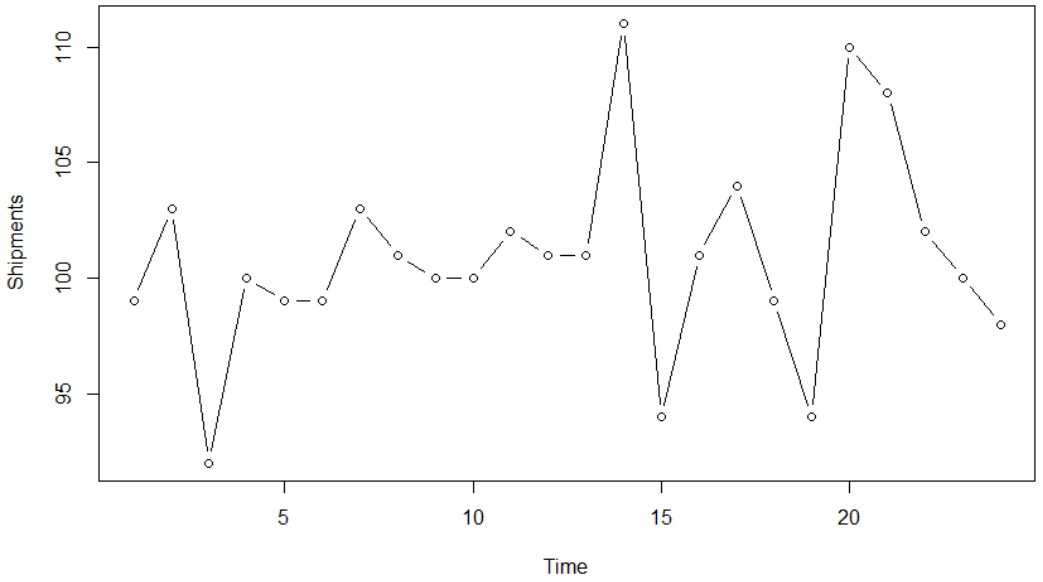
Stationary Series: A series free from trend and seasonal patterns.

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



INTRODUCTION

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

ADF Test

R Code

```
> library(tseries)
> adf.test(mydata)
```

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

INTRODUCTION

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

KPSS test

R Code

```
> kpss.test(mydata)
```

Statistic	Value
KPSS Level	0.1967
P value	0.1

Since p value = 0.1 >= 0.05, the data is stationary

INTRODUCTION

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

Exercise 1 : Check whether the GDP data is stationary. File GDP.csv?

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

INTRODUCTION

Differencing: A method for making series stationary

A differenced series is the series of difference between each observation Yt and the previous observation Yt-1

$$Yt' = Yt - Yt\text{-}1$$

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

A series with seasonality can be made stationary with seasonal differencing

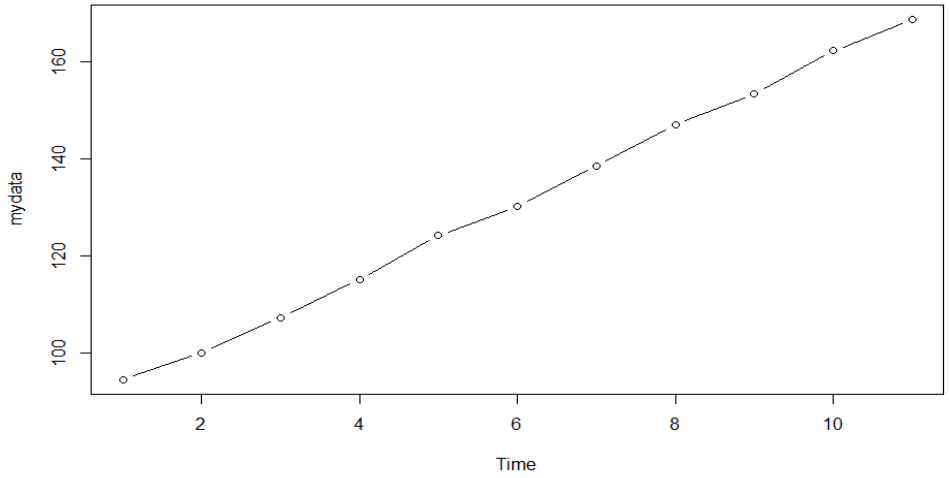
Example: Is it possible to make the GDP data given in GDP.csv stationary?

INTRODUCTION

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	1.1121
P value	0.01

Conclusion
Series has a linear trend
KPSS test (p value < 0.05) shows data is not stationary

INTRODUCTION

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

```
> library(forecast)
> ndiffs(mydata)
```

Differencing required is 1

Yt' = Yt – Yt-1

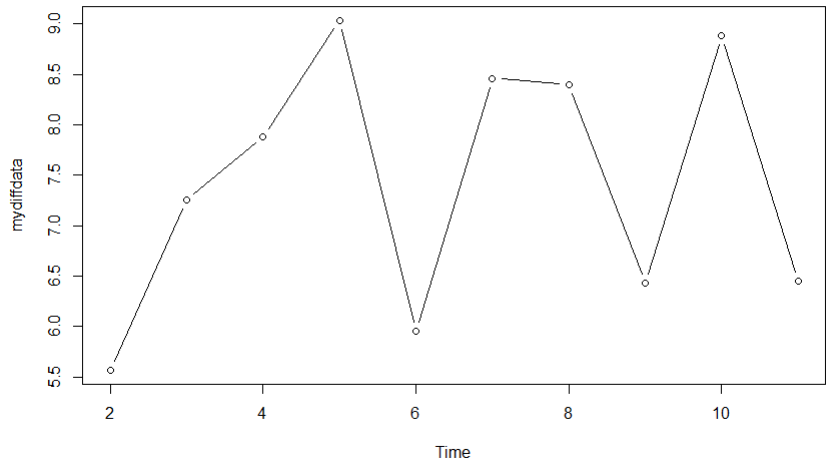
R Code

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

INTRODUCTION

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	0.0963	0.1

Conclusion: Series became stationary after1st differencing

SINGLE EXPONENTIAL SMOOTHING

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

SINGLE EXPONENTIAL SMOOTHING

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

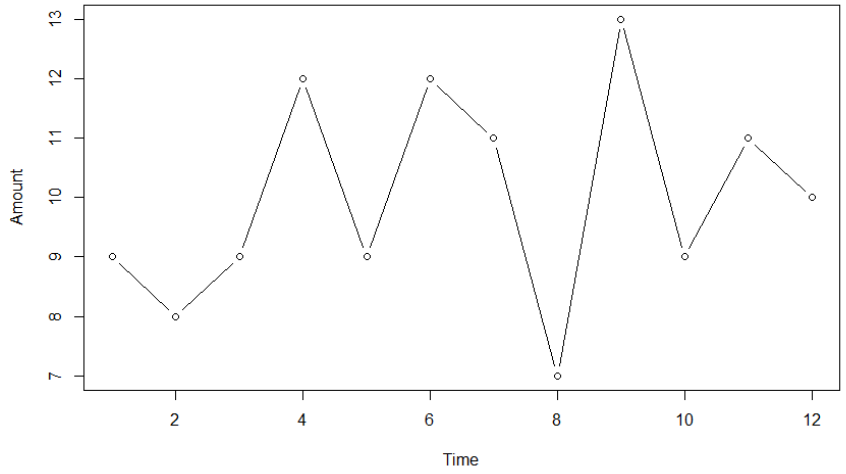
SINGLE EXPONENTIAL SMOOTHING

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



SINGLE EXPONENTIAL SMOOTHING

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(mydata)
> kpss.test(mydata)
```

Test	Statistic	P value
ADF	-2.3285	0.4472
KPSS	0.1157	0.1

KPSS shows that the series is stationary

SINGLE EXPONENTIAL SMOOTHING

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(mydata, beta = FALSE, gamma = FALSE)
> mymodel
```

Smoothing parameter	value
alpha	0.1285076

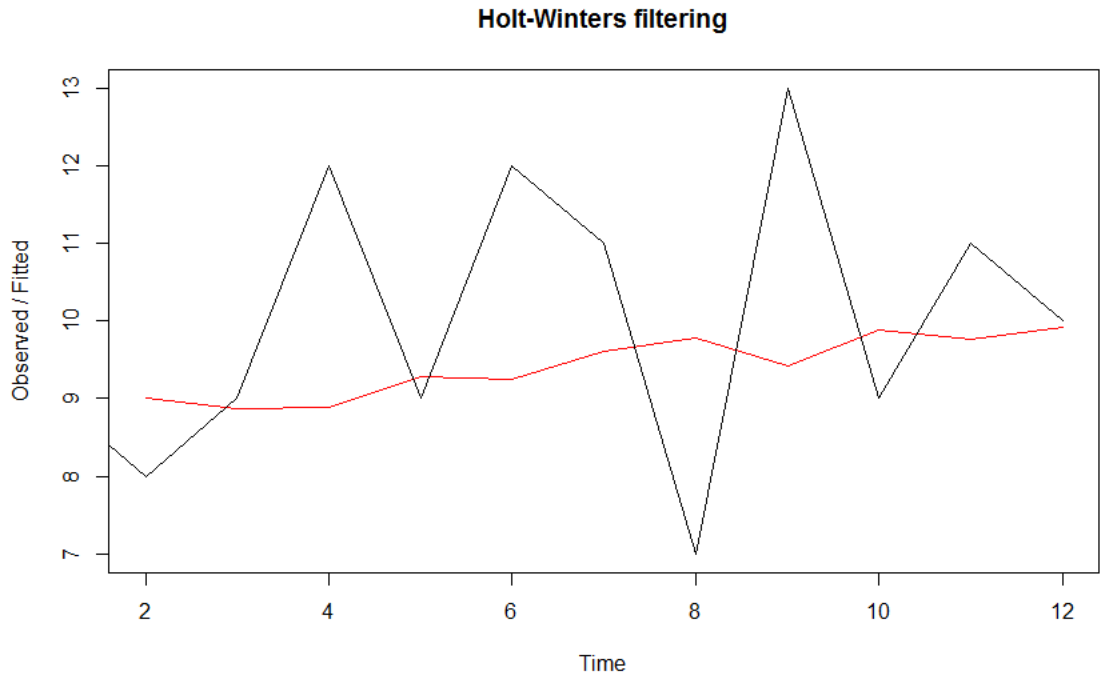
SINGLE EXPONENTIAL SMOOTHING

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)
```



SINGLE EXPONENTIAL SMOOTHING

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(mydata, pred[,1], res)
> write.csv(outputdata, "D:/Deloittee/Material_Part2/outputdata.csv")
```

