Paper D03-2009

Time Series Analysis 101: an introduction using Base SAS and SAS STAT

David Corliss, University of Toledo Department of Physics and Astronomy, Toledo, OH

ABSTRACT

The SAS system has many powerful tools for modeling events that change over time. This class offers an a basic introduction to time series analysis, what it can you for you and how to get the procedures in Base SAS and SAS Stat up and running. Techniques covered in this paper include Regression on Time Series, Autoregressive and Moving Average models. Source code supporting basic ARIMA models – both Autoregressive and Moving Average – using only Base SAS and SAS STAT are given. The strengths, weaknesses and optimal situations for each method are compared. Additional capabilities available with the SAS add-on ETS, including PROC ARIMA and PROC FORECAST, are briefly described.

INTRODUCTION

A Time Series is any set of data describing a function or property that changes over time. Time Series Analysis studies this changing data, often with the intention of predicting what will happen in the future. This paper provides a basic introduction to Time Series Analysis, its distinctive terminology and techniques, along with source code to provide basic functionality in these methods. The SAS ETS package provides advanced capabilities for Time Series Analysis but this is beyond the scope of this introduction – the intention here is to get tools for Time Series Analysis into the hands of more analysts, enabling them to advance to more advanced tools in the future.

TIME SERIES REGRESSION

Regression methods can be applied to a Time Series just like any other data. Here are some examples from the SAS/STAT Guide chapter on PROC REG; also available from the SAS Institute Sample Library

(http://ftp.sas.com/techsup/download/sample/samp_lib/statsampDocumentation_Examples_for_Proc_.html):

```
/* Autocorrelation in Time Series Data */

/* Output 28.35 */

  proc reg data=uspop;
     model pop=year yearsq / dw;
  run;


/* ---Example 1: Population Growth Trends--- */

/* Output 28.36 */

  proc reg data=uspop;
     var yearsq;
     model pop=year / r cli clm;
     plot r.*p.;
     add yearsq;
     print;
     plot;
  run;
```

LIMITATIONS TO TIME SERIES ANALYSIS USING ORDINARY REGRESSION

Regression is often used in a very general to identify a relationship between different variable and parameters in order to use several known quantities to predict the value something that is unknown. This leads to two different kinds of limitations. Firstly, Time Series data is usually used to make a forecast and therefore attempts to predict a value outside the domain of the existing data set. Normally, regression is used to find the f(x) associated with input values *within the extent of the original data*. By contrast, Time Series Analysis is used to predict what happens in the future, *beyond* the existing data. Therefore, a technique that is specifically designed for forecasting is best.

Time Series Regression also affected by all the same limitations as any regression analysis. Utilizing the Central Limit Theorem, regression analysis requires that the input data be normally distributed: strongly skewed data renders the technique invalid. Further, the data must consist of *independent* random trials, without autocorrelation. Time series data, however, is often strongly autocorrelated because events in the immediate past often have a much stronger influence on the immediate future than points further in the past. Under these circumstances, ARIMA models may provide a better choice.

ARIMA MODELS

ARIMA stands for Auto-Regressive Integrated Moving Average. This statistical technique examines trends in data by using several successive terms to predict the next value (or next several values) in a series. This prediction can be made by either of two techniques, one focusing on Auto-Regressive properties and the using a Moving Average.

ARIMA MODELS USING AN AUTOREGRESSIVE TERM

ARIMA models are usually classified using a three part numbering system. The first digit is the number of Auto-Regressive terms in the model. The second number gives the number of random or White Noise terms in the model and the third digit is the number of Moving Average terms. So, a (1,1,0) ARIMA model combines Auto-Regression and a random element *without* the use of a Moving Average component.

(1,1,0) ARIMA models are sometimes referred to as Box – Jenkins models, following the seminal paper by George Box and Gwilym Jenkins in the 1970's (see Box, George and Jenkins, Gwilym, 1970, *Time series analysis: Forecasting and control*, San Francisco: Holden-Day). These models are very widely used in economic forecasting today. Rather than being invalidated by significant autocorrelation in the data, these models leverage the autocorrelation to produce more accurate forecasts. It is important to remember that these models work best in situations where there is strong autocorrelation, such as organic growth. The example below is a classic application of a (1,1,0) ARIMA model: used in econometrics to forecast the future price of a commodity, in this case, the US average retail price per gallon of regular gasoline.

