Forecast – Response Variable

When the values in a response column are ordered sequentially over time, it is often of interest to forecast their behavior beyond the end of the data. This procedure fits a parametric ARIMA time series model to a sample of data and forecasts its future behavior.

The data for this analysis consist of *n* sequentially ordered observations. Let

x_t = observation at time *t*, *t* = 1, 2, …, *n*
n = number of observations

Access

Highlight: one *Response* column. A *Time index* column may also be selected to position the points along the X axis.

Select: *Forecast* from the main menu.

Output Page 1: Forecasts from the fitted model.

Output Page 2: Autocorrelation function (ACF) for the residuals from the fitted model.

Output Page 3: Plot of the residuals.

Options

An ARIMA time series model is used to forecast the data. The general form of this model is most easily expressed in terms of the backwards operator *B*, which operates on the time index of a data value such that $B^j Y_t = Y_{t-j}$. Using this operator, the model takes the form

$$\left(1 - B - B^2 - ... - B^p\right)\left(1 - B^s - B^{2s} - ... - B^{Ps}\right)(1 - B)^d (1 - B^s)^D Z_t$$

$$= \left(1 - B - B^2 - ... - B^q\right)\left(1 - B^s - B^{2s} - ... - B^{Qs}\right)a_t \qquad (1)$$

where

$$Z_t = Y_t - \mu \qquad (2)$$

and a_t is a random error or shock to the system at time *t*, usually assumed to be random observations from a normal distribution with mean 0 and standard deviation σ_a. For a stationary series, μ represents the process mean. Otherwise, it is related to the slope of the forecast function. μ is sometimes assumed to equal 0.

The above model is often referred to as an ARIMA(p,d,q)x(P,D,Q)s model. It consists of several terms:

1. A nonseasonal autoregressive term of order p.
2. Nonseasonal differencing of order d.
3. A nonseasonal moving average term of order q.
4. A seasonal autoregressive term of order P
5. Seasonal differencing of order D.
6. A seasonal moving average term of order Q.

While the general model looks formidable, the most commonly used models are relatively simple special cases. These include:

AR(1) – autoregressive of order 1

The observation at time t is expressed as a mean plus a multiple of the deviation from the mean at the previous time period plus a random shock:

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + a_t \tag{3}$$

AR(2) – autoregressive of order 2

The observation at time t is expressed as a mean plus multiples of the deviations from the mean at the 2 previous time periods plus a random shock:

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + a_t \tag{4}$$

MA(1) – moving average of order 1

The observation at time t is expressed as a mean plus a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t = \mu + a_t - \theta_1 a_{t-1} \tag{5}$$

MA(2) – moving average of order 2

The observation at time t is expressed as a mean plus a random shock at the current time period plus multiples of the random shocks at the 2 previous time periods:

$$Y_t = \mu + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \tag{6}$$

ARMA(1,1) – mixed model with 2 first order terms

The observation at time t is expressed as a mean plus a multiple of the deviation from the mean at the previous time period plus a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + a_t - \theta_1 a_{t-1} \tag{7}$$

ARIMA(0,1,1) – moving average of order 1 applied to the first differences
The difference between the current period and the previous period is expressed as a random shock at the current time period plus a multiple of the random shock at the previous time period:

$$Y_t - Y_{t-1} = a_t - \theta_1 a_{t-1} \quad (8)$$

It can be shown that this model is equivalent to a *Simple Exponential Smoothing* model.

ARIMA(0,1,1) with constant – moving average of order 1 applied to the first differences with a constant
This model adds a constant to the model defined above:

$$Y_t - Y_{t-1} = \mu + a_t - \theta_1 a_{t-1} \quad (9)$$

The addition of a constant causes the forecasts to follow a linear trend.

ARIMA(0,2,2) – moving average of order 2 applied to the second differences
The difference of the first differences is expressed as a random shock at the current time period plus multiples of the random shocks at the 2 previous time periods:

$$(Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \quad (10)$$

This model is equivalent to *Holt's Linear Exponential Smoothing* model.

