TIME SERIES ANALYSIS

L.M. BHAR AND V.K.SHARMA
Indian Agricultural Statistics Research Institute
Library Avenue, New Delhi-110 012
lmb@iasri.res.in

1. Introduction

Time series (TS) data refers to observations on a variable that occur in a time sequence. Mostly these observations are collected at equally spaced, discrete time intervals. When there is only one variable upon which observations are made then we call them a single time series or more specifically a univariate time series. A basic assumption in any time series analysis/ modeling is that some aspects of the past pattern will continue to remain in the future. Also under this set up, the time series process is based on past values of the main variable but not on explanatory variables which may affect the variable/ system. So the system acts as a black box and we may only be able to know about 'what' will happen rather than 'why' it happens. So if time series models are put to use, say, for instance, for forecasting purposes, then they are especially applicable in the 'short term'. Here it is assumed that information about the past is available in the form of numerical data. Ideally, at least 50 observations are necessary for performing TS analysis/ modeling, as propounded by Box and Jenkins who were pioneers in TS modeling.

2. Time Series Components and Decomposition

An important step in analysing TS data is to consider the types of data patterns, so that the models most appropriate to those patterns can be utilized. Four types of time series components can be distinguished. They are

(i) Horizontal – when data values fluctuate around a constant value
(ii) Trend – when there is long term increase or decrease in the data
(iii) Seasonal – when a series is influenced by seasonal factor and recurs on a regular periodic basis
(iv) Cyclical – when the data exhibit rise and falls that are not of a fixed period

Many data series include combinations of the preceding patterns. After separating out the existing patterns in any time series data, the pattern that remains unidentifiable form the 'random' or 'error' component. Time plot (data plotted over time) and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data. A crude yet practical way of decomposing the original data (ignoring cyclical pattern) is to go for a seasonal decomposition either by assuming an additive or multiplicative model viz.

$$Y_t = T_t + S_t + E_t \text{ or } Y_t = T_t \,.\, S_t \,.\, E_t ,$$

where

Y_t - Original TS data
T_t - Trend component
S_t – Seasonal component
E_t – Error/ Irregular component

If the magnitude of a TS varies with the level of the series then one has to go for a multiplicative model else an additive model. This decomposition may enable one to study the TS components separately or will allow workers to de-trend or to do seasonal adjustments if needed for further analysis.

3. Moving Averages and Exponential Smoothing Methods

3.1 Simple Moving Averages

A Moving Average (MA) is simply a numerical average of the last N data points. There are prior MA, centered MA etc. in the TS literature. In general, the moving average at time t, taken over N periods, is given by

$$M_t^{[1]} = \frac{Y_t + Y_{t-1} + \ldots + Y_{t-N+1}}{N}$$

where Y_t is the observed response at time t. Another way of stating the above equation is

$$M_t^{[1]} = M_{t-1}^{[1]} + (Y_t - Y_{t-N}) / N$$

At each successive time period the most recent observation is included and the farthest observation is excluded for computing the average. Hence the name 'moving' averages.

3.2 Double Moving Averages

The simple moving average is intended for data of constant and no trend nature. If the data have a linear or quadratic trend, the simple moving average will be misleading. In order to correct for the bias and develop an improved forecasting equation, the double moving average can be calculated. To calculate this, simply treat the moving averages $M_t^{[1]}$ over time as individual data points and obtain a moving average of these averages.

3.3 Simple Exponential Smoothing (SES)

Let the time series data be denoted by $Y_1, Y_2, \ldots, Y_t$. Suppose we wish to forecast the next value of our time series Y_{t+1} that is yet to be observed with forecast for Y_t denoted by F_t. Then the forecast F_{t+1} is based on weighting the most recent observation Y_t with a weight value α and weighting the most recent forecast F_t with a weight of $(1-\alpha)$ where α is a smoothing constant/ weight between 0 and 1. Thus the forecast for the period t+1 is given by

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

The choice of α has considerable impact on the forecast. A large value of α (say 0.9) gives very little smoothing in the forecast, whereas a small value of α (say 0.1) gives considerable smoothing. Alternatively, one can choose α from a grid of values (say α=0.1,0.2,…,0.9) and choose the value that yields the smallest MSE value.

If the above model is expanded recursively then F_{t+1} will come out to be a function of α, past y_t values and F_1. So, having known values of α and past values of y_t our point of concern relates to initializing the value of F_1. One method of initialization is to use the first observed value Y_1 as the first forecast ($F_1=Y_1$) and then proceed. Another possibility

would be to average the first four or five values in the data set and use this as the initial forecast. However, because the weight attached to this user-defined F_1 is minimal, its effect on F_{t+1} is negligible.

3.4 Double Exponential Smoothing (Holt)

This is to allow forecasting data with trends. The forecast for Holt's linear exponential smoothing is found by having two more equations to SES model to deal with – one for level and one for trend. The smoothing parameters (weights) α and β can be chosen from a grid of values (say, each combination of α=0.1,0.2,…,0.9 and β=0.1,0.2,…,0.9) and then select the combination of α and β which correspond to the lowest MSE.

3.5 Triple Exponential Smoothing (Winters)

This method is recommended when seasonality exists in the time series data. This method is based on three smoothing equations – one for the level, one for trend, and one for seasonality. It is similar to Holt's method, with one additional equation to deal with seasonality. In fact there are two different Winter's methods depending on whether seasonality is modeled in an additive or multiplicative way.

4. Stationarity of a TS process

A TS is said to be stationary if its underlying generating process is based on a constant mean and constant variance with its autocorrelation function (ACF) essentially constant through time. Thus, if different subsets of a realization are considered (time series 'sample') the different subsets will typically have means, variances and autocorrelation functions that do not differ significantly.

A statistical test for stationarity is the most widely used Dickey Fuller test. To carry out the test, estimate by OLS the regression model

$$y'_t = \phi y_{t-1} + b_1 y'_{t-2} + \ldots + b_p y'_{t-p}$$

where y'_t denotes the differenced series ($y_t - y_{t-1}$). The number of terms in the regression, p, is usually set to be about 3. Then if ϕ is nearly zero the original series y_t needs differencing and if $\phi < 0$ then y_t is already stationary.

