TIME SERIES ANALYSIS

(A case study demonstrating the applications of Box-Jenkins Methodology involving trend, seasonality and cyclicity analysis, stationarity, modelling and forecasting using ARIMA models for some macro-economic and business time series.)

Submitted by: SAIKAT GHOSH

Due Date: 11th April, 2016

Date of Submission: 11th April, 2016

Instructor's Remarks:

2. Time Series Analysis (17 Cases)

2.1. Estimation and Removal of Deterministic Components

2.1.1. Testing the Presence of Trend, it's Estimation and Removal

2.1.1.1. Testing for the presence of trend, its estimate and removal for consumption expenditure (in million dollars) for the United States for 1944 to 2000 obtaining trend forecasts.

2.1.1.2. Testing for the presence of trend, its estimate and removal for the Following world development indicators for India and obtaining trend forecasts.

A. Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD,
B. Population Total,
C. Gross Domestic Product (GDP) (current USD),
D. Gross Domestic Product (GDP) Growth (annual %) and
E. Life Expectancy at birth (years)

2.1.1.3. Testing for the presence of trend, its estimate and removal for the annual sales measured in million USD for a trading company for 1994-2013 obtaining trend forecasts.

2.1.2. Testing the Presence of Seasonality, it's Estimation and Removal

2.1.2.1. Testing the Presence of Seasonality, its Estimation and Removal for monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India. Obtaining additive decomposition and forecasting based on deterministic components.

2.1.2.2. Testing the Presence of Seasonality, its Estimation and Removal for monthly World Airline Passengers from 1949-1960. Obtaining additive decomposition and forecasting based on deterministic components.

2.1.2.3. Testing the Presence of Seasonality, its Estimation and Removal for the quarterly demand for an industrial good measured in thousand units for a manufacturing company for 2001-2005. Obtaining additive decomposition and forecasting based on deterministic components.

2.2. Modeling the Random Component using Auto Regressive Integrated Moving Averages (ARIMA)

2.2.1. Testing for Stationarity and making the Series Stationary if it is not

2.2.1.1. Testing Stationarity of the de-trended series for consumption expenditure (in million dollars) for the United States for 1944 to 2000.

2.2.1.2. Testing Stationarity of the de-trended series for the Following world development indicators for India:

A. Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD,
B. Population Total,
C. Gross Domestic Product (GDP) (current USD),
D. Gross Domestic Product (GDP) Growth (annual %) and
E. Life Expectancy at birth (years)

2.2.1.3. Testing Stationarity of the de-trended series for the annual sales measured in million USD for a trading company for 1994-2013 obtaining trend forecasts.

2.2.1.4. Testing Stationarity of the Estimated Random Component for monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India.

2.2.1.5. Testing Stationarity of the Estimated Random Component for monthly World Airline Passengers from 1949-1960.

2.2.1.6. Testing Stationarity of the Estimated Random Component for the quarterly demand for an industrial good measured in thousand units for a manufacturing company for 2001-2005.

2.2.1.7. Testing Stationarity of the Estimated Random Component for a simulated AR(1) time Series

2.2.2. Identification of the Order of the ARIMA Model

2.2.2.1. Identifying the order of the ARIMA model for a simulated AR(1) time Series

2.2.2.2. Identifying the order of the ARIMA model for a simulated MA(1) time Series

2.2.3. Building ARIMA Model and Forecasting

2.2.3.1. Modeling a simulated Gaussian AR(1) time Series using ARIMA Model while doing the following objectives:

- Test for stationarity of the data using the Augmented Dickey Fuller (ADF) Test. Make the series stationary if it is not.
- Fit an 'appropriate' order (identify it using sample Correlogram and sample Partial Correlogram) of ARMA model.
- Check the goodness of the model by using the following:
    a. Stationary R-Square
    b. Root Mean Square Error (RMSE)
    c. Absolute Percentage Error (MAPE)

- Validate the assumption of driving Gaussian White Noise using the following:
    a. Ljung–Box Test for White Noise
    b. ACF and PACF for White Noise
    c. Q-Q Plot for Normality
- Assess the goodness of model built on simulated data by checking if the estimates are close to the parameters?
- Apply the model and forecast for next 20 time points.

2.2.3.2. Modeling a simulated Gaussian AR(1) time Series using ARIMA Model while doing the following objectives:

- Test for stationarity of the data using the Augmented Dickey Fuller (ADF) Test. Make the series stationary if it is not.
- Fit an 'appropriate' order (identify it using sample Correlogram and sample Partial Correlogram) of ARMA model.
- Check the goodness of the model by using the following:
    a. Stationary R-Square
    b. Root Mean Square Error (RMSE)
    c. Absolute Percentage Error (MAPE)

- Validate the assumption of driving Gaussian White Noise using the following:
    a. Ljung–Box Test for White Noise
    b. ACF and PACF for White Noise
    c. Q-Q Plot for Normality
- Assess the goodness of model built on simulated data by checking if the estimates are close to the parameters?
- Apply the model and forecast for next 20 time points.

Inroduction to Time Series :

Arrangement of statistical data in chronological order is called the Time Series. Time series analysis comprises methods for analyzing time series in order to extract meaningful statistics and other characteristics of the data. A time series is a sequence of data points made over a continuous time interval using equal spacing between every two consecutive measurements. Examples of time series are ocean tides, counts of sunspots.

Time series data have a natural temporal ordering. This makes the time series analysis distinct from cross-sectional studies, in which there is no natural ordering of the observations. Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations

A time series is a sequence of data points made:

- over a continuous time interval
- out of successive measurements across that interval
- using equal spacing between every two consecutive measurements
- with each time unit within the time interval having at most one data point

One simple method of describing a series is that of classical decomposition. The notion is that the series can be decomposed into four elements:

- Trend (Tt) — long term movements in the mean;
- Seasonal effects (St) — cyclical fluctuations related to the calendar;
- Cycles (Ct) — other cyclical fluctuations (such as a business cycles);
- Residuals (Et) — other random or systematic fluctuations.

The idea is to create separate models for these four elements and then combine them :

- additively $Xt = Tt + It + Ct + Et$
- multiplicatively $Xt = Tt \cdot It \cdot Ct \cdot Et$.

Most time series patterns can be described in terms of two basic classes of components: trend and seasonality. The former represents a general systematic linear or (most often) nonlinear component that changes over time and does not repeat or at least does not repeat within the time range captured by our data (e.g., a plateau followed by a period of exponential growth). The latter may have a formally similar nature (e.g., a plateau followed by a period of exponential growth), however, it repeats itself in systematic intervals over time. Those two general classes of time series components may coexist in real-life data.

Estimation and Removal of Deterministic Components

Testing the Presence of Trend

1. Relative Ordering Test for the Presence of Trend:

Let $y1, \dots, yn$ be the observed time series and we are interested in testing H_0 of no trend. Define:

qij = 1 if $yi > yj$ when $i < j$

0 otherwise

Let Q = Σ Σ qij : # of decreasing points/ # of discordants.

$i<j$

Under H_0 of no trend dicordants and concordants will be equally probable, i.e. $(qij = 0) = 1/2 = (qij = 1)$, hence, $(qij) = 1/2$

$\Rightarrow E(Q) = \Sigma\Sigma (qij) = n(n-1)/4$

$i<j$

If $Q \gg E(Q)$ then it is indicative of a falling trend, otherwise if $Q \ll (Q)$ then it is indicative of a rising trend.

Calculate $\tau = 1 - 4Q/(n-1)$

Clearly $(\tau) = 0$ and $(\tau) = 2(2n+5)/9(n-1)$

Define the test statistic as follows:

$Z = \frac{\tau-(\tau)}{\sqrt{V(\tau)}}$ ~N(0,1) asymptotically under H_0.

Hence we can use the normal test. If H_0 is rejected then on the basis of relation between Q and $E(Q)$

we can decide if the trend is present or not. After this we can move towards estimation and

elimination of trend.

2. Friedman Test for testing the Presence of Seasonality

Let $y_1, \dots, y_n$ be the observed time series and we are interested in testing H_0 of no seasonality. Friedman described the following test procedure for monthly seasonality:

Step 1: Remove trend if necessary.

Step 2: Rank the de-trended data within each year from smallest (1) to largest (12).

Let M_{ij} be the rank of ith month in jth year, then under the assumption of no seasonality for a particular year j, $\{M_{1j}, \dots, M_{12j}\}$ can be any random permutation of $\{1, \dots, 12\}$.

Step 3: Obtain the monthly totals of the ranks across different years and call them as $M_1, \dots, M_{12}$.

The average of the sequence $1,2, \dots, r$ is $\frac{(r+1)}{2}$, hence $\frac{c(r+1)}{2}$ represents the expected rank sum under no seasonality where r is no. of months in an year and c is the total number of years.

Step 4: Calculate the test statistic:

$$X = \frac{2\sum_{j=1}^{r}\left(M_j - \frac{c(r+1)}{2}\right)^2}{c(r+1)} \sim \chi^2_{r-1}$$

Hence we can use the chi-square test. After this we can move towards estimation and elimination of seasonality.

ESTIMATION AND REMOVAL OF TREND

- ## Least square estimation.

Consider the additive decomposition

$$y_t = m_t + s_t + c_t + e_t$$

Equivalently $y_t = m_t + \delta_t$

Where $\delta_t = s_t + c_t + e_t$ consists of seasonal composition, cyclic component and random component. Least square estimation of trend involves estimating m_t such that

$$\sum_{t=1}^{n} \delta_t^2 = \sum_{t=1}^{n} (y_t - m_t)^2$$

Is minimized. Appropriate shape of m_t can be guessed by means of plotting the available time series.

- ## Moving Average Method

This method consists in measurement of trend by smoothing out by fluctuations of the data by means of a moving average trend. Consider the additive decomposition.

$y_t = m_t + s_t + c_t + e_t$

Equivalently $y_t = m_t + \delta_t$

Where $\delta_t = s_t + c_t + e_t$ consists of seasonal composition, cyclic component and random component.

Moving average trend estimation involves estimating m_t by moving average of the observed time series with an appropriate span.

If the appropriate span is an odd integer, i.e. of the form 2k+1 where k is a positive integer, then moving average trend estimate is given by,

$$\hat{m}_t = \frac{\sum_{i=-k}^{k} y_{t+i}}{2K+1}$$

when appropriate span is an odd integer then problems is that the moving average is to be taken as an estimate of trend at which time point as there is no midpoint of the span. In that case moving average trend estimate is given by moving average with span 2 of the moving averages which removes the aforesaid problem.

If the appropriate span is an even integer, i.e. of the form 2k where k is a positive integer, then moving average trend estimate is given by,

$$\hat{m}_t = \frac{1}{2}\left(\frac{1}{2k}\sum_{i=-k}^{k-1} y_{t+i} + \frac{1}{2k}\sum_{i=-(k-i)}^{k} y_{t+i}\right)$$

the above trend estimate is also called as centered moving average.

The span of the moving average determines the smoothness of the trend-cycle estimate. In general, a larger order means a smoother curve. The appropriate span of the moving average can be taken as arithmetic average of the periods between different peaks in the observed time series. After estimating the trend using an appropriate method, trend eliminated series can be obtained as $y_t - \hat{m}_t$ where $\hat{m}_t$ is the estimated trend.

Seasonality analysis

Seasonality is defined as recurring pattern in the data generally within a year. Consider the additive decomposition: $y_t = m_t + s_t + c_t + e_t$ and arrange the observed time series by years and months. Let $y_{j,k}$ represents observation corresponding to j^{th} year in k^{th} season, j=1,2,...,J and k=1,2,...,n_s.

Testing for presence of seasonality – friedman(JASA) test

Let $y_1,...,y_n$ be the observed time series and we are interested in testing H_0 of no seasonality. Friedman described the following test produce for monthly seasonality.

