Investigating Seasonality in BLS Data Using PROC ARIMA

Joseph Earley, Loyola Marymount University
Los Angeles, California

Abstract

This paper illustrates how the SAS System® statistical procedure PROC ARIMA may be used to detect, estimate and forecast time series characterized by seasonal patterns. The series selected for analysis are maintained by the Bureau of Labor Statistics[1]. In particular, ARIMA models are developed for urate, the civilian labor force unemployment rate for workers 16 years and older (Series:LFU21000000) and for eduservices, the number of employees in educational services (Series:EEU80820001). The paper illustrates how the autocorrelation and partial correlation functions are used to detect and model seasonality in the series. Finally, results determined from this methodology are compared with those derived by using the ARIMA option from the SAS procedure PROC X11.

Introduction

The purpose of this paper is to provide an example of how to use the SAS PROC ARIMA time series procedure in the presence of seasonality. Using several macroeconomic time series, the paper explains how to identify, estimate and forecast with the appropriate Arima(p,d,q)(P,D,Q) TRANSFORMATION model. The paper begins with a discussion of the univariate time series model. Following this is a discussion of seasonality, and how it influences the detection and estimation of the model. Finally, the above mentioned series are used to illustrate how the procedure may be applied.

Univariate Non-Seasonal ARIMA Modeling

A univariate ARIMA (p,d,q) model may be represented as[2]:

$$\phi(B)\, Z_t = \delta + \theta(B)\, \varepsilon_t$$

where: $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B_p$

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

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

Process Identification

The appropriate Arima model is determined by examining the ACF, IACF and PACF functions.[3] 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 us 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. The IACF is generally thought to be useful in determining whether or not a series has been overdifferenced. 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. 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.

Parameter Estimation

Based on the results from the identification phase, 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. 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.

Forecasting the Time Series

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. The SAS default presents the forecast, standard error of the forecast and upper and lower 95 percent confidence intervals.

Stationarity Adjustment

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 not stationary with respect to this definition. In order to use theoretically known Arima models, a stationary series can frequently by obtained by taking the first-difference in the time series. A way to detect 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. A general rule of thumb is to observe the autocorrelation for the first five lags. If the value of the autocorrelation is still large (i.e. 0.7 or greater), then the series may be considered nonstationary for practical purposes[4].

Univariate Seasonal ARIMA Modeling

A univariate time series model which also has components of seasonality may be concisely expressed as:

$(p,d,q)(P,D,Q)_S$ transformation

where p is the number of regular autoregressive parameters
d is the number of regular differences
q is the number of regular moving average parameters
P is the number of seasonal autoregressive parameters
D is the number of seasonal differences
Q is the number of seasonal

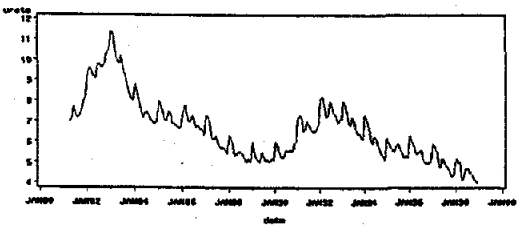
moving average parameters
S is the level of seasonality
transformation is type of transformation

Modeling a seasonal univariate ARIMA time series involves the same distinct steps as for a non-seasonal model, recognizing in addition that seasonal patterns occur at the seasonal lags. For example, purely seasonal AR processes will tail off exponentially in the ACF while showing spikes at the seasonal lags in the PACF. Likewise, purely seasonal MA model processes will exhibit spikes in the ACF at the seasonal lag while showing exponential decay in the PACF. Mixed purely seasonal models will tail off exponentially in both the ACF and the PACF.

Modeling the Unemployment Rate

Following is a plot of the civilian labor force unemployment rate for workers 16 years and older(Series:LFU21000000) for data from January, 1948 to June, 1999. Analysis of the ACF, IACF and PACF of the undifferenced urate series indicated that both a first and a seasonal difference transformation were necessary to achieve stationarity. Following are the ACF, IACF and PACF functions with periods of differencing 1 and 12.

