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: Vishal Kumar

Due Date: 11th April, 2016

Date of Submission: 10th 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.

Time Series Analysis

A time series is a sequence of data points, typically consisting of successive measurements made over a time interval. Time series are used in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, communications engineering, and largely in any domain of applied science and engineering which involves temporal measurements.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.

The additive model of the time series is given as:

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

The multiplicative model of the time series is given as:

$$y_t = m_t \text{ x } s_t \text{ x } c_t \text{ x } e_t$$

taking logarithm on both sides:

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

which is again the additive model of time series.

where $t = t_1, t_2, \ldots\ldots t_n, \ldots$ denotes the time points.

$m_t =$ Trend or long term time series

$s_t =$ Seasonal component of time series

$c_t =$ Cyclic component of time series

$e_t =$ Irregular or random component

The first three components are deterministic components and the fourth component is a random component.

One of the most famous methods of studying a time series is **Box-Jenkins method**.

This method involves the following steps:

1. Estimation and removal of deterministic components from the process.
2. Testing for stationarity of the process and making the series stationary if not.
3. Auto Regressive Integrated Moving Average modelling.

There are several elementary methods of estimating and eliminating the deterministic components.

To detect presence of Trend we have following methods:

Relative Ordering Test:

Let $y_1, y_2, \ldots\ldots\ldots y_n$ be the observed time series and we are interested in testing

$$H_o: \text{No trend is present.}$$

$$H_1: T\text{rend is present.}$$

Define:

$$q_{ij} = \begin{cases} 1 & if\, y_i > y_j when\ i < j \\ 0 & otherwise \end{cases}$$

Let $= \sum_{i<j} \sum q_{ij}$: # of decreasing points/ # of discordants. Under H_o of no trend dicordants and concordant will be equally probable, i.e. $P(q_{ij} = 0) = 1/2 = P(q_{ij} = 1), H\text{ence } (q_{ij}) = 1/2$.

$$\Rightarrow E(Q) = \sum\sum_{i<j} E(q_{ij}) = \frac{n(n-1)}{4}$$

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

Calculate τ = 1 − $\frac{4Q}{n(n-1)}$ .Clearly E(τ) = 0 and E(τ) =$\frac{2(2n+5)}{9n(n-1)}$

Define the test statistic as follows:

$$Z = \frac{\tau - \mathrm{E}(\tau)}{\sqrt{V(\tau)}} \sim \mathrm{N}(0,1) \text{ asymptotically under } H_o$$

Hence we can use the normal test. If H_o 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.

To remove Trend we have following methods:

1. **Graphical method**:

In this method we plot the line graph of the data provided against the time. On seeing the curve formed we can clearly state whether the data has increasing trend or decreasing trend. If the curve is rising then it is an increasing trend and if the curve is falling then it is a decreasing trend. Sometimes, it is not possible to state anything about trend using the graph. So we have other methods.

2. **Least square method**:

Given a set of data and the desire to produce some kind of model of those data, there are a variety of functions that can be chosen for the fit. If there is no prior understanding of the data, then the simplest function to fit is a straight line with the data plotted vertically and values of time ($t = 1, 2, 3, \ldots$) plotted horizontally.

Once it has been decided to fit a straight line, there are various methods to proceed, but the most usual choice is a least-squares fit. This method minimises the sum of the squared errors in the data series, denoted the y variable. Given a set of points in time t, and data values y_t observed for those points in time, values of a and b are chosen so that

$$\sum_t \left\{ [(at + b) - y_t]^2 \right\}$$

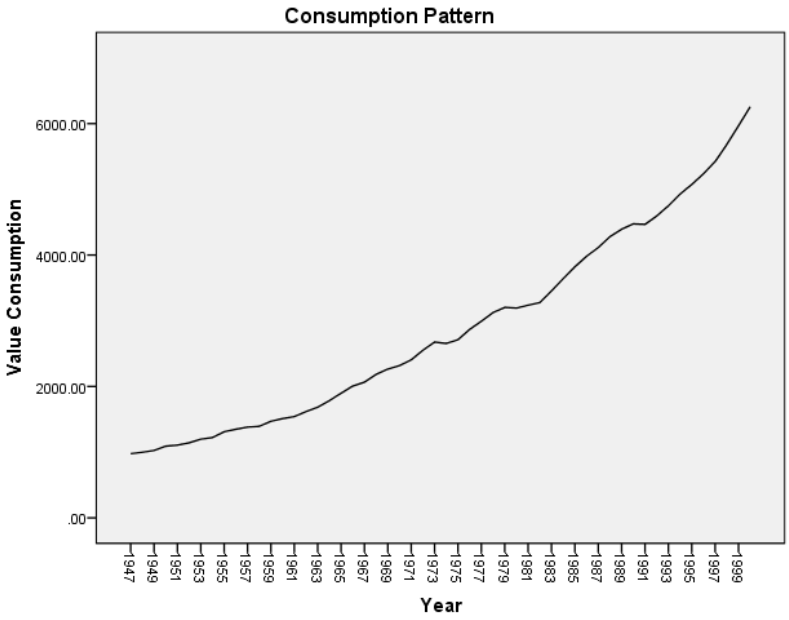
is minimised. Here at + b is the trend line, so the sum of squared deviations from the trend line is what is being minimised. This can always be done in closed form since this is a case of simple linear regression.

We are not able to use the method of least squares if our residuals are correlated. So we use following methods:

Moving average method: Exponential smoothing is a technique that can be applied to time series data, either to produce smoothed data for presentation, or to make forecasts. The time series data themselves are a sequence of observations. The observed phenomenon may be an essentially random process, or it may be an orderly, but noisy, process. Whereas in the simple moving average the past observations are weighted equally, exponential smoothing assigns exponentially decreasing weights over time.

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.

Solution: First we will plot the data to get an estimate of the trend presence in our data.



Conclusion

We see there is an increasing pattern the consumption expenditure for the United States for 1944 to 2000.

We will now perform a relative ordering test to confirm the presence of an increasing trend in the data.

Performing relative ordering test in R software, we get the following output:

```
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
```

Conclusion

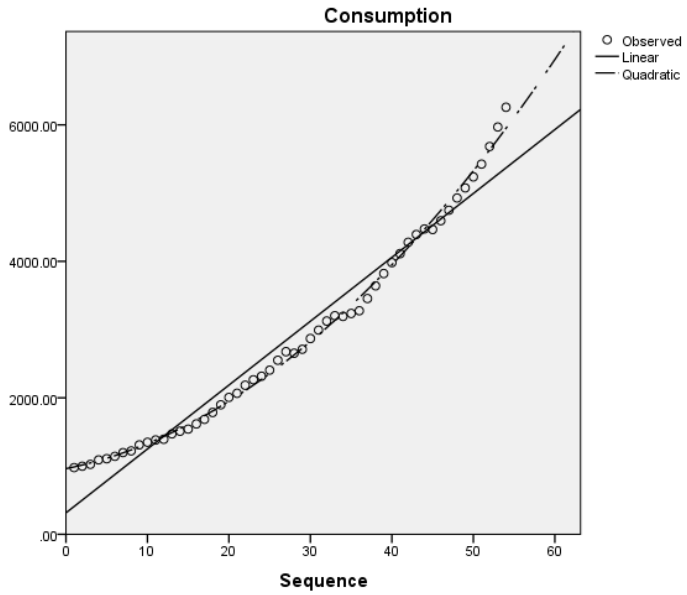
We reject the null hypothesis and conclude that an increasing trend is present. And now we will proceed further to estimate the trend presence in our data.

We will try to estimate the trend by curve estimation and see which type of curve fits the data well.

Model Summary and Parameter Estimates

Dependent Variable: Consumption

	Model Summary					Parameter Estimates		
Equation	R Square	F	df1	df2	Sig.	Constant	b1	b2
Linear	.963	1340.287	1	52	.000	314.216	93.605	
Quadratic	.997	7553.391	2	51	.000	962.181	24.180	1.262



Conclusion

We see that a quadratic pattern is observed in the consumption pattern, hence we will go through estimation of trend with a 2nd order polynomial regression.

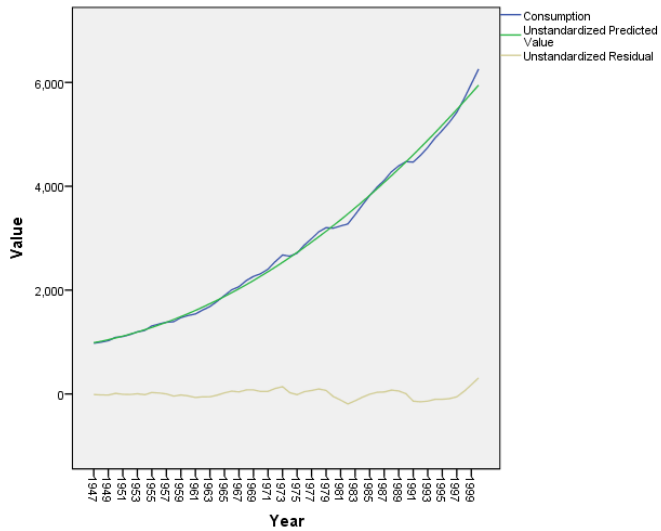
Trend will be estimated by the following model:

$$\hat{y_t} = 962.181 + 24.18x_t + 1.262x_t^2$$

The detrended values are:

YEAR	Detrended	YEAR	Detrended	YEAR	Detrended	YEAR	Detrended	YEAR	Detrended
1947	-11.2233	1958	-41.11069	1969	78.7327	1980	-50.4932	1991	-139.7883
1948	-17.4904	1959	-19.14768	1970	47.92577	1981	-118.77	1992	-150.9351
1949	-20.782	1960	-37.30922	1971	49.59431	1982	-193.071	1993	-138.1064
1950	11.80186	1961	-67.69529	1972	106.3383	1983	-130.597	1994	-103.0023
1951	-7.53882	1962	-54.90591	1973	140.6578	1984	-63.1478	1995	-102.1227
1952	-10.304	1963	-54.04106	1974	24.85267	1985	-4.22286	1996	-89.36764
1953	3.906199	1964	-21.60075	1975	-14.07695	1986	32.17759	1997	-54.63713
1954	-14.5081	1965	20.31502	1976	45.26888	1987	37.95349	1998	50.96885
1955	28.35306	1966	55.40625	1977	67.29018	1988	75.10486	1999	178.9503
1956	18.58968	1967	39.57294	1978	96.18694	1989	57.83168	2000	309.1072
1957	0.901767	1968	79.11509	1979	68.45916	1990	4.633969		

A plot of original series, estimated trend and de-trended values is given below:



Now a simple trend based on forecast for consumption expenditure for the next five years is given by:

