Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots

John Y. Campbell and Pierre Perron

PRINCETON UNIVERSITY AND NBER/PRINCETON UNIVERSITY AND CRDE

*Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots**

1. Introduction

The field of macroeconomics has its share of econometric pitfalls for the unwary applied researcher. During the last decade, macroeconomists have become aware of a new set of econometric difficulties that arise when one or more variables of interest may have unit roots in their time series representations. Standard asymptotic distribution theory often does not apply to regressions involving such variables, and inference can go seriously astray if this is ignored. In this paper we survey unit root econometrics in an attempt to offer the applied macroeconomist some reliable guidelines. Unit roots can create opportunities as well as problems for applied work. In some unit root regressions, coefficient estimates converge to the true parameter values at a faster rate than they do in standard regressions with stationary variables. In large samples, coefficient estimates with this property are robust to many types of misspecification, and they can be treated as known in subsequent empirical exercises. On the other hand, such estimates may have poor finite-sample properties. A second goal of this paper is to indicate how applied researchers can exploit unit root econometric opportunities in finite samples of the size typically encountered in macroeconomics.

The early literature on unit roots concentrated on the univariate properties of macroeconomic time series. The seminal paper of Nelson and

*This paper was presented at the NBER Macroeconomics Conference, Cambridge, MA, March 8–9, 1991. We are grateful to Kevin Carey for research assistance, and to Olivier Blanchard, Jon Faust, Bruce Hansen, David Hendry, Søren Johansen, Rick Mishkin, Joon Park, and James Stock for helpful comments on the first version.

Plosser (1982), for example, carried out tests for unit roots in 14 individual time series. There is a great deal of subsequent work in this spirit, concerned, for example, with the persistence of fluctuations in real GNP (Campbell and Mankiw, 1987; Christiano and Eichenbaum, 1989; Cochrane, 1988; Perron, 1989a). We begin this paper with a thorough review of univariate unit root econometrics in Section 2.

It is characteristic of macroeconomics, however, that different time series are related by identities or behavioral models; therefore we emphasize multivariate unit root methods in this survey. Consider for example the system of five variables $(m_t, y_t, p_t, i_t, b_t)$, where m_t is the log nominal money stock, y_t is log nominal output, p_t is the log price level, i_t is the short-term nominal interest rate, and b_t is a long-term nominal bond yield. The variables in this system can be combined to form an ex post real interest rate $r_t \equiv i_t - \Delta p_t$, the nominal interest rate less the inflation rate. Now suppose that one wishes to analyze the unit root properties of the nominal interest rate, the inflation rate, and the real interest rate. Because of the identity linking these variables, if any two of them are stationary then the third variable must also be stationary. Univariate unit root tests cannot take account of this fact, which complicates inference.

The five variables listed above may also be linked together by behavioral relationships. Most practical work in macroeconometrics has the objective of estimating these relationships and testing hypotheses about them. Three obvious examples are as follows. First, one may wish to estimate a money demand function by regressing the log nominal money stock on the log price level, log nominal output, and the nominal interest rate, or by regressing the log real money stock on log real output and the nominal interest rate. Second, one may wish to test for Granger causality from money to output (either in nominal or real terms), in systems that may or may not include the nominal interest rate. Third, one may wish to test a hypothesis about the relationship between short-term and long-term nominal interest rates, such as the expectations theory of the term structure. It turns out that each of these empirical exercises is importantly affected by unit root econometric issues. At the end of this survey we return to these examples and show how the principles we discuss apply to them.

The organization of the paper is as follows. After the univariate discussion in Section 2, we review multivariate methods in Section 3. Both the univariate and multivariate sections of the paper first discuss alternative representations of time series with unit roots, and then discuss testing procedures. Section 3 also discusses in some detail how one can estimate cointegrated models. We do not attempt to provide a complete theoretical

review and we do not give full details of the various procedures.[1] Rather we discuss intuitively the main econometric procedures that are currently available and their relative strengths and weaknesses. We give extensive references to sources where further details can be found. We also occasionally state some "rules" to help structure the discussion. These should be viewed as useful guidelines or rules of thumb, and not as formal propositions. Throughout the paper we emphasize two themes. First, the proper handling of deterministic trends is a vital prerequisite for dealing with unit roots. Second, there are serious conceptual difficulties in distinguishing unit root processes from stationary processes in finite samples. Despite this fact, we argue that unit root econometric methods have many practical uses.

