SAS® GLOBAL FORUM 2020

MARCH 29 – APRIL 1
WASHINGTON, DC

USERS PROGRAM

Simulating Time Series Analysis Using SAS® - Part II Cointegration

Ismail E. Mohamed, Senior Financial Analyst

Federal Housing Finance Agency, National Mortgage Database (NMDB®)

The views expressed in this article are those of the author and are not necessarily those of the Federal Housing Finance Agency (FHFA)

Simulating Time Series Analysis Using SAS® Part II Cointegration

Ismail Mohamed

Ismail Mohamed currently Sr. Financial Analyst with the Federal Housing Finance Agency (FHFA) - National Mortgage Database (NMDB®). He joined the National Mortgage Database team from the U.S Department of Housing and Urban Development (HUD) where he had worked as part of L-3 Communications (L3 Technologies, Inc) contractor team for 13 years. He has over 20 years of solid technical expertise in statistical software, computer applications development, data analysis, and applied economic research.

The Problem

Many analysts mistakenly use the framework of linear regression (OLS) models to model variables of Time series data and to predict change over time or extrapolate from present conditions to future conditions.

Time series data are different from randomly selected data and time series techniques must be used to analyze time series data

Random Sample *vs* Time Series Data

Random Sample

- Randomly sampled
- *No dependency*
- Assumptions: errors are independent, variance of errors is constant
- Unless these assumptions are satisfied, results from sample data cannot be used to make inference on the relationship between population parameters.

Time Series

- **NOT** randomly sampled
- Observations come in a very particular order - time ordering, ordered time intervals
- Errors correlated over time
- Errors from one time period are carried over into future time periods (Serial correlation/auto-correlation)
- Trending data over time data series may look like they are related, but really is 'spurious' (biased coefficients)

Random Sample Data *vs* Time Series Data

Random Sample Analysis Techniques

- Simple regression - OLS technique, is primarily used to predict the relationship among population parameters

Time Series Analysis Techniques

- Autoregressive moving average ARIMA model. The general model introduced by Box and Jenkins (1976)
- when using non stationary variables in OLS you run into the potentially fatal issue of *spurious regression*
- Check for stationarity- checking for stationarity isn't about improving the accuracy of the model per se, it is about keeping the model stable

Random Sample Data *vs* Time Series Data

Analysis Techniques

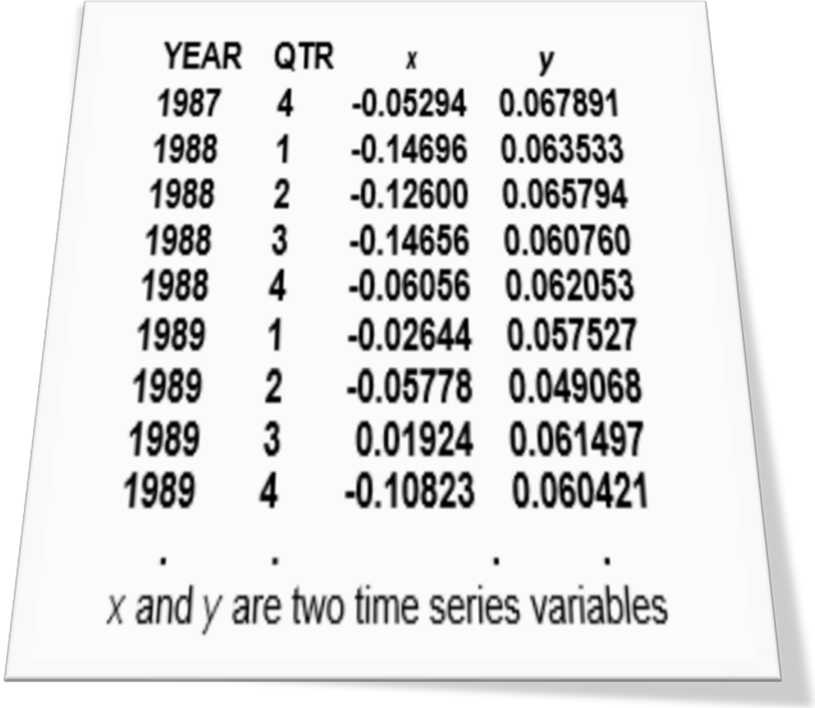
- Simple regression - OLS technique, is primarily used to predict the relationship among population parameters