Unemployment Rate



The ARIMA Procedure
Name of Variable = Urate
Period(s) of Differencing 1,12
Mean of Working Series -0.00234
Standard Deviation 0.338751

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	Std Error
0	0.114752	1.00000	|********************|	0
1	0.022332	0.19461	. |**** .	0.040658
2	0.026302	0.29698	. |****** .	0.043279
3	0.021803	0.19002	. |**** .	0.045744
4	0.015115	0.13172	. |*** .	0.047060
5	0.012649	0.12152	. |** .	0.047582
6	0.0075468	0.06579	. |*. .	0.048175
7	-0.0054365	-.04739	. *| .	0.048295
8	-0.0018980	-.01655	. . | .	0.048404
9	-0.0062963	-.05488	. *| .	0.048413
10	-0.018507	-.16472	. ***| .	0.048515
11	-0.0057561	-.05015	. *| .	0.049231
12	-0.055272	-.48188	**********| .	0.049316
13	-0.015037	-.13105	. ***| .	0.052777
14	-0.021327	-.18588	. ****| .	0.053240
15	-0.017449	-.15208	. ***| .	0.055315
16	-0.014975	-.13052	. ***| .	0.055974
17	-0.010867	-.09469	. **| .	0.056454
18	-0.012091	-.10538	. **| .	0.056892
19	-0.000172	-.00150	. | .	0.056999
20	-0.0011470	-.01000	. | .	0.056998
21	-0.0003419	-.00297	. | .	0.057001
22	-0.0030881	-.02701	. *| .	0.057002
23	-0.0091299	-.07971	. **| .	0.057023
24	-0.0050380	-.04441	. *| .	0.057225

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	-0.11673	. **| .
2	-0.08368	. **| .
3	-0.06151	. *| .
4	-0.04282	. *| .
5	-0.01811	. | .
6	0.00245	. | .
7	0.01342	. | .
8	-0.05420	. *| .
9	-0.05840	. *| .
10	0.00128	. | .
11	-0.12362	. **| .
12	0.52562	. |**********
13	-0.04081	. *| .
14	-0.01908	. | .
15	-0.00297	. | .
16	-0.00166	. | .
17	0.00464	. | .
18	0.01753	. | .
19	-0.01068	. | .
20	-0.05021	. *| .
21	-0.04115	. *| .
22	0.02298	. | .
23	-0.05844	. *| .
24	0.18557	. |****

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	0.19461	. |****
2	0.27042	. |*****
3	0.10718	. |**
4	0.01505	. | .
5	0.02791	. |*.
6	-0.00928	. | .
7	-0.12212	. **| .
8	-0.03745	. *| .
9	-0.02243	. | .
10	-0.12184	. **| .
11	0.01696	. | .
12	-0.45666	*********| .
13	0.02567	. |*.
14	0.06790	. |*.
15	0.01388	. | .
16	-0.01600	. | .
17	0.02579	. |*.
18	-0.01520	. | .
19	-0.00545	. | .
20	0.02712	. |*.
21	0.00112	. | .
22	-0.14295	. ***| .
23	0.01753	. | .
24	-0.28812	. ******| .

The spikes at lags 1, 2 and 12 in the ACF function indicate that regular moving average components of lags 1 and 2 may be appropriate with a seasonal MA component at lag 12. The spikes at lags 1 and 2 of the PACF indicate that an AR component may be used for lags 1 and 2. The autocorrelation check for white noise indicates that the urate series is not white noise.

Autocorrelation Check for White Noise

To	Chi Square	Pr > ChiSq	Autocorrelations					
6	126.37	<.0001	0.195	0.298	0.191	0.132	0.122	0.066
12	388.49	<.0001	-0.047	-0.017	-0.055	-0.148	-0.055	-0.487
18	392.99	<.0001	-0.134	-0.188	-0.132	-0.128	-0.092	-0.105
24	1183.54	<.0001	0.002	-0.010	-0.033	-0.027	-0.011	-0.004

After numerous estimations, the following model was found to fit the data well, as judged by statistically significant t-values.