SINGLE EXPONENTIAL SMOOTHING

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

SINGLE EXPONENTIAL SMOOTHING

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

SINGLE EXPONENTIAL SMOOTHING

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/mydata
```

SINGLE EXPONENTIAL SMOOTHING

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

SINGLE EXPONENTIAL SMOOTHING

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)	15.23%

Criteria

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

SINGLE EXPONENTIAL SMOOTHING

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(error)
> qqline(error)
> shapiro.test(error)
> mean(error)
```

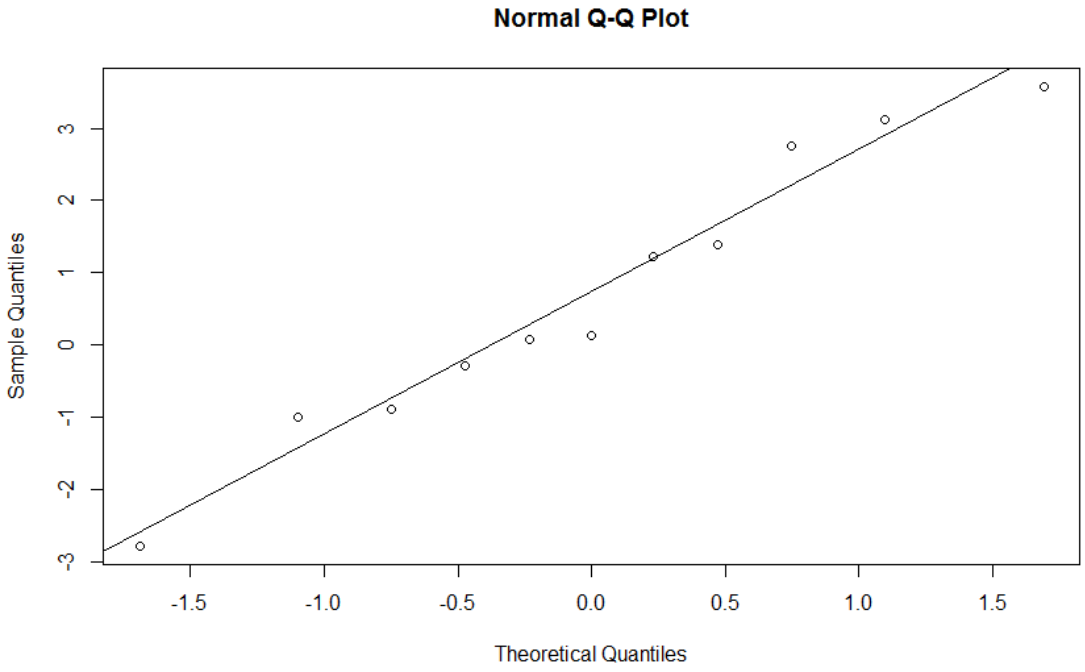
Statistic (w)	P value
0.962	0.7963

Error Mean	0.6638

SINGLE EXPONENTIAL SMOOTHING

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



SINGLE EXPONENTIAL SMOOTHING

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 $S_{13} = 0.1285076 \times 10 + (1 - 0.1285076) \times 9.9293 = 9.9383$

SINGLE EXPONENTIAL SMOOTHING

Forecast

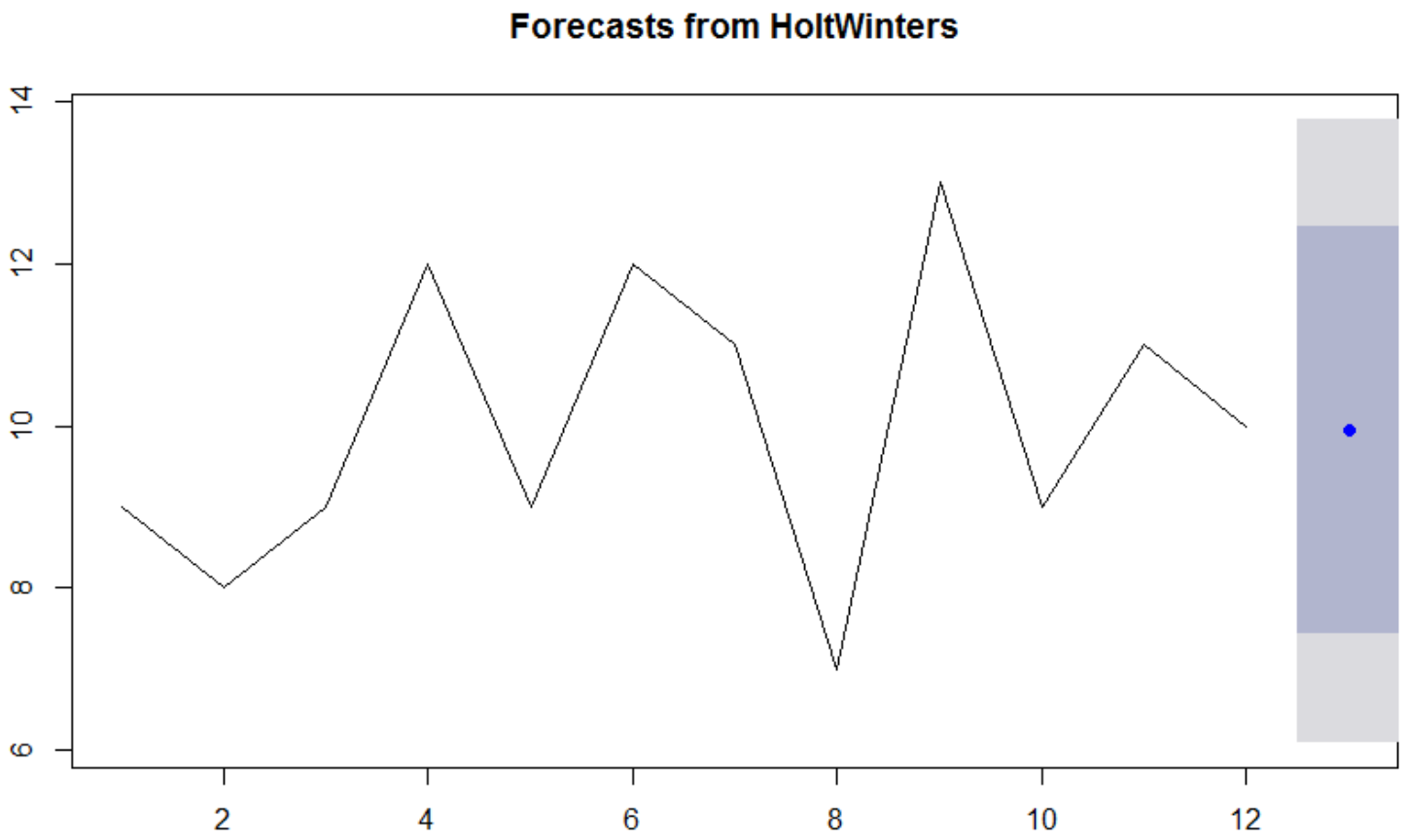
R Code

```
> library(forecast)
> forecast = forecast(mymodel, 1)
> forecast
> plot.forecast(forecast)
```

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

SINGLE EXPONENTIAL SMOOTHING

Forecast Plot



TIME SERIES MODELING

General form of linear model

y is modeled in terms of x's

$$Y = a + b_1x_1 + b_2x_2 + - - - + b_kx_k$$

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

$$y_t = a + b_1y_{t-1} + b_2y_{t-2} + - -$$

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

correlation between y and previous values of y are called autocorrelation

TIME SERIES MODELING

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

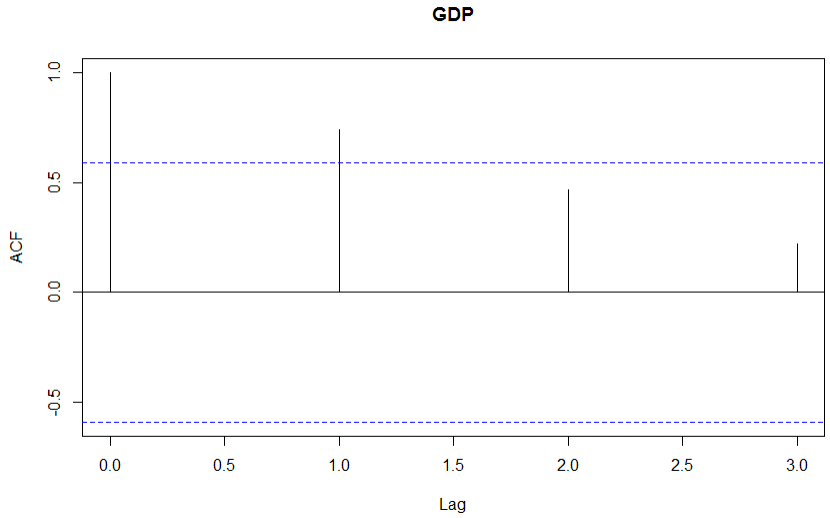
Lag	variables	Auto Correlation
1	y_t vs y_{t-1}	0.9985
2	y_t vs y_{t-2}	0.9984
3	y_t vs y_{t-3}	0.9981

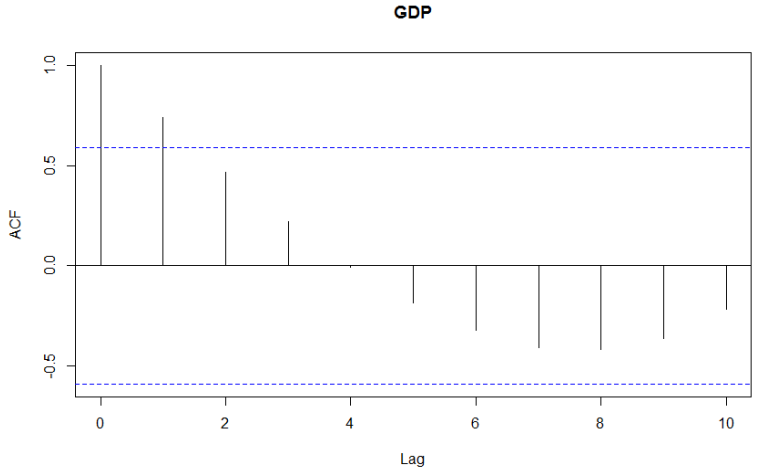
TIME SERIES MODELING

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

R Code

```
> mydata = ts(mydata)
> acf(mydata, 3)
> acf(mydata)
```





TIME SERIES MODELING

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

FORECAST METHODS

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})

