Paper 205-2009

Simulating Time Series Testing Using SAS® - Part I
The Augmented Dickey-Fuller (ADF) Test

Ismail E. Mohamed, L3 Communications-ETIS, Reston, VA

ABSTRACT

The purpose of this series of articles is to present simple discussion and to present SAS programming techniques specifically designed to simulate the steps involved in time series data analysis. Part I of this series will cover the Augmented Dickey-Fuller (ADF) test of time series variables (stationarity test). Part II will continue the discussion on how to move further beyond the ADF testing and will focus discussion on examining time series variables long-run relationships (cointegration). A third part of this series is intended and will discuss how to develop an error correction model (ECM), a mechanism and concept discussed by many authors including Granger (1983), and Banergee et al 1993, that is utilized by many to determine time series short-run deviations from long-run equilibrium. The simplified SAS techniques covered in all 3 parts of this series can be used with the more complex SAS routines such as PROC ARIMA**,** which require high level of research and analysis expertise (Bails & Peppers, 1982).

INTRODUCTION

Time series data analysis has many applications in many areas including studying relationship between wages and house prices, profits and dividends, and consumption and GDP. Many analysts erroneously use the framework of linear regression (OLS) models to predict change over time or extrapolate from present conditions to future conditions. Extreme caution is needed when interpreting the results of regression models estimated using time series data. Statisticians and analysts working with time series data uncovered a serious problem with standard analysis techniques applied to time series. Estimation of parameters of the Ordinary Least Square Regression (OLS) model produces statistically significant results between time series that contain a trend and are otherwise random. This finding led to considerable work on how to determine what properties a time series must possess if econometric techniques are to be used. One basic conclusion was that any times series used in econometric applications must be stationary (Granger and Newbold, 1974). This paper will discuss a simple SAS framework to assist SAS programmers in understanding and modeling time series data on a univariate series.

$$Y_t = \alpha + \beta X_t + \varepsilon_t \qquad (1)$$

Basics and Terminology

Time series datasets are different from other ordinary datasets in that their observations are recorded sequentially over equal time increments (daily, weekly, monthly, quarterly, annually …etc). A simple example of a time series dataset (RawData) is illustrated below. Each of x and y is called a series, while the combination of the 2 variables YEAR and QTR represent the sequential equal time increments. If x and y series are both non-stationary random processes (integrated), then modeling the x, y relationship as a simple OLS relationship as in equation 1 will only generate a spurious regression. Granger and Newbold (1974) introduced the notion of a spurious regression which they argued "produces statistically significant results between series that contain a trend and are otherwise random". Time series stationarity is the statistical characteristics of a series such as its mean and variance over time. If both are constant

YEAR	QTR	*X*	*Y*
1987	4	-0.05294	0.067891
1988	1	-0.14696	0.063533
1988	2	-0.12600	0.065794
1988	3	-0.14656	0.060760
1988	4	-0.06056	0.062053
1989	1	-0.02644	0.057527
1989	2	-0.05778	0.049068
1989	3	0.01924	0.061497
1989	4	-0.10823	0.060421
.	.	.	.

x and y are two time series variables

over time, then the series is said to be a stationary process (i.e. is not a random walk/has no unit root), otherwise, the series is described as being a non-stationary process (i.e. a random walk/has unit root). Differencing techniques are normally used to transform a time series from a non-stationary to stationary by subtracting each datum in a series from its predecessor. As such the set of observations that correspond to the initial time period (t) when the measurement was taken is describes as the series level. Differencing a series using differencing operations produces other sets of observations such as the first-differenced values, the second-differenced values and so on.

x level	x_t
x 1st-diiferenced value	$x_t - x_{t-1}$
x 2nd-diiferenced value	$x_t - x_{t-2}$