5. Autocorrelation Functions

5.1 Autocorrelation

Autocorrelation refers to the way the observations in a time series are related to each other and is measured by the simple correlation between current observation (Y_t) and observation from p periods before the current one (Y_{t-p}). That is for a given series Y_t, autocorrelation at lag p = correlation (Y_t, Y_{t-p}) and is given by

$$r_p = \frac{\sum_{t=1}^{n-p}(Y_t - \overline{Y})(Y_{t-p} - \overline{Y})}{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}$$

It ranges from –1 to +1. Box and Jenkins has suggested that maximum number of useful r_p are roughly n/4 where n is the number of periods upon which information on Y_t is available.

5.2 Partial Autocorrelation

Partial autocorrelations are used to measure the degree of association between Y_t and Y_{t-p} when the Y-effects at other time lags 1, 2, 3,…,p-1 are removed.

5.3 Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)

Theoretical ACFs and PACFs (Autocorrelations versus lags) are available for the various models chosen. Thus compare the correlograms (plot of sample ACFs versus lags) with these theoretical ACF/ PACFs, to find a reasonably good match and tentatively select one or more ARIMA models. The general characteristics of theoretical ACFs and PACFs are as follows:- (here 'spike' represents the line at various lags in the plot with length equal to magnitude of autocorrelations)

Model	ACF	PACF
AR	Spikes decay towards zero	Spikes cutoff to zero
MA	Spikes cutoff to zero	Spikes decay to zero
ARMA	Spikes decay to zero	Spikes decay to zero

6. Description of ARIMA Representation

6.1 ARIMA Modeling

In general, an ARIMA model is characterized by the notation ARIMA (p,d,q) where, p, d and q denote orders of auto-regression, integration (differencing) and moving average respectively. In ARIMA, TS is a linear function of past actual values and random shocks. For instance, given a time series process $\{Y_t\}$, a first order auto-regressive process is denoted by ARIMA (1,0,0) or simply AR(1) and is given by

$$Y_t = \mu + \phi_1 Y_{t-1} + \varepsilon_t$$

and a first order moving average process is denoted by ARIMA (0,0,1) or simply MA(1) and is given by

$$Y_t = \mu - \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Alternatively, the model ultimately derived, may be a mixture of these processes and of higher orders as well. Thus a stationary ARIMA (p, q) process is defined by the equation

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \ldots - \theta_q \varepsilon_{t-q} + \varepsilon_t$$

where ε_t's are independently and normally distributed with zero mean and constant variance σ^2 for t = 1,2,...n. The values of p and q, in practice lie between 0 and 3.

6.2 Seasonal ARIMA Modeling

Identification of relevant models and inclusion of suitable seasonal variables are necessary for seasonal modeling and their applications, say, forecasting production of crops. Seasonal forecasts of production of principal crops are of greater utility for

planners, administrators and researchers alike. Agricultural seasons vary significantly among the states of India. For example, Tamil Nadu has unique three-season cropping pattern for paddy crop whereas two-season paddy rules elsewhere in the country. Thus seasonal forecasts of crop production can also be made using seasonal ARIMA models.

The Seasonal ARIMA i.e. ARIMA (p,d,q) $(P,D,Q)_s$ model is defined by

$$\phi_p(B)\varnothing_P(B^s)\nabla^d\nabla_s^D Y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t,$$

where

$$\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p, \quad \theta_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$$

$$\varnothing_P(B^s) = 1 - \varnothing_1 B^s - \ldots - \varnothing_P B^{sP}, \quad \Theta_Q(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{sQ}$$

B is the backshift operator (i.e. $B\,Y_t = Y_{t-1}$, $B^2 Y_t = Y_{t-2}$ and so on), 's' the seasonal lag and 'ε_t' a sequence of independent normal error variables with mean 0 and variance σ^2. $\varnothing$'s and ϕ's are respectively the seasonal and non-seasonal autoregressive parameters. Θ's and θ's are respectively seasonal and non-seasonal moving average parameters. p and q are orders of non-seasonal autoregression and moving average parameters respectively whereas P and Q are that of the seasonal auto regression and moving average parameters respectively. Also d and D denote non-seasonal and seasonal differences respectively.

7. The Art of ARIMA Model Building

7.1 Identification

The foremost step in the process of modeling is to check for the stationarity of the series, as the estimation procedures are available only for stationary series. There are two kinds of stationarity, viz., stationarity in 'mean' and stationarity in 'variance'. A look at the graph of the data and structure of autocorrelation and partial correlation coefficients may provide clues for the presence of stationarity. Another way of checking for stationarity is to fit a first order autoregressive model for the raw data and test whether the coefficient 'ϕ_1' is less than one. If the model is found to be non-stationary, stationarity could be achieved mostly by differencing the series. Or use a Dickey Fuller test (see section 4). Stationarity in variance could be achieved by some modes of transformation, say, log transformation. This is applicable for both seasonal and non-seasonal stationarity.

Thus, if 'X_t' denotes the original series, the non-seasonal difference of first order is

$$Y_t = X_t - X_{t-1}$$

followed by the seasonal differencing (if needed)

$$Z_t = Y_t - Y_{t-s} = (X_t - X_{t-1}) - (X_{t-s} - X_{t-s-1})$$

The next step in the identification process is to find the initial values for the orders of seasonal and non-seasonal parameters, p, q, and P, Q. They could be obtained by looking for significant autocorrelation and partial autocorrelation coefficients (see section 5.3). Say, if second order auto correlation coefficient is significant, then an AR (2), or MA (2) or ARMA (2) model could be tried to start with. This is not a hard and fast rule, as sample autocorrelation coefficients are poor estimates of population autocorrelation coefficients.

Still they can be used as initial values while the final models are achieved after going through the stages repeatedly.

7.2 Estimation

At the identification stage, one or more models are tentatively chosen that seem to provide statistically adequate representations of the available data. Then precise estimates of parameters of the model are obtained by least squares as advocated by Box and Jenkins. Standard computer packages like SAS, SPSS etc. are available for finding the estimates of relevant parameters using iterative procedures.

7.3 Diagnostics

Different models can be obtained for various combinations of AR and MA individually and collectively. The best model is obtained with following diagnostics:

7.3.1 Low Akaike Information Criteria (AIC)/ Bayesian Information Criteria (BIC)/ Schwarz-Bayesian Information Criteria (SBC)

AIC is given by AIC = (-2 log L + 2 m) where m=p+ q+ P+ Q and L is the likelihood function. Since -2 log L is approximately equal to $\{n (1+\log 2\pi) + n \log \sigma^2\}$ where σ^2 is the model MSE, AIC can be written as AIC=$\{n (1+\log 2\pi) + n \log \sigma^2 + 2 m\}$and because first term in this equation is a constant, it is usually omitted while comparing between models. As an alternative to AIC, sometimes SBC is also used which is given by SBC = $\log \sigma^2 + (m \log n) /n$.