FORECAST METHODS

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

FORECAST METHODS

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)

FORECAST METHODS

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

FORECAST METHODS

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

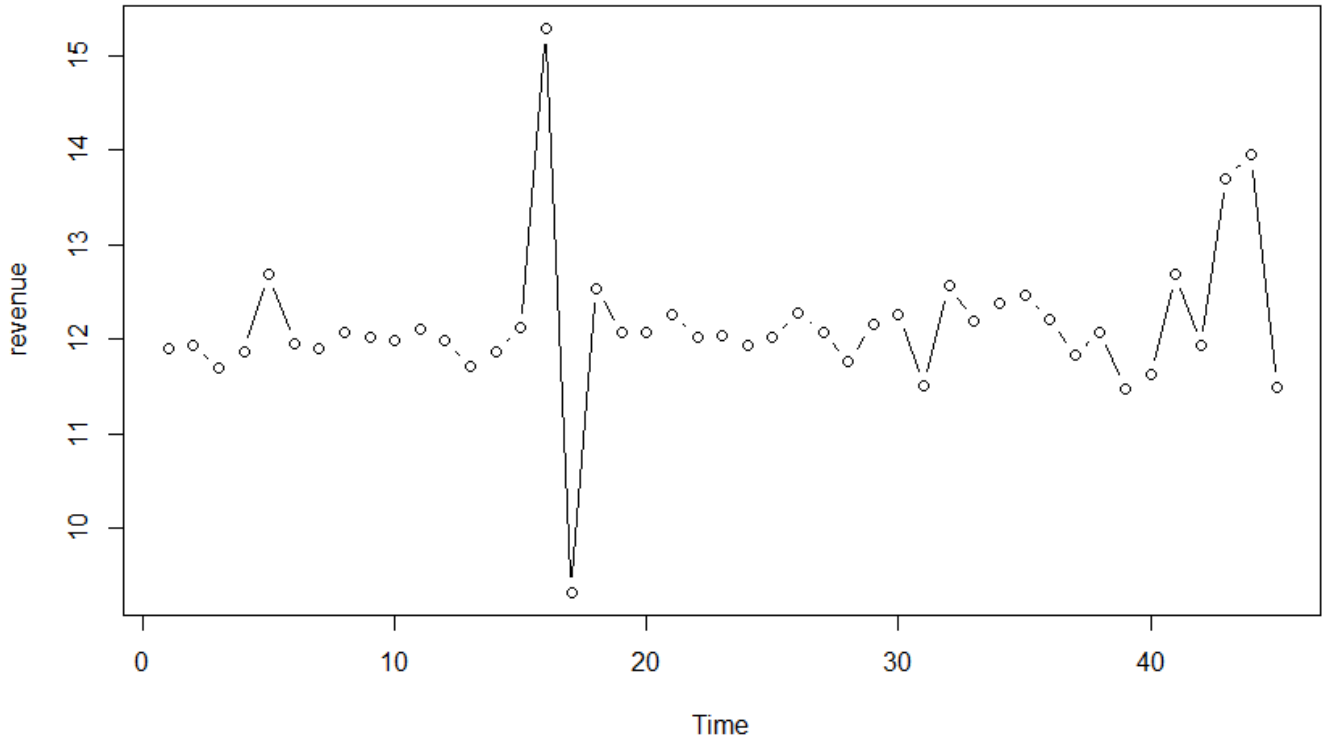
Example: The data daily revenues is given in revenue.csv. Fit Forecasting model using ARIMA?

FORECAST METHODS

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

Step 1: Read and plot the series

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



FORECAST METHODS

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

Step 2: Descriptive Statistics

> summary(mydata)

Statistic	Value
Minimum	9.33
Quartile 1	11.90
Median	12.134
Mean	12.13
Quartile 3	12.26
Maximum	15.28

FORECAST METHODS

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

Step 3: Check whetehr the series is stationary

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

Test	Statistic	P value
ADF	-3.6273	0.04172
KPSS	0.1642	0.1

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

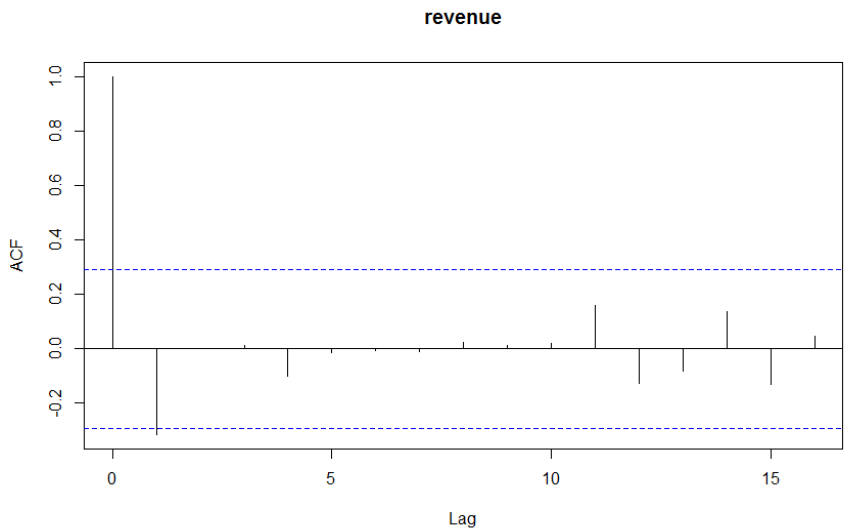
FORECAST METHODS

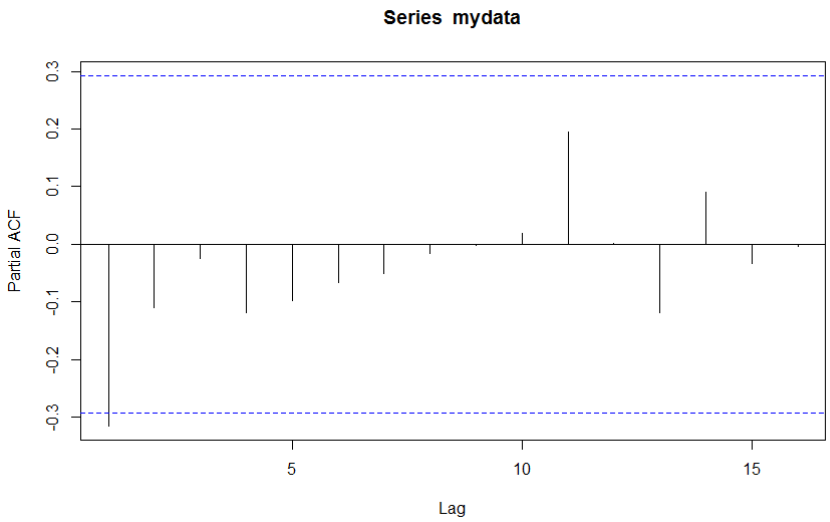
Auto Regressive Integrated Moving Average Models (ARIMA (p,d,q))

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

FORECAST METHODS

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(0,0,1)	-49.9	105.79	111.21

Model Parameters	Value
Intercept	12.1349
MA1	-0.3777

FORECAST METHODS

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

Step 6: Identification of model manually

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

Model		Log likelihood	AIC	BIC
p=1,q=0	ARIMA(1,0,0)	-50.25	106.5	111.92
p=0,q=1	ARIMA(0,0,1)	-49.90	105.79	111.21
p=1,q=1	ARIMA(1,0,1)	-48.84	105.67	112.90

Conclusion:

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

Identified automatically

FORECAST METHODS

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

Step 7: Estimation of parameters

ARIMA(0,0,1) Parameters	Value	Std Error
Intercept	12.1349	0.0689
MA1	-0.3777	0.1651

The model is $y_t = a + \theta_1 e_{t-1}$

$$y_t = 12.135 - 0.3778 e_{t-1}$$

FORECAST METHODS

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

Step 8: Model Diagnostics

> summary(mymodel)

Statistic	Description	Value
ME	Residual average	-0.0042
MAE	Average of absolute residuals	0.4230
RMSE	Root mean square of residuals	0.7321
MAPE	Mean absolute percent error	3.4021

FORECAST METHODS

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

Step 8: Model Diagnostics

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

Normality check on Residuals

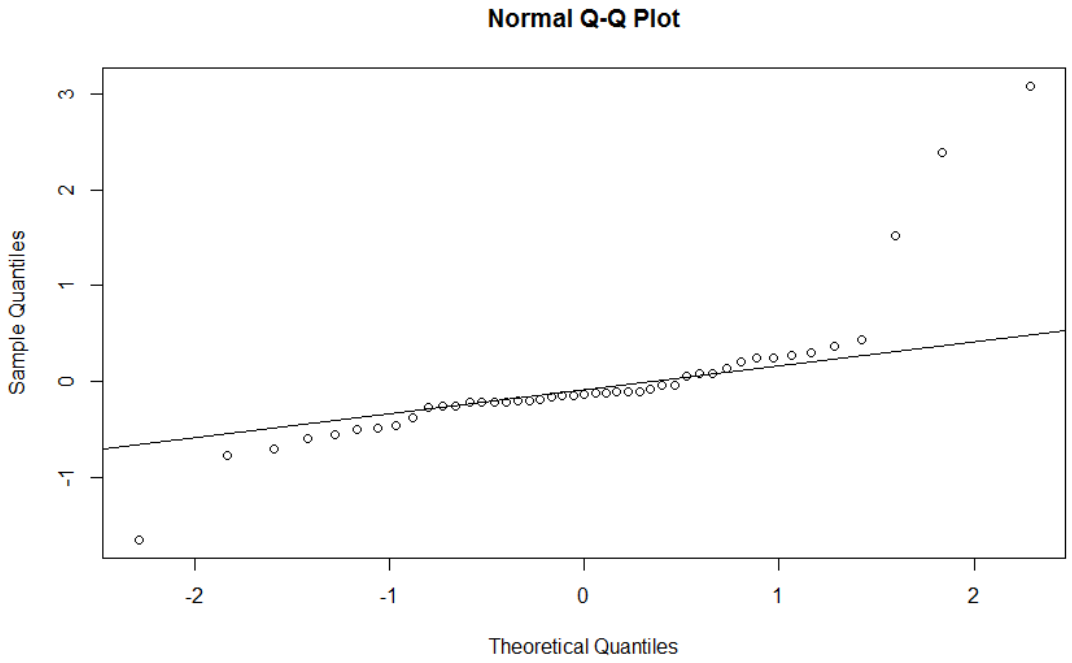
```
> qqnorm(res)
> qqline(res)
> shapiro.test(res)
> hist(res, col = "grey")
```