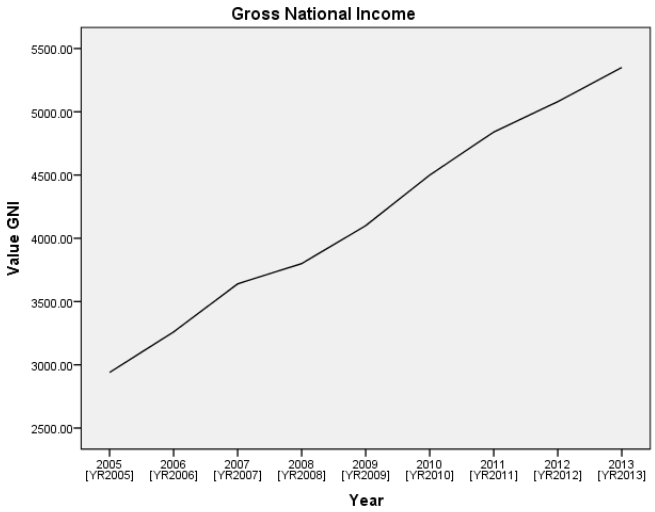
Year	Forecast
2001	6109.631
2002	6273.893
2003	6440.679
2004	6609.989
2005	6781.823

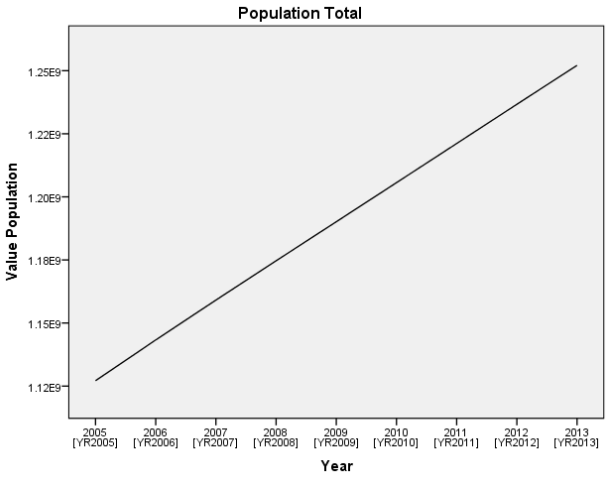
Case 2: Test for the presence of trend and estimate it if it's present for the following world development indicators for India: (time period)

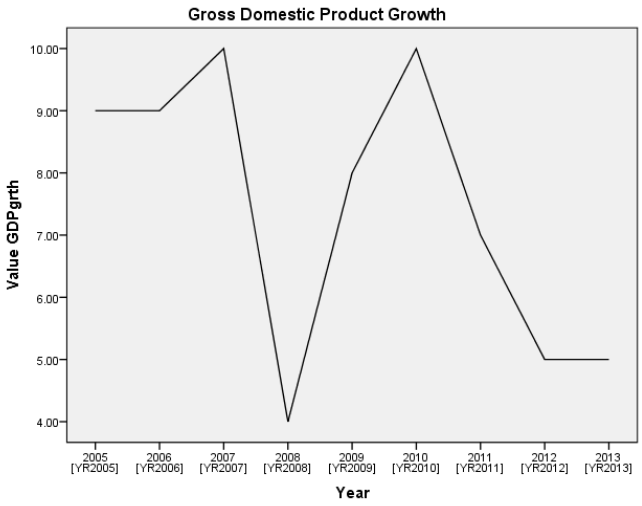
i. Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD.
ii. Population Total.
iii. Gross Domestic Product (GDP) (current USD).
iv. Gross Domestic Product (GDP) Growth (annual %).
v. Life Expectancy at birth (years)

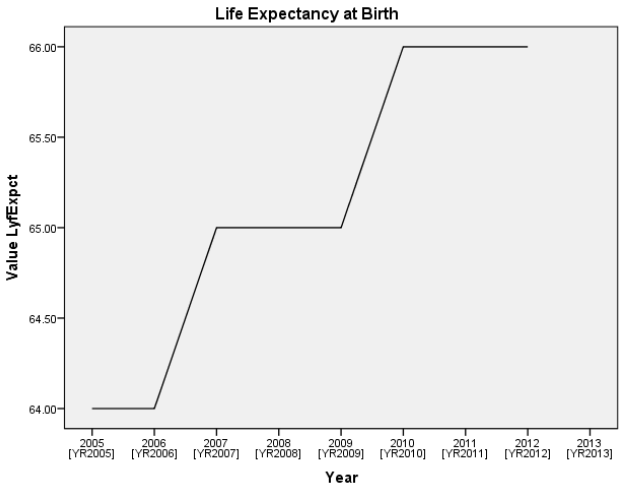
Obtain the de-trended indicators.

Solution: First we will plot the data to get an estimate of the trend presence in our data.









Conclusion

We see there is an increasing trend for Gross National Income, Population Total, Gross Domestic Product and Life Expectancy. However trend pattern is not clear in Gross Domestic Product Growth.

Applying the relative ordering test increasing trend exists for all the development parameter:

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

Conclusion

We reject null hypothesis and conclude that **GNI** shows **increasing** trend.

Now estimating the trend by curve estimation:

ANOVA[a]

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

a. Dependent Variable: GNI

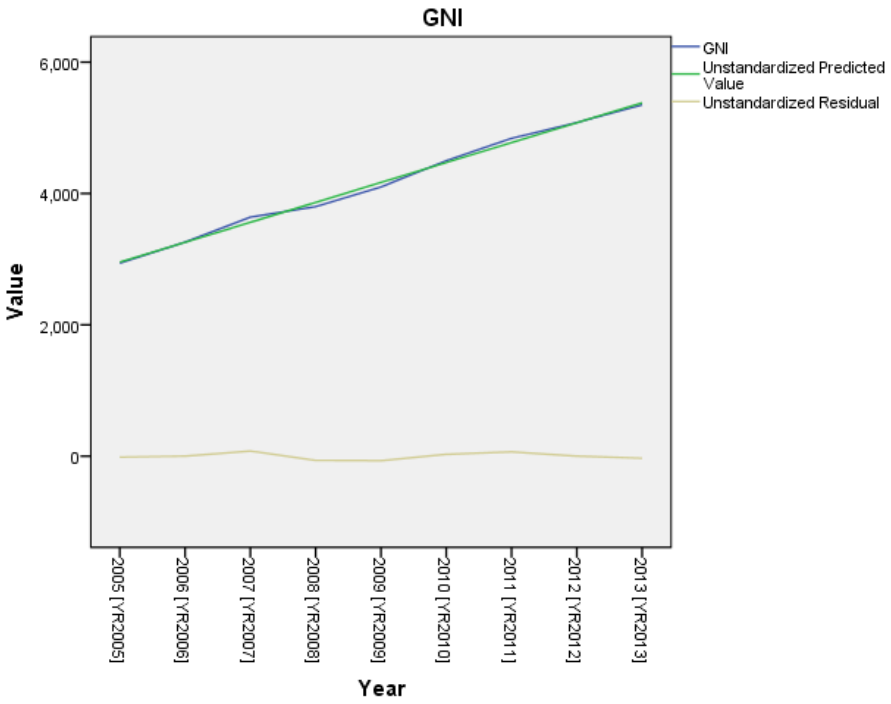
b. Predictors: (Constant), S.no

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2651.111	40.064		66.172	.000
	S.no	303.333	7.120	.998	42.606	.000

a. Dependent Variable: GNI

$$\hat{y}_t = 2651.111 + 303.333x_t$$



De-trended values are:

Year	GNI
2005	-14.4444
2006	2.22222
2007	78.88889
2008	-64.4444
2009	-67.7778
2010	28.88889
2011	65.55556
2012	2.22222
2013	-31.1111

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

Conclusion

We reject null hypothesis and conclude that **Population Total** shows **increasing** trend.

Now estimating the trend by curve estimation:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14578545755190400.000	1	14578545755190400.000	219472.584	.000[b]
	Residual	464977531953.175	7	66425361707.596		
	Total	14579010732722350.000	8			

a. Dependent Variable: Pop

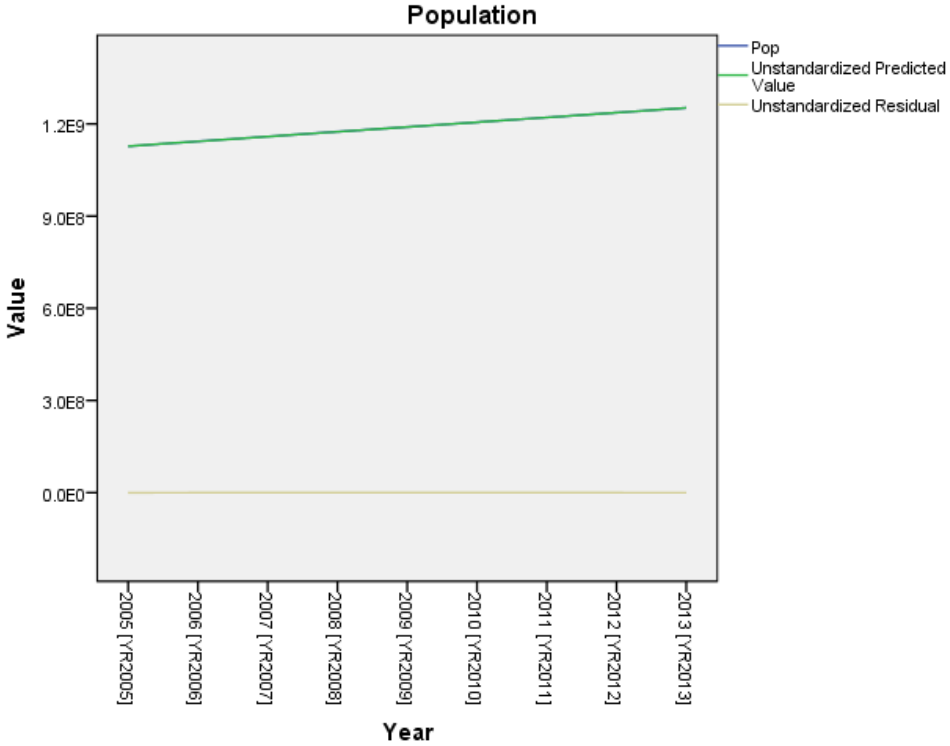
b. Predictors: (Constant), S.no

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1112054472.611	187237.362		5939.277	.000
	S.no	15587679.833	33272.952	1.000	468.479	.000

a. Dependent Variable: Pop

$$\hat{y_t} = 1112054472.611 + 15587679.833x_t$$



De-trended values are:

Year	Population
2005	-498604
2006	59517.72
2007	277737.9
2008	257142.1
2009	145197.2
2010	44096.39
2011	-11912.4
2012	-69179.3
2013	-203995

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

Conclusion

We reject null hypothesis and conclude that **GDP** shows **increasing** trend.