7.3.2 Non-significance of auto correlations of residuals via Portmonteau tests (Q-tests based on Chisquare statistics)-Box-Pierce or Ljung-Box texts

After tentative model has been fitted to the data, it is important to perform diagnostic checks to test the adequacy of the model and, if need be, to suggest potential improvements. One way to accomplish this is through the analysis of residuals. It has been found that it is effective to measure the overall adequacy of the chosen model by examining a quantity Q known as Box-Pierce statistic (a function of autocorrelations of residuals) whose approximate distribution is chi-square and is computed as follows:

$$Q = n \Sigma r^2 (j)$$

where summation extends from 1 to k with k as the maximum lag considered, n is the number of observations in the series, r (j) is the estimated autocorrelation at lag j; k can be any positive integer and is usually around 20. Q follows Chi-square with $(k-m_1)$ degrees of freedom where m_1 is the number of parameters estimated in the model. A modified Q statistic is the Ljung-box statistic which is given by

$$Q = n(n+2) \Sigma r^2 (j) / (n-j)$$

The Q Statistic is compared to critical values from chi-square distribution. If model is correctly specified, residuals should be uncorrelated and Q should be small (the probability value should be large). A significant value indicates that the chosen model does not fit well.

All these stages require considerable care and work and they themselves are not exhaustive.

EXERCISE

(a) Identify a series of ARIMA (p,1,q) models (p, q ranging from 0 to 5) that might be useful in describing the following time series data. Which of your models is the best according to their AIC values?

(b) For the best model perform diagnostic tests upon residuals using (i) ACF of forecast errors, (ii) Portmanteau tests

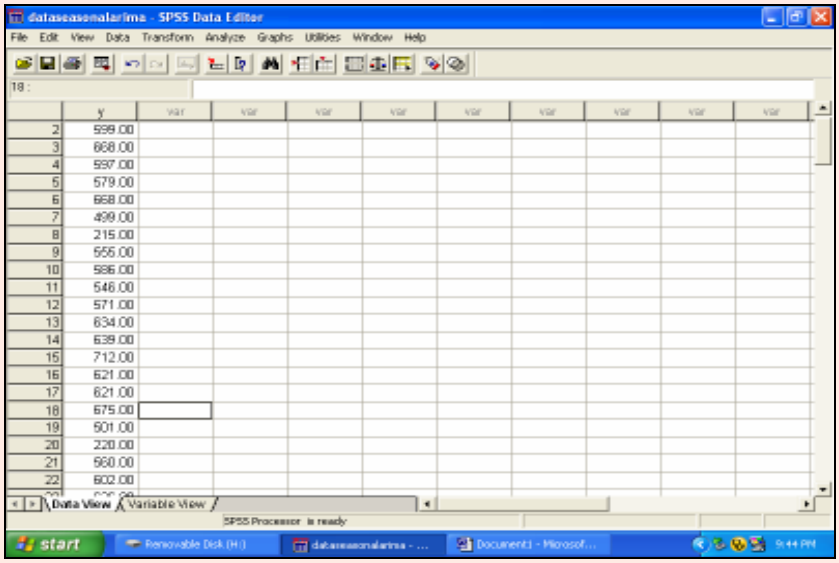
(c) Write this model in terms of backshift operator and then without using backshift operator.

(d) Forecast for ten lead periods ahead.

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1981	562	599	668	597	579	668	499	215	555	586	546	571
1982	634	639	712	621	621	675	501	220	560	602	626	605
1983	646	658	712	687	723	707	629	237	613	730	734	651
1984	676	748	816	729	701	790	594	230	617	691	701	705
1985	747	773	813	766	728	749	680	241	680	708	694	772
1986	795	788	889	797	751	821	691	290	727	868	812	799
1987	843	847	941	804	840	871	656	370	742	847	731	898
1988	778	856	938	813	783	823	657	310	780	860	780	807
1989	895	856	893	875	835	934	832	300	791	900	781	880
1990	875	992	976	968	871	1006	832	345	849	913	868	993