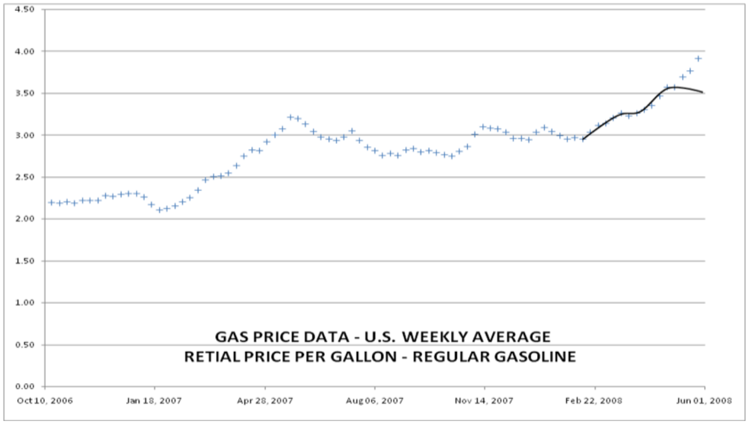

Figure 1: Weekly U.S. Average Retail Gasoline Prices

Because (1,1,0) ARIMA models rely on autocorrelation to make a forecast, they tend to work best when there is a smooth change over time. Chaotic changes introduce a large random element, decreasing the relative amount of autocorrelation and weakening the model. The failure of this very gas price model was the subject of an article in Crain’s Detroit Business in 2008; it was subsequently learned that the model performed very well so long as the changes were smooth and failed with the advent of chaotic fluctuations in gas prices in late spring of 2008.

To produce such a model in Base SAS / SAS STAT, use PROC REG recursively, stepping forward at each iteration by delta t:

```
**** (1,1,0) ARIMA MODEL IN BASE SAS ****;
**** DAVID CORLISS, UNIVERSITY OF TOLEDO DEPT. OF PHYSICS AND ASTRONOMY, 2009 ****;

DATA WORK.TSERIES;
   SET PROJECT.CRYER;
   IF MONTH = 1;
   DUMMY = 1;
   ATTRIB T INFORMAT=8.0 FORMAT=8.0;
   T = YEAR;
   ATTRIB Y INFORMAT=8.0 FORMAT=8.1;
   Y = VOLUME;
RUN;


%GLOBAL LAMBDA SIGMA;

PROC MEANS DATA=WORK.TSERIES NOPRINT;
   VAR VOLUME;
   OUTPUT OUT=WORK.RANDOM_TERM;
RUN;

DATA _NULL_; SET WORK.RANDOM_TERM;
   IF _STAT_ = MEAN THEN CALL SYMPUT ('LAMBDA' PUT(VOLUME, 8.));
   IF _STAT_ = STD THEN CALL SYMPUT ('SIGMA' PUT(VOLUME, 8.));
RUN;


%MACRO AC(N);

PROC SORT DATA=WORK.TSERIES;
   BY DUMMY;
RUN;

DATA WORK.LAST;
   SET WORK.TSERIES;
   BY DUMMY;
   IF LAST.DUMMY;
   RECENT = _N_ - &N. + 1;
   KEEP DUMMY RECENT;
RUN;

DATA  WORK.RECENT;
   MERGE WORK.TSERIES WORK.LAST;
   BY DUMMY;
   IF _N_ GE RECENT;
   DROP RECENT;
RUN;


PROC REG DATA=WORK.RECENT NOPRINT;
   MODEL Y=T;
   OUTPUT OUT=WORK.TREND PREDICTED=AUTOREG_TERM RESIDUAL=RESIDUAL;
RUN;
```

```
DATA WORK.TREND;
   SET WORK.TREND;
   OUTPUT;
   T_PREVIOUS = T;
   RANDOM_TERM = ((-1)**RAND(BERNOULLI,))* RAND('NORMAL',LAMBDA,SIGMA);
   Y_PREVIOUS = AUTOREG_TERM + RANDOM_TERM;
   RETAIN T_PREVIOUS Y_PREVIOUS;
RUN;

DATA WORK.NEW;
   SET WORK.TREND;
   BY DUMMY;
   IF LAST.DUMMY;
   DELTA_T = T - T_PREVIOUS;
   T = T + DELTA_T;
   DELTA_Y = Y - Y_PREVIOUS + 1;
   Y = Y + DELTA_Y;
   KEEP T Y DUMMY;
RUN;

DATA WORK.TSERIES;
   SET WORK.TSERIES WORK.NEW;
RUN;


%MEND AC;
```