Now estimating the trend by curve estimation:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1257732816666666500000000	1	1257732816666666500000000	109.569	.000[b]
	Residual	80352072222222340000000	7	11478867460317478000000		
	Total	1338084888888889000000000	8			

a. Dependent Variable: GDP

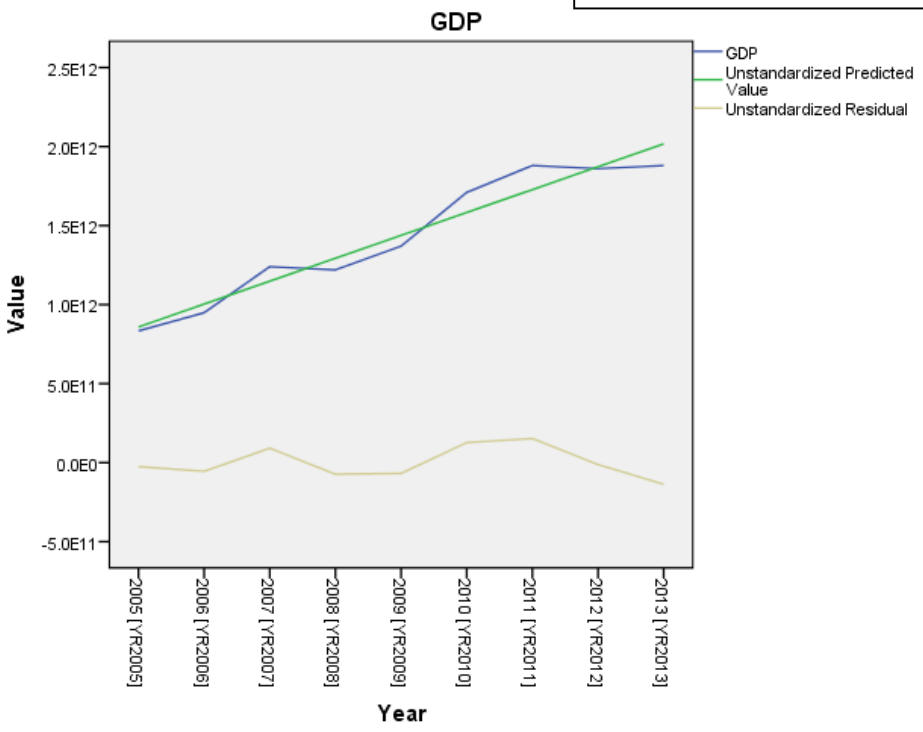
b. Predictors: (Constant), S.no

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	714194444444.444	77835025275.335		9.176	.000
	S.no	144783333333.333	13831646961.659	.970	10.468	.000

a. Dependent Variable: GDP

$$\hat{y}_t = 714194444444.444 + 144783333333.333x_t$$



De-trended values are:

Year	GDP
2005	-2.5E+10
2006	-5.5E+10
2007	9.15E+10
2008	-7.3E+10
2009	-6.8E+10
2010	1.27E+11
2011	1.52E+11
2012	-1.2E+10
2013	-1.4E+11

```
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
```

Conclusion

We fail reject null hypothesis and conclude that **GDP(growth)** shows **absence** of trend.

Hence the original series is free from trend.

```
Relative Ordering Test for Presence of Trend

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

Test Statistic: 3.4641
p_value: 3e-04
No. of Discordants: 0
Expected No. of Discordants: 14
```

Conclusion

We reject null hypothesis and conclude that **Life Expectancy** shows **increasing** trend.

We find small trend here i.e.

$$y_t = m_t$$

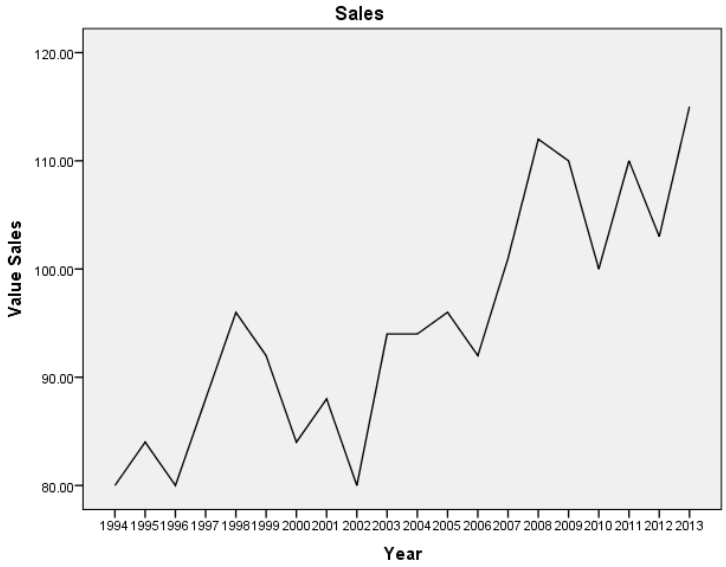
Trend estimation by simple linear regression is not possible, hence we cannot obtain de-trended values. However we can go for Moving average process to get an estimate of trend

De-trended Values are:

Year	Life Expectancy
2005	
2006	-0.32143
2007	0.35714
2008	0.03571
2009	-0.28571
2010	0.39286
2011	0.07143
2012	-0.25
2013	

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.

Solution: First we will plot the data to get an estimate of the trend presence in our data.



Conclusion

We see there is an increasing pattern in the Sales value across 1994-2013.

We will now perform a relative ordering test to conform the presence of an increasing in the data.

Performing relative ordering test in R software, we get the following output:

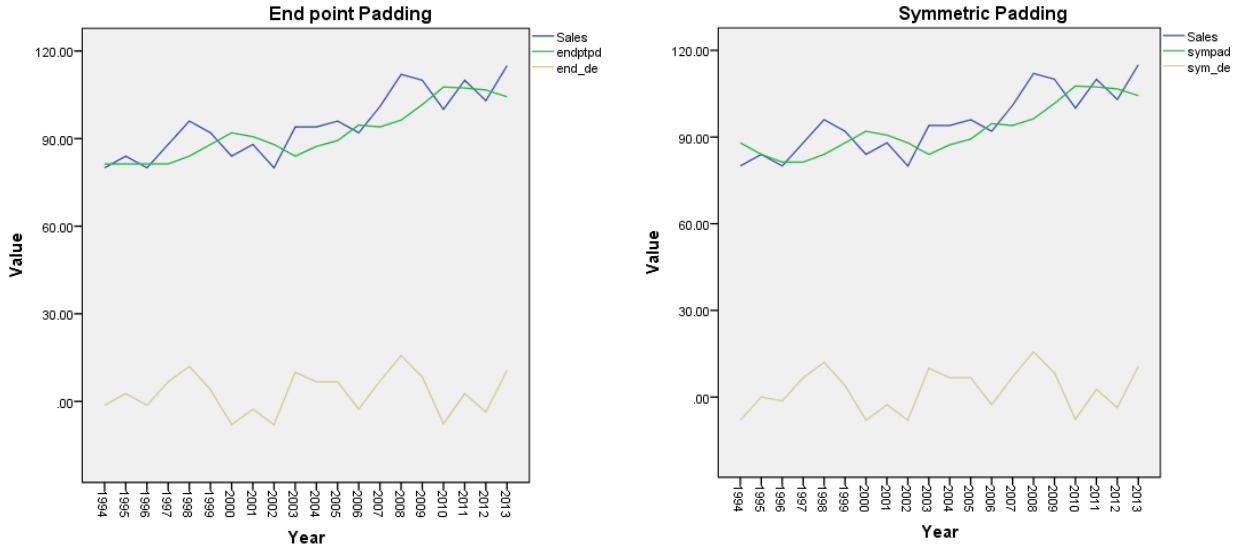
```
Relative Ordering Test for Presence of Trend

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
```

Conclusion:

We reject the null hypothesis and conclude that an **increasing** trend is present in the Sales value across 1994-2013. And now we will proceed further to estimate the trend presence in our data.

We will estimate the trend by **moving average** method. We see peaks after every 3 years, hence we will go for a prior moving average (because we need to forecast for the next 3 years) of the order 3.

We will plot the Moving averages for 2 padding scheme i.e. end point padding and symmetric padding:



Conclusion: We see end point padding scheme yields better results, hence we will keep de-trended values from end point padding scheme.

Year	Sales (in million $)	Prior MA	Detrended	Year	Sales (in million $)	Prior MA	Detrended
1994	80	81.33	-1.33	2004	94	87.33	6.67
1995	84	81.33	2.67	2005	96	89.33	6.67
1996	80	81.33	-1.33	2006	92	94.67	-2.67
1997	88	81.33	6.67	2007	101	94	7
1998	96	84	12	2008	112	96.33	15.67
1999	92	88	4	2009	110	101.67	8.33
2000	84	92	-8	2010	100	107.67	-7.67
2001	88	90.67	-2.67	2011	110	107.33	2.67
2002	80	88	-8	2012	103	106.67	-3.67
2003	94	84	10	2013	115	104.33	10.67

To detect the presence of Seasonality we have the following method:

Freidman's Test for the Presence of Seasonality: Let $y_1, y_2, \dots \dots . y_n$ be the observed time series and we are interested in testing H_o of no seasonality. Friedman described the following test procedure for monthly seasonality:

1. Remove trend if necessary.
2. Rank the de-trended data within each year from smallest (1) to largest (12).
3. Let M_{ij} be the rank of i^{th} month in j^{th} year, then under the assumption of no seasonality for a particular year j, $\{M_{1j}, M_{2j}, \dots \dots M_{12j}\}$ can be any random permutation of $\{1, \dots, 12\}$.
4. Obtain the monthly totals of the ranks across different years and call them as $M_1, M_2, \dots \dots, M_{12}$.The average of the sequence 1,2,…,r is $\frac{(r+1)}{2}$, hence $\frac{c(r+1)}{12}$ 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.
5. Calculate the test statistic:

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

Hence we can use the chi-square test.

If calculated chi-square is > tabulated chi-square then we reject the null hypothesis of no seasonality and conclude that seasonality is present in the data.

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 a deterministic components based forecast for the monthly Whole Sale Price Index for the next 5 months.

Solution: First we will plot the data to get an estimate of the trend presence in our data.

Conclusion

From plot we don't suspect trend present in the data, this is because there is a sharp decrease in WPI-inflation around 2009 due to crash in stock exchange. This point might be the reason of no trend.

We will now perform a relative ordering test to detect presence of trend in our data.

Performing relative ordering test in R software, we get the following output:

```
Relative Ordering Test for Presence of Trend

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

Test Statistic: 1.0881
p_value: 0.1383
No. of Discordants: 3107
Expected No. of Discordants: 3335
```

Conclusion

We fail to reject the null hypothesis and conclude that no trend is present.