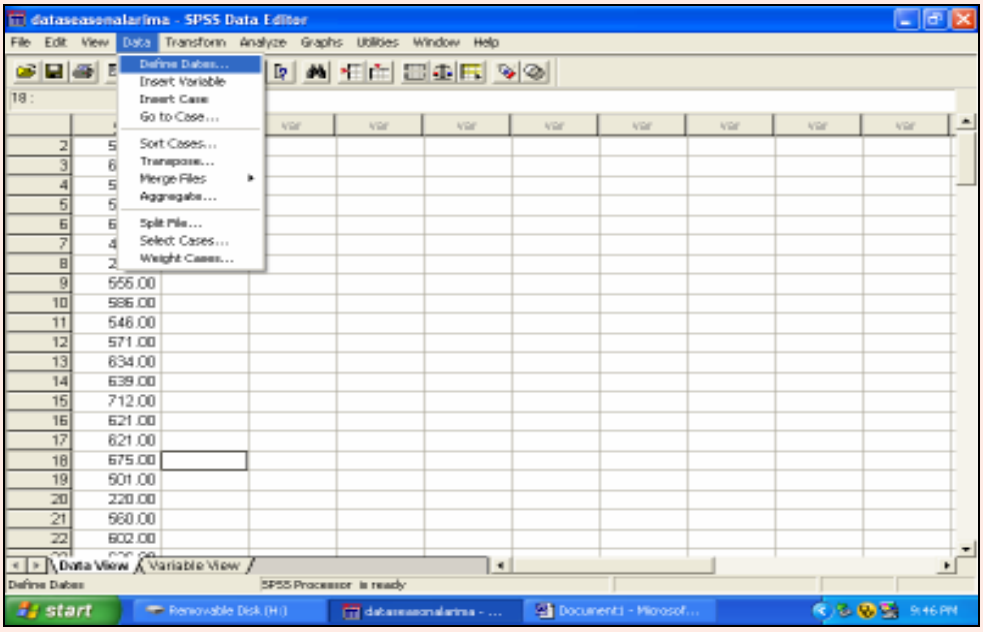
Steps for Analysis using SPSS

Data Entry

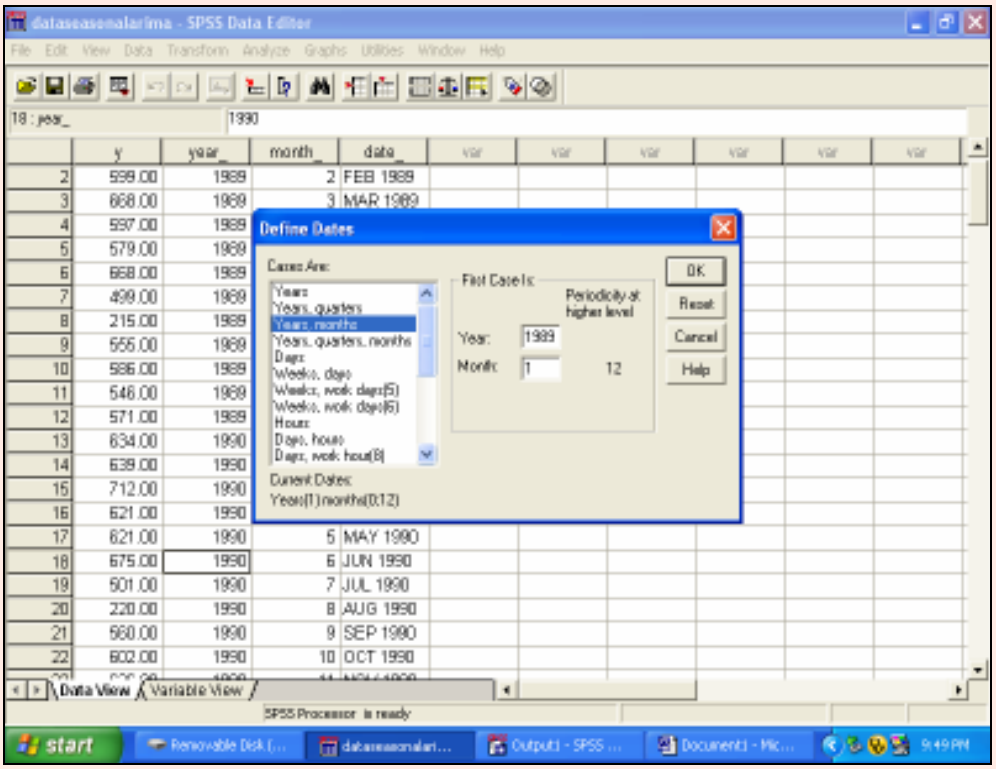


- Define time series data
- Here seasonality parameter s is 12, since there are 12 months in each season that is year
- Go to data
- Then define date as follows:

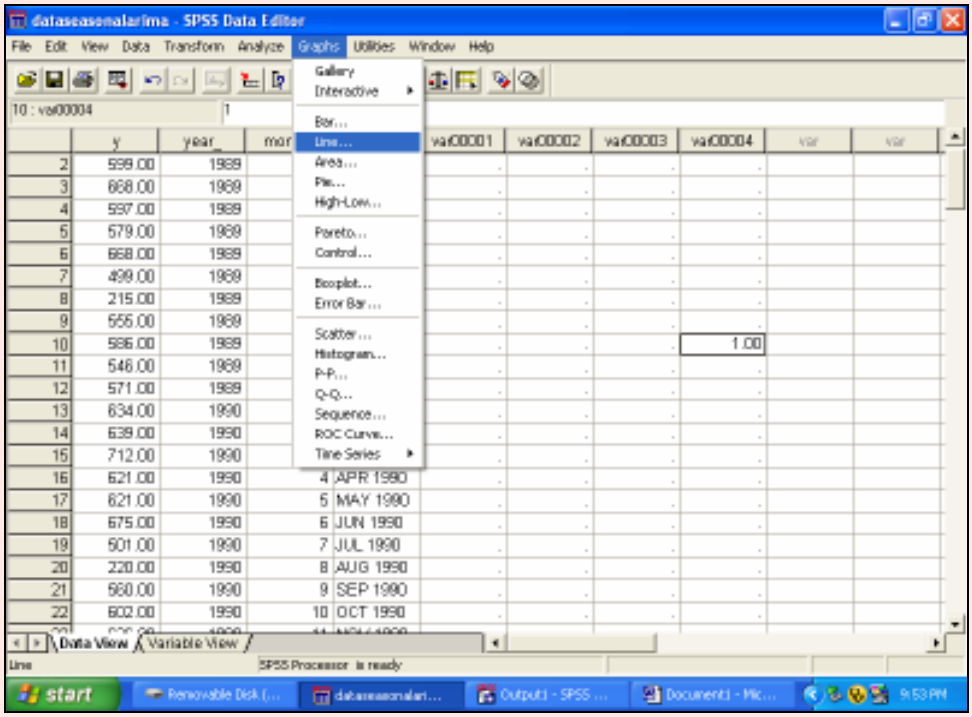
Click | Data | → | Date | → | Year, Month |



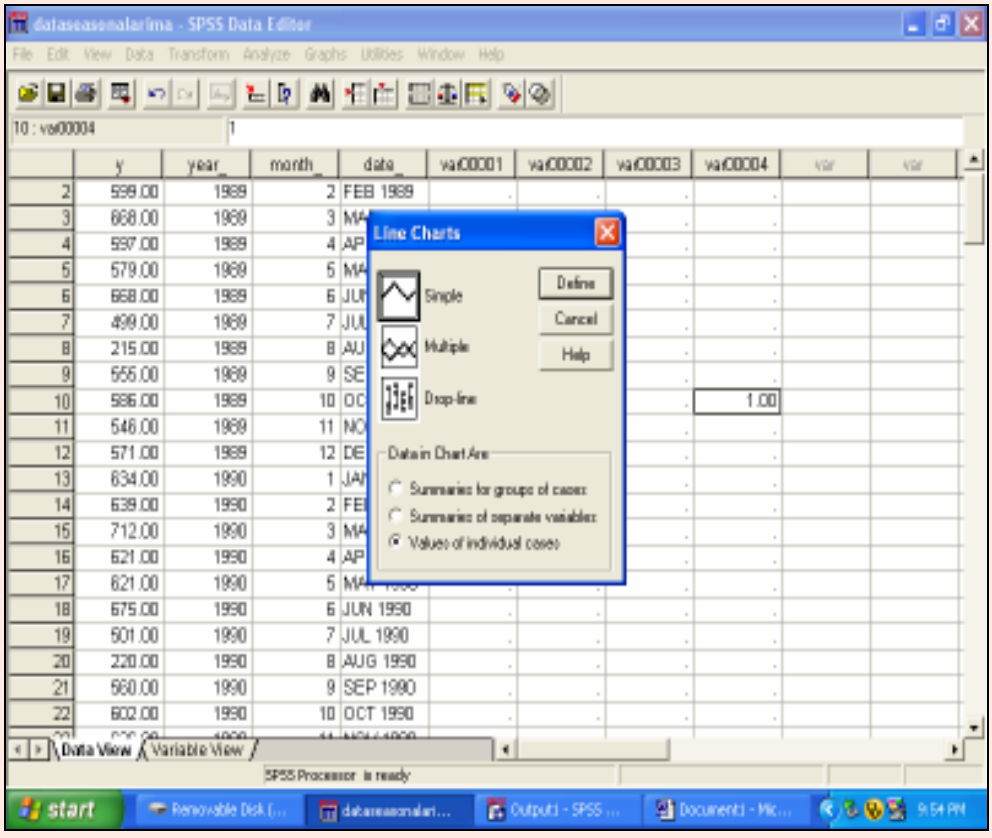
- Put the initial year and month like 1989 and 1 for the month of January



