Time Series Analysis Using SAS®

Part I
The Augmented Dickey-Fuller (ADF) Test

By Ismail E. Mohamed

ABSTRACT
The purpose of this series of articles is to discuss SAS programming techniques specifically designed to simulate the steps involved in time series data analysis. The first part of this series will cover the Augmented Dickey-Fuller (ADF) test of time series (stationarity test). The second part will cover cointegration and error correction. The SAS techniques presented in both parts can be used with the more complex SAS routines such as PROC ARIMA, which require a high level of research and analysis expertise (Bails & Peppers, 1982).

INTRODUCTION
Time series data analysis has many applications in many areas including studying the 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 produced 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 as a univariate series (Eq 1).

\[ Y_t = \alpha + \beta X_t + \epsilon_t \] (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.

<table>
<thead>
<tr>
<th>YEAR</th>
<th>QTR</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>4</td>
<td>-0.05294</td>
<td>0.067891</td>
</tr>
<tr>
<td>1988</td>
<td>1</td>
<td>-0.14686</td>
<td>0.063533</td>
</tr>
<tr>
<td>1988</td>
<td>2</td>
<td>-0.12600</td>
<td>0.065794</td>
</tr>
<tr>
<td>1988</td>
<td>3</td>
<td>-0.14656</td>
<td>0.060760</td>
</tr>
<tr>
<td>1988</td>
<td>4</td>
<td>-0.06056</td>
<td>0.062053</td>
</tr>
<tr>
<td>1989</td>
<td>1</td>
<td>-0.02644</td>
<td>0.057527</td>
</tr>
<tr>
<td>1989</td>
<td>2</td>
<td>-0.05778</td>
<td>0.049068</td>
</tr>
<tr>
<td>1989</td>
<td>3</td>
<td>0.01924</td>
<td>0.061497</td>
</tr>
<tr>
<td>1989</td>
<td>4</td>
<td>-0.10823</td>
<td>0.060421</td>
</tr>
<tr>
<td>1990</td>
<td>1</td>
<td>-0.04056</td>
<td>0.050771</td>
</tr>
<tr>
<td>1990</td>
<td>2</td>
<td>-0.03390</td>
<td>0.036702</td>
</tr>
<tr>
<td>1990</td>
<td>3</td>
<td>-0.06903</td>
<td>0.016959</td>
</tr>
<tr>
<td>1990</td>
<td>4</td>
<td>0.07547</td>
<td>0.002585</td>
</tr>
</tbody>
</table>

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 a statistical characteristic of a series’ mean and variance over time. If both are constant 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 describes 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.

\[
x \text{ level} \\
x \text{ 1}\text{st}-\text{differenced value} = x_t - x_{t-1} \\
x \text{ 2}\text{nd}-\text{differenced value} = x_t - x_{t-2}
\]

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 still smaller effect in time ‘t+2’, etc. 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 generalized auto-regression model formulated in the
following regression equation (Dickey and Fuller (1981))

\[
\Delta x_{i,t} = \kappa x_{i,t-1} + \sum_{k=1}^{5} \sigma_{i,k} x_{i,t-k} + \varepsilon_{i,k,t} \quad (2)
\]

The model hypotheses of interest are: The Series is

\(H_0: \text{Non-stationary} \quad H_a: \text{Stationary}\)

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
\(\sum_{k=1}^{5} \sigma_{i,k} x_{i,t-k} + \varepsilon_{i,k,t} \) These are the 1st, 2nd, 3rd, 4th, & 5th-lagged of 1st-differenced of values of x
\(+ \varepsilon_{i,k,t} \) 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. A similar step is needed to create the same variables from 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 DIFn function works the same way as LAGn, 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.

```sas
DATA TimeSeries;
SET RawData;
  x_1^st_LAG = LAG1(x);
  x_1^st_DIFF = DIF1(x);
  x_1^st_DIFF_1^st_LAG = DIF1(LAG1(x));
  x_1^st_DIFF_2^nd_LAG = DIF1(LAG2(x));
  x_1^st_DIFF_3^rd_LAG = DIF1(LAG3(x));
  x_1^st_DIFF_4^th_LAG = DIF1(LAG4(x));
  x_1^st_DIFF_5^th_LAG = DIF1(LAG5(x));
RUN;
```

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

<table>
<thead>
<tr>
<th>YEAR</th>
<th>QTR</th>
<th>X</th>
<th>LAG</th>
<th>DIFF</th>
<th>1^st_LAG</th>
<th>2^st_LAG</th>
<th>3^rd_LAG</th>
<th>4^th_LAG</th>
<th>5^th_LAG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>4</td>
<td>-.05294</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1988</td>
<td>1</td>
<td>-.14696</td>
<td>-.05294</td>
<td>-.09402</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1988</td>
<td>2</td>
<td>-.12000</td>
<td>-.14696</td>
<td>0.02096</td>
<td>-.09402</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1989</td>
<td>1</td>
<td>-.09804</td>
<td>-.06056</td>
<td>0.03412</td>
<td>-0.02057</td>
<td>0.02096</td>
<td>-.09402</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1990</td>
<td>2</td>
<td>-.05778</td>
<td>-.02644</td>
<td>-.03134</td>
<td>0.03412</td>
<td>0.08600</td>
<td>-.02057</td>
<td>0.09086</td>
<td>-.09402</td>
</tr>
</tbody>
</table>

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). SAS REG procedure for Unit Root Test at level, with fixed 5 Lag Length and a Constant:

```sas
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;
QUIT;
```

**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 hypotheses that the x series is:

- **H₀**: Non-Stationary
- **H₁**: 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 a 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.
SAS Output – Regression Analysis (Unit Root Test) –Level with 5 Lags

**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

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>0.08731</td>
<td>0.01455</td>
<td>21.25</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>65</td>
<td>0.04451</td>
<td>0.00068479</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>71</td>
<td>0.13182</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Root MSE</td>
<td></td>
<td>0.02617</td>
<td>R-Square</td>
<td>0.6623</td>
<td></td>
</tr>
<tr>
<td>Dependent Mean</td>
<td></td>
<td>0.00172</td>
<td>Adj R-Sq</td>
<td>0.6312</td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td></td>
<td>1518.81011</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

| Variable        | DF | Parameter Estimate | Standard 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|

**EVIEWS® CODE AND OUTPUT FOR COMPARISON**
Similarly, EVIEWS or other SAS time series tools can be used to carry out the same test. The following EVIEWS Code can be used to carry out the ADF test. Results of this code are shown in the next page.

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

---

1 EVIEWS® is an econometrics & Time Series Analysis software package by Quantitative Micro Software. http://www.eviews.com/index.html
The hypothesis that x has a unit root cannot be rejected.