Now we will proceed further to estimate the seasonality present in the data, we will use the Friedman test to detect seasonality:

The data available to us is from April, 2005 to November, 2014, data points are not a multiple of 12, hence towards the end we will drop 8 data points and apply Friedman test:

```
Freidman (JASA) Test for Presence of Seasonality

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

Test Statistic: 12.7692 (Chi Sqaure with 11 df)
p_value: 0.3087
```

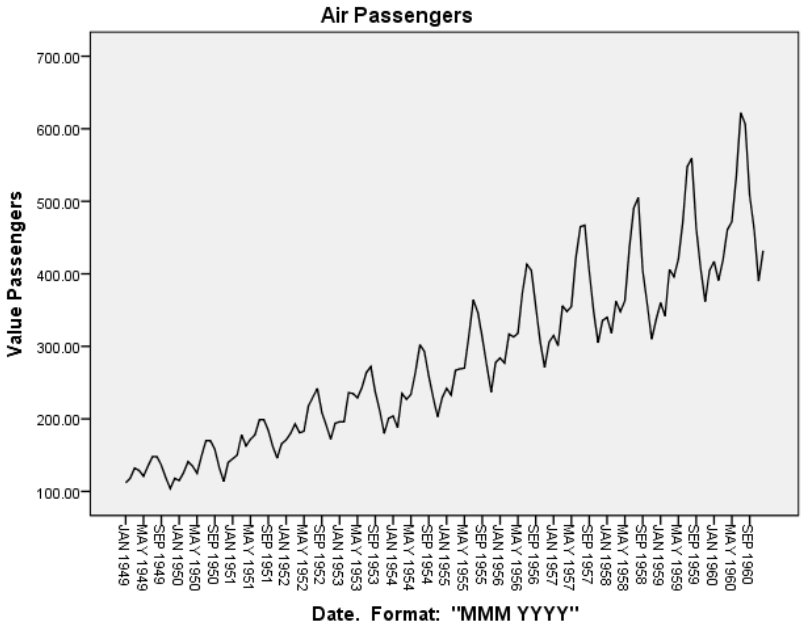
Conclusion

We see p value > 0.05, hence we fail to reject null hypothesis and conclude that no seasonality is present in the data.

Conclusion: We are in a situation that we are unable to extract deterministic component out of the data, hence we stop at this point. For further analysis we can suggest either analysing the series in two parts to estimate deterministic component, before 2009 and after 2009, or we can ignore data points from year 2009 and then analyse the data to estimate deterministic component.

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 a deterministic components based forecast for the monthly World Airline Passengers for the next 5 months.

Solution: First we will plot the data to get an estimate of the trend presence in our data.



Conclusion

Form the plot we see **increasing** trend in the no. of passengers flying between 1949 and 1960.

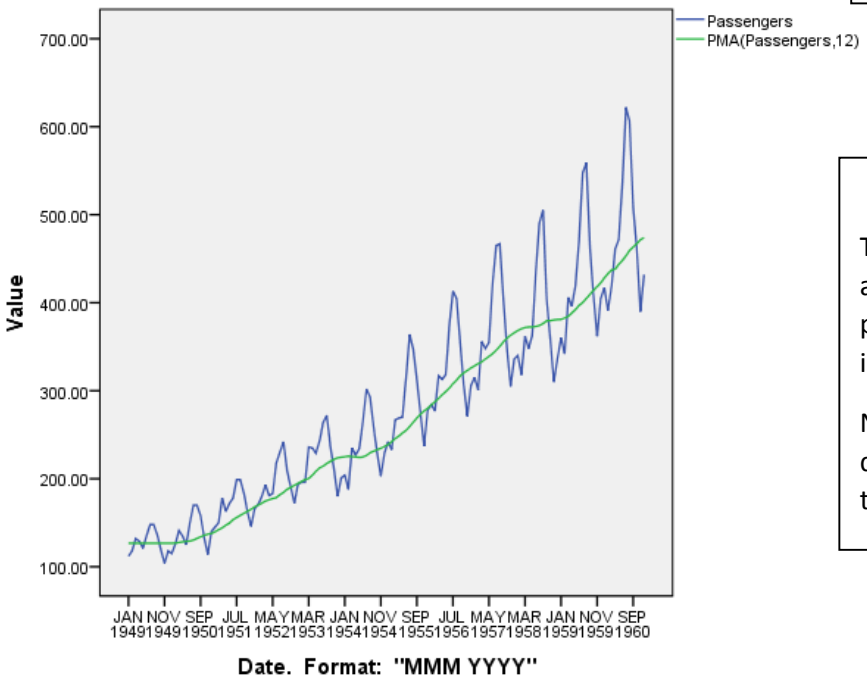
We will now perform a relative ordering test to detect presence of trend in our data.

Performing relative ordering test in R software, we get the following output:

```
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
```

Conclusion

We reject the null hypothesis and conclude that trend is present in the data. And now we will proceed further to estimate the trend presence in our data.



Conclusion

Trend is estimated by Prior moving average of span 12 and shown in the plot, we have found an **increasing** trend in the data.

Now we will remove the trend from the data and test for seasonality present in the data by Friedman test.

```
Freidman (JASA) Test for Presence of Seasonality

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

Test Statistic: 239.5641 (Chi Sqaure with 11 df)
p_value: 0
```

Conclusion

We reject the null hypothesis and conclude that seasonality is present in the data. We will now proceed to estimate seasonality present in the data by seasonal decomposition method.

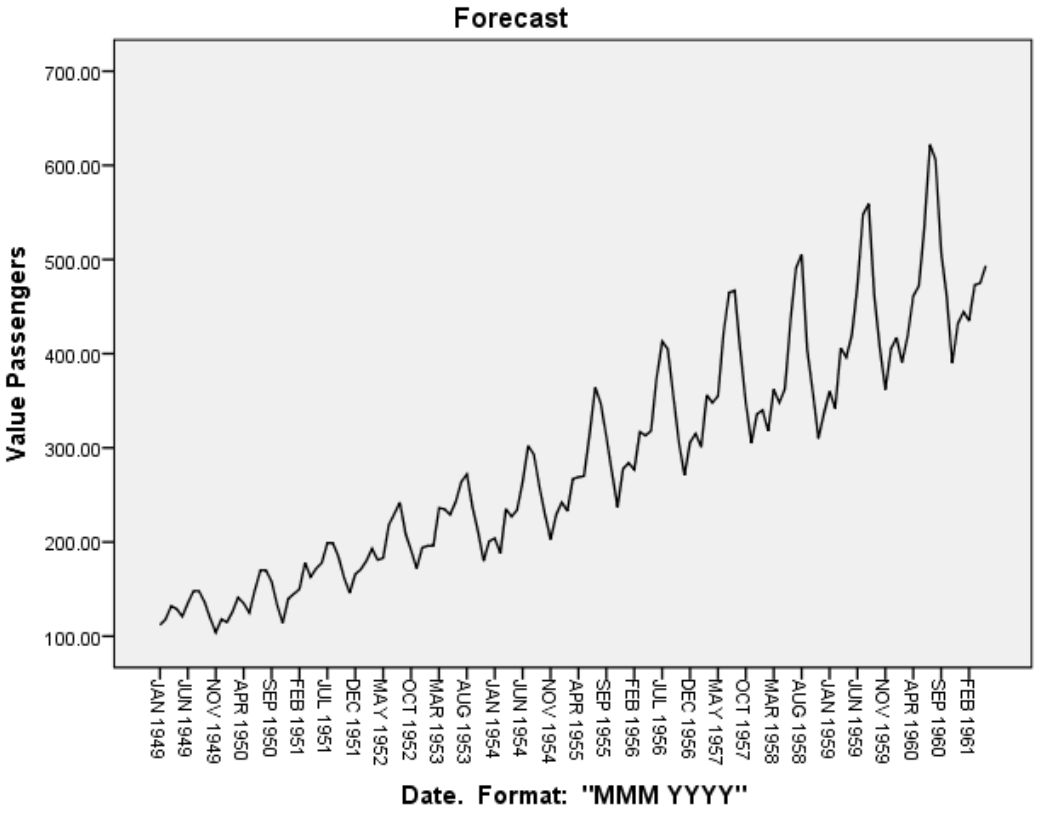
After calculating all the deterministic components from the time series data, deterministic components based forecast for the monthly World Airline Passengers for the next 5 months is given by:

$$y_t = m_t + s_t + c_t$$

To forecast trend for the next month we will take average of trend present in the data for the last 3 months. To forecast seasonality we will take average of seasonality present in the for the last 3 data point of the same month in last 3 years and same for forecasting cyclicity.

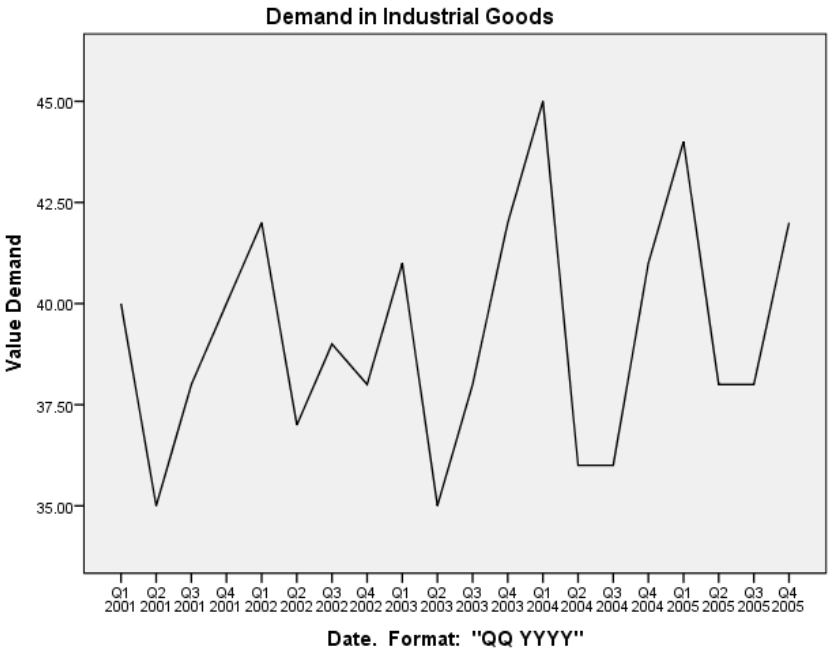
Deterministic components based forecast for the monthly World Airline Passengers for the next 5 months:

Date	Trend	Seasonality	Cyclicity	Forecast
Jan, 1961	470.86	-25.0034	-1.33322	444.5234
Feb, 1961	472.12	-36.3398	-0.88646	434.8937
March, 1961	472.3	-2.26597	2.984057	473.0181
April, 1961	471.76	-7.7988	11.07034	475.0315
May, 1961	472.06	-4.41686	25.73815	493.3813



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 in 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 a deterministic components based forecast for the quarterly demand for the industrial good for the next 2 quarters.

Solution: First we will plot the data to get an estimate of the trend presence in our data.



Conclusion

From the plot we see slightly increasing trend, we will apply relative ordering test to confirm whether trend is present in the data or not.