ARIMA MODELS USING A MOVING AVERAGE

In the standard classification system for ARIMA models, the third digit is the number of Moving Average terms. So, (0,1,1) ARIMA model combines Moving Average and a random element *without* the use of a Auto-Regression component.

While (1,1,0) ARIMA models tend to work best when there is a smooth change over time, Moving Average models often are better suited to data with chaotic fluctuations. Following the failure of the (1,1,0) ARIMA model used in the earlier example, a (0,1,1) model to forecast gas prices was successfully implemented in the summer of 2008. At present, both the (1,1,0) and (0,1,1) ARIMA models are in use, each making their own forecast for gas prices.

The first step in producing a (0,1,1) ARIMA model in Base SAS is to create a moving average with an appropriate number of points. This code develops a Moving Average routine from the SAS institute to create a moving average using eleven points.

```
**** MOVING AVERAGE MACRO - USES MOVING AVERAGE CODE FROM THE SAS INSTITUTE ****;

DATA TEMP;
    DO X=1 TO 4 BY 0.1;  SUMX+X;  OUTPUT;  END;
 RUN;


DATA NEW;
   IF _N_=1 THEN DO;
      DO N=1 TO 11;
         SET TEMP;  AVERAGE=SUMX/N;  OUTPUT;  END;  END;
   ELSE DO;
      MERGE TEMP TEMP(FIRSTOBS=12 RENAME=(SUMX=SUMX2));
      IF SUMX2 ^= . ;
      AVERAGE=(SUMX2-SUMX)/11;  OUTPUT;  END;
RUN;
```

Next, this Moving Average routine is used as part of a macro for creating a (0,1,1) ARIMA model in base SAS. The number of successive points to be used in the moving average is a parameter in the macro.

```
**** (0,1,1) ARIMA MODEL IN BASE SAS ****;

**** DAVID CORLISS, UNIVERSITY OF TOLEDO DEPT. OF PHYSICS AND ASTRONOMY, 2009****;

DATA WORK.TSERIES;
   DO X=1 TO 15 BY 0.1;  OUTPUT;  END;
RUN;



%MACRO MA(N);

DATA WORK.TEMP;
   SET WORK.TSERIES;
   SUMX+X;
RUN;

DATA WORK.TREND;
%LET M = %EVAL(&N+1);
   IF _N_=1 THEN DO;
      DO N=1 TO &N.;
         SET WORK.TEMP;  AVERAGE = SUMX / N;
      OUTPUT;  END;  END;
   ELSE DO;
      MERGE WORK.TEMP WORK.TEMP(FIRSTOBS=&M. RENAME=(SUMX=SUMX2));
      IF SUMX2 ^= . ;
      AVERAGE = (SUMX2 - SUMX) / &N.;
   OUTPUT;  END;
   MA_TERM = AVERAGE;
   RETAIN MA_TERM;
RUN;

DATA WORK.TREND;
   SET WORK.TREND;
   DUMMY = 1;
   RESIDUAL = MA_TERM - X;
   FORECAST = MA_TERM - RESIDUAL;
RUN;

DATA WORK.NEW;
   SET WORK.TREND;
   BY DUMMY;
   IF LAST.DUMMY;
   X = X - (RESIDUAL / ((&N. + 1) / 2));
   KEEP X;
RUN;

DATA WORK.TSERIES;
   SET WORK.TSERIES WORK.NEW;
   KEEP X;
RUN;

%MEND MA;
```

