CHAPTER 4

ESTIMATION OF BEHAVIOURAL EQUATIONS: COINTEGRATION ANALYSIS IN ECONOMETRIC MODELLING

Applied econometrics cannot be done mechanically: it needs understanding, intuition and skill. (Cuthbertson et al. 1992: v)

4.1 INTRODUCTION

This chapter provides a brief overview of the relevant econometrics literature for the estimation of dynamic models. Aspects like non-stationarity and the problem of spurious regressions, integrating and cointegrating properties of the data, as well as the estimation of single equations and multivariate systems using cointegration techniques will be covered. The Engle-Granger framework will be taken as point of departure, followed by the Engle-Yoo extension of the basic model, and finally the Johansen approach. Links between the short and the long run will be explored, the concept of an error correction model (ECM) will be discussed as well as the role of diagnostic testing and dynamic simulation in econometric modelling.

The Johansen technique has been employed in this study for the empirical estimation of the long-run equilibrium relationships in the behavioural equations that are reported in Chapter 5, namely a learning model for private consumption expenditure in South Africa. The single equation techniques are reported merely to emphasise the limitations of employing these techniques as opposed to a multivariate estimation technique like the Johansen approach.

4.2 THE PROBLEM OF NON-STATIONARITY AND SPURIOUS REGRESSIONS

Early reference to the problem of spurious correlations is made by Yule in a 1926 paper, noting that in analysing time-series data, it is possible to observe spurious correlations.

Darnell (1994:378) defines a spurious correlation as an observed sample correlation between series which, though appearing to be statistically significant, is a reflection of a common trend rather than a reflection of any underlying association. In the context of regressions using time-series data, it is possible to regress a variable y_t on another variable x_t, obtain a high R^2 statistic, large computed t-values and a very low Durbin-Watson statistic. This combination of statistics is a classic symptom of a spurious regression: one which has the superficial appearance of a good fit, especially when the model has been re-estimated using some adjustment for autocorrelation, such as Cochrane-Orcutt.

It is generally the presence of trends in the underlying data generation process that leads to the spurious identification of relationships between variables, simply because the trends in the data series generate statistical association, rather than meaningful causal relationships. The fundamental problem with regressing non-stationary series is that t- and F-tests no longer have the standard distributions associated with stationary series. With non-stationary series, there is a tendency to reject the null hypothesis of no association between individual variables, as well as for all regressors jointly. Moreover, this tendency increases with sample size. Ordinary least squares (OLS) estimation therefore presents problematic inferences when the data set contains non-stationary data. Detrending the data series would solve the problem only for trend stationary data and simply differencing the data to remove the non-stationary stochastic trend would not be appropriate, since the use of differenced variables, although avoiding the spurious regression problem, will also remove any long-run information (Banarjee *et al.* 1993, 82-84).

When considering long-run relationships, it becomes necessary to consider the underlying properties of the process that generates time series variables, i.e. a distinction must be made between stationary and non-stationary variables since, as suggested above, failure to do so can lead to spurious regression. The next section reports proposed techniques to establish integrating and cointegrating properties of the data.

4.3 INTEGRATING AND COINTEGRATING PROPERTIES OF THE DATA

Stationarity is a key concept in cointegration analysis. A definition for stationarity will be given in this section, followed by an exposition of a testing procedure to establish the order of integration of an individual time series. This will be followed by an introduction to techniques to apply when data series are non-stationary, involving the concept of cointegration.

4.3.1 Stationarity

A (weakly) stationary variable may be defined as a series with a constant mean and constant, finite variance[5]. Thus, a time series (x_t) is stationary if its mean, $E(x_t)$, is independent of t, and its variance, $E[x_t-E(x_t)]^2$, is bounded by some finite number and does not vary systematically with time. A non-stationary series on the other hand, will have a time-varying mean, or variance, so that any reference to the mean or variance should include reference to the particular time period under consideration.