Analysis Techniques

- Apply Time series analysis techniques
- Test series stationarity, a common assumption in time series techniques is that the data are stationary - For useful issues associated with stationarity please refer to Mohamed, Ismail E. (2008) Time Series Analysis Using SAS-Part I: The Augmented Dickey- Fuller (ADF) Test, 21st Annual Conference of the NESUG
- Deal with periodic fluctuations
- ..
- Model the relationship

Univariate Time Series

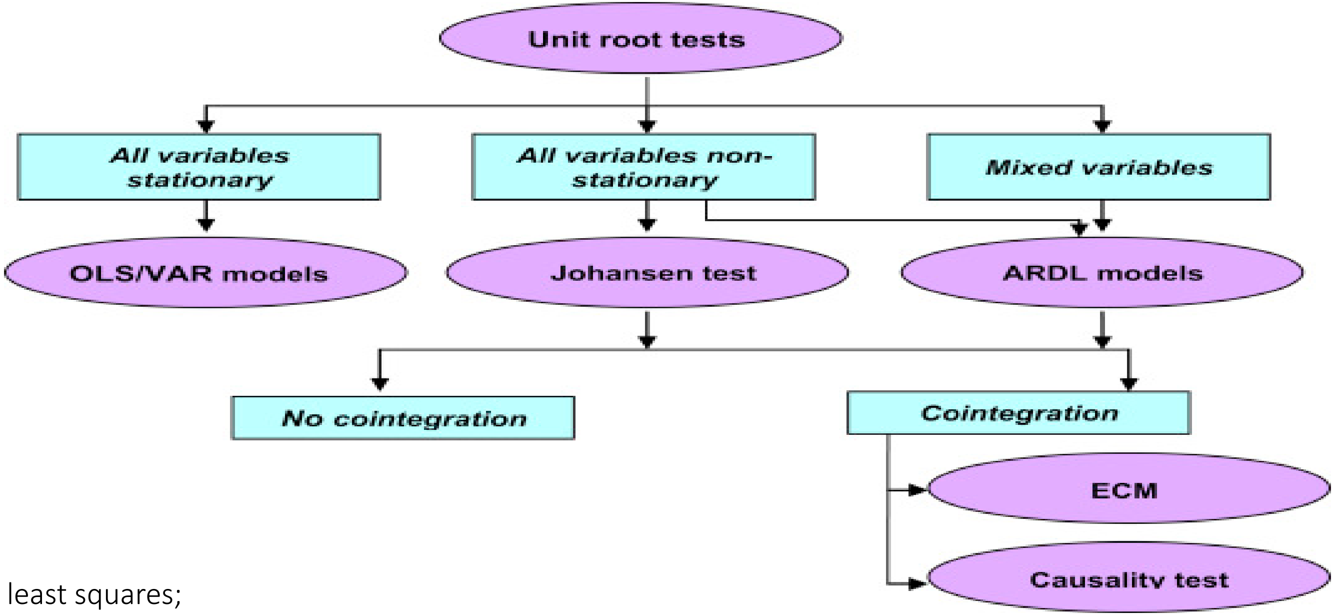
Time series data means that data is in a series of particular time periods or intervals.



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
.	.	.	.

