Chapter 471

ARIMA (Box-Jenkins)

Introduction

Although the theory behind ARIMA time series models was developed much earlier, the systematic procedure for applying the technique was documented in the landmark book by Box and Jenkins (1976). Since then, ARIMA forecasting and Box-Jenkins forecasting usually refer to the same set of techniques. In this chapter, we will document the running of the ARIMA program. The methodology put forth by Box and Jenkins will be outlined in another chapter, since it uses several time series procedures.

ARIMA time series modeling is complex. You will want to become familiar with the details of the methodology before you place a lot of confidence in your forecasts. Our intent is to provide you with the tools you need to become proficient in the Box-Jenkins method.

Data Structure

The data are entered in a single variable.

Missing Values

When missing values are found in the series, they are either replaced or omitted. The replacement value is the average of the nearest observation in the future and in the past or the nearest non-missing value in the past.

If you do not feel that this is a valid estimate of the missing value, you should manually enter a more reasonable estimate before using the algorithm. These missing value replacement methods are particularly poor for seasonal data. We recommend that you replace missing values manually before using the algorithm.

Example 1 – Fitting an ARIMA Model

This section presents an example of how to fit an ARIMA model to a time series. The Intel_Close variable in the Intel dataset will be fit with an ARMA(2,0,0) model.

Setup

To run this example, complete the following steps:

1 Open the Intel example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Intel** and click **OK**.

2 Specify the ARIMA (Box-Jenkins) procedure options

- Find and open the **ARIMA (Box-Jenkins)** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Time Series Variable ..... **Intel_Close**
Regular AR **2**
Regular MA..... **0**

Reports Tab

Forecasts..... **Data and Forecasts**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Iterations (Minimization Phase)

Iterations (Minimization Phase)

Iteration Number	Error Sum of Squares	Lambda	AR(1)	AR(2)
0	85.89011	0.1	0.1	0.1
1	22.34787	0.1	0.9479776	-0.1168852
2	17.8811	0.04	1.292394	-0.4623865
3	17.57868	0.016	1.390741	-0.5737677
4	17.57362	0.0064	1.403895	-0.589241
5	17.57359	0.00256	1.40471	-0.5902494

Normal convergence.

This report displays the algorithms progress toward a solution.

Error Sum of Squares

The sum of the squared residuals. This is the value that is being minimized by the algorithm.

Lambda

The value of Marquart's lambda parameter.

AR(...), MA(...)

The values of the autoregressive and moving average parameters. Note that if there are more parameters in the model than will fit on a single report line, only the first few parameters are displayed.

Model Description

Model Description

Series	Intel_Close - MEAN
Model	Regular(2,0,0) Seasonal(No seasonal parameters)
Mean	64.45625
Observations	20
Missing Values	None
Iterations	5
Pseudo R-Squared	84.653036
Residual Sum of Squares	17.57359
Mean Square Error	0.9763107
Root Mean Square	0.9880844

This report displays summary information about the solution.

Series

The name of the variable being analyzed.

Model

The phrase *Regular (p,d,q)* gives the highest order of the regular ARIMA parameters. The *Seasonal(P,D,Q)* gives the highest order of the seasonal ARIMA parameters if they were used.

- *p* Highest order autoregression parameter in the model.
- *d* Number of times the series was differenced.
- *q* Highest order moving average parameter in the model.
- *P* Highest order seasonal autoregression parameter in the model.
- *D* Number of times the series was seasonally differenced.
- *Q* Highest order seasonal moving average parameter in the model.

Mean

The average of the series.

Observations

The number of observations (rows) in the series.

Missing Values

If missing values were found, this option lists the method used to estimate them.

Iterations

The number of iterations before the algorithm converged or was aborted.

