Advanced Forecasting Techniques and Models: ARIMA

Short Examples Series

using

Risk Simulator

For more information please visit:
www.realoptionsvaluation.com
or contact us at:
admin@realoptionsvaluation.com

Forecasting – Time-Series ARIMA

File Name: *Forecasting – Time-Series ARIMA*
Location: *Modeling Toolkit | Forecasting | ARIMA*
Brief Description: *This sample model illustrates how to run an econometric model called the Box-Jenkins ARIMA, which stands for autoregressive integrated moving average, an advanced forecasting technique that takes into account historical fluctuations, trends, seasonality, cycles, prediction errors, and nonstationarity of the data*
Requirements: *Modeling Toolkit, Risk Simulator*

The *Data* worksheet in the model contains some historical time-series data on money supply in the United States, denoted M1, M2, and M3. M1 is the most liquid form of money (cash, coins, savings accounts, and so forth); M2 and M3 are less liquid forms of money (bearer bonds, certificates of deposit, and so forth). These data sets are useful examples of long-term historical time-series data where ARIMA can be applied.

Briefly, ARIMA econometric modeling takes into account historical data and decomposes it into an *Autoregressive* (AR) process, where there is a memory of past events (e.g., the interest rate this month is related to the interest rate last month, and so forth, with a decreasing memory lag); an *Integrated* (I) process, which accounts for stabilizing or making the data stationary and ergodic, making it easier to forecast; and a *Moving Average* (MA) of the forecast errors, such that the longer the historical data, the more accurate the forecasts will be, as it learns over time. ARIMA models therefore have three model parameters, one for the AR(p) process, one for the I(d) process, and one for the MA(q) process, all combined and interacting among each other and recomposed into the ARIMA (p,d,q) model.

There are many reasons why an ARIMA model is superior to common time-series analysis and multivariate regressions. The common 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. 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: $y_t = a_1 y_{t-1} + ... + a_p y_{t-p} + e_t + b_1 e_{t-1} + ... + b_q e_{t-q}$.

In interpreting the results of an ARIMA model, most of the specifications are identical to the multivariate regression analysis. However, there are several additional sets of results specific to the ARIMA analysis. 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 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.

Running an ARIMA Forecast

To run this model, simply:

1. Go to the *Data* worksheet and select **Risk Simulator | Forecasting | ARIMA**.
2. Click on the **LINK** icon beside the *Time Series Variable* input box, and link in **C7:C442**.
3. Enter in the relevant *P, D, Q* inputs, forecast periods, maximum iterations, and so forth (Figure 1) and click **OK**.

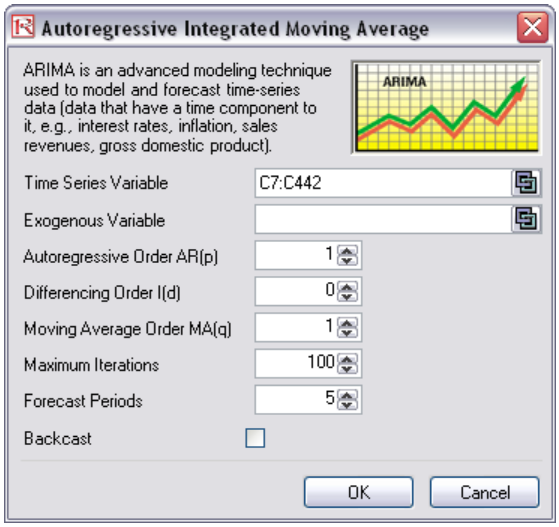


Figure 1: Running a Box-Jenkins ARIMA model

The nice thing about using Risk Simulator is the ability to run its AUTO-ARIMA module. That is, instead of needing advanced econometric knowledge, the AUTO-ARIMA module can automatically test all most commonly used models and rank them from the best fit to the worst fit. Figure 2 illustrates the results generated from an AUTO-ARIMA module in Risk Simulator and Figure 3 shows the best-fitting ARIMA model report.

AUTO-ARIMA (Autoregressive Integrated Moving Average)

	Adjusted R-Squared	Akaike Information Criterion (AIC)	Schwarz Criterion (SC)	Durbin-Watson Statistic (DW)	Number of Iterations	Model Rank
P=2, D=0, Q=0	0.9999	4.5624	4.6044	2.1254	0	1
P=1, D=0, Q=1	0.9999	4.6213	4.6632	1.8588	5	2
P=1, D=0, Q=0	0.9999	4.8908	4.9187	0.9211	0	3
P=0, D=0, Q=1	0.7309	12.9116	12.9395	0.1188	62	4
P=1, D=1, Q=2	0.5025	4.496	4.552	2.0424	15	5
P=2, D=1, Q=1	0.4977	4.4966	4.5527	1.9676	20	6
P=1, D=1, Q=1	0.4865	4.5301	4.5721	1.7916	15	7
P=2, D=1, Q=0	0.4514	4.587	4.6291	2.0969	0	8
P=1, D=1, Q=0	0.434	4.6297	4.6577	2.242	0	9
P=0, D=1, Q=1	0.3242	4.7944	4.8224	1.7167	8	10
P=1, D=2, Q=2	0.2689	4.4883	4.5444	1.9968	23	11
P=2, D=2, Q=2	0.2672	4.5009	4.5711	2	28	12
P=0, D=2, Q=2	0.2593	4.5118	4.5538	2.0413	11	13
P=1, D=2, Q=1	0.2526	4.5127	4.5547	1.9681	8	14
P=2, D=2, Q=0	0.2057	4.5862	4.6284	2.0417	0	15
P=0, D=1, Q=0	0	5.1877	5.2016	0.6802	0	16
P=0, D=2, Q=0	0	4.8166	4.8305	2.6443	0	17
P=0, D=0, Q=2	N/A	N/A	N/A	N/A	N/A	18
P=2, D=0, Q=1	N/A	N/A	N/A	N/A	N/A	19
P=2, D=0, Q=2	N/A	N/A	N/A	N/A	N/A	20

