Paper 204-2009

Simulating Time Series Analysis Using SAS® - Part II Cointegration

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

ABSTRACT

The purpose of this series of articles is to present simple discussion and SAS programming techniques specifically designed to simulate the steps involved in time series data analysis. Part I of this series covered 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 examine the time series variables long-run relationships (cointegration). A third part of this series is intended and will discuss how to develop an error correction mechanism (ECM), a concept discussed by many authors including Granger (1983), and Banergee et al 1993, that used to determine time series short-run deviations from long-run equilibrium. The simple 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

Empirical research in financial economics and many others sectors is largely based on time series data. Such data represents a phenomenon is collected over long and different periods of time. Many analysts erroneously use the framework of linear regression (OLS) models to model variables of such data and to predict change over time or extrapolate from present conditions to future conditions. Part I of this series cautioned against interpretation of the results of regression models estimated using time series data and suggested a simple framework to assist SAS programmers in understanding, modeling, and carrying out Stationarity testing (ADF) using a time series data on a univariate series (Mohamed, 2008)

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

If ‘*x*’ is	And ‘*y*’ is	Model the relationship As
Stationary	Stationary	OLS Regression
non-Stationary	non-Stationary	Co-integration
Stationary	non-Stationary	Logically Inconsistent[1]
non-Stationary	Stationary	Logically Inconsistent

Part I introduced a simple discussion on how to use a simple SAS data step together with SAS PROC REG to conduct a staionarity test on time series variables (ADF test). It also examined and discussed the regression results and related them to the ADF testing in details. Part II will start by suggesting a general framework (shown in the above table) to determine the path forward with time series analysis after concluding the ADF testing. For useful issues associated with unit root testing please refer to Phillips and Xiao (1998). First we will establish some criteria to differentiate three possible outcomes of the ADF results which either will lead us to move with our

[1] A set of statements are logically consistent if and only if it is possible for all of them to be true at the same time.

analysis forward to the next step or will simply suggest to us to stop (refer to Table 1 above for simple summary)

i. If both x and y determined to be stationary at their level I(0), it is not necessary to proceed further since standard time series methods apply to stationary variables i.e. we will simply apply a classical OLS regression analysis.

ii. If both x and y determined to be non-stationary at their level, and further examination revealed they are integrated of same order; usually integrated of the first order I(1), that is they become stationary after we apply a first differencing, we will have to examine the nature of x and y 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. In this case we say both x and y are integrated of the same order and consequently, we have to proceed with our analysis further to estimate x and y long-run equilibrium relationship

iii. If x and y are integrated of different order, it is inevitable to conclude that they are not cointegrated (refer to Table 1 and footnote on page 1)

BASICS AND TERMINOLOGY

Part I introduced and defined the concepts of time series, spurious regression, stationarity, differencing techniques, order of integration, and stationarity. Part II will start by reiterating that 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). For simplicity we introduce this example of a time series dataset (REG_SERIES).

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

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, introduced by Granger and Newbold (1974) who argued that "spurious regression 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 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).

Stationarity testing uses 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) and was introduced and discussed in simple details in Part I.

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

In this Part we introduce the concept 'cointegration' in simple terms by simply stating that if there exists a stationary linear combination between 2 non-stationary time series, the 2 variables combined are said to be 'cointegrated' (Granger (1981, 1983). In other words, the 2 series are cointegrated when each has been differenced once and both become stationary at that point, and the 2 variables move together in the long-run (co-move)