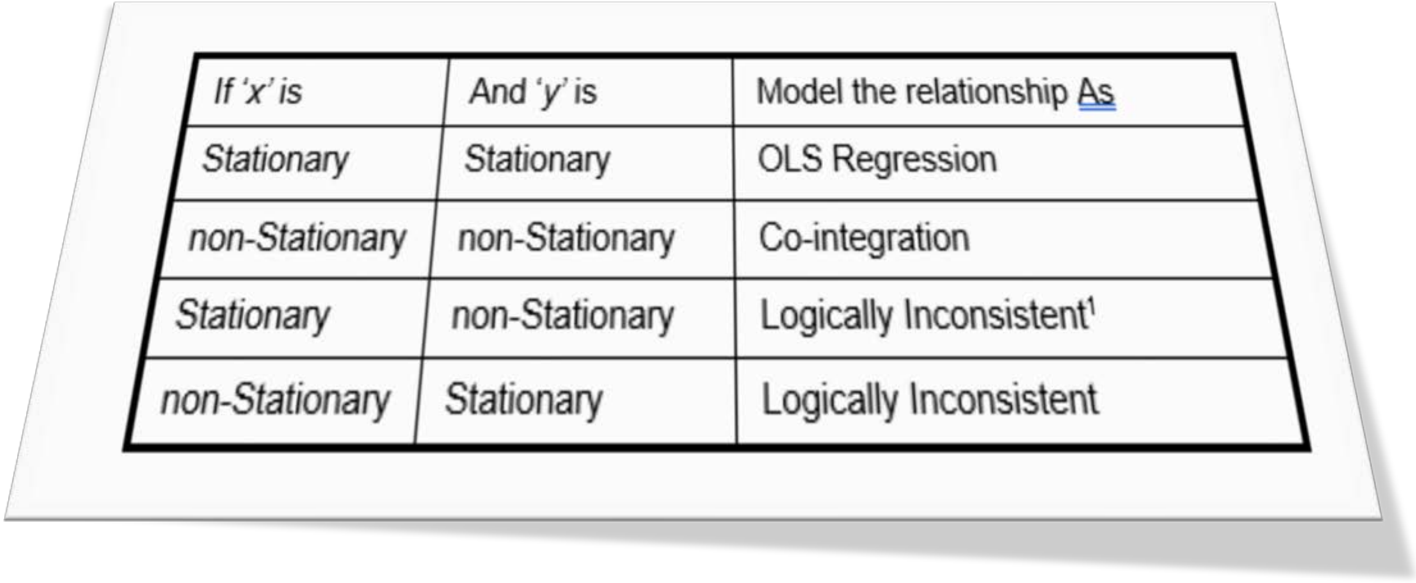
x and *y* are two time series variables

Time Series Analysis Techniques



OLS: Ordinary least squares;
VAR: Vector autoregressive;
ARDL: Autoregressive distributed lags;
ECM: Error correction models.
Source: Shresthaa and Bhatta (2018). Selecting appropriate methodological framework for time series data analysis. The Journal of Finance and Data Science Volume 4, Issue71:89

Cointegration



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

If two or more series are individually integrated (in the time series sense) but some linear combination of them has a lower order of integration, then the series are said to be cointegrated

Why Time Series data is different?

The stationarity or otherwise of a time series can strongly influence its behavior and properties

If the variables in the regression model are not stationary, then it can be proved that the standard assumptions for asymptotic analysis will not be valid. In other words, the usual "t-ratios" will not follow a t-distribution, so we cannot validly undertake hypothesis tests about the regression parameters.

TECHNIQUES

To present simple discussion and SAS programming coding techniques specifically designed to simulate the steps involved in time series data analysis specifically, modelling long-run relationship and examining time series variables long-run relationships (cointegration).

