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.)**

**2. TIME Series Analysis (17 Cases)**

* 1. **Estimation and Removal of Deterministic Components**

**2.1.1 Testing the Presence of Trend, it’s Estimation and Removal**

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.

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

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 %) and
5. Life Expectancy at birth (years)
   * + 1. 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.
     1. **Testing the Presence of Seasonality, it’s Estimation and Removal**
        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. 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.
        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.

* 1. **Modeling the Random Component using Auto Regressive Integrated Moving Averages (ARIMA)** 
     1. **Testing for Stationarity and making the Series Stationary if it is not**
        1. Testing Stationarity of the de-trended series for consumption expenditure (in million dollars) for the United States for 1944 to 2000.

* + - 1. Testing Stationarity of the de-trended series for the Following world development indicators for India:

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 %) and
5. Life Expectancy at birth (years)
   * + 1. 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. Testing Stationarity of the Estimated Random Component for monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India.
       3. Testing Stationarity of the Estimated Random Component for monthly World Airline Passengers from 1949-1960.
       4. 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.
       5. Testing Stationarity of the Estimated Random Component for a simulated AR(1) time Series
     1. **Identification of the Order of the ARIMA Model**
        1. Identifying the order of the ARIMA model for a simulated AR(1) time Series
        2. Identifying the order of the ARIMA model for a simulated MA(1) time Series

**2.2.3 Building ARIMA Model and Forecasting**

* + - 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  model.
* Check the goodness of the model by using the following:
  1. Stationary R-Square
  2. Root Mean Square Error (RMSE)
  3. Absolute Percentage Error (MAPE)
* Validate the assumption of driving Gaussian White Noise using the following:
  1. Ljung–Box Test for White Noise
  2. ACF and PACF for White Noise
  3. 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.
  + - 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  model.
* Check the goodness of the model by using the following:
  1. Stationary R-Square
  2. Root Mean Square Error (RMSE)
  3. Absolute Percentage Error (MAPE)
* Validate the assumption of driving Gaussian White Noise using the following:
  1. Ljung–Box Test for White Noise
  2. ACF and PACF for White Noise
  3. 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.

**Introduction**

A time series is a sequence of observations ordered by a time parameter. Time series may be measured continuously or discreetly and is obtained by measuring a single variable regularly over a period of time. A series showing the market share of a product might consist of weekly market share taken over a few years. A series of total sales figures might consist of one observation per month for many years. What each of these examples has in common is that some variable was observed at regular, known intervals over a certain length of time. Thus, the form of the data for a typical time series is a single sequence or list of observations representing measurements taken at regular intervals.

**Reasons for analyzing Time Series**

One reason to collect time series data is to try to discover systematic patterns in the series so a mathematical model can be built to explain the past behavior of the series. The discovery of a strong seasonal pattern, for example, might help explain large fluctuations in the data. Explaining a variable’s past behavior can be interesting and useful, but often one wants to do more than just evaluate the past. One of the most important reasons for doing time series analysis is to forecast future values of the series.

The parameters of the model that explained the past values may also predict whether and how much the next few values will increase or decrease. The ability to make such predictions successfully is obviously important to any business or scientific field.

Another reason for analyzing time series data is to evaluate the effect of some event that intervenes and changes the normal behavior of a series. Intervention analysis examines the pattern of a time series before and after the occurrence of such an event. The goal is to see if the outside event had a significant impact on the series pattern. If it did, there is a significant upward or downward shift in the values of the series after the occurrence of the event.

**A Model-Building Strategy**

The most popular strategy for building a model is the one developed by **Box and Jenkins** (1976), who defined three major stages of model building: identification, estimation, and diagnostic checking. Although Box and Jenkins originally demonstrated the usefulness of this strategy specifically for ARIMA model building, the general principles can be extended to all model building.

**Identification** involves selecting a tentative model type with which to work. This tentative model type includes initial judgements about the number and kind of parameters involved and how they are combined. In making these judgements, you should be parsimonious. The methods usually employed at this stage include plotting the series and its autocorrelation function to find out whether the series shows any upward or downward trend, whether some sort of data transformation might simplify analysis, and whether any kind of seasonal pattern is apparent.

**Estimation** is the process of fitting the tentative model to the data and estimating its parameters. This stage usually involves using a computerized model-fitting routine to estimate the parameters and test them for significance. The estimated parameters can then be used to see how well they would have predicted the observed values. If the parameter estimates are unsatisfactory on statistical grounds, you return to the identification stage, since the tentative model could not satisfactorily explain the behavior of the series.

**Diagnosis** is the stage in which you examine how well the tentative model fits the data. Methods used at this stage include plots and statistics describing the residual, or error, series. This information tells you whether the model can be used with confidence, or whether you should return to the first stage and try to identify a better model.

**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

**The first step in analyzing a time series is to plot it**. A plot gives you a general idea of how the series behaves:

• Does it have an overall trend (a persistent tendency to increase or decrease over time)?

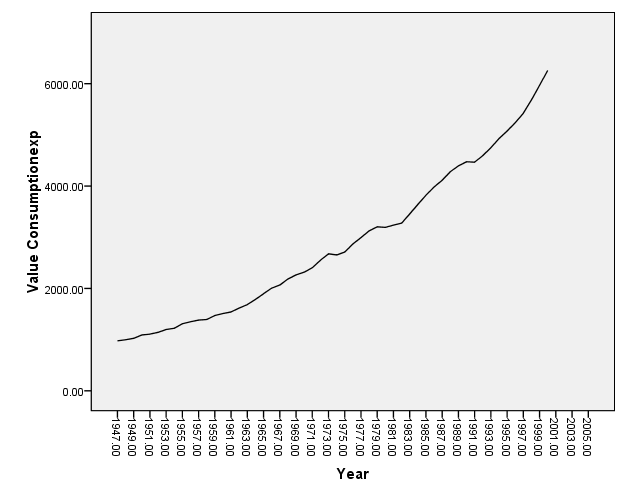
• Does it show seasonality (a cyclical pattern that repeats over and over, typically every year)? • Does it vary smoothly from one period to the next, or is it choppy?

• Is there a break or sudden change in the behavior of the series, or does it look much the same from beginning to end?

• Is the short-term variation about the same throughout the series? Does short-term variation increase or decrease with time? With the overall level of the series?

• Are there outliers? (Such points are often due to unique events, and must be excluded when you search for the process underlying the series as a whole.)

**Graph**



The time series plot produced by SPSS reveals the presence of an upward trend in the consumption expenditure for the United States for 1994-2000.

Now we check the presence of trend by using Relative ordering test in R. The R code and the result are given below.

>#Testing for Presence of Trend using Relative Ordering Test.

> #ro.test(y)

> # y is vector representing the observed time series with n observations.

> ro.test(data[,2])

**RELATIVE ORDERING**

Null Hypothesis: Absence of Trend, and

Alternative Hypothesis: Presence of Trend.

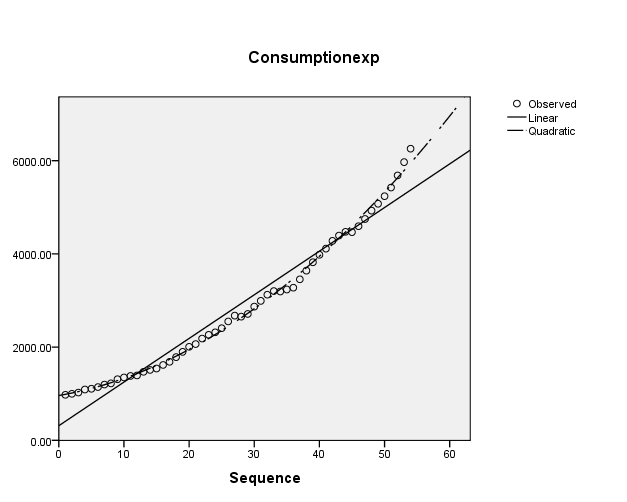
Test Statistic: 10.6311

p\_value: 0

No. of Discordants: 3

Expected No. of Discordants: 715.5

As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series.



* The curve fit chart gives you a quick visual assessment of the fit of each model to the observed values. From this plot, it appears that the Quadratic model better follows the shape of the data.
* In particular, the linear model seems to overestimate consumption expenditure for cases with small or large values of cases and underestimate consumption expenditure for cases in the middle.

The dependent and independent variables should be quantitative. If you select Time from the active dataset as the independent variable (instead of selecting a variable), the Curve Estimation procedure generates a time variable where the length of time between cases is uniform. If Time is selected, the dependent variable should be a time-series measure. Time-series analysis requires a data file structure in which each case (row) represents a set of observations at a different time and the length of time between cases is uniform

| Model Summary and Parameter Estimates | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dependent Variable:Consumptionexp | | | | | | | | |
| Equation | Model Summary | | | | | Parameter Estimates | | |
| 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 |

The model summary table reports the strength of the relationship between the model and the dependent variable.

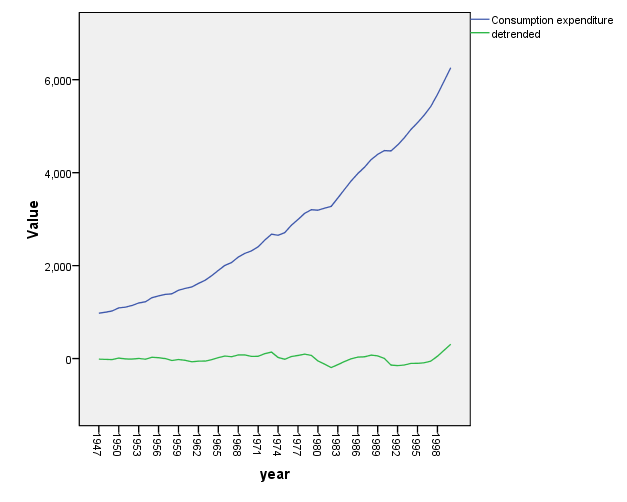
* The R Square statistic is a better measure of the strength of relationship. The R Square statistic is a measure of the strength of association between the observed and model-predicted values of the dependent variable.
* It shows that 96.3% of the variation in time is explained by the linear model. It shows that 99.7% of the variation in time is explained by the quadratic model.
* While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable, it does not directly address the strength of that relationship.

Using the Curve Estimation procedure, we have compared Linear and Quadratic models for the relationship between consumption expenditure and year. From the graph we see that the graph fits quadratic line well.

* The Quadratic model states that the expected consumption expenditure is equal to 962.181 + 24.180\*year – 1.262\*squared year. The positive value for b2 means that this model suggests that past a certain point, increased year would actually increase consumption.
* The significance value of the F statistic is less than 0.05 for quadratic model, which means that the variation explained by the model is not due to chance.

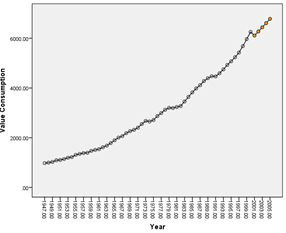
| **Case Summariesa** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Year | Consumption expenditure | Fit for Consumption from CURVEFIT, MOD\_1 QUADRATIC | Detrended |
| 1 | | 1947 | 976.4000 | 987.62333 | -11.22 |
| 2 | | 1948 | 998.1000 | 1015.59040 | -17.49 |
| 3 | | 1949 | 1025.3000 | 1046.08200 | -20.78 |
| 4 | | 1950 | 1090.9000 | 1079.09814 | 11.80 |
| 5 | | 1951 | 1107.1000 | 1114.63882 | -7.54 |
| 6 | | 1952 | 1142.4000 | 1152.70404 | -10.30 |
| 7 | | 1953 | 1197.2000 | 1193.29380 | 3.91 |
| 8 | | 1954 | 1221.9000 | 1236.40810 | -14.51 |
| 9 | | 1955 | 1310.4000 | 1282.04694 | 28.35 |
| 10 | | 1956 | 1348.8000 | 1330.21032 | 18.59 |
| 11 | | 1957 | 1381.8000 | 1380.89823 | .90 |
| 12 | | 1958 | 1393.0000 | 1434.11069 | -41.11 |
| 13 | | 1959 | 1470.7000 | 1489.84768 | -19.15 |
| 14 | | 1960 | 1510.8000 | 1548.10922 | -37.31 |
| 15 | | 1961 | 1541.2000 | 1608.89529 | -67.70 |
| 16 | | 1962 | 1617.3000 | 1672.20591 | -54.91 |
| 17 | | 1963 | 1684.0000 | 1738.04106 | -54.04 |
| 18 | | 1964 | 1784.8000 | 1806.40075 | -21.60 |
| 19 | | 1965 | 1897.6000 | 1877.28498 | 20.32 |
| 20 | | 1966 | 2006.1000 | 1950.69375 | 55.41 |
| 21 | | 1967 | 2066.2000 | 2026.62706 | 39.57 |
| 22 | | 1968 | 2184.2000 | 2105.08491 | 79.12 |
| 23 | | 1969 | 2264.8000 | 2186.06730 | 78.73 |
| 24 | | 1970 | 2317.5000 | 2269.57423 | 47.93 |
| 25 | | 1971 | 2405.2000 | 2355.60569 | 49.59 |
| 26 | | 1972 | 2550.5000 | 2444.16170 | 106.34 |
| 27 | | 1973 | 2675.9000 | 2535.24225 | 140.66 |
| 28 | | 1974 | 2653.7000 | 2628.84733 | 24.85 |
| 29 | | 1975 | 2710.9000 | 2724.97695 | -14.08 |
| 30 | | 1976 | 2868.9000 | 2823.63112 | 45.27 |
| 31 | | 1977 | 2992.1000 | 2924.80982 | 67.29 |
| 32 | | 1978 | 3124.7000 | 3028.51306 | 96.19 |
| 33 | | 1979 | 3203.2000 | 3134.74084 | 68.46 |
| 34 | | 1980 | 3193.0000 | 3243.49316 | -50.49 |
| 35 | | 1981 | 3236.0000 | 3354.77002 | -118.77 |
| 36 | | 1982 | 3275.5000 | 3468.57142 | -193.07 |
| 37 | | 1983 | 3454.3000 | 3584.89736 | -130.60 |
| 38 | | 1984 | 3640.6000 | 3703.74784 | -63.15 |
| 39 | | 1985 | 3820.9000 | 3825.12286 | -4.22 |
| 40 | | 1986 | 3981.2000 | 3949.02241 | 32.18 |
| 41 | | 1987 | 4113.4000 | 4075.44651 | 37.95 |
| 42 | | 1988 | 4279.5000 | 4204.39514 | 75.10 |
| 43 | | 1989 | 4393.7000 | 4335.86832 | 57.83 |
| 44 | | 1990 | 4474.5000 | 4469.86603 | 4.63 |
| 45 | | 1991 | 4466.6000 | 4606.38828 | -139.79 |
| 46 | | 1992 | 4594.5000 | 4745.43508 | -150.94 |
| 47 | | 1993 | 4748.9000 | 4887.00641 | -138.11 |
| 48 | | 1994 | 4928.1000 | 5031.10228 | -103.00 |
| 49 | | 1995 | 5075.6000 | 5177.72269 | -102.12 |
| 50 | | 1996 | 5237.5000 | 5326.86764 | -89.37 |
| 51 | | 1997 | 5423.9000 | 5478.53713 | -54.64 |
| 52 | | 1998 | 5683.7000 | 5632.73115 | 50.97 |
| 53 | | 1999 | 5968.4000 | 5789.44972 | 178.95 |
| 54 | | 2000  **2001**  **2002**  **2003**  **2004**  **2005** | 6257.8000  **6109.631**  **6273.893**  **6440.679**  **6609.989**  **6781.823** | 5948.69283 | 309.11 |
| Total | N | 54 | 54 | 54 | 54 |
| a. Limited to first 100 cases. | | | | | |