1) Remove the trend if necessary.
2) Rank the detrended data within each year from smallest to largest. Let Mij be the rank of ith month in jth year, then under the assumption of no seasonality for a particular year j,{M1j,...,M12j} can be any random permutation of {1,2,...12}.
3) Obtain the monthly totals of the ranks across different years and call them as M1,...,M12. The averages of the sequence 1,2,...,r is $\frac{r+1}{2}$, hence $\frac{c(r+1)}{2}$ represents the expected rank sum under no seasonality where r is no. of months in an year and c is the total number of years.
4) Calculate the test statistic:

$$X = \frac{2\sum_{j=1}^{r}\left(M_j - \frac{c(r+1)}{2}\right)^2}{c(r+1)} \sim \chi^2_{r-1}$$

Hence we can use the chi-square test. After this we can move towards estimation and elimination of seasonality.

Estimation and elimination of seasonality

- **<u>Small trend method</u>**

This method works well in case if the trend is small, i.e. constant within a year.

1) Estimate the trend as the average of observations within the period of seasonality:

$$\hat{m}_{j,k} = \frac{\sum_{k=1}^{n_s} y_{j,k}}{n_s} \quad j=1,...,J \text{ and } k=1,... n_s$$

2) Estimate the seasonal decomposition as follows:

$$S_k = \frac{\sum_{k=1}^{n_s}(y_{j,k} - \hat{m}_{j,k})}{j}; \quad k=1,...,n_s$$

After estimating the seasonality using an appropriate method, seasonality eliminated series can be obtained as $y_t - \hat{s}_t$ where $\hat{s}_t$ is the estimated seasonality.

- **<u>Rapidly changing trend method</u>**

This method works well in case if the trend is volatile, i.e. changes even within a year.

1) Estimate the trend using moving averages with span equal to the period of seasonality.
2) Detrended the data by subtraction.
3) Let Dj,k be the detrended data then define the estimate of seasonality as follows:

$$s_k = D_{j,k} - \frac{\sum_{k=1}^{n_s} D_{j,k}}{n_s}; \quad k = 1,...,n_s$$

after estimating the seasonality using an appropriate method, seasonality eliminated series can be obtained as $y_t - \hat{s}_t$ where $\hat{s}_t$ is the estimated seasonality.

<u>Estimation and elimination of cyclicity-residual approach</u>

It is a crude method of estimating the cyclic component which is described as follows.

1) Estimate the trend and seasonality and then eliminate them, cyclic and random component are left.
2) Average out the random component using some low-pass filter like moving average. The left series will be an estimate of the cyclic component.

Case 1: Test for the presence of trend and estimate it if it's present for consumption expenditure (in million dollars) for the United States for 1944 to 2000 using appropriate test and method. Obtain the de-trended consumption series. Also provide a simple trend based forecast for the consumption expenditure for the next 5 years.

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 10.6311

p value: 0

No. of Discordants: 3

Expected No. of Discordants: 715.5

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

For Consumption:

Linear

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.981	.963	.962	292.837

Cubic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.997	.997	84.659

Quadratic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.997	.997	88.751

ANOVA

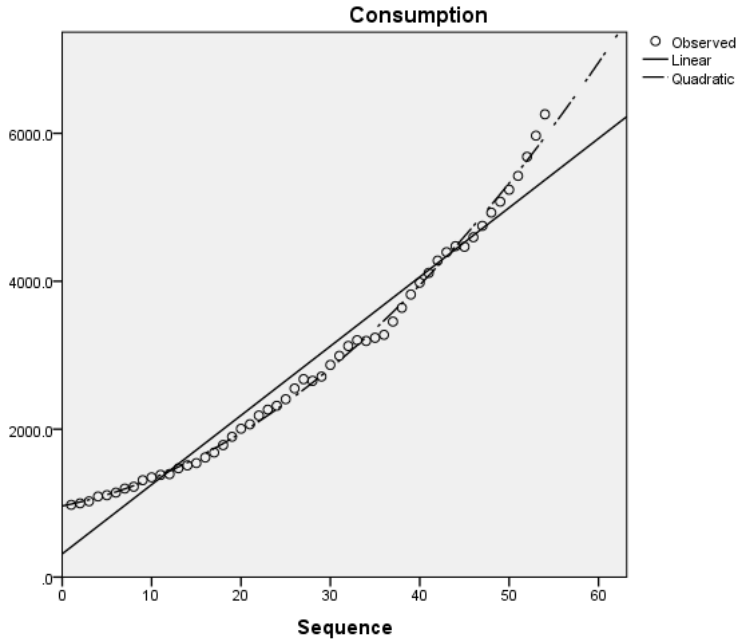
	Sum of Squares	df	Mean Square	F	Sig.
Regression	118991850.747	2	59495925.373	7553.391	.000
Residual	401712.587	51	7876.717		
Total	119393563.333	53			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	24.180	3.155	.253	7.663	.000
Case Sequence ** 2	1.262	.056	.751	22.696	.000
(Constant)	962.181	37.617		25.578	.000

The equation of trend is $m_t = 962.181 + 24.180t + 1.262t^2$

Inference: We observe that the value of Adjusted R square of Non-linear Regression model is greater than Linear Regression model (.997>.963). Thus, Quadratic model (or we can choose cubic) fits the data better.



Inference: On fitting the quadratic, cubic and linear polynomial to the given data, we observe that the Quadratic model has the highest value of adj. R square. Thus for Quadratic Regression Model, the coefficient of determination is 0.997 which implies that there is 99.7% variation in consumption expenditure. Estimation of trend can be done by using least squares method for quadratic model.

Now, we obtain the De trended time series by subtracting the trend value from original data:

Case Summaries[a]

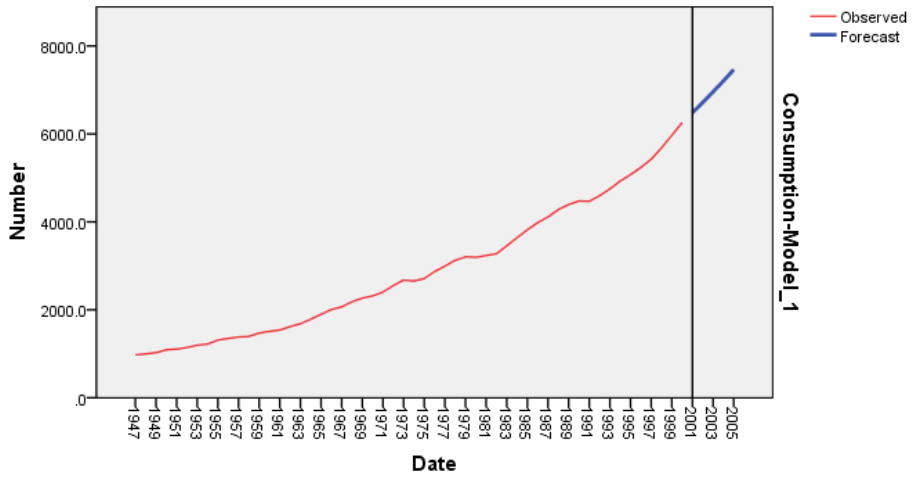
	Year	Consumption	Estimated Trend Value	De trended Value
1	1947	976.4	987.62333	-11.22333
2	1948	998.1	1015.59040	-17.49040
3	1949	1025.3	1046.08200	-20.78200
4	1950	1090.9	1079.09814	11.80186
5	1951	1107.1	1114.63882	-7.53882
6	1952	1142.4	1152.70404	-10.30404
7	1953	1197.2	1193.29380	3.90620
8	1954	1221.9	1236.40810	-14.50810
9	1955	1310.4	1282.04694	28.35306
.				
.	.	.	.	.
.				
40	1986	3981.2	3949.02241	32.17759
41	1987	4113.4	4075.44651	37.95349
42	1988	4279.5	4204.39514	75.10486
43	1989	4393.7	4335.86832	57.83168
44	1990	4474.5	4469.86603	4.63397
45	1991	4466.6	4606.38828	-139.78828
46	1992	4594.5	4745.43508	-150.93508
47	1993	4748.9	4887.00641	-138.10641
48	1994	4928.1	5031.10228	-103.00228
49	1995	5075.6	5177.72269	-102.12269

50		1996	5237.5	5326.86764	-89.36764
51		1997	5423.9	5478.53713	-54.63713
52		1998	5683.7	5632.73115	50.96885
53		1999	5968.4	5789.44972	178.95028
54		2000	6257.8	5948.69283	309.10717
Total	N	54	54	54	54

a. Limited to first 100 cases.

Forecasted trend for the next 5 years:

Year	Forecasted Trend
2001	6126.131
2002	6290.693
2003	6457.779
2004	6627.389
2005	6799.523



Case 2: Test for the presence of trend and estimate it if it's present for the following world development indicators for India: (time period)

1. Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD.

2. Population Total

3. Gross Domestic Product (GDP) (current USD)

4. Gross Domestic Product (GDP) Growth (annual %)

5. Life Expectancy at birth (years)

Obtain the de-trended indicators.

For GNI per capita:

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 3.7533

p value: 1e-04

No. of Discordants: 0

Expected No. of Discordants: 18

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

Quadratic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.996	.995	59.469

Cubic

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.996	.994	63.876

Linear

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.998	.996	.996	55.148

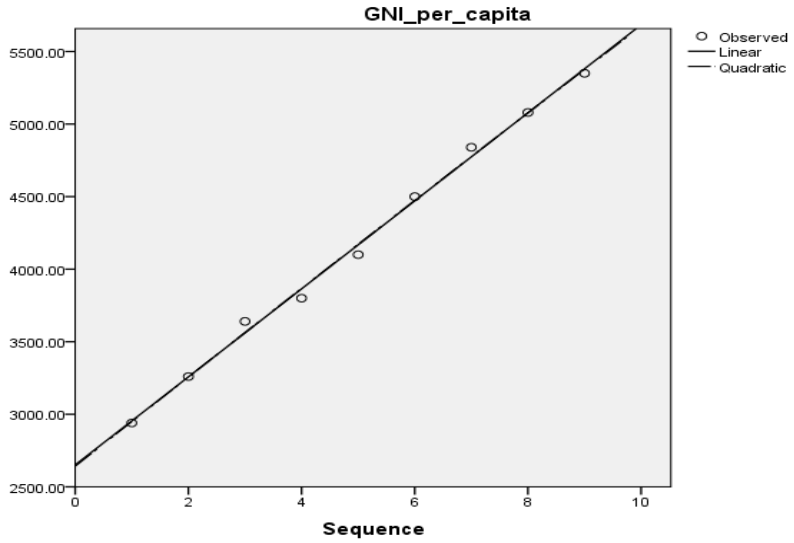
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	5520666.667	1	5520666.667	1815.251	.000
Residual	21288.889	7	3041.270		
Total	5541955.556	8			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	303.333	7.120	.998	42.606	.000
(Constant)	2651.111	40.064		66.172	.000

Inference: We observe that the value of Adjusted R square of Linear Regression model is greater than Non-Linear Regression model (.996 >.995). Thus, linear model fits the data better.



Inference: On fitting the quadratic, cubic and linear polynomial to the given data, we observe that the Linear model has the highest value of R square. Thus for Linear Regression Model, the coefficient of determination is 0.996 which implies that there is 99.6% variation in Gross National income (GNI). Estimation of trend can be done by using least squares method for linear model.

Now, we obtain the de trended time series by subtracting the trend value from original data:

Case Summaries[a]

	Series_name	GNI_per_capita	Estimated Trend Value	De trended_Value
1	2005	2940.00	2954.44444	-14.44444
2	2006	3260.00	3257.77778	2.22222
3	2007	3640.00	3561.11111	78.88889
4	2008	3800.00	3864.44444	-64.44444
5	2009	4100.00	4167.77778	-67.77778
6	2010	4500.00	4471.11111	28.88889
7	2011	4840.00	4774.44444	65.55556
8	2012	5080.00	5077.77778	2.22222
9	2013	5350.00	5381.11111	-31.11111
Total N	9	9	9	9

a. Limited to first 100 cases.

