Box-Jenkins ARIMA Advanced Time Series

www.realoptionsvaluation.com

ROV Technical Papers Series: Volume 25

In This Issue

1. Learn about *Risk Simulator's* ARIMA and Auto ARIMA modules.
2. Find out why an ARIMA model is superior to common time-series analysis and multivariate regressions.

"How is ARIMA forecasting different from multivariate regression?"

Contact Us

Real Options Valuation, Inc.

4101F Dublin Blvd., Ste. 425,
Dublin, California 94568 U.S.A.

admin@realoptionsvaluation.com
www.realoptionsvaluation.com
www.rovusa.com

Theory

One very powerful advanced times-series forecasting tool is the ARIMA or *Auto-Regressive Integrated Moving Average* approach, which assembles three separate tools into a comprehensive model. The first tool segment is the autoregressive or "AR" term, which corresponds to the number of lagged value of the residual in the unconditional forecast model. In essence, the model captures the historical variation of actual data to a forecasting model and uses this variation or residual to create a better predicting model. The second tool segment is the integration order or the "I" term. This integration term corresponds to the number of differencing the time series to be forecasted goes through to make the data stationary. This element accounts for any nonlinear growth rates existing in the data. The third tool segment is the moving average or "MA" term, which is essentially the moving average of lagged forecast errors. By incorporating this lagged forecast errors component, the model in essence learns from its forecast errors or mistakes and corrects for them through a moving average calculation.

The ARIMA model follows the Box-Jenkins methodology with each term representing steps taken in the model construction until only random noise remains. Also, ARIMA modeling uses correlation techniques in generating forecasts. ARIMA can be used to model patterns that may not be visible in plotted data. In addition, ARIMA models can be mixed with exogenous variables, but you must make sure that the exogenous variables have enough data points to cover the additional number of periods to forecast. Finally, be aware that ARIMA cannot and should not be used to forecast stochastic processes or time-series data that are stochastic in nature—use the *Stochastic Process* module to forecast instead.

There are many reasons why an ARIMA model is superior to common time-series analysis and multivariate regressions. The usual finding in time-series analysis and multivariate regression is that the error residuals are correlated with their own lagged values. This serial correlation violates the standard assumption of regression theory that disturbances are not correlated with other disturbances. The primary problems associated with serial correlation are:

- Regression analysis and basic time-series analysis are no longer efficient among the different linear estimators. However, as the error residuals can help to predict current error residuals, we can take advantage of this information to form a better prediction of the dependent variable using ARIMA.
- Standard errors computed using the regression and time-series formula are not correct and are generally understated. If there are lagged dependent variables set as the regressors, regression estimates are biased and inconsistent but can be fixed using ARIMA.

Autoregressive Integrated Moving Average, or ARIMA(p,d,q), models are the extension of the AR model that uses three components for modeling the serial correlation in the time-series data. As previously noted, the first component is the autoregressive (AR) term. The AR(p) model uses the p lags of the time series in the equation. An AR(p) model has the form: $y_t = a_1 y_{t-1} + \ldots + a_p y_{t-p} + e_t$. The second component is the integration (d) order term. Each integration order corresponds to differencing the time series. I(1) means differencing the data once; I(d) means differencing the data d times. The third component is the moving average (MA) term. The MA(q) model uses the q lags of the forecast errors to improve the forecast. An MA(q) model has the form: $y_t = e_t + b_1 e_{t-1} + ... + b_q e_{t-q}$. Finally, an ARMA(p,q) model has the combined form: $y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + e_t + b_1 e_{t-1} + ... + b_q e_{t-q}$.

Procedure

- Start Excel and enter your data or open an existing worksheet with historical data to forecast (Figure 1 uses the example file *Time-Series Forecasting*).
- Click on *Risk Simulator | Forecasting | ARIMA* and select the time-series data.
- Enter the relevant *p*, *d*, and *q* parameters (positive integers only) and enter the number of forecast periods desired, and click *OK*.