The simple trend based forecast for the consumption expenditure for the next 5 years has been highlighted above. The detrended consumption series is shown in last column of the above table.



* From the graph above we see the original curve (blue line with increasing trend) and detrended curve (horizontal green line).
* After removal of trend the graph shows a horizontal pattern. Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting

.



* The simple trend based forecast for the consumption expenditure for the next 5 years has been highlighted above.
* In this case we have plotted the quadratic model against the year and we see that the actual series remains bracketed by the quadratic model.

**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 %) and

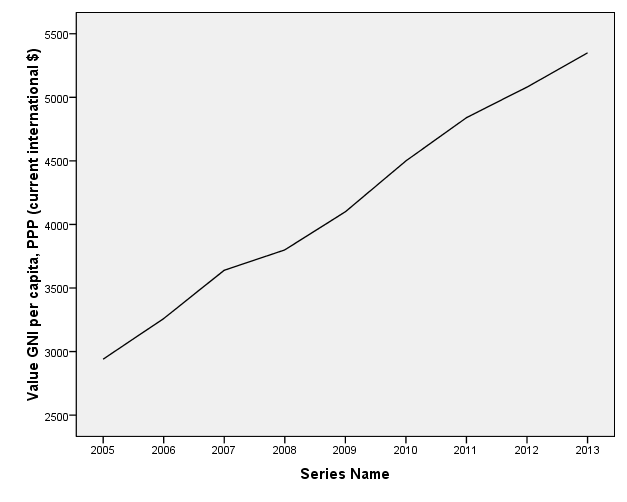
v. Life Expectancy at birth (years)

Obtain the de-trended indicators.

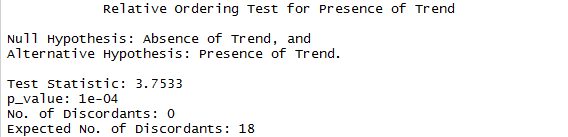
**We first plot the first variable** Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD,

**Graph**

* Regression may be used to test and model a trend. First one plots the series against time. If the trend appears linear, one can regress it against a measure of time.
* If one finds a significant and/or a substantial relationship with time, the magnitude of coefficient of time is evidence of a linear trend. Alternatively some trends may appear to be nonlinear.
* The time series plot produced by SPSS reveals the presence of an upward trend in the world development indicators for India: (Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD).



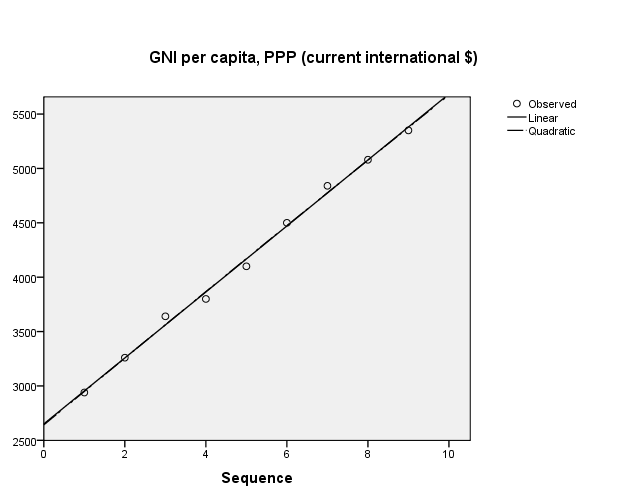
Now we check the presence of trend by using Relative ordering test in R. The R code and the result is as follows;



* As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series
* To test the existence of a statistically significant quadratic trend, a regression model was specified with both the linear and quadratic time component. For example time and time squared.
* The dependent variable is the GNI and the independent variables is a count of the number of years and the squared count of the number of years since the inception of the series. Then see which of the predictor variables is statistically significant.

The signs of the linear coefficients determine whether the curve is slopping downwards or upwards. The signs of the quadratic coefficients determine whether the function is curved upwards or downwards

**Curve Fit**



* The curve fit chart gives you a quick visual assessment of the fit of each model to the observed values. From this plot, both (linear and quadratic models) appear to follow the shape of the data.

| **Model Summary and Parameter Estimates** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Dependent Variable: GNI per capita, PPP (current international $) | | | | | | | | |
| Equation | Model Summary | | | | | Parameter Estimates | | |
| R Square | F | df1 | df2 | Sig. | Constant | b1 | b2 |
| Linear | .996 | 1815.251 | 1 | 7 | .000 | 2651.111 | 303.333 |  |
| Quadratic | .996 | 780.535 | 2 | 6 | .000 | 2642.381 | 308.095 | -.476 |

* The quadratic model is summarized on the second line, the line with quadratic in the equation column. The coefficients of the equation appear in the columns labeled b0 (the constant), b1 (the linear term). The best-fitting linear curve is given **by GNI=2651.111+ 303.333\* case**

Where case is the sequential case number.

* The best-fitting quadratic curve is given **by GNI=2642.381+ 308.095\* case-0.476\*case square**

Where case is the sequential case number

The quadratic term is quite small; this quadratic curve is almost a straight line.

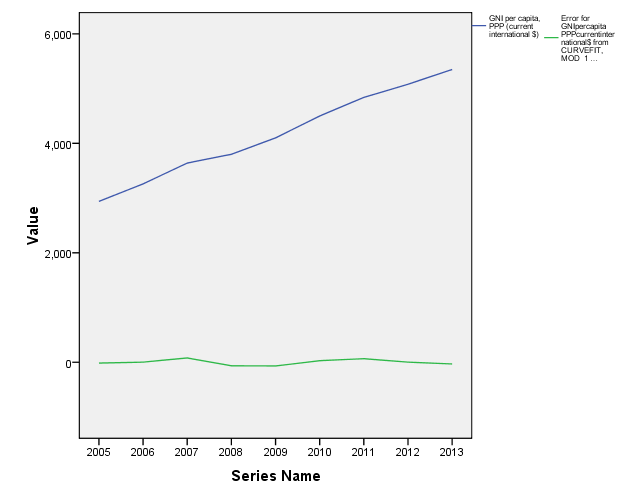
* The model summary table reports the strength of the relationship between the model and the dependent variable. The R Square statistic is a better measure of the strength of relationship. The R Square statistic is a measure of the strength of association between the observed and model-predicted values of the dependent variable.
* It shows that 99.6% of the variation in time is explained by the linear model. It shows that 99.6% of the variation in time is explained by the quadratic model.
* While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable, it does not directly address the strength of that relationship.

| **Quadratic Coefficients** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|  | B | Std. Error | Beta |
| Case Sequence | 308.095 | 34.744 | 1.014 | 8.868 | .000 |
| Case Sequence \*\* 2 | -.476 | 3.389 | -.016 | -.141 | **.893** |
| (Constant) | 2642.381 | 75.669 |  | 34.920 | .000 |

* From the graph we can’t conclude whether linear or quadratic fit is better. So we check the statistical significance of β2.Asits p value>0.05 we may accept the null hypothesis (that is Beta coefficient is not significant.)

| **Linear 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 |

From the above table we can conclude that the linear model fits the data adequately as its coefficients are statistically significant.



* (The upper line indicates original data (GNI) and lower green line shows the detrended line).Finally we choose a linear fit.
* A statistically significant linear trend curving upwards has been found to characterize Gross National Income (GNI) per capita based on Purchasing Power Parity (PPP) Exchange Rates (ER) measured in current USD,

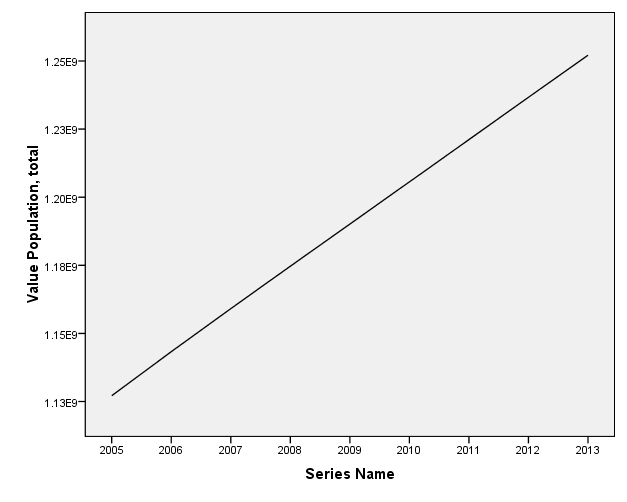
| **Case Summaries** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Series Name | GNI per capita, PPP (current international $) | Fit for GNIpercapitaPPPcurrentinternational$ from CURVEFIT, MOD\_1 LINEAR | Error for GNIpercapitaPPPcurrentinternational$ from CURVEFIT, MOD\_1 LINEAR  (detrended) |
| 1 | | 2005 | 2940 | 2954.44444 | -14.44444 |
| 2 | | 2006 | 3260 | 3257.77778 | 2.22222 |
| 3 | | 2007 | 3640 | 3561.11111 | 78.88889 |
| 4 | | 2008 | 3800 | 3864.44444 | -64.44444 |
| 5 | | 2009 | 4100 | 4167.77778 | -67.77778 |
| 6 | | 2010 | 4500 | 4471.11111 | 28.88889 |
| 7 | | 2011 | 4840 | 4774.44444 | 65.55556 |
| 8 | | 2012 | 5080 | 5077.77778 | 2.22222 |
| 9 | | 2013 | 5350 | 5381.11111 | -31.11111 |
| Total | N | 9 | 9 | 9 | 9 |
|  | | | | | |

The detrended indicator is shown in the last column.

Now we move to **the second variable population total.**

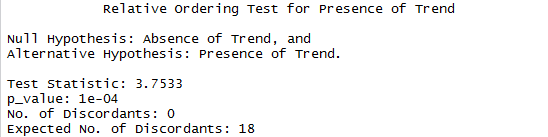
**Graph**

The time series plot produced below by SPSS reveals the presence of an upward linear trend in the world development indicators for India: (Population Total)



* Regression may be used to test and model a trend. First one plots the series against time. If the trend appears linear, one can regress it against a measure of time.
* If one finds a significant and/or a substantial relationship with time, the magnitude of coefficient of time is evidence of a linear trend. Alternatively some trends may appear to be nonlinear.

Now we check the presence of trend by using Relative ordering test in R. The R code and the result is as follows;



* As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series.
* To test the existence of a statistically significant linear trend, a regression model was specified with the linear time component.
* The dependent variable is the population total and the independent variables is a count of the number of years. The signs of the linear coefficients determine whether the curve is slopping downwards or upwards.