FORECAST METHODS

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

Step 8: Model Diagnostics

Normality check on Residuals : Normal Q – Q Plot

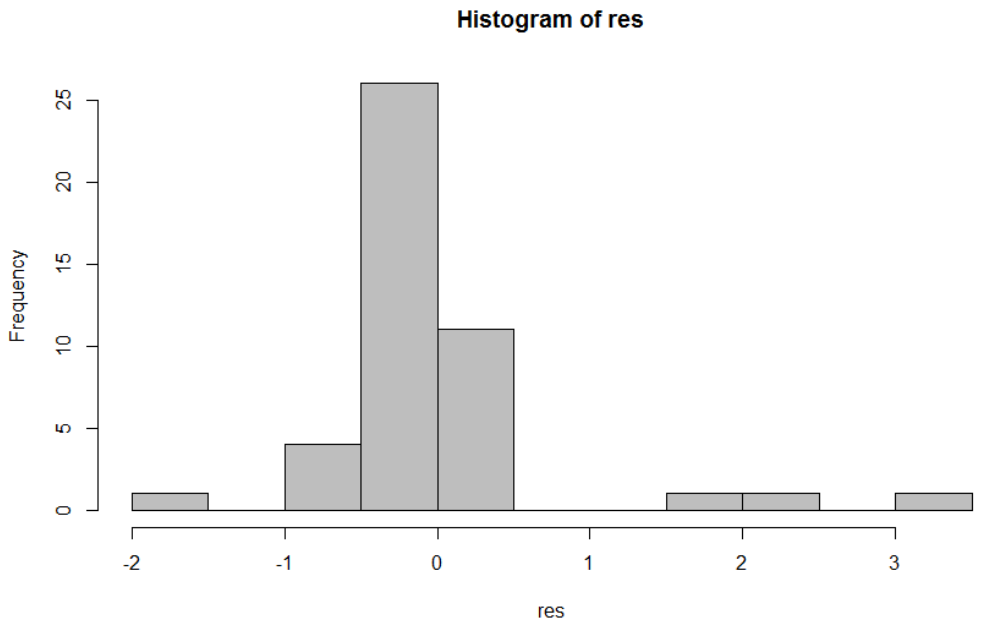


FORECAST METHODS

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

Step 8: Model Diagnostics

Normality check on Residuals: Histogram of Residuals



FORECAST METHODS

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
0.7153	0.000

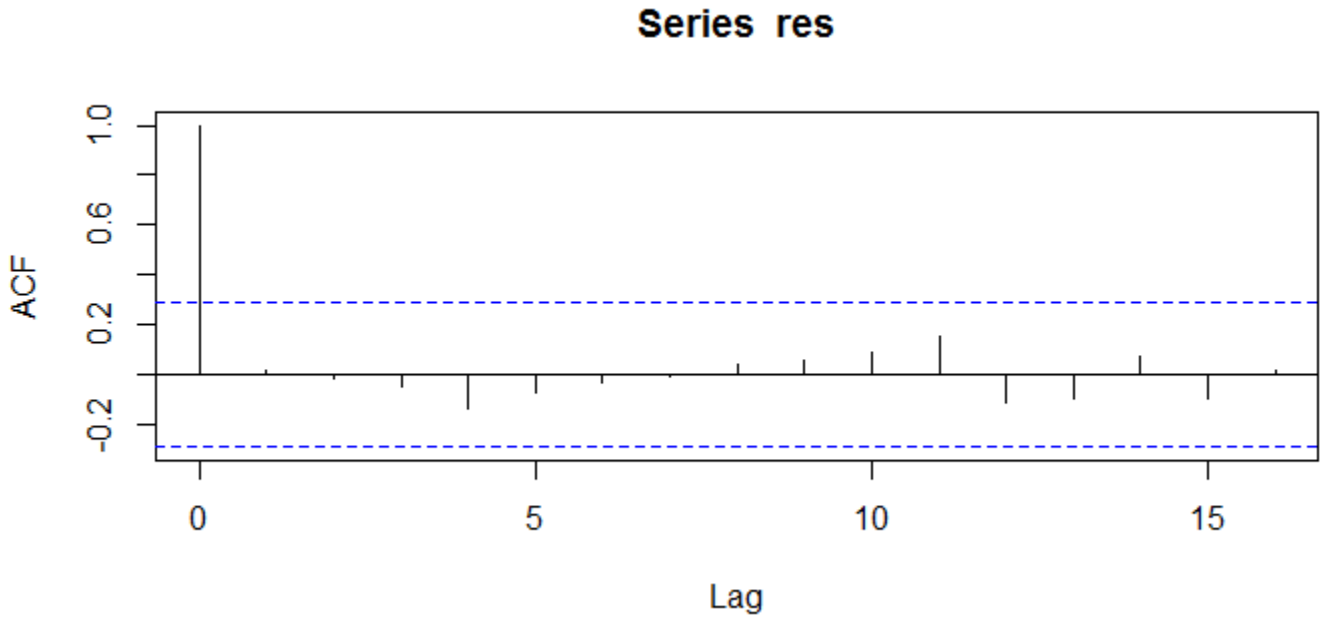
P < 0.05, Residuals are not normal
Ideally residuals should be normally distributed

FORECAST METHODS

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

FORECAST METHODS

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 $\geq$ 0.05, then the residuals are not auto correlated and independent

Portmanteau Test

Another test for autocorrelation of residuals
More powerful than Ljung – Box test
If p value $\geq$ 0.05, then the residuals are not auto correlated and independent or
Residuals are white noise

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

Step 8: Model diagnostics

Portmanteau and Ljung & Box Test

```
> Box.test(res, lag = 15, type = "Ljung-Box")
> library(portes)
> portest(res)
```

Test	Lag	Statistic	df	p value
Ljung & Box	15	6.024	15	0.9793

Since the p value ≥ 0.05, The residuals are not auto correlated

The residuals are white noise

FORECAST METHODS

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

Step 8: Model diagnostics

Portmanteau and Ljung & Box Test

Test	Lag	Statistic	df	p value
Portmanteau	5	0.6409	4.0909	0.9311
	10	1.3224	7.8571	0.9710
	15	2.5557	11.6129	0.9680
	20	3.4440	15.3659	0.9840
	25	4.1059	19.1176	0.9920
	30	5.3043	22.8689	0.9920

Since the p values for both test > 0.05, The model fits the data

The residuals are not auto correlated

The residuals are white noise

FORECAST METHODS

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

Step 9: Forecasting upcoming values

```
> forecast = forecast.Arima(mymodel, h = 3)
> forecast
```

Point	Forecast	80% Prediction Interval		95% Prediction Interval	
		Lower	Upper	Lower	Upper
46	12.03364	11.09541	12.97187	10.59875	13.46853
47	12.13492	11.13199	13.13785	10.60107	13.66877
48	12.13492	11.13199	13.13785	10.60107	13.66877

FORECAST METHODS

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

Exercise 1: 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

FORECAST METHODS

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

Exercise 2: 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

MULTIVARIATE TIME SERIES ANALYSIS

MULTIVARIATE TIME SERIES ANALYSIS

Used to forecast a vector of multiple characteristics or components

Generalization of univariate time series

Denoted by $\mathbf{y}_t = (y_{1t}, y_{2t,} - - -, y_{kt})$

Not only $\{\mathbf{y}_t\}$ have serial dependence within each component $\{y_{it}\}$ but also interdependence between components $\{y_{it}\}$ and $\{y_{jt}\}$,$i \neq j$

Challenges

- As number of components in $\{\mathbf{y}_t\}$ increase, the number of parameters to be estimated also increases
- May not be possible to identify a unique ARIMA model
- Difficult to estimate MA terms
- AR terms can be estimated – VAR modelling

VECTOR AR MODELS

Vector AR or VAR time series model for a time series vector $\{\mathbf{y}_t\}$ of k components is

$$y_{t+1} = \nu + \Phi_1 y_t + \Phi_2 y_{t-1} + - - - + z_t$$

Where $\nu = (\nu 1, \nu 2, - - -, \nu k)'$ is fixed vector of intercept terms and