Figure 2: AUTO-ARIMA results

AUTO-ARIMA (Autoregressive Integrated Moving Average)

Regression Statistics

R-Squared (Coefficient of Determination)	0.9999	Akaike Information Criterion (AIC)	4.5624
Adjusted R-Squared	0.9999	Schwarz Criterion (SC)	4.6044
Multiple R (Multiple Correlation Coefficient)	1.0000	Log Likelihood	-990.0409
Standard Error of the Estimates (SEy)	297.5328	Durbin-Watson (DW) Statistic	2.1254
Number of Observations	434	Number of Iterations	0

Autoregressive Integrated Moving Average or 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 the percent variation in the dependent variable that 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. However, under some ARIMA modeling circumstances (e.g., with nonconvergence models), the R-Squared 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 correlation 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 and 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	AR(1)	AR(2)
Coefficients	-0.0025	1.5454	-0.5429
Standard Error	0.2020	0.0407	0.0410
t-Statistic	-0.0122	37.9479	-13.2551
p-Value	0.9902	**0.0000**	**0.0000**
Lower 5%	0.3304	1.6125	-0.4754
Upper 95%	-0.3354	1.4782	-0.6104

Degrees of Freedom	
Degrees of Freedom for Regression	2
Degrees of Freedom for Residual	431
Total Degrees of Freedom	433

Hypothesis Test	
Critical t-Statistic (99% confidence with df of 431)	2.5873
Critical t-Statistic (95% confidence with df of 431)	1.9655
Critical t-Statistic (90% confidence with df of 431)	1.6484

The Coefficients provide the estimated regression intercept and slopes. For instance, the coefficients are estimates of the true; population b values in the following regression equation Y = β0 + β1X1 + β2X2 + ... + βnXn. The Standard Error measures 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 coefficients is statistically 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 significant 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 90% confidence or 0.10 alpha level, while those highlighted in red indicate that they are not statistically significant at any other alpha levels.

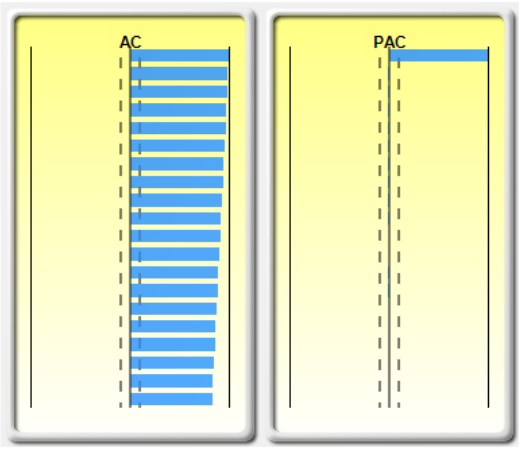
Analysis of Variance

	Sums of Squares	Mean of Squares	F-Statistic	p-Value	Hypothesis Test	
Regression	38329215.66	19164607.83	3392629.5	0	Critical F-statistic (99% confidence with df of 2 and 431)	4.6547
Residual	2434.67	5.65			Critical F-statistic (95% confidence with df of 2 and 431)	3.0167
Total	38331650.33	19164613.48			Critical F-statistic (90% confidence with df of 2 and 431)	2.3149

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 Coefficients' 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 by comparing the calculated F-Statistic with the critical F values at various significance levels.