Pseudo R-Squared

This value generates a statistic that acts like the R-Squared value in multiple regression. A value near zero indicates a poorly fitting model, while a value near one indicates a well-fitting model. The statistic is calculated as follows:

$$R^2 = 100\left(1 - \frac{SSE}{SST}\right)$$

where *SSE* is the sum of square residuals and *SST* is the total sum of squares after correcting for the mean.

Residual Sum of Squares

The sum of the squared residuals. This is the value that is being minimized by the algorithm.

Mean Square Error

The average squared residual (MSE) is a measure of how closely the forecasts track the actual data. The statistic is popular because it shows up in analysis of variance tables. However, because of the squaring, it tends to exaggerate the influence of outliers (points that do not follow the regular pattern).

Root Mean Square

The square root of MSE. This statistic is popular because it is in the same units as the time series.

Model Estimation

Model Estimation

Parameter				
Name	Estimate	Standard Error	T-Value	P-Value
AR(1)	1.40471	0.2065638	6.8004	0.000000
AR(2)	-0.5902494	0.2330099	-2.5332	0.011304

Parameter Name

The is the name of the parameter that is reported on this line.

- AR(i) The ith-order autoregressive parameter.
- MA(i) The ith-order moving average parameter.
- SAR(i) The ith-order seasonal autoregressive parameter.
- SMA(i) The ith-order seasonal moving average parameter.

Parameter Estimate

This is the estimated parameter value.

Standard Error

A large sample (N>50) estimate of the standard error of the parameter value.

T-Value

The t-test value testing whether the parameter is statistically significant (different from zero). The degrees of freedom is equal to the N minus the number of model parameters and differences.

P-Value

The p-value for the above test. If you were testing at the alpha = 0.05 level of significance, this value would have to be less than 0.05 in order for the parameter to be considered statistically different from zero. When the highest order parameter is not significant, you should decrease the order by one and rerun. When a nonsignificant parameter is not the highest order, you should not delete it.

Asymptotic Correlation Matrix of Parameters

Asymptotic Correlation Matrix of Parameters

	AR(1)	AR(2)
AR(1)	1.000000	-0.881734
AR(2)	-0.881734	1.000000

This report gives the asymptotic estimates of the correlation between the parameter estimates. If some of the correlations are greater than 0.9999, you should consider removing appropriate parameters.

Parameter Name

The is the name of the parameter that is reported on this line.

AR(i) The ith-order autoregressive parameter.