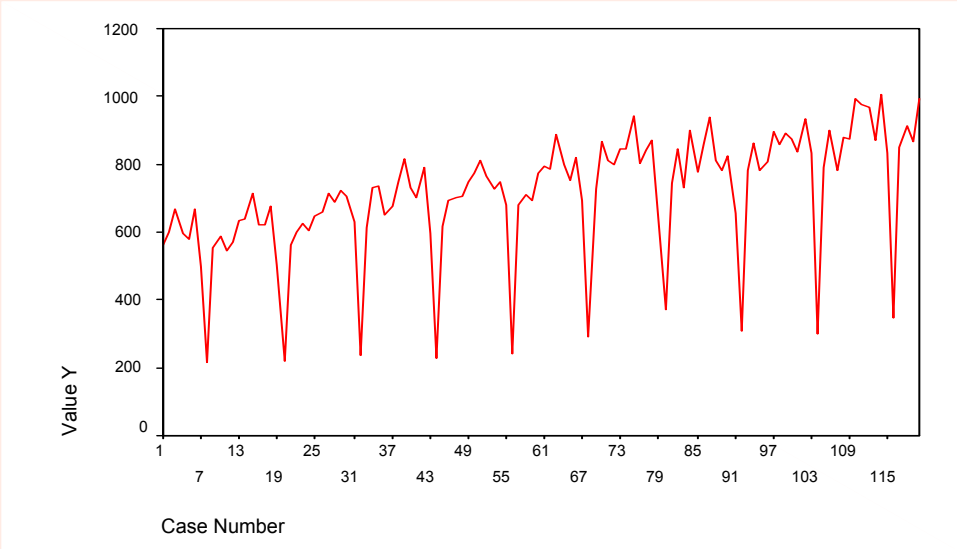
- Then go to graph to check Stationarity