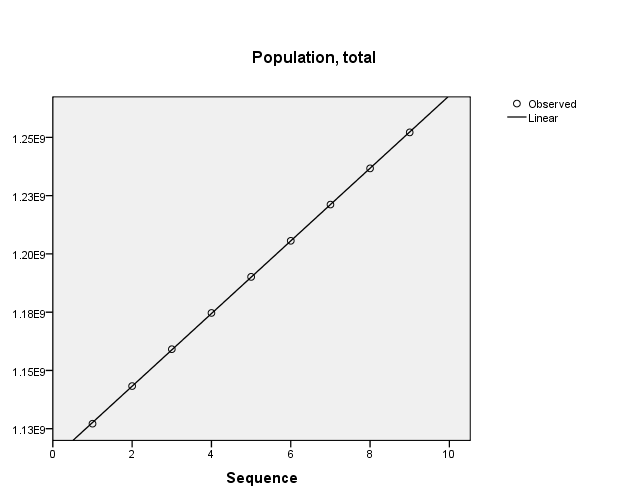
| **Model Summary and Parameter Estimates** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dependent Variable: Population, total | | | | | | | |
| Equation | Model Summary | | | | | Parameter Estimates | |
| R Square | F | df1 | df2 | Sig. | Constant | b1 |
| Linear | 1.000 | 219472.584 | 1 | 7 | .000 | 1.112E9 | 1.559E7 |

* The linear model is summarized above. The coefficients of the equation appear in the columns labeled b0 (the constant), b1 (the linear term). The best-fitting linear curve is given by

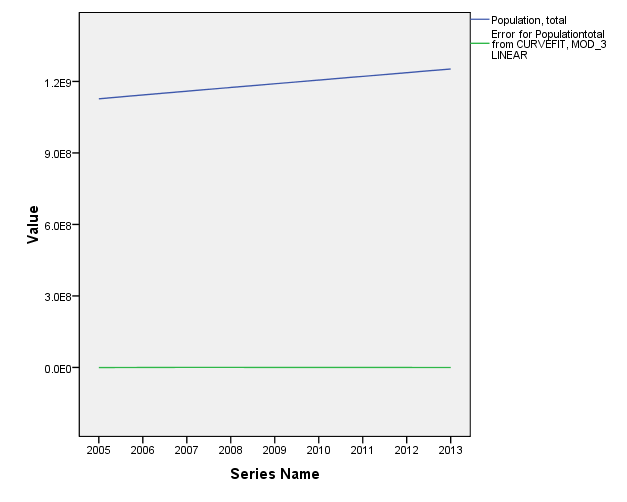
Population Total=1.112\*10^9 + (1.559\*10^7)\* case

Where case is the sequential case number.

* The model summary table reports the strength of the relationship between the model and the dependent variable.
* The R Square statistic is a better measure of the strength of relationship. The R Square statistic is a measure of the strength of association between the observed and model-predicted values of the dependent variable.
* It shows that 99.6% of the variation in time is explained by the linear model. It shows that 99.6% of the variation in time is explained by the quadratic model.
* While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable, it does not directly address the strength of that relationship.



* The curve fit chart gives you a quick visual assessment of the fit of each model to the observed values. From this plot, (linear model) appears to follow the shape of the data quite well.
* The upper line (blue line) represents population total and lower green line represents the detrended line. From the graph above we see the original curve and detrended curve.
* After removal of trend the line is parallel to x axis. Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting.

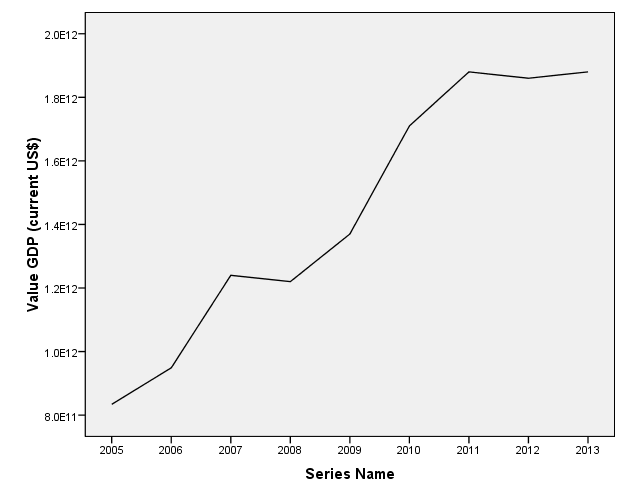


| **Case Summaries** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Series Name | Population, total | Fit for Populationtotal from CURVEFIT, MOD\_3 LINEAR | Error for Populationtotal from CURVEFIT, MOD\_3 LINEAR(**detrended**) |
| 1 | | 2005 | 1127143548 | 1.12764E9 | -4.98604E5 |
| 2 | | 2006 | 1143289350 | 1.14323E9 | 59517.72222 |
| 3 | | 2007 | 1159095250 | 1.15882E9 | 2.77738E5 |
| 4 | | 2008 | 1174662334 | 1.17441E9 | 2.57142E5 |
| 5 | | 2009 | 1190138069 | 1.18999E9 | 1.45197E5 |
| 6 | | 2010 | 1205624648 | 1.20558E9 | 44096.38889 |
| 7 | | 2011 | 1221156319 | 1.22117E9 | -11912.44444 |
| 8 | | 2012 | 1236686732 | 1.23676E9 | -69179.27778 |
| 9 | | 2013 | 1252139596 | 1.25234E9 | -2.03995E5 |
| Total | N | 9 | 9 | 9 | 9 |

Detrended indicator is shown in the last column.

Now we **proceed towards the 3rd variable**.

**Graph**

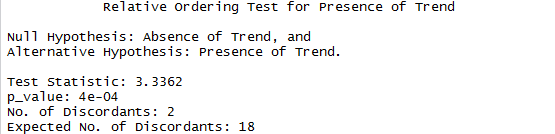


* Regression may be used to test and model a trend. First one plots the series against time. If the trend appears linear, one can regress it against a measure of time.
* If one finds a significant and/or a substantial relationship with time, the magnitude of coefficient of time is evidence of a linear trend. Alternatively some trends may appear to be nonlinear.
* The time series plot produced by SPSS reveals the presence of an upward linear trend in the world development indicators for India :( GDP (current)).

Now we check the presence of trend by using Relative ordering test in R. The R code and the result is as follows;

* As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series.

To test the existence of a statistically significant quadratic trend, a regression model was specified with both the linear and quadratic time component. For example time and time squared



The model summary table reports the strength of the relationship between the model and the dependent variable.

Adjusted R Square is a "corrected" R Square statistic that penalizes models with large numbers of parameters.

| **Model Summary** | | | |
| --- | --- | --- | --- |
|  | | | |
| R | R Square | Adjusted R Square | Std. Error of the Estimate |
| .970 | .940 | .931 | 1.071E11 |

* R Square, the coefficient of determination, is the squared value of the multiple correlation coefficient. It shows that 94% of the variation in time is explained by the model.
* R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value indicates a strong relationship.

| **ANOVA** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Regression | 1.258E24 | 1 | 1.258E24 | 109.569 | .000 |
| Residual | 8.035E22 | 7 | 1.148E22 |  |  |
| Total | 1.338E24 | 8 |  |  |  |

The ANOVA table tests the acceptability of the model from a statistical perspective. The Regression row displays information about the variation accounted for by your model. The Residual row displays information about the variation that is not accounted for by your model.

* The regression sum of squares is considerably larger than the residual sum of squares, which indicates that most of the variation in the proportion of GDP (current) is explained by the model. The significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance.

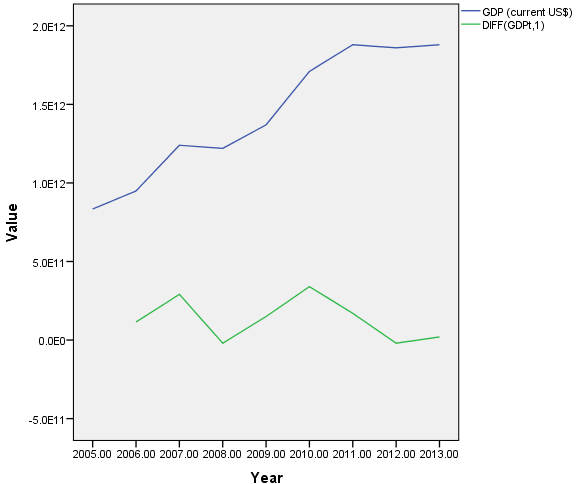
| **Coefficients** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|  | B | Std. Error | Beta |
| Case Sequence | 1.448E11 | 1.383E10 | .970 | 10.468 | .000 |
| (Constant) | 7.142E11 | 7.784E10 |  | 9.176 | .000 |

The coefficients of the equation appear in the columns labeled b0 (the constant), b1 (the linear term). The best-fitting linear curve is given by

**GDP (Current) = (7.142\*10^11)+ (1.448\*10^11)\* case**  Where case is the sequential case number



* The curve fit chart gives you a quick visual assessment of the fit of each model to the observed values. From this plot, (linear model) appears to follow the shape of the data quite well.

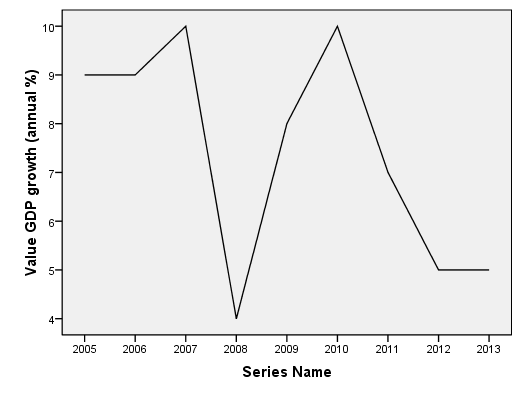
 The detrended estimator is given in the last column

* The upper line (blue line) represents GDP (current) and lower green line represents the detrended line. From the graph we see the original curve and detrended curve.
* After removal of trend the graph shows a cyclical pattern. Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting.

| **Case Summaries** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Series Name | GDP (current US$) | Fit for GDPcurrentUS$ from CURVEFIT, MOD\_6 LINEAR | Error for GDPcurrentUS$ from CURVEFIT, MOD\_6 LINEAR |
| 1 | | 2005 | 8.E11 | 8.58978E11 | -2.49778E10 |
| 2 | | 2006 | 9.E11 | 1.00376E12 | -5.47611E10 |
| 3 | | 2007 | 1.E12 | 1.14854E12 | 9.14556E10 |
| 4 | | 2008 | 1.E12 | 1.29333E12 | -7.33278E10 |
| 5 | | 2009 | 1.E12 | 1.43811E12 | -6.81111E10 |
| 6 | | 2010 | 2.E12 | 1.58289E12 | 1.27106E11 |
| 7 | | 2011 | 2.E12 | 1.72768E12 | 1.52322E11 |
| 8 | | 2012 | 2.E12 | 1.87246E12 | -1.24611E10 |
| 9 | | 2013 | 2.E12 | 2.01724E12 | -1.37244E11 |
| Total | N | 9 | 9 | 9 | 9 |

Now we proceed towards the 4 th variable

**Graph**

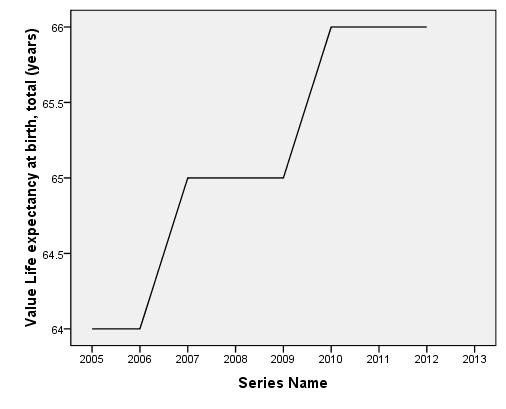


The time series plot produced by SPSS reveals the absence of trend in the world development indicators for India :( GDP (growth)).

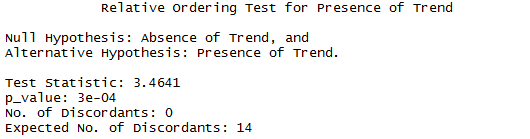
Trend has to be removed in order to use the time series for forecasting.

Now we proceed towards 5 th variable (Life Expectancy)

Missing data at the beginning or end of the series poses no particular problem. They simply shorten the useful length of the series.



We see a different kind of increasing trend in the given graph. It is constant for the year 2005-2006, then it increases by a single point between 2006-2007.It is again constant for a period of 2 years. Again it rises in the period 2009-2010 and is constant after that.



As p value <0.05 we reject our null hypothesis and conclude that there is presence of trend.

| **Model Summary** | | | |
| --- | --- | --- | --- |
| R | R Square | Adjusted R Square | Std. Error of the Estimate |
| .943 | .890 | .872 | .299 |

* The model summary table reports the strength of the relationship between the model and the dependent variable.
* Adjusted R Square is a "corrected" R Square statistic that penalizes models with large numbers of parameters
* R Square, the coefficient of determination, is the squared value of the multiple correlation coefficient. It shows that 89% of the variation in time is explained by the model. R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value indicates a strong relationship

| **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 |  |  |  |

* The ANOVA table tests the acceptability of the model from a statistical perspective. The Regression row displays information about the variation accounted for by your model. The Residual row displays information about the variation that is not accounted for by your model.
* The regression sum of squares is considerably larger than the residual sum of squares, which indicates that most of the variation in the proportion of life expectancy is explained by the model. The significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance.