ARIMA(0,1,1)x(0,1,1)s – seasonal and nonseasonal MA terms of order 1
The observation at time *t* is expressed as a combination of the observation one season ago plus the difference between the observation last period and its counterpart one season ago plus multiple of the shocks to hit the system this period, last period, and two periods one season ago:

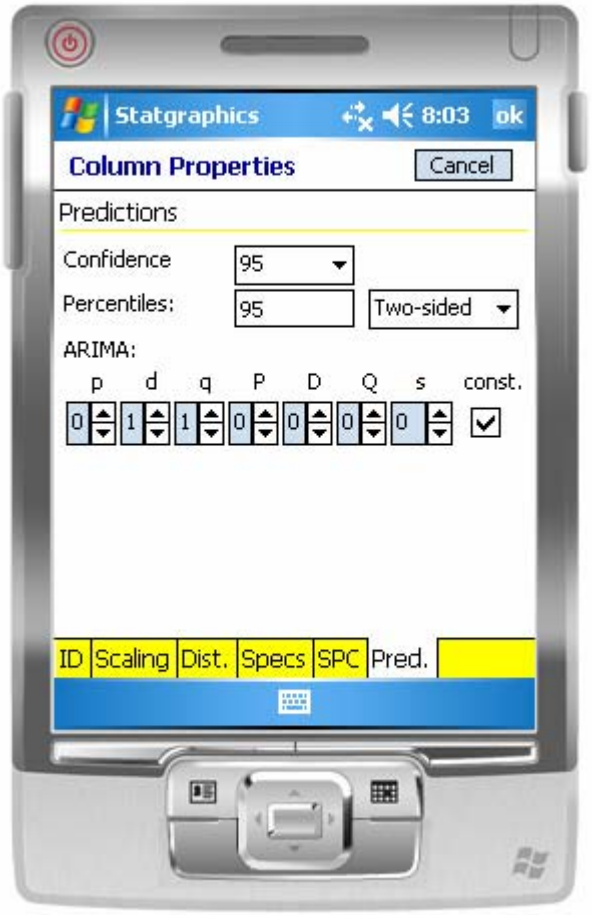
$$Y_t = Y_{t-s} + Y_{t-1} - Y_{t-s-1} + a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-s} + \theta_1 \Theta_1 a_{t-s-1} \quad (11)$$

Many economic time series with a seasonal component can be well represented by this model.

The type of ARIMA model to be fit is specified by:

1. Accessing the *Properties* dialog box for the *Response* column containing the time series data.

2. On the *Pred.* tab, select the desired model.



Specify the order for each term in the model, and the length of seasonality *s* if the data is seasonal. Also specify in the *Percentiles* field the type of forecast limits desired (percentage and type).

It is sometimes helpful to transform the data using a logarithm or some other power transformation. To specify a transformation:

1. Access the *Properties* dialog box column containing the time series data by double-clicking on the column header.

2. On the *Dist.* tab, select the assumed distribution. The default selection assumes that the data with no transformation follow a normal distribution. If you select *Lognormal*, the logarithms of the data will be assumed to follow a normal distribution. If you select *Power normal*, the data will be assumed to follow a normal distribution after raising them to the indicated *power*. When forecasting the data, the observations are transformed to the metric in which they are normally distributed, forecasts and forecast limits are calculated in the transformed metric, and the inverse transformation is made to the forecasts to put them back in their original units.

Sample Data

The file *price.sgm* contains a time series of $n = 62$ observations, spaced 15 minutes apart. The first several observations are shown below:

Row	*Time*	*Price*
1	0:00	444.5
2	0:15	445
3	0:30	446.1
4	0:45	447.1
5	1:00	446.7
6	1:15	446.2
7	1:30	445.6
8	1:45	446.4
9	2:00	446.3
10	2;15	446.5

Forecasts

The initial page displays the forecasts generated from the fitted model.

It plots:

1. Point symbols for the observations used to fit the model.

2. A line through the observations showing the one-ahead forecasts. The one-ahead forecasts are the forecasts for each time period made using all of the information available at the previous time period. If

   $F_t(k)$ = forecast for time $t+k$ made at time t

   then the one-ahead forecast error is $F_t(1)$.

3. Forecasts $F_n(k)$ for several periods beyond the last data value.

4. Predictions limits for the forecasts. These limits are created at the *Percentage* level specified on the *Pred.* tab of the *Column Properties* dialog box.