Forecasts

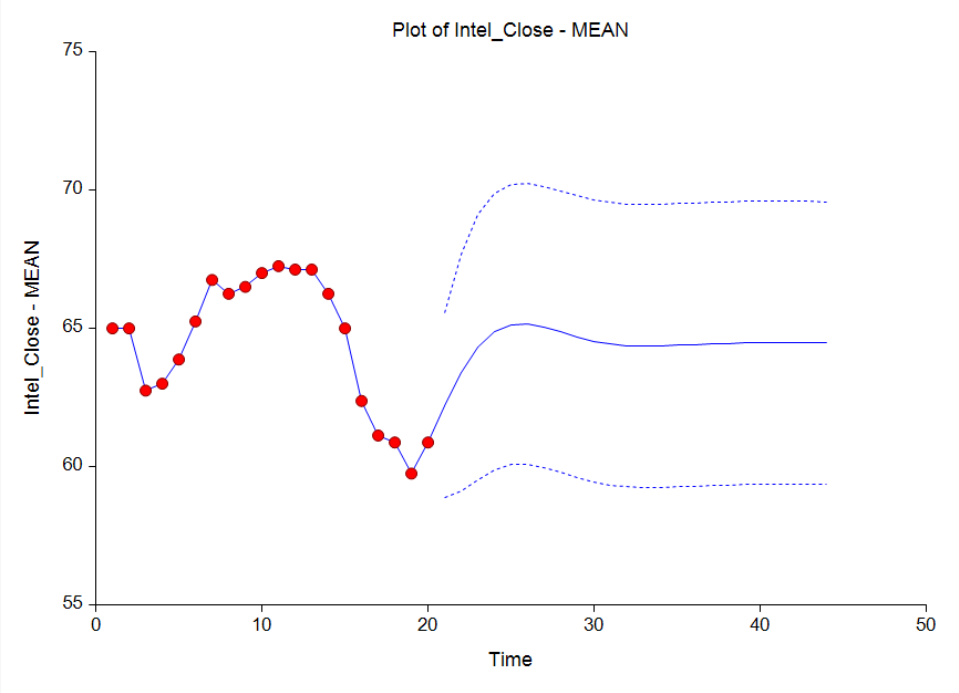
Forecasts of Intel_Close

					95% Prediction Interval Limits	
Row	Date	Actual	Residual	Forecast	Lower	Upper
1	1	65.0	0.1	64.9	61.6	68.2
2	2	65.0	0.0	65.0	61.6	68.3
3	3	62.8	-2.1	64.9	61.6	68.2
4	4	63.0	1.3	61.7	58.4	65.1
5	5	63.9	0.5	63.4	60.1	66.8
6	6	65.3	0.8	64.5	61.2	67.8
7	7	66.8	0.8	65.9	62.6	69.3
8	8	66.3	-1.0	67.2	63.9	70.5
9	9	66.5	0.9	65.6	62.3	69.0
10	10	67.0	0.7	66.3	62.9	69.6
11	11	67.3	0.4	66.8	63.5	70.2
12	12	67.1	0.2	66.9	63.5	70.2
13	13	67.1	0.6	66.6	63.2	69.9
14	14	66.3	-0.4	66.6	63.3	70.0
15	15	65.0	-0.4	65.4	62.1	68.7
16	16	62.4	-1.8	64.2	60.8	67.5
17	17	61.1	-0.1	61.2	57.9	64.6
18	18	60.9	-0.1	61.0	57.7	64.3
19	19	59.8	-1.6	61.4	58.1	64.7
20	20	60.9	0.9	60.0	56.6	63.3
21	21			62.2	58.9	65.5
22	22			63.4	59.1	67.7
23	23			64.3	59.5	69.1
24	24			64.9	59.9	69.9
25	25			65.1	60.1	70.2
26	26			65.2	60.1	70.2
27	27			65.0	60.0	70.1
28	28			64.9	59.8	69.9
29	29			64.7	59.6	69.8
30	30			64.5	59.4	69.6
31	31			64.4	59.3	69.5
32	32			64.4	59.3	69.5
33	33			64.4	59.2	69.5
34	34			64.4	59.2	69.5
35	35			64.4	59.3	69.5
36	36			64.4	59.3	69.5
37	37			64.4	59.3	69.6
38	38			64.5	59.3	69.6
39	39			64.5	59.3	69.6
40	40			64.5	59.4	69.6
41	41			64.5	59.4	69.6
42	42			64.5	59.3	69.6
43	43			64.5	59.3	69.6
44	44			64.5	59.3	69.6

This section presents the forecasts, the residuals, and the 100(1 – alpha)% prediction limits.

Forecast and Data Plot

Forecast and Data Plot



This section displays a plot of the data values, the forecasts, and the prediction limits. It lets you determine if the forecasts are reasonable.