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

* The coefficients of the equation appear in the columns labeled b0 (the constant), b1 (the linear term). The best-fitting linear curve is given by

Life Expectancy = 63.679+ 0.321\* case

Where case is the sequential case number

* We fitted a linear model because the coefficients are significant, the linear model also fits the data well.



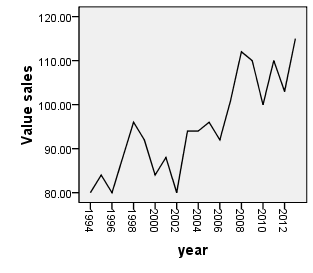
* From the graph we conclude that the green line is the original series and blue line is the detrended series.
* The detrended line is parallel to x-axis.
* Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting.

| **Case Summaries** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Series Name | Life expectancy at birth, total (years) | Fit for Lifeexpectancyatbirthtotalyears from CURVEFIT, MOD\_7 LINEAR | Error for Lifeexpectancyatbirthtotalyears from CURVEFIT, MOD\_7 LINEAR(**detrended**) |
| 1 | | 2005 | 64 | 64.00000 | .00000 |
| 2 | | 2006 | 64 | 64.32143 | -.32143 |
| 3 | | 2007 | 65 | 64.64286 | .35714 |
| 4 | | 2008 | 65 | 64.96429 | .03571 |
| 5 | | 2009 | 65 | 65.28571 | -.28571 |
| 6 | | 2010 | 66 | 65.60714 | .39286 |
| 7 | | 2011 | 66 | 65.92857 | .07143 |
| 8 | | 2012 | 66 | 66.25000 | -.25000 |

The detrended indicator is given in the last column.

**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.

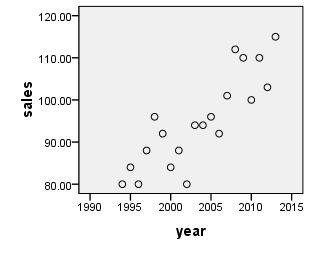
**Graph**

**Graph**

* Regression may be used to test and model a trend. First one plots the series against time. If the trend appears linear, one can regress it against a measure of time.

If one finds a significant and/or a substantial relationship with time, the magnitude of coefficient of time is evidence of a linear trend. Alternatively some trends may appear to be nonlinear.

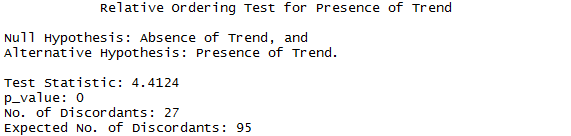
The time series plot produced by SPSS reveals the presence of an upward trend in the annual sales measured in million USD for a trading company for 1994-2013.



**Trend estimation** is a [statistical](https://en.wikipedia.org/wiki/Statistics) technique to aid interpretation of data. When a series of measurements of a process are treated as a [time series](https://en.wikipedia.org/wiki/Time_series), trend estimation can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred.

We can see an increasing trend pattern in the following graph.

Now we check for the presence of trend using Relative Ordering Test.



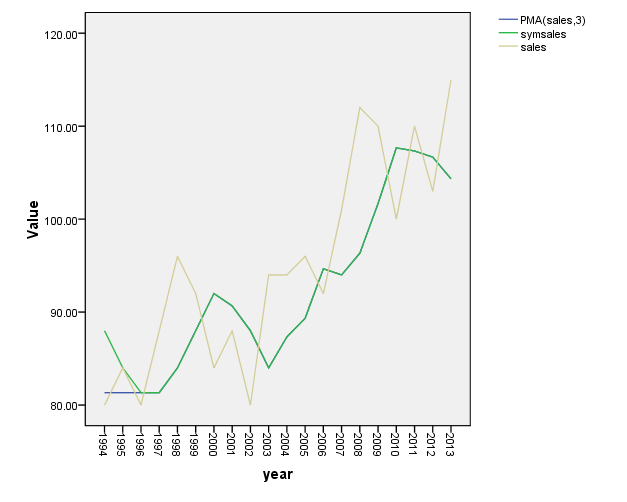
* As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series. A moving average is a technique to get an overall idea of the trends in a data set; it is an [average](http://www.statisticshowto.com/average/)of any subset of numbers. The moving average is extremely useful for **forecasting long-term trends**

Now we estimate trend by using prior moving average method.

| **Created Series** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | Case Number of Non-Missing Values | |  | |
|  | Series Name | First | Last | N of Valid Cases | Creating Function |
| 1 | sales\_1 | 4 | 21 | 18 | PMA(sales,3) |

PMA creates new series based on the prior moving averages of existing series. The prior moving average for each case in the original series is computed by averaging the values of a span of cases preceding it.

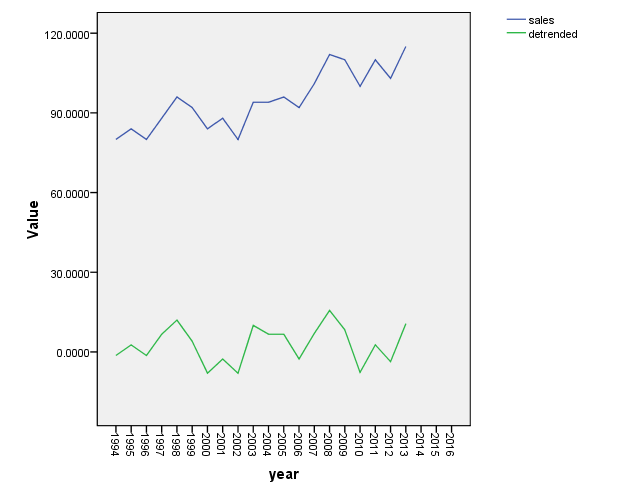
We **use prior moving average** as it facilitates the ultimate task of forecasting. The choice of padding for the missing moving average estimates for the initial time points is to be made as follows:



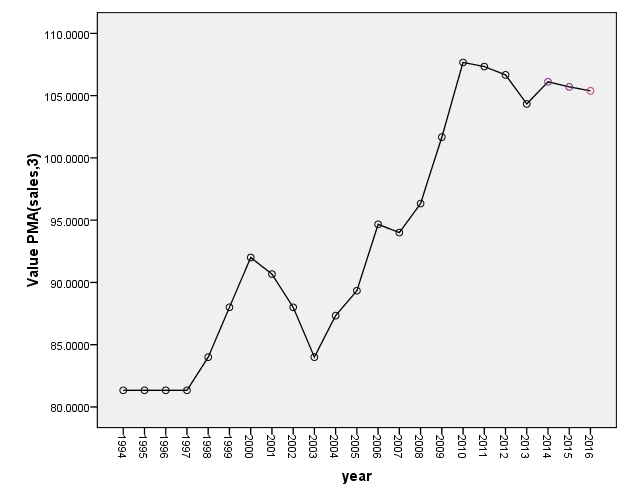
First of all we obtain the prior moving average. Then we pad the missing data by using both an endpoint padding scheme and a symmetric padding scheme

Make a plot of the original data and both the moving averages from the Step 2, and select the one that best approximates the trend in for the initial time points

* The yellow line indicates the original sales data .The green line indicates series obtained after symmetric padding.
* The blue line indicates series obtained after endpoint padding. As end point padding approximates the trend for the initial time points better than symmetric padding. So we choose end point padding.



* The upper line (blue line) represents sales and lower green line represents the detrended line.
* After removal of trend we observe a zigzag pattern which indicates that other deterministic components may be present in the model. Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting.



The simple trend based forecast is highlighted in the given table.

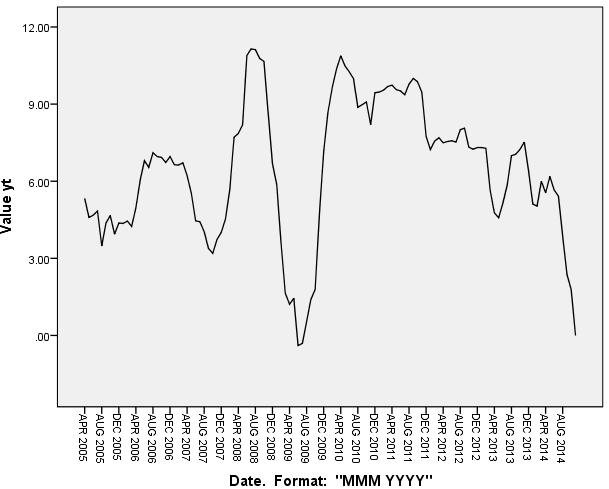
The last column gives us the series after removing trend.

We have used prior moving averages for the purpose of forecasting.

* The simple trend based forecast for the annual sales for the next 3 years has been highlighted.
* In this case we have plotted the model obtained after estimating trend using prior moving averages with end point padding. Against the year and we see that the actual series remains bracketed by the model.

| **Case Summariesa** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  |  | year | sales | PMA(sales,3) | detrended |
| 1 | | 1994 | 80.0000 | 81.3333 | -1.3333 |
| 2 | | 1995 | 84.0000 | 81.3333 | 2.6667 |
| 3 | | 1996 | 80.0000 | 81.3333 | -1.3333 |
| 4 | | 1997 | 88.0000 | 81.3333 | 6.6667 |
| 5 | | 1998 | 96.0000 | 84.0000 | 12.0000 |
| 6 | | 1999 | 92.0000 | 88.0000 | 4.0000 |
| 7 | | 2000 | 84.0000 | 92.0000 | -8.0000 |
| 8 | | 2001 | 88.0000 | 90.6667 | -2.6667 |
| 9 | | 2002 | 80.0000 | 88.0000 | -8.0000 |
| 10 | | 2003 | 94.0000 | 84.0000 | 10.0000 |
| 11 | | 2004 | 94.0000 | 87.3333 | 6.6667 |
| 12 | | 2005 | 96.0000 | 89.3333 | 6.6667 |
| 13 | | 2006 | 92.0000 | 94.6667 | -2.6667 |
| 14 | | 2007 | 101.0000 | 94.0000 | 7.0000 |
| 15 | | 2008 | 112.0000 | 96.3333 | 15.6667 |
| 16 | | 2009 | 110.0000 | 101.6667 | 8.3333 |
| 17 | | 2010 | 100.0000 | 107.6667 | -7.6667 |
| 18 | | 2011 | 110.0000 | 107.3333 | 2.6667 |
| 19 | | 2012 | 103.0000 | 106.6667 | -3.6667 |
| 20 | | 2013 | 115.0000 | 104.3333 | 10.6667 |
| 21 | | 2014 | . | **106.1100** | . |
| 22 | | 2015 | . | **105.7033** | . |
| 23 | | 2016 | . | **105.3811** | . |
| Total | N | 23 | 20 | 23 | 20 |
| a. Limited to first 100 cases. | | | | | |

**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 component based forecast for the monthly Whole Sale Price Index for the next 5 months



Regression may be used to test and model a trend. First one plots the series against time. If the trend appears linear, one can regress it against a measure of time.

If one finds a significant and/or a substantial relationship with time, the magnitude of coefficient of time is evidence of a linear trend. The time series plot produced by SPSS shows the presence of an upward trend along with a seasonal pattern in the WPI data.

The series shows a sharp dip at Aug 2009.Then it again rises fast and reaches the same level by April 2010.

As we cannot conclude from the figure so we proceed further and use Relative ordering test.

Relative Ordering Test for Presence of Trend

Null Hypothesis: Absence of Trend, and

Alternative Hypothesis: Presence of Trend.

Test Statistic: 0.4195

p\_value: 0.3374

No. of Discordants: 2810

Expected No. of Discordants: 2889

Freidman (JASA) Test for Presence of Seasonality

Null Hypothesis: Absence of Seasonality,

Alternative Hypothesis: Presence of Seasonality.

Test Statistic: 4.9744 (Chi Sqaure with 11 df)

p\_value: 0.9324

As p value>0.05 so we may accept null hypothesis and conclude that there is no trend.

As p value>0.05 so we may accept null hypothesis and conclude that there is no seasonality.

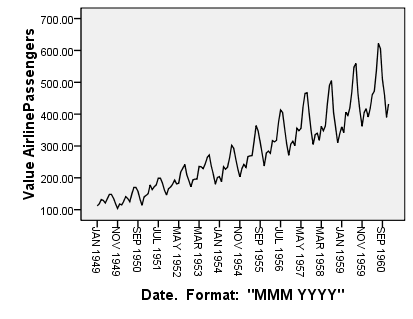
As there is no trend and no seasonality we do not proceed further.

**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 component based forecast for the monthly World Airline Passengers for the next 5 months.

In order to uncover any real trend in the data we first need to account for the variation due to seasonal effects. The Seasonal Decomposition procedure requires the presence of a periodic date component in the active dataset—for example, a yearly periodicity of 12 (months), a weekly periodicity of 7 (days), and so on. It's a good idea to plot your time series first, because viewing a time series plot often leads to a reasonable guess about the underlying periodicity (How many observations are there in a season of the series).We take the periodicity 12 from the date variable defined for the series. The Seasonal Decomposition procedure normally treats a series as a product of seasonal trend and cycle components.

**Trend estimation** is a [statistical](https://en.wikipedia.org/wiki/Statistics) technique to aid interpretation of data. When a series of measurements of a process are treated as a [time series](https://en.wikipedia.org/wiki/Time_series), trend estimation can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred.

We can see an increasing trend pattern in the following graph.



