CHAPTER 3

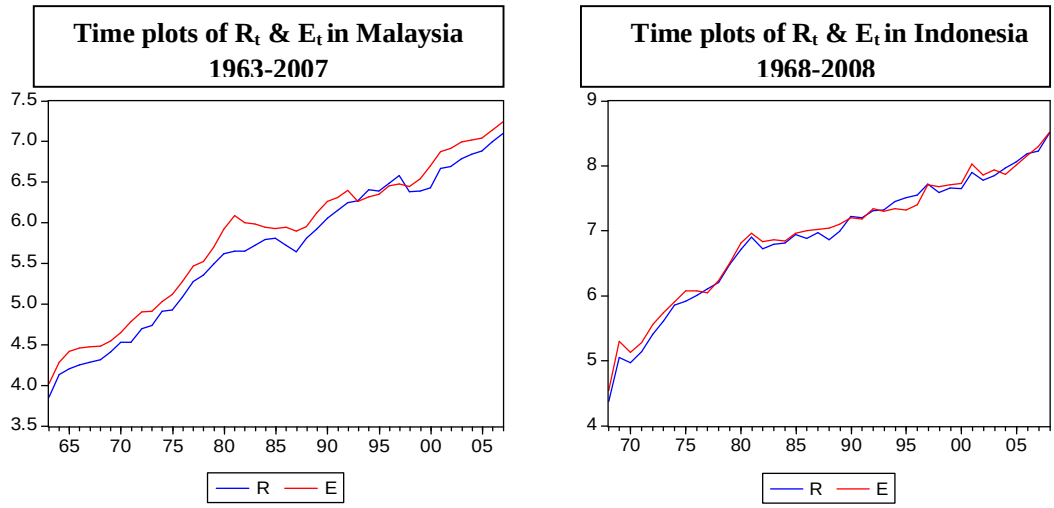
DATA AND RESEARCH METHODOLOGY

3.1 Data

The present study uses annual data on government revenue and government expenditure for Indonesia over the period 1968 to 2008, Malaysia during the period 1963 to 2007, Singapore from 1966 to 2007 and Thailand over the period 1953 to 2007. All data have been obtained from different issues of International Financial Statistics published by IMF. The wholesale price index (Year 2000) is chosen to be price deflator. Data on Wholesale Price Index (WPI) are obtained from different issues of International Financial Statistics. All data are transformed into real terms. The logarithm of the real government revenue (R_t) and government expenditure (E_t) have been used in the present study. The transformation of the series to Natural logarithms is necessary to avoid the problem of heteroscedasticity.

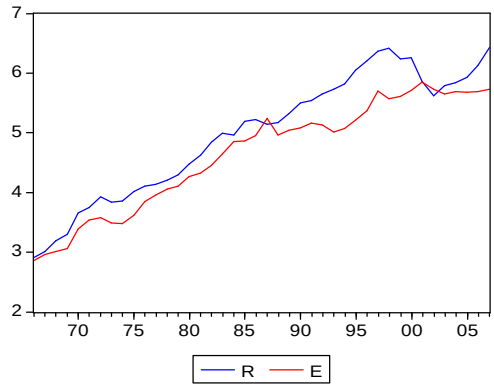
3.2 Time plots of Government Revenue and Government Expenditure:

Figure- 3.1

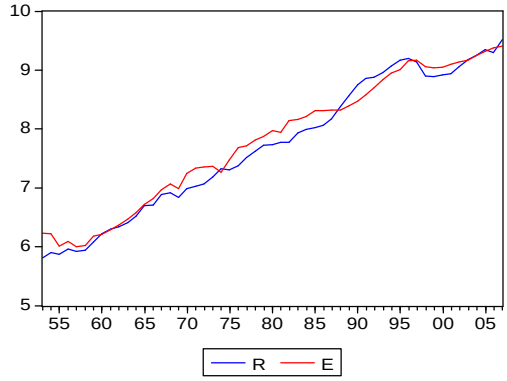


Time plots of R_t & E_t in Malaysia during 1963-2007 and for Indonesia over the period 1968 to 2008 indicate that there is a clear upward trend of plots of original data against time and indicting that that both the series are non-stationary. The non-stationary of the series has further been examined through ADF and PP unit root tests in the subsequent chapter.