Whether a variable is stationary depends on whether is has a unit root. Comparing stationary and non-stationary variables is also related to the different types of time trends that can be found in variables. Non-stationary variables contain stochastic, or random, trends, while stationary variables contain deterministic, or fixed, trends. It is data with random trends which often leads to spurious correlations. Next, a testing procedure for the presence of unit roots in data series will be presented. When data series are found to contain one or more unit roots, i.e. series are non-stationary, it is important to establish cointegration between variables, in order to infer a non-spurious causal long-run relationship between the series under consideration, as is discussed in section 4.3.3.

---

[5] Strict stationarity demands not only that the mean and variance of the series are independent of time, but also that all other higher moments are independent of time. In the cointegration framework, the requirement of weak stationarity generally suffices.

4.3.2 Unit roots and order of integration

As already noted above, non-stationarity implies the presence of a unit root in the time series under consideration. Testing for a unit root can be used to establish the order of integration.

If a series must be differenced d times before it becomes stationary, then it is said to be integrated of order d, denoted I(d). Thus a series x_t is I(d) if x_t is non-stationary but $\Delta^d x_t$ is stationary. That means that the series has d unit roots (solutions) associated with its ARIMA(p,d,q) representation, $(1-L)^d \phi(L)x_t = \theta(L)e_t$, for some p and q with $\phi(L)$ and $\theta(L)$ polynomials in the lag operator (L) and e_t a stationary process (Cuthbertson *et al.* 1992:130).

In order to test for the presence of unit roots, and hence for the degree of integration of individual series, a number of statistical tests may be used. The most popular of these are based on the class of tests developed by Dickey and Fuller (1979, 1981). Surveys by Diebold and Nerlove (1988); Pagan and Wickens (1989); Dolado *et al.* (1990); and Muscatelli and Hurn (1995)), amongst others, provide accounts of these tests.

The testing strategy followed in this study to determine the order of integration of individual time series is the one suggested by Dolado *et al.* (1990), employing the augmented Dickey-Fuller (ADF) test. A practical application of the Dolado testing strategy is provided by Sturm and De Haan (1995:69).

Dickey and Fuller (1981) test the null hypothesis of non-stationarity versus stationarity, suggesting ordinary least squares estimation of

$$\Delta Y_t = \eta_0 + \eta_1 \text{Trend} + \eta_2 Y_{t-1} + \sum_{i=1}^{m} \eta_{2+i} \Delta Y_{t-i} + \varepsilon_t \qquad (4.1)$$

where Y_t is the series being tested, m is the number of lags in the testing equation and ε_t is the residual. Lagged values of the dependent variable are included to take account of any serial correlation, and m is chosen so as to ensure that the residuals are white noise.

Dolado *et al.* suggest commencing the test with the specification of (4.1). The test is implemented through the usual t-statistic of $\hat{\eta}_2$, denoted as τ_τ. Under the null hypothesis, τ_τ will not follow the standard t-distribution; adjusted values as computed by MacKinnon (1990) have to be used for evaluation. If τ_τ is significant, the null of non-stationarity is rejected, and the series is stationary. This then concludes the test.

If τ_τ is insignificant however, the joint null hypothesis that $\eta_1=\eta_2=0$ using the F-statistic, denoted as Φ_3, is tested. The relevant critical values from Dickey and Fuller (1981) are used. If Φ_3 is significant, the test for a unit root must be conducted again, in this instance using the critical values of the standard t-distribution.

If the trend is not significant in the maintained model, the next step would be to estimate equation (4.1) without a trend ($\eta_1\equiv 0$). Once again the unit root test must be conducted, now denoting the t-statistic of $\hat{\eta}_2$ as τ_μ and using the relevant critical values from MacKinnon. If the null hypothesis is rejected, there is again no need to continue.

If the null is not rejected, the joint null hypothesis $\eta_0=\eta_2=0$ with use of the F-statistic, denoted as Φ_1, is tested, employing the critical values reported by Dickey and Fuller. Again, if it is significant, the unit root test must be conducted, using the standardised normal distribution.

If not, the constant must be removed from the testing equation as well ($\eta_0\equiv\eta_1\equiv 0$). The new statistic is called τ. MacKinnon also reports relevant critical values for this t-statistic. The last step is to examine whether the null hypothesis is rejected or not, i.e. whether the series is stationary or not.