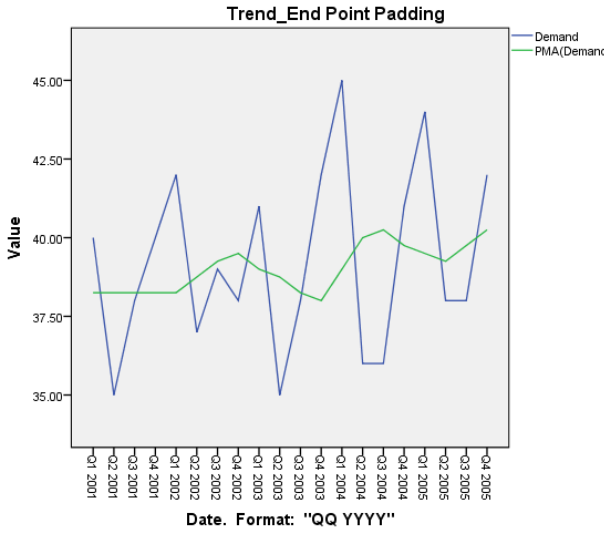
Performing relative ordering test in R software, we get the following output:

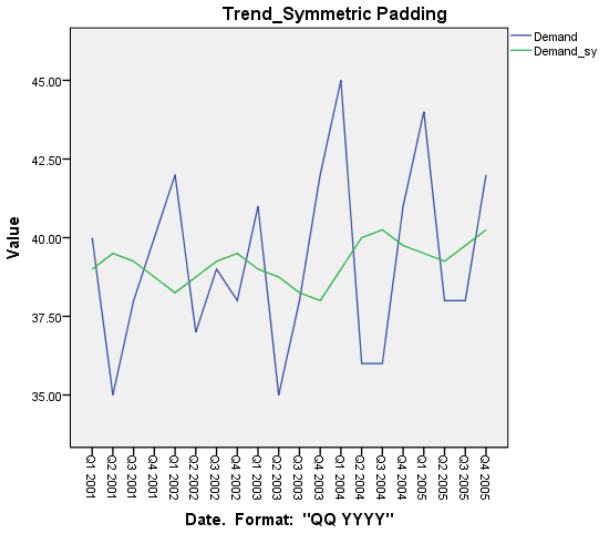
```
Relative Ordering Test for Presence of Trend

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
```

Conclusion

We see p-value is close to 0.05, we will still take this to be significant p-value, we will reject the null hypothesis and conclude that trend is present in the data.

Now we will proceed further to estimate the trend presence in our data. We will use Prior Moving Average of order 4 to estimate trend present in the data.





Conclusion: We see trend estimation by end point padding gives close estimate of trend to the real data, hence we will obtain the de-trended data end point moving average and test for seasonality present in the data by Friedman test:

```
Freidman (JASA) Test for Presence of Seasonality

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

Test Statistic: 7.48 (Chi Sqaure with 3 df)
p_value: 0.0581
```

Conclusion

We see p-value is close to 0.05, we will still reject the null hypothesis and conclude that seasonality is present in the data.

We will now proceed to estimate seasonality present in the data by seasonal decomposition method.

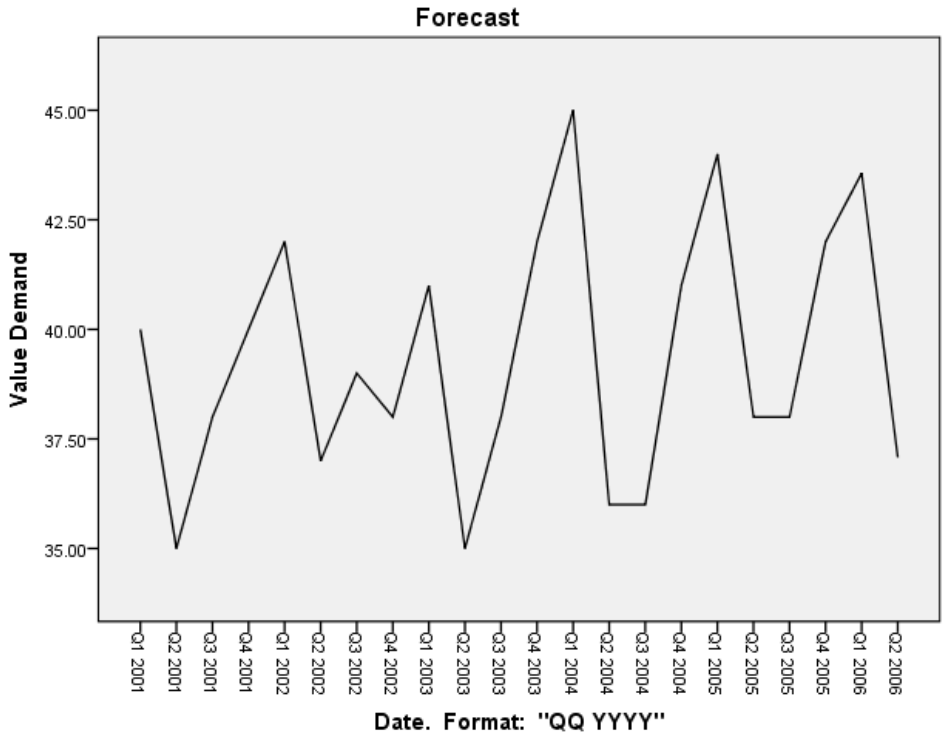
We have calculated all the deterministic components from the time series data, a deterministic components based forecast for the quarterly demand for the industrial good for the next 2 quarters.

$$y_t = m_t + s_t + c_t$$

To forecast trend for the next quarter we will take average of trend present in the data for the last 3 months. To forecast seasonality we will take average of seasonality present in the for the last 3 data point of the same month in last 3 years and same for forecasting cyclicity.

Deterministic components based forecast for the quarterly demand for the industrial good for the next 2 quarters:

Time	Trend	Seasonality	Cyclicity	Forecast
Q1 2006	39.75	3.62031	0.190337	43.56065
Q2 2006	39.91667	-3.03594	0.198437	37.07916



Stationarity

Stationarity means that a statistical properties of a process do not change over time. In statistics and econometrics, an augmented Dickey–Fuller test (ADF) is a test for a unit root in a time series sample. It is an augmented version of the Dickey–Fuller test for a larger and more complicated set of time series models. The augmented Dickey–Fuller (ADF) statistic, used in the test, is a negative number. The more negative it is, the stronger the rejection of the hypothesis that there is a unit root at some level of confidence.

Let us consider an AR(p) process:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \ \dots\dots + \Phi_p X_{t-p} + \epsilon_t$$

Where $\epsilon_t \sim WN(0, \sigma^2)$. We can write,

$$\nabla X_t = \beta X_{t-1} + \sum_{i=1}^{p-1} \alpha_i \nabla X_{t-p+1} + \epsilon_t$$

Where β and α_i's are appropriately defined.

Defining the null hypothesis is:

$$H_o: \beta = 0 \, , the\ AR(p)\ process\ is\ not\ stationary$$

Vs

$$H_1: \beta < 0 \, , the\ AR(p)\ process\ is\ stationary$$

Estimation of β is done using ordinary least square technique. The dickey fuller test statistics is given by:

$$DF = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Follows a τ statistics whose values have already been calculated using Monte Carlo Simulations.

We will use R software to test stationarity using ADF test:

R software has an inbuilt test for adf test in package "tseries" which has the null hypothesis that there exists a unit root i.e. the time series is not stationary and the alternative hypothesis that there is no unit root i.e. the time series is stationary.

Case 7: Consider the de-trended series for consumption expenditure (in million dollars) for the United States for 1944 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.

Solution: Performing ADF test on the de-trended series, we get:

```
Augmented Dickey-Fuller Test

data:  x[, 1]
Dickey-Fuller = -2.2273, Lag order = 3, p-value = 0.483
alternative hypothesis: stationary
```

Conclusion: We see p-value > 0.05, hence we fail to reject null hypothesis and conclude that given series is **not stationary**. So we go for differencing to make it stationary.

```
Augmented Dickey-Fuller Test

data:  diff(x[, 1])
Dickey-Fuller = -2.7359, Lag order = 3, p-value = 0.2782
alternative hypothesis: stationary
```

Conclusion: We see p-value > 0.05, hence we fail to reject null hypothesis and conclude that given series is **not stationary**. So we again go for differencing to make it stationary.

```
Augmented Dickey-Fuller Test

data:  diff(x[, 1], differences = 2)
Dickey-Fuller = -5.1833, Lag order = 3, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: We see p-value < 0.05, hence we reject null hypothesis and conclude that given series is **stationary**.

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

i. Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD,
ii. Population Total,
iii. Gross Domestic Product (GDP) (current USD),
iv. Gross Domestic Product (GDP) Growth (annual %) and
v. 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 they are not.

Solution: Performing ADF test on the de-trended series, we get:

1. **Gross National Income (GNI) per capita based on Purchasing Power Parity.**

```
Augmented Dickey-Fuller Test

data:  x[, 2]
Dickey-Fuller = -26.722, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

2. **Population Total**

```
Augmented Dickey-Fuller Test

data:  x[, 3]
Dickey-Fuller = -0.025305, Lag order = 2, p-value = 0.99
alternative hypothesis: stationary
```

Conclusion: P-value is > 0.05 hence we fail reject the null hypothesis and conclude that the series is **not stationary**.

To make it stationary we will go for first order differencing:

```
Augmented Dickey-Fuller Test

data:  diff(x[, 3])
Dickey-Fuller = -16.041, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

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

```
Augmented Dickey-Fuller Test

data:  x[, 4]
Dickey-Fuller = -3.2277, Lag order = 2, p-value = 0.1047
alternative hypothesis: stationary
```

Conclusion: P-value is > 0.05 hence we fail reject the null hypothesis and conclude that the series is **not stationary.** To make it stationary we will go for first order differencing.

```
Augmented Dickey-Fuller Test

data:  diff(x[, 4])
Dickey-Fuller = -2.0387, Lag order = 1, p-value = 0.5576
alternative hypothesis: stationary
```

Conclusion: P-value is > 0.05 hence we fail reject the null hypothesis and conclude that the series is **not stationary**.

To make it stationary we will go for second order differencing:

```
Augmented Dickey-Fuller Test

data:  diff(x[, 4], differences = 2)
Dickey-Fuller = -9.4042, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

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

We were unable to detect the trend in the data, hence we consider the current series to be free from trend and we will perform ADF test on the given series only:

```
Augmented Dickey-Fuller Test

data:  x[, 5]
Dickey-Fuller = -2.7033, Lag order = 2, p-value = 0.3045
alternative hypothesis: stationary
```

Conclusion: P-value is > 0.05 hence we fail reject the null hypothesis and conclude that the series is **not stationary**.

To make it stationary we will go for first order differencing:

```
Augmented Dickey-Fuller Test

data: diff(x[, 5])
Dickey-Fuller = -2.4143, Lag order = 1, p-value = 0.4145
alternative hypothesis: stationary
```

Conclusion: P-value is > 0.05 hence we fail reject the null hypothesis and conclude that the series is **not stationary**.