Time plots of R_t & E_t in Singapore 1966-2007



Time plots of R_t & E_t in Thailand 1953-2007



Time plots of R_t & E_t in Singapore during 1966-2007 and for Thailand over the period 1953 to 2007 indicate that both series exhibit stochastic trends with some growth and suggesting that both the series are non-stationary. The non-stationarity of the series is further confirmed by ADF and PP unit root tests in the subsequent chapter.

3.3 Research Methodology

3.3.1 Methodological Issues: Stationarity Test

A common feature of many economic time series is that they have a positive trend, of the kind displayed by the stylized random walk with drift. A characterisation of these time series is that they are stationary-or integrated-series in which shocks have persistent effects. An alternative and contrasting view is that these series contain a deterministic trend, which accounts for the sustained increase in the series over time, and a random disturbance term which is stationary. Such a series is described as being trend stationary. A trend stationary series is contrasted here with an I(1) series which is stationary after being differenced once and is, therefore , described as being difference stationary.

The result of unit-root tests are very sensitive with the included assumptions about the time series, such as trend or intercept or both trend and intercept etc Therefore, to confirm about the nature of the underlying series, we can plot them graphically and confined the particular nature (at levels and after differencing).

Test of Stationarity: Augmented Dickey-Fuller Test

The problem of formally testing for unit roots first received a systematic treatment at the hands of Fuller (1976) and Dickey and Fuller (1979, 1981), Consider a variable y(t) exhibiting a linear trend $(\alpha + \beta t)$ Let x(t) denote the deviations from trend of y(t) i.e.

3.3.2 Test of Stationarity: The Correlogram Analysis

The random walk process for the time series implies non-stationarity in the sense that the series has been growing over time so that mean and other moments of the series are time dependent. The confirmation of the existence of non-stationarity in any series requires that the plots of *Autocorrelation Functions* (PACF) of different lags with corresponding Q-statistic are called *Correlogram.* Therefore, *Correlogram* test is a simple test of stationarity, which is based on the *Autocorrelation Function* (ACF) and Partial *Autocorrelation Functions* (PACF).

The *Correlogram* of the stationary and the non-stationary time series have some distinguishing features. If in any *Correlogram* the autocorrelation coefficients start from a very high level and decline very slowly towards zero as the lags lengthen, we can conclude that the time series in question is non-stationary and it follows a *random walk process.* So in case of non-stationary time series we see that solid spikes of the initial lags are very large and these decline very slowly. On the other hand, in case of stationary series the *Correlogram* has only few solid spikes at lower lags. Thus stationarity of the variables can be confirmed by examining the *Correlogram* of the time series concerned.

3.4 Cointegration: Meaning and Relevance

Random walks process attains stationarity after differencing. If a test fails to reject the hypothesis of a random walk, one can difference the series before using them in regression. Since many economic times series follow random walk, variables are subject to first differencing before using them in a regression.

However, differencing the date has a cost. The cost arises from the fact that differencing may result is a loss of information about the long-run relationship between variables concerned. This occurs because the models, estimated with differenced data, do not have a long-run solution. Moreover, the level of a variable and its first difference will typically be very

different in terms of mean, variance etc. The theory of *Correlogram* , therefore, may use as a diagnostics for linear regression.

Engle and Granger (1987) hold that non-stationary random walk time series data can still be used for the study of long-run equilibrium relationship among the variables concerned, provided that the variables are *Correlogram*. In many cases two or more variables follow random walk processes but the linear combinations of these variables are found to be stationary. If this be the case, then these variables are called *Correlogram* variables. *Correlogram* provides a method for eliminating the cost of differencing by retaining terms in levels but only in linear combination, which are stationary.