Results Interpretation

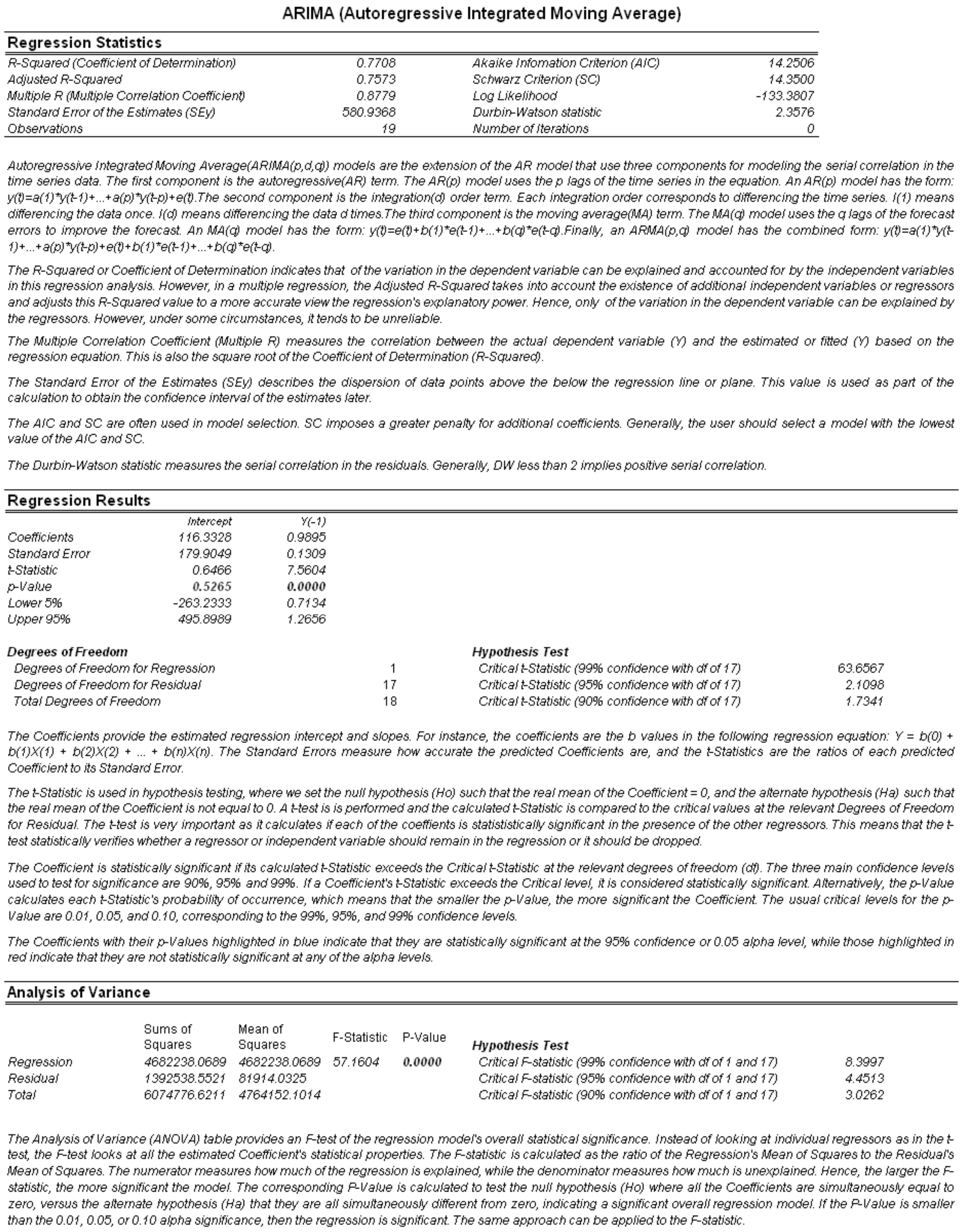
In interpreting the results of an ARIMA model, most of the specifications are identical to the multivariate regression analysis (see Chapter 9, Using the Past to Predict the Future, in *Modeling Risk*, Second Edition, for more technical details about interpreting the multivariate regression analysis and ARIMA models). However, there are several additional sets of results specific to the ARIMA analysis as seen in Figure 1. The first is the addition of Akaike Information Criterion (AIC) and Schwarz Criterion (SC), which are often used in ARIMA model selection and identification. That is, AIC and SC are used to determine if a particular model with a specific set of p, d, and q parameters is a good statistical fit. SC imposes a greater penalty for additional coefficients than the AIC, but generally the model with the lowest AIC and SC values should be chosen. Finally, an additional set of results called the autocorrelation (AC) and partial autocorrelation (PAC) statistics are provided in the ARIMA report.