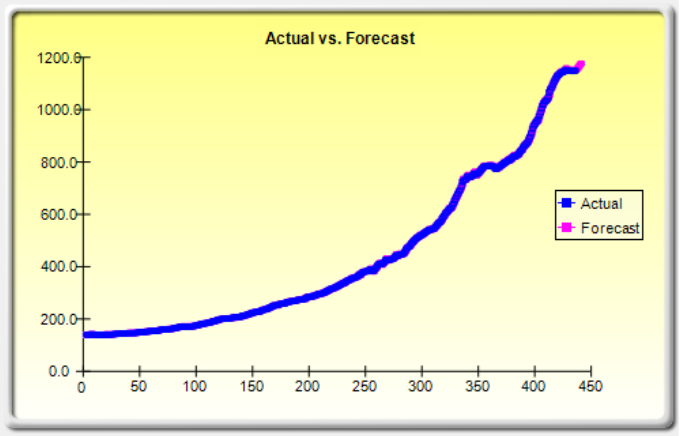
Autocorrelation

Time Lag	AC	PAC	Lower Bound	Upper Bound	Q-Stat	Prob
1	0.9921	0.9921	(0.0958)	0.0958	430.1374	-
2	0.9841	(0.0105)	(0.0958)	0.0958	854.3510	-
3	0.9760	(0.0109)	(0.0958)	0.0958	1,272.5766	-
4	0.9678	(0.0142)	(0.0958)	0.0958	1,684.7083	-
5	0.9594	(0.0098)	(0.0958)	0.0958	2,090.7005	-
6	0.9509	(0.0113)	(0.0958)	0.0958	2,490.4913	-
7	0.9423	(0.0124)	(0.0958)	0.0958	2,884.0058	-
8	0.9336	(0.0147)	(0.0958)	0.0958	3,271.1421	-
9	0.9247	(0.0121)	(0.0958)	0.0958	3,651.8357	-
10	0.9156	(0.0139)	(0.0958)	0.0958	4,026.0019	-
11	0.9066	(0.0049)	(0.0958)	0.0958	4,393.6732	-
12	0.8975	(0.0068)	(0.0958)	0.0958	4,754.8599	-
13	0.8883	(0.0097)	(0.0958)	0.0958	5,109.5382	-
14	0.8791	(0.0087)	(0.0958)	0.0958	5,457.6999	-
15	0.8698	(0.0064)	(0.0958)	0.0958	5,799.3668	-
16	0.8605	(0.0056)	(0.0958)	0.0958	6,134.5714	-
17	0.8512	(0.0062)	(0.0958)	0.0958	6,463.3408	-
18	0.8419	(0.0038)	(0.0958)	0.0958	6,785.7337	-
19	0.8326	(0.0003)	(0.0958)	0.0958	7,101.8507	-
20	0.8235	0.0002	(0.0958)	0.0958	7,411.7987	-



Forecasting

Period	Actual (Y)	Forecast (F)	Error (E)
3	139.699997	140.0142	(0.3142)
4	139.699997	140.2063	(0.5063)
5	140.699997	140.0435	0.6565
6	141.199997	141.5888	(0.3888)
7	141.699997	141.8186	(0.1186)
8	141.899994	142.3199	(0.4199)
9	141	142.3575	(1.3575)
10	140.5	140.8581	(0.3581)
11	140.399994	140.5740	(0.1740)
12	140	140.6909	(0.6909)
13	140	140.1271	(0.1271)
14	139.899994	140.3442	(0.4442)
15	139.800003	140.1897	(0.3897)
16	139.600006	140.0894	(0.4894)
17	139.600006	139.8347	(0.2347)
18	139.600006	139.9432	(0.3432)
19	140.199997	139.9432	0.2568
20	141.300003	140.8704	0.4296
21	141.199997	142.2446	(1.0446)
22	140.899994	141.4929	(0.5929)
23	140.899994	141.0836	(0.1836)
24	140.699997	141.2464	(0.5464)
25	141.100006	140.9374	0.1626
26	141.600006	141.6641	(0.0641)
27	141.899994	142.2196	(0.3196)
28	142.100006	142.4118	(0.3118)
29	142.699997	142.5580	0.1420
30	142.899994	143.3766	(0.4766)
31	142.899994	143.3600	(0.4600)
32	143.5	143.2514	0.2486
33	143.800003	144.1786	(0.3786)
34	144.100006	144.3165	(0.2165)
35	144.800003	144.6172	0.1828
36	145.199997	145.5361	(0.3361)
37	145.199997	145.7742	(0.5742)
38	145.699997	145.5571	0.1429
39	146	146.3298	(0.3298)
40	146.399994	146.5219	(0.1219)
41	146.800003	146.9772	(0.1772)
42	146.600006	147.3782	(0.7782)



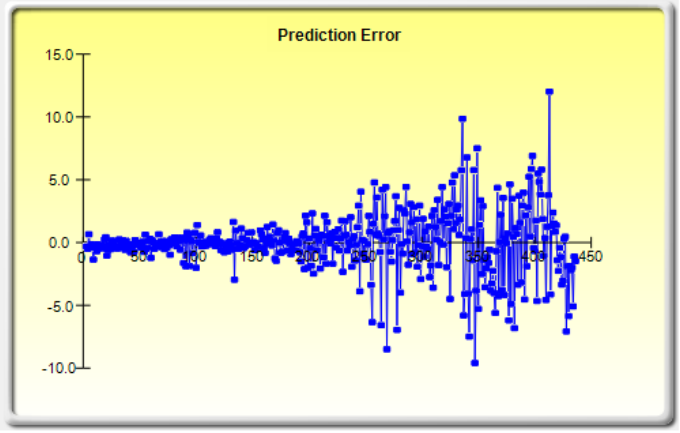


Figure 3: Best-Fitting AUTO-ARIMA results