* This graph shows the total of international air passengers from 1949-1960.From the graph we observe an increasing trend pattern along with a seasonal effect.
* We notice that the series shows a marked seasonal pattern since travel is at its highest in the late summer months, while a secondary peak occurs in the spring.
* In general we say that a series exhibits periodic behavior with period s, when similarities in the series occur after s basic time intervals.
* In this graph the basic time interval is 1 month and the period is s 12 months. Many other series particularly sales data shows similar characteristics.

Now we check for the presence of trend using Relative Ordering Test.

As p value<0.05 we may reject null hypothesis and conclude that there is presence of trend in the given series. A moving average is a technique to get an overall idea of the trends in a data set; it is an [average](http://www.statisticshowto.com/average/)of any subset of numbers. The moving average is extremely useful for **forecasting long-term trends. We further estimate trend by using prior moving averages as it facilitates in the ultimate task of forecasting.**

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

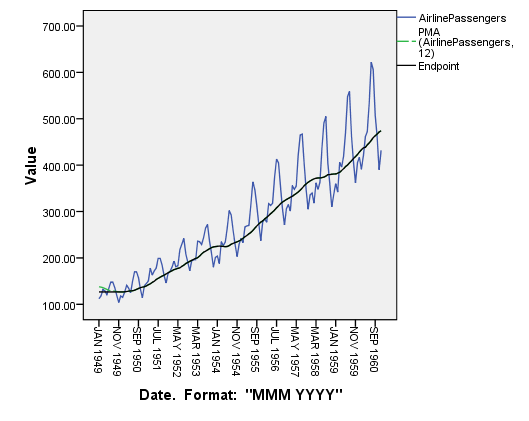
One of the deficiencies in the analysis of time series in the past has been a confusion between fitting a series and forecasting it.

Suppose a time series has shown a tendency to increase over a particular period and also to follow a seasonal pattern. A common method of analysis is to decompose a series arbitrarily into 3 components.

The Seasonal Decomposition procedure decomposes a series into a seasonal component, a combined trend and cycle component, and an "error" component.

* The Seasonal Decomposition procedure removes periodic fluctuations from time series, such as annual or seasonal highs or lows. It is used primarily as a preliminary tool when attempting to analyze trend in such series.

Referring to the airline data, the seasonal effect implies that an observation for a particular month say April is related to the observations for previous Aprils. The error components in these models would not in general be uncorrelated. For example the total of airline passengers in April 1960,while related to previous April totals, would also be related to totals in march 1960,Feb 1960,Jan 1960 and so on.



First of all we obtain the prior moving average. Then we pad the missing data by using both an endpoint padding scheme and a symmetric padding scheme.

Make a plot of the original data and both the moving averages from the Step 2, and select the one that best approximates the trend in for the initial time points.

* The blue line indicates the original Air passenger’s data .The green line indicates series obtained after symmetric padding.
* The black line indicates series obtained after endpoint padding. As end point padding approximates the trend for the initial time points better than symmetric padding. So we choose end point padding. For the purpose of forecasting also end point padding is preferred.
* The Seasonal Decomposition procedure creates four new variables for each of the original variables analyzed by the procedure. By default, the new variables are added to the active data set. The new series have names beginning with the following prefixes:
* SAF. Seasonal adjustment factors, representing seasonal variation. For the additive model, the value 0 represents the absence of seasonal variation.

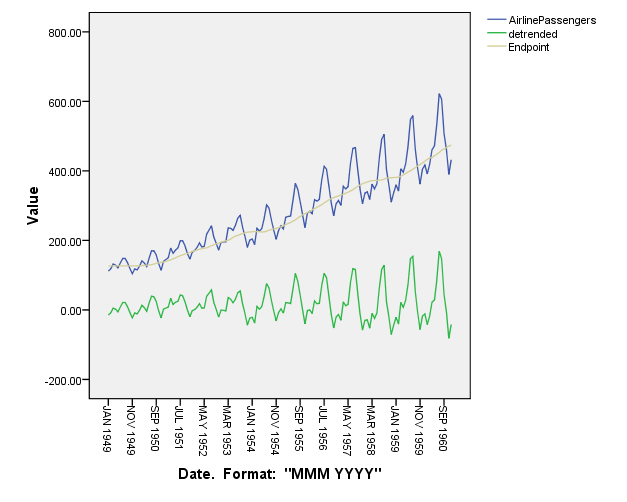
SAS. Seasonally adjusted series, representing the original series with seasonal variations removed. Working with a seasonally adjusted series, for example, allows a trend component to be isolated and analyzed independent of any seasonal component.

STC. Smoothed trend-cycle component, which is a smoothed version of the seasonally adjusted series that shows both trend and cyclic components.

ERR. The residual component of the series for a particular observation.

* The seasonal factor is the average deviation from the other components for each month of Airline passenger data.
* Period 1 averaged about 25 units below the deseasonalised data. And as we can see the periods 7 and 8 have the highest Air passenger level. While periods 11 and 2 have the lowest level.

|  | |
| --- | --- |
| Period | Seasonal Factor |
| 1 | -25.00335 |
| 2 | -36.33983 |
| 3 | -2.26597 |
| 4 | -7.79880 |
| 5 | -4.41686 |
| 6 | 35.55221 |
| **7** | **70.39412** |
| **8** | **61.83883** |
| 9 | 14.79653 |
| 10 | -22.32090 |
| 11 | -55.02102 |
| 12 | -29.41496 |



* The figure shows the forecast for the next 5 months. The additive model faithfully reproduces the seasonal pattern and supplies excellent forecasts.
* When changes occur in the seasonal pattern these will be appropriately projected into the forecast.
* Simple trend based forecast for the airline passenger data for the next 5 years has been highlighted by a different color.
* The graph illustrates growing variation in annual international Airline data. The upper blue line indicates the original Air Passengers data with trend and seasonality.
* The lower green line indicates the detrended data.
* After removal of trend we observe a zigzag pattern which indicates that other deterministic components that is seasonal variation may be present in the model. Trends have to be considered for extracting, fitting and forecasting otherwise model becomes unamenable for forecasting.

Thus a deterministic component based forecast for the monthly World Airline Passengers for the next 5 months is as follows:

The columns given below are yt(forecast series),mt(forecasted trend),st(forecasted seasonality),ct(forecasted cyclicity) respectively.

1 JAN 1961 430.42 473.92 -25.00335 -18.50051

2 FEB 1961 419.08 473.92 -36.33983 -18.50051

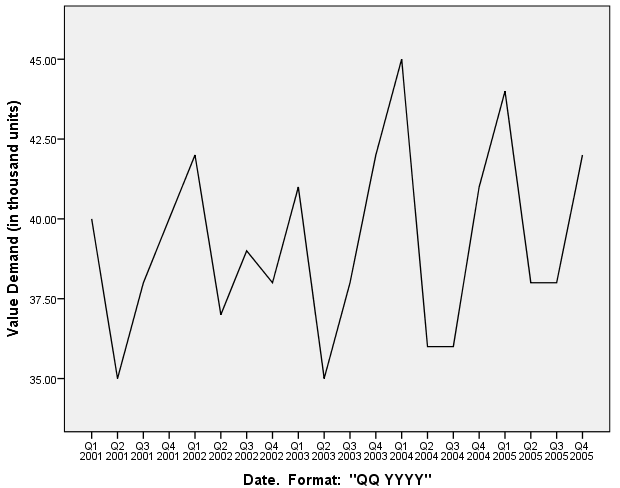
3 MAR 1961 453.15 473.92 -2.26597 -18.50051

4 APR 1961 447.62 473.92 -7.79880 -18.50051

5 MAY 1961 451.00 473.92 -4.41686 -18.50051

**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.

First we all we plot the series.



This series is a series of quarterly demand for an industrial good measured in thousand units for a manufacturing company for 2001-2005.

Repeated wave like patterns that span periods longer than 1 year are called cycles.So hence in this graph we see cyclical patterns.

friedman.test(data[,3],4)

Freidman (JASA) Test for Presence of Seasonality

Null Hypothesis: Absence of Seasonality, and

Alternative Hypothesis: Presence of Seasonality.

Test Statistic: 7.48 (Chi Square with 3 df)

p\_value: 0.0581

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

We see that p value>0.05 we may accept null hypothesis and conclude the absence of trend.

As p value(JASA test) >0.05 we may accept null hypothesis and conclude the absence of seasonality.

**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.

Augmented Dickey-Fuller Test

data: data[, 5]

Dickey-Fuller = -2.2273, Lag order = 3, p-value = 0.483

alternative hypothesis: stationary

An **Augmented Dickey–Fuller test (ADF)** is a test for a [unit root](https://en.wikipedia.org/wiki/Unit_root) in a [time series](https://en.wikipedia.org/wiki/Time_series) [sample](https://en.wikipedia.org/wiki/Sample_(statistics)). It is an augmented version of the [Dickey–Fuller test](https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_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.[[](https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test#cite_note-1)

As P value>0.05 we may accept null hypothesis and conclude that process is not stationary .This is for the first 2 boxes where p values are greater than 0.05.

.

Augmented Dickey-Fuller Test

data: data[, 6]

Dickey-Fuller = -2.788, Lag order = 3, p-value = 0.2571

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: data[, 7]

Dickey-Fuller = -5.3562, Lag order = 3, p-value = 0.01

alternative hypothesis: stationary

|  |  |  |  |
| --- | --- | --- | --- |
| year | detrended | diff1 | diff 2 |
| 1947 | -11.2233 | 0 | 0 |
| 1948 | -17.4904 | -6.26707 | 0 |
| 1949 | -20.782 | -3.2916 | 2.97547 |
| 1950 | 11.80186 | 32.58386 | 35.87546 |
| 1951 | -7.53882 | -19.3407 | -51.9245 |
| 1952 | -10.304 | -2.76522 | 16.57546 |
| 1953 | 3.9062 | 14.21024 | 16.97546 |
| 1954 | -14.5081 | -18.4143 | -32.6245 |
| 1955 | 28.35306 | 42.86116 | 61.27546 |
| 1956 | 18.58968 | -9.76338 | -52.6245 |
| 1957 | 0.90177 | -17.6879 | -7.92453 |
| 1958 | -41.1107 | -42.0125 | -24.3246 |
| 1959 | -19.1477 | 21.96301 | 63.97547 |
| 1960 | -37.3092 | -18.1615 | -40.1246 |
| 1961 | -67.6953 | -30.3861 | -12.2245 |
| 1962 | -54.9059 | 12.78938 | 43.17545 |
| 1963 | -54.0411 | 0.86485 | -11.9245 |
| 1964 | -21.6008 | 32.44031 | 31.57546 |
| 1965 | 20.31502 | 41.91577 | 9.47546 |
| 1966 | 55.40625 | 35.09123 | -6.82454 |
| 1967 | 39.57294 | -15.8333 | -50.9245 |
| 1968 | 79.11509 | 39.54215 | 55.37546 |
| 1969 | 78.7327 | -0.38239 | -39.9245 |
| 1970 | 47.92577 | -30.8069 | -30.4245 |
| 1971 | 49.59431 | 1.66854 | 32.47547 |
| 1972 | 106.3383 | 56.74399 | 55.07545 |
| 1973 | 140.6578 | 34.31945 | -22.4245 |
| 1974 | 24.85267 | -115.805 | -150.125 |
| 1975 | -14.077 | -38.9296 | 76.87546 |
| 1976 | 45.26888 | 59.34583 | 98.27545 |
| 1977 | 67.29018 | 22.0213 | -37.3245 |
| 1978 | 96.18694 | 28.89676 | 6.87546 |
| 1979 | 68.45916 | -27.7278 | -56.6245 |
| 1980 | -50.4932 | -118.952 | -91.2245 |
| 1981 | -118.77 | -68.2769 | 50.67546 |
| 1982 | -193.071 | -74.3014 | -6.02454 |
| 1983 | -130.597 | 62.47406 | 136.7755 |
| 1984 | -63.1478 | 67.44952 | 4.97546 |
| 1985 | -4.22286 | 58.92498 | -8.52454 |
| 1986 | 32.17759 | 36.40045 | -22.5245 |
| 1987 | 37.95349 | 5.7759 | -30.6246 |
| 1988 | 75.10486 | 37.15137 | 31.37547 |
| 1989 | 57.83168 | -17.2732 | -54.4246 |
| 1990 | 4.63397 | -53.1977 | -35.9245 |
| 1991 | -139.788 | -144.422 | -91.2245 |
| 1992 | -150.935 | -11.1468 | 133.2755 |
| 1993 | -138.106 | 12.82867 | 23.97547 |
| 1994 | -103.002 | 35.10413 | 22.27546 |
| 1995 | -102.123 | 0.87959 | -34.2245 |
| 1996 | -89.3676 | 12.75505 | 11.87546 |
| 1997 | -54.6371 | 34.73051 | 21.97546 |
| 1998 | 50.96885 | 105.606 | 70.87547 |
| 1999 | 178.9503 | 127.9814 | 22.37545 |
| 2000 | 309.1072 | 130.1569 | 2.17546 |

As P value<0.05 we may reject null hypothesis and conclude that process is stationary.

The probability of x having unit root and therefore being non-stationary is 0.01, the test tells me that there is a very high probability that x is stationary

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

The columns obtained after differencing

Are shown in the last 2 columns.

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.

GNI

Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -26.7218, Lag order = 2, p-value = 0.01

alternative hypothesis: stationary

.