For instance, if autocorrelation AC(1) is nonzero, it means that the series is first-order serially correlated. If AC dies off more or less geometrically with increasing lags, it implies that the series follows a low-order autoregressive process. If AC drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. In contrast, PAC measures the correlation of values that are *k* periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than *k*, then the partial autocorrelation at lag *k* will be close to zero. The Ljung-Box Q-statistics and their p-values at lag *k* are also provided, where the null hypothesis being tested is such that there is no autocorrelation up to order *k*. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at approximately the 5% significance level. Finding the right ARIMA model takes practice and experience. These AC, PAC, SC, and AIC elements are highly useful diagnostic tools to help identify the correct model specification. Finally, the ARIMA parameter results are obtained using sophisticated optimization and iterative algorithms, which means that although the functional forms look like those of a multivariate regression, they are not the same. ARIMA is a much more computationally intensive and advanced econometric approach.

Auto ARIMA (Box–Jenkins ARIMA Advanced Time-Series)

Theory

This tool provides analyses identical to the ARIMA module except that the Auto-ARIMA module automates some of the traditional ARIMA modeling by automatically testing multiple permutations of model specifications and returns the best-fitting model. Running the *Auto-ARIMA* module is similar to running regular ARIMA forecasts. The differences being that the *p, d, q* inputs are no longer required and that different combinations of these inputs are automatically run and compared.



ARIMA (Autoregressive Integrated Moving Average)

Regression Statistics

R-Squared (Coefficient of Determination)	0.7708	Akaike Infomation Criterion (AIC)	14.2506
Adjusted R-Squared	0.7573	Schwarz Criterion (SC)	14.3500
Multiple R (Multiple Correlation Coefficient)	0.8779	Log Likelihood	-133.3807
Standard Error of the Estimates (SEy)	580.9368	Durbin-Watson statistic	2.3576
Observations	19	Number of Iterations	0

*Autoregressive Integrated Moving Average(ARIMA(p,d,q)) models are the extension of the AR model that use three components for modeling the serial correlation in the time series data. The first component is the autoregressive(AR) term. The AR(p) model uses the p lags of the time series in the equation. An AR(p) model has the form: y(t)=a(1)*y(t-1)+...+a(p)*y(t-p)+e(t).The second component is the integration(d) order term. Each integration order corresponds to differencing the time series. I(1) means differencing the data once. I(d) means differencing the data d times.The third component is the moving average(MA) term. The MA(q) model uses the q lags of the forecast errors to improve the forecast. An MA(q) model has the form: y(t)=e(t)+b(1)*e(t-1)+...+b(q)*e(t-q).Finally, an ARMA(p,q) model has the combined form: y(t)=a(1)*y(t-1)+...+a(p)*y(t-p)+e(t)+b(1)*e(t-1)+...+b(q)*e(t-q).*

The R-Squared or Coefficient of Determination indicates that of the variation in the dependent variable can be explained and accounted for by the independent variables in this regression analysis. However, in a multiple regression, the Adjusted R-Squared takes into account the existence of additional independent variables or regressors and adjusts this R-Squared value to a more accurate view the regression's explanatory power. Hence, only of the variation in the dependent variable can be explained by the regressors. However, under some circumstances, it tends to be unreliable.

