Exploring Economic Time Series with PROC ARIMA

Joseph Earley, Loyola Marymount University, Los Angeles, California

Abstract

This paper explains how the SAS procedure PROC ARIMA may be used to identify, estimate and forecast economic time series. The ACF, IACF and PACF functions are explained and used to identify an appropriate Arima(p,d,q) model. Also discussed is differencing for stationarity. To illustrate these concepts, several correlograms are presented for the economic time series Gross Domestic Product (=GDP) and Unemployment (=U).

Univariate ARIMA Modeling

A univariate ARIMA (p,d,q) model may be represented as [1]:

$$\phi(B) Z_t = \delta + \theta(B)\epsilon_t$$

where:

$$\phi(B) = 1 - \phi_1 B - \dots - \phi_p B_p$$

$$\theta(B) = 1 - \theta_1 B - \dots - \theta_q B_q$$

δ is a constant

$$Z_t = \begin{cases} \nabla^d Y_t & d > 0 \\ Y_t & d = 0 \end{cases}$$

with ∇ as the difference operator and B the backshift operator

$$\nabla = 1 - B$$

$$\nabla Y_t = Y_t - Y_{t-1}$$

$$B^k Y_t = Y_{t-k}$$

ϵ_t are random shocks (errors) assumed to be normally and independently distributed with mean zero and constant variance

Modeling a univariate ARIMA time series involves three distinct steps:

1. Identification of the type of time series process. This is done by examining the ACF, IACF and PACF functions.
2. Estimation of the parameters.
3. Forecasting the time series into the future.

Identification of Univariate Arima Model

In order to identify the appropriate Arima model for a time series, the first step is to run the PROC ARIMA SAS procedure with the IDENTIFY statement. This yields three important statistical functions which are integral to the selection of the proper model. These functions are the ACF, autocorrelation function, IACF, inverse autocorrelation function, and the PACF, partial autocorrelation function. A visual examination of these plotted functions gives a preliminary idea as to the best fitting model. The autocorrelation function is a plotted diagram varying from -1.0 to 1.0 on the horizontal axis and 0 to 24 (the default) on the vertical axis. It measures the autocorrelation (correlation of the series with itself) at successive lags. The dotted lines indicate confidence limits which are two standard errors for the sample autocorrelation at each lag p, derived from the null hypothesis that a pure MA process of p-1 generated the series. If the asterisks appear outside the dotted lines, the autocorrelation at that lag is called a spike. If the autocorrelations after lag p are not significant, it is said to cut off or drop off after lag p. Alternatively, a time series may show a pattern in which the autocorrelations decline in an exponential pattern. It is then said to decay, dampen, tail off or die down. This pattern of decay may also be transcendental such as a sine wave pattern. The inverse autocorrelation function is the autocorrelation function of an inverted model. For example, an MA(1) model such as:

$$Z_t = \epsilon_t - \theta_1 \epsilon_{t-1}$$

may be written as:

$$Z_t = -\theta_1 Z_{t-1} - \theta_2 Z_{t-2} - \theta_3 Z_{t-3} - \theta_4 Z_{t-4} - \dots + \epsilon_t$$

which is an infinite number of autoregressive terms and no lagged error terms. The importance of this is that the sample IACF operates like the sample partial autocorrelation function. The IACF is generally thought to be useful in determining whether or not a series has been overdifferenced.[2] For example, if a series has been overdifferenced, the IACF will appear as an ACF for a nonstationary series.

Finally, the partial autocorrelation function measures the correlation between time series observations p units apart after controlling for the correlation or effects of the intervening observations at intermediate lags. Essentially, the PACF allows us to summarize the effects of all the information in the ACF in a small number of parameters. For example we need observe only p partial autocorrelations in order to identify an AR process, whereas the autocorrelations for the process could stretch to infinity. Like the ACF plot, the PACF plot also shows dots indicating confidence intervals at each lag p used to evaluate the null hypothesis that the model is a pure AR process of p-1 order.