To make it stationary we will go for second order differencing:

```
Augmented Dickey-Fuller Test

data:  diff(x[, 5], differences = 2)
Dickey-Fuller = -3.7967, Lag order = 1, p-value = 0.03595
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

5. **Life Expectancy at birth (years).**

```
Augmented Dickey-Fuller Test

data:  na.omit(x[, 5])
Dickey-Fuller = -215140, Lag order = 1, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

Case 9: Consider the de-trended series for annual sales measured in million USD for a trading company for 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.

Solution: Performing ADF test on the de-trended series, we get:

```
Augmented Dickey-Fuller Test

data:  x[, 1]
Dickey-Fuller = -3.8629, Lag order = 2, p-value = 0.03122
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the 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.

Solution: Performing ADF test on the de-trended series, we get:

```
Augmented Dickey-Fuller Test

data:  x[, 1]
Dickey-Fuller = -5.5305, Lag order = 4, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the 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.

Solution: Performing ADF test on the de-trended series, we get:

```
Augmented Dickey-Fuller Test

data:  x[, 1]
Dickey-Fuller = -7.0277, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the 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.

Solution: Performing ADF test on the de-trended series, we get:

```
Augmented Dickey-Fuller Test

data:  x[, 1]
Dickey-Fuller = -5.1698, Lag order = 2, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: P-value is < 0.05 hence we reject the null hypothesis and conclude that the series is **stationary**.

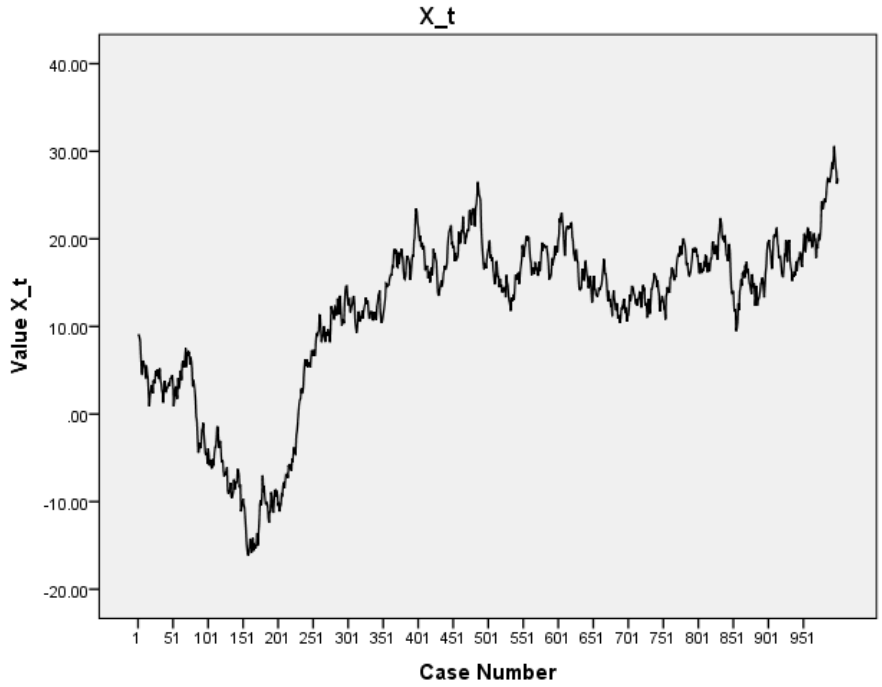
Case 13: Generate 1000 data points from the following AR(1) process:

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

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.

Solution: After generating data points we will plot this series:



Now we will test for stationary using ADF test in R:

```
Augmented Dickey-Fuller Test

data:  x[, 3]
Dickey-Fuller = -2.0423, Lag order = 9, p-value = 0.5604
alternative hypothesis: stationary
```