Below the graph are the forecasted values for the next 4 periods beyond the end of the data, and summary statistics showing how well the model was able to predict the observed values. The statistics are from the one-ahead forecast errors e_t in each period, which equal

$$e_t = Y_t - F_{t-1}(1) \tag{12}$$

Depending on the type of model selected, some number of initial values m must be determined from the data before forecasts can be generated. The one-ahead forecast errors are thus defined for $t = m+1, m+2, \ldots, n$. The error statistics displayed are:

- Root mean squared error

$$RMSE = \sqrt{\frac{\sum_{t=m+1}^{n} e_t^2}{n-m}} \tag{13}$$

- Mean error

$$ME = \frac{\sum_{t=m+1}^{n} e_t}{n-m} \tag{14}$$

- Mean absolute percentage error

$$MAPE = 100\left(\frac{\sum_{t=m+1}^{n} |e_t / Y_t|}{n-m}\right)\% \tag{15}$$

- Mean percentage error

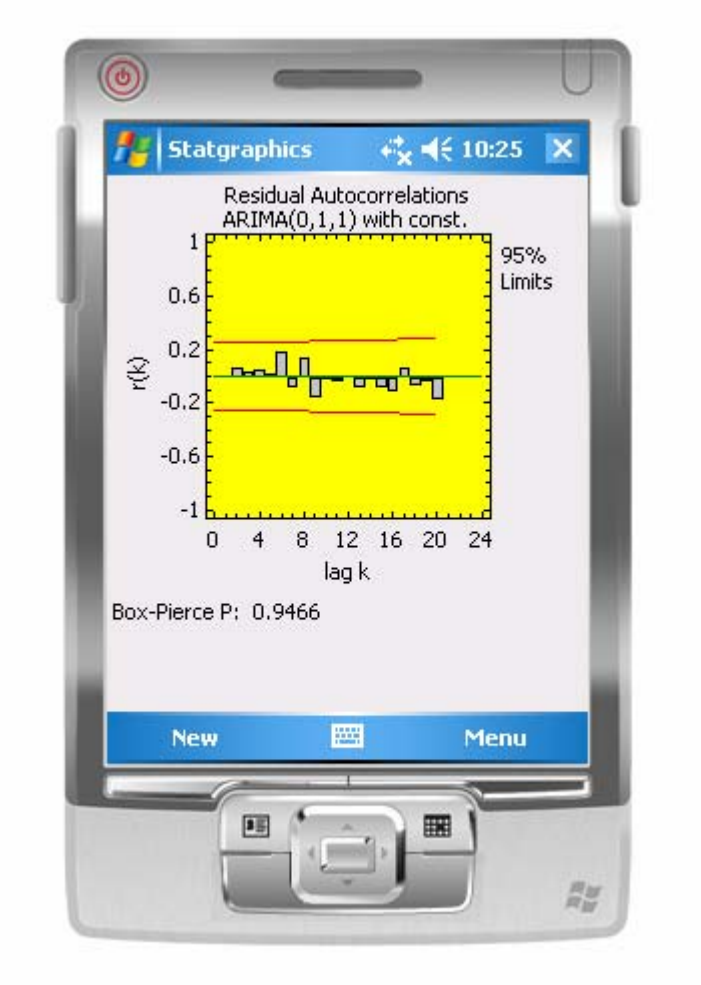
$$MPE = 100\left(\frac{\sum_{t=m+1}^{n} (e_t / Y_t)}{n-m}\right)\% \tag{16}$$

The MAPE and MPE are only displayed if all data values are positive.

If the example above, the MAPE indicates that the error in forecasting the next data value averages about 1%.

ACF (Autocorrelation Function)

The procedure also plots the estimated autocorrelation function of the one-ahead forecast errors.



The autocorrelation at lag *k*, defined by

$$r_k = \frac{\sum_{t=m+1}^{n-k}(e_t - \bar{e})(e_{t+k} - \bar{e})}{\sum_{t=m+1}^{n}(e_t - \bar{e})^2}, \qquad k = 1, 2, \ldots, w \tag{17}$$