The number of lags used in the estimated equations may be determined as suggested by Perron (1989). Perron suggested starting with eight lags. If the last lag is insignificant at a 10 per cent level (using the standard normal distribution), it is omitted. Next, seven lags are included. Again it is tested whether the last lag is significant (or there are no lags left, in which case the test is called the Dickey-Fuller (DF) test). This large significance level is taken because, as Perron (1989:1384) pointed out, 'including too many regressors of lagged

first-differences does not affect the size of the test but only decreases its power. Including too few lags may have a substantial effect on the size of the test'. Furthermore, Molinas (1986) noticed that 'a rather large number of lags has to be taken in the ADF test in order to capture the essential dynamics of the residuals'. Alternatively, the lag truncation parameter can be selected in order to minimise Akaike's information criterion (AIC) and to obtain stationarity of the residuals.

The result of applying the above test procedure to data series employed in this study is reported in Chapter 5, section 5.3.2. Tables 5.2, 5.3 and 5.4 contain values of test statistics for data in levels and first and second differenced form where relevant.

In cases where the ADF test proves to be inconclusive, other tools may be relied upon, for example graphical representations of the data in levels and first and second differenced form. It is also common to investigate whether a series is stationary by visual inspection of the graph of the sample autocorrelations against time, known as the *correlogram*, calculated by dividing the sample autocovariances by the sample variance. Alternative tests may also be considered. Phillips and Perron (1988) and Perron (1988) for example have suggested a non-parametric procedure in order to take account of the serial correlation in the model. Their procedure yields a number of 'modified' DF-type statistics, also known as Z-statistics. The advantage of these modified Z-statistics is that asymptotically, they eliminate the nuisance parameters that are present in the DF-statistics when the errors are not independently and identically distributed (IID). However, the main drawback in computing these Z-statistics is that the researcher has to decide *a priori* on the number of residual autocovariances which are to be used in implementing the corrections suggested by Phillips and Perron (Muscatelli and Hurn 1995:175).

4.3.3 The concept of cointegration

According to Harris (1995:6), the economic interpretation of cointegration states that if two (or more) series are linked to form an equilibrium relationship spanning the long run, then even though the series themselves may contain stochastic trends and thus be non-stationary, they will nevertheless move closely together over time and the difference between them will be stable (i.e. stationary).

The formal definition of cointegration of two variables, developed by Engle and Granger (1987:253) is as follows: time series x_t and y_t are said to be *cointegrated of order d,b* where $d \geq b \geq 0$, written as

$$x_t, y_t \sim CI(d,b),$$

if

(i) Both series are integrated of order d,

(ii) There exists a linear combination of these variables, say $\alpha_1 x_t + \alpha_2 y_t$, which is integrated of order d–b.

The vector $[\alpha_1, \alpha_2]$ is called the *cointegrating vector*.

A straightforward generalisation of the above definition for the case of n variables is as follows: if x_t denotes an $(n \times 1)$ vector of series $x_{1t}, x_{2t}, \ldots, x_{nt}$ and

(i) each x_{it} is I(d),

(ii) there exists an $(n \times 1)$ vector α such that $x'_t \alpha \sim I(d-b)$, then

$x'_t \alpha \sim CI(d, b)$.

Condition (i) in the above definition can be relaxed, as follows: if a linear combination of any two time series y_t and x_t is formed and each is integrated of a different order, then the resulting series will be integrated at the higher of the two orders of integration. Thus if $y_t \sim I(1)$ and $x_t \sim I(0)$ (or $y_t \sim I(0)$ and $x_t \sim I(1)$), these two series cannot possibly be cointegrated as the I(0) series has a constant mean while the I(1) series tends to drift over time. Consequently the error $(y_t - \alpha x_t \sim I(1))$ between them would not be stable over time. It is however possible to obtain cointegration between 3 or more series even if all series are not integrated of the same order. Pagan and Wickens (1989) pointed out that, in this instance, a subset of the higher-order series must cointegrate to the order of the lower-order series. If $y_t \sim I(1)$, $x_t \sim I(2)$ and $z_t \sim I(2)$, and a cointegration relationship between x_t and z_t is found, such that $v_t\ (= x_t - \lambda z_t) \sim I(1)$, then this result can potentially cointegrate with y_t, to obtain $w_t\ (= y_t - \gamma v_t) \sim I(0)$.