For Population Total:

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 3.7533

p value: 1e-04

No. of Discordants: 0

Expected No. of Discordants: 18

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

Quadratic:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
1.000	1.000	1.000	154183.457

Cubic:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
1.000	1.000	1.000	73225.437

Linear:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
1.000	1.000	1.000	257731.181

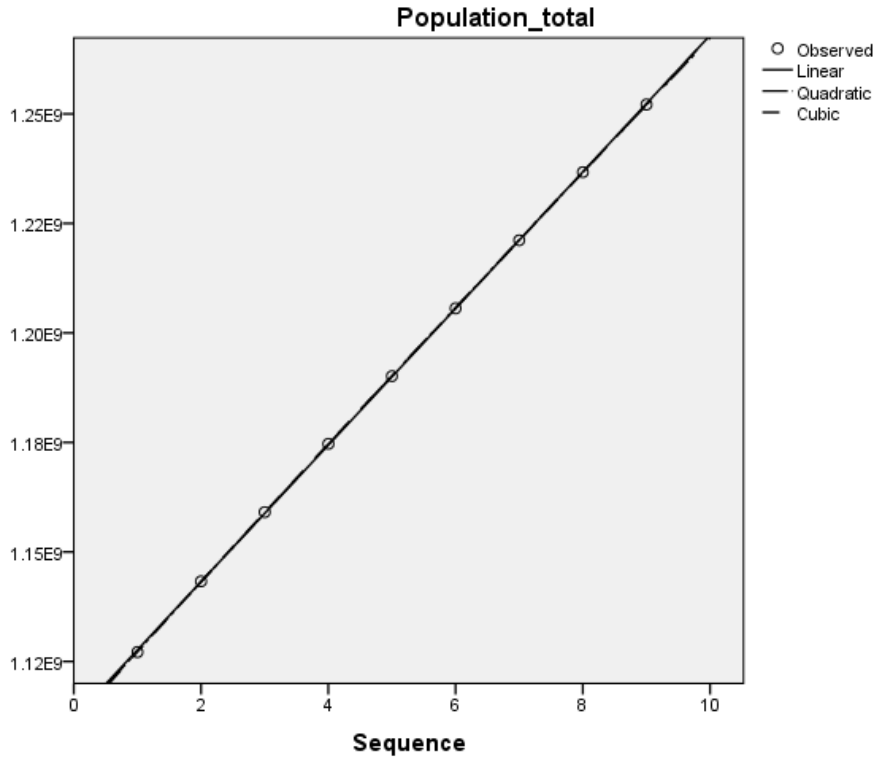
ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	1457854575519 0390.000	1	1457854575519 0390.000	219472.584	.000
Residual	464977531959. 649	7	66425361708.5 21		
Total	1457901073272 2350.000	8			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
Case Sequence	15587679.833	33272.952	1.000	468.479	.000
(Constant)	1112054472.61 1	187237.362		5939.277	.000

Inference: We observe that the value of Adjusted R square of Linear Regression and Non-Linear Regression model is equal. Thus, we choose a linear model to fit the data.



Inference: On fitting the quadratic, cubic and linear polynomial to the given data, we observe that all models have the same value of R square. Thus for Linear Regression Model, the coefficient of determination is 1.000 which implies that there is 100% variation in Population Total. Estimation of trend can be done by using least squares method for linear model.

Now, we obtain the de trended time series by subtracting the trend value from original data:

Case Summaries[a]

	Series_name	Population_total	Estimated_Trend_Value	De trended Value
1	2005	1127143548.00	1127642152.44444	-498604.44444
2	2006	1143289350.00	1143229832.27778	59517.72222
3	2007	1159095250.00	1158817512.11111	277737.88889
4	2008	1174662334.00	1174405191.94444	257142.05556
5	2009	1190138069.00	1189992871.77778	145197.22222
6	2010	1205624648.00	1205580551.61111	44096.38889
7	2011	1221156319.00	1221168231.44444	-11912.44444
8	2012	1236686732.00	1236755911.27778	-69179.27778
9	2013	1252139596.00	1252343591.11111	-203995.11111
Total N	9	9	9	9

a. Limited to first 100 cases.

For GDP (current USD):

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 3.3362

p value: 4e-04

No. of Discordants: 2

Expected No. of Discordants: 18

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

Linear:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.970	.940	.931	107139476666.248

Quadratic:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.975	.951	.935	104115274248.740

Cubic:

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
.983	.966	.946	94784362998.207

ANOVA

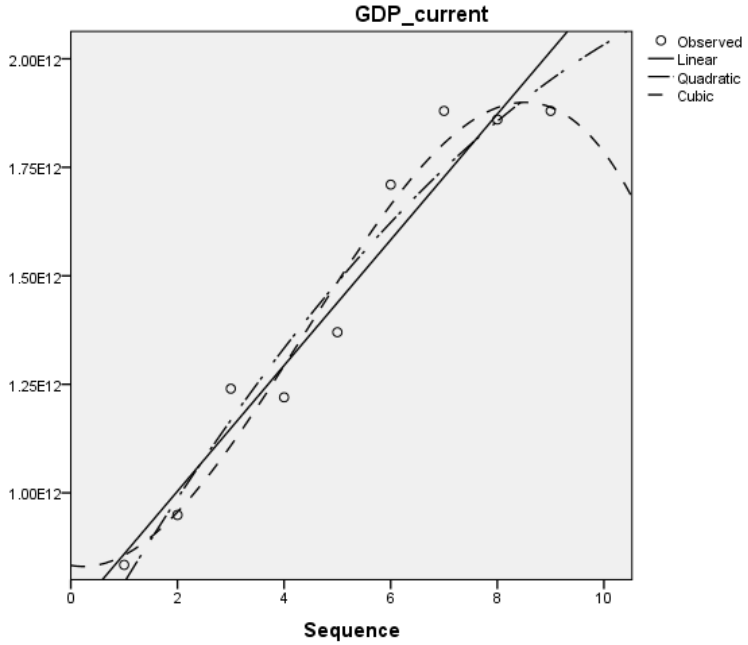
	Sum of Squares	df	Mean Square	F	Sig.
Regression	1293164511544009400000000.000	3	4310548371813365000000000.000	47.980	.000
Residual	4492037734487935000000000.000	5	898407546897587000000000.000		
Total	1338084888888888600000000000.000	8			

Coefficients

	Unstandardized Coefficients	Standardized Coefficients	t	Sig.

	B	Std. Error	Beta		
Case Sequence	-22133597883.579	168042216552.863	-.148	-.132	.900
Case Sequence ** 2	49300144300.140	38040902726.750	3.385	1.296	.252
Case Sequence ** 3	-3756734006.734	2510370671.229	-2.348	-1.496	.195
(Constant)	832873015872.996	204931665947.818		4.064	.010

Inference: We observe that the value of Adjusted R square of Non-Linear Regression model is greater than Linear Regression model (.946 >.935). Thus, cubic model fits the data better.



Inference: On fitting the quadratic, cubic and linear polynomial to the given data, we observe that the cubic (Non-linear) model has the highest value of R square. Thus for Non-linear Regression Model, the coefficient of determination is 0.966 which implies that there is 96.6% variation in GDP (current USD). Estimation of trend can be done by using least squares method for cubic model.

Now, we obtain the de trended time series by subtracting the trend value from original data:

Case Summaries[a]

	Series_name	GDP_current	Estimated_Trend_Value	De_trended_value
1	2005	834000000000.00	856282828282.82310	-22282828282.82312
2	2006	949000000000.00	955752525252.52770	-6752525252.52771
3	2007	1240000000000.00	1108741702741.70730	131258297258.29272

4	2008	1220000000000.00	1292709956709.95970	-72709956709.95972
5	2009	1370000000000.00	1485116883116.88300	-115116883116.88306
6	2010	1710000000000.00	1663422077922.07450	46577922077.92554
7	2011	1880000000000.00	1805085137085.13200	74914862914.86792
8	2012	1860000000000.00	1887565656565.65380	-27565656565.65381
9	2013	1880000000000.00	1888323232323.23700	-8323232323.23706
Total N		9	9	9	9

a. Limited to first 100 cases.

For GDP growth (annual %):

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

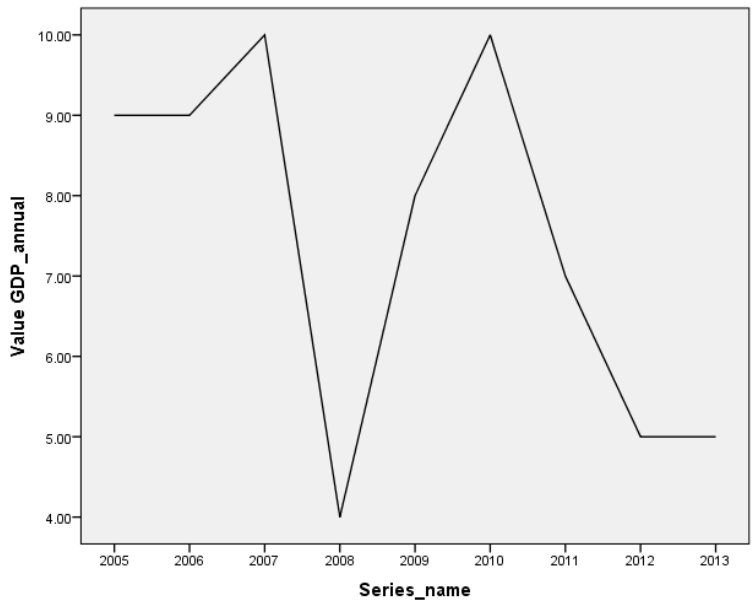
Test Statistic: -1.0426

p value: 0.1486

No. of Discordants: 23

Expected No. of Discordants: 18

Inference: Since, p-value (0.148) > 0.05, therefore null hypothesis is accepted and we conclude that there is absence of trend.



For Life Expectancy at birth (years):

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 3.3362

p value: 4e-04

No. of Discordants: 2

Expected No. of Discordants: 18

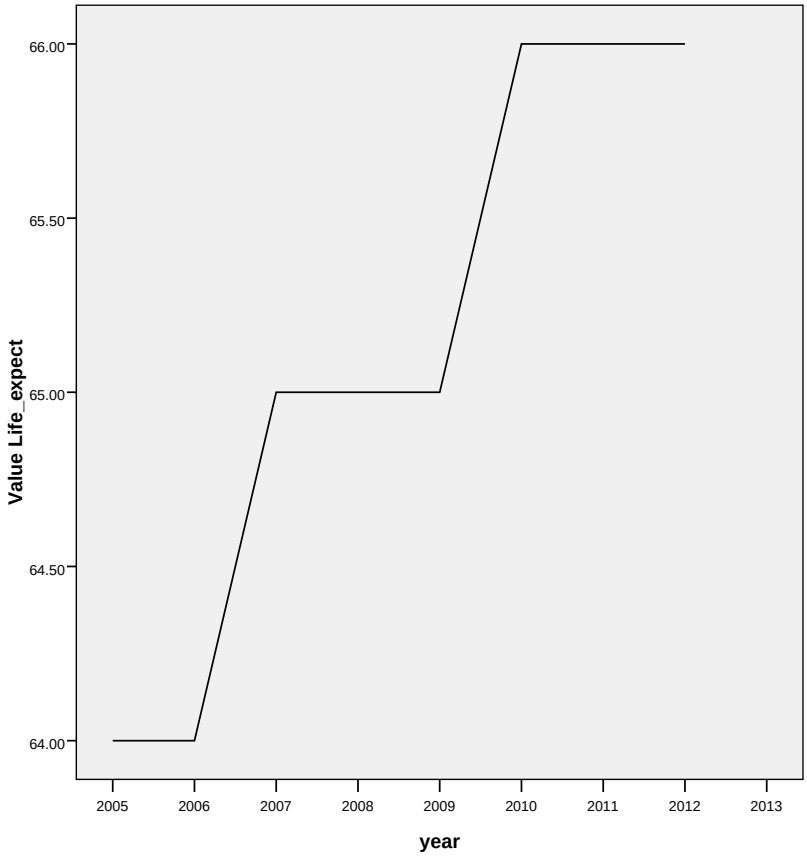
Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

Q=	0
E(Q)=	18
T=	1
V(T)=	0.070988

Test Statistic:

Z=	3.753259

Inference: As Z statistic is greater than 1.96, we reject the null hypothesis. Hence it indicates presence of trend. As Q<<E(Q), thus it is indicative of rising trend.