If a series is stationary without any differencing it is designated as I(0), or integrated of order **0.** On the other hand, a series that has stationary first differences is designated I(1), or integrated of order 1. Stationarity of a series is an important phenomenon because it can influence its behavior. For example, the term 'shock' is used frequently to indicate an unexpected change in the value of a variable (or error). For a stationary series a shock will gradually die away. That is, the effect of a shock during time '*t*' will have a smaller effect in time '*t+1*', a smaller effect in time '*t+2*', etc. Since the data used in this paper assumed to represents time series data. Each series in equation 1 namely, *x* and *y* requires examinations at level for stationarity before proceeding further to investigate the relationship between the two variables (the OLS regression analysis). In this specification, because the data used by the paper is a quarterly series, stationarity testing will be conducted at level for up to 5-lagged periods. The stationarity test will utilize the Augmented Dickey-Fuller (ADF) technique (Dickey and Fuller (1981) which is a general auto-regression model formulated in the following regression equation (Dickey and Fuller (1981)

$$\Delta x_{i,t} = \kappa x_{i,t-1} + \sum_{k=1}^{5} \varpi_{i,k} \Delta xi_{,t-k} + \varepsilon_{k,t} \qquad (2)$$

The model hypotheses of interest are: The Series is

H_O: Non-stationary

H_A: Stationary

ADF Statistics is compared to Critical values to draw conclusions about Stationarity (see Dickey and Fuller, 1979 for the critical values)

An anatomy of an ADF Equation

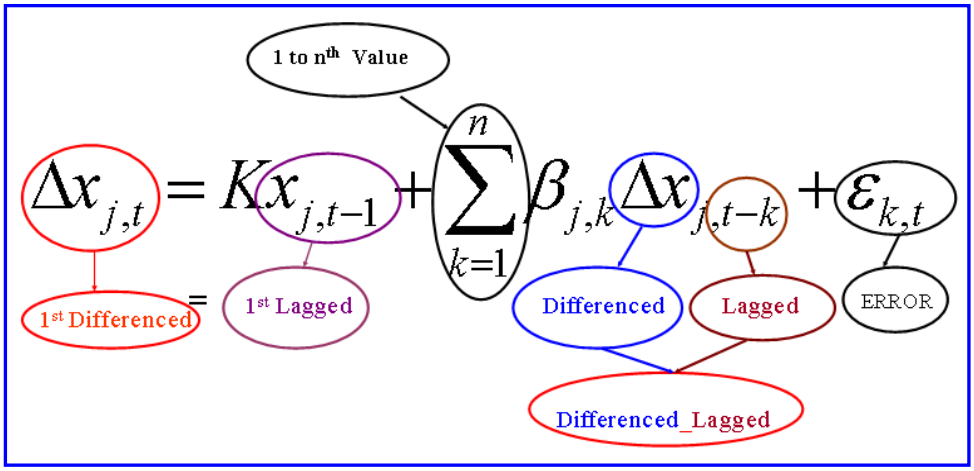
$\Delta x_{i,t} =$ This is the *1st-differenced value* of x

$\kappa x_{i,t-1} +$ This is the *1st-lagged value* of x

$+\sum_{k=1}^{5} \varpi_{i,k} \Delta x_{i,t-k}$ These are the *1st, 2nd, 3rd, 4th, & 5th-lagged* of *1st-differenced of* values of x

$+\varepsilon_{k,t}$ This is the *error* term

The above elements can be easily seen in the following chart.



SAS Techniques

As it was mentioned earlier that our sample data is quarterly spaced, this dictates that five lagged differences have to be included in testing of stationarity of both series (x and y) for more explanatory power. The following SAS Data step creates the first lagged, the first differenced and the five lagged-differenced values of the x series. Similar step is needed to create the same variables form the y series. The SAS Data step exploits the power of SAS LAG and DIF functions to create the set of the lagged and differenced values of x. SAS *LAG* function simply looks back in the dataset nth number of records and allows you to obtain a previous value of a variable and store it in the current observation. '*n*' refers to the number of records back in the data and can be an integer from 1 to 99. Many times the only thing you want to do with a previous value of a variable is to compare it with the current value to compute the difference. It is always recommended that the LAG and DIF functions not to be executed conditionally because they could cause unexpected results. If you have to use them with conditional processing of a dataset, first execute the functions and assign their results to a new variable, then use the new variable for the conditional processing.