As P value<0.05 we may reject null hypothesis and conclude that process is stationary.

As P value>0.05 we may accept null hypothesis and conclude that process is not stationary

Population Total

Augmented Dickey-Fuller Test

data: data[, 2]

Dickey-Fuller = 0.0677, Lag order = 2, p-value = 0.99

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -15.8565, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

As P value<0.05 we may reject null hypothesis and conclude that process is stationary.

As P value>0.05 we may accept null hypothesis and conclude that process is not stationary

As P value>0.05 we may accept null hypothesis and conclude that process is not stationary

As P value<0.05 we may reject null hypothesis and conclude that process is stationary

As P value<0.05 we may reject null hypothesis and conclude that process is stationary

(GDPC) Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -3.2855, Lag order = 2, p-value = 0.09368

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -2.0611, Lag order = 1, p-value = 0.5491

alternative hypothesis: stationary

(GDPC) Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -9.0811, Lag order = 1, p-value = 0.01

alternative hypothesis: stationary

Lifeexp Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -215139.5, Lag order = 1, p-value = 0.01

alternative hypothesis: 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.

As P value<0.05 we may reject null hypothesis and conclude that process is stationary.

Augmented Dickey-Fuller Test

data: data[, 3]

Dickey-Fuller = -3.8635, Lag order = 2, p-value = 0.03118

alternative hypothesis: 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.

Since there is no trend and no seasonality (as proved in case 4 using standard tests) so additional decomposition of WPI could not be done. As a result we do not have the estimated random component for monthly Wholesale Price Index (WPI) – Inflation, Base year 2004-05 for India. Hence we cannot proceed further.

**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

AIRLINE PASSENGERS

Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -7.0281, Lag order = 5, p-value = 0.01

alternative hypothesis: stationary

As P value<0.05 we may reject null hypothesis and conclude that process 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.

Since there is no trend and no seasonality (as proved in case 6 using standard tests) so additional decomposition of data could not be done. As a result we do not have the estimated random component for the quarterly demand for an industrial good measured in thousand units for a manufacturing company for 2001-2005 . Hence we cannot proceed further.

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



where 𝜖𝑡 is a Gaussian 𝑊𝑁(0,1) and 𝑋0=10.

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

As P value>0.05 we may accept null hypothesis and conclude that process is not stationary.

As P value<0.05 we may reject null hypothesis and conclude that process is stationary

Augmented Dickey-Fuller Test

Dickey-Fuller = -2.0115, Lag order = 9, p-value = 0.5734

alternative hypothesis: stationary

Augmented Dickey-Fuller Test

Dickey-Fuller = -10.9091, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

The **Ljung–Box test** (named for [Greta M. Ljung](https://en.wikipedia.org/w/index.php?title=Greta_M._Ljung&action=edit&redlink=1) and [George E. P. Box](https://en.wikipedia.org/wiki/George_E._P._Box)) is a type of statistical of whether any of a group of [autocorrelations](https://en.wikipedia.org/wiki/Autocorrelation) of a [time series](https://en.wikipedia.org/wiki/Time_series) are different from zero. Insteadoftesting [randomness](https://en.wikipedia.org/wiki/Randomness) at each distinct lag, it tests the "overall" randomness based on a number of lags and is therefore a [portmanteau test](https://en.wikipedia.org/wiki/Portmanteau_test) .

**NOTE Portmanteau statistic might inflate the autocorrelation under conditions of short lag times or short series for which reason the modified Ljung box statistiuc to provide better significant tests**.

The Ljung–Box test may be defined as:

**H0:** 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).

**Ha:** The data are not independently distributed; they exhibit serial correlation.

The test statistic is:[[2]](https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test#cite_note-LB-2)

Q = n\left(n+2\right)\sum_{k=1}^h\frac{\hat{\rho}^2_k}{n-k}

where *n* is the sample size, \hat{\rho}_k is the sample autocorrelation at lag *k*, and *h* is the number of lags being tested. Under H_0 the statistic Q follows a \chi^2_{(h)}. For [significance level](https://en.wikipedia.org/wiki/Significance_level) α, the[critical region](https://en.wikipedia.org/wiki/Critical_region) for rejection of the hypothesis of randomness is

Q > \chi_{1-\alpha,h}^2

where \chi_{1-\alpha,h}^2 is the α-[quantile](https://en.wikipedia.org/wiki/Quantile) of the [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) with *h* degrees of freedom.

The Ljung–Box test is commonly used in [autoregressive integrated moving average](https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average)(ARIMA) modeling. Note that it is applied to the [residuals](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) 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. When testing the residuals of an estimated ARIMA model, the degrees of freedom need to be adjusted to reflect the parameter estimation. For example, for an ARIMA(p,0,q) model, the degrees of freedom should be set to h - p - q.[[3]](https://en.wikipedia.org/wiki/Ljung%E2%80%93Box_test#cite_note-3)

**Case 14:** Generate 1000 data points from the following MA(1) process: 𝑋𝑡=𝜖𝑡−𝜖𝑡−1

where 𝜖𝑡 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.

Once the series has been rendered stationary,the ACF and PACF are examined to determine the type and order of the model.If the ACF of the AR process is positive then the PACF spikes will be positive and vice versa.

MA processes have short term finite memories.Consequently the ACF of an MA(1) process usually has only 1 significant spike ,whereas the PACF of the MA models generally exhibits gradual attenuation.An MA process exhibits a different characteristic ACF pattern. The characteristic pattern consists of sharp spikes up to and including the lag, indicating the order of the MA (q) process under consideration. Hence the moving average is shown to spike at the lag of its order and then dropped to 0.

Consequently, the memory of the moving average process is finite and limited to the order of its process.

| **Created Series** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | Series Name | Case Number of Non-Missing Values | | N of Valid Cases | Creating Function |
|  | First | Last |
| 1 | random\_1 | 2 | 1001 | 1000 | LAGS(random,1) |

We first create a lagged variable for the residual series generated by spss.

| **Autocorrelations** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| Lag | Autocorrelation | Std. Errora | Box-Ljung Statistic | | |
| Value | df | Sig.b |
| 1 | -.487 | .032 | 237.577 | 1 | .000 |
| 2 | -.006 | .038 | 237.612 | 2 | .000 |
| 3 | -.013 | .038 | 237.782 | 3 | .000 |
| 4 | .027 | .038 | 238.531 | 4 | .000 |
| 5 | -.056 | .038 | 241.734 | 5 | .000 |
| 6 | .052 | .039 | 244.442 | 6 | .000 |
| 7 | .021 | .039 | 244.885 | 7 | .000 |
| 8 | -.071 | .039 | 249.998 | 8 | .000 |
| 9 | .015 | .039 | 250.227 | 9 | .000 |
| 10 | .005 | .039 | 250.257 | 10 | .000 |
| 11 | .017 | .039 | 250.538 | 11 | .000 |
| 12 | .025 | .039 | 251.191 | 12 | .000 |
| 13 | -.029 | .039 | 252.048 | 13 | .000 |
| 14 | -.007 | .039 | 252.103 | 14 | .000 |
| 15 | .002 | .039 | 252.108 | 15 | .000 |
| 16 | .033 | .039 | 253.213 | 16 | .000 |
| **a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used.** | | | | | |
| **b. Based on the asymptotic chi-square approximation.** | | | | | |

* The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.
* **Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation
* The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.
* **Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation

| Partial Autocorrelations | | |
| --- | --- | --- |
| Lag | Partial Autocorrelation | Std. Error |
| 1 | -.487 | .032 |
| 2 | -.319 | .032 |
| 3 | -.251 | .032 |
| 4 | -.167 | .032 |
| 5 | -.202 | .032 |
| 6 | -.137 | .032 |
| 7 | -.057 | .032 |
| 8 | -.122 | .032 |
| 9 | -.135 | .032 |
| 10 | -.142 | .032 |
| 11 | -.116 | .032 |
| 12 | -.048 | .032 |
| 13 | -.059 | .032 |
| 14 | -.067 | .032 |
| 15 | -.065 | .032 |
| 16 | -.012 | .032 |

**Sample Autocorrelation Function** When we analyze the ARIMA process we find several functions that are of considerable analytical utility. The Autocorrelation function defines the autoregressive process as the expectation of the current observation times that of the previous observation

The PACF is not useful in identifying the order of the MA process as it is in identifying the order of the AR process.For the first order MA model,the PACF gradually attenuates(diminishes) as time passes.If the shock is positive then the PACF will be negative in sign and will be exponential in decline of size.



* An MA process exhibits a different characteristic ACF pattern. The characteristic pattern consists of sharp spikes up to and including the lag, indicating the order of the MA (q) process under consideration. Hence the moving average is shown to spike at the lag of its order and then dropped to 0.

The ACF will indicate the order of the model.There will be as many significant spikes as the model order.

The ACF of an MA(1) process usually has only 1 significant spike ,whereas the PACF of the MA models generally exhibits gradual attenuation.

**Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model ,after observing the graph (ACF)**

* **we conclude that as there is 1 negative spike at lag 1**
* **Exponential decay of negative spikes,it is an ARMA(0,1) or an MA(1) model.**

**Case 15:** Generate 1000 data points from the following (1) process: 𝑋𝑡=0.9𝑋𝑡−1+𝜖𝑡

where 𝜖𝑡 is a Gaussian 𝑊𝑁(0,1) and 𝑋0=10.

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

In general the AR process is identified by the characteristic patterns of its ACf and PACF .The ACf has a gradual attenuation and the PACF possesses the same number of spikes as the order of the model.The strength of the autocorrelation of the stationary Autoregressive process exponentially diminishes over time as long as the magnitude of the autoregressive parameter remains less than 1.

With this exponential attenuation the decline in magnitude approaches 0 as the time lag becomes infinite. This exponential decline in magnitude of the parameter forms the characteristic pattern of the ACF for the AR Process.

| **Autocorrelations** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Series:xt15 | | | | | |
| Lag | Autocorrelation | Std. Errora | Box-Ljung Statistic | | |
| Value | df | Sig.b |
| 1 | .881 | .032 | 778.017 | 1 | .000 |
| 2 | .766 | .051 | 1367.658 | 2 | .000 |
| 3 | .662 | .061 | 1807.979 | 3 | .000 |
| 4 | .570 | .068 | 2134.343 | 4 | .000 |
| 5 | .489 | .072 | 2374.868 | 5 | .000 |
| 6 | .428 | .076 | 2559.672 | 6 | .000 |
| 7 | .368 | .078 | 2696.581 | 7 | .000 |
| 8 | .312 | .080 | 2794.976 | 8 | .000 |
| 9 | .280 | .081 | 2874.100 | 9 | .000 |
| 10 | .256 | .082 | 2940.334 | 10 | .000 |
| 11 | .232 | .083 | 2994.749 | 11 | .000 |
| 12 | .203 | .083 | 3036.615 | 12 | .000 |
| 13 | .166 | .084 | 3064.602 | 13 | .000 |
| 14 | .138 | .084 | 3083.893 | 14 | .000 |
| 15 | .117 | .084 | 3097.897 | 15 | .000 |
| 16 | .104 | .085 | 3108.926 | 16 | .000 |
| a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used. | | | | | |
| b. Based on the asymptotic chi-square approximation. | | | | | |



* The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.
* **Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation

The characteristic pattern of an AR process is seen to be one of gradual attenuation of the magnitude of autocorrelation.

The attenuation begins after first lag.

We cant identify the order of the process.

AR(1) has an exponentially declining function as the lag k increases.

* The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.
* **Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation.

| **Partial Autocorrelations** | | |
| --- | --- | --- |
| Lag | Partial Autocorrelation | Std. Error |
| 1 | .881 | .032 |
| 2 | -.042 | .032 |
| 3 | -.020 | .032 |
| 4 | -.007 | .032 |
| 5 | -.005 | .032 |
| 6 | .040 | .032 |
| 7 | -.034 | .032 |
| 8 | -.020 | .032 |
| 9 | .072 | .032 |
| 10 | .018 | .032 |
| 11 | -.010 | .032 |
| 12 | -.032 | .032 |
| 13 | -.052 | .032 |
| 14 | .026 | .032 |
| 15 | .013 | .032 |
| 16 | .012 | .032 |

The PACF used in conjunction with the ACF can be used to distinguish a first order from a higher order AR process.



* For an AR process the PACF exhibits diminishing spikes through the lag of the process after which those spikes disappear. We know that for an AR(1) model there will be 1 spike in the PACF
* . Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model ,from the figure we infer that the autocorrelation is positive, as the PACF is exhibiting a positive spike at lag 1.Because the model is only that of an AR(1) process there will be no partial spikes at higher lags.

Therefore the PACF very clearly indicates the order of the AR process.

Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model, So hence we conclude that it is an ARMA(1,0) process after looking at ACF plot and PACF plot.(The ACF plot shows exponential decay and the presence of positive spikes while the PACF shows 1 positive spike at lag 1.)

The **root-mean-square deviation (RMSD)** or **root-mean-square error (RMSE)** is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. The RMSD represents the [sample standard deviation](https://en.wikipedia.org/wiki/Sample_standard_deviation) of the differences between predicted values and observed values. These individual differences are called [residuals](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) when the calculations are performed over the data sample that was used for estimation, and are called *prediction errors* when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision), but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent

The RMSD of an [estimator](https://en.wikipedia.org/wiki/Estimator) \hat{\theta} with respect to an estimated parameter \theta is defined as the square root of the [mean square error](https://en.wikipedia.org/wiki/Mean_square_error):

\operatorname{RMSD}(\hat{\theta}) = \sqrt{\operatorname{MSE}(\hat{\theta})} = \sqrt{\operatorname{E}((\hat{\theta}-\theta)^2)}.