Furthermore, if there are $n>2$ variables in the model, there may be more than one cointegrating vector. It is possible for up to $n-1$ linearly independent cointegration vectors to exist, which has implications for testing and estimating cointegration relationships.

4.4 MODELLING COINTEGRATED SERIES THROUGH ERROR CORRECTION MODELS

This section will describe how cointegrated non-stationary variables can be used to formulate and estimate a model with an error correction mechanism. The fact that variables are cointegrated implies that there is some adjustment process which prevents the errors in the long-run relationship from becoming increasingly larger. Engle and Granger (1987:255-258) have shown that for any set of variables that are cointegrated of order 1,1, that is CI(1,1), there exists a valid error correction representation of the data. The *Granger Representation Theorem* formalises this theoretical connection between cointegration and error correction.

Different approaches towards establishing cointegration between variables and estimating the long-run relationship, and the subsequent specification of an error correction model representing the short-run adjustment towards equilibrium, will be discussed below. The discussion will commence with single equation cointegration techniques, namely the Engle-Granger (1987) approach and the Engle-Yoo (1989) extension of this procedure, followed by a multivariate cointegration technique known as the Johansen (1988, 1989) approach.

4.4.1 Engle-Granger estimation

The first approach was originally proposed by Engle and Granger (1987). Consider the long-run relation for the bivariate case (the extension to the multivariate case is direct) with the form:

$$y_t = \beta x_t + u_t, \tag{4.2}$$

where both y_t and x_t are I(1), and with β an unknown coefficient. The Engle-Granger two-step procedure would entail the following steps: first, estimate equation (4.2) by ordinary

least squares and test for stationarity of the residuals. This entails testing whether $u_t \sim I(1)$ against the alternative that $u_t \sim I(0)$. Second, if the null hypothesis of no cointegration can be rejected, equation (4.3) below can be estimated, replacing β by its previously computed OLS estimate, $\hat{\beta}$:

$$\Delta y_t = \phi(L)\Delta y_{t-1} + \Theta(L)\Delta x_t + \alpha(y_{t-1} - \hat{\beta}x_{t-1}) + \varepsilon_t, \qquad (4.3)$$

where ε_t is an error term and α is negative. Δy_t, Δx_t and $(y_{t-1} - \hat{\beta}x_{t-1})$ are all I(0) and consequently, provided that the model is properly specified, ε_t is also I(0). Equation (4.3) represents the *error correction model*, containing the long-run cointegration relationship in the form of the lagged residual obtained from the estimated long-run cointegration equation (called the *error correction mechanism*) as well as the short-run dynamic structure allowing for adjustment towards equilibrium. Engle and Granger (*op. cit.*:254) motivate the inclusion of the error correction mechanism as follows: "The idea is simply that a proportion of the disequilibrium from one period is corrected in the next period". The slope coefficient α in equation (4.3) indicates the speed of adjustment towards equilibrium.

Testing for cointegration in the first step of the procedure entails the following. As with univariate unit root tests, the unit root in the residual (implying no cointegration between the variables) is based on a t-test with a non-normal distribution. However, unless the value of β is already known (not estimated by equation (4.2)), it is also not possible to use the standard Dickey-Fuller tables of critical values. There are two major reasons for this. First, because of the way it is constructed, the least square estimator chooses the parameter vector which minimises residual variance, even if the variables are not cointegrated, causing the error, u_t, to appear as stationary as possible. Thus the standard DF distribution would tend to over-reject the null. Second, the distribution of the test statistic under the null is affected by the number of regressors included in equation (4.2). MacKinnon (1991:273-75) uses response surface analysis to obtain approximate finite sample critical values for the conventional ADF-test on residuals of the long-run relationship, and must be used in this instance (Harris 1995:54).