zt is a k x k matrix with mean vector zero and covariance matrix $\sum$

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

R code : Reading the variables as time series data

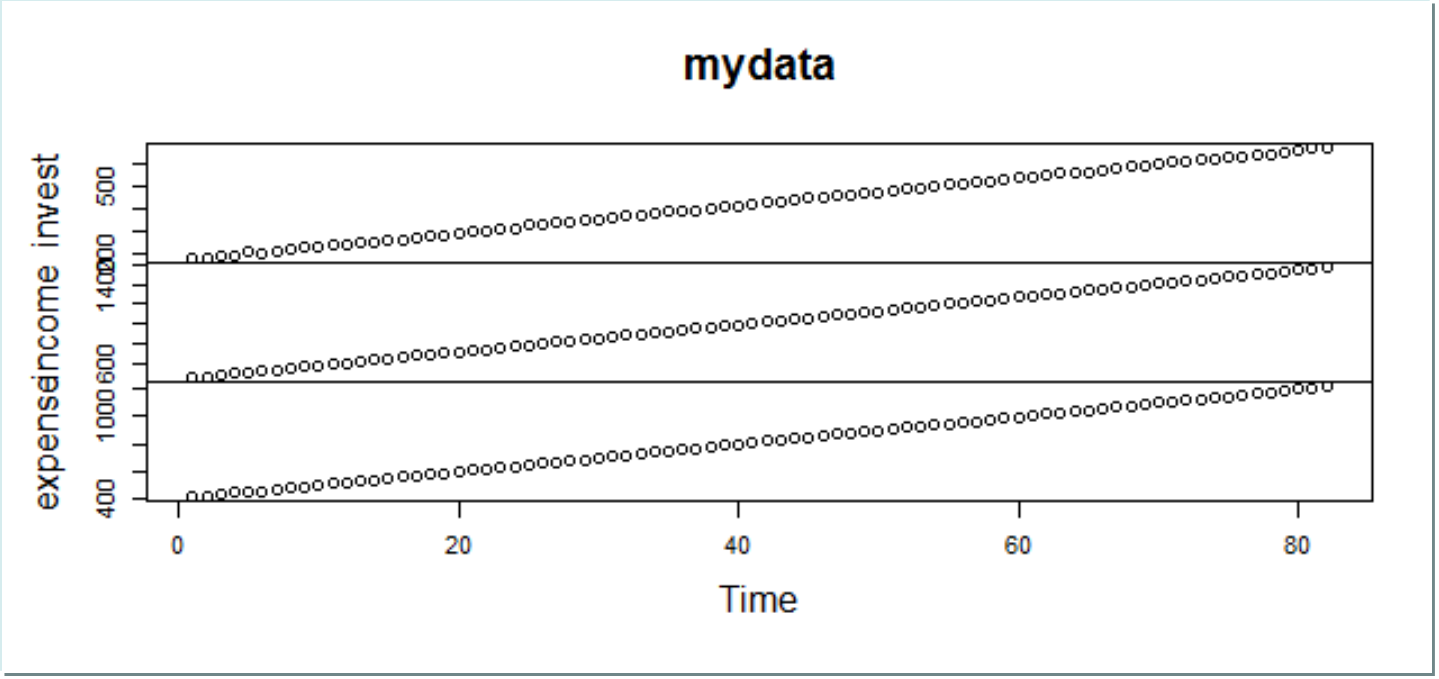
```
> invest = ts(VAR_Data$Investment)
> income = ts(VAR_Data$Disposible_Income)
> expense = ts(VAR_Data$Consumption_Expenditure)
```

Joining the variables as a time series vector

```
> mydata = ts.union(invest, income, expense)
> plot(mydata, type = "b")
```

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?



VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Summary Statistics

```
> colMeans(mydata)
> cov(mydata)
> cor(mydata)
```

Statistics	invest	income	expense
Mean	422.3902	1012.8902	811.2927

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Covariance Matrix

	invest	income	expense
invest	21044.64	47850.85	33812.51
income	47850.85	108840.27	76907.27
expense	33812.5	76907.27	54346.83

Correlation Matrix

	invest	income	expense
invest	1.0000000	0.9998253	0.9998150
income	0.9998253	1.0000000	0.9999666
expense	0.9998150	0.9999666	1.0000000

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Checking for differences required and differencing the series

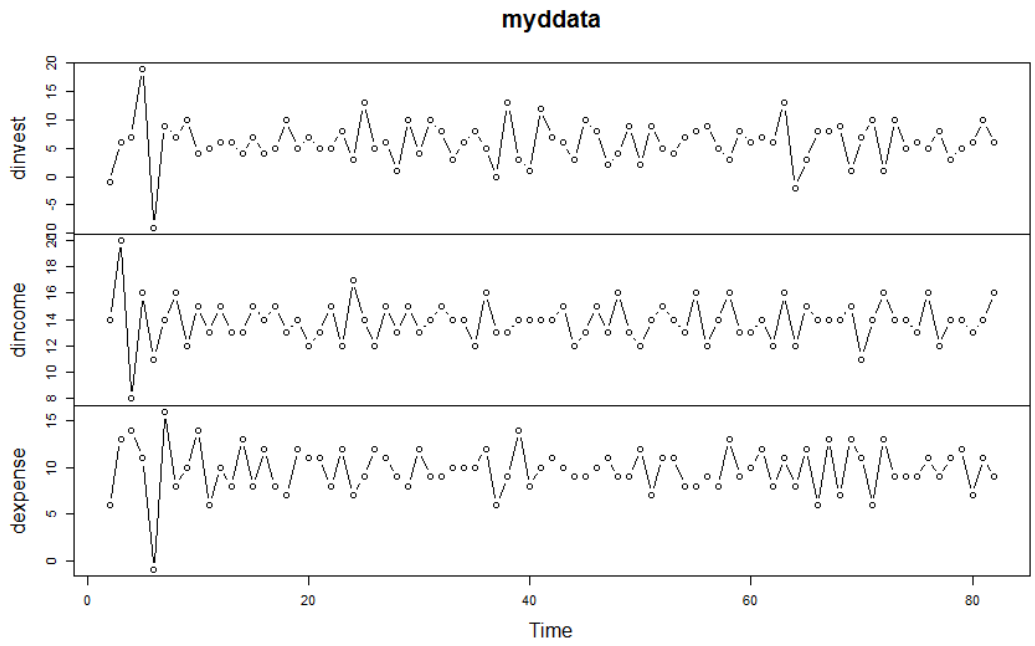
```
> ndiffs(mydata)
```

Number of differences required = 1

```
> dinvest = diff(invest, difference = 1)
> dincome = diff(income, difference = 1)
> dexpense = diff(expense, difference = 1)

> myddata = ts.union(dinvest, dincome, dexpense)
> plot(myddata, type = "b")
```

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

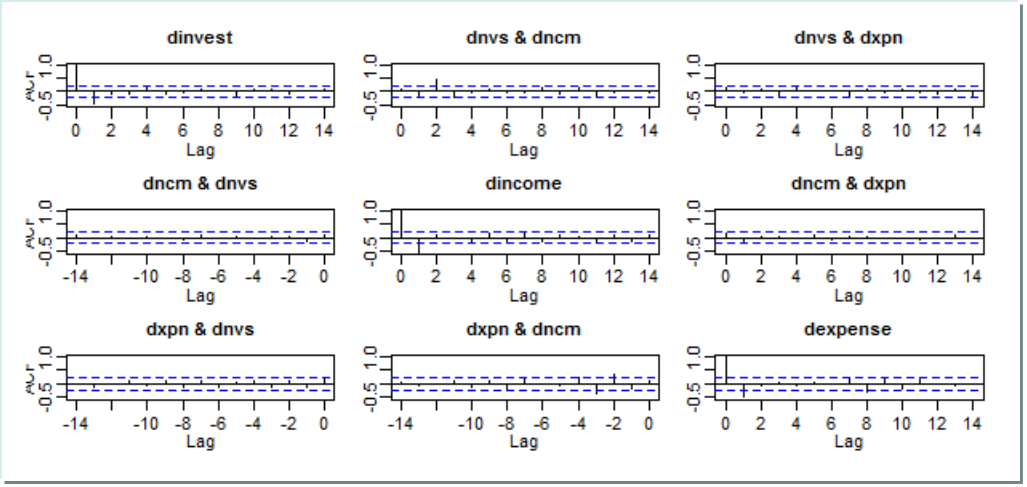


VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Checking auto correlation after differencing

> acf(myddata)



VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Developing the model

```
> mymodel = ar(myddata)
> mymodel$order
> mymodel
```

Best model is VAR(4)

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Model Parameters

Φ_1	dinvest	dincome	dexpense
dinvest	-0.6789	-0.1733	0.0021
dincome	-0.0629	-0.7221	-0.1802
dexpense	-0.0474	-0.1068	-0.7235

Φ_2	dinvest	dincome	dexpense
dinvest	-0.565	0.632	0.079
dincome	-0.013	-0.454	-0.140
dexpense	-0.064	0.228	-0.681

Φ_3	dinvest	dincome	dexpense
dinvest	-0.3965	0.1937	-0.0561
dincome	-0.0405	-0.2196	-0.1415
dexpense	-0.1125	-0.1158	-0.4024

Φ_4	dinvest	dincome	dexpense
dinvest	-0.0971	-0.1537	0.0870
dincome	-0.0385	-0.3317	-0.0700
dexpense	-0.0489	0.1191	-0.2724

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Model Parameters :$\sum$

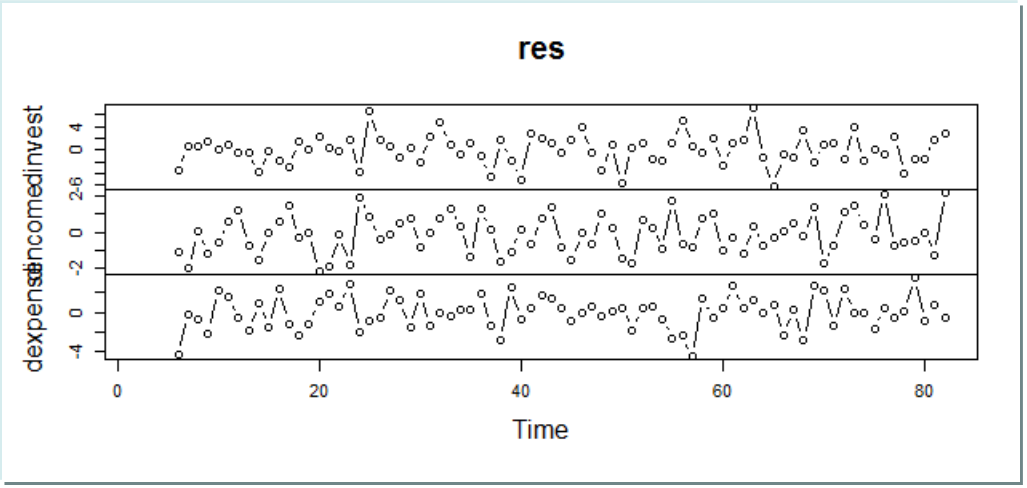
	dinvest	dincome	dexpense
dinvest	9.1154	-0.1968	-0.0112
dincome	-0.1968	1.7633	-0.0504
dexpense	-0.0112	-0.0504	3.7200

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Residual Analysis

```
> res = mymodel$resid
> plot(res, type = "b")
```



VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Forecasting values

```
> forecast = predict(mymodel, n.ahead = 2)
> forecast
```

Point	dinvest	dincome	dexpense
83	3.9310	12.6811	9.7803
84	7.2815	14.1846	10.7613
85	5.7738	13.3762	8.4348

VECTOR AR MODELS

Example: The data on quarterly investment, disposable income and consumption expenditure of an European country for twenty years is given in Var_Data.csv. Develop a time series model to forecast the variables investment, disposable income and consumption expenditure ?