For an [unbiased estimator](https://en.wikipedia.org/wiki/Unbiased_estimator), the RMSD is the square root of the variance, known as the[standard error](https://en.wikipedia.org/wiki/Standard_error_(statistics)).

The RMSD of predicted values \hat y_t for times *t* of a [regression's](https://en.wikipedia.org/wiki/Regression_analysis) [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) y_t is computed for *n* different predictions as the square root of the mean of the squares of the deviations:

\operatorname{RMSD}=\sqrt{\frac{\sum_{t=1}^n (\hat y_t - y_t)^2}{n}}.

The **mean absolute percentage error** (**MAPE**), also known as **mean absolute percentage deviation** (**MAPD**), is a measure of prediction accuracy of a forecasting method in [statistics](https://en.wikipedia.org/wiki/Statistics), for example in [trend estimation](https://en.wikipedia.org/wiki/Trend_estimation). It usually expresses accuracy as a percentage, and is defined by the formula:

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

where *At* is the actual value and *Ft* is the forecast value.

The difference between *At* and *Ft* is divided by the Actual value *At* again. The absolute value in this calculation is summed for every forecasted point in time and divided 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: 𝑋𝑡=5+0.50𝑋𝑡−1+𝜖𝑡

where 𝜖𝑡 is a Gaussian 𝑊𝑁(0,1) and 𝑋0=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.

An **Augmented Dickey–Fuller test (ADF)** is a test for a [unit root](https://en.wikipedia.org/wiki/Unit_root) in a [time series](https://en.wikipedia.org/wiki/Time_series) [sample](https://en.wikipedia.org/wiki/Sample_(statistics)). It is an augmented version of the [Dickey–Fuller test](https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_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.[[](https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test#cite_note-1)

Augmented Dickey-Fuller Test

data: data[, 3]

Dickey-Fuller = -9.4532, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

As P value<0.05 we may reject null hypothesis and conclude that process is stationary.

The probability of x having unit root and therefore being non-stationary is 0.01, the test tells me that there is a very high probability that x is stationary.

2. Fit an ‘appropriate’ order (identify it using sample Correlograms and sample Partial Correlograms) of 𝐴𝑅𝑀𝐴 model

A more complex situation as compared to AR and MA process is that of the MIXED AR and MA process. The implications of the ACF in ARMA process are interesting. These models may have ACF s that taper off. The magnitude of the ACF is modulated by the order of the MA parameter in the denominator.

Although the magnitude and relative magnitude of ACF are important the standard error and confidence interval are essential for proper inference.

The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.

**Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation.

| **Autocorrelations** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Lag | Autocorrelation | Std. Errora | Box-Ljung Statistic | | |
| Value | df | Sig.b |
| 1 | .509 | .032 | 259.757 | 1 | .000 |
| 2 | .240 | .039 | 317.791 | 2 | .000 |
| 3 | .097 | .040 | 327.256 | 3 | .000 |
| 4 | .027 | .041 | 327.983 | 4 | .000 |
| 5 | -.024 | .041 | 328.547 | 5 | .000 |
| 6 | -.017 | .041 | 328.844 | 6 | .000 |
| 7 | -.050 | .041 | 331.332 | 7 | .000 |
| 8 | -.088 | .041 | 339.083 | 8 | .000 |
| 9 | -.043 | .041 | 340.977 | 9 | .000 |
| 10 | .008 | .041 | 341.046 | 10 | .000 |
| 11 | .040 | .041 | 342.670 | 11 | .000 |
| 12 | .030 | .041 | 343.603 | 12 | .000 |
| 13 | -.013 | .041 | 343.772 | 13 | .000 |
| 14 | -.031 | .041 | 344.721 | 14 | .000 |
| 15 | -.030 | .041 | 345.668 | 15 | .000 |
| 16 | -.034 | .041 | 346.839 | 16 | .000 |
| a. The underlying process assumed is MA with the order equal to the lag number minus one. The Bartlett approximation is used. | | | | | |
| b. Based on the asymptotic chi-square approximation.  Nonstationary series have an ACF that remains significant for half a dozen or more lags, rather than quickly declining to zero. We must difference such a series until it is stationary before you can identify the process. | | | | | |



* Autoregressive processes have an exponentially declining ACF and spikes in the first one or more lags of the PACF. The number of spikes indicates the order of the auto regression.
* Moving average processes have spikes in the first one or more lags of the ACF and an exponentially declining PACF. The number of spikes indicates the order of the moving average.

Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model, So hence we conclude that it is an ARMA(1,0) process after looking at ACF plot and PACF plot.(The ACF plot shows exponential decay and the presence of positive spikes while the PACF shows 1 positive spike at lag 1.)

.



For an AR process the PACF exhibits diminishing spikes through the lag of the process after which those spikes disappear.We know that for an AR(1) model there will be 1 spike in the PACF . Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model ,from the figure we infer that the autocorrelation is positive,as the PACF is exhibiting a positive spike at lag 1.Because the model is only that of an AR(1) process there will be no partial spikes at higher lags.

Therefore the PACF very clearly indicates the order of the AR process.

| **Partial Autocorrelations** | | |
| --- | --- | --- |
| Lag | Partial Autocorrelation | Std. Error |
| 1 | .509 | .032 |
| 2 | -.025 | .032 |
| 3 | -.021 | .032 |
| 4 | -.013 | .032 |
| 5 | -.035 | .032 |
| 6 | .019 | .032 |
| 7 | -.054 | .032 |
| 8 | -.057 | .032 |
| 9 | .045 | .032 |
| 10 | .034 | .032 |
| 11 | .025 | .032 |
| 12 | -.015 | .032 |
| 13 | -.048 | .032 |
| 14 | -.006 | .032 |
| 15 | -.007 | .032 |
| 16 | -.020 | .032 |

* The significance of the autocorrelation coefficient is determined by the box-Ljung statistic.

**Since the p value<0.05 we may reject null hypothesis and conclude that the**  data are not independently distributed; they exhibit serial correlation

After estimation of the models the Box Jenkins model building strategy entails a diagnosis of the adequacy of the model. This entails being sure that the model converged upon a minimum sum of squared errors. Although there is no single absolute level above which the model is unacceptable, the smaller the measure of error, the better the model fits the data.

The Model Fit table provides fit statistics calculated across all of the models. It provides a concise summary of how well the models, with re-estimated parameters, fit the data. For each statistic, the table provides the mean, standard error (SE), minimum, and maximum value across all models.

While a number of statistics are reported, we will focus on two: MAPE (mean absolute percentage error) and RMSE.

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)

**Model Summary** **(ARIMA(1,0,0))**

| **Model Fit** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Fit Statistic | Mean | SE | Minimum | Maximum | Percentile | | | | | | |
| 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| Stationary R-squared | .259 | . | .259 | .259 | .259 | .259 | .259 | .259 | .259 | .259 | .259 |
| R-squared | .259 | . | .259 | .259 | .259 | .259 | .259 | .259 | .259 | .259 | .259 |
| RMSE | .963 | . | .963 | .963 | .963 | .963 | .963 | .963 | .963 | .963 | .963 |
| MAPE | 7.814 | . | 7.814 | 7.814 | 7.814 | 7.814 | 7.814 | 7.814 | 7.814 | 7.814 | **7.814** |

* The model statistics table provides summary information and goodness-of-fit statistics for each estimated model Absolute percentage error is a measure of how much a dependent series varies from its model-predicted level. By examining the mean and maximum across all models, you can get an indication of the uncertainty in your predictions.
* RMSE serves as a good basis of comparison of different models .It is a common criterion of lack of fit.
* The sum of squared residuals should be quite small.so that the R2 of the model will be quite large It also contains percentile values that provide information on the distribution of the statistic across models. For each percentile, that percentage of models have a value of the fit statistic below the stated value
* For instance 95% of the models have a value of MAPE that is less than 7.814.Absolute percentage error is a measure of how much a dependent series varies from its model-predicted level and provides an indication of the uncertainty in your predictions The mean absolute percentage error has a fixed value of 7.814% across all models.

4. Validate the assumption of driving Gaussian White Noise using the following:

a. Ljung–Box Test for White Noise

It provides a concise summary of how well the models fit the data

| **Model Statistics** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Number of Predictors | Model Fit statistics | | | Ljung-Box Q(18) | | | Number of Outliers |
| Stationary R-squared | RMSE | MAPE | Statistics | DF | Sig. |
| Xt-Model\_1 | 0 | .259 | .963 | 7.814 | 18.432 | 17 | .362 | 0 |

* The Ljung-Box statistic, also known as the modified Box-Pierce statistic, provides an indication of whether the model is correctly specified. A significance value > 0.05 **And conclude that** 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).Hence we conclude that there are no outliers.
* Larger values of stationary R-squared (up to a maximum value of 1) indicate better fit. A value of 0.259 means that the model poorly explains the observed variation in the series. This statistic provides an estimate of proportion of the total variation in the series that is explained by the model and is preferable to ordinary R2 when there is a trend or seasonal pattern.

b. ACF and PACF for White Noise

Model diagnosis entails residual analysis as well. If the model is properly specified and the model parameters account for all of the systematic variance, then the residuals should resemble white noise.

Residual analysis is performed with the ACF and PACF. These correlograms can be examined with reference to modified portmanteau tests with their associated significance levels.

White noise residuals do not have significant p values. These White noise p values of the residuals should not be< 0.05.

Graphically white noise residuals have associated spikes that do not extend beyond the confidence interval limits. The ACF and PACF plots reveal these limits as dotted lines spreading out from the midpoint of the plot. Indication of significant ACF or PACF residual spikes is empirical evidence of lack of fit



As spikes do not protrude beyond the limit of 2 S.E.”s on each side of the central vertical axis of no autocorrelation. Then the autocorrelation or partial autocorrelation of the residuals do not have significant spikes with p value<0.05.

So there is no indication of lack of fit.

5. 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. |
| Xt-Model\_1 | Xt | No Transformation | Constant | | 10.033 | .062 | 162.063 | **.000** |
| AR | Lag 1 | .509 | .027 | 18.677 | **.000** |

This table displays values of all the parameters in the model, with an entry for each estimated model labelled by the model identifier.

* Hence we have listed all of the variables in the model (dependent and independent variables) that were significant.
* As p values are < 0.05 we conclude that coefficients are significant.
* As the standard errors are low so it indicates a good model fit.

Note:When we fit an 𝐴𝑅𝐼𝑀𝐴 model with a constant term the estimated constant is not the intercept term included in the model rather it’s the estimated process mean. We have already seen that for an 𝐴𝑅𝑀𝐴(𝑝,𝑞) model, i.e. 𝑋𝑡 = 𝛿 + 𝜙1𝑋𝑡−1 + 𝜙2𝑋𝑡−2 + ⋯+ 𝜙𝑝𝑋𝑡−𝑝 + 𝜖𝑡 + 𝜃1𝜖1 + ⋯+ 𝜃𝑞𝜖𝑞

where 𝜙𝑝 ≠ 0,𝜃𝑞 ≠ 0, and {𝜖𝑡} is 𝑊𝑁(0,𝜎2) ∋ 𝐶𝑜𝑣(𝜖𝑡,𝑋𝑡−𝑗) = 0 ∀ 𝑗 > 0. We have,

𝐸𝑋𝑡 = 𝛿 + 𝜙1𝐸𝑋𝑡−1 + 𝜙2𝐸𝑋𝑡−2 + ⋯+ 𝜙𝑝𝐸𝑋𝑡−𝑝 + 0 = 𝜇

⇒ 𝛿 + 𝜙1𝜇 + 𝜙2𝜇 + ⋯+ 𝜙𝑝𝜇 = 𝜇

⇒ (1 − 𝜙1 − ⋯− 𝜙𝑝)𝜇 = 𝛿

Hence the estimated intercept is to be obtained by solving the following:

(1 − 𝜙 ̂1 − ⋯− 𝜙 ̂𝑝)𝜇̂ = 𝛿 ̂

Consider the case where we have an 𝐴𝑅(1) process 𝑋𝑡 = 5 + 0.5𝑋𝑡−1 + 𝜖𝑡 and the estimated constant (mean) will be somewhere around 10 and the estimated 𝜙 will be around 0.5 and hence the estimated intercept will be

**(1 − 0.5) × 10 = 5**

Note that the logic remains same for 𝐴𝑅 process as 𝑀𝐴 part doesn’t contribute to the process mean. This issue will not arise for a proper 𝑀𝐴 process as there mean is same as the intercept.

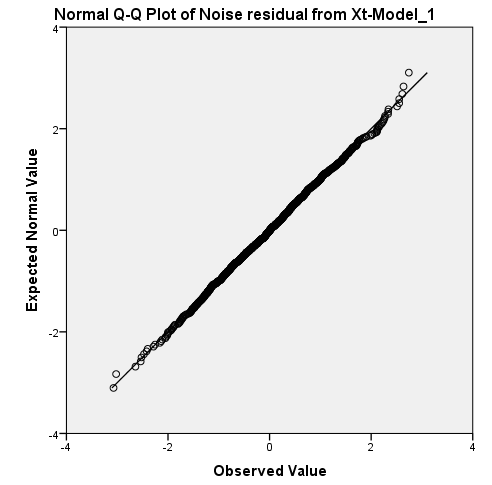
intercept will be =(10.033 \* (1- .509))=4.9262. So the fitted AR model is 𝑋𝑡 = 4.9262 + 0.509𝑋𝑡−1

QQ plot plots the quantiles of a variable distribution against the quantiles of any number of test distributions.