Parameter	Estimate	Approx Std Error	t Value	Pr > |t|	Lag
MA1,1	-0.14671	0.03911	-3.75	0.0002	1
MA1,2	0.36273	0.14609	2.48	0.0133	2
MA2,1	0.73996	0.02781	26.61	<.0001	12
AR1,1	0.54519	0.13337	4.09	<.0001	2

Model for variable urate
Period(s) of Differencing: 1,12
No mean term in this model.
Autoregressive Factors:
Factor 1: 1 - 0.54519 B**(2)
Moving Average Factors:
Factor 1: 1 + 0.14671 B**(1) - 0.36273 B**(2)
Factor 2: 1 - 0.73996 B**(12)

The following residual check for the fitted model shows that the null hypothesis of white noise cannot be rejected.

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations
6	5.58	2	0.1667	-0.009 0.015 0.002 -0.025 0.058 0.051
12	8.41	8	0.3942	-0.007 0.009 -0.011 -0.003 0.049 -0.010
18	20.23	14	0.1222	-0.018 -0.079 -0.003 -0.064 0.085 -0.045
24	21.65	20	0.3695	0.004 -0.009 -0.021 -0.009 -0.026 -0.004
30	25.85	26	0.4713	0.002 -0.047 -0.049 0.053 0.025 0.025
36	37.19	32	0.2431	0.019 -0.072 -0.004 0.097 -0.006 0.022
42	49.56	38	0.0986	-0.009 0.060 -0.016 -0.019 -0.025 0.035
48	52.75	44	0.1717	-0.004 -0.047 -0.076 0.048 0.002 0.009

Forecasts for the series were calculated using the above model. Forecast diagnostics indicate that the model predicts the future of the time series well.

The X11 Procedure

The SAS procedure, PROC X11, is used in the seasonal adjustment of time series. When the ARIMA option is used with PROC X11, five pre-defined Arima models are estimated, with the best model then selected for use in the seasonal adjustment. For comparison, results of using PROC X11 with the ARIMA option are presented. While the variance from the user-determined and the pre-defined models are almost exactly the same, the AIC and SBC for the user determined model is smaller.

Seasonal Adjustment of: urate

Parameter	Estimate	Approx. Std Error	t Value	Lag
MU	-0.001863	0.0051024	-0.37	0
MA1,1	0.3360963	0.4484797	0.75	1
MA1,2	0.0461744	0.3175263	0.15	2
MA2,1	0.7368309	0.0280209	26.30	12
AR1,1	0.4645349	0.4450258	1.04	1
AR1,2	0.1844157	0.3761195	0.49	2

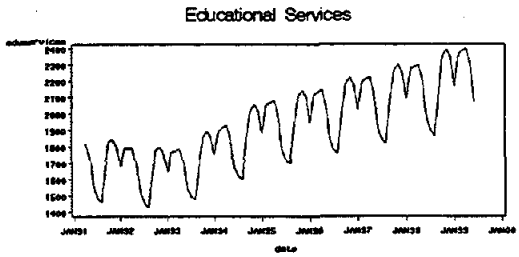
Conditional Least Squares Estimation
Variance Estimate = 0.0665424
Std Error Estimate = 0.2579581

AIC = 82.618406
SBC = 108.98998

Criteria Summary for Model 5: (2,1,2)(0,1,1)s, No Transformation
Box-Ljung Chi-square: 19.58 with 19 df Prob= 0.42 (Criteria prob > 0.05)
Test for over-differencing:sum of MA parameters = 0.74 (must be < 0.90)
MAPE - Last Three Years:2.74(Must be< 15.00 %)
- Last Year: 3.14
- Next to Last Year: 3.01
- Third from Last Year: 2.07

Modeling Educational Services

Following is a plot of the number of employees in educational services (Series:EEU80820001) for data from January, 1948 to June, 1999.
Analysis of the ACF, IACF and PACF of the undifferenced eduservices series again indicated that both a regular and a seasonal difference transformation were necessary to achieve stationarity. In addition, a logarithmic transformation was found to be helpful. Following are the ACF, IACF and PACF functions with periods of differencing 1 and 12, where leduservices is the natural logarithm of the education services series.