Forecasting actual values

Point	Invest	Income	Expense
83	674	1589	1218
84	681	1603	1229
85	687	1616	1237

COINTEGRATION MODELS

Many variables may be correlated in time

A time series can not only have serial correlation with past values but also can be correlated with other series.

Cointegration develops models based on autocorrelation and correlation with other series

Used to develop a forecasting model with AR and MA terms as well as other factors

Definition

Two non stationary time series $\{x_t\}$ and $\{y_t\}$ are conitegrated if some linear combination of $ax_t + by_t$ with a and b constants is a stationary series

COINTEGRATION MODELS

Steps

1. Check whether the series $\{y_t\}$ and $\{x_t\}$ are non stationary using ADF and KPSS test
2. If yes, check whether the variables $\{y_t\}$ and $\{x_t\}$ are cointegrated using Philips – Ouliaris Cointegration test (PO Test)
3. If p value of PO test is small (< 0.1), the variables are cointegrated
4. Develop a model for $\{y_t\}$ with $\{x_t\}$ as factors using OLS regression without intercept
5. Compute the residuals of the model
6. Fine tune the model prediction by fitting a ARIMA model to the residuals
7. Use the fitted ARIMA model to forecast the residuals
8. With the predicted residuals and value of $\{x_t\}$, forecast the series $\{y_t\}$

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Read Data

```
> sales = Sales$Sales
> rate = Sales$Intrest.Rates
```

Check whether series sales is stationary

```
> library(tseries)
> adf.test(sales)
> kpss.test(sales)
```

Test	Statistics	p value
Dickey-Fuller	-1.4795	0.7707
KPSS	0.5297	0.03498

Both the tests shows that the series is non stationary

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Check whether series interest rate is stationary

```
> adf.test(rate)
> kpss.test(rate)
```

Test	Statistics	p value
Dickey-Fuller	-1.6136	0.7196
KPSS	0.3806	0.0855

ADF test (p value ≥ 005 and KPSS test (p value < 0.1) shows that the series is non stationary

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Check for cointegration of variables

```
> po.test(cbind(sales,rate))
```

Test	Statistics	p value
PO	-22.4542	0.03521

Since p value < 0.05, the series sales and interest rates are cointegrated

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Develop a model for sales in terms of interest rate

```
> mymodel = lm(sales ~ rate)
> summary(mymodel)
```

Statistic	Value
Multiple R-squared	0.9872
Adjusted R-squared	0.9866
Residual standard error	4226
F-statistic	1768
p-value	0.000

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Develop a model for sales in terms of interest rate

```
> mymodel = lm(sales ~ rate + 0)
> summary(mymodel)
```

	Coefficients	Std Error	t statistic	p value
Rate	4775.3	113.6	42.05	0.00

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Compute residuals store as time series

```
> res =residuals(mymodel)
> mydata = ts(res)
```

Check whetehr residuals are stationary

```
> adf.test(mydata)
 > kpss.test(mydata)
```

Test	Statistics	p value
Dickey-Fuller	-1.4722	0.7734
KPSS	0.4854	0.0449

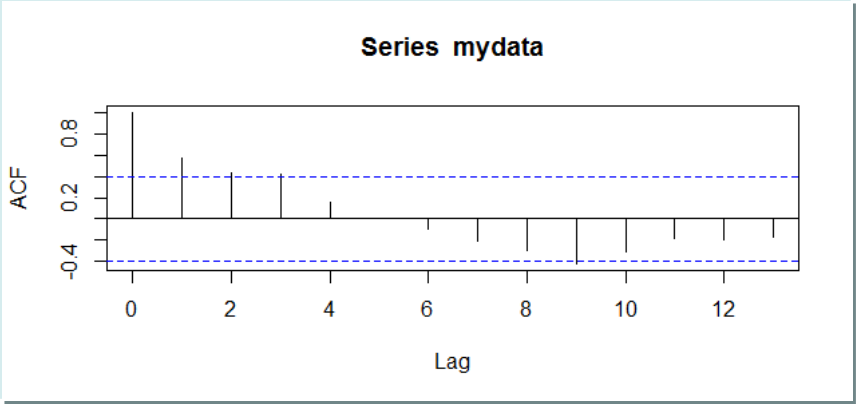
COINTEGRATION MODELS

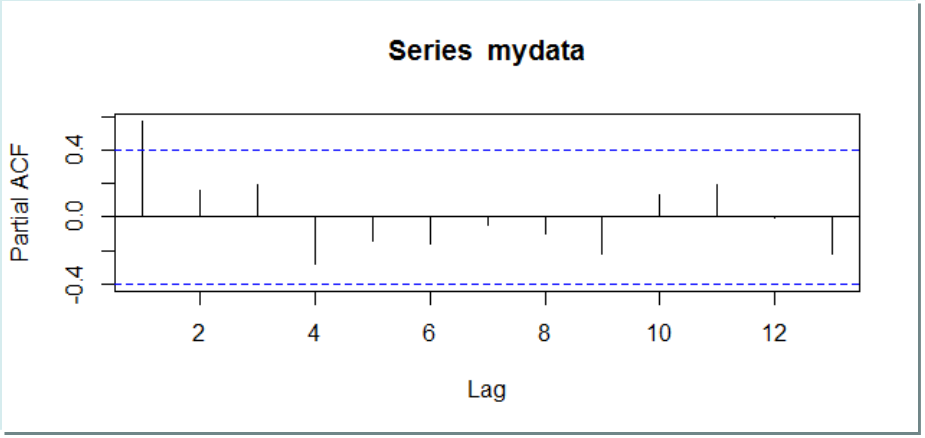
Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Check auto correlation and partial auto correlation functions

> acf(mydata)

> pacf(mydata)





COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Identify the order of differencing required

```
> ndiffs(mydata)
```

Differencing required = 1

Difference the series

```
> myddata = diff(mydata, difference = 1)
```

COINTEGRATION MODELS

Example: Develop a forecasting model for predicting the monthly sales volume a branded car. The data is given in Sales.csv. The data on monthly interest rates for car loans offered by the banks is also given in the file.

Identify best ARIMA model

```
> mymodel = auto.arima(myddata)
> summary(mymodel)
```

Model	ARIMA(0,0,1)
MA(1) Coefficient	-0.4035
Std Error	0.1923
AIC	444.45
BIC	446.72

PANEL DATA ANALYSIS

PANEL DATA ANALYSIS

Panel Data

Also known as longitudinal or cross – sectional time series data

A Dataset in which the behavior of entities are observed across time

Entities could be states, companies, individuals, countries, etc

Allows to control variable that change over time but not across entities (i.e. national policies, federal regulations, international agreements, etc)

PANEL DATA ANALYSIS

Panel Data

Use when interested in analyzing the impact of variables that vary over time

Explores the relationship between predictor and outcome variables within an entity (country, person, company, etc)

Assumptions

Use when interested in analyzing the impact of variables that vary over time

The entity effect need to be controlled otherwise it may affect the predictor variable

Modeling remove the effect of the time invariant entity effect so we can asses the net effect of the predictors on the outcome variable

PANEL DATA ANALYSIS

Panel Data Model

$$y_{it} = \alpha_i + \beta_1 x_{it} + u_{it}$$

Where

α_i : the unknown intercept for each entity, i = 1,2, - - ,n

y_{it}: dependant variable

x_{it}: independent variables

β_i: coefficients of independent variables

u_{it} : error term

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Read data

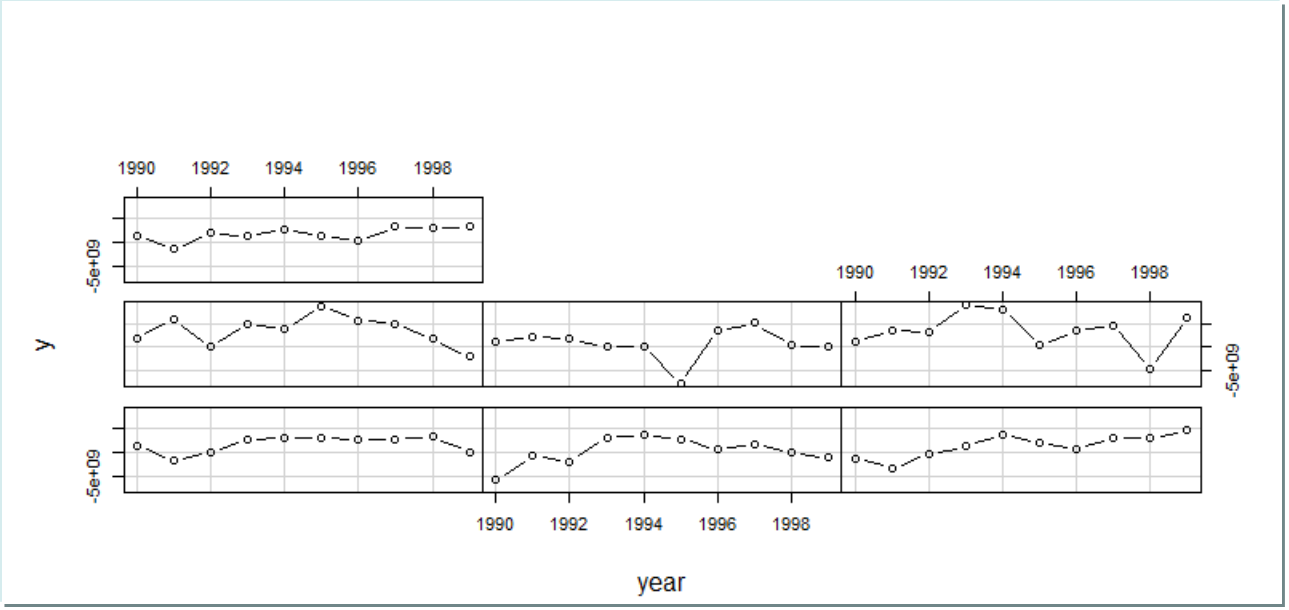
```
> y = mydata$y
> x1 = mydata$x1
> x2 = mydata$x2
> x3 = mydata$x3
> country = mydata$country
> year = mydata$year
```