The DIF$_n$ function works the same way as LAG$_n$, but rather than simply assigning a value, it assigns the difference between the current value and a previous value of a variable. The statement

$$A_t = DIF_n(X)$$

tells SAS that A_t, should equal the current value of x minus the value x had nth number of records back in the time.

Both LAG and DIF functions should only be used on the right hand side of assignment statements and again should not be executed conditionally.

```
DATA TimeSeries;
     SET RawData;

             x_1st_LAG               = LAG1(x);
             x_1st_DIFF              = DIF1(x);
             x_1st_DIFF_1st_LAG      = DIF1(LAG1(x));
             x_1st_DIFF_2nd_LAG      = DIF1(LAG2(x));
             x_1st_DIFF_3rd_LAG      = DIF1(LAG3(x));
             x_1st_DIFF_4th_LAG      = DIF1(LAG4(x));
             x_1st_DIFF_5th_LAG      = DIF1(LAG5(x));
RUN;
```

Simple SAS Data Step creates lagged, differenced, and differenced lagged variables from x series

SAS Output – (partial): 1st_lagged, 1st_differenced, and the 1st – 5th_lagged values of the 1st_differenced value of x

YEAR	QTR	X	x_1st_ LAG	x_1st_ DIFF	x_1st_ DIFF_ 1st_LAG	x_1st_ DIFF_ 2nd_LAG	x_1st_ DIFF_ 3rd_LAG	x_1st_ DIFF_ 4th_LAG	x_1st_ DIFF_ 5th_LAG
1987	4	-0.05294	.	.	.	.	.	.	.
1988	1	-0.14695	-0.05294	-0.09402	.	.	.	.	.
1988	2	-0.12600	-0.14695	0.02096	-0.09402	.	.	.	.
1988	3	-0.14655	-0.12600	-0.02057	0.02096	-0.09402	.	.	.
1988	4	-0.06055	-0.14655	0.08600	-0.02057	0.02096	-0.09402	.	.
1989	1	-0.02644	-0.06055	0.03412	0.08600	-0.02057	0.02096	-0.09402	.
1989	2	-0.05778	-0.02644	-0.03134	0.03412	0.08600	-0.02057	0.02096	-0.09402
1989	3	0.01924	-0.05778	0.07702	-0.03134	0.03412	0.08600	-0.02057	0.02096
1989	4	-0.10823	0.01924	-0.12748	0.07702	-0.03134	0.03412	0.08600	-0.02057
1990	1	-0.04056	-0.10823	0.06767	-0.12748	0.07702	-0.03134	0.03412	0.08600
1990	2	-0.03390	-0.04056	0.00666	0.06767	-0.12748	0.07702	-0.03134	0.03412
1990	3	-0.06903	-0.03390	-0.03513	0.00666	0.06767	-0.12748	0.07702	-0.03134
1990	4	0.07547	-0.06903	0.14451	-0.03513	0.00666	0.06767	-0.12748	0.07702
1991	1	0.03567	0.07547	-0.03981	0.14451	-0.03513	0.00666	0.06767	-0.12748
1991	2	0.09819	0.03567	0.06252	-0.03981	0.14451	-0.03513	0.00666	0.06767

Next the SAS REG procedure, one of many regression procedures in the SAS System is used in the analysis to regress the lagged and differenced values of x generated by the above data step. The regression model used here was set as a relationship in which the value of x at the preceding time period (lagged value of x) is the dependent variable and the independent variables are the set of 5 previous-differenced values of the x series. This analysis provides a "best-fit" mathematical equation for the relationship exhibited in Eq (2).

```
PROC REG DATA = TimeSeries;
         MODEL x_1st_DIFF = x_1st_LAG
                            x_1st _DIFF_1st _LAG
                            x_1st _DIFF_2nd_LAG
                            x_1st _DIFF_3rd_LAG
                            x_1st _DIFF_4th_LAG
                            x_1st _DIFF_5th_LAG;
RUN;
```