Conclusion: We fail to reject null hypothesis and conclude that given time series is **not stationary**. Now we will use first order differencing to make this series stationary. This is justified by the fact $|\varphi| = 1$ which is not less than 1. Hence the series is not to be stationary.

```
Augmented Dickey-Fuller Test

data:  diff(x[, 3])
Dickey-Fuller = -10.908, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: We reject the null hypothesis and conclude that given time series is **stationary**.

Autocorrelation Functions

Time domain analysis is the approach based on the autocorrelation functions to make inference from an observed time series to estimate the model. The autocorrelation function is an extremely important tool to describe the properties of a stationary stochastic process. The theoretical ACF is defined as follows:

$$\rho(k) = corr(x_t, x_{t-k}) = \frac{cov(x_t, x_{t-k})}{\sqrt{var(x_t)}\sqrt{var(x_{t-k})}}$$

And for a stationary process it takes the form

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

The properties of an ACF is given as:

$$\rho(0) = 1$$

1. ACF is an even function i.e. $\rho(-k) = \rho(k) \quad \forall\, k$
2. ACF is negative non definite.

Although a given stochastic process has a unique covariance structure, the converse is generally not true i.e. it is possible to find more than one stochastic process with the same ACF. Hence uniqueness property does not hold for ACF.

For identification of order of the ARMA process we use the following table:

MODEL	ACF	PACF
White noise	Cuts off with one significant spike at lag 0	Cuts off with one significant spike at lag 0
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

Sample ACF is the sample counter part of ACF based on the sample values, given as:

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$$

Partial ACF function has the similar properties for AR processes as the ACF function has for the MA process. The Partial ACF for AR(p) process cuts off after q lags.

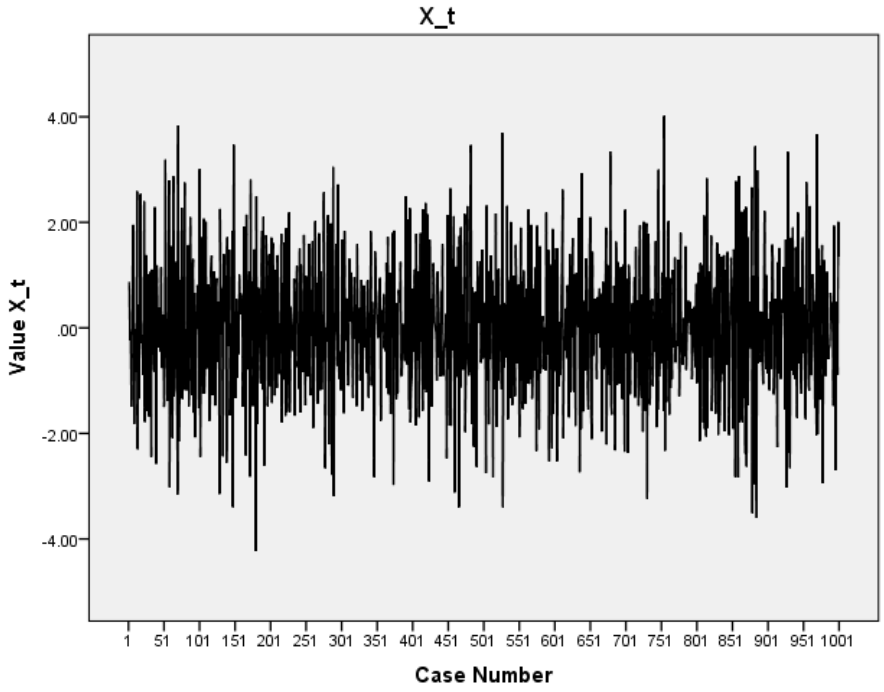
Case 14: Generate 1000 data points from the following (1) process:

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

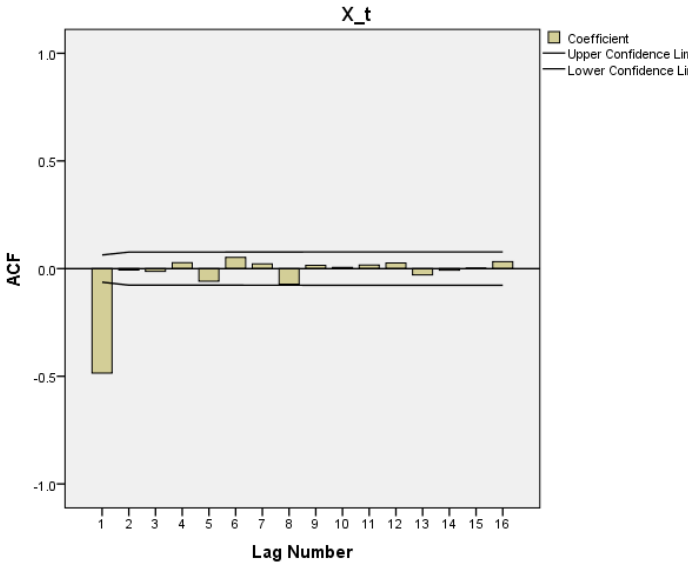
where ϵ_t is a Gaussian $WN(0,1)$.

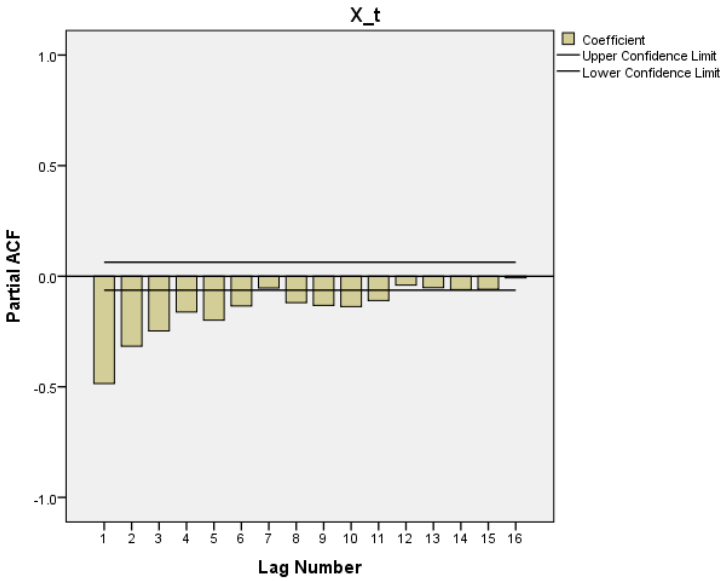
Assuming the generated data as a sample from some $ARMA$ model, identify its order using sample Correlogram and sample Partial Correlogram.

Solution: After generating the series we will plot it.



Now we will plot ACF and PACF:





Conclusion: In ACF plot we see a significant spike at lag 1 and in PACF plot we see spikes are tailing off, hence we conclude that process may be coming from **ARMA(0, 1)** model.

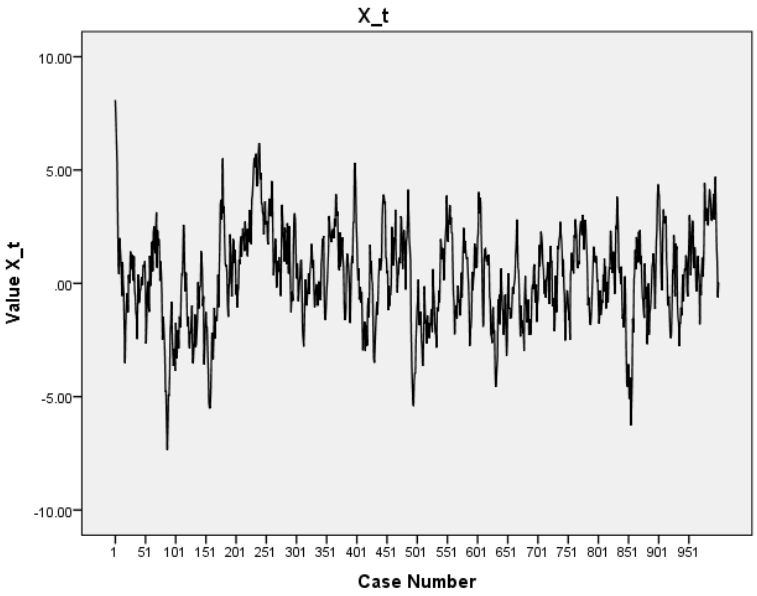
Case 15: Generate 1000 data points from the following $AR(1)$ process:

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

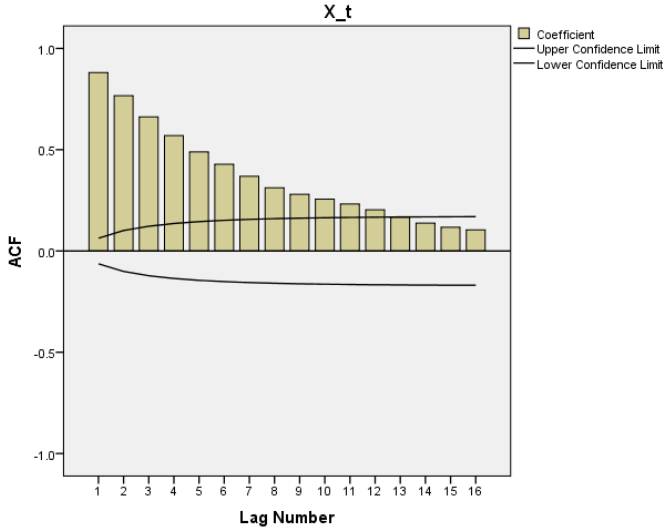
where ϵ_t is a Gaussian $WN(0,1)$ and $X_0 = 10$.

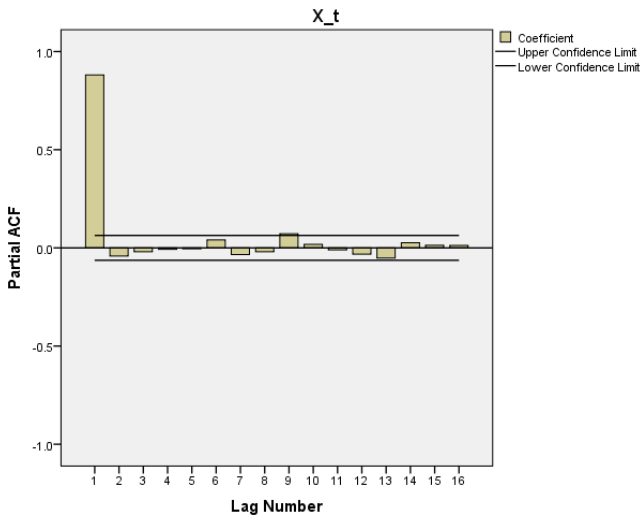
Assuming the generated data as a sample from some $ARMA$ model, identify its order using sample Correlogram and sample Partial Correlogram.

Solution: After generating the data we will plot the data:



Now we will plot ACF and PACF:





Conclusion: In PACF plot we see a significant spike at lag 1 and in ACF plot we see spikes are tailing off, hence we conclude that process may be coming from **ARMA(1, 0)** model.

ARIMA Modelling

Stationary R squared: A measure that compares the stationary part of the model to a simple mean model. This measure is preferable to ordinary R square 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.

Root Mean Square Error: The square root of mean square error. A measure of how much a dependent series varies from its model-predicted level, expressed in the same units as the dependent series.

Mean Absolute Percentage Error: A measure of how much a dependent series varies from its model-predicted level. It is independent of the units used and can therefore be used to compare series with different units

Ljung–Box test: This is a type of statistical test of whether any of a group of autocorrelations of a time series are different from zero. Instead of testing randomness at each distinct lag, it tests the "overall" randomness based on a number of lags.

The Ljung–Box test can be defined as follows.

$\boldsymbol{H_0}$: The data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process).

$\boldsymbol{H_1}$: The data are not independently distributed; they exhibit correlation.

The Ljung–Box test is commonly used in autoregressive integrated moving average (ARIMA) modeling. Note that it is applied to the residuals of a fitted ARIMA model, not the original series, and in such applications the hypothesis actually being tested is that the residuals from the ARIMA model have no autocorrelation.

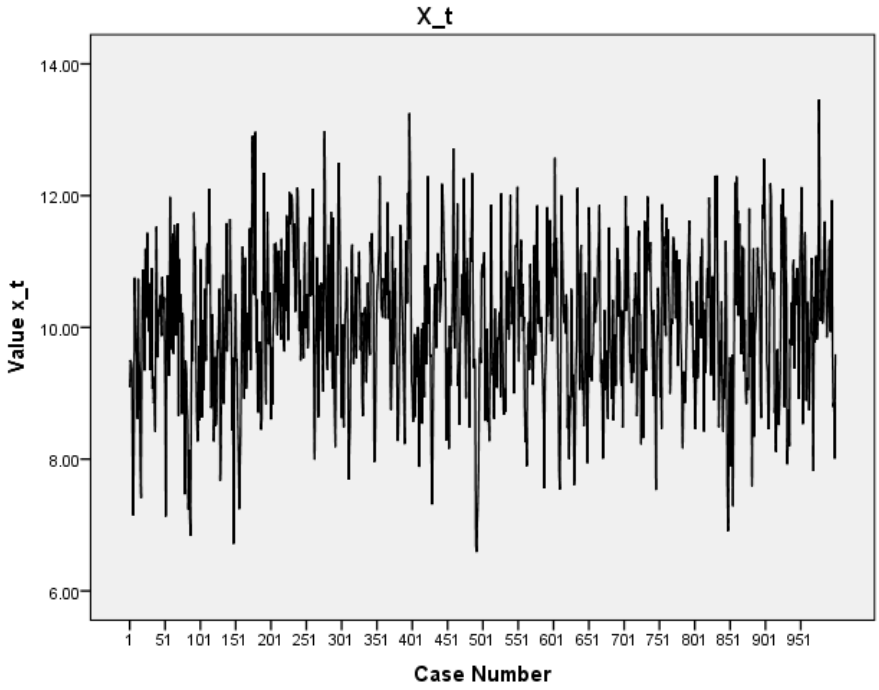
Case 16: Generate 1000 data points from the following (1) process:

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

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) of $ARMA$ 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.

Solution: We have generated the data from AR(1) process. We will plot the data:

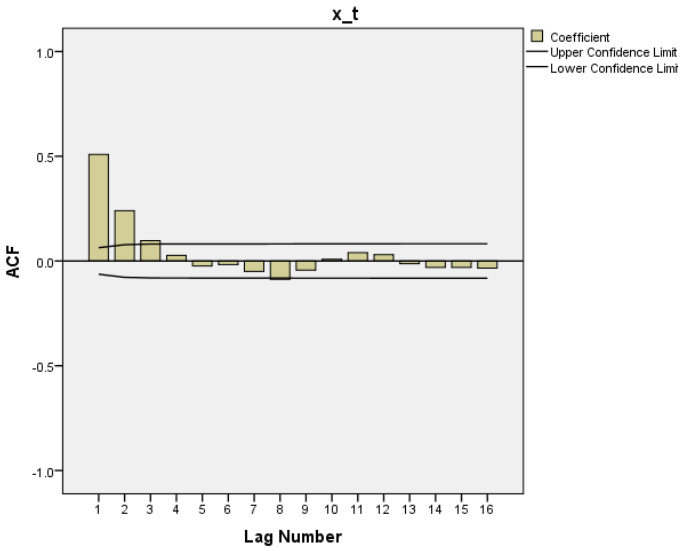


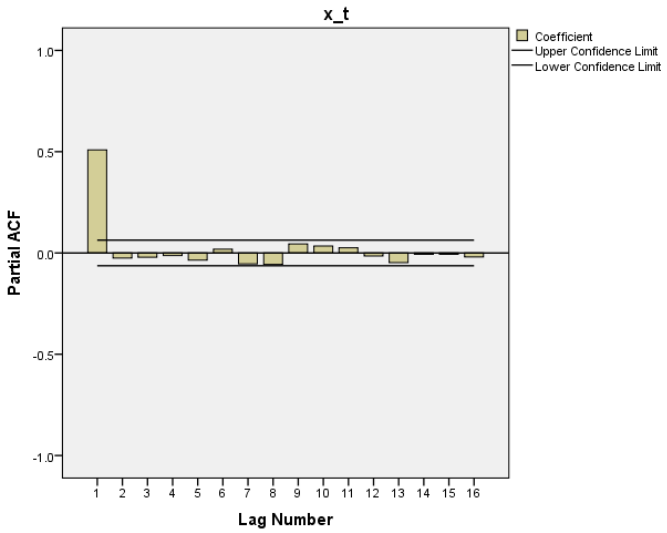
We will check stationarity of the data using the Augmented Dickey Fuller (ADF) Test:

```
Augmented Dickey-Fuller Test

data:  x[, 3]
Dickey-Fuller = -9.4532, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: We see p-value is < 0.05, hence we reject the null hypothesis and conclude that process is stationary.

We will identify the appropriate the order using sample Correlogram and sample Partial Correlogram:





Conclusion: In PACF plot we see a significant spike at lag 1 and in ACF plot we see spikes are tailing off, hence we conclude that process may be coming from **ARMA(1, 0)** model.