The Multiple Correlation Coefficient (Multiple R) measures the correlation between the actual dependent variable (Y) and the estimated or fitted (Y) based on the regression equation. This is also the square root of the Coefficient of Determination (R-Squared).

The Standard Error of the Estimates (SEy) describes the dispersion of data points above the below the regression line or plane. This value is used as part of the calculation to obtain the confidence interval of the estimates later.

The AIC and SC are often used in model selection. SC imposes a greater penalty for additional coefficients. Generally, the user should select a model with the lowest value of the AIC and SC.

The Durbin-Watson statistic measures the serial correlation in the residuals. Generally, DW less than 2 implies positive serial correlation.

Regression Results

	Intercept	Y(-1)
Coefficients	116.3328	0.9895
Standard Error	179.9049	0.1309
t-Statistic	0.6466	7.5604
p-Value	**0.5265**	**0.0000**
Lower 5%	-263.2333	0.7134
Upper 95%	495.8989	1.2656

Degrees of Freedom		**Hypothesis Test**	
Degrees of Freedom for Regression	1	Critical t-Statistic (99% confidence with df of 17)	63.6567
Degrees of Freedom for Residual	17	Critical t-Statistic (95% confidence with df of 17)	2.1098
Total Degrees of Freedom	18	Critical t-Statistic (90% confidence with df of 17)	1.7341

The Coefficients provide the estimated regression intercept and slopes. For instance, the coefficients are the b values in the following regression equation: Y = b(0) + b(1)X(1) + b(2)X(2) + ... + b(n)X(n). The Standard Errors measure how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error.

The t-Statistic is used in hypothesis testing, where we set the null hypothesis (Ho) such that the real mean of the Coefficient = 0, and the alternate hypothesis (Ha) such that the real mean of the Coefficient is not equal to 0. A t-test is is performed and the calculated t-Statistic is compared to the critical values at the relevant Degrees of Freedom for Residual. The t-test is very important as it calculates if each of the coeffients is statististically significant in the presence of the other regressors. This means that the t-test statistically verifies whether a regressor or independent variable should remain in the regression or it should be dropped.

The Coefficient is statistically significant if its calculated t-Statistic exceeds the Critical t-Statistic at the relevant degrees of freedom (df). The three main confidence levels used to test for significance are 90%, 95% and 99%. If a Coefficient's t-Statistic exceeds the Critical level, it is considered statistically significant. Alternatively, the p-Value calculates each t-Statistic's probability of occurrence, which means that the smaller the p-Value, the more significant the Coefficient. The usual critical levels for the p-Value are 0.01, 0.05, and 0.10, corresponding to the 99%, 95%, and 99% confidence levels.