MOVING AVERAGE EXAMPLE

The Cryer Milk Data (Cryer, J.D., *Time Series Analysis*, Duxbury Press, Belmont, 1986, p. 269) is a well-known dataset often used to illustrate techniques in Time Series Analysis. This data is a compilation of dairy farm’s monthly milk per cow, tracked over a fourteen year period from 1962 to 1975.

```
**** EXAMPLE: CRYER MILK DATA ****;

DATA WORK.CRYER;
   INFILE '/ford/.u_01/dcorlis3/Cryer.txt' DSD DLM='09'X LRECL=80 FIRSTOBS=2
   OBS=145 TRUNCOVER;
   INPUT
      DAY    :8.0
      MONTH  :8.0
      YEAR   :8.0
      VOLUME :8.0
      ;
   DATE = MDY(MONTH,DAY,YEAR);
   X = VOLUME;
   KEEP DATE X;
RUN;

PROC FREQ DATA=WORK.CRYER;
   TABLES DATE;
   FORMAT DATE YYMM.;
RUN;

DATA WORK.X;
   SET WORK.CRYER;
   KEEP X;
RUN;

%MA(12);


PROC PRINT DATA=WORK.TREND;
   VAR FORECAST;
RUN;
```

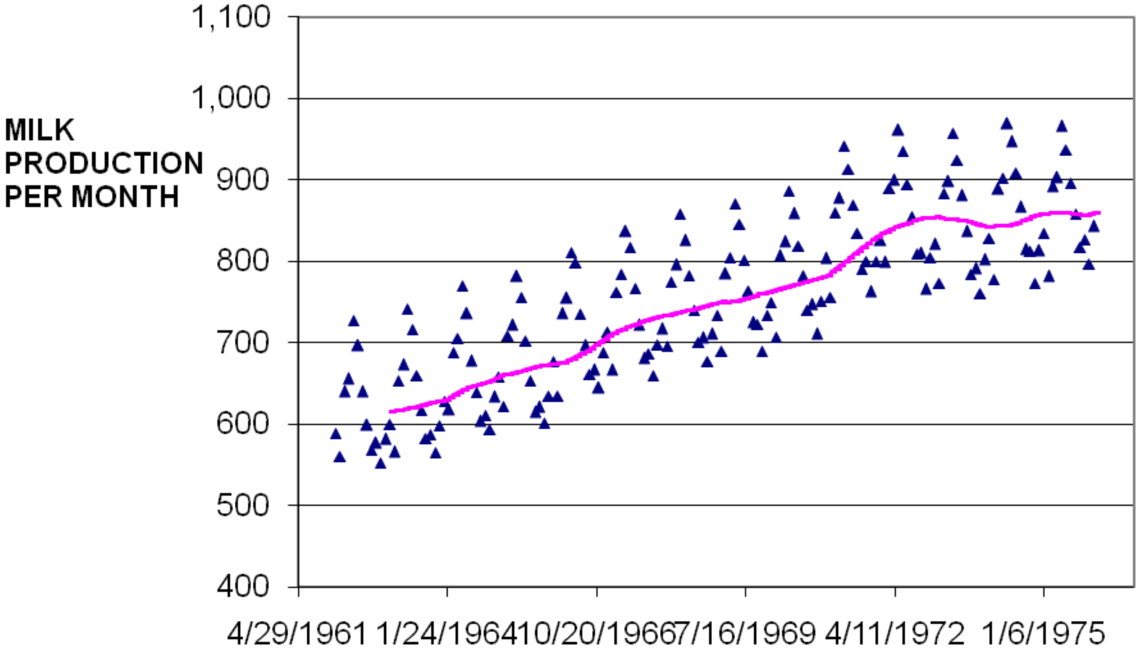

Figure 2: Cryer Milk Data (Cryer, J.D., *Time Series Analysis*, Duxbury Press, Belmont, 1986, p. 269). The observed values from the Cryer data are marked with blue triangles; a (0,1,1) ARIMA model (pink line) plots the long-term trend without correction for seasonal variation.