Engle and Granger (1987) also suggest the use of a second test for cointegration, namely the cointegrating regression Durbin-Watson (CRDW) test, proposed by Sargan and

Bhargava (1983). This test is computed from the cointegrating regression. This statistic should be compared with the critical values from either Table II or III (*op. cit.*:269-70); under the null of non-cointegration, CRDW should be close to zero and so the null is rejected if the statistic exceeds the critical value. According to Engle and Granger, this statistic should however be used only for a quick approximate result.

The Engle-Granger two-step approach has the advantage of relative simplicity. Where there is a unique cointegrating vector, it allows for the use of the superconsistency property of OLS to obtain consistent estimates of the cointegrating vector. However, there are a number of critical limitations to the technique, which severely limits its usefulness and applicability. One important limitation pertains to the fact that inferences cannot be drawn using standard t-statistics about the significance of the parameters of the static long-run model, since the distribution of the estimators of the cointegrating vector is generally non-normal. Furthermore, while the static regression gives consistent estimates of the cointegrating vector, these estimates are not fully efficient.

Engle and Yoo (1989) proposed a third step to the Engle and Granger two-step estimation to overcome the above-mentioned problems.

4.4.2 Engle-Yoo estimation

The Engle-Yoo estimation procedure only provides an extension (third step) to the two-step Engle-Granger approach. The third step provides a correction of the parameter estimates of the first stage static regression, which provides a set of standard errors allowing the valid calculation of standard t-tests.

The third stage simply consists of a further regression of the conditioning variables from the static regression multiplied by minus the error correction parameter, regressed on the errors from the second-stage error correction model. The coefficients from this model are the corrections to the parameter estimates while their standard errors are the relevant standard errors for the inference (Cuthbertson *et al.*:141).

The three steps are then: first, estimate the standard cointegrating regression of the form:

$$y_t = \beta x_t + u_t, \tag{4.4}$$

where u_t is the OLS residual to give first-stage estimates for β, namely β^1. The second-stage dynamic model has to be estimated next, using the residuals from the cointegrating regression to impose the long-run constraint:

$$\Delta y_t = \phi(L)\Delta y_{t-1} + \Theta(L)\Delta x_t + \alpha u_{t-1} + \varepsilon_t. \tag{4.5}$$

The third stage consists of the regression

$$\varepsilon_t = \delta(-\alpha)x_t + v_t. \tag{4.6}$$

The correction of the first stage estimates then is

$$\beta^3 = \beta^1 + \delta \tag{4.7}$$

and the correct standard errors for β^3 are given by the standard errors for δ in the third-stage regression.

4.4.3 Problems associated with the single equation approach

A number of problems related to a single equation approach towards modelling the static long-run equilibrium relationship and the dynamic short-run properties of the underlying data generating process have been pointed out in the preceding sections. More limitations will subsequently be highlighted, followed by a strategy known as the Johansen maximum likelihood estimation procedure to address these deficiencies.

One important limitation that has been addressed pertains to the fact that standard t-statistics cannot be used to draw inferences about the significance of the parameters of the static long-run model, due to the non-normal distribution of the estimators of the cointegrating vector.

Perhaps the most severe limitation of the single equation model follows from the possibility of more than one cointegrating vector present in the data when the cointegration regression

contains more than two variables. Including n variables in the equation may yield (n–1) equilibrium relationships governing the joint evolution of the variables, and hence there may exist (n–1) cointegrating relationships in the data. Single equation techniques assume that only one cointegrating vector exists in the data. Should this not be the case, the consequence will be inefficiency in the estimation, since the cointegrating vector will be a linear combination of all cointegrating relationships present in the data.

Furthermore, the Engle-Granger approach effectively ignores the short-run dynamics when estimating the long-run equilibrium relationship, or the cointegrating vector. Estimating the model by means of only including the long-run equilibrium relationship (the data in levels) effectively shifts the short-run dynamics to the error term, subjecting the residuals to serial correlation.