SAS REG procedure showing Stationarity test at level, with fixed 5 Lag Length and a Constant

Discussion

The 'x_1st_LAG' t-value generated by the above regression model corresponds to the Augmented Dickey-Fuller test (ADF) Statistics. Compare this t-value to the Critical Values (see Dickey and Fuller, 1979 for the critical values) to test the 2 Hypothesis that the *x* series is:

H_O: Non-Stationary
H_A: Stationary

In our example the t-value of **(-1.83)** is greater than the Critical Values (CVs) at 1%, 5%, and 10% significant level (-3.524233, -2.902358, and -2.588587 respectively). We would fail to reject the null hypothesis and conclude that the *x* series is a non-stationary process when tested at level.

What is Next?

If we fail to reject the null hypothesis, and concluded that x and perhaps y are non-stationary series, we would have to difference each series once, create set of lagged and differenced variables as shown in the earlier SAS data step this time from the differenced-values of each series, and finally carry out the ADF test (testing the series stationarity at its first-differenced value). Differencing of a series normally transforms it from non-stationarity to stationarity. A differenced stationary series is said to be *integrated* and is denoted as I(d) where 'd' is the order of integration. The order of integration is the number of unit roots contained in the series, or the number of differencing operations it takes to make the series stationary. For our purpose here, since we will difference our example series once, there is one unit root, so it is an I(1) series. Once both *x* and *y* determined non-stationary at their level, we will move further to examine the nature of their linear combination. Specifically we will be interested in examining the linear combination between the non-stationary x and y, if such a linear combination exists, then x and y series are said to be cointegrated. The linear combination between them is the cointegrating equation and may be interpreted as the long-run equilibrium relationship among the 2 variables. Fortunately, this test can also be accomplished using the Augmented Dickey-Fuller test and will be the subject of discussion of the second part of this series of articles.

```
NULL HYPOTHESIS:  'x' has a unit root
LAG LENGTH: 5 (FIXED)
AUGMENTED DICKEY-FULLER TEST STATISTICS, TEST CRITICAL VALUES:
1% LEVEL T-STATISTICS  = -3.524233
5% LEVEL T-STATISTICS  = -2.902358
10% LEVEL T-STATISTICS = -2.588587
LEVEL WITH 5 LAGS
The REG Procedure
Model: MODEL1
Dependent Variable: x_1st_DIFF
Number of Observations Read                          78
Number of Observations Used                          72
Number of Observations with Missing Values            6
                          Analysis of Variance
                                   Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F

Model                       6      0.08731        0.01455      21.25    <.0001
Error                      65      0.04451     0.00068479
Corrected Total            71      0.13182
Root MSE              0.02617    R-Square     0.6623
Dependent Mean        0.00172    Adj R-Sq     0.6312
Coeff Var          1518.81011


                       Parameter Estimates

                                Parameter       Standard
Variable                   DF    Estimate          Error    t Value    Pr > |t|

Intercept                   1     0.00916        0.00422       2.17      0.0338
x_1st_LAG                   1    -0.16361        0.08960      -1.83      0.0724
x_1st_DIFF_1st_LAG          1    -0.43485        0.13151      -3.31      0.0015
x_1st_DIFF_2nd_LAG          1     0.11255        0.10735       1.05      0.2983
x_1st_DIFF_3rd_LAG          1     0.23609        0.10676       2.21      0.0305
x_1st_DIFF_4th_LAG          1    -0.42082        0.10964      -3.84      0.0003
x_1st_DIFF_5th_LAG          1    -0.12741        0.10698      -1.19      0.2380
```