Now we will check the **goodness of the model:**

Model	Model Fit statistics				Ljung-Box Q(18)		
	Stationary R-squared	R-squared	RMSE	MAPE	Statistics	DF	Sig.
x_t-Model_1	.259	.259	.963	7.814	18.432	17	.362

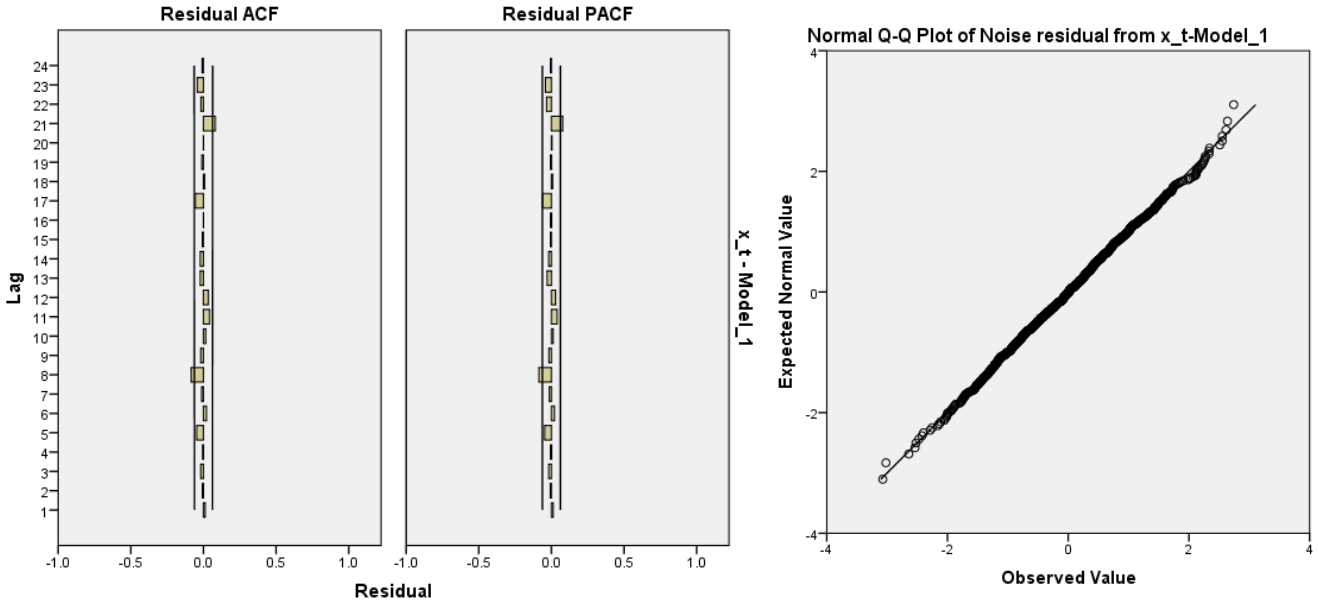
Conclusion: Here the Stationary R squared is 0.259. Since the Stationary R squared is positive hence this stationary model is better than the baseline model.

RMSE = 0.963 shows how much this stationary series deviate from the predicted series.

MAPE is 7.814, can be used to compare with other models.

Validating the assumption of driving **Gaussian White Noise**:

The p value of Ljung test is 0.362. It is greater than 0.05. Hence we fail to reject H_o at 5% level of significance. Hence the data are **independently distributed.**



The Q-Q plot is given above and we can see that there is no deviation from normality. The Residual noise is approximately normal.

We will assess the goodness of model built on simulated data by checking if the estimates are close to the parameters:

ARIMA Model Parameters

					Estimate	SE	t	Sig.
x_t-Model_1	x_t	No Transformation	Constant		10.033	.062	162.063	.000
			AR	Lag 1	.509	.027	18.677	.000

Here the model parameter i.e. $\hat{\varphi}$ = 0.509 which is very close to the original model parameter.

Also $\hat{\mu} = \frac{\delta}{1-\hat{\varphi}}$ => $10.033 = \frac{\delta}{1-0.509}$ => $\hat{\delta} = 4.926203$

which is close to 5. Hence the model parameter estimates are very close to the model parameter.

Forecast for the next 20 time points is:

Time Point	Forecast	Time Point	Forecast
1	9.81	11	10.03
2	9.92	12	10.03
3	9.97	13	10.03
4	10	14	10.03
5	10.02	15	10.03
6	10.03	16	10.03
7	10.03	17	10.03
8	10.03	18	10.03
9	10.03	19	10.03
10	10.03	20	10.03

Penalized log likelihood: Sometimes it is not possible to identify the order of the process using ACF or Partial ACF. So we consider another approach known as penalized log likelihood, where a penalty term is attached to the log of the likelihood function. There are various penalty terms defined by various statistician, of which one is Normalized BIC.

Normalized Bayesian Information Criterion(BIC) : A general measure of the overall fit of a model that attempts to account for model complexity. It is a score based upon the mean square error and includes a penalty for the number of parameters in the model and the length of the series. The penalty removes the advantage of models with more parameters, making the statistic easy to compare across different models for the same series.

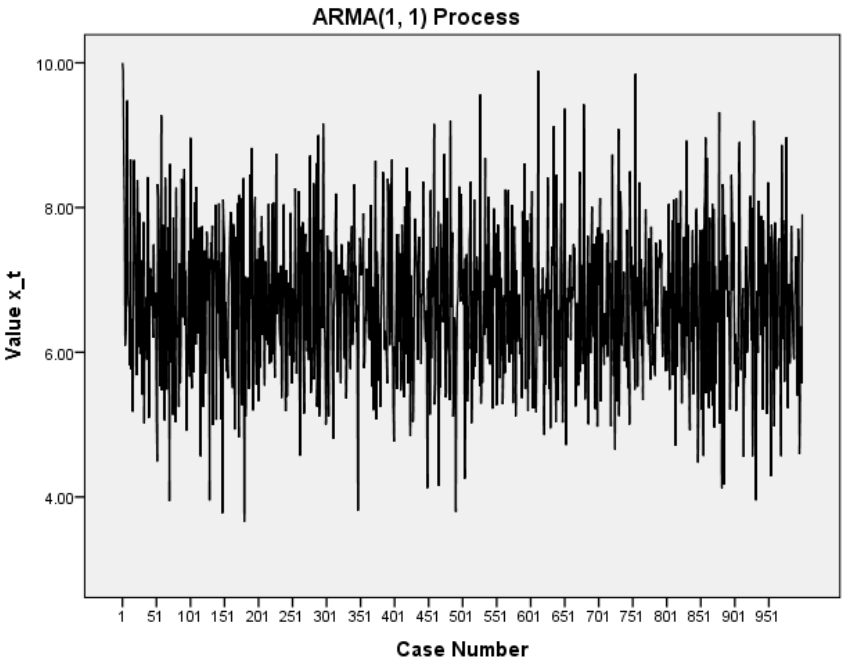
Case 17: Generate 1000 data points from the following $ARMA(1, 1)$ process:

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

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) of $ARMA$ 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.

Solution: We have generated the data from ARMA(1, 1) process. We will plot the data:



We will check stationarity of the data using the Augmented Dickey Fuller (ADF) Test:

```
Augmented Dickey-Fuller Test

data:  x[, 4]
Dickey-Fuller = -15.501, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Conclusion: We see p-value is < 0.05, hence we reject the null hypothesis and conclude that process is **stationary**.

We will find the appropriate order of the process by **Penalized log likelihood**, taking values of p = 0,1,2,3 and q = 0,1,2,3 we get 16 combinations and correspondingly obtain 16 normalized BIC.

The values are given as follows:

q/p	0	1	2	3
0	0.117	0.115	0.112	0.109
1	0.112	0.043	0.051	0.058
2	0.102	0.051	0.058	0.066
3	0.091	0.058	0.066	0.074

Clearly, ARIMA(1,0,1) has the least normalized BIC. So the process is ARMA(1, 1).

Now we will check the goodness of the model:

	Model Fit statistics			Ljung-Box Q(18)		
Model	Stationary R-squared	RMSE	MAPE	Statistics	DF	Sig.
x_t-Model_1	.086	1.011	12.383	32.038	16	.010

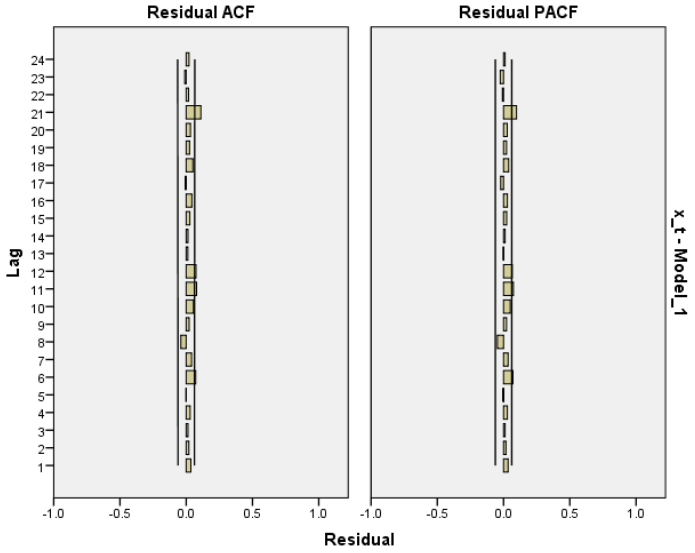
Conclusion: Here the Stationary R squared is 0.086. Since the Stationary R squared is positive hence this stationary model is better than the baseline model.

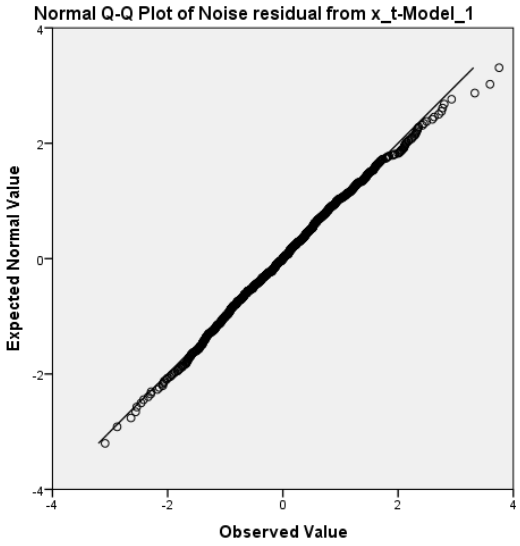
RMSE = 1.011 shows how much this stationary series deviate from the predicted series.

MAPE is 12.383, can be used to compare with other models.

Validating the assumption of driving **Gaussian White Noise**:

The p-value of Ljung test is 0.010 < 0.05. Hence we reject H_o and conclude that data are **not independently distributed**.





Conclusion: The Q-Q plot is given above and we can see that there is no deviation from normality. The Residual noise is approximately normal.

We will assess the **goodness of model** built on simulated data by checking if the estimates are close to the parameters:

ARIMA Model Parameters

					Estimate	SE	t	Sig.
x_t-Model_1	x_t	No Transformation	Constant		6.669	.002	3814.804	.000
			AR	Lag 1	.763	.024	32.128	.000
			MA	Lag 1	.988	.007	150.116	.000

Here the model parameter i.e. $\hat{\varphi}$ = 0.763 which is close to the original model parameter $\varphi = 0.7$.

Also $\hat{\mu} = \frac{\delta}{1-\hat{\varphi}}$ => $6.669 = \frac{\delta}{1-0.763}$ => $\hat{\delta} = 1.580553$

which is close to 2. Hence the model parameter estimates are close to the model parameter.

Forecast for the next 20 time points is:

Time Point	Forecast	Time Point	Forecast
1	6.99	11	6.69
2	6.91	12	6.68
3	6.86	13	6.68
4	6.81	14	6.68
5	6.78	15	6.68
6	6.75	16	6.67
7	6.73	17	6.67
8	6.72	18	6.67
9	6.71	19	6.67
10	6.7	20	6.67