Inference: We compare the R-square of linear regression model and non linear regression model and we find that R-square of non linear regression is no better than R-square of former. Thus we choose a linear regression model to fit the data.

Linear:

Model Summary

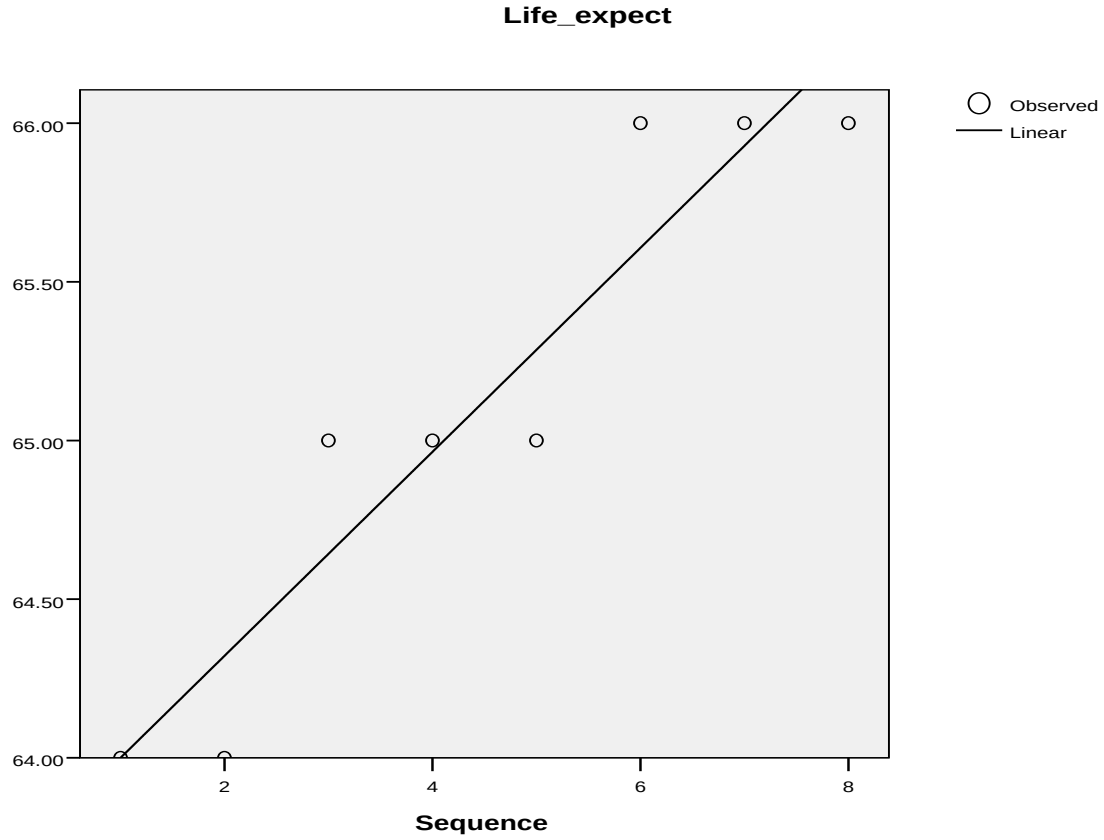
R	R Square	Adjusted R Square	Std. Error of the Estimate
.943	.890	.872	.299

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	4.339	1	4.339	48.600	.000
Residual	.536	6	.089		
Total	4.875	7			

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
Case Sequence	.321	.046	.943	6.971	.000
(Constant)	63.679	.233		273.500	.000



Inference: On fitting higher order polynomial functions(starting from the cubic polynomial) to the given data, it was observed that all the coefficients of the model were significant, but the R^2 value was not increasing further. Thus we consider linear regression model. The value of coefficient of determination as 0.890 this means that this model explains 89% of the variation in life expectancy at birth.

Case 3: Test for the presence of trend and estimate it if it's present for the annual sales measured in million USD for a trading company for 1994-2013. Obtain the de-trended sales. Also provide a simple trend based forecast for the annual sales for the next 3 years.

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 4.4124

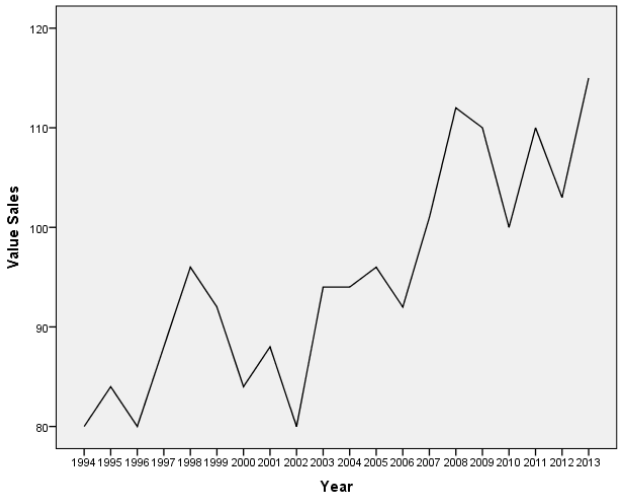
p value: 0

No. of Discordants: 27

Expected No. of Discordants: 95

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

For Sales:



Inference: We observe from the above graph that there are peaks and drops in the data. Therefore Moving Average method is used to estimate the deterministic components.

Year	1995-1998	1998-2005	2005-2007	2007-2008	2008-2013
No. of peaks	2	7	2	1	5

Span of the moving average = $\frac{2+7+2+1+5}{5} = 3$

Now, we obtain the de trended time series by subtracting the trend value from original data:

Case Summaries[a]

	Year	Sales	MA(Sales,3,3)	Detrended_value
1	1994	80	81.30	-1.30

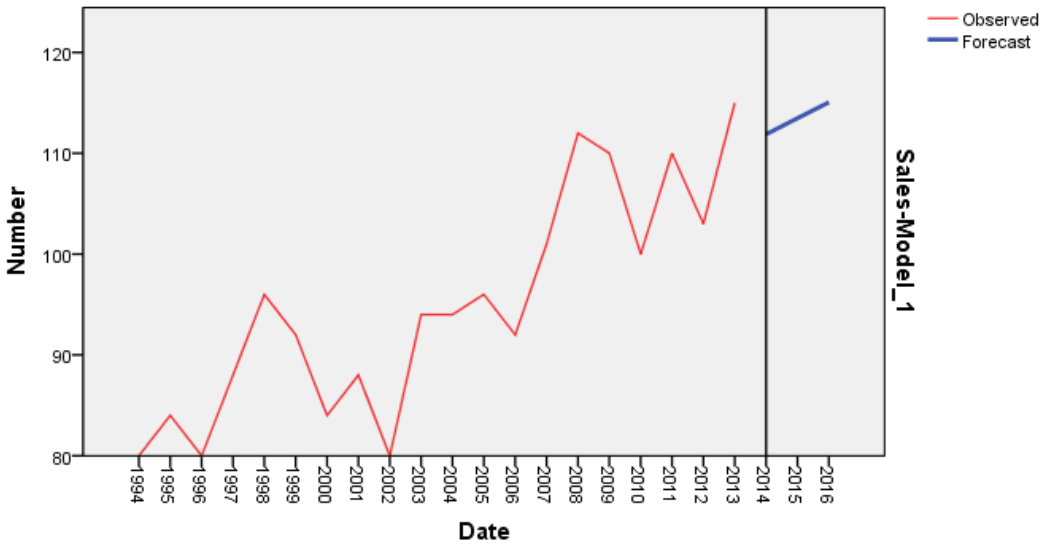
2	1995	84	81.33	2.67
3	1996	80	84.00	-4.00
4	1997	88	88.00	.00
5	1998	96	92.00	4.00
6	1999	92	90.67	1.33
7	2000	84	88.00	-4.00
8	2001	88	84.00	4.00
9	2002	80	87.33	-7.33
10	2003	94	89.33	4.67
11	2004	94	94.67	-.67
12	2005	96	94.00	2.00
13	2006	92	96.33	-4.33
14	2007	101	101.67	-.67
15	2008	112	107.67	4.33
16	2009	110	107.33	2.67
17	2010	100	106.67	-6.67
18	2011	110	104.33	5.67
19	2012	103	109.33	-6.33
20	2013	115	109.33	5.67
Total N	20	20	20	20

a. Limited to first 100 cases.

Inference: We observe from the graph of time series that most of the fluctuations are smoothened out using the Moving Average Method.

Forecasted Trend for the next 3 years:

Year	Forecasted Trend
2014	111.88
2015	113.47
2016	115.06



Additive Decomposition

Case 4: Test for the presence of trend and seasonality, and estimate them if they are present for the monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India using appropriate tests and methods. Obtain the additive decomposition of the original series, viz. estimated trend, estimated seasonality, estimated cyclicity and estimated random component. Give deterministic components based forecast for the monthly Whole Sale Price Index for the next 5 months.

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: -1.0785

p value: 0.1404

No. of Discordants: 3561

Expected No. of Discordants: 3335

Inference: Since, p-value (0.140) > 0.05, therefore null hypothesis is accepted and we conclude that there is absence of trend and hence estimation and elimination of trend is not possible.

Freidman (JASA) Test for Presence of Seasonality

Null Hypothesis: Absence of Seasonality, and

Alternative Hypothesis: Presence of Seasonality.

Test Statistic: 10.191 (Chi Sqaure with 11 df)

p_value: 0.5133

Inference: Since, p-value (0.5133) > 0.05, therefore null hypothesis is accepted and we conclude that there is absence of seasonality and hence estimation and elimination of seasonal component is not possible.

Case 5: Test for the presence of trend and seasonality, and estimate them if they are present for the monthly World Airline Passengers from 1949-1960 using appropriate tests and methods. Obtain the additive decomposition of the original series, viz. estimated trend, estimated seasonality, estimated cyclicity and estimated random component. Give deterministic components based forecast for the monthly World Airline Passengers for the next 5 months.

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 14.4294

p value: 0

No. of Discordants: 971

Expected No. of Discordants: 5148

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the trend is present.

Freidman (JASA) Test for Presence of Seasonality

Null Hypothesis: Absence of Seasonality, and

Alternative Hypothesis: Presence of Seasonality.

Test Statistic: 241.5256 (Chi Sqaure with 11 df)

p_value: 0

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that the seasonality is present.