SAS Output – Regression Analysis (Unit Root Test) –Level with 5 Lags

EVIEWS[1]® code and output for comparison

Code *Uroot(adf,const,lag=5,save=mout)*

```
Null Hypothesis: X has a unit root
Exogenous: Constant
Lag Length: 5 (Fixed)

                                                t-Statistic   Prob.*

Augmented Dickey-Fuller test statistic          1.826060     0.3662
Test critical values:     1% level             -3.524233
                          5% level             -2.902358
                          10% level            -2.588587

*MacKinnon (1996) one-sided p-values.


Augmented Dickey-Fuller Test Equation
Dependent Variable: D(X)
Method: Least Squares
Date: 04/30/00   Time: 15:36
Sample (adjusted): 1988Q3 2006Q2
Included observations: 72 after adjustments

     Variable      Coefficient     Std. Error     t-Statistic     Prob.

         X(-1)       0.033676       0.018446       1.826060     0.0724
       D(X(-1))      0.434090       0.101500       4.277041     0.0001
       D(X(-2))      0.112580       0.107952       1.042816     0.2981
       D(X(-3))      0.236089       0.106755       2.211500     0.0305
       D(X(-4))     -0.420817       0.109643      -3.838076     0.0003
       D(X(-5))     -0.127411       0.106876      -1.191923     0.2380
              C      0.009161       0.004224       2.168795     0.0338

R-squared            0.662320     Mean dependent var      0.001723
Adjusted R-squared   0.631159     S.D. dependent var      0.043088
S.E. of regression   0.026168     Akaike info criterion  -4.356357
Sum squared resid    0.044571     Schwarz criterion      -4.135015
Log likelihood       163.8289     F-statistic             21.24918
Durbin-Watson stat   1.991931     Prob(F-statistic)       0.000000
```

[1] EVIEWS® is an econometrics & Time Series Analysis software package by Quantitative Micro Software. http://www.eviews.com/index.html

References

Bails, Dale G. and Larry C. Peppers (1982) Business Fluctuations: Forecasting Techniques and Applications, Englewood Cliffs NJ: Prentice-Hall Inc.
Banergee, A. Dolado, J., Galbraith, J.W. and Hendry, D.F. (1993), "Co-integration, Errorcorrection, and the Econometric Analysis of Time Series," Oxford University Press, Oxford.
Dickey, D. and W. Fuller (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root, Journal of the American Statistical Association, 74, 427-431.
Fuller, W. (1996). Introduction to Statistical Time Series, Second Edition. John Wiley, New York.
Granger, C.W.J., and P. Newbold(1974). Spurious regressions in econometrics Journal of Econometrics, 2, 111-120.
Granger, Clive W.J., 1983, Co-integrated variables and error-correcting models, University of California, San Diego, Department of Economics Working Paper: 83-13
Hamilton (1994). Time Series Analysis, Princeton University Press.
Phillips, P.C.B. (1987). Time Series Regression with a Unit Root, Econometrica, 55, 227-301.

Acknowledgements

My sincere thanks to everyone I have had the pleasure of exchanging time Series analysis related ideas with in recent years. Special thanks to Theresa Diventi, Ian Keith both with the Financial Institutions Regulation Division, Kee N. Cheung with the Housing Finance Analysis Division of the U.S. Department of Housing and Urban Development, and Ronald Hanson with L3 Communications, Enterprise IT Solutions (EITS), for their constructive suggestions which added much to this paper. My sincere appreciation goes to Eric Wolf and Karen Cinibulk both with L3 Communications, Enterprise IT Solutions Division (EITS) for their continuous encouragement and support.

Trademarks

 To obtain the entire SAS code and sample data, please e-mail the author. The author welcomes and encourages any questions, corrections, improvements, feedback, remarks, both on- and off-topic via email.

Contact Information

The author welcomes and encourages any questions, corrections, improvements, feedback, remarks, both on- and off-topic via email.
Name: Ismail E. Mohamed, Ph.D, Software Engineer 5
Enterprise: L3 Communications, Enterprise IT Solutions (EITS), U.S. Department of Housing & Urban Development
Address: 451 7th Street, SW, Room 8212,
City, State ZIP: Washington, DC 20410
Work Phone: (202)-402-5884
E-mail: ismail.mohamed@L-3com.com; Ismail.Mohamed@hud.gov