The justification of the *Correlogram* study is that in equilibrium relationship of a set of variables, the individual variables cannot move independently. Therefore, equilibrium relationship among a set of non-stationary variables implies that there stochastic trends must bear some relation to the current deviation from the equilibrium relationship The connection between the change in a variable and the deviation from equilibrium needs to be examined in detail.

Methods of Cointegration

Correlogram among the macro-economic time series is studied by applying different methods, namely,

(i) Engle-Granger Two Step Method , and
(ii) Johansen Method.

3.4.1 Engle-Granger Method of Cointegration

Engle-Granger Method of *Correlogram* study is suitable for bi-variate analysis. The method involves two steps of estimation as given below

Let there be two non-stationary variables, E_t and R_t, Then in the first step, one variable is regressed on the other such that:

$$E_t = \alpha_1 + \beta_1 R_t + \vartheta_t \quad \ldots\ldots\ldots\ldots (1)$$

$$R_t = \alpha_1 + \beta_2 E_t + \omega_t \quad \ldots\ldots\ldots\ldots (2)$$

From the equation (1) and (2) the residuals are obtained such that

$$\vartheta_t = E_t - \alpha_1 - \beta_1 R_t \ \ldots\ldots (3)$$

$$\omega_t = R_t - \alpha_2 - \beta_2 E_t \ \ldots\ldots.(4)$$

Here, ϑ_t and ω_t represent the linear combination of E_t and R_t.

(ii) In the second step stationarity of the residuals ϑ_t and ω_t is examined. If E_t and R_t are *cointegrated,* and linear combination of these variables would generate stationary residuals. For the purpose of testing stationarity Augmented Dickey Fuller (ADF) Test, Phillips-Perron Test etc. may be applied. Again, the stationarity of the residuals can be confirmed through the examination of their *Correlograms*.

(iii) If residuals ϑ_t and ω_t exhibit random walk, the time series are subject to filtering like differencing.

(iv)The cointegration equation has be re-estimated through the use of filtered (first differenced) data. For example, if and be the filtered series, then the estimable equations are

$$\Delta E_t = \alpha_1 + \beta_1 \Delta R_t + \varepsilon_t$$

$$\Delta R_t = \alpha_2 + \beta_2 \Delta E_t + \epsilon_t$$

(vi) Residuals of the re-equation are

$$\varepsilon_t = E_t - \alpha_1 - \beta_1 R_t$$

$$\epsilon_t = R_t - \alpha_2 - \beta_2 E_t$$

The residuals again are subject to ADF, PP test for ascertaining if the residuals are white noise. Again the stationarity of residuals can be confirmed through Correlogram.

These procedures have to be repeated until the residuals of the estimated equation are free from random walk. Thus the order of *Cointegration* will be determined.

3.4.2 Johansen Cointegration Test:

The Johansen (1988.1992) and Stock and Watson (1988) procedures are similar. Here we will describe only the Johansen test. To carry out the test, we first formulate the VAR.

$$y_t = \tau_1 y_{t-1} + \tau_2 y_{t-2} + \ldots\ldots + \tau_p y_{t-p} + \varepsilon_t \ldots\ldots\ldots.(5)$$

The order of the model, p, must be determined in advance. Now, let z_t denotes the vector of

M (p-1) variables,

$$z_t = [\Delta y_{t-1}, y_{t-2}, \ldots \Delta y_{t-p+1}]$$

That is, z_t contains the lags 1 to p-1 of the first differences of all M variables. Now, using the T available observations, we obtain two T*M matrices of least sqaures residuals:

D = the residuals in the regressions of Δy_t on z_t,

E = the residuals in the regressions of Δy_{t-p} on z_t,

We now require the M squared canonical correlations between the columns in D and those in E. to continue, we shall digress briefly to define the canonical correlations. Let d_1^* denote a linear combination of the columns of D, and let e_1^* denote the same from E. we wish to choose these two linear combinations so as to maximize the correlation between them. These pair of variables are the first canonical variables, and their correlation r_1^* is the first canonical correlation. Now with d_1^* and e_1^* in hand, we seek a second pair of variables d_2^* and e_2^* to maximize their correlation, subject to the constraint that this second variable in each pair be orthogonal to the first. This procedure continues for all M pairs of variables. The squared canonical correlations are simply the ordered characteristic roots of the matrix

$$R^* = R_{DD}^{-1/2} R_{DE} R_{EE}^{-1} R_{ED} R_{DD}^{-1/2}$$

Where R_{ij} is the correlation matrix between variables in set I and in set j , for i, j = D, E.

Finally, the null hypothesis that there are r or fewer cointgrating vectors is tested using the test statist

TRACE TEST = $-T \sum_{i=r+1}^{M} LOG\ [1 - (r_i^*)]$

If the correlations based on actual disturbances had been observed instead of estimated, then we would refer this statistic to the chi-squared distribution with M-r degrees of freedom.

3.5 The Vector Error Correction Mechanism

The time path of any variables is influenced by the extent of its deviations from the long-run equilibrium level. The Vector Error Correction (VEC) specification restricts the long-run behavior of the endogenous variables to converge to their *Cointegrating* relationships while

allowing for a wide range of short-run dynamics. The Cointegration term is known as the *error correction* term since the deviation from the long run equilibrium is corrected gradually through a series of partial short-run adjustments. Therefore, VEC modeling has given important information about the short-run relationship between the *Cointegrating* variables.

Vector Error Correction (VEC) Model

The Vector Correction (VEC) may be applied to analyze the short-run dynamics between two variables. Let the variables be Government revenue (R_t) and Government expenditure (E_t). Then the relevant VEC equations are:

$$\Delta E_t = \alpha_1 + \rho_1 Z_{1t-1} + \sum_{i=1}^{m} \beta_{1i} \Delta E_{t-i} + \sum_{i-1}^{m} \gamma_{1i} \Delta R_{t-i} + \theta_1 \ldots\ldots (6)$$

$$\Delta R_t = \alpha_2 + \rho_1 Z_{2t-1} + \sum_{i=1}^{m} \beta_{2i} \Delta R_{t-i} + \sum_{i-1}^{m} \gamma_{2i} \Delta E_{t-i} + \mu_1 \ldots\ldots (7)$$

Where, i = the number of lags included in the model. It can be determined empirically on the basis of Akaike Information Criterion and Schwartz Information Criterion so that the number of lags will be minimum. $\theta_1 \, and \, \mu_1$ are white noise error terms. Z_{it-1} and Z_{2t-1} are error correction terms.

The focus of the vector error correction analysis is son the lagged Z_{lt-1} and Z_{2t-1} terms. These lagged terms are the residuals from the previously estimated *Cointegrating* equations. In the present case the residuals from two lag specifications of the *Cointegrating* equations have been used in the vector error correction estimates. These Z_{lt-1} and Z_{2t-1} terms provide an explanation of short-run deviations from the long-run equilibrium for the test equations above. Lagging these terms means that disturbance of the last period impacts upon the current time period. In general, finding a statistically insignificant coefficient of the Z_{it-1}, Z_{2t-1} terms imply that the system under investigation is in the short-run equilibrium. If the coefficients of Z_{it-1} and Z_{2t-1} terms are found to be statistically significant, then the system is in the state of the short-run disequilibrium. In such a case the sign of Z_{it-1} and Z_{2t-1} terms give an indication of the causality direction between the two test variables.