Additive Decomposition of the series:

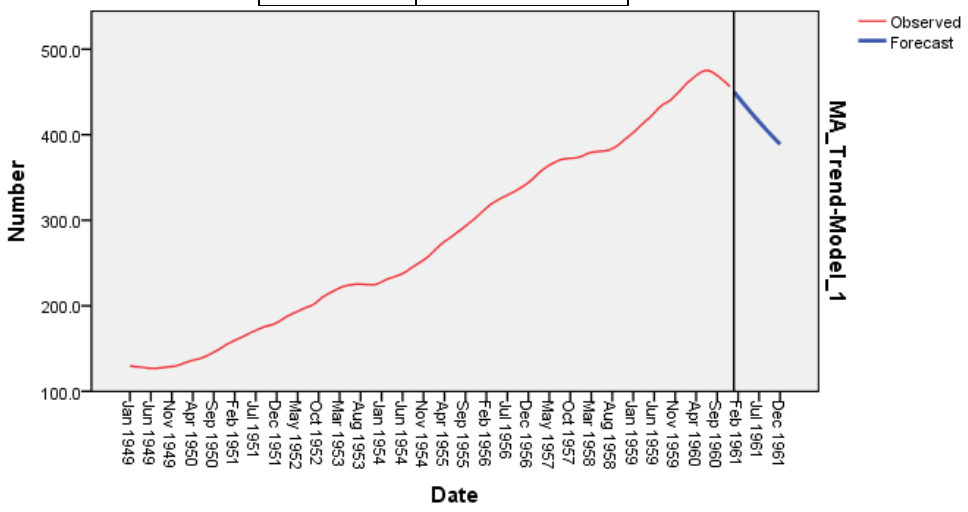
Case Summaries[a]

	YEAR_	MONTH_	Airline_Passengers	Trend	Seasonal	Cyclic	Residual
1	1949	1	112	129.8	-24.84380	13.64	-6.59596
2	1949	2	118	129.0	-36.34380	12.97	12.37374
3	1949	3	132	128.6	-2.72259	10.43	-4.30808
4	1949	4	129	128.0	-8.63168	4.18	5.45286
5	1949	5	121	127.3	-5.31350	-8.67	7.68820
6	1949	6	135	126.8	34.70165	-23.09	-3.41232
7	1949	7	148	126.8	69.89547	-31.95	-16.73695
8	1949	8	148	127.3	61.96680	-27.76	-13.46114
9	1949	9	136	128.0	15.71680	-11.32	3.64696
.							
.	.	.	.	.	.	.	.
.							
130	1959	10	407	441.0	-21.27562	-2.88	-9.80556
131	1959	11	362	445.8	-54.15441	-16.71	-12.97138
132	1959	12	405	450.6	-28.99532	-20.22	3.59512
133	1960	1	417	456.3	-24.84380	-24.55	10.05640
134	1960	2	391	461.4	-36.34380	-26.62	-7.41077
135	1960	3	419	465.2	-2.72259	-23.18	-20.30808
136	1960	4	461	469.3	-8.63168	-9.93	10.23064
137	1960	5	472	472.8	-5.31350	10.10	-5.53402
138	1960	6	535	475.0	34.70165	33.11	-7.85676
139	1960	7	622	475.0	69.89547	48.84	28.26305
140	1960	8	606	472.8	61.96680	49.81	21.42775
141	1960	9	508	469.3	15.71680	33.56	-10.57527
142	1960	10	461	465.2	-21.27562	15.33	1.75000
143	1960	11	390	461.4	-54.15441	1.08	-18.32071
144	1960	12	432	456.3	-28.99532	-2.85	7.54545
Total N	144	144	144	144	144	144	144

a. Limited to first 150 cases.

Deterministic Components for the next 5 months:

Months	Forecasted Trend
Jan	450.1
Feb	444.1
Mar	438.1
Apr	432.2
May	426.4



Case 6: Test for the presence of trend and seasonality, and estimate them if they are present for the quarterly demand for an industrial good measured on thousand units for a manufacturing company for 2001-2005 using appropriate tests and methods. Obtain the additive decomposition of the original series, viz. estimated trend, estimated seasonality, estimated cyclicity and estimated random component. Give deterministic components based forecast for the quarterly demand for the industrial good for the next 2 quarters.

Relative Ordering Test for Presence of Trend:

Null Hypothesis: Absence of Trend, and
Alternative Hypothesis: Presence of Trend.

Test Statistic: 1.4275

p value: 0.0767

No. of Discordants: 73

Expected No. of Discordants: 95

Inference: Since, p-value (0.076) > 0.05, therefore null hypothesis is accepted and we conclude that there is absence of trend.

Freidman (JASA) Test for Presence of Seasonality

Null Hypothesis: Absence of Seasonality, and

Alternative Hypothesis: Presence of Seasonality.

Test Statistic: 33.6769 (Chi Square with 11 df)

p_value: 4e-04

Inference: Since, p-value (0.000) < 0.05, therefore null hypothesis is rejected and we conclude that there is presence of seasonality.

Additive Decomposition:

Case Summaries[a]

	DATE_	Demand	Trend	Seasonal	Cyclic	Residual
1	Q1 2001	40	39.0	3.53438	-1.32	-1.21250
2	Q2 2001	35	38.5	-2.96563	-.51	-.02500
3	Q3 2001	38	38.5	-1.54063	.12	.92500
4	Q4 2001	40	39.0	.97188	.00	.02500
5	Q1 2002	42	39.4	3.53438	-.10	-.80833
6	Q2 2002	37	39.3	-2.96563	.08	.63611
7	Q3 2002	39	38.9	-1.54063	.19	1.48056
8	Q4 2002	38	38.5	.97188	-.16	-1.30833
9	Q1 2003	41	38.1	3.53438	-.07	-.58611
10	Q2 2003	35	38.5	-2.96563	-.06	-.47500
11	Q3 2003	38	39.5	-1.54063	.00	.03611
12	Q4 2003	42	40.1	.97188	.10	.80278
13	Q1 2004	45	40.0	3.53438	.16	1.30278
14	Q2 2004	36	39.6	-2.96563	-.07	-.58611
15	Q3 2004	36	39.4	-1.54063	-.20	-1.63056
16	Q4 2004	41	39.5	.97188	.06	.46944
17	Q1 2005	44	40.0	3.53438	.05	.41389
18	Q2 2005	38	40.4	-2.96563	.07	.52500
19	Q3 2005	38	40.4	-1.54063	.11	-.97083
20	Q4 2005	42	40.0	.97188	.55	.48125
Total N		20	20	20	20	20	20

Deterministic Components for the next 2 quarters:

Quarters	Forecasted Trend
Q1 2006	39.6
Q2 2006	39.5

Modelling the Random Component using Auto Regressive Integrated Moving Average (ARIMA) Model

Testing for Stationarity and making the Series Stationary if it is not:

After estimating and eliminating the trend component, the given time series is tested for stationarity using **Augmented Dickey**

Fuller test. Since MA process is always stationary, hence it is the AR process we need to check for stationarity.
Given an observed time series y_1 y_2...y_n, Dickey and Fuller considered the following differential form autoregressive equation to test the presence of unit root:

$$\Delta y_t = \beta y_{t-1} + \sum_{j=1}^{p} \delta_j \Delta y_{t-j} + \epsilon_t$$

Where, t is the time index, β is the coefficient presenting process root i.e. the focus of testing, p is the lag order of the first differences autoregressive process. $\epsilon_t \sim WN(0,\sigma^2)$

The null and alternative hypothesis corresponding to above model is:

H_0: β = 0 i.e. the AR (p) process is not stationary.

H_1: β < 0 i.e. the process is stationary.

ADF testing technique involves Ordinary Least Squares (OLS) method to find the coefficients of model equation. To estimate the significance of the coefficients in focus, Dickey-Fuller statistic is computed and compared with the relevant critical value.

ADF test is applied to the observed time series values using R.

If the test leads to acceptance of null hypothesis i.e. if the test indicates that the given time series is not stationary, then the series can be made stationary using methods such as differencing.

Case 7: Consider the de trended series for consumption expenditure (in million dollars) for the United States for 1947 to 2000 from case 1 and test for its stationarity using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t7[, 4]
Dickey-Fuller = -2.2273, Lag order = 3, p-value = 0.483

alternative hypothesis: stationary

Inference: Since, p-value (0.483) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Now, make the series stationary by order differencing:

Case Summaries[a]

	Year	Consumption	Detrended	Ist_order_differencing	IInd_order_differencing
1	1947	976.4	-11.22333		
2	1948	998.1	-17.49040	-6.26707	
3	1949	1025.3	-20.78200	-3.2916	2.97547
4	1950	1090.9	11.80186	32.58386	35.87546
5	1951	1107.1	-7.53882	-19.34068	-51.92454
6	1952	1142.4	-10.30404	-2.76522	16.57546
7	1953	1197.2	3.90620	14.21024	16.97546
8	1954	1221.9	-14.50810	-18.4143	-32.62454
9	1955	1310.4	28.35306	42.86116	61.27546
.					
.	.	..	..	..	.
.					
45	1991	4466.6	-139.78828	-144.42225	-91.22454
46	1992	4594.5	-150.93508	-11.1468	133.27545
47	1993	4748.9	-138.10641	12.82867	23.97547
48	1994	4928.1	-103.00228	35.10413	22.27546
49	1995	5075.6	-102.12269	0.87959	-34.22454
50	1996	5237.5	-89.36764	12.75505	11.87546
51	1997	5423.9	-54.63713	34.73051	21.97546
52	1998	5683.7	50.96885	105.60598	70.87547
53	1999	5968.4	178.95028	127.98143	22.37545
54	2000	6257.8	309.10717	130.15689	2.17546
Total N	54	54	54	54	54

a. Limited to first 150 cases.

Augmented Dickey-Fuller Test: (Ist Order Differencing)

data: t7[2:54, 5]

Dickey-Fuller = -2.7359, Lag order = 3, p-value = 0.2782

alternative hypothesis: stationary

Inference: Since, p-value (0.278) > 0.05, therefore null hypothesis is accepted and we conclude that again the series is not stationary.

Augmented Dickey-Fuller Test: (IInd Order Differencing)

data: t7[3:54, 6]

Dickey-Fuller = -5.1833, Lag order = 3, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we now conclude that the series is stationary.

Case 8: Consider the following de-trended series from the case 2:

1. **Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD.**
2. **Population Total.**
3. **Gross Domestic Product (GDP) (current USD).**
4. **Gross Domestic Product (GDP) Growth (annual %).**
5. **Life Expectancy at birth (years).**

Test for stationarity of the all de-trended series using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

1) Augmented Dickey-Fuller Test: (Gross National Income)

data: t8[, 7]
Dickey-Fuller = -26.7218, Lag order = 2, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

2) Augmented Dickey-Fuller Test: (Population Total)

data: t8[, 8]
Dickey-Fuller = -0.0253, Lag order = 2, p-value = 0.99

alternative hypothesis: stationary

Inference: Since, p-value (0.99) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Now, make the series stationary by order differencing:

Case Summaries[a]

	Series_name	Population_total	de_trended	Differencing_1
1	2005	1127143548.00	-498604.44	
2	2006	1143289350.00	59517.72	558122.1667
3	2007	1159095250.00	277737.89	218220.1667
4	2008	1174662334.00	257142.06	-20595.8333
5	2009	1190138069.00	145197.22	-111944.8334
6	2010	1205624648.00	44096.39	-101100.8333
7	2011	1221156319.00	-11912.44	-56008.83334
8	2012	1236686732.00	-69179.28	-57266.83333
9	2013	1252139596.00	-203995.11	-134815.8333

10		.	.	.	
Total	N	9	9	9	10

a. Limited to first 100 cases.

Augmented Dickey-Fuller Test: (Differencing_1)

data: t8[2:9, 9]
Dickey-Fuller = -16.0404, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that now the series is stationary.

