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.
Procedure Options
This section describes the options available in this procedure.

Variables Tab
Specify the variable on which to run the analysis.

Time Series Variable
Specify the variable on which to run the analysis.

Use Logarithms
Specifies that the log (base 10) transformation should be applied to the values of the variable.

Missing Values
Choose how missing (blank) values are processed.

The algorithm used in this procedure cannot tolerate missing values since each row is assumed to represent the next point in a time sequence. Hence, when missing values are found, they must be removed either by imputation (filling in with a reasonable value) or by skipping the row and pretending it does not exist.

Whenever possible, we recommend that you replace missing values manually.

Here are the available options.

Average the Adjacent Values
Replace the missing value with the average of the nearest values in the future (below) and in the past (above).

Carry the Previous Value Forward
Replace the missing value with the first non-missing value immediately above (previous) this value.

Omit Row from Calculations
Ignore the row in all calculations. Analyze the data as if the row was not on the database.

Forecasting Options

Number of Forecasts
This option specifies the number of forecasts to be generated.

Data Adjustment Options

Remove Mean
Checking this option indicates that the series average should be subtracted from the data. This is almost always done.

Remove Trend
Checking this option indicates that the least squares trend line should be subtracted from the data. This is sometimes done, although differencing is often used to remove trends instead.

Regular Differencing
Specify the number of times to difference the series. You can enter 0, 1, or 2.
**Seasonal Differencing**
Specify the number of times to seasonally difference the series. The number of seasons per year is specified later.

---

**Seasonality Options**

**Number of Seasons**
Specify the number of seasons per year in the series. Use ‘4’ for quarterly data or ‘12’ for monthly data.

**First Season**
Specify the first season of the series. This value is used to format the reports and plots. For example, if you have monthly data beginning with March, you would enter a ‘3’ here.

**First Year**
Specify the first year of the series. This value is used to format the reports and plots.

---

**ARIMA Model Options**

**Regular AR**
Specify the highest order of the autoregressive parameters. For example, if you specify ‘2’ here, both the lag one and lag two autoregressive parameters will be included in the model.

**Regular MA**
Specify the highest order of the moving average parameters. For example, if you specify ‘2’ here, both the lag one and lag two moving average parameters will be included in the model.

**Seasonal AR**
Specify the highest order of the seasonal autoregressive parameters. The number of seasons per year is specified later.

**Seasonal MA**
Specify the highest order of the seasonal moving average parameters. The number of seasons per year is specified later.

---

**ARIMA Model Options**

**Max Iterations**
The nonlinear estimation procedure will not converge for every model and data combination. This parameter lets you set the maximum number of iterations before the estimation algorithm is terminated.

**Convergence**
As the nonlinear estimation proceeds through each step, the residual sum of squares is calculated. When the ratio of the residual sum of squares from the current step to the residual sum of squares from the last step is less than this amount, the estimation procedure concludes. Hence, decreasing this amount will cause the procedure to go through more iterations, while increasing this amount will cause it to run fewer iterations.

**Lambda**
Lambda is a parameter from Marquart’s nonlinear estimation procedure. This value of lambda was suggested by Marquart and we suggest that you leave it at the default value.
Reports Tab
The following options control which reports are displayed.

Select Additional Reports

Minimization Report - Portmanteau Test Report
Each of these options specifies whether the indicated report is displayed.

Forecast Report
This option specifies which parts of the series are listed on the numeric reports: the original data and forecasts, just the forecasts, or neither.

Alpha Level
The value of alpha for the asymptotic prediction limits of the forecasts. Usually, this number will range from 0.001 to 0.1. A common choice for alpha is 0.05, but this value is a legacy from the age before computers when only printed tables were available. You should determine a value appropriate for your needs.

Report Options
Decimals
Specifies the number of decimal places to use when displaying the forecasts.

Precision
Specify the precision of numbers in the report. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy. Note that all reports are formatted for single precision only.

Variable Names
Specify whether to use variable names or (the longer) variable labels in report headings.

Plots Tab
This section controls the forecast and autocorrelation plots.

Select Plots

Forecast Plot - Autocorrelation Plot
Each of these options specifies whether the indicated plot is displayed. Click the plot format button to change the plot settings.

Plot Options
Large Plots
When checked, the plots displayed are larger (about five inches across) than normal (about two inches across).
Horizontal Axis Variable

Horizontal Variable
This option controls the spacing on the horizontal axis when missing or filtered values occur.
Your choices are

Actual Row Number
Use the actual row number of each row from the dataset along the horizontal axis.

Constructed Date
Construct a date value from the sequence (relative row) number and the Seasonality Options settings. Any missing or filtered values are skipped when forming the sequence number.

Storage Tab

The forecasts, prediction limits, and residuals may be stored on the current database for further analysis. This group of options lets you designate which statistics (if any) should be stored and which variables should receive these statistics. The selected statistics are automatically stored to the current database.

Note that existing data is replaced. Be careful that you do not specify variables that contain important data.

Data Storage Columns

Forecasts, Residuals, Lower Prediction Limits, and Upper Prediction Limits
The forecasts, residuals (Y-forecast), lower 100(1-alpha) prediction limits, and upper 100(1-alpha) prediction limits may be stored in the columns specified here.
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.