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

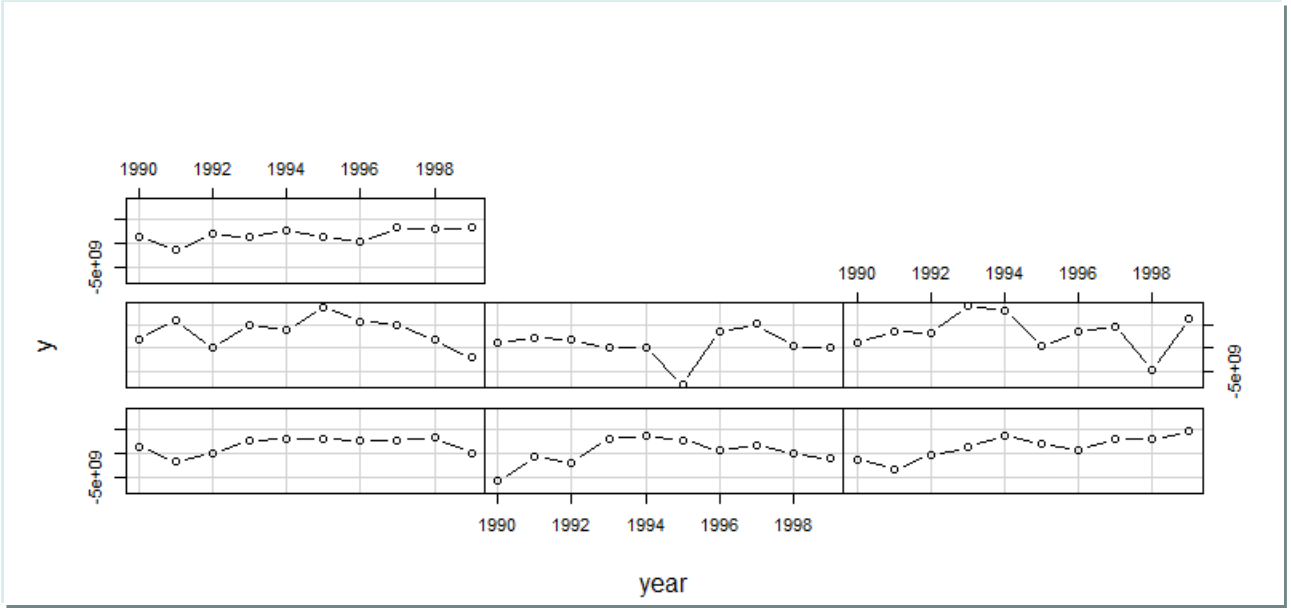
Plot y across years for different countries

> coplot(y ~ year | country, Type = “b”)



PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.



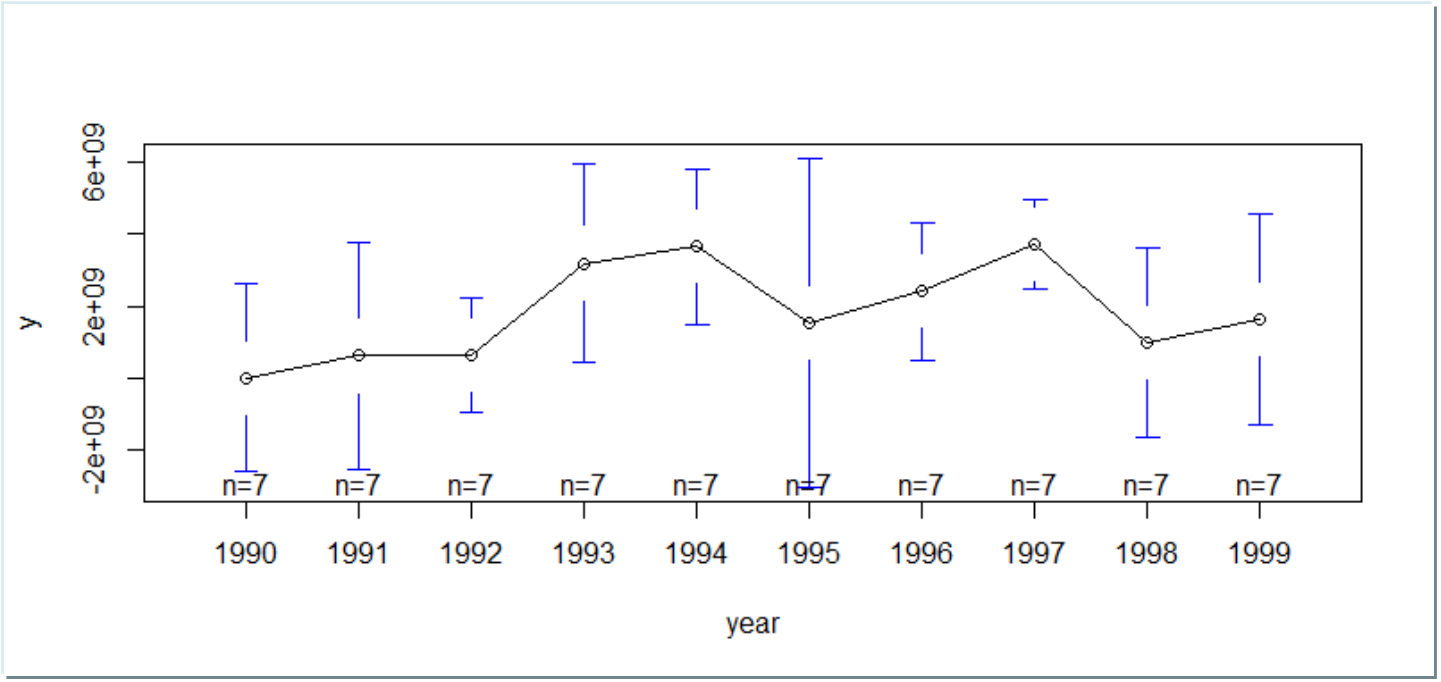
Variation of y across years depends on country.
Country's effect need to be studied or blocked

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Exploring the variation in y across years

> plotmeans(y ~ year)

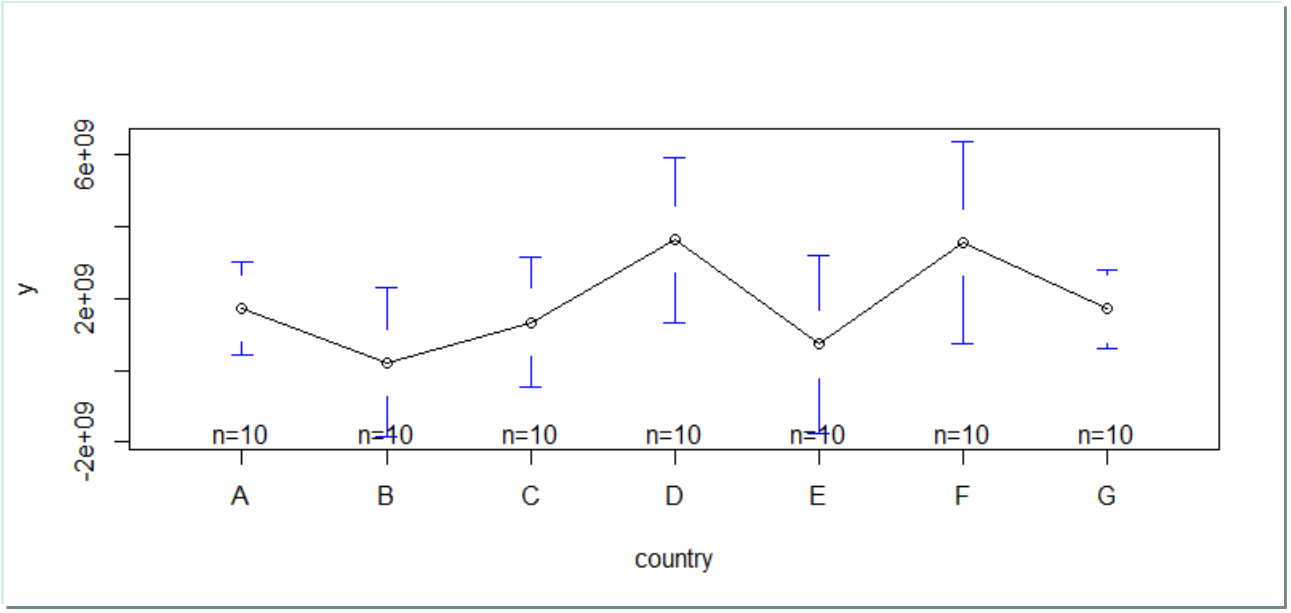


PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Exploring the variation in y across countries

> plotmeans(y ~ country)



PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Fitting an ordinary regression model

```
> regmodel = lm(y ~ x1 + x2 + x3)
> summary (regmodel)
```

Statistic	Value
R square	0.007624
Adjusted R square	-0.03748

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Fitting an ordinary regression model

```
> regmodel = lm(y ~ x1 + x2 + x3)
> summary (regmodel)
```

	Estimate	Std. Error	t	p value
(Intercept)	1.40E+09	7.62E+08	1.837	0.0707
x1	5.59E+08	9.16E+08	0.611	0.5436
x2	8.75E+07	3.50E+08	0.25	0.8032
x3	9.26E+07	2.94E+08	0.315	0.7534

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Fitting panel data model

```
> library(plm)
> plmmodel = plm(y ~ x1 + x2 + x3, data = mydata, index = c("country", "year"), model = "within")
> summary(plmmodel)
```

Type	Code	Value
Groups / Panels	n	7
Years	T	10
Total	N	70

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Fitting panel data model

	Estimate	Std. Error	t	p value
x1	2424529175	1156516240	2.096	0.04
x2	1822699666	2028055945	0.899	0.372
x3	309718343	368552481	0.84	0.404

Only x1 has significant effect on y

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Fitting the best panel data model

```
> plmmodel = plm(y ~ x1 , data = mydata, index = c("country", "year"), model = "within")
> summary(plmmodel)
```

	Estimate	Std. Error	t	p value
x1	2475617828	1106675594	2.237	0.02889

ANOVA

F Statistic	df 1	df 2	p value
5.00411	1	62	0.028892

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Identifying the effect of each country

> fixef(plmmodel)

Country	Effect
A	880542404
B	-1057858363
C	-1722810755
D	3162826897
E	-602622000
F	2010731793
G	-984717494

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Checking whetehr panel data model is better than ordinary regression model

```
> pFtest( plmmodel, regmodel)
```

F Test

F Statistic	df 1	df 2	p value
4.4137	4	62	0.003331

Since p value < 0.05, the panel datal model is better than regression model

PANEL DATA ANALYSIS

Example: The data on an outcome variable (y) and three independent factors (x_1, x_2, x_3) for different countries and across the years is given in panel_data.csv. Develop a model to predict y in terms of x variables.

