## 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.

# 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.

## Setup

To run this example, complete the following steps:

**1 Open the SeriesA example dataset**

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

**2 Specify the Automatic ARMA procedure options**

- Find and open the **Automatic ARMA** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

| Option | Value |
|---|---|
| **Variables Tab** | |
| Time Series Variable | **SeriesA** |
| **Reports Tab** | |
| Decimals | **3** |

**3 Run the procedure**

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

## 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 | 0.00102101 |
| Selected Model | ARMA(6,5) | Root Mean Square | 0.03195325 |
| 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) | 0.03503909 |
| AR(5) | 0.01653267 |
| AR(6) | 0.1440598 |
| MA(1) | 0.01811239 |
| MA(2) | -0.05549401 |
| MA(3) | 0.004643534 |
| MA(4) | 0.002481859 |
| MA(5) | -0.02905689 |

### 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

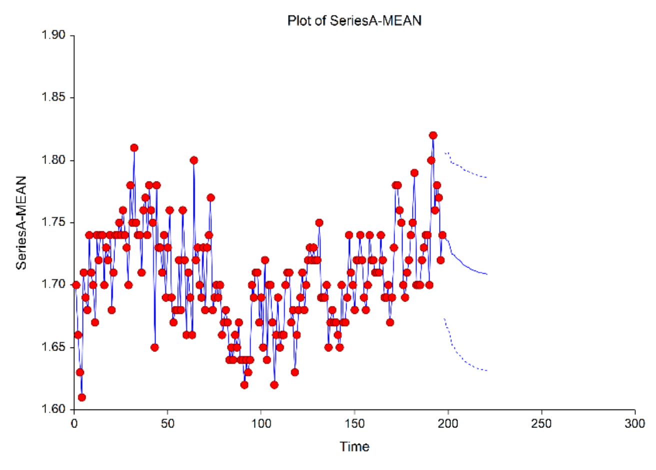
**Forecast Section of SeriesA**

| Row | Date | Forecast | Lower 95% Limit | Upper 95% Limit |
|---|---|---|---|---|
| 198 | 198 | 1.740 | 1.673 | 1.806 |
| 199 | 199 | 1.736 | 1.666 | 1.805 |
| 200 | 200 | 1.736 | 1.665 | 1.807 |
| 201 | 201 | 1.732 | 1.661 | 1.804 |
| 202 | 202 | 1.725 | 1.652 | 1.797 |
| 203 | 203 | 1.724 | 1.650 | 1.798 |
| 204 | 204 | 1.722 | 1.648 | 1.797 |
| 205 | 205 | 1.721 | 1.646 | 1.796 |
| 206 | 206 | 1.720 | 1.644 | 1.796 |
| 207 | 207 | 1.718 | 1.642 | 1.795 |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |

This section presents the forecasts and the 100(1-alpha)% prediction limits.

# Forecast and Data Plot Section

**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 Section

**Autocorrelations of Residuals of SeriesA-MEAN**

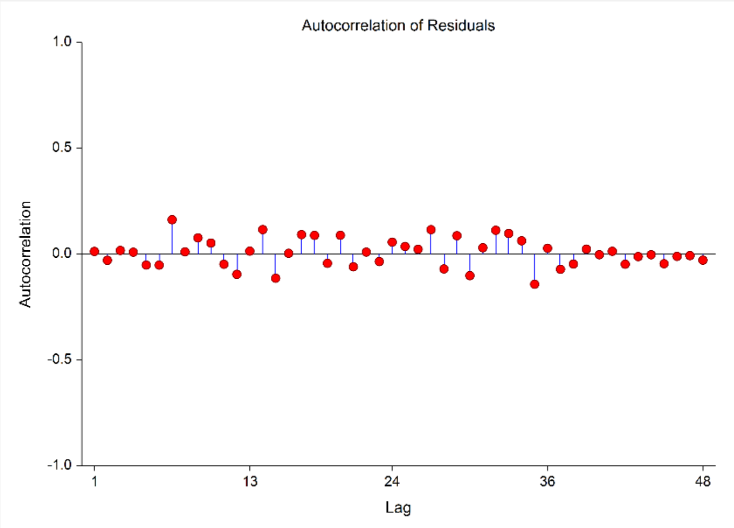
| Lag | Correlation | Lag | Correlation | Lag | Correlation | Lag | Correlation |
|---|---|---|---|---|---|---|---|
| 1 | 0.012330 | 13 | 0.013916 | 25 | 0.035760 | 37 | -0.071720 |
| 2 | -0.029505 | 14 | 0.115632 | 26 | 0.022786 | 38 | -0.047004 |
| 3 | 0.017103 | 15 | -0.114148 | 27 | 0.115409 | 39 | 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

**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.