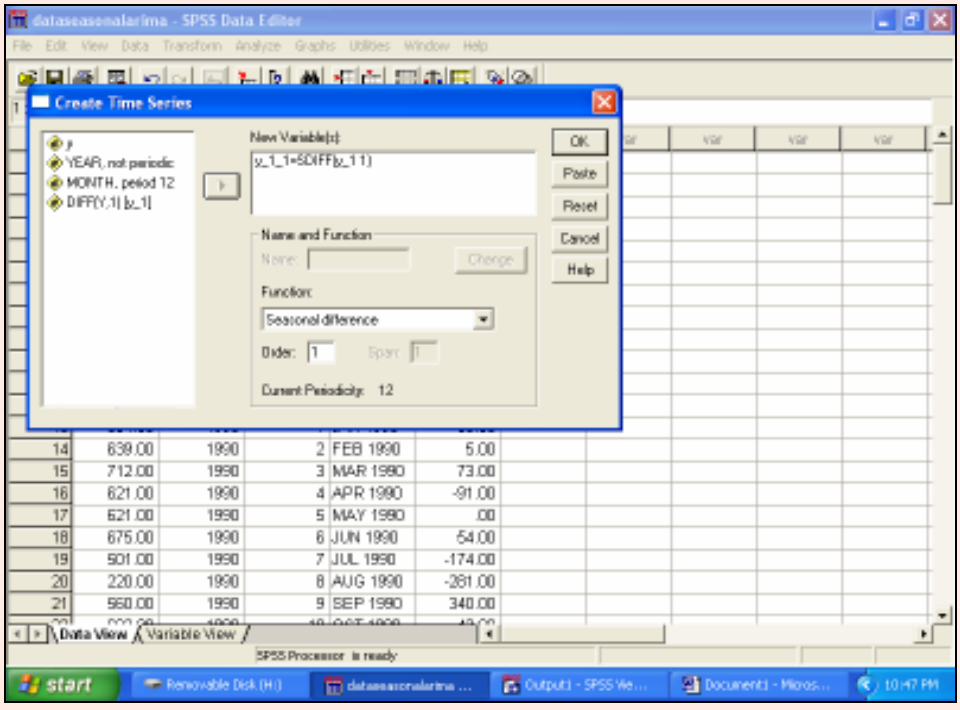
- Select line and values



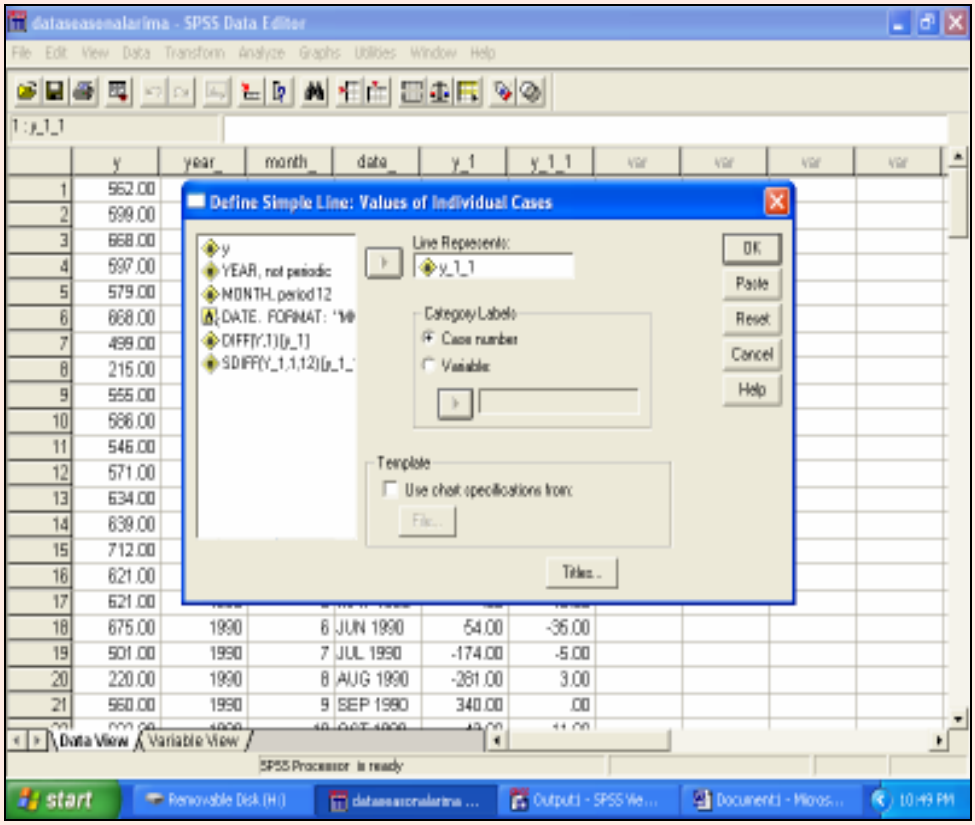
- For the present example the graph would look like this:



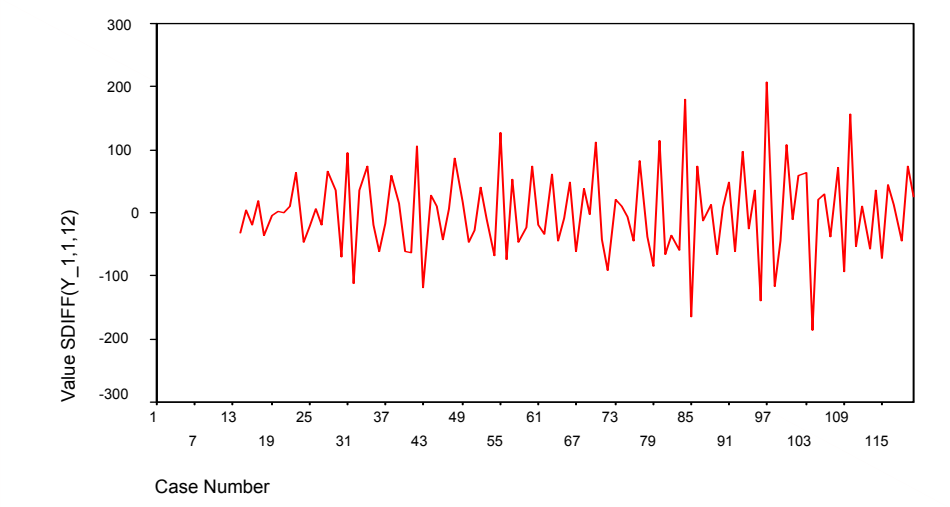
- Which shows clearly that the series is not stationary and there is seasonal variation
- We apply difference and seasonal difference once
- For that go to transform MENU and select Create time series.



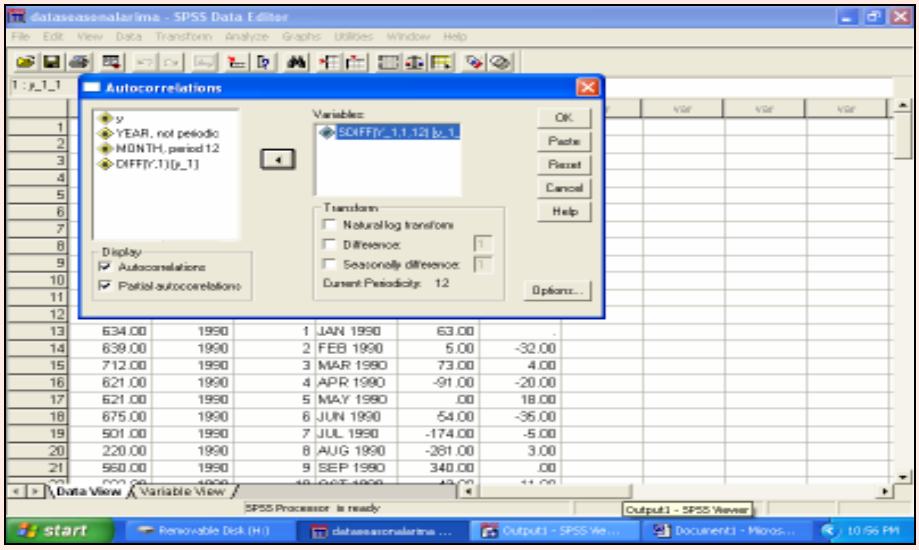
- Created new variable SDIFF (Y_1,1,12) should now be tested for stationarity



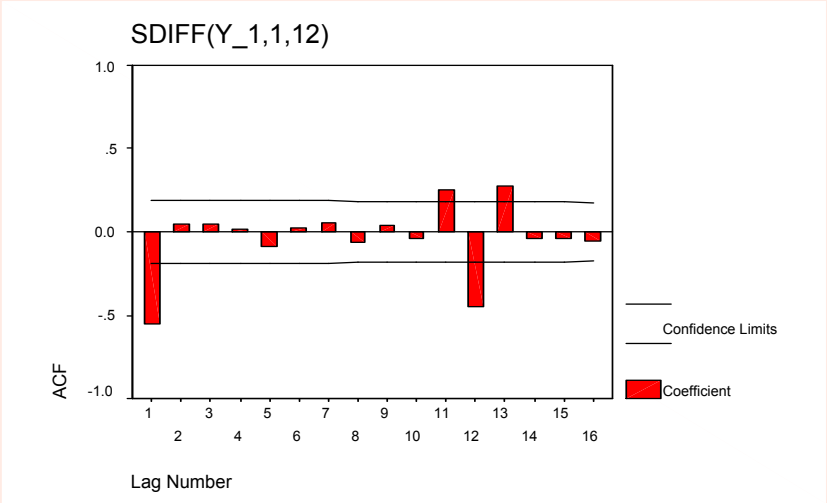
- On the same line as before we get the graph as below:

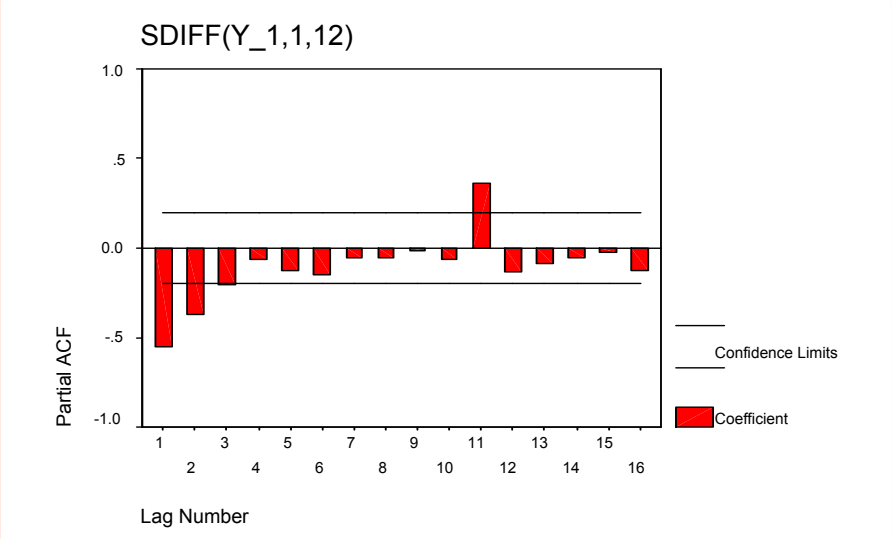


- It seems that data is now stationary.
- Now check for ARIMA (p, d, q)
- Go to graph →time series → auto correlations..



- See the ACF and PACF



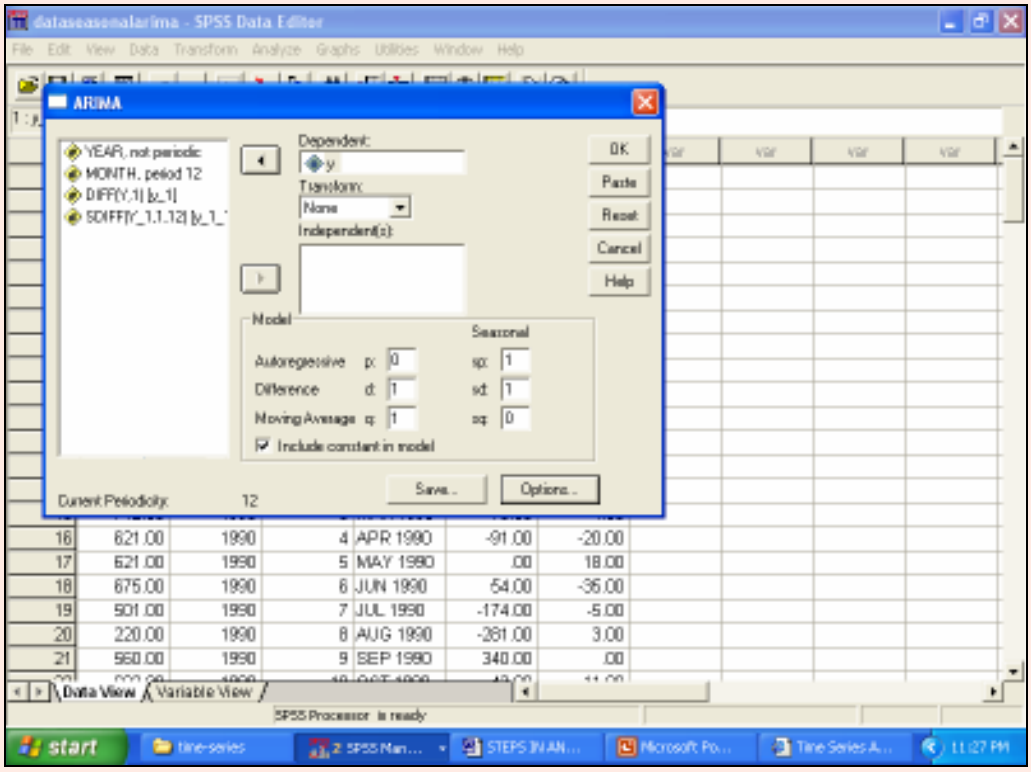


- Determine p and q on the basis of the following table:

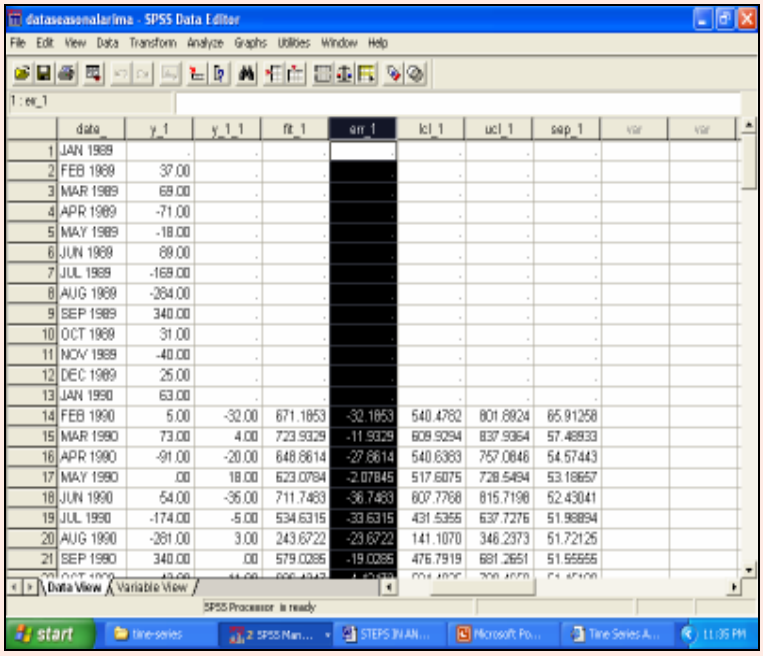
Type of model	Typical pattern of ACF	Typical pattern of PACF
AR(p)	Decays exponentially or with damped sine wave pattern or both	Significant spikes through lags p
MA(q)	Significant spikes through lags q	Declines exponentially
ARMA(p,q)	Exponential decay	Exponential decay