Autocorrelations of Residuals

Autocorrelations of Residuals of Intel_Close - MEAN

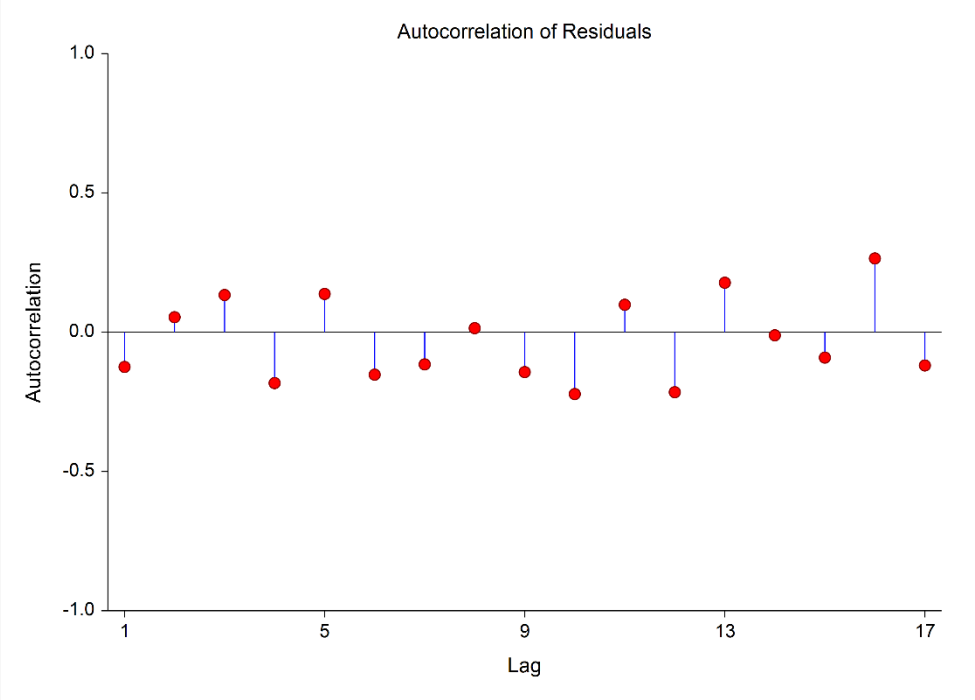
Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	-0.124727	6	-0.152506	11	0.098453	16	0.264817
2	0.053877	7	-0.115625	12	-0.215545	17	-0.119435
3	0.133499	8	0.014388	13	0.177615		
4	-0.182724	9	-0.143282	14	-0.011190		
5	0.137313	10	-0.222115	15	-0.091297		

Significant if |Correlation| > 0.447214

If the residuals are white noise, these autocorrelations should all be nonsignificant. If significance is found in these autocorrelations, the model should be changed.

Autocorrelation Plot

Autocorrelation Plot



This plot is the key diagnostic to determine if the model is adequate. If no pattern can be found here, you can assume that your model is as good as possible and proceed to use the forecasts. If large autocorrelations or a pattern of autocorrelations is found in the residuals, you will have to modify the model.

Portmanteau Test

Portmanteau Test of Intel_Close - MEAN

Lag	DF	Test Statistic Value (Q)	P-Value	Decision at Alpha = 0.05
3	1	0.89	0.344802	Adequate Model
4	2	1.81	0.404407	Adequate Model
5	3	2.36	0.500420	Adequate Model
6	4	3.09	0.542103	Adequate Model
7	5	3.55	0.616258	Adequate Model
8	6	3.55	0.736666	Adequate Model
9	7	4.38	0.735595	Adequate Model
10	8	6.55	0.586230	Adequate Model
11	9	7.02	0.634975	Adequate Model
12	10	9.58	0.478457	Adequate Model
13	11	11.56	0.397699	Adequate Model
14	12	11.57	0.480965	Adequate Model
15	13	12.30	0.503088	Adequate Model
16	14	20.02	0.129650	Adequate Model
17	15	22.11	0.105010	Adequate Model

The Portmanteau Test (sometimes called the Box-Pierce-Ljung statistic) is used to determine if there is any pattern left in the residuals that may be modeled. This is accomplished by testing the significance of the autocorrelations up to a certain lag. In a private communication with Dr. Greta Ljung, we have learned that this test should only be used for lags between 13 and 24. The test is computed as follows:

$$Q(k) = N(N+2)\sum_{j=1}^{k}\frac{r_j^2}{N-j}$$

Q(k) is distributed as a Chi-square with *(K-p-q-P-Q)* degrees of freedom.