This data is highly periodic, varying over a period of 12 months: there is marked seasonal variation. Here, use of a (0,1,1) ARIMA model without correction for seasonality is able to the long-term trend. Because of this, variations in the trend, previously obscured by seasonal variation, are now apparent (e.g., the above-average performance in 1972).

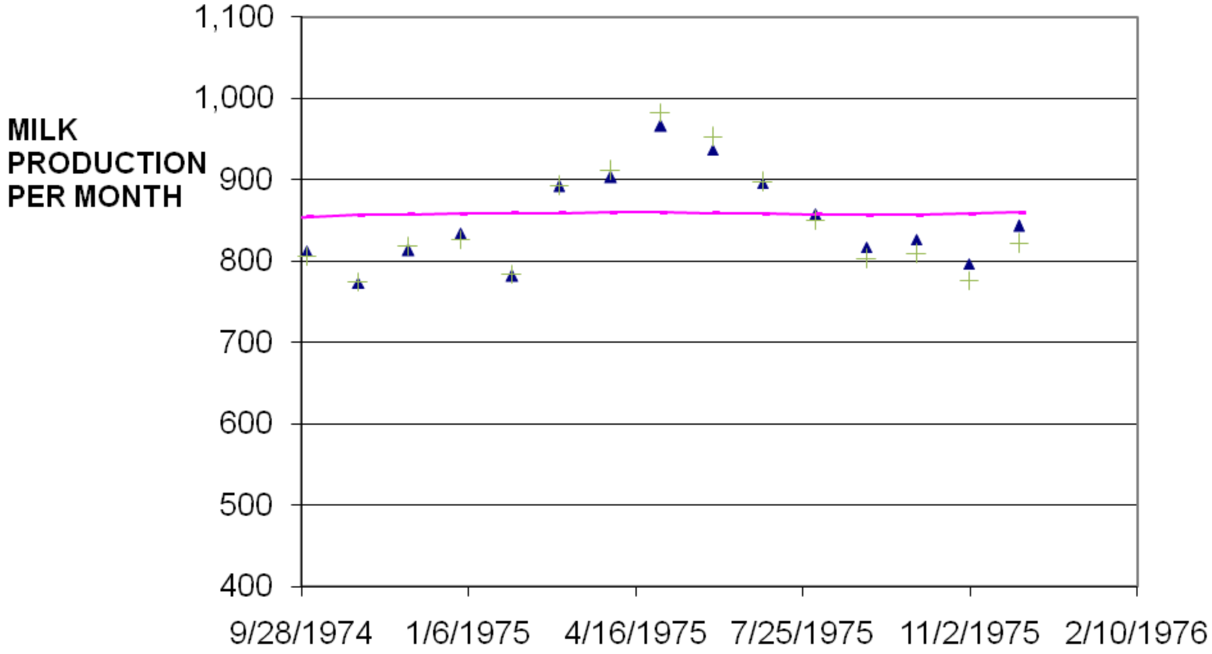

Figure 2: Cryer Milk Data (Cryer, J.D., *Time Series Analysis*, Duxbury Press, Belmont, 1986, p. 269). The observed values from are marked with blue triangles; a (0,1,1) ARIMA model (pink line) plots the long-term trend without correction for seasonal variation; the seasonally-corrected forecast is indicated by the green crosses.

After adding a correction for seasonal variation, the (0,1,1) ARIMA model is able to incorporate both the long-term trend and the monthly variation, resulting in a highly accurate forecast.

CONCLUSION

Time Series Regression, Auto-Regressive and Moving Average ARIMA models provide powerful tools for time series analysis, with each methodology is optimal for different circumstances. These techniques can be performed at a beginning level using only Base SAS and SAS / STAT, providing a firm grounding in the methodology and many useful results. Use of these techniques can also provide a path to the use of the advanced functionality of SAS / ETS.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the author at:

David Corliss
Marketing Associates
777 Woodward Avenue, Suite 500
Detroit, MI 48226
(313) 202 - 6323
dcorliss@marketingassociates.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies. 5053