Techniques – Step 1

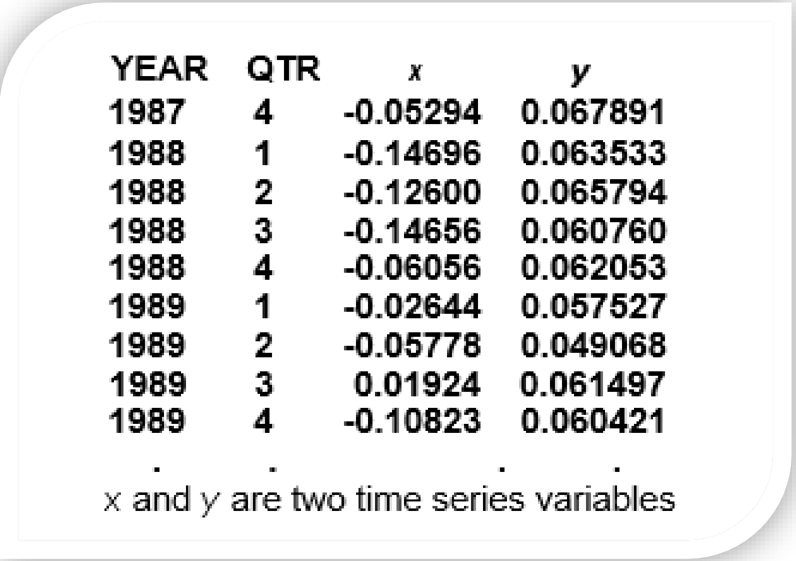
Estimate the long-run relationship

$$y_t = a + bx_t$$

and get the residuals series (e_t) of the regression

Techniques – Step 1

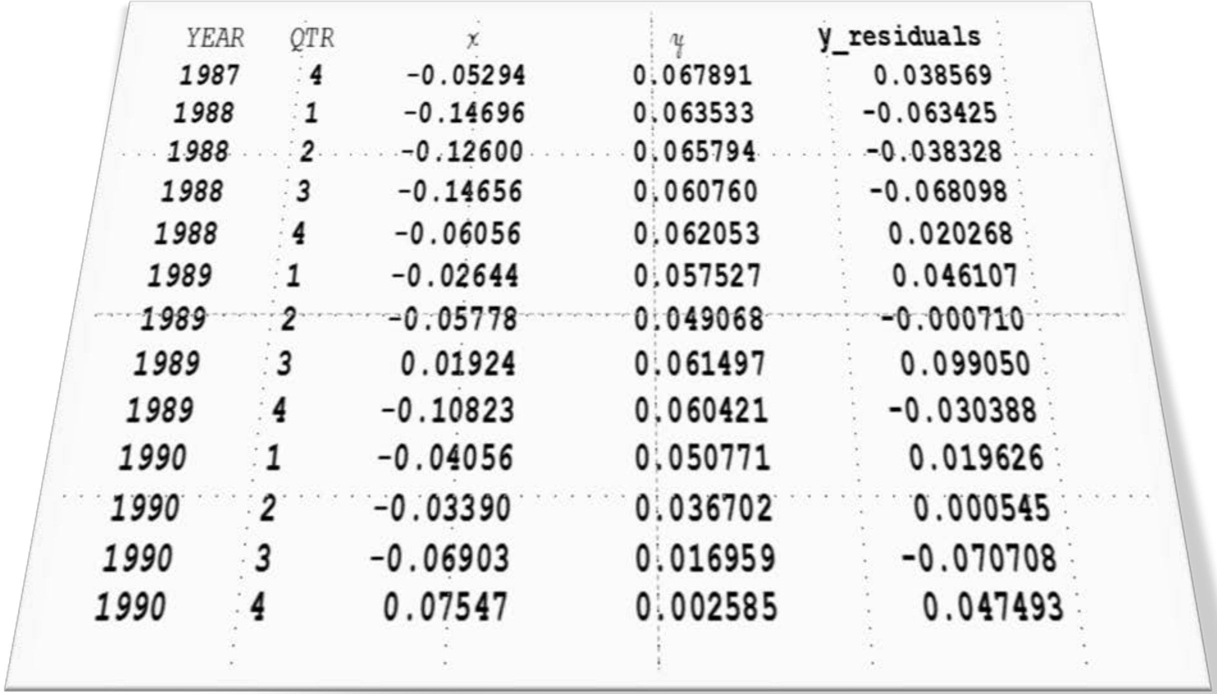
```
*SAS code;
PROC REG DATA= REG_SERIES; MODEL y = x;
OUTPUT OUT = RESIDS
R = y_residuals;
RUN; QUIT;
```