Estimation of the Model

Based on the results from the identification of the model, the SAS ESTIMATE statement is then used to fit the appropriate model. SAS uses an iterative process to estimate model parameters. There are three methods available: conditional least squares (CLS), unconditional least squares (ULS) and maximum likelihood (ML), with CLS being the default. The particular choice of method is placed in the ESTIMATE statement with the METHOD = option. See the SAS/ETS User's Guide for more details on the methods. It is generally thought that with a sufficiently long time series and an appropriate model specification, all methods yield nearly the same results. Observing the autocorrelation check for white noise, sometimes referred to as Q-statistics, we determine whether or not the null hypothesis of white noise can be rejected. This statistic is a chi-square type statistic using the Ljung and Box form. Each row gives the value of the Q statistic up to the appropriate number of lags. Simply observing the p-values for the Q statistic indicates whether or not to reject the null hypothesis of white noise for lags up to that number.

Forecast for the Model

Having estimated the model, the final stage is to forecast the series using the FORECAST statement. Forecasts for each Arima model are based on the results from the ESTIMATE stage. The model selected from this stage is then extrapolated into the future using a forward shifting of the time series. In its output, SAS presents the forecast, standard error of the forecast and upper and lower 95% confidence intervals.

Differencing for Stationarity

A time series is said to be stationary if both its mean and variance are constant over time. Many economic time series such as GDP, Consumption and the Money Supply are obviously not stationary with respect to this definition. In order to use the theoretically known Arima models, a stationary series can by derived by taking the difference in the time series. One should be sure, however, not to over-difference the time series. A way to detect the existence of stationarity is to observe the pattern of the ACF function. A nonstationary time series will exhibit significant and large spikes for a long lag period. An example of a stationary time series would be that of a pure white noise model:

$$Y_t = \epsilon_t$$

If this series were over-differenced it would have the form:

$$Y_t - Y_{t-1} = \epsilon_t - \epsilon_{t-1}$$

which in difference form would be:

$$Z_t = \epsilon_t - \epsilon_{t-1}$$

which is an MA(1) model with $\theta_1 = 1.0$.[3]

AR(1) Model for Unemployment

To illustrate the Arima process we examine an economic time series which may be modeled as an AR(1) Arima model. The series is annual data for Unemployment of all workers over 16 years from 1950 to 1992. Examination of the visual functions shows that the ACF appears to tail off exponentially. The IACF and the PACF both show one statistically significant spike at lag 1. This along with the ACF pattern indicates that an AR(1) model may be an appropriate representation for the time series. The Q statistics in the autocorrelation check for residuals indicate that the white noise hypothesis may not be rejected. The general form for an AR(1) model is:

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \epsilon_t$$

An alternative widely used way to write the AR(1) model is:

$$Y_t = \theta_o + \phi_1 Y_{t-1} + \epsilon t$$

where the constant parameter $\theta_o = \mu(1 - \phi_1)$

This version of the AR(1) model emphasizes that the value of the time series Y_t is a constant plus a proportion of the previous value Y_t plus random error. Numerous response variables in economics tend toward this type of adaptive response behavior.