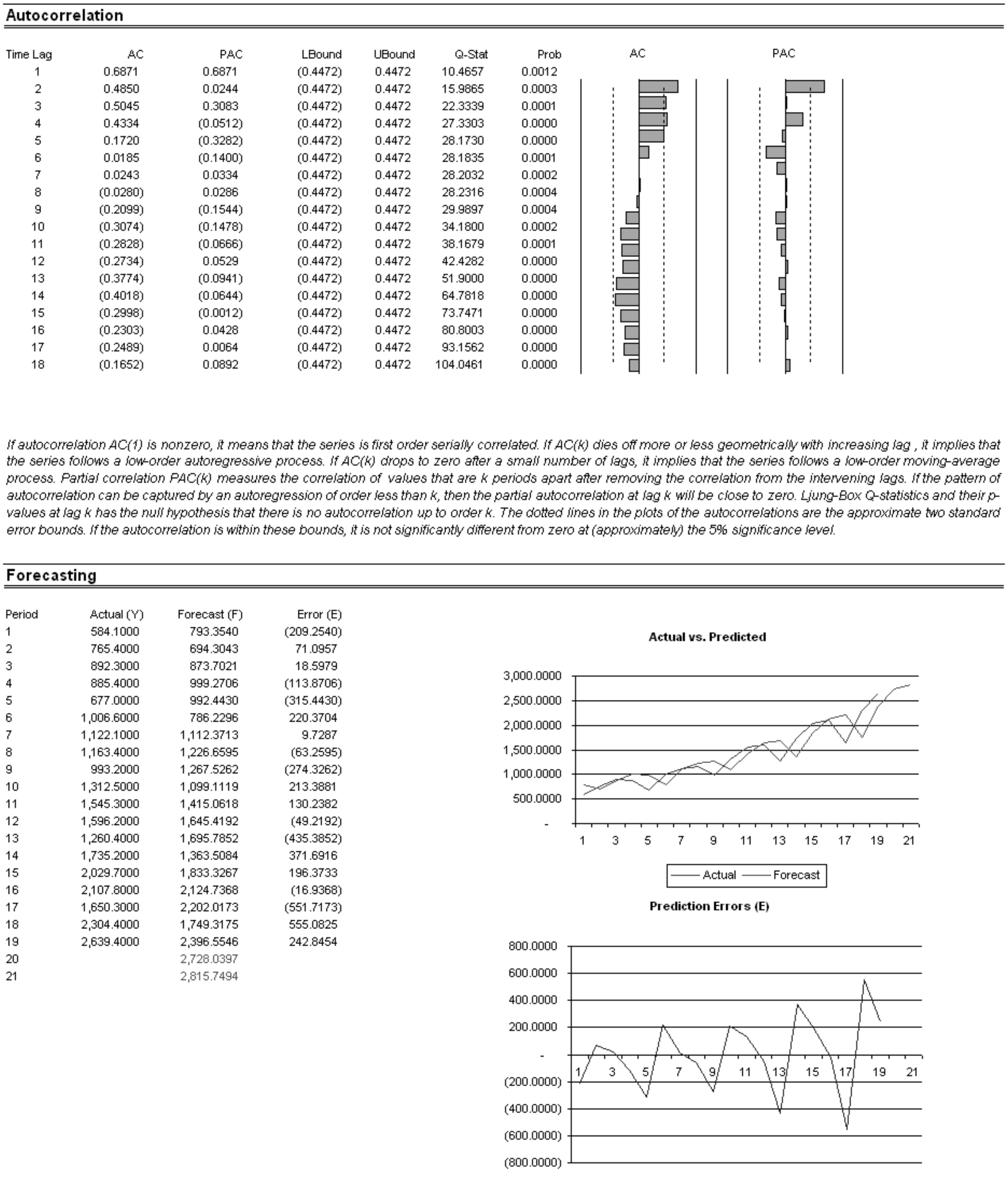
The Coefficients with their p-Values highlighted in blue indicate that they are statistically significant at the 95% confidence or 0.05 alpha level, while those highlighted in red indicate that they are not statistically significant at any of the alpha levels.

Analysis of Variance

	Sums of Squares	Mean of Squares	F-Statistic	P-Value	**Hypothesis Test**	
Regression	4682238.0689	4682238.0689	57.1604	**0.0000**	Critical F-statistic (99% confidence with df of 1 and 17)	8.3997
Residual	1392538.5521	81914.0325			Critical F-statistic (95% confidence with df of 1 and 17)	4.4513
Total	6074776.6211	4764152.1014			Critical F-statistic (90% confidence with df of 1 and 17)	3.0262

The Analysis of Variance (ANOVA) table provides an F-test of the regression model's overall statistical significance. Instead of looking at individual regressors as in the t-test, the F-test looks at all the estimated Coefficient's statistical properties. The F-statistic is calculated as the ratio of the Regression's Mean of Squares to the Residual's Mean of Squares. The numerator measures how much of the regression is explained, while the denominator measures how much is unexplained. Hence, the larger the F-statistic, the more significant the model. The corresponding P-Value is calculated to test the null hypothesis (Ho) where all the Coefficients are simultaneously equal to zero, versus the alternate hypothesis (Ha) that they are all simultaneously different from zero, indicating a significant overall regression model. If the P-Value is smaller than the 0.01, 0.05, or 0.10 alpha significance, then the regression is significant. The same approach can be applied to the F-statistic.