Yet another problem with residual based tests worth noting, pertains to the endo-exogenous division of variables. Charemza and Deadman (1997:151) point out that the distinction between the types of variables appearing in a multiple equation system is in stark contrast with single equation structural modelling. Usually, what is on the left-hand side of a single equation structural model is simply treated as endogenous and what is on the right-hand side as exogenous. Charemza and Deadman (*op. cit.*:150) refer to a model of aggregate income, Y_t, and aggregate consumption, C_t, to demonstrate that a rise in income will lead to a rise in consumption. However due to the income identity, it is impossible to change the value of C_t, without influencing Y_t. Both variables would thus be regarded as endogenous variables and be described as jointly dependent variables. This type of observed simultaneity between variables is disregarded in a single equation approach.

Multivariate estimation techniques are able to address the above problems by detecting all possible long-run cointegration relationships present in the data, accommodating short-run properties when estimating the long-run relationships and by taking simultaneity between variables into account. Harris (1995:61) demonstrates that information is lost unless endogenous variables appear on the left-hand side of the estimated equations in a multivariate model, except for the case where all the variables in the cointegration vector are weakly exogenous. The Johansen approach, a multivariate estimation technique, namely, is addressed in the next section.

4.4.4 The Johansen approach

What has come to be termed the Johansen approach emerged from the work of Johansen (1988, 1989) and Johansen and Juselius (1990). One means of understanding the impulse of the approach comes from a contribution of Sims (1980), who noted that in macroeconomic systems many variables are likely to be interdependent, rendering exogeneity of variables rare. Sims's suggestion for dealing with this problem was to use vector autoregressive (VAR) estimation.

An exposition of the Johansen approach will be presented next. The method begins by expressing the data generation process of a vector z_t of n potentially endogenous variables, as an unrestricted vector autoregression (VAR) in the levels of the variables, involving up to k lags of z_t:

$$z_t = A_1 z_{t-1} + \ldots + A_k z_{t-k} + u_t \tag{4.8}$$

where z_t is $(n\times 1)$ and each of the A_i is an $(n\times n)$ matrix of parameters. The system of equations (4.8) can be reparameterised in ECM form:

$$\Delta z_t = \Gamma_1 \Delta z_{t-1} + \ldots + \Gamma_{k-1} \Delta z_{t-k+1} + \Pi z_{t-k} + u_t \tag{4.9}$$

where $\Gamma_i = -(I - A_1 - \ldots - A_i)$, $i=1,\ldots,k-1$ and $\Pi = -(I - A_1 - \ldots - A_k)$. This way of specifying the system contains information on both the short and long-run adjustment to changes in z_t, via the estimation of $\hat{\Gamma}_i$ and $\hat{\Pi}$. Thus, Π defines the long-run levels solution to equation (4.8). It can be shown that $\Pi = \alpha\beta'$, where α represents the speed of adjustment to equilibrium, while β is a matrix of long-run coefficients such that the term $\beta' z_{t-k}$ embedded in (4.9) represents up to (n–1) cointegration relationships in the multivariate model which ensures that the z_t converge to their steady-state solutions. Assuming that z_t is a vector of non-stationary I(1) variables, then all the terms in (4.9) which involve Δz_{t-i} are I(0), while Πz_{t-k} must also be stationary for $u_t \sim I(0)$ to be white noise. Stationarity of the term Πz_{t-k},

i.e. a set of I(1) variables which form linear combinations that are I(0), implies that it contains the long-run *cointegrating relationships between the variables in levels*.

The Johansen technique therefore reveals the exact number of cointegrating relationships present between variables in the model, as well as the nature of these relationships. The way in which the information is extracted involves a method known as *reduced rank regression*. Rewriting equation (4.9) yields:

$$\Delta z_t + \alpha\beta' z_{t-k} = \Gamma_1 \Delta z_{t-1} + \ldots + \Gamma_{k-1} \Delta z_{t-k+1} + u_t . \tag{4.10}$$

It is possible to correct for short-run dynamics (i.e. eliminate their effect) by regressing Δz_t and z_{t-k} separately on the right-hand side of (4.10). That is, the vectors R_{0t} and R_{kt} are obtained from:

$$\Delta z_t = P_1 \Delta z_{t-1} + \ldots + P_{k-1} \Delta z_{t-k+1} + R_{0t} \tag{4.11}$$

$$z_{t-k} = T_1 \Delta z_{t-1} + \ldots + T_{k-1} \Delta z_{t-k+1} + R_{kt} . \tag{4.12}$$