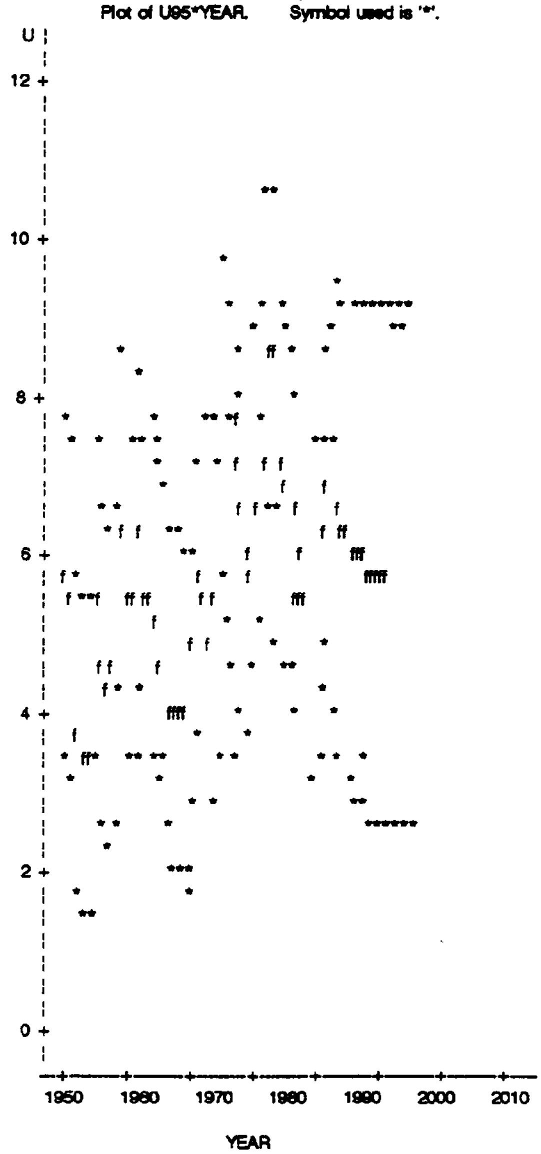
```
Name of variable = U
Mean of working series = 5.634884
Standard deviation     = 1.596345
Number of observations =      43

Autocorrelations
Lag Covariance Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
 0  2.548318   1.00000 |                    |********************|
 1  1.928879   0.75892 |                .   |***************     |
 2  1.306552   0.51271 |            .   .   |**********         |
 3  0.966497   0.37927 |             .      |********  .       |
 4  0.829919   0.32567 |             .      |*******   .       |
 5  0.618869   0.24285 |             .      |*****     .       |
 6  0.408424   0.16027 |             .      |***       .       |
 7  0.276059   0.10833 |             .      |**        .       |
 8  0.153554   0.06026 |             .      |*         .       |
 9  0.141379   0.05548 |             .      |*         .       |
10  0.121945   0.04785 |             .      |*         .       |
                        "." marks two standard errors

Inverse Autocorrelations

Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
 1   -0.58758 |    ************|  .                |
 2    0.10123 |             .  |**  .              |
 3    0.09420 |             .  |**  .              |
 4   -0.14901 |            . ***|  .              |
 5    0.04329 |             .  |*   .              |
 6    0.05800 |             .  |*   .              |
 7   -0.11153 |            .  **|  .              |
 8    0.12038 |             .  |**  .              |
 9   -0.06546 |             .  *|  .              |
10    0.01267 |             .  |   .              |

Partial Autocorrelations

Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
 1    0.75892 |             .  |***************    |
 2   -0.14101 |            . ***|  .              |
 3    0.10319 |             .  |**  .              |
 4    0.07843 |             .  |**  .              |
 5   -0.09000 |            .  **|  .              |
 6   -0.00662 |             .  |   .              |
 7    0.00639 |             .  |   .              |
 8   -0.05549 |            .   *|  .              |
 9    0.08191 |             .  |**  .              |
10   -0.02407 |             .  |   .              |

Autocorrelation Check for White Noise

 To   Chi               Autocorrelations
Lag  Square DF  Prob
 6   55.37  6  0.000  0.757  0.513  0.379  0.326  0.243  0.160

Conditional Least Squares Estimation
Parameter   Estimate   Std Error   T Ratio  Lag
MU          5.81913    0.59653     9.42     0
AR1,1       0.77679    0.10158     7.65     1
Constant Estimate  = 1.25422458
Variance  Estimate = 1.10162504
Std Error Estimate = 1.04958327
AIC                = 128.142523*
SBC                = 131.664923*
Number of Residuals=       43

Autocorrelation Check of Residuals

 To   Chi               Autocorrelations
Lag  Square DF  Prob
 6    2.31  5  0.805  0.119 -0.114 -0.102  0.088  0.012 -0.053
12    3.62 11  0.980  0.010 -0.071 -0.023  0.037  0.108  0.059
18    7.30 17  0.979 -0.052 -0.019 -0.073  0.036  0.189 -0.069
24   15.77 23  0.865 -0.114 -0.065  0.197  0.102 -0.124 -0.096
Model for variable U
Estimated Mean = 5.61912621
```

Autoregressive Factors
Factor 1: 1 - 0.77679 B**(1)

Forecasts for variable U

Obs	Forecast	Std Error	Lower 95%	Upper 95%
44	6.9248	1.0496	4.8677	8.9820
45	6.6334	1.3290	4.0285	9.2383
46	6.4070	1.4722	3.5215	9.2925
47	6.2311	1.5523	3.1886	9.2735
48	6.0945	1.5986	2.9613	9.2277
49	5.9884	1.6259	2.8017	9.1752
50	5.9060	1.6422	2.6873	9.1246
51	5.8420	1.6519	2.6042	9.0797
52	5.7922	1.6578	2.5430	9.0414
53	5.7536	1.6613	2.4975	9.0097
54	5.7236	1.6634	2.4633	8.9838
55	5.7003	1.6647	2.4375	8.9630

Plot of U*YEAR. Symbol used is ' '.
Plot of FORECAST*YEAR. Symbol used is 'f'.
Plot of L95*YEAR. Symbol used is '*'.
Plot of U95*YEAR. Symbol used is '*'.



