Chapter 474

Automatic ARMA

Introduction

The ARIMA (or Box-Jenkins) method is often used to forecast time series of medium (N over 50) to long lengths. It requires the forecaster to be highly trained in selecting the appropriate model. The procedure discussed here automates the ARIMA forecasting process by having the program select the appropriate model.

The Method

The Automatic ARMA program uses methodology from several authors to find and estimate an appropriate forecasting model. The method may be outlined as follows:

1. Using the model selection theory of Pandit and Wu (1983), any deterministic trend is removed from the series.
2. A set of models of increasing complexity is fit. These are *ARIMA(1,0,0), ARIMA(2,0,1), ARIMA(4,0,3), ARIMA(6,0,5),* and so on, increasing both *p* and *q* by two at each step. The most complex model tried is specified in the Maximum Order box. The residual sum of squares is calculated for each model and the minimum is noted.
3. Using the minimum residual sum of squares as the criterion, the models are again arranged from simplest to most complex. The first model to be within the user-defined percentage of the minimum sum of squares is selected and used.
4. Once this model has been determined, one final attempt is made to find a model of smaller order that is within the specified percentage of the minimum. Suppose the previous steps lead to an *ARIMA(4,3)* model. This step would fit an *ARIMA(3,0,2)* model and check to see if the residual sum of squares was within the specified percentage. If it was, the *ARIMA(3,0,2)* model would be used. If not, the *ARIMA(4,3)* model would be used.

Because the procedure has to fit so many models, several of which are of large order, we use a sub-optimal (but much faster) model estimation algorithm. We chose the least squares modified Yule-Walker technique described in Marple (1987), section 10.4. This method is fast and seems to provide reasonable estimates of the residual sum of squares.

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

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 option should be used if a trend is apparent in the data.

ARIMA Model Options

Maximum Order

The largest number of AR parameters that will be tried. If you are using seasonal data, this should be two more than the length of any seasonal pattern. Hence, for monthly data you would try fourteen, for quarterly data you would use six, and for annual data you would use four or six.

Even-order models are tried up to this size. For example, if you enter a six here, the program will fit the ARIMA models *ARIMA(2,0,1), ARIMA(4,0,3),* and *ARIMA(6,0,5).* The residual sum of squares is noted, and the simplest model is used for forecasting.

Percent of Best

Once the program has found the residual sum of squares for each of the models designated by the Maximum Order, it finds the smallest of these values. It then searches through models, calculating the percent increase in the residual sum of squares of the current model over that of the best model. It selects the simplest (smallest in number of parameters) model that is less than this criterion.

Hence, the larger the percentage you enter here, the simpler will be the model. Normally, the value of five is sufficient.

Autoregressive Terms

When this value is greater than zero, no search is conducted. Instead, a model with this specific autoregressive order is calculated.

Moving Average Terms

When this value is greater than zero, no search is conducted. Instead, a model with this specific moving average order is calculated.

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.

Reports Tab

The following options control which reports are displayed.

Select Additional Reports

Search 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 plot and the autocorrelation plot.

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 dataset 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 dataset.

Note that existing data is replaced. Be careful that you do not specify columns 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 Automatic ARMA Model

This section presents an example of how to fit an Automatic ARMA model. The SeriesA variable in the SeriesA dataset will be fit.

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 Automatic ARMA window.

1 Open the SeriesA dataset.

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

2 Open the Automatic ARMA window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Automatic ARMA** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Automatic ARMA window, select the **Variables tab**.
- Double-click in the **Time Series Variable** box. This will bring up the variable selection window.
- Select **SeriesA** from the list of variables and then click **Ok**.

4 Specify the reports.

- On the Automatic ARMA window, select the **Reports tab**.
- Enter **3** in the **Decimals** box.

5 Run the procedure.

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

Model Search Results Section

Model Search Results Section

No.	AR Order (P)	MA Order (Q)	Sum of Squares	Pseudo R-Squared	Percent Change From Last
0	0	0	0.3124203	0.00	0.00
1	1	0	0.2104635	32.63	-32.63
2	2	1	0.1966642	37.05	-6.56
3	4	3	0.1953752	37.46	-0.66
4	**6**	**5**	**0.1899079**	**39.21**	**-2.80**
5	8	7	0.1827622	41.50	-3.76

This report displays information about the various models that were fit during the search. In this case, we note that the selected model is ARIMA(6,0,5). The individual definitions are as follows:

AR Order (P)

The number of autoregressive parameters in the model.

MA Order (Q)

The number of moving average parameters in the model.

Sum Squares

The sum of the squared residuals. The smaller this amount, the better the precision of the model.

Psuedo 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.

Percent Change From Last

The percent change in the sum of squares from model immediately above.

Model Description Section

Model Description Section

Series	SERIESA-MEAN	R-Squared	39.213981
Observations	197	Sum Squares Error	0.1899079
Mean	1.706244	Mean Square Error	1.02101E-03
Selected Model	ARMA(6,5)	Root Mean Square	3.195325E-02
Missing Values	None		

This report displays summary information about the solution.

Series

The name of the variable being analyzed.

Observations

The number of observations (rows) in the series.

Trend Equation

The trend equation that was fit and removed from the series before the ARMA models were fit.

Selected Model

The phrase *ARMA (p,q)* gives the highest order of the regular ARMA parameters.

- *p* Number of autoregression parameters in the model.
- *q* Number of moving average parameters in the model.

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.