3.6 Vector Autoregression Model (VAR)

Christopher Sims has developed the Vector Autoregresive Model (VAR) where true simultaneity among a set of variables necessitates that these are to be treated on equal footing. There should not be any a prior distinction between endogenous and exogenous variables. Economic theories contain behavioral, structural, andor reduced form relationship

that can be incorporated into a VAR analysis. The VAR model is designed for forecasting of the interrelated variables by analyzing the dynamic impact of the random disturbances of the system.

The model of Vector Autoregression (VAR) for the two variables E_t and R_t may be presented through the following equations:

$$\Delta E_t = \alpha_2 + \sum_{i=}^{n} \beta_{2i} \Delta R_{t-i} + \sum_{i=}^{n} \gamma_{2i} \Delta E_{t-i} + \vartheta_{2t} \ldots\ldots\ldots\ldots\ldots (8)$$

$$\Delta R_t = \alpha_1 + \sum_{i=}^{n} \beta_{1i} \Delta E_{t-i} + \sum_{i=}^{n} \gamma_{1i} \Delta R_{t-i} + \vartheta_{1t} \ldots\ldots\ldots\ldots\ldots (9)$$

Where ϑ_{1t} and ϑ_{2t} are the stochastic error terms, called impulse or innovations or shocks.

Thus equation for any endogenous variable (i) the auto regression term of that endogenous variable along with the distributed lag terms of other endogenous variable/s. In this VAR model

(1) the variables (like E_t and R_t) must be stationary, and

(2) ϑ_{1t} and ϑ_{2t} must be white-noise terms

3.7 Selection of Lag Length for the VAR Estimation

In the estimation of any VAR model, the selection of maximum lag length (k) is important since inclusions of too many lagged terms consume degrees of freedom and consequently, problem of multicollinearity may arise. On the other hand, inclusion of few lags may lead to selection errors.

Akaike Information Criterion (AIC), Schwartz Bayesian Criterion (SBC), Hannan-Quinn Criterion (HQ), Sequential Modified LR test statistic, etc are generally used for the determination of optimum lag lengthy. Another popular method has been developed by Enders (1995). Under this method, model has to be estimated with high lag (specific number depends on the nature of the underlying time series i.e., whether t is annual or quarterly time series etc).

Then lags are to be reduced by one and carried out the estimation, given that estimated t-static for the coefficient involved is insignificant. Finally, the lag corresponding to the estimated model with maximum number of significant coefficients has to be undertaken for analysis.

AIC and SBC corresponding to any lag length may be calculated as follows

AIC = T In (sum of squired residuals) + 2n

SBC = T In (sum of squired residuals) + n In (T)

Where n= number of parameters estimated

T = no of usable observations

A distributed lag econometric model is called good fit and, therefore, valid inferences can be drawn from such model if AIC and SBC values obtained from the model are negative and as small as possible. In other words, fit of the model improves, the AIC and SBC approach - ∞.

Of the two criteria, the SBC has superior large sample properties. It is asymptotically consistent. On the other hand, AIC is biased towards selecting an over parameterized model. However, in small samples, the AIC can work better than SBC.

3.8 Test of Structural Stability

3.8.1Chow Test

A series of data can often contain a structural break due to a change in policy or sudden shock to the economy. In order to test for a structural break, we often use the chow test. The model in effect uses F-test to determine whether a single regression equation is more efficient than two separate regressions involving splitting the data into two sub-samples.

In the first case we have just a single regression line to fit the data points. It can be expressed as:

$$E_t = \alpha_0 + \alpha_1 R_t + u_t \ldots\ldots\ldots\ldots (10)$$

In the second case, where is a structural break, we have two separate models, expressed as:

$$E_t = \beta_1 + \beta_2 R_t + u_{1t}$$

$$E_t = \delta_1 + \delta_2 R_t + u_{2t}$$

This suggests that model 1 applies before the break at time t, then model 2 applies after the structural break. If the parameters in the above models are the same i.e $\beta_1 = \delta_1, \beta_2 = \delta_2$, then models 1 and 2 can be expressed as a single model as in case 1,where there is a single

regression line. The chow test basically tests whether the single regression line or the two separate regression lines fit the data best. The stages in running the chow test are:

(1) Firstly run the regression using all the data , before and after the structural break, collect RSS_C

(2) Run two separate regressions on the data before and after the structural break, collecting the RSS in both cases, giving RSS_1 and RSS_2.