MA(1) Model for GDP

The next example of the Arima procedure is that of a moving average model. The time series selected is annual data for Gross Domestic Product in constant dollars from 1947 to 1992. A moving average model may in be described as:

$$Y_t = \theta_o + \epsilon_t - \theta_1 \epsilon_{t-1} - \theta_2 \epsilon_{t-2} \dots - \epsilon_{t-q}$$

Thus an MA(1) model may be written as:

$$Y_t = \theta_o + \epsilon_t - \theta_1 \epsilon_{t-1}$$

meaning that the time series is a constant value plus random error plus a proportion of random error from the previous time period. Initially the series was shown to be nonstationary. After differencing, the identification of the moving average model relies on the fact that the ACF of the differenced series has one spike at lag 1 and both the IACF and the PACF appear to tail off exponentially.

Name of variable = GDP
Period(s) of Differencing = 1.
Mean of working series = 65.04921
Standard deviation = 93.81904
Number of observations = 63
NOTE: The first observation was eliminated by differencing.

Autocorrelations

```
Lag Covariance Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  0   8802.013    1.00000 |                    |********************|
  1   3935.201    0.44708 |              .     |*********          |
  2    233.008    0.02647 |              .     |*   .              |
  3  -1555.380   -0.17871 |            . ****|      .             |
  4  -1980.737   -0.22503 |            .*****|      .             |
  5  -1209.946   -0.13748 |             .  ***|      .             |
  6   -481.964   -0.05476 |              .   *|      .             |
  7    687.947    0.07816 |              .    |**    .             |
  8   1547.338    0.17579 |              .    |****   .            |
  9  45.204351    0.00514 |              .    |       .            |
 10   -723.809   -0.06223 |              .  **|       .            |
 11  26.101973    0.00297 |              .    |       .            |
 12    329.458    0.03743 |              .    |*      .            |
 13 -48.198915   -0.00548 |              .    |       .            |
 14    826.019    0.09384 |              .    |**     .            |
 15   1112.450    0.12639 |              .    |***     .           |
                        "." marks two standard errors
```

Inverse Autocorrelations

```
Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1   -0.42670 |        **********|  .        |
  2   -0.01348 |            .    |    .       |
  3    0.12744 |            .    |***  .      |
  4    0.09737 |            .    |**   .      |
  5   -0.15757 |            . ***|     .      |
  6    0.10582 |            .    |**   .      |
  7    0.09516 |            .    |**   .      |
  8   -0.23219 |            *****|     .      |
  9    0.09769 |            .    |**   .      |
 10    0.08885 |            .    |**   .      |
 11   -0.08356 |            .  **|     .      |
 12   -0.11289 |            .  **|     .      |
 13    0.17459 |            .    |***  .      |
 14   -0.11331 |            .  **|     .      |
 15   -0.01371 |            .    |     .      |
```

Partial Autocorrelations

```
Lag Correlation -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1   0.44708 |               .    |**********        |
  2  -0.21673 |              .****|   .              |
  3  -0.12355 |               . **|  .              |
  4  -0.09914 |               . **|  .              |
  5  -0.00893 |               .   |  .             |
  6  -0.03918 |               .  *|  .             |
  7   0.08712 |               .   |** .            |
  8   0.08695 |               .   |** .            |
  9  -0.17875 |              .****|   .             |
 10   0.00858 |               .   |  .             |
 11   0.13075 |               .   |*** .            |
 12  -0.00923 |               .   |  .             |
 13  -0.07299 |               .  *|  .             |
 14   0.19145 |               .   |****.            |
 15   0.02054 |               .   |  .             |
```