Computing fitted and residual values

```
> pred = predict(plmmodel)
> res = residuals(plmmodel)
```

MARKET BASKET ANALYSIS

MARKET BASKET ANALYSIS

A modeling technique based upon the logic that if a customer buy a certain group of items, he is more (or less) likely to buy another group of items

Example:

Those who buy cigarettes are more likely to buy match box also.

MARKET BASKET ANALYSIS

Association Rule Mining:

Developing rules that predict the occurrence of of an item based on the occurrence of other items in the transaction

Example

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

{Milk, Bread} → {Biscuits} with probability = 2 / 3

MARKET BASKET ANALYSIS

Itemset:

A collection of one or more items

k – itemset

An itemset consisting of k items

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Support count:

Frequency of occurrence of an itemset

Example

{Milk, Bread, Biscuits} = 2

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Support :

Proportion or fraction of transaction that contain an itemset

Example

{Milk, Bread, Biscuits} = 2 / 5

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

Frequent Itemset

An itemset whose support is greater than or equal to minimum support

MARKET BASKET ANALYSIS

Confidence

Conditional probability that an item will appear in transactions that contain another items

Example

Confidence that Toys will appear in transaction containing Milk & Biscuits

= {Milk, Biscuits, Toys} / {Milk, Biscuits} = 2 / 3 = 0.67

Id	Items
1	Milk, Bread
2	Bread, Biscuits, Toys, Eggs
3	Milk, Biscuits, Toys, Fruits
4	Bread, Milk, Toys, Biscuits
5	Milk, Bread, Biscuits, Fruits

MARKET BASKET ANALYSIS

Association Rule Mining

1. Frequent Itemset Generation

   Fix minimum support value

   Generate all itemsets whose support $\geq$ minimum support

2. Rule Generation

   Fix minimum confidence value

   Generate high confidence rules from each frequent itemset

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

a. Fix minimum support count
b. Generate all itemsets of length = 1
c. Calculate the support for each itemset
d. Eliminate all itemsets with support count < minimum support count
e. Repeat steps c & d for itemsets of length = 2, 3, ---

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Id	Items
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E
5	A,E
6	A,C,E

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 1:

Generate itemsets of length = 1 & calculate support

Item	Support count
A	4
B	3
C	4
D	1
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
D	1
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 2:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A	4
B	3
C	4
E	5

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 3:

generate itemsets of length = 2

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, B	1
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 4:

eliminate itemsets with support count < minimum support count (2)

Item	Support count
A, C	3
A,E	3
B, C	2
B, E	3
C,E	3

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 5:

generate itemsets of length = 3

Item	Support count
A, C, E	2
B, C, E	2

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Step 6:

generate itemsets of length = 4

Itemset	Support Count
A, B, C, E	1

MARKET BASKET ANALYSIS

Frequent Itemset Generation: Apriori Algorithm

Example:

Minimum Support count = 2

Result:

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

MARKET BASKET ANALYSIS

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support count	Support
A, C, E	2	0.33
B, C, E	2	0.33
A , C	3	0.50
A , E	3	0.50
B,C	2	0.33
B,E	3	0.50
C,E	3	0.50

MARKET BASKET ANALYSIS

Association Rule Mining: Apriori Algorithm

Example:

Minimum Support = 0.50

Minimum Confidence = 0.5

Item	Support	Confidence
A → C	0.50	0.75
A → E	0.50	0.75
B → E	0.50	1.00
C → E	0.50	0.75
C → A	0.50	0.75
E → A	0.50	0.60
E → B	0.50	0.60
E → C	0.50	0.60

MARKET BASKET ANALYSIS

Association Rule Mining: Other Measures

Lift

Lift (A $\rightarrow$ C) = Confidence (A $\rightarrow$ C) / Support (C)

Example

Item	Confidence	Support	Lift
A $\rightarrow$ C	0.75	C = 0.67	1.12
A $\rightarrow$ E	0.75	E = 0.83	0.93

Criteria : Lift $\geq$ 1

Lift (A , C) = 1.12 > Lift (A , E) indicates that A has a greater impact on the frequency of C than it has on the frequency of E

MARKET BASKET ANALYSIS

R Code

Reading the file and variables

```
>target = mydata$items
>ident = mydata$Id
```

Making transactions

```
>library(arules)
>transactions = as(split(target, ident),"transactions")
```

Generating rules

```
>library(arules)
>myrules = apriori(transactions, parameter = list(support = 0.5, confidence = 0.25, minlen = 2))
```

MARKET BASKET ANALYSIS

R Code

Displaying rules

```
>myrules
>inspect(myrules)
```

MARKET BASKET ANALYSIS

Exercise 1:

The market basket Software data set contains the details of transaction at a software product company.

1. Identify the frequent product types with a support of minimum 25% ?
2. Also identify the association of products with a confidence of minimum 50% ?
3. What is the chance that Operating System and Office Suite will be purchased together?
4. What is the chance that Operating System and Visual Studio will be purchased together?
5. Estimate the chance that the customers who buy Operating System will also purchase Office Suite ?
6. Estimate the chance that the customers who buy Operating System will also purchase Visual Studio?

CONJOINT ANALYSIS

CONJOINT ANALYSIS

- Application of DoE in Marketing
- A marketing research technique used to determine the desirable features of a product or service

CONJOINT ANALYSIS

Example:

A marketing company wants to enter into the marketing of credit cards. They want to understand the customer preferences on credit cards.

Use Conjoint Analysis to find out the credit card with customer preferred features?

Step 1:

Identify the attributes of the product (Factors)

- Brand
- Service Charge
- Credit limit
- Card Type

CONJOINT ANALYSIS

Example:

A marketing company wants to enter into the marketing of credit cards. They want to understand the customer preferences on credit cards.

Use Conjoint Analysis to find out the credit card with customer preferred features?

Step 2:

Identify 2 or 3 options (Levels) for each attribute

Attribute	Option 1	Option 2	Option 3
Brand	Visa	Master Card	Diners
Servce Charge	Rs. 0	Rs.500	Rs. 1000
Credit Limit	Rs. 50,000	Rs. 1,00,000	Rs. 1,50,000
Card Type	Platinum	Gold	Silver

CONJOINT ANALYSIS

Step 3:

Design the Survey

Attribute	No. of Options	df
Brand	3	2
Servce Charge	3	2
Credit Limit	3	2
Card Type	3	2
	Total df	8

Minimum number of Experiments required = 8 + 1 = 9

Nearest OA = L_9

CONJOINT ANALYSIS

Step4:

Design the Survey

Exp No.	A	B	C	D
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

CONJOINT ANALYSIS

Step4:

Design the Survey

Exp No.	Brand	Service Charge	Credit Limit	Card Type
1	Visa	Rs. 0	Rs. 50,000	Platinum
2	Visa	Rs.500	Rs. 1,00,000	Gold
3	Visa	Rs. 1000	Rs. 1,50,000	Silver
4	Master Card	Rs. 0	Rs. 1,00,000	Silver
5	Master Card	Rs.500	Rs. 1,50,000	Platinum
6	Master Card	Rs. 1000	Rs. 50,000	Gold
7	Diners	Rs. 0	Rs. 1,50,000	Gold
8	Diners	Rs.500	Rs. 50,000	Silver
9	Diners	Rs. 1000	Rs. 1,00,000	Platinum

CONJOINT ANALYSIS

Step 5: Conduct the Survey

Get the rating for each combination from the customers on 1 to 10

10: Most preferred & 1: Least preferred

Exp No.	Brand	ervice Charg	Credit Limit	Card Type	Rating									
1	Visa	Rs. 0	Rs. 50,000	Platinum	9	10	9	9	9	8	9	7	9	9
2	Visa	Rs.500	Rs. 1,00,000	Gold	8	7	8	9	8	7	8	8	7	8
3	Visa	Rs. 1000	Rs. 1,50,000	Silver	3	3	3	4	3	2	3	4	3	3
4	Master Card	Rs. 0	Rs. 1,00,000	Silver	7	7	8	8	6	7	7	9	7	7
5	Master Card	Rs.500	Rs. 1,50,000	Platinum	5	4	5	5	6	7	5	6	7	4
6	Master Card	Rs. 1000	Rs. 50,000	Gold	3	3	4	2	1	3	3	3	4	3
7	Diners	Rs. 0	Rs. 1,50,000	Gold	6	6	6	7	5	5	6	5	6	6
8	Diners	Rs.500	Rs. 50,000	Silver	7	7	9	9	6	7	7	8	8	6
9	Diners	Rs. 1000	Rs. 1,00,000	Platinum	3	2	3	4	3	3	3	3	2	1

CONJOINT ANALYSIS

Step 6: Analyze the Survey Results

Part Worth Utility (Level Averages)

Brand	Utility
Visa	6.6
Master Card	5.2
Diners	5.3

Service Charge	Utility
Rs. 0	7.3
Rs. 500	6.9
Rs. 1000	2.9

Credit Limit	Utility
Rs. 50,000	6.4
Rs. 1,00,000	5.9
Rs. 1,50,000	4.8

Card Type	Utility
Platinum	5.6
Gold	5.5
Silver	5.9

CONJOINT ANALYSIS

Step 7: Calculate Importance Score

Importance Score = Highest Utility Score – Smallest Utility Score

Eg; Brand = 6.6 – 5.2 = 1.4

Attribute	Importance Score	Importance %
Brand	1.4	17.52
Service Charge	4.4	56.41
Credit Limit	1.6	20.51
Card Type	0.4	5.56
Total	7.8	

CONJOINT ANALYSIS

Exercise:

Through Conjoint Analysis identify the critical attributes for improving Employee Satisfaction?