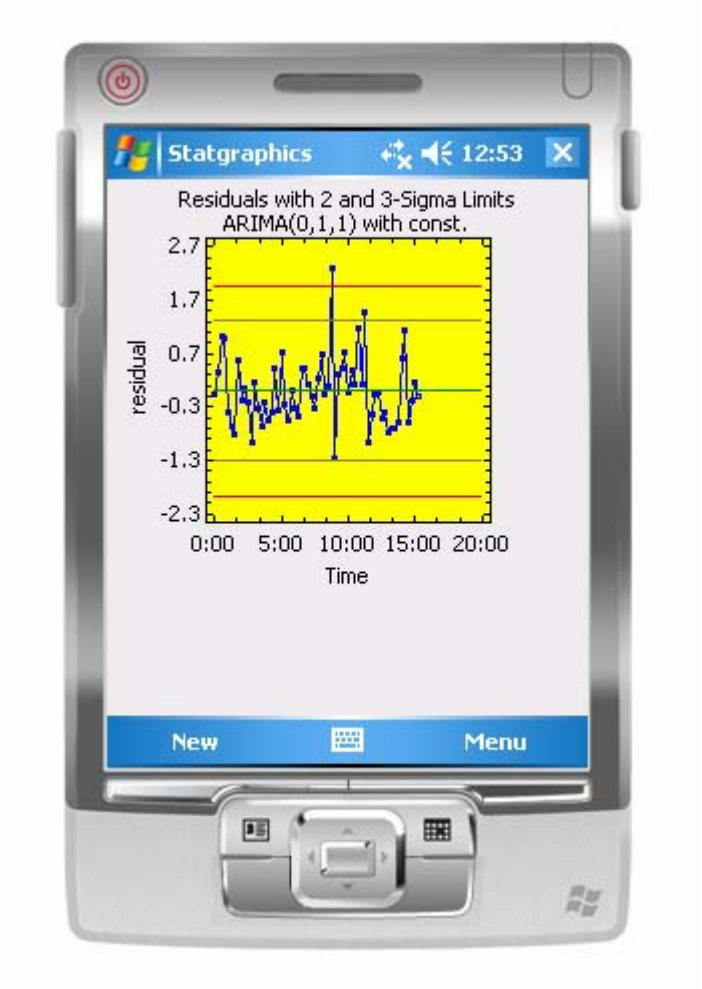
estimates the correlation between forecast errors *k* units apart. If the model fits the data well, then all of the bars should fall within the probability limits, which are drawn as horizontal lines around 0. In addition, a P-value is displayed for the Box-Pierce statistic

$$Q = n\sum_{i=1}^{w} r_i^2 \tag{18}$$

where *w* is the number of autocorrelations that have been estimated. A small P-value (below 0.05) would indicate significant autocorrelation remaining in the errors, which might necessitate using a different ARIMA model.

Residuals

This plot shows the one-ahead forecast errors.



Horizontal lines are drawn at 2-sigma and 3-sigma to help in identifying outliers.

Missing Values

ARIMA models require that data be available at equally spaced intervals of time. Consequently, it is assumed that the time or distance between each row is the same (i.e., 15 minutes, one day, one business day, one inch, etc.) Consequently, special techniques are needed to handle missing data.

The rules for handling missing data are as follows:

- Missing values at the top of the column are ignored.
- Missing values at the end of the column are treated as points for which forecasts are desired.
- Missing values in the middle of the column are replaced with interpolated values according to the following rule, as long as there are not too many missing values close together:

    1. If y_t, the observation at time t, is missing, find the two observations in the same season that precede time t (y_{t-s} and y_{t-2s}) and the two observations in the same season that come after time t (y_{t+s} and y_{t+2s}).

    2. If none of the four observations are missing, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 12y_{t-s} + 12y_{t+s} - 3y_{t+2s}}{18} \tag{19}$$

    3. If y_{t+2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-y_{t-2s} + 3y_{t-s} + y_{t+s}}{3} \tag{20}$$

    4. If y_{t+s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 8y_{t-s} + y_{t+s}}{6} \tag{21}$$

    5. If y_{t-s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 8y_{t+s} - 3y_{t+2s}}{6} \tag{22}$$

    6. If y_{t-2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + 3y_{t+s} - y_{t+s}}{3} \tag{23}$$

    7. If y_{t+s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = -y_{t-2s} + 2y_{t-s} \quad (24)$$

8. If y_{t-s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 2y_{t+s}}{3} \quad (25)$$

9. If y_{t-s} and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + y_{t+2s}}{2} \quad (26)$$

10. If y_{t-2s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + y_{t+s}}{2} \quad (27)$$

11. If y_{t-2s} and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{2y_{t-s} + y_{t+2s}}{3} \quad (28)$$

12. If y_{t-2s} and y_{t-s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = 2y_{t+s} - y_{t+2s} \quad (29)$$

If more than 2 of the four observations are missing, an error message will be displayed and the analysis will not be performed.

The interpolated values are designed to perfectly reproduce a quadratic trend (if only one observation is missing) or a linear trend (is two observations are missing), provided no noise is present.