The above can then be used to form residual (product moment) matrices:

$$S_{ij} = T^{-1} \sum_{t=1}^{T} R_{it} R'_{jt} \qquad i,j=0,k. \tag{4.13}$$

The maximum likelihood estimates of β is obtained as the eigenvectors corresponding to the r highest eigenvalues from solving the equation

$$| \lambda S_{kk} - S_{k0} S_{00}^{-1} S_{0k} | = 0 \tag{4.14}$$

which gives the n eigenvalues $\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots > \hat{\lambda}_n$ and the corresponding eigenvectors $\hat{V} = (\hat{v}_1, \ldots, \hat{v}_n)$. The r elements in $\hat{V}$ which determine the linear combinations of stationary relationships can be denoted $\hat{\beta} = (\hat{v}_1, \ldots, \hat{v}_r)$, that is, these are the cointegration vectors. Johansen (1992b) points out that a test for the rank of Π equal to 1 (r=1), for example, is essentially a test that $\hat{\lambda}_2 = \hat{\lambda}_3 = \ldots = \hat{\lambda}_n = 0$, whereas $\hat{\lambda}_1 > 0$.

Thus, testing for cointegration amounts to a consideration of the rank of Π, that is, finding the number of r linearly independent columns in Π. If Π has full rank (there are r=n

linearly independent columns) then the variables z_t are I(0), while if the rank of Π is zero, there are no cointegration relationships. The interesting case is when Π has reduced rank, i.e. when there are $r \leq (n-1)$ cointegration vectors present.

To test the null hypothesis that there are at most r cointegration vectors amounts to:

$$H_0: \lambda_i = 0 \qquad i = r+1,\ldots,n \tag{4.15}$$

versus H_A: $\lambda_i \neq 0$, where only the first r eigenvalues are non-zero. The restriction can be imposed for different values of r. The log of the maximised likelihood function for the restricted model is then compared to the log of the maximised likelihood function of the unrestricted model and a standard likelihood ratio test is computed (although with a non-standard distribution). That is, it is possible to test the null hypothesis using what has become known as the *trace* statistic:

$$\lambda_{trace} = -2\log(Q) = -T\sum_{i=r+1}^{n}\log(1-\hat{\lambda}_i) \qquad r = 0,1,2,\ldots,n\text{-}2,n\text{-}1. \tag{4.16}$$

where Q=(restricted maximum likelihood ÷ unrestricted maximum likelihood). Asymptotic critical values are provided in Osterwald-Lenum (1992). Harris (1995:88) points out that when only a small sample of observations on z_t is available, there are likely to be problems with the power and size properties of the above test when using asymptotic critical values.

Another test of the significance of the largest λ_r, is the so-called maximal-eigenvalue or λ_{max} statistic:

$$\lambda_{max} = -T\log(1-\hat{\lambda}_{r+1}) \qquad r = 0,1,2,\ldots,n\text{-}2,n\text{-}1.$$

This tests that there are r cointegration vectors present against the alternative that r+1 such vectors exist.

This section is concluded by presenting a number of distinct steps in the practical estimation process as set out by Harris (1995:76):

(i) Test the order of integration of each variable entering the multivariate model.

(ii) Select the correct lag length, k, of the vector autoregressive (VAR) model, to ensure that the (vector) error correction model has Gaussian errors. Model selection criteria as the Akaike information criterion (AIC) or the Schwarz Bayesian criterion may be used, as well as F tests for the hypothesis that the i-period lag is zero.

(iii) Determine whether the system needs to be conditioned on any I(0) variables, including dummy variables.

(iv) Test for the reduced rank of the system. The Johansen cointegration test may be used to test whether $\Pi(=\alpha\beta')$ has reduced rank; and to determine the value of r, with $r \leq (n-1)$ the number of cointegration vectors present in β.