Figure 1. Box Jenkins ARIMA Forecast Report (*continues*)



Autocorrelation

Time Lag	AC	PAC	LBound	UBound	Q-Stat	Prob
1	0.6871	0.6871	(0.4472)	0.4472	10.4657	0.0012
2	0.4850	0.0244	(0.4472)	0.4472	15.9865	0.0003
3	0.5045	0.3083	(0.4472)	0.4472	22.3339	0.0001
4	0.4334	(0.0512)	(0.4472)	0.4472	27.3303	0.0000
5	0.1720	(0.3282)	(0.4472)	0.4472	28.1730	0.0000
6	0.0185	(0.1400)	(0.4472)	0.4472	28.1835	0.0001
7	0.0243	0.0334	(0.4472)	0.4472	28.2032	0.0002
8	(0.0280)	0.0286	(0.4472)	0.4472	28.2316	0.0004
9	(0.2099)	(0.1544)	(0.4472)	0.4472	29.9897	0.0004
10	(0.3074)	(0.1478)	(0.4472)	0.4472	34.1800	0.0002
11	(0.2828)	(0.0666)	(0.4472)	0.4472	38.1679	0.0001
12	(0.2734)	0.0529	(0.4472)	0.4472	42.4282	0.0000
13	(0.3774)	(0.0941)	(0.4472)	0.4472	51.9000	0.0000
14	(0.4018)	(0.0644)	(0.4472)	0.4472	64.7818	0.0000
15	(0.2998)	(0.0012)	(0.4472)	0.4472	73.7471	0.0000
16	(0.2303)	0.0428	(0.4472)	0.4472	80.8003	0.0000
17	(0.2489)	0.0064	(0.4472)	0.4472	93.1562	0.0000
18	(0.1652)	0.0892	(0.4472)	0.4472	104.0461	0.0000

If autocorrelation AC(1) is nonzero, it means that the series is first order serially correlated. If AC(k) dies off more or less geometrically with increasing lag , it implies that the series follows a low-order autoregressive process. If AC(k) drops to zero after a small number of lags, it implies that the series follows a low-order moving-average process. Partial correlation PAC(k) measures the correlation of values that are k periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k, then the partial autocorrelation at lag k will be close to zero. Ljung-Box Q-statistics and their p-values at lag k has the null hypothesis that there is no autocorrelation up to order k. The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at (approximately) the 5% significance level.