(3) Using these three values, calculate the test statistic from the following formula:

$$F = \frac{RSS_c - (RSS_1 + RSS_2)/k}{RSS_1 + RSS_2 / n - 2k}$$

(4) Find the critical values in the F-test table, in this case it has F (k, n-2k) degrees of freedom.

(5) Conclude, the null hypothesis is that there is no structural break.

3.8.2 Recursive Estimate

For a linear model $E_t = R_t \beta + u_t$ (t= 1,2,…..,T) with x_t (K* 1) and $\widehat{\beta}_{(t)}$ denoting the OLS estimator based on the first t observations only, that is, $\widehat{\beta} = (\sum_{t=1}^{r} R_t R_t^l)^{-1} \sum_{t=1}^{t} R_t E_t$ t≥ K

the recursive residuals are defined as

$$\hat{u}_t^{(r)} = \frac{E_t - R_t^{'} \widehat{\beta}_{t-1}}{[1 + R_t^{'} (\sum_{t=1}^{t-1} R R_t^{'})^{-1} R_t]^{1/2}} \quad t = K + 1\text{……..},T.$$

If x_t consists of fixed nonstochastic regressors, the forecast error $E_t - R_t^{'} \hat{\beta}_{(t-1)}$ is known to have mean zero and variance

$$\delta_u^2 [1 + R_t^{'} (\sum_{t=1}^{t-1} R_t R_t^{'})^{-1} R_t]$$

Hence the recursive residuals have constant variance δ_u^2. Therefore, even if some of the regressors are stochastic, the recursive residuals are often plotted with $\pm 2 \hat{\delta}_u$ bounds, where

$$\hat{\delta}_u^2 = (T - K)^{-1} \sum_{t=1}^{t-1} \hat{u}_t^2$$

is the usual residual variance estimator. Here $\hat{u}_t$'s are obtained from OLS estimation of the model based on all T observations. In other words,

$\hat{u}_t = E_t - R_t' \widehat{\beta_T}$. The recursive residuals exists only if the inverse of $\sum_{t=1}^{r} R_t R_t'$ exists for all t= K+1,……, T. Thus they may not be available in the presence of dummy variables.

3.8.3 CUSUM Test-

The so-called CUSUM, that is, the cumulative sum of recursive residuals defined as

$$\text{CUSUM}_t = \sum_{t=k+1}^{r} \hat{u}_t^{(r)} / \hat{\delta}_u,$$

can also reveal structural changes and is therefore often plotted for t= K+ 1…….T in checking a model. The CUSUM was proposed for this purpose by Brown, Durbin, & Evans (1975). If the CUSUM wanders off too far from the zero line, this is evidence against structural stability of the underlying model. A test with a significance level of about 5% is obtained by rejecting stability if CUSUM_t crosses the lines

$$\pm 0.948 \left[\sqrt{T-K} + 2(t-K)/\sqrt{T-K}\right]$$

This test is designed to detect a nonzero mean of the recursive residuals due to shifts in the model parameters. The test may not have much power if there is not only one parameter shift but various shifts that may compensate their impacts on the means of the recursive residuals. In that case the **CUSUM Squares** plot based on

$$\text{CUSUM} - SQt = \sum_{t=k+1}^{r} (\widehat{u_t}^{(r)})^2 / \sum_{t=k+1}^{T} (\widehat{u_t}^{(r)})^2$$

may be more informative. If these quantities cross the lines given by $\pm c + (t-K)/(T-K)$, a structural instability is diagnosed. The constant c depends on the desired significance level, the sample size T, and the number of regressors in the model.