SAS TECHNIQUES

The stationarity of a time series has important implication for regression analysis since the classical tests of regression analysis, such as the t-test and f-test, are based on the assumption that time series are stationary. Consequently, the validity of coefficients on explanatory variables is based on stationary series. If, however, a time series process exhibit non-stationarity, standard test statistics are no longer valid and concerns arise over interpreting coefficients that are spurious. If the ADF testing indicates that both x and y series are non-stationary then modeling the cointegration between non-stationary variables such as in equation 3, provides one approach for obtaining useful regression results. Despite the two variables x and y being individually non-stationary, a linear combination of the two can be stationary. In this case a conintegrating link is said to exist and suggests there is a long run, or equilibrium, relationship between the two variables. After testing the variables for their order of integration using the Augmented Dickey-Fuller tests (ADF) discussed in details in Part I we will use Engle and Grager (1987) two-step procedure to test for cointegration between the two variables x and y. In the first step we will model relationship between the two variables (cointegration equation) such as in equation (1) In the second step, the ADF test will be used to test for stationarity of the residual or the leftover deviations resulted after fitting the regression model. The significance of this is that, if y and x are linked by a long-run relationship, the coefficient of the regression is valid though slightly bias. To summarize the Engle and Grager (1987) 2-step procedure:

i. Estimate a relationship' $y_t = \alpha + \beta x_t$ and get the residuals series (ϵ_t) of the regression
ii. Apply stationarity test on the residuals series (ϵ_t): If (ϵ_t) series is non-stationary then we will reject cointegration.

For illustration purposes we will use the hypothetical dataset REG_SERIES (partial data table)
Step 1: Estimate the long-run relationship $y_t = \alpha + \beta x_t$ and extract the residuals (ϵ_t)

```
PROC REG DATA= REG_SERIES;
MODEL y = x;
OUTPUT OUT = RESIDS
R = y_residuals;
RUN;
QUIT;
```

SAS codes to model the long-run relationship' $y_t = \alpha + \beta x_t$

The above **PROC REG** with OUTPUT OUT = RESIDS option will create a SAS data set RESIDS that will save residuals **'y_residuals'**, calculated as 'actual' minus 'predicted' and produces the following dataset (partial table):

```
YEAR    QTR          x              y          y_residuals
1987     4       -0.05294       0.067891        0.038569
1988     1       -0.14696       0.063533       -0.063425
1988     2       -0.12600       0.065794       -0.038328
1988     3       -0.14656       0.060760       -0.068098
1988     4       -0.06056       0.062053        0.020268
1989     1       -0.02644       0.057527        0.046107
1989     2       -0.05778       0.049068       -0.000710
1989     3        0.01924       0.061497        0.099050
1989     4       -0.10823       0.060421       -0.030388
1990     1       -0.04056       0.050771        0.019626
1990     2       -0.03390       0.036702        0.000545
1990     3       -0.06903       0.016959       -0.070708
1990     4        0.07547       0.002585        0.047493
```

Estimated Residual series resulted from fitting the x and y regression in step 1

For each observation in our original dataset, we now have a corresponding residual - **y_residuals** (or an error term). Remember that this residual represents the unexplained (or residual) variation after fitting the regression y and x model. It is the difference (or left over) between the observed value of the variable and the value suggested by the regression model for each observation. It is what is binding our 2 series x and y in their long life journey. Now since we have Year and Quarter we can look at this error at different time points (residuals as a function of time). If these deviations from the long-run equilibrium are found to be stationary, then x_t and y_t are said to be cointegrated.

Step 2: stationarity test on the residuals series (ϵ_t) - residual ADF testing
In order to determine if x and y are actually cointegrated, denote the estimated residual sequence from this equation by ϵ_t. Thus ϵ_t is the series of the estimated residuals of the long-run relationship. The fact is since the residual ϵ_t is also a time series, then we can perform an ADF test of stationarity on it. The form of the ADF test is similar to the one that was discussed in Part I and can be expressed as follows:

$$\Delta\varepsilon_{i,t} = \kappa\varepsilon_{i,t-1} + \sum_{k=1}^{5} \varpi_{i,k}\,\Delta\varepsilon_{i',t-k} + \varepsilon rror_{k,t} \qquad (3)$$