Autocorrelation Check for White Noise

```
 To   Chi                Autocorrelations
Lag  Square DF  Prob
  6  20.44  6  0.002  0.447  0.026 -0.177 -0.225 -0.137 -0.055
 12  23.83 12  0.021  0.078  0.176  0.005 -0.082  0.003  0.037
```

Conditional Least Squares Estimation

```
                      Approx.
Parameter  Estimate   Std Error  T Ratio  Lag
MU         64.11972   15.65359     4.10   0
MA1,1      -0.48619    0.11436    -4.25   1
Constant Estimate  = 64.1197169
Variance  Estimate = 7096.09148
Std Error Estimate = 84.2363018
AIC                = 739.393684*
SBC                = 743.679954*
Number of Residuals=       63
```

Autocorrelation Check of Residuals

```
 To   Chi                Autocorrelations
Lag  Square DF  Prob
  6   3.74   5  0.588  0.050  0.058 -0.150 -0.145 -0.061 -0.036
 12   7.40  11  0.766  0.026  0.190 -0.032 -0.059 -0.017  0.082
 18  11.83  17  0.810 -0.092  0.122  0.035  0.141 -0.007 -0.084
 24  14.91  23  0.898 -0.042 -0.046  0.080  0.074 -0.017  0.120
```

Model for variable GDP
Estimated Mean = 64.1197169
Period(s) of Differencing = 1.
Moving Average Factors
Factor 1: 1 + 0.48619 B**(1)

Forecasts for variable GDP

Obs	Forecast	Std Error	Lower 95%	Upper 95%
65	5026.0235	84.2363	4860.9195	5191.1276
66	5090.1433	150.8963	4794.3920	5385.8945
67	5154.2630	196.0696	4769.9737	5538.5522
68	5218.3827	232.6303	4762.4357	5674.3297
69	5282.5024	264.1788	4764.7214	5800.2834
70	5346.6221	292.3423	4773.6417	5919.6026
71	5410.7418	318.0214	4787.4313	6034.0524
72	5474.8616	341.7766	4804.9918	6144.7313
73	5538.9813	363.9646	4825.5845	6252.3780
74	5603.1010	384.9135	4848.6844	6357.5176
75	5667.2207	404.7616	4873.9025	6460.5389
76	5731.3404	423.6610	4900.9410	6561.7399

Conclusion

The purpose of this paper was to present an introduction to univariate time series methods using PROC ARIMA. While the univariate model is a useful statistical tool, PROC ARIMA has more advanced capabilities which allow us to model vastly more complicated time series. PROC ARIMA may be used to model a time series with explanatory variables. It allows the user to identify the error process before fitting the model. PROC AUTOREG, in comparison, makes an implicit assumption about the error process, i.e. that it is autoregressive in nature. PROC ARIMA also allows for the use of Transfer Function models, in which forecasted values of the explanatory variables themselves are used to forecast the dependent variable. Finally PROC ARIMA allows the user to develop what is called an Intervention model, which is analogous to a regression model in which one or more of the regressors is a dummy variable.

References

Box, G.E.P. and G.M. Jenkins. Time Series Analysis: Forecasting and Control 2nd. edition. San Francisco, Holden-Day, 1976.

eDATA: A General Purpose Economic Database and Data Retrieval Program. Applied Data Associates, Portland, Oregon. 1994.

Nelson, Charles R., Applied Time Series Analysis. San Francisco: Holden-Day, 1973.

SAS Institute Inc., SAS/ETS User's Guide. Version 6. Second Edition,, Cary, N.C. 1993.

SAS Institute Inc., SAS/ETS Software: Applications Guide 1, Version 6, First Edition. Cary, North Carolina, 1991.

SAS Institute Inc., SAS System for Forecasting Time Series, 1986 Edition.

Notes

[1] Notation used in this paper follows that used in SAS/ETS Software: Applications Guide 1. Version 6, First Edition. The SAS Institute, Cary, North Carolina, USA.

[2] SAS/ETS User's Guide Second Edition, p. 136. See Cleveland(1972), Chatfield (1980) and Priestly (1981) for more information concerning the IACF.

[3] Vandaele, Applied Time Series and Box-Jenkins Models, p. 71.

Author' Address:

Joseph Earley
Loyola Marymount University
80th and Loyola Blvd.
Los Angeles, CA 90045
e-mail: jearley@lmumail.lmu.edu