You may follow along here by making the appropriate entries or load the completed template Example 1 by clicking on Open Example Template from the File menu of the ARIMA (Box-Jenkins) window.

1. **Open the Intel dataset.**
   - From the File menu of the NCSS Data window, select **Open Example Data**.
   - Click on the file **Intel.NCSS**.
   - Click **Open**.

2. **Open the ARIMA (Box-Jenkins) window.**
   - Using the Analysis menu or the Procedure Navigator, find and select the ARIMA (Box-Jenkins) procedure.
   - On the menus, select File, then New Template. This will fill the procedure with the default template.

3. **Specify the variables.**
   - On the ARIMA (Box-Jenkins) window, select the Variables tab.
   - Double-click in the Time Series Variable box. This will bring up the variable selection window.
   - Select Intel_Close from the list of variables and then click Ok.
   - Enter 2 in the Regular AR box.
   - Enter 0 in the Regular MA box.

4. **Specify the reports.**
   - On the ARIMA (Box-Jenkins) window, select the Reports tab.
   - Select Data and Forecasts in the Forecast Report.

5. **Run the procedure.**
   - From the Run menu, select Run Procedure. Alternatively, just click the green Run button.

### Minimization Phase Section

<table>
<thead>
<tr>
<th>In No.</th>
<th>Error Sum of Squares</th>
<th>Lambda</th>
<th>AR(1)</th>
<th>AR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>85.89011</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>1</td>
<td>22.34787</td>
<td>0.1</td>
<td>0.9479776</td>
<td>-0.1168852</td>
</tr>
<tr>
<td>2</td>
<td>17.8811</td>
<td>0.04</td>
<td>1.292394</td>
<td>-0.4623865</td>
</tr>
<tr>
<td>3</td>
<td>17.5762</td>
<td>0.016</td>
<td>1.390741</td>
<td>-0.589241</td>
</tr>
<tr>
<td>4</td>
<td>17.5736</td>
<td>0.0064</td>
<td>1.403895</td>
<td>-0.5902494</td>
</tr>
<tr>
<td>5</td>
<td>17.57359</td>
<td>0.00256</td>
<td>1.40471</td>
<td>-0.5902494</td>
</tr>
</tbody>
</table>

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 Section**

<table>
<thead>
<tr>
<th>Model Description Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Series</td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Observations</td>
</tr>
<tr>
<td>Missing Values</td>
</tr>
<tr>
<td>Iterations</td>
</tr>
<tr>
<td>Pseudo R-Squared</td>
</tr>
<tr>
<td>Residual Sum of Squares</td>
</tr>
<tr>
<td>Mean Square Error</td>
</tr>
<tr>
<td>Root Mean Square</td>
</tr>
</tbody>
</table>

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 Section

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>T-Value</th>
<th>Prob Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>1.40471</td>
<td>0.2065638</td>
<td>6.8004</td>
<td>0.000000</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-0.5902494</td>
<td>0.2330099</td>
<td>-2.5332</td>
<td>0.011304</td>
</tr>
</tbody>
</table>

**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.

**Prob Level**

The probability level 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 significance, 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

<table>
<thead>
<tr>
<th></th>
<th>AR(1)</th>
<th>AR(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR(1)</td>
<td>1.000000</td>
<td>-0.881734</td>
</tr>
<tr>
<td>AR(2)</td>
<td>-0.881734</td>
<td>1.000000</td>
</tr>
</tbody>
</table>

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.

**Forecast Section**

<table>
<thead>
<tr>
<th>Row</th>
<th>Date</th>
<th>Actual</th>
<th>Residual</th>
<th>Forecast</th>
<th>Lower 95% Limit</th>
<th>Upper 95% Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>65.0</td>
<td>0.1</td>
<td>64.9</td>
<td>61.6</td>
<td>68.2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>65.0</td>
<td>0.0</td>
<td>65.0</td>
<td>61.6</td>
<td>68.3</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>66.9</td>
<td>0.9</td>
<td>60.0</td>
<td>56.6</td>
<td>63.3</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td></td>
<td></td>
<td>62.2</td>
<td>58.9</td>
<td>65.5</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td></td>
<td></td>
<td>63.4</td>
<td>59.1</td>
<td>67.7</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td></td>
<td></td>
<td>64.3</td>
<td>59.5</td>
<td>69.1</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td></td>
<td></td>
<td>64.9</td>
<td>59.9</td>
<td>69.9</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td></td>
<td></td>
<td>65.1</td>
<td>60.1</td>
<td>70.2</td>
</tr>
</tbody>
</table>

This section presents the forecasts, the residuals, and the 100(1-alpha)% prediction limits.
Forecast and Data Plot Section

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 Section

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 Section**

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 Section**

<table>
<thead>
<tr>
<th>Lag</th>
<th>DF</th>
<th>Test Value</th>
<th>Prob</th>
<th>Decision (0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>0.89</td>
<td>0.344802</td>
<td>Adequate Model</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.81</td>
<td>0.404407</td>
<td>Adequate Model</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>2.36</td>
<td>0.500420</td>
<td>Adequate Model</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>3.09</td>
<td>0.542103</td>
<td>Adequate Model</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>3.55</td>
<td>0.616298</td>
<td>Adequate Model</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.