These plots are generally used to determine whether the distribution of a variable matches a given distribution (normal in this case)

If the data really are from a normal distribution the plot should approximate a straight line.

Hence through the given figure we can validate the assumption of the white noise being Gaussian.



We consider the following hypothesis:

Ho:Data is normally distributed.

H1: Data is not normally distributed

As p value is >0.05 we may accept null hypothesis and conclude that Data is normally distributed

|  | Kolmogorov-Smirnova | | |
| --- | --- | --- | --- |
|  | Statistic | df | Sig. |
| Noise residual from Xt-Model\_1 | .015 | 1000 | .200\* |

.



* Superimposed on the plot of the original series we have the ARIMA predictions.
* The graph accurately capture the amount of variation that we should expect from the given model.

Predicted values are represented by the middle green line.

Our forecasted values are plotted in the graph below. Red line at the centre shows the Observed values.

As we cannot see the forecasted blue line which is merging fully with the observed values so it indicates a perfect fit.

6. Apply the model and forecast for next 20 time points

9.81 9.92 9.97 10.00 10.02 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03 10.03

**Case 17:** Generate 1000 data points from the following (1,1) process: 𝑋𝑡=2+0.7𝑋𝑡−1+𝜖𝑡−𝜖𝑡−1

where 𝜖𝑡 is a Gaussian 𝑊𝑁(0,1) and 𝑋0=6.2. Assuming the generated data as a sample from some 𝐴𝑅𝑀𝐴 model do the following:

The probability of x having unit root and therefore being non-stationary is 0.01, the test tells me that there is a very high probability that x is stationary.

Augmented Dickey-Fuller Test

data: data[, 1]

Dickey-Fuller = -15.5304, Lag order = 9, p-value = 0.01

alternative hypothesis: stationary

BIC

**USE OF MODEL SELECTION CRITERIA –BIC**

In the implementation of this approach a, range of potential ARIMA models are estimated by maximum likelihood methods and for each a criterion such as BIC (or AIC) is evaluated. In the information Criteria Approach models that yield **a minimum value for** the criterion are to be preferred, the BIC values are compared among various models as the basis for selection of the model. Hence since the BIC criterion imposes a greater penalty for the number of estimated model parameters than AIC. Use of minimum BIC for model selection would always result in a chosen model whose number of parameters is no greater than that chosen under AIC.

| **Model ARIMA(1,0,2)** | Model Fit statistics |
| --- | --- |
| Normalized BIC |
| Xt-Model\_1 | -.039 |

| **Model ARIMA(2,0,2**) | Model Fit |
| --- | --- |
| Normalized BIC |
| Xt-Model\_1 | -.032 |

| **Model(ARIMA(1,0,1)** | Model Fit |
| --- | --- |
| Normalized BIC |
| Xt-Model\_1 | -.046 |

Hence after comparing all the BIC values we select ARIMA (1, 0, 1) as it has minimum BIC value.

| **Model Statistics** | | | | | |
| --- | --- | --- | --- | --- | --- |
| Model ARIMA(1,0,1) |  | Model Fit statistics | | | |
| Number of Predictors | Stationary R-squared | RMSE | MAPE | Normalized BIC |
| Xt-Model\_1 | 0 | .144 | .967 | 11.968 | -.046 |

* The model statistics table provides summary information and goodness-of-fit
* statistics for each estimated model Absolute percentage error is a measure of how much a dependent series varies from its model-predicted level. By examining the mean and maximum across all models, you can get an indication of the uncertainty in your predictions.
* RMSE serves as a good basis of comparison of different models .It is a common criterion of lack of fit. The sum of squared residuals should be quite small.so that the R2 of the model will be quite large It also contains percentile values that provide information on the distribution of the statistic across models. For each percentile, that percentage of models have a value of the fit statistic below the stated value For instance 95% of the models have a value of MAPE that is less than 11.968.
* Larger values of stationary R-squared (up to a maximum value of 1) indicate better fit. A value of 0.259 means that the model poorly explains the observed variation in the series. This statistic provides an estimate of proportion of the total variation in the series that is explained by the model and is preferable to ordinary R2 when there is a trend or seasonal pattern.

| **Model Fit** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Fit Statistic | Percentile | | | | | | |
| 5 | 10 | 25 | 50 | 75 | 90 | 95 |
| Stationary R-squared | .144 | .144 | .144 | .144 | .144 | .144 | .144 |
| R-squared | .144 | .144 | .144 | .144 | .144 | .144 | .144 |
| RMSE | .967 | .967 | .967 | .967 | .967 | .967 | .967 |
| MAPE | 11.968 | 11.968 | 11.968 | 11.968 | 11.968 | 11.968 | 11.968 |
| Normalized BIC | -.046 | -.046 | -.046 | -.046 | -.046 | -.046 | -.046 |

| Mean | Minimum | Maximum |
| --- | --- | --- |
| .144 | .144 | .144 |
| .144 | .144 | .144 |
| .967 | .967 | .967 |
| 11.968 | 11.968 | 11.968 |
| -.046 | -.046 | -.046 |

* Absolute percentage error is a measure of how much a dependent series varies from its model-predicted level and provides an indication of the uncertainty in your predictions .The mean absolute percentage error has a fixed value of 11.968% across all models.

| Model Statistics | | | | |
| --- | --- | --- | --- | --- |
| Model | Ljung-Box Q(18) | | |  |
| Statistics | DF | Sig. | Number of Outliers |
| Xt-Model\_1 | 17.776 | 16 | .337 | 0 |

The Ljung-Box statistic, also known as the modified Box-Pierce statistic, provides an indication of whether the model is correctly specified. A significance value > 0.05 **And conclude that** 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).Hence we conclude that there are no outliers

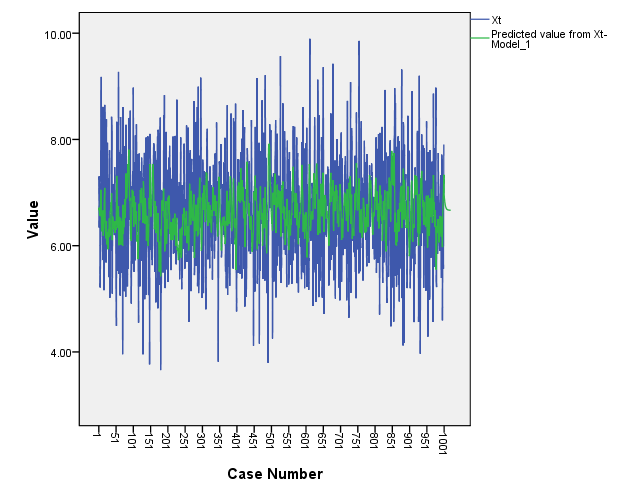
.

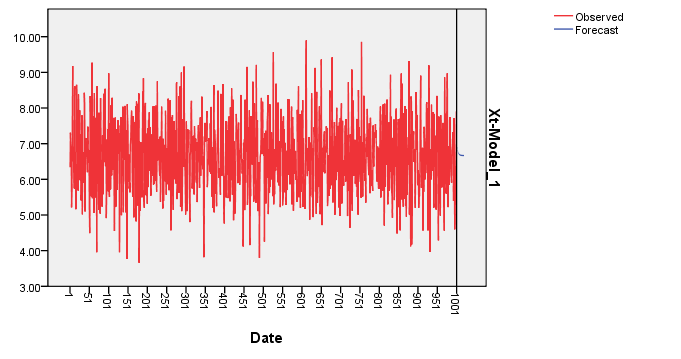
| **ARIMA Model Parameters** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | Estimate | SE |
| Xt-Model\_1 | Xt | No Transformation |  | Constant | 6.667 | .000 |
| AR | Lag 1 | .703 | .023 |
| MA | Lag 1 | 1.000 | .060 |

| t | Sig. |
| --- | --- |
| 16914.876 | .000 |
| 30.436 | .000 |
| 16.670 | .000 |

* As the p value<0.05 we may reject null hypothesis (coefficients are 0) and conclude that the parameter estimates are significant.
* The ARIMA model parameters table displays values for all of the parameters in the model, with an entry for each estimated model labeled by the model identifier. For our purposes, it will list all of the variables in the model, including the dependent variable and any independent variables that were significant.

Intercept is = (6.667\* (1- .703)) =1.98. **So the fitted ARMA model is 𝑋𝑡 = 1.98 + 0.703𝑋𝑡−1 – εt**





* The figure in the LHS graphically presents the observed values along with the values predicted by the fitted model. A line of demarcation separates the actual from the predicted values. Superimposed on the plot of the original series we have the ARIMA predictions.
* The graph accurately capture the amount of variation that we should expect from the given model.

Predicted values are represented by the middle green line. Our forecasted values are plotted in the graph below. Red line at the Centre shows the Observed values.

As we cannot see the forecasted blue line which is merging fully with the observed values so it indicates a perfect fit

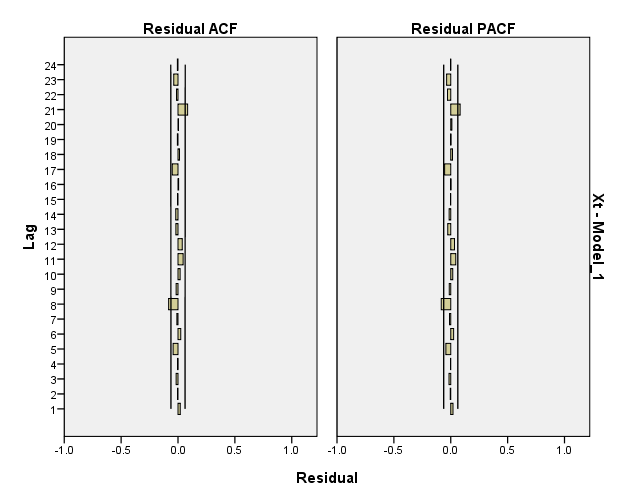
b. **ACF and PACF for White Noise**

Model diagnosis entails residual analysis as well. If the model is properly specified and the model parameters account for all of the systematic variance, then the residuals should resemble white noise.

Residual analysis is performed with the ACF and PACF. These correlograms can be examined with reference to modified portmanteau tests with their associated significance levels.

White noise residuals do not have significant p values. These White noise p values of the residuals should not be< 0.05.

Graphically white noise residuals have associated spikes that do not extend beyond the confidence interval limits. The ACF and PACF plots reveal these limits as dotted lines spreading out from the midpoint of the plot. Indication of significant ACF or PACF residual spikes is empirical evidence of lack of fit



As spikes do not protrude beyond the limit of 2 S.E.”s on each side of the central vertical axis of no autocorrelation. Then the autocorrelation or partial autocorrelation of the residuals do not have significant spikes with p value<0.05.

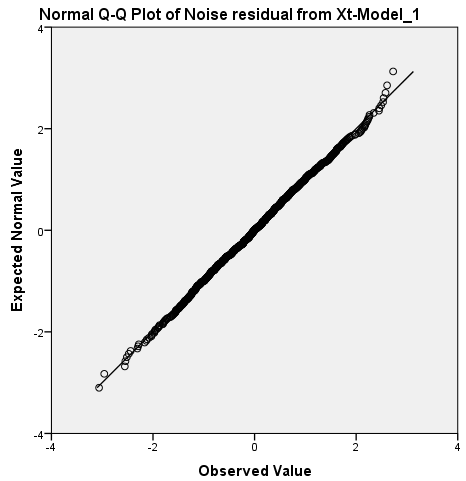
So there is no indication of lack of fit.

If the model is reasonable the residuals will look like white noise (independent 0 mean random variables) which means that the ACF and PACF are theoretically 0 atall non zero lags.

The sample estimates in our case indicates that the residuals do not have significant lagged ACF s and PACF s.So that is good for the model.

| Estimated Distribution Parameters | | |
| --- | --- | --- |
|  |  | Xt |
| Normal Distribution | Location | 6.6679 |
| Scale | 1.04428 |
| The cases are unweighted. | | |

* QQ plot plots the quantiles of a variable distribution against the quantiles of any number of test distributions.
* These plots are generally used to determine whether the distribution of a variable matches a given distribution (normal in this case)
* If the data really are from a normal distribution the plot should approximate a straight line. Hence through the given figure we can validate the assumption of the white noise being Gaussian.



We consider the following hypothesis:

Ho:Data is normally distributed.

H1: Data is not normally distributed

As p value is >0.05 we may accept null hypothesis and conclude that Data is normally distributed

|  | Kolmogorov-Smirnova | | |
| --- | --- | --- | --- |
|  | Statistic | df | Sig. |
| Noise residual from Xt-Model\_1 | .021 | 1000 | .200\* |

**The forecast for the next 20 time points is 6.97,6.88,6.82,6.77,6.74,6.72,6.70,6.69,6.69,6.68,6.68,6.67,6.67,6.67,6.67,6.67, 6.67,6.67,6.67,6.67**