3) Augmented Dickey-Fuller Test: (GDP current USD)

data: t8[, 10]
Dickey-Fuller = -3.2855, Lag order = 2, p-value = 0.09368

alternative hypothesis: stationary

Inference: Since, p-value (0.09) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Now, make the series stationary by order differencing:

Case Summaries[a]

	Series_name	GDP_current	De_trended	Differencing_1	Differencing_2
1	2005	834000000000.00	-24977777778.00		
2	2006	949000000000.00	-54761111111.00	-297833	
3	2007	1240000000000.00	91455555556.00	1.46217	1.76E+1
4	2008	1220000000000.00	-73327777778.00	-1.6478	-3.11E+
5	2009	1370000000000.00	-68111111111.00	5216666	1.7E+11
6	2010	1710000000000.00	127000000000.00	1.95217	1.9E+11
7	2011	1880000000000.00	152000000000.00	2521600	-1.7000
8	2012	1860000000000.00	-12461111111.00	-1.6478	-1.8999
9	2013	1880000000000.00	-137000000000.00	-1.2478	4000022

Total	N	9	9	9	9	9

a. Limited to first 100 cases.

Augmented Dickey-Fuller Test: (Differencing_1)

data: t8[2:9, 13]
Dickey-Fuller = -2.0611, Lag order = 1, p-value = 0.5491

alternative hypothesis: stationary

Inference: Since, p-value (0.54) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Augmented Dickey-Fuller Test: (Differencing_2)

data: t8[3:9, 14]
Dickey-Fuller = -9.0811, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that now the series is stationary.

4) Augmented Dickey-Fuller Test: (GDP annual growth)

data: t8[, 13]
Dickey-Fuller = -2.7059, Lag order = 2, p-value = 0.3035

alternative hypothesis: stationary

Inference: Since, p-value (0.30) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Now, make the series stationary by order differencing:

Case Summaries[a]

	Series_name	GDP_annual	De_trended	Differencing_1	Differencing_2
1	2005	9.00	-.31027	.	
2	2006	9.00	.15564	.47	
3	2007	10.00	1.62179	1.47	1.00024
4	2008	4.00	-3.91183	-5.53	-6.99977
5	2009	8.00	.55478	4.47	10.00023
6	2010	10.00	3.02162	2.47	-1.99977
7	2011	7.00	.48870	-2.53	-4.99976
8	2012	5.00	-1.04399	-1.53	1.00023
9	2013	5.00	-.57645	.47	2.00023
10	.	.	.	.	
Total N	9	9	9	8	10

a. Limited to first 100 cases.

Augmented Dickey-Fuller Test: (Differencing_1)

data: t8[2:9, 14]
Dickey-Fuller = -2.4143, Lag order = 1, p-value = 0.4145

alternative hypothesis: stationary

Inference: Since, p-value (0.41) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Augmented Dickey-Fuller Test: (Differencing_2)

data: t8[3:9, 15]
Dickey-Fuller = -3.7967, Lag order = 1, p-value = 0.03595

alternative hypothesis: stationary

Inference: Since, p-value (0.03) < 0.05, therefore null hypothesis is rejected and we conclude that now the series is stationary.

5) Augmented Dickey-Fuller Test: (Life Expectancy)

data: t8[, 16]
Dickey-Fuller = -215139.5, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Case 9: Consider the de trended series for annual sales measured in million USD for a trading company from 1994-2013 from case 3 and test for its stationarity using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t9[, 3]
Dickey-Fuller = -4.3967, Lag order = 2, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Case 10: Consider the estimated random component for monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India from case 4 and test for its stationarity using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t10[, 2]

Dickey-Fuller = -4.3378, Lag order = 4, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Case 11: Consider the estimated random component for monthly World Airline Passengers from 1949-1960 from case 5 and test for its stationarity using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t11[, 4]

Dickey-Fuller = -6.975, Lag order = 5, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Case 12: Consider the estimated random component for monthly the quarterly demand for an industrial good measured in thousand units for a manufacturing company for 2001-2005 from case 6 and test for its stationarity using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t12[, 4]

Dickey-Fuller = -3.8083, Lag order = 2, p-value = 0.03512

alternative hypothesis: stationary

Inference: Since, p-value (0.03) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Case 13: Generate 1000 data points from the following (1) process:

$$\mathbf{X_t = X_{t-1} + \epsilon_t}$$

where ϵ_t is a Gaussian WN$(0, 1)$ and $X_0 = 10$.

Test for stationarity of the data using the Augmented Dickey Fuller (ADF) Test. Make the specific series stationary if it is not.

Augmented Dickey-Fuller Test:

data: t13[, 3]

Dickey-Fuller = -2.0423, Lag order = 9, p-value = 0.5604

alternative hypothesis: stationary

Inference: Since, p-value (0.5) > 0.05, therefore null hypothesis is accepted and we conclude that the given series is not stationary.

Now, make the series stationary by order differencing:

Augmented Dickey-Fuller Test:

data: t13[2:1001, 4]
Dickey-Fuller = -10.9079, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

Identification of the Order of the ARIMA Model:

In Box-Jenkins approach of time series modelling, the stationary time series is modelled as a linear process using some ARMA model. After testing for stationarity and then removing non-stationarity if it was there, the next step is to identify the order of the ARMA model to be fitted.

Order Identification Using ACF:

Plot of auto correlation function (ACF) against non-negative lags is called as correlogram while its sample counterpart is termed as sample correlogram which is useful to identify the order of an MA process.

Order Identification Using PACF:

Plot of partial auto correlation function(PACF) against non-negative lags is called as partial correlogram while its sample counterpart is termed as sample partial correlogram which is useful to identify the order of an AR process.

Sample correlogram and sample partial correlogram with appropriate acceptance bands for insignificance of autocorrelation and partial autocorrelation resp. can be used to visualize the order of the model.

The following table gives a summary of the behaviour of ACF & PACF plot for ARMA process:

MODEL	ACF	PACF
White Noise	Cuts off with one significant spike at lag 0.	Cuts off with no significant spike.

MA (q)	Cuts off with significant spikes at the most up to q lags.	Tails off indefinitely.
AR (p)	Tails off indefinitely.	Cuts off with significant spikes at the most up to p lags.
ARMA (p,q)	Tails off indefinitely.	Tails off indefinitely.

In case the process is a proper ARMA, then neither ACF nor PACF gives information about the model order. The next approach of model order identification is an information theoretic method.

Order Identification using BIC:

The Bayesian Information Criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based on the likelihood function. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over fitting. BIC resolves this problem by introducing a penalty term for the number of parameters in the model. The BIC is formally defined as:

$$\mathbf{BIC} = -\mathbf{2}.\mathbf{ln}\hat{\mathbf{L}} + \mathbf{k}.\mathbf{ln}(\mathbf{n})$$

Case 14: Generate 1000 data points from the following $\mathbf{MA}(1)$ process.

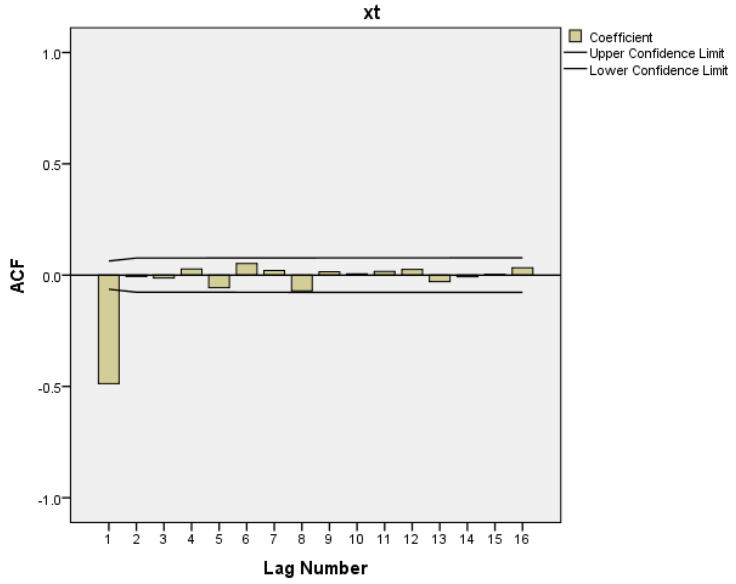
$$\mathbf{X_t} = \boldsymbol{\epsilon_t} - \boldsymbol{\epsilon_{t-1}}$$

where $\boldsymbol{\epsilon_t}$ is a Gaussian $\mathbf{WN}(\mathbf{0}, \mathbf{1})$.

Assuming the generated data as a sample from some ARMA model, identify its order using sample Correlogram and sample Partial Correlogram.

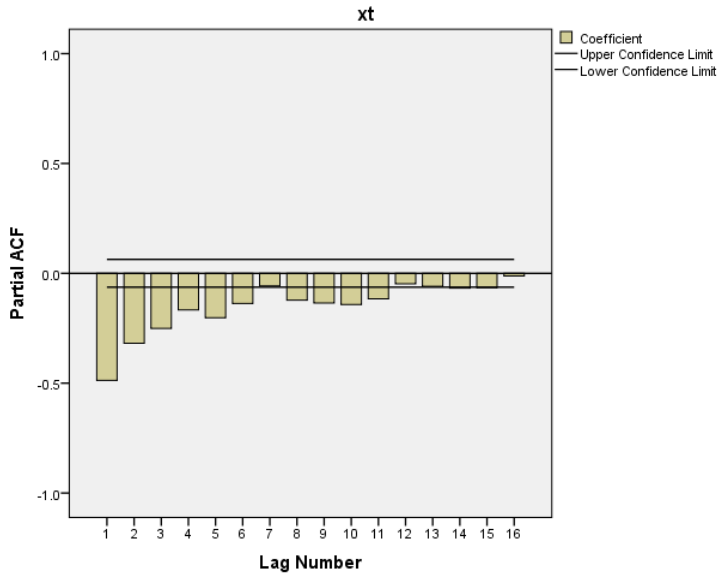
Since the given process is MA, therefore there is no need to check if it is stationary or not as an MA process is always stationary. Order of the MA process is identified using ACF and PACF plot.

ACF Plot:



Inference: From the ACF Plot, we see that the sample correlogram exhibits significant spike at lag 1 and there is no significant spike after lag 1, thus it is indicative of an MA(1) process i.e. order is 1.

PACF Plot:



Inference: From the PACF plot, we see that the PACF tails off indefinitely (or does not cut off after any finite lags).

Thus ACF & PACF plot for the given series indicate that the data points have been generated from an MA (1) process.

Case 15: Generate 1000 data points from the following AR(1) process.

$$\mathbf{X_t = 0.9X_{t-1} + \epsilon_t}$$

where ϵ_t is a Gaussian WN$(0, 1)$ and $X_0 = 10$.

Assuming the generated data as a sample from some ARMA model, identify its order using sample Correlogram and sample Partial Correlogram.