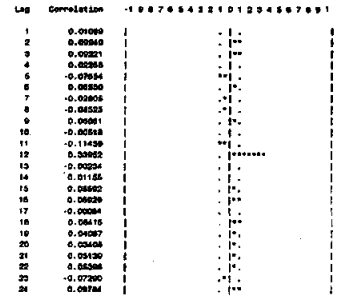


The ARIMA Procedure
Name of Variable = leduservices
Period(s) of Differencing: 1,12
Mean of Working Series -0.00002
Standard Deviation 0.009749
Number of Observations 485

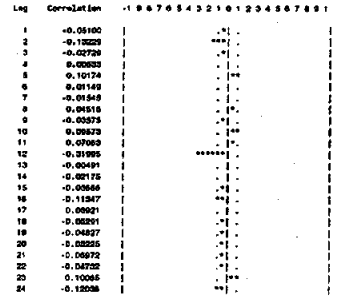
Autocorrelations

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.00009504	1.00000	|********************|	0
1	-4.561E-6	-.04799	.*| .	0.045406
2	-0.0000123	-.12936	***| .	0.045629
3	-1.188E-6	-.01250	. | .	0.046257
4	2.429E-6	0.02557	. |* .	0.046294
5	9.777E-6	0.10287	. |** .	0.046313
6	-4.467E-7	-.00521	. | .	0.046752
7	-4.838E-6	-.05091	. *| .	0.046753
8	4.2067E-6	0.04425	. |* .	0.046863
9	-2.733E-6	-.02876	. *| .	0.046949
10	9.1603E-6	0.09639	. |** .	0.046965
11	5.0283E-6	0.05294	. |* .	0.047391
12	-0.0000324	-.34043	*******| .	0.047565
13	1.7186E-6	0.01808	. | .	0.052849
14	5.0529E-6	0.05316	. |* .	0.052861
15	-5.504E-7	-.00580	. | .	0.052874
16	-9.891E-6	-.10407	. **| .	0.052874
17	+4.32E-6	-.04550	. *| .	0.052998
18	-1.994E-6	-.02099	. | .	0.053074
19	+1.377E-6	-.01449	. | .	0.053096
20	-5.274E-6	-.05549	. *| .	0.053204
21	-3.358E-6	-.03729	. *| .	0.053329
22	-8.919E-6	-.07291	. *| .	0.053417
23	5.2616E-6	0.05535	. |* .	0.053582
24	2.0978E-7	0.00211	. | .	0.053600

Inverse Autocorrelations



Partial Autocorrelations



The spikes at lags 2 and 12 in the ACF function indicate that regular moving average component of lag 2 may be appropriate with a seasonal MA component at lag 12. The autocorrelation check for white noise indicates that the log of educational services series is not white noise.

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	15.07	6	0.0197	-0.051	-0.129	-0.013	0.026	0.165	-0.005
12	31.93	12	<.0001	-0.042	0.044	-0.029	0.096	0.053	-0.349
18	39.15	18	<.0001	0.016	0.053	-0.005	-0.104	-0.045	-0.021
24	56.60	24	<.0001	-0.014	-0.058	-0.037	-0.073	0.056	0.002

An Arima (0,1,1)(0,1,1) log transform model was estimate with the following results:

Parameter	Estimate	Approx Std Error	t Value	Pr > |t|	Lag
MA1,1	0.12331	0.04523	2.73	0.0066	2
MA2,1	0.42813	0.04127	10.37	<.0001	12