The model hypotheses of interest are: The residuals **s**eries ϵ_t is

H_O: ϵ_t is Non-stationary
H_A: ϵ_t is Stationary

ADF Statistics is compared to Critical values to draw conclusions about Stationarity (see Dickey and Fuller, 1979 for the critical values).
Similar to what we did in Part I we will use the same SAS techniques to conduct the ADF test. The SAS Data step below creates the first lagged, the first differenced and the five lagged-differenced values of the y_residuals series. The SAS Data step exploits the

```
DATA TimeSeries;
     SET RESIDS;
         y_residuals_1st_LAG          = LAG1 (y_residuals);
         y_residuals_1st_DIFF         = DIF1 (y_residuals);
         y_residuals_1st_DIFF_1st_LAG = DIF1 (LAG1(y_residuals));
         y_residuals_1st_DIFF_2nd_LAG = DIF1 (LAG2(y_residuals));
         y_residuals_1st_DIFF_3rd_LAG = DIF1 (LAG3(y_residuals));
         y_residuals_1st_DIFF_4th_LAG = DIF1 (LAG4(y_residuals));
         y_residuals_1st_DIFF_5th_LAG = DIF1 (LAG5(y_residuals));
RUN;
```

SAS LAG and DIF functions to create the set of the lagged and differenced values of y_residuals

Next we will use SAS **PROC REG** again in the analysis to regress the lagged and differenced values of y_residuals generated in the above step. The regression model used here was set as a relationship in which the value of ϵ_t at the preceding time period (lagged value of y_residuals) is the dependent variable and the independent variables are the set of 5 previous-differenced values of the y_residuals series. This analysis provides a "best-fit" mathematical equation for the relationship exhibited in Eq (2).

```
PROC REG DATA = TimeSeries;
     MODEL y_residuals_1st_DIFF =   y_residuals_1st_LAG
                                    y_residuals_1st_DIFF_1st_LAG
                                    y_residuals_1st_DIFF_2nd_LAG
                                    y_residuals_1st_DIFF_3rd_LAG
                                    y_residuals_1st_DIFF_4th_LAG
                                    y_residuals_1st_DIFF_5th_LAG;
RUN;
QUIT;
```

SAS **PROC REG** for residuals ADF (stationarity) test at level, with fixed 5 Lag Length and a constant

YEAR	QTR	X	Y	y_residuals	y_residuals_ 1st_LAG	y_residuals_ 1st_DIFF	y_residuals_ 1st_DIFF_ 1st_LAG	y_residuals_ 1st_DIFF_ 2nd_LAG	y_residuals_ 1st_DIFF_ 3rd_LAG	y_residuals_ 1st_DIFF_ 4th_LAG	y_residuals_ 1st_DIFF_ 5th_LAG
1987	4	-0.05294	0.067891	0.038569	.	.	.	.	.	.	.
1988	1	-0.14696	0.063533	-0.063425	0.038569	-0.10199	.	.	.	.	.
1988	2	-0.126	0.065794	-0.038328	-0.063425	0.0251	-0.10199	.	.	.	.
1988	3	-0.14656	0.06076	-0.068098	-0.038328	-0.02977	0.0251	-0.10199	.	.	.
1988	4	-0.06056	0.062053	0.020268	-0.068098	0.08837	-0.02977	0.0251	-0.10199	.	.
1989	1	-0.02644	0.057527	0.046107	0.020268	0.02584	0.08837	-0.02977	0.0251	-0.10199	.
1989	2	-0.05778	0.049068	-0.00071	0.046107	-0.04682	0.02584	0.08837	-0.02977	0.0251	-0.10199
1989	3	0.01924	0.061497	0.09905	-0.00071	0.09976	-0.04682	0.02584	0.08837	-0.02977	0.0251
1989	4	-0.10823	0.060421	-0.030388	0.09905	-0.12944	0.09976	-0.04682	0.02584	0.08837	-0.02977
1990	1	-0.04056	0.050771	0.019626	-0.030388	0.05001	-0.12944	0.09976	-0.04682	0.02584	0.08837
1990	2	-0.0339	0.036702	0.000545	0.019626	-0.01908	0.05001	-0.12944	0.09976	-0.04682	0.02584

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

DISCUSSION

The ‘y_residuals_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 ϵ_t **(**y_residuals**)** series is:

H_O: ϵ_t is Non-stationary
H_A: ϵ_t is Stationary

In our example the t-value of (-4.24) - form SAS Output – Regression Analysis (Unit Root Test) – at Level with 5 Lags for ϵ_t (y_residuals) series is smaller than the Critical Values (CVs) at 1%, 5%, and 10% significant level (-3.524233, -2.902358, and -2.588587 respectively). We would reject the null hypothesis and conclude that the ϵ_t (y_residuals) series - is a stationary process when tested at level. Since ϵ_t, is the series of the estimated residuals of the long-run relationship between x_t and y_t (deviations from x and y long-run equilibrium) are found to be stationary, then x_t and y_t are said to be cointegrated. Now that this being said, let’s ask the following two important questions: is cointegration correlation? And if not, then what is the difference between cointegration and correlation? A simple answer would suggest that If the two time series variables x and y are really correlated, when x goes up one day, y would likely go up also on the same day, and vice versa. So it seems that x and y daily (or weekly, monthly, quarterly, or yearly) behavior would have risen or fallen in synchrony. But that’s not what this is about. If we claim that x and y are cointegrated, we mean that the two series cannot wander off in opposite directions for very long without coming back to a mean distance eventually (Carol 2001). But it doesn’t mean that on a daily basis the two series have to move in synchrony at all.

WHAT IS NEXT?

If the time series variables are found to be cointegrated, the residuals from the equilibrium regression can be used to estimate the error-correction, a convenient model or mechanism that measures the correction from disequilibrium of the previous period to analyze the long-run and short-run effects of the 2 variables. We have seen in our discussion how to test this ‘disequilibrium error’ term for stationarity, and if that is the case, this implies that there is some adjustment process that prevents this ‘errors’ in the long-run relationship becoming larger and larger. Fortunately, this error-correction mechanism can be accomplished by using a simple regression modeling technique and will be the subject of our discussion in Part III of this series of articles.

```
 NULL HYPOTHESIS:  'ϵ' 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: y_residuals_1st_DIFF

                            Analysis of Variance

                                     Sum of           Mean
Source                   DF         Squares         Square    F Value    Pr > F

Model                     6         0.36886        0.06148     104.11    <.0001
Error                    51         0.03012     0.00059050
Corrected Total          57         0.39898


Root MSE              0.02430    R-Square     0.9245
Dependent Mean    -0.00066944    Adj R-Sq     0.9156
Coeff Var          -3629.95450


                                       Parameter Estimates

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

Intercept                                  1        0.00141        0.00328       0.43      0.6696
y_residuals_1st_LAG                        1       -4.53398        1.06971      -4.24      <.0001
y_residuals_1st_DIFF_1st_LAG               1        2.19868        0.97870       2.25      0.0290
y_residuals_1st_DIFF_2nd_LAG               1        1.17839        0.79060       1.49      0.1423
y_residuals_1st_DIFF_3rd_LAG               1        0.69251        0.57193       1.21      0.2315
y_residuals_1st_DIFF_4th_LAG               1        0.32332        0.34131       0.95      0.3480
R1 y_residuals_1st_DIFF_5th_LAG            1        0.13422        0.13457       1.00      0.3233
```

The t-value is smaller than any critical value at 1%, 5%, and 10%, the hypothesis that ϵ is non-stationary is rejected

SAS Output – Regression Analysis (Stationarity Test) –Level with 5 Lags (residuals series)

REFERENCES

Alexander, Carol (2001). Market Models: A Guide to Financial Data Analysis. John Wiley & Sons

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, Error correction, 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., 1981, some properties of time series data and their use in econometric model specification, Journal of Econometrics, 16, 121-30

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.

Mohamed, Ismail E. (2008) Time Series Analysis Using SAS-Part I: The Augmented Dickey-Fuller (ADF) Test Poster presentation, 21st Annual Conference of the North East SAS User Group (NESUG), Pittsburgh, Pennsylvania, 14-17 September, 2008.

Phillips, P.C.B. (1987). Time Series Regression with a Unit Root, Econometrica, 55, 227-301.

Phillips, P.C.B. and Z. Xiao (1998). A Primer on Unit Root Testing, Journal of Economic Surveys, 12, 423-470

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, of Office policy Development and Research - US Department of Housing & Urban Development, and Ian Keith with the Federal Housing Finance Agency (FHFA), 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

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