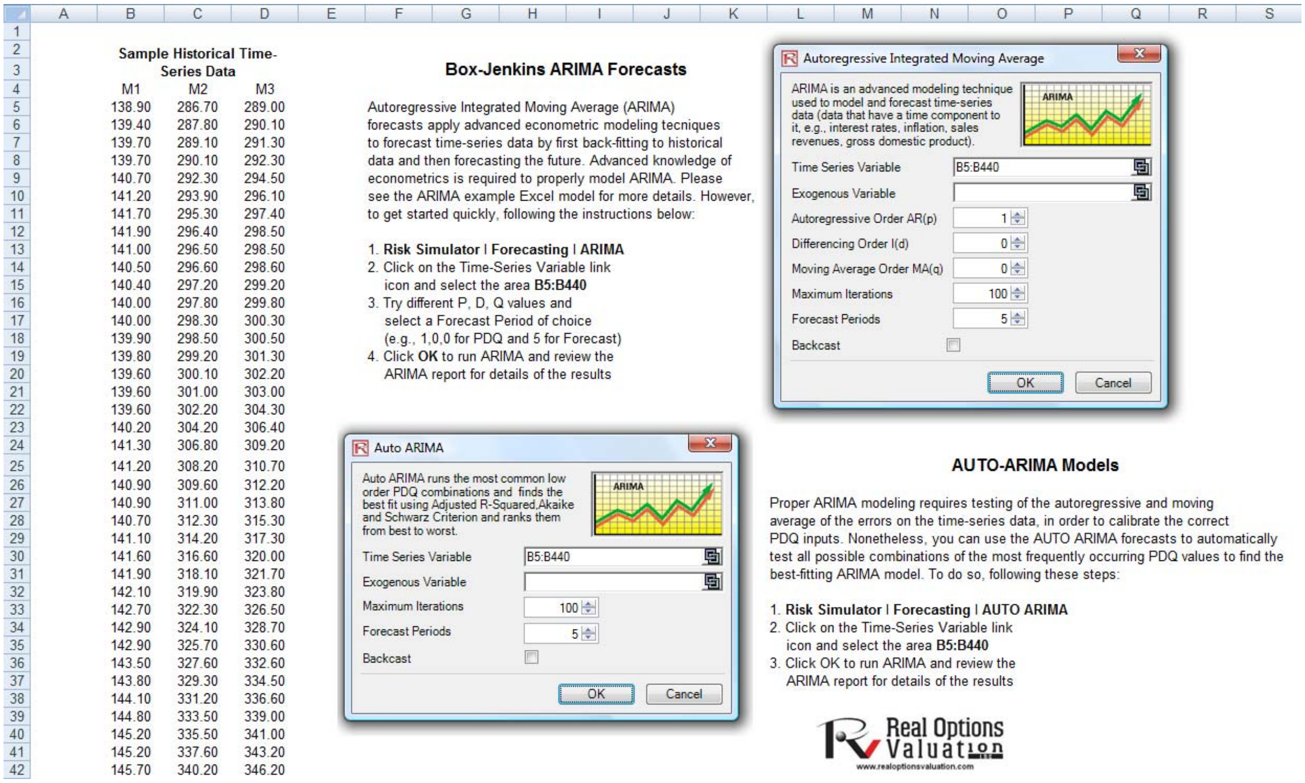
Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
1	584.1000	793.3540	(209.2540)
2	765.4000	694.3043	71.0957
3	892.3000	873.7021	18.5979
4	885.4000	999.2706	(113.8706)
5	677.0000	992.4430	(315.4430)
6	1,006.6000	786.2296	220.3704
7	1,122.1000	1,112.3713	9.7287
8	1,163.4000	1,226.6595	(63.2595)
9	993.2000	1,267.5262	(274.3262)
10	1,312.5000	1,099.1119	213.3881
11	1,545.3000	1,415.0618	130.2382
12	1,596.2000	1,645.4192	(49.2192)
13	1,260.4000	1,695.7852	(435.3852)
14	1,735.2000	1,363.5084	371.6916
15	2,029.7000	1,833.3267	196.3733
16	2,107.8000	2,124.7368	(16.9368)
17	1,650.3000	2,202.0173	(551.7173)
18	2,304.4000	1,749.3175	555.0825
19	2,639.4000	2,396.5546	242.8454
20		2,728.0397	
21		2,815.7494	

Figure 1. Box Jenkins ARIMA Forecast Report (*continued*)

Procedure

- Start Excel and enter your data or open an existing worksheet with historical data to forecast (the illustration shown in Figure 2 uses the example file *Advanced Forecasting Models* in the *Examples* menu of ***Risk Simulator***).
- In the *Auto ARIMA* worksheet, select *Risk Simulator | Forecasting | AUTO-ARIMA*. You can also access the method through the Forecasting icons ribbon or right-clicking anywhere in the model and selecting the forecasting shortcut menu.
- Click on the link icon and link to the existing time-series data, enter the number of forecast periods desired, and click *OK*.