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

Step 1: SAS data output



```
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

Techniques – Step 1

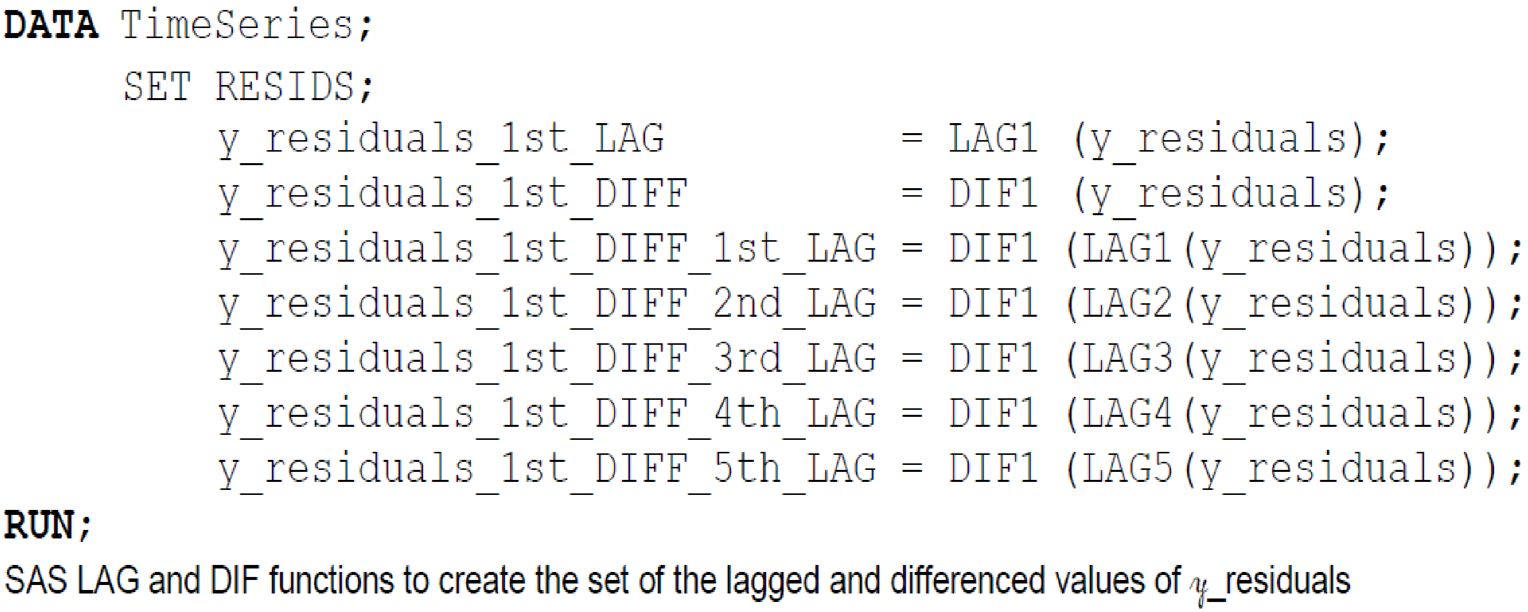
Apply stationarity test on the residuals series (e_t): If (e_t) series is non-stationary then we will reject cointegration.