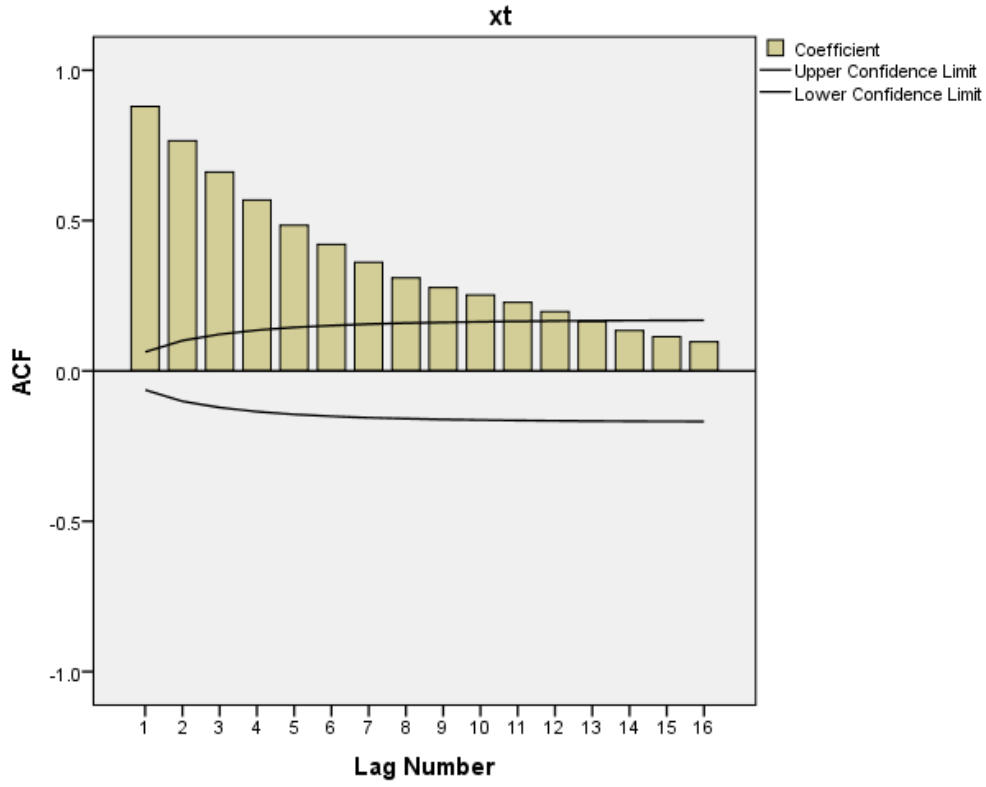
Augmented Dickey-Fuller Test:

data: t13[, 5]
Dickey-Fuller = -6.2695, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

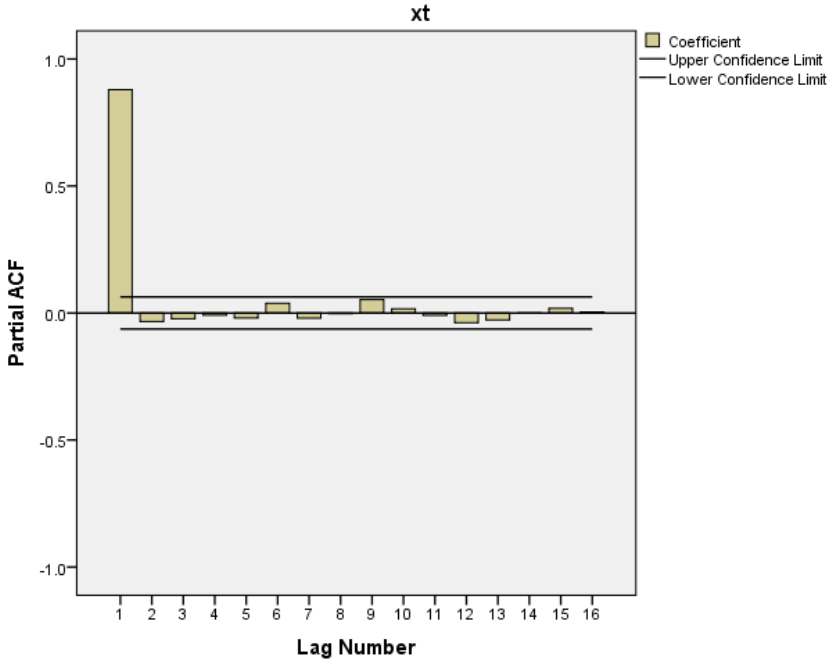
Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given series is stationary.

ACF Plot:



Inference: From the ACF Plot, we see that the ACF tails off indefinitely (or does not cut off after any finite lags).

PACF Plot:



Inference: From the PACF plot, we can see that the sample partial autocorrelation at lag 1 falls outside the acceptance band of insignificance but the sample partial autocorrelation after lag 1 falls inside the acceptance bands of insignificance.
Thus ACF & PACF plot for the given series indicate that the data points have been generated from an AR (1) process.

Building ARIMA Model and Forecasting:

R-squared is an estimate of proportion of total variation in the series which is explained by the model and measure is useful when the series is stationary. Stationary R-squared is a measure that compares stationary part of the model to a simple mean model and is preferable to ordinary R-squared when there is a trend or seasonal pattern. Stationary R-squared can be negative with a range of negative infinity to 1. Negative values mean that the model under consideration is worse than the baseline model. Positive values

mean that the model under consideration is better than the baseline model

Mean absolute percentage error (MAPE):

MAPE is a measure of accuracy of a method for constructing fitted time series values in statistics, specifically in trend estimation. It usually expresses accuracy as a percentage, and is defined by the formula:

$$\mathrm{M} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|,$$

where A_t is the actual value and F_t is the forecast value.

The difference between A_t and F_t is divided by the Actual value A_t again. The absolute value in this calculation is summed for every fitted or forecasted point in time and divided again by the number of fitted points n. Multiplying by 100 makes it a percentage error.

Case 16: Generate 1000 data points from the following AR(1) process.

$$\mathbf{X_t = 5 + 0.50X_{t-1} + \epsilon_t}$$

Where ϵ_t is a Gaussian WN$(0,1)$ and $X_0 = 10$. Assuming the generated data as a sample from some ARMA model do the following:

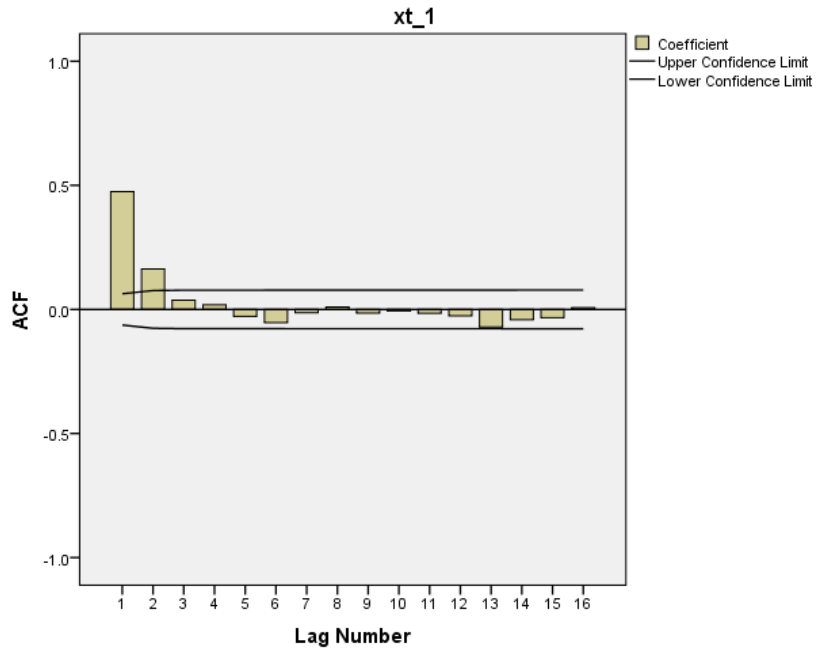
1. **Test for stationarity of the data using the Augmented Dickey Fuller (ADF) Test. Make the series stationary if it is not.**
2. **Fit an 'appropriate' order (identify it using sample Correlogram and sample Partial Correlogram) ARMA model.**
3. **Check the goodness of the model by using the following:**
    a. **Stationary R-Square**
    b. **Root Mean Square Error (RMSE)**
    c. **Mean Absolute Percentage Error (MAPE)**
4. **Validate the assumption of driving Gaussian White Noise using the following:**
    a. **Ljung–Box Test for White Noise**
    b. **ACF and PACF for White Noise**
    c. **Q-Q Plot for Normality**
5. **Assess the goodness of model built on simulated data by checking if the estimates are close to the parameters?**
6. **Apply the model and forecast for next 20 time points.**

1) Augmented Dickey-Fuller Test:

data: t16[, 3]
Dickey-Fuller = -9.4468, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

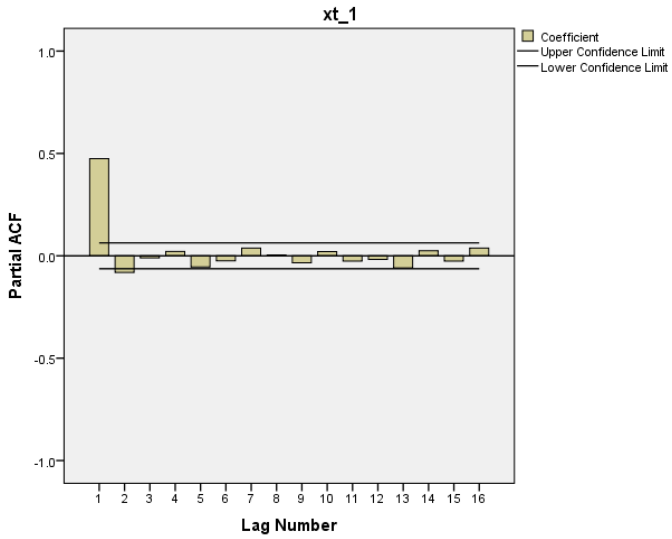
Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given AR(1) series is stationary.

2) ACF Plot:



Inference: From the ACF plot, we can see that the sample partial autocorrelation at lag 1 and lag 2 falls outside the acceptance band of insignificance but the sample partial autocorrelation after lag 2 falls inside the acceptance bands of insignificance.

PACF Plot:



Inference: From the PACF plot, it can be observed that the sample partial autocorrelation at lag 1 falls outside the acceptance band of insignificance but the sample partial autocorrelations after lag 1 fall inside the acceptance bands of insignificance.

Thus ACF & PACF plot for the given series do not clearly indicate the order of the process. Therefore we try identifying the order of the process using penalized log likelihood approach.

Model Type	BIC
ARIMA(0,0,1)	-0.007
ARIMA(0,0,2)	-0.046
ARIMA(0,0,3)	-0.045
ARIMA(1,0,0)	-0.062

ARIMA(1,0,1)	-0.055
ARIMA(1,0,2)	-0.047
ARIMA(1,0,3)	-0.040
ARIMA(2,0,0)	-0.055
ARIMA(2,0,1)	-0.049
ARIMA(2,0,2)	-0.041
ARIMA(2,0,3)	-0.033
ARIMA(3,0,0)	-0.048
ARIMA(3,0,1)	-0.041
ARIMA(3,0,2)	-0.033
ARIMA(3,0,3)	-0.026

From the above table, it can be observed that BIC corresponding to ARIMA (1, 0, 0) i.e. AR (1) is minimum. Thus it can be inferred that for the given series, the data points have been generated from an AR (1) process.

3) Checking the goodness of the model:

Model Statistics

Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	RMSE	MAPE	Normalized BIC	Statistics	DF	Sig.	
xt_1-Model_1	0	.206	1.017	8.230	.048	48.695	17	.000	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
xt_1-Model_1	xt_1	No Transformation	Constant		10.001	.046	216.669	.000
			MA	Lag 1	-.438	.028	-15.377	.000

Stationary R-Squared

Since stationary r squared is 0.206 > 0, therefore the model under consideration is better than the baseline model.

Root Mean Squared Error (RMSE)

RMSE value is low indicating a good fit. Also the dependent series is close with its model-predicted level.

Mean Absolute Percentage Error (MAPE)

MAPE is used when assessing forecast accuracy. For the given case MAPE is 8.230 → % of error involved in forecasting was 8%. Hence the forecasted values are close to the actual values.