Variance Estimate	0.000081
Std Error Estimate	0.008979
AIC	-3193.09
SBC	-3184.72

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.62	4	0.2291	-0.034	0.091	-0.042	-0.004	0.089	-0.024
12	11.93	10	0.2895	-0.005	0.024	-0.029	0.066	0.071	0.021
18	20.40	16	0.2029	-0.004	-0.041	-0.056	-0.104	-0.015	-0.055
24	26.97	22	0.2123	-0.009	-0.073	-0.064	-0.016	0.087	-0.031
30	35.03	28	0.1691	-0.045	-0.111	-0.032	0.015	0.004	-0.007
36	47.69	34	0.0594	0.046	-0.020	0.003	0.130	-0.045	-0.042
42	52.90	40	0.0632	-0.038	0.046	-0.019	0.021	-0.044	0.053
48	57.94	46	0.1114	0.036	-0.024	-0.033	-0.050	0.027	-0.041

Model for variable leduservices

Period(s) of Differencing: 1,12
No mean term in this model.
Moving Average Factors
Factor 1: 1 - 0.12331 B**(2)
Factor 2: 1 - 0.42813 B**(12)

Again, the PROC X11 seasonal adjustment was used for comparison with the author-derived model. Using the Arima option under PROC X11 resulted in an Arima (0,1,2)(0,1,1) log transform model. The t-values for that model were not significant for the both the constant and the MA(1) terms. Both the AIC and the SBC statistics were smaller for the author derived model.

The X11 Procedure

Seasonal Adjustment of - Eduservices

Parameter	Estimate	Approx. Std Error	t- Value	Lag
MU	-0.000017	0.0001992	-0.08	0
MA1,1	0.0375033	0.0453552	0.83	1
MA1,2	0.1246365	0.0453069	2.75	2
MA2,1	0.4270633	0.0414703	10.30	12

Conditional Least Squares Estimation

Variance Estimate	=	0.0000808
Std Error Estimate	=	0.0089914
AIC	=	-3189.79
SBC	=	-3173.053

Criteria Summary for Model 2: (0,1,2)(0,1,1)s, Log Transform Box-Ljung Chi-square: 27.36 with 21 df Prob= 0.16 (Criteria prob > 0.05)
Test for over-differencing: sum of MA parameters = 0.43 (must be < 0.90)
MAPE - Last Three Years:0.25(Must be<15.00 %)
- Last Year: 0.31
- Next to Last Year: 0.11
- Third from Last Year: 0.32

Conclusion

The primary conclusion derived from this paper is that users of the PROC ARIMA procedure should be experienced in using the classic Box-Jenkins methodology. While forecasts can be made mechanically from the SAS/ETS® Time Series Forecasting System, it was shown that careful analysis of the ACF, IACF and PACF may help the user develop a better fitting model. The author recommends that PROC X11 with the Arima option be used as a screening tool whenever seasonality appears to exist in a time series.

Fine-tuning the automatically selected model may yield a more parsimonious model for the series.

References

Bernstein, Jake Seasonality: Systems, Strategies, and Signals New York: John Wiley & Sons, 1998.

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

Gujarati, Damodar (1995). Basic Econometrics, Third Edition, New York: McGraw-Hill, Inc.

Jaditz, Ted, "Seasonality: Economic Data and Model Estimation", Monthly Labor Review, December 1994, pp. 17-22.

McIntire, Robert J. "Revision of Seasonally Adjusted Labor Force Series", Bureau of Labor Statistics. Http://stats.bls.gov/cpsrs.htm.

Miron, Jeffrey A., The Economics of Seasonal Cycles Cambridge, Massachusetts: The MIT Press, 1996.

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

Nerlove, Marc, Grether, David M., and Carvalho, Jose L. Analysis of Economic Time Series: A Synthesis, New York: Academic Press, 1979.

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

SAS Institute Inc., SAS/ETS Software: Applications Guides 1 and 2, Version 6, First Edition, 1993.

SAS Institute Inc., SAS/STAT User's Guide, Vols I and II, 1990 Version 6, Fourth Edition. Cary, N.C.: SAS Institute Inc.

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

Shiller, Robert J. Market Volatility Cambridge, Massachusetts: The MIT Press, 1989.

Notes

1 See:http//stats.bls.gov/blshome.htm.

2 Vandaele, Applied Time Series and Box-Jenkins Models, pp. 52-55.

3 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.

4 Vandaele, Applied Time Series and Box -Jenkins Models, p. 71.