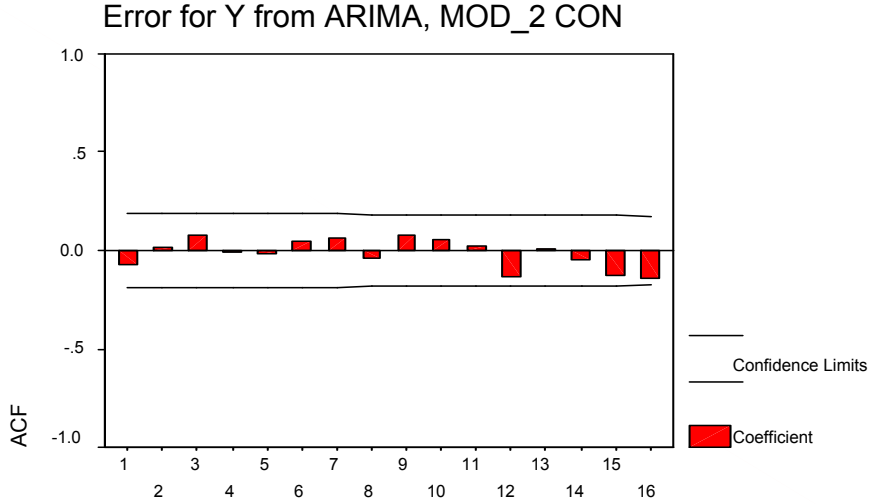
  - Similarly determine P and Q (Seasonal components)
  - For the present case q=1, p=0 d=1, D=1, P=1 and Q=0

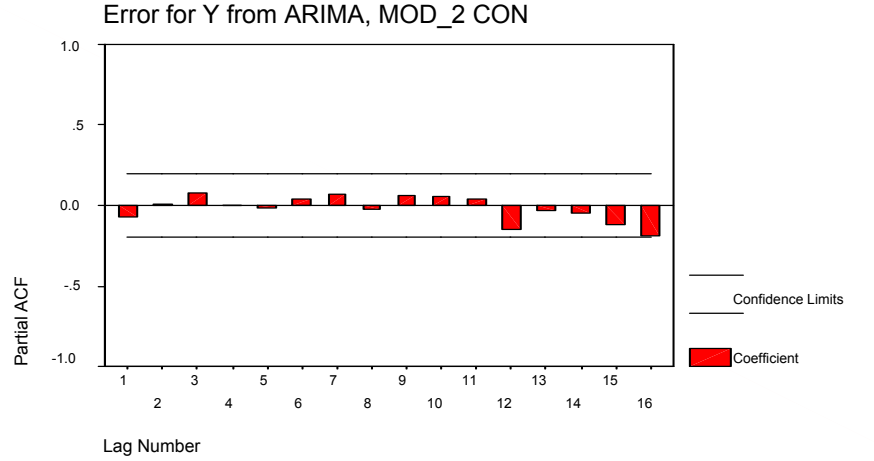
- Go to Analyse → time series → ARIMA.



- Model is fitted
- Observe the AIC or SBC values for the present model
- Fit different models in the neighborhood of p and q
- Observe the AIC or SBC values for all these models
- Choose the model on the basis of least AIC or SBC value
- For diagnostic check see the ACF and PACF of the errors of the fitted model


dataseasonalarima - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : er_1
data_ y_1 y_1_1 fit_1 er_1 lcl_1 ucl_1 sep_1 var var
1 JAN 1989 . . . .
2 FEB 1989 37.00 . . . .
3 MAR 1989 69.00 . . . .
4 APR 1989 -71.00 . . . .
5 MAY 1989 -18.00 . . . .
6 JUN 1989 89.00 . . . .
7 JUL 1989 -169.00 . . . .
8 AUG 1989 -284.00 . . . .
9 SEP 1989 340.00 . . . .
10 OCT 1989 31.00 . . . .
11 NOV 1989 -40.00 . . . .
12 DEC 1989 25.00 . . . .
13 JAN 1990 63.00 . . . .
14 FEB 1990 5.00 -32.00 671.1853 -32.1853 540.4782 801.8924 65.91258
15 MAR 1990 73.00 4.00 723.9329 -11.9329 609.9294 837.9364 57.48833
16 APR 1990 -91.00 -20.00 648.8614 -27.8614 540.6383 757.0846 54.57443
17 MAY 1990 .00 18.00 623.0784 -2.07845 517.6075 728.5494 53.18657
18 JUN 1990 54.00 -36.00 711.7483 -36.7483 607.7768 815.7198 52.43041
19 JUL 1990 -174.00 -5.00 534.6315 -33.6315 431.5355 637.7276 51.98894
20 AUG 1990 -281.00 3.00 243.6722 -23.6722 141.1070 346.2373 51.72125
21 SEP 1990 340.00 .00 579.0285 -19.0285 476.7919 681.2651 51.55555
Data View Variable View
SPSS Processor is ready
start
time-series
2 SPSS Man...
STEPS IN AN...
Microsoft Po...
Time Series A...
11:05 PM


Error for Y from ARIMA, MOD_2 CON
ACF
1.0
.5
0.0
-.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag Number
Confidence Limits
Coefficient


Error for Y from ARIMA, MOD_2 CON
Partial ACF
1.0
.5
0.0
-.5
-1.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Lag Number
Confidence Limits
Coefficient

- In the present case the ACF and PACF are lying within the limits.
- So the present model is fitted well.
- The final model is: ARIMA (0, 1, 1) $(1, 1,0)^{12}$
- The model can be written as

$$(1- \phi \mathbf{B}^{12})\,(1- \mathbf{B})(1- \mathbf{B}^{12})\,\mathbf{Y_t} = (1- \theta \mathbf{B})\,\mathbf{e_t}$$

where $\mathbf{Y_t}$ is the study variable at t^{th} time period, $\mathbf{e}_t$ is the error term, **B** is the back shift operator, that is, $\mathbf{B}\ \mathbf{Y_t} = \mathbf{Y_{t-1}}$, $\mathbf{B}^{12}\ \mathbf{Y_t} = \mathbf{Y_{t-12}}$ and so on. The parameter estimates $\boldsymbol{\theta}$ (Seasonal MA(1) component) is 0.811 and $\boldsymbol{\varphi}$ (Nonseasonal AR(1) component) is – 0.431.

References and Suggested Reading

Blank, D.S. (1986). SAS system for forecasting time series, SAS Institute Inc., USA

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (1994). Time series analysis : Forecasting and control, Pearson Education, Delhi.

Makridakis, S., Wheelwright, S.C. and Hyndman, R.J. (1998). Forecasting: Methods and Applications, John Wiley, New York.

Pankratz, A. (1983). Forecasting with univariate Box – Jenkins models: concepts and cases, New york: John Wiley & Sins.