(v) Determine whether the system is to be estimated with deterministic variables (constant and trend) or not. There are three possible options: (1) the VAR may be specified without any constant term; (2) the VAR may have a restricted constant term which appears only as a part of the cointegrating vectors so that the ECM from (4.9) contains any constants within the term Πz_{t-k} only; (3) the VAR may have an unrestricted constant (Cuthbertson *et al.* 1995:148).

(vi) Test for weak exogeneity. Weakly exogenous variables may be removed from the left-hand side of the equation, while remaining in the long-run model.

(vii) Test the linear hypothesis on cointegrating relationships as well as for unique cointegrating vectors. This entails imposing restrictions motivated by economic arguments (e.g. that some of the β_{ij}s are zero, or that homogeneity restrictions are needed such as $\beta_{1j}=-\beta_{2j}$) and then testing whether the columns of β are identified.

(viii) Impose joint tests of restrictions on the α loading matrix, (i.e. the speed-of-adjustment parameters) and the β cointegrating vector.

4.5 THE ROLE OF DIAGNOSTIC TESTING AND DYNAMIC SIMULATION IN ECONOMETRIC MODELLING

According to Hendry (1980:403), the three golden rules of econometrics are test, test, test. Diagnostic checking is therefore a very important part of the whole process of model selection – "Rigorously tested models, which adequately describe the available data, encompass previous findings and were derived from well based theories would greatly

enhance any claim to be scientific." (*op. cit.*:403). Thus, whether the error correction model has been derived using single equation estimation techniques, or a more sophisticated multivariate estimation technique, in order to assess the validity of the model, it must be subjected to a battery of diagnostic tests.

Diagnostic tests, or mis-specification tests, are designed to test the adequacy of the specification of a regression equation. Darnell (1994:93) summarises the possible ways in which an equation might be mis-specified: (i) the set of regressors may be incomplete – some variables may have been omitted; (ii) the parameter vector may not be constant; (iii) the functional form may be incorrect; (iv) one or more of the regressors may not be exogenous; (v) the error term may be autocorrelated; (vi) the error term may be heteroscedastic; and (vii) the error term may be non-normally distributed.

Tests developed to test for mis-specification include the following: tests of omitted variables may be carried out as a linear hypothesis test – additional variables are included in the model and their exclusion is tested as an F-test; constancy of the parameter vector may be examined using a Chow test or a test of predictive failure or may be examined using recursive residuals; the functional form may be examined using Ramsey's RESET test or a Box-Cox test; exogeneity may be examined by Hausman's test; autocorrelation and heteroscedasticity may be examined by a number of tests, including the Lagrange multiplier test and the Box-Pierce and Lung-Box tests for serial correlation, the Breuch-Pagan test for heteroscedasticity and Engle's test for autoregressive conditional heteroscedasticity; and the normality assumption may be tested by the Bera-Jarque statistic and by testing for outliers. The detail of these tests will not be discussed here, but can been found in the literature (amongst others Godfrey 1988; Darnell 1993; Cuthbertson *et al.* 1995; Greene 1997).

Cuthbertson *et al.* (1995:106) point out that often an ECM is constructed so that it passes a set of diagnostic tests. Tests for parameter constancy and encompassing tests then become of increasing importance in testing competing models.

Dynamic, in-sample simulation and deterministic analysis of the response characteristics of the model, testing whether short and long-run response characteristics correspond to

theoretical priors and long-run equilibrium properties of the data often prove helpful in assessing the validity of the model. The process would consist of conducting a dynamic baseline forecast for each stochastic equation. An exogenous shock is applied to the system and the adjustment path towards a new equilibrium is then determined. Dynamic out-of-sample simulation can be useful in establishing the forecasting performance of the model.

4.6 CONCLUSION

This chapter highlighted the problem of spurious regression when modelling non-stationary data series and reported econometrics literature addressing this issue. Single equation estimation techniques were discussed to point out the limitations thereof as opposed to a multivariate estimation technique like the Johansen approach. The Johansen technique however poses it own demands, and the discussion included the presentation of a number of distinct steps to consider when applying this technique. Finally, the importance of diagnostic testing was re-emphasised.