4) Validation of assumptions of driving Gaussian white noise:

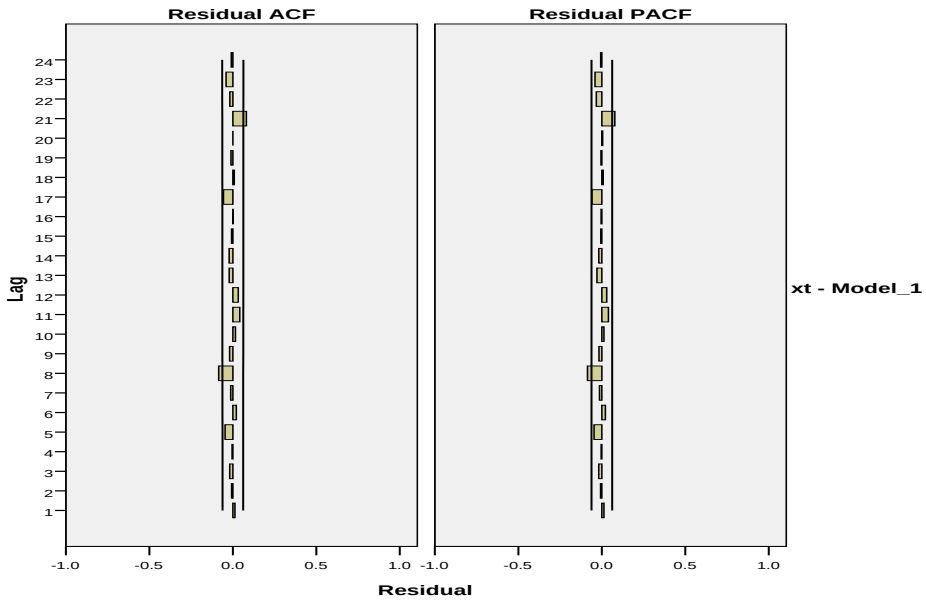
- **Ljung–Box Test for White Noise:**

  Now we test the randomness of the residuals using the Box-Ljung test.

H_0: The residuals are random.
H_1: The residuals are not random.
Test statistic: Q = 18.432

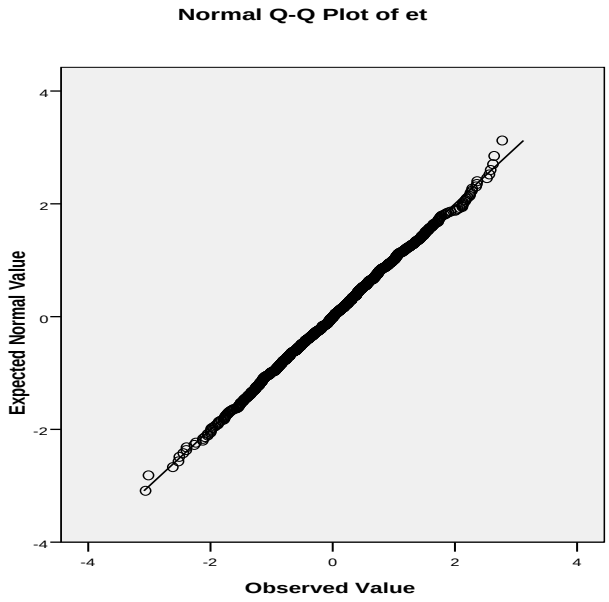
Inference: Since 0.05 <0.362, therefore we may accept H_0 at 5% level of significance. Hence we may conclude that the residuals for the fitted model are random i.e. are independently distributed.

- **ACF and PACF for White Noise:**



Inference: Since both the ACF and PACF for residuals tail off indefinitely, it can be inferred that the residuals ϵ_t have been generated from white noise process.

- **Q-Q Plot for Normality:**



Inference: Since the Q–Q plot follows the 45° line y = x, it can be inferred that the driving white noise is normally distributed.

Now we assess the goodness of model built on simulated data by checking if the estimates are close to the parameters. For testing this, the null and alternative hypotheses are:

H_0: The process mean is zero.

H_1: The process mean differs significantly from zero.

H_0': Model parameter $\phi = 0$.

H_1': Model parameter differs significantly from zero.

Inference: Since 0.00 < 0.05, therefore we reject H_0 & H_0' at 5 %level of significance and conclude that the process mean and model parameter ϕ differ significantly from zero.

6) Forecast for next 20 time points:

Observation	Forecast
1	10.58
2	10.21
3	10.08
4	10.03
5	10.01
6	10
7	10
8	10
9	10
10	10
11	10
12	10
13	10
14	10
15	10
16	10
17	10
18	10
19	10
20	10

Case 17: Generate 1000 data points from the following ARMA$(1, 1)$ process.

$$\mathbf{X_t = 2 + 0.7X_{t-1} + \epsilon_t - \epsilon_{t-1}}$$

where ϵ_t is a Gaussian WN$(0, 1)$ and $X_0 = 6.2$. Assuming the generated data as a sample from some ARMA model do the following:

1. **Test for stationarity of the data using the Augmented Dickey Fuller (ADF) Test. Make the series stationary if it is not.**
2. **Fit an 'appropriate' order (identify it using Penalized Log Likelihood) ARMA model.**
3. **Check the goodness of the model by using the following:**
    a. **Stationary R-Square**
    b. **Root Mean Square Error (RMSE)**
    c. **Mean Absolute Percentage Error (MAPE)**
4. **Validate the assumption of driving Gaussian White Noise using the following:**

a. **Ljung–Box Test for White Noise**
b. **ACF and PACF for White Noise**
c. **Q-Q Plot for Normality**

5. **Assess the goodness of model built on simulated data by checking if the estimates are close to the parameters?**
6. **Apply the model and forecast for next 20 time points.**

1) Augmented Dickey-Fuller Test:

data: t17[, 3]
Dickey-Fuller = -15.5765, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary

Inference: Since, p-value (0.01) < 0.05, therefore null hypothesis is rejected and we conclude that the given AR(1) series is stationary.

2) Order Identification:

Model Type	BIC
ARIMA(0,0,1)	0.086
ARIMA(0,0,2)	0.065
ARIMA(0,0,3)	0.023
ARIMA(1,0,0)	0.091
ARIMA(1,0,1)	-0.043
ARIMA(1,0,2)	-0.036
ARIMA(1,0,3)	-0.029
ARIMA(2,0,0)	0.084
ARIMA(2,0,1)	-0.036
ARIMA(2,0,2)	-0.029
ARIMA(2,0,3)	-0.021
ARIMA(3,0,0)	0.078
ARIMA(3,0,1)	-0.029
ARIMA(3,0,2)	-0.021
ARIMA(3,0,3)	-0.013

From the above table, it can be observed that BIC corresponding to ARIMA (1, 0, 1) i.e. ARMA (1, 1) is m minimum. Thus it can be inferred that for the given series, the data points have been generated from an ARMA (1, 1) process.

3) Checking the goodness of the model:

Model Statistics

Model	Number of Predictors	Model Fit statistics				Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	RMSE	MAPE	Normalized BIC	Statistics	DF	Sig.	

xt_1-Model_1	0	.145	.969	11.982	-.043	18.498	16	.0296	0

ARIMA Model Parameters

					Estimate	SE	t	Sig.
Xt-Model_1	Xt	No Transformation	Constant		6.667	.000	16839.276	.000
			AR	Lag 1	.702	.023	30.255	.000
			MA	Lag 1	1.000	.068	14.809	.000

Note: For MA we get the negative of the parameter of the original model.

Stationary R-Squared

Since stationary r squared is 0.145 > 0, therefore the model under consideration is better than the baseline model.

Root Mean Squared Error (RMSE)

RMSE value is low indicating a good fit. Also the dependent series is close with its model-predicted level.

Mean Absolute Percentage Error (MAPE)

MAPE is used when assessing forecast accuracy. For the given case MAPE is 11.982 → % of error involved in forecasting was 12%. Hence the forecasted values are close to the actual values.

4) Validation of assumptions of driving Gaussian white noise:

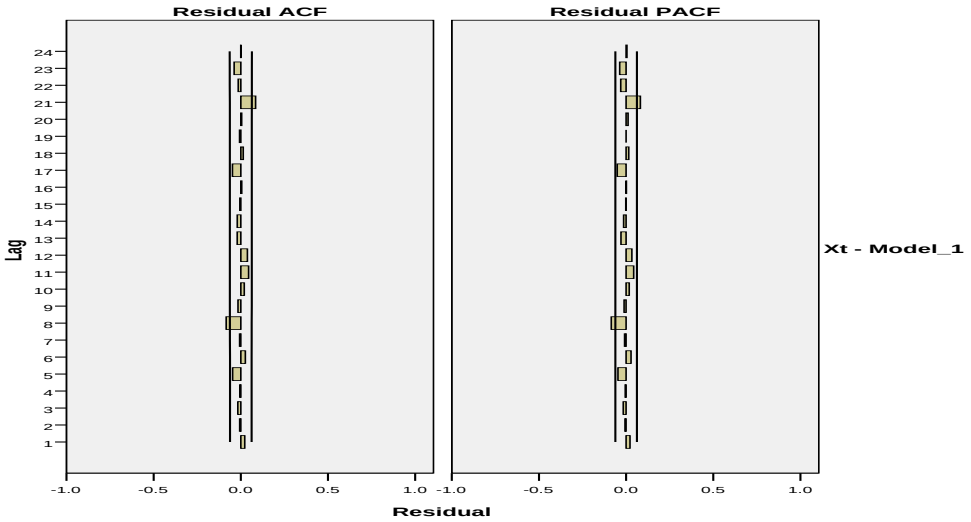
- **Ljung–Box Test for White Noise:**

  Now we test the randomness of the residuals using the Box-Ljung test.

  H_0: The residuals are random.
  H_1: The residuals are not random.
  Test statistic: Q = 18.498

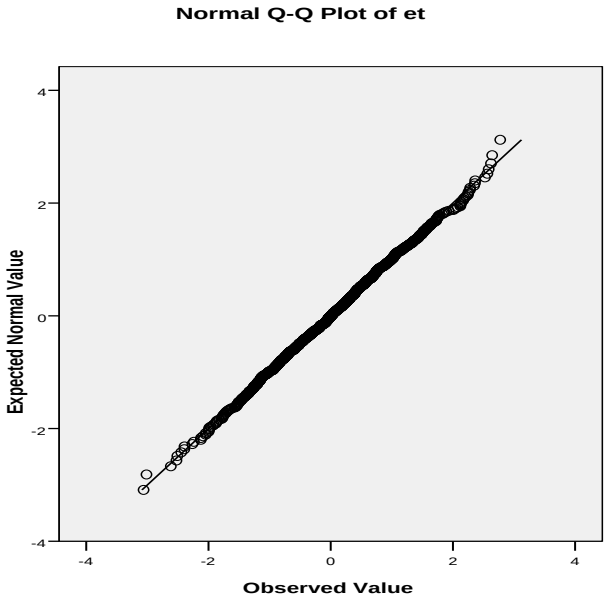
Inference: Since 0.296 > 0.05, therefore we may accept H_0 at 5% level of significance. Hence we may conclude that the residuals for the fitted model are random i.e. are independently distributed.

- **ACF and PACF for White Noise:**



Inference: Since both the ACF and PACF for residuals tail off indefinitely, it can be inferred that the residuals ϵ_t have been generated from white noise process.

- **Q-Q Plot for Normality:**



Inference: Since the Q–Q plot follows the 45° line y = x, it can be inferred that the driving white noise is normally distributed.

Now we assess the goodness of model built on simulated data by checking if the estimates are close to the parameters. For testing this, the null and alternative hypotheses are:

H_0: The process mean is zero.

H_1: The process mean differs significantly from zero.

H_0': Model parameter $\phi = 0$.

H_1': Model parameter differs significantly from zero.

Inference: Since 0.00 < 0.05, therefore we reject H_0 & H_0' at 5 %level of significance and conclude that the process mean and model parameter ϕ differ significantly from zero.

Forecasted value for next 20 years :

Year	Forecasted Values	Year	Forecasted Values
1001	7.014132	1011	6.676738
1002	6.909801	1012	6.673814
1003	6.836828	1013	6.671769
1004	6.785788	1014	6.670339
1005	6.750089	1015	6.669339
1006	6.725119	1016	6.668639
1007	6.707654	1017	6.668149
1008	6.695438	1018	6.667807
1009	6.686894	1019	6.667568
1010	6.680918	1020	6.6674