Techniques – Step 2

Step 2: stationarity test on the residuals series (et) - residual ADF testing

$$\Delta \varepsilon_{i,t} = \kappa \varepsilon_{i,t-1} + \sum_{k=1}^{5} \varpi_{i,k} \Delta \varepsilon_{i' t-k} + \varepsilon_{k,t}$$

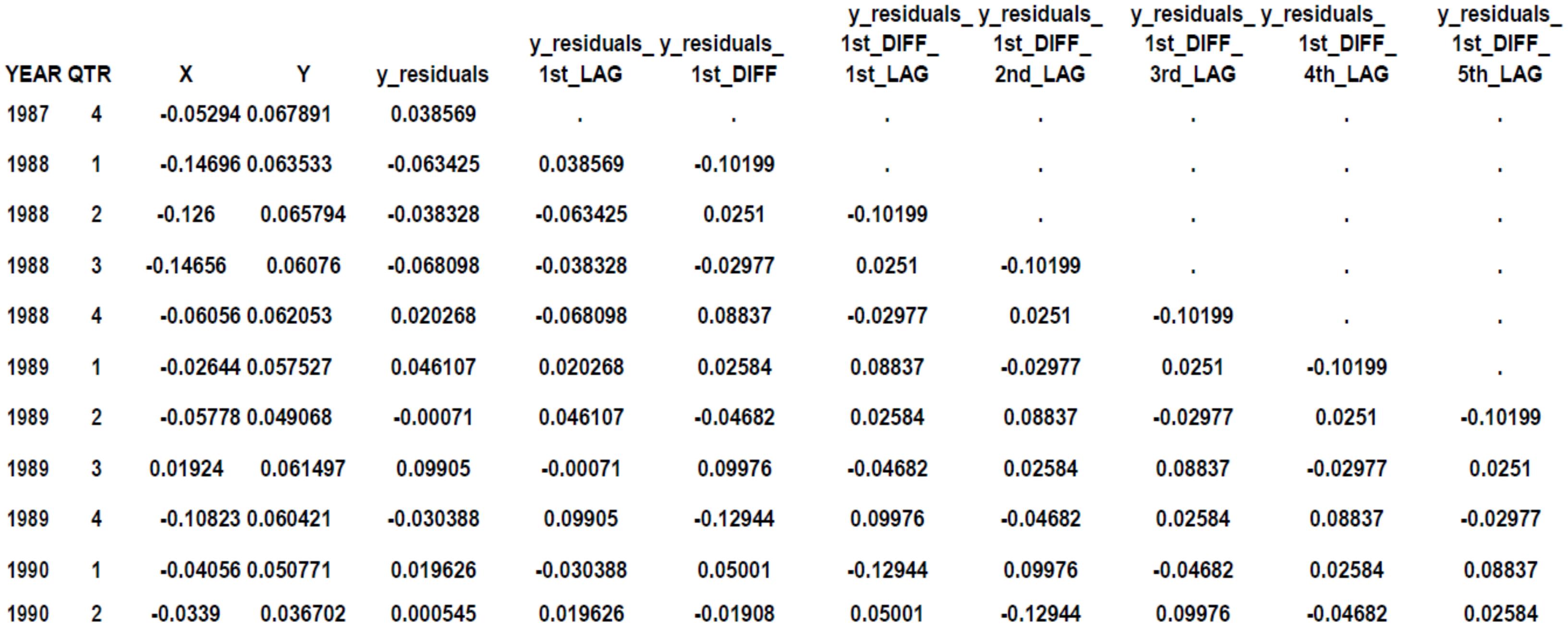
SAS Code