Sum of Squares Error

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

Model Estimation Section

Parameter Name	Parameter Estimate
AR(1)	0.3655652
AR(2)	0.1581511
AR(3)	0.0183087
AR(4)	3.503909E-02
AR(5)	1.653267E-02
AR(6)	0.1440598
MA(1)	1.811239E-02
MA(2)	-5.549401E-02
MA(3)	4.643534E-03
MA(4)	2.481859E-03
MA(5)	-2.905689E-02

Parameter Name

This 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.

Parameter Estimate

This is the estimated parameter value.

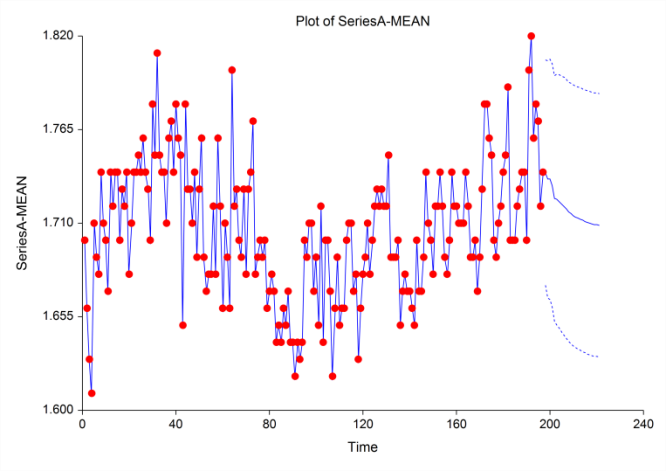
Forecast Section

Forcast Section of SeriesA

Row	Date	Forecast	Lower 95% Limit	Upper 95% Limit
198	17 6	1.740	1.673	1.806
199	17 7	1.736	1.666	1.805
200	17 8	1.736	1.665	1.807
201	17 9	1.732	1.661	1.804
202	17 10	1.725	1.652	1.797
203	17 11	1.724	1.650	1.798
204	17 12	1.722	1.648	1.797
205	18 1	1.721	1.646	1.796
206	18 2	1.720	1.644	1.796
207	18 3	1.718	1.642	1.795
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

This section presents the forecasts 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

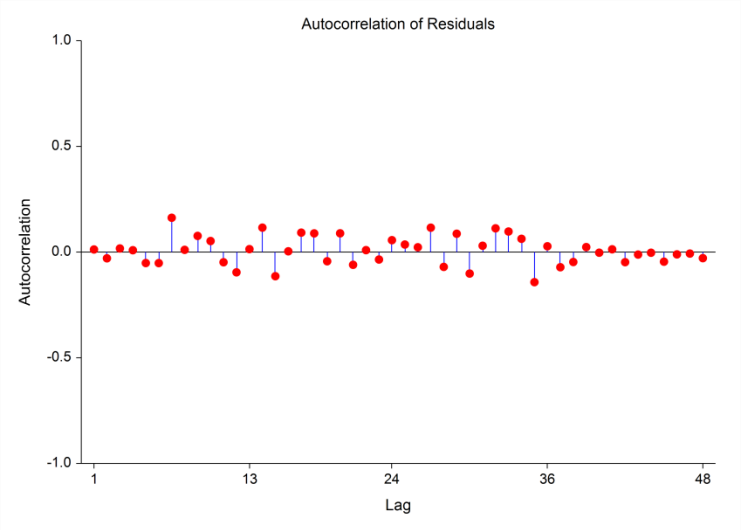
Autocorrelations of Residuals of SeriesA-MEAN

Lag	Correlation	Lag	Correlation	Lag	Correlation	Lag	Correlation
1	0.012330	13	0.013916	25	0.035760	37	-0.071720
2	-0.029505	14	0.115632	26	0.022786	38	-0.047004
3	0.017103	15	-0.114148	27	0.115409	39	0.023569
4	0.008808	16	0.003723	28	-0.070129	40	-0.002934
5	-0.051858	17	0.091939	29	0.086803	41	0.013134
6	-0.052182	18	0.088479	30	-0.101699	42	-0.047757
7	0.162349	19	-0.043432	31	0.029961	43	-0.011785
8	0.010603	20	0.088817	32	0.112346	44	-0.003118
9	0.076537	21	-0.060232	33	0.097306	45	-0.045457
10	0.052340	22	0.009222	34	0.062783	46	-0.010849
11	-0.048122	23	-0.035292	35	-0.142371	47	-0.007085
12	-0.095803	24	0.056370	36	0.027464	48	-0.028750

Significant if |Correlation|> 0.142494

If the residuals are white noise, these autocorrelations should all be non-significant. 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

Portmanteau Test Section SeriesA-MEAN

Lag	DF	Portmanteau Test Value	Prob Level	Decision (0.05)
12	1	11.08	0.000873	Inadequate Model
13	2	11.12	0.003849	Inadequate Model
14	3	13.98	0.002927	Inadequate Model
15	4	16.79	0.002122	Inadequate Model
16	5	16.79	0.004908	Inadequate Model
17	6	18.63	0.004827	Inadequate Model
18	7	20.35	0.004862	Inadequate Model
19	8	20.76	0.007799	Inadequate Model
20	9	22.51	0.007390	Inadequate Model
21	10	23.32	0.009624	Inadequate Model
22	11	23.34	0.015825	Inadequate Model
23	12	23.62	0.022901	Inadequate Model
24	13	24.34	0.028143	Inadequate 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)* degrees of freedom.