Sample Historical Time-Series Data

M1	M2	M3
138.90	286.70	289.00
139.40	287.80	290.10
139.70	289.10	291.30
139.70	290.10	292.30
140.70	292.30	294.50
141.20	293.90	296.10
141.70	295.30	297.40
141.90	296.40	298.50
141.00	296.50	298.50
140.50	296.60	298.60
140.40	297.20	299.20
140.00	297.80	299.80
140.00	298.30	300.30
139.90	298.50	300.50
139.80	299.20	301.30
139.60	300.10	302.20
139.60	301.00	303.00
139.60	302.20	304.30
140.20	304.20	306.40
141.30	306.80	309.20
141.20	308.20	310.70
140.90	309.60	312.20
140.90	311.00	313.80
140.70	312.30	315.30
141.10	314.20	317.30
141.60	316.60	320.00
141.90	318.10	321.70
142.10	319.90	323.80
142.70	322.30	326.50
142.90	324.10	328.70
142.90	325.70	330.60
143.50	327.60	332.60
143.80	329.30	334.50
144.10	331.20	336.60
144.80	333.50	339.00
145.20	335.50	341.00
145.20	337.60	343.20
145.70	340.20	346.20

Box-Jenkins ARIMA Forecasts

Autoregressive Integrated Moving Average (ARIMA) forecasts apply advanced econometric modeling tecniques to forecast time-series data by first back-fitting to historical data and then forecasting the future. Advanced knowledge of econometrics is required to properly model ARIMA. Please see the ARIMA example Excel model for more details. However, to get started quickly, following the instructions below:

1. **Risk Simulator | Forecasting | ARIMA**
2. Click on the Time-Series Variable link icon and select the area **B5:B440**
3. Try different P, D, Q values and select a Forecast Period of choice (e.g., 1,0,0 for PDQ and 5 for Forecast)
4. Click **OK** to run ARIMA and review the ARIMA report for details of the results

Autoregressive Integrated Moving Average

ARIMA is an advanced modeling technique used to model and forecast time-series data (data that have a time component to it, e.g., interest rates, inflation, sales revenues, gross domestic product).

Time Series Variable: B5:B440
Exogenous Variable:
Autoregressive Order AR(p): 1
Differencing Order I(d): 0
Moving Average Order MA(q): 0
Maximum Iterations: 100
Forecast Periods: 5
Backcast ☐

Auto ARIMA

Auto ARIMA runs the most common low order PDQ combinations and finds the best fit using Adjusted R-Squared,Akaike and Schwarz Criterion and ranks them from best to worst.

Time Series Variable: B5:B440
Exogenous Variable:
Maximum Iterations: 100
Forecast Periods: 5
Backcast ☐

AUTO-ARIMA Models

Proper ARIMA modeling requires testing of the autoregressive and moving average of the errors on the time-series data, in order to calibrate the correct PDQ inputs. Nonetheless, you can use the AUTO ARIMA forecasts to automatically test all possible combinations of the most frequently occurring PDQ values to find the best-fitting ARIMA model. To do so, following these steps:

1. **Risk Simulator | Forecasting | AUTO ARIMA**
2. Click on the Time-Series Variable link icon and select the area **B5:B440**
3. Click OK to run ARIMA and review the ARIMA report for details of the results

Figure 2. AUTO-ARIMA Module

ARIMA and AUTO ARIMA Note

For *ARIMA* and *Auto ARIMA*, you can model and forecast future periods either by using only the dependent variable (*Y*), that is, the *Time Series Variable* by itself, or you can insert additional exogenous variables (X_1, X_2,…, X_n) just as in a regression analysis where you have multiple independent variables. You can run as many forecast periods as you wish if you only use the time-series variable (*Y*). However, if you add exogenous variables (*X*), be sure to note that your forecast periods are limited to the number of exogenous variables' data periods minus the time-series variable's data periods. For example, you can only forecast up to 5 periods if you have time-series historical data of 100 periods and only if you have exogenous variables of 105 periods (100 historical periods to match the time-series variable and 5 additional future periods of independent exogenous variables to forecast the time-series dependent variable).