```
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

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



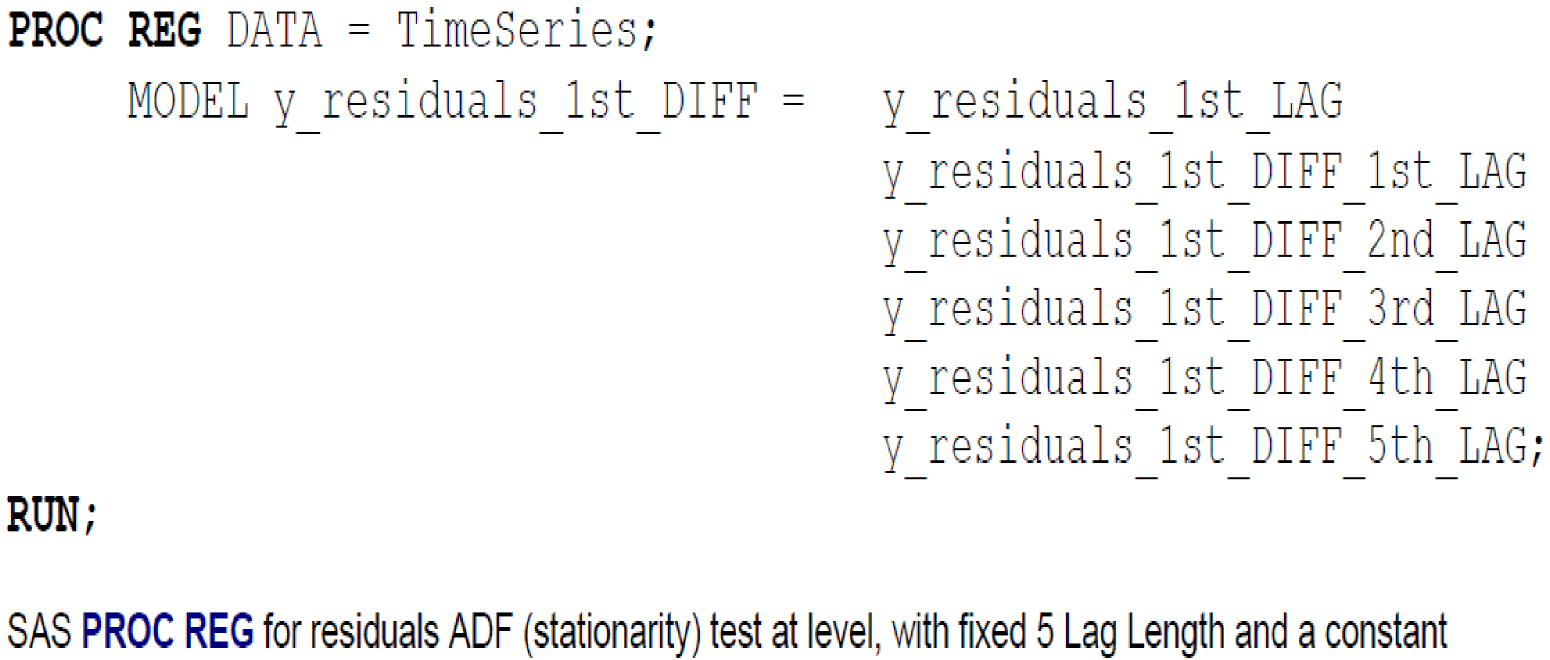
							y_residuals_	y_residuals_	y_residuals_	y_residuals_	y_residuals_
					y_residuals_	y_residuals_	1st_DIFF_	1st_DIFF_	1st_DIFF_	1st_DIFF_	1st_DIFF_
YEAR	QTR	X	Y	y_residuals	1st_LAG	1st_DIFF	1st_LAG	2nd_LAG	3rd_LAG	4th_LAG	5th_LAG
1987	4	-0.05294	0.067891	0.038569	.	.	.	.	.	.	.
1988	1	-0.14696	0.063533	-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

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 e_t (y_residuals) series is:

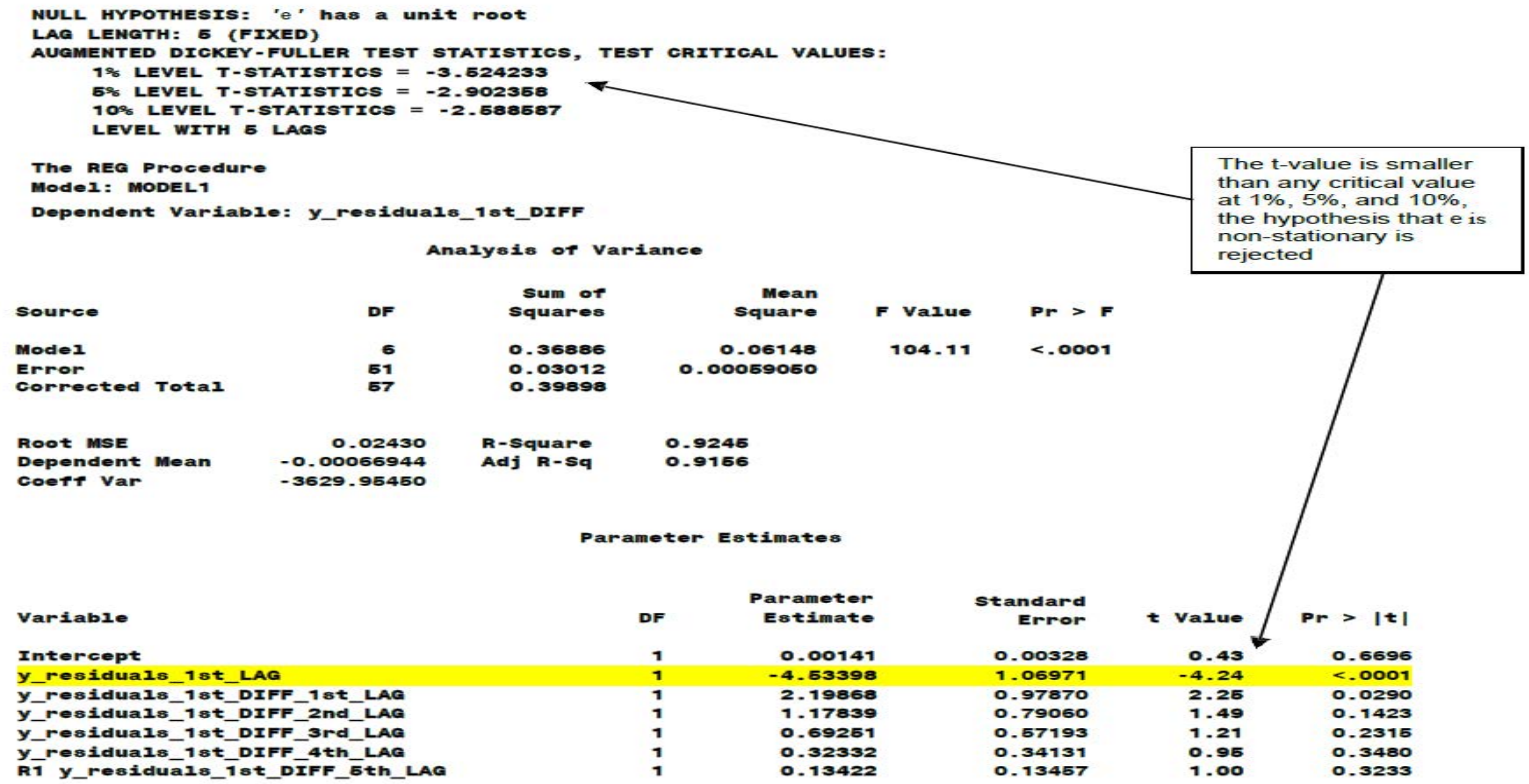
H_O: e_t is Non-stationary

H_A: e_t is Stationary



```
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;
```

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



```
NULL HYPOTHESIS:  'e' 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 e is non-stationary is rejected

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

Thank you!

Contact Information
Ismail.Mohamed@fhfa.gov

Reminder:

Complete your session survey in the conference mobile app.

SAS® GLOBAL FORUM 2020

USERS PROGRAM

MARCH 29 – APRIL 1 | WASHINGTON, DC | #SASGF