2. *Review of Univariate Procedures and Issues*

2.1 REPRESENTATION OF A TIME SERIES WITH AND WITHOUT A UNIT ROOT

It is often useful to think of a macroeconomic time series y_t as the sum of several components with different properties. We begin by writing

$$y_t = TD_t + Z_t. \tag{2.1}$$

Here TD_t is a deterministic trend in y_t and Z_t is the noise function or stochastic component of y_t. The unit root hypothesis concerns the behavior of the noise function, but the specification of the deterministic trend is crucial in testing this hypothesis. In principle a wide variety of specifications are possible, but the leading postulate is that TD_t is linear in time t, that is

$$TD_t = \kappa + \delta t. \tag{2.2}$$

We shall work primarily with the specification (2.2), but below we discuss some alternatives that have recently been proposed. For simplicity, we assume that the noise function Z_t can be described by an autoregressive-moving average process:

$$A(L)Z_t = B(L)e_t, \tag{2.3}$$

1. Recent theoretical surveys include Dickey, Bell, and Miller (1986), Perron (1988), Diebold and Nerlove (1990), Dolado, Jenkinson, and Sosvilla-Rivero (1990), and Phillips and Loretan (1991). Stock and Watson (1988b) is a particularly readable introduction with a macroeconomic perspective.

where $A(L)$ and $B(L)$ are polynomials in the lag operator L of order p and q, respectively, and e_t is a sequence of i.i.d. innovations.[2] The noise function Z_t is assumed to have mean zero, as the deterministic trend TD_t includes the mean of y_t. We also assume that the moving average polynomial $B(L)$ has roots strictly outside the unit circle. Equation (2.3) summarizes the univariate dynamics of the process Z_t. In this section we refer to the system (2.1)–(2.3) as a data-generating process (DGP) even though it may simply summarize the univariate implications of a more complex multivariate system.

We can now distinguish two alternative models for y_t. In the *trend-stationary* model the roots of $A(L)$, the autoregressive polynomial, are strictly outside the unit circle so that Z_t is a stationary process and y_t is stationary around a trend. In the *difference-stationary* model, Z_t has one unit autoregressive root and all other roots strictly outside the unit circle. In this case $\Delta Z_t \equiv (1 - L)Z_t$ is a stationary process and Δy_t is stationary around a fixed mean. The unit root hypothesis is that y_t is difference-stationary. The trend-stationary and difference-stationary models are often referred to as zeroth-order and first-order integrated models, or $I(0)$ and $I(1)$ models, respectively.[3]

To understand the meaning of the unit root hypothesis, it is useful to further decompose the noise function Z_t into a cyclical component C_t and a stochastic trend TS_t.[4] The cyclical component is assumed to be a mean-zero stationary process. The stochastic trend incorporates all random shocks that have permanent effects on the level of y_t. The sum of the deterministic trend TD_t and the stochastic trend TS_t is the overall trend. It is common in empirical macroeconomics to try to isolate the cyclical component C_t by subtracting from y_t the trend components TD_t and TS_t.

In the trend-stationary model, the decomposition of Z_t into stochastic trend and cycle is trivial, because Z_t is already assumed to be stationary so it satisfies the conditions assumed for the cycle C_t. In this case the stochastic trend TS_t is zero and the cycle C_t equals the noise Z_t. In the difference-stationary model, things are more complicated. When the

2. One could, of course, allow more general processes to characterize the noise function, such as "mixing type conditions," which permit some degree of heterogeneity and a richer class of serial correlation (see, e.g., Phillips, 1987; Phillips and Perron, 1988). However, the issues involved are easier to illustrate using the traditional ARMA(p,q) framework.

3. For simplicity we focus our discussion on the case of $I(0)$ versus $I(1)$ variables, which is the main case of interest to macroeconomists. This excludes the possibility of multiple unit roots, but most of the issues we discuss apply equally well to that case. For a general testing procedure allowing an arbitrary number of unit roots see Pantula (1989) and Dickey and Pantula (1987).

4. Note that the possibility of stochastic seasonal nonstationarity is beyond the scope of this paper.

polynomial $A(L)$ in (2.3) has a unit root, we can write $A(L) = (1 - L)A^*(L)$ where $A^*(L)$ has roots strictly outside the unit circle. The first difference ΔZ_t follows the stationary ARMA process $A^*(L)\Delta Z_t = B(L)e_t$. Following Beveridge and Nelson (1981), we can construct the following decomposition. Let $\psi(L) = A^*(L)^{-1}B(L)$ be the moving-average representation of the first difference of Z_t. The notation $\psi(1)$ denotes the sum of the moving-average coefficients. We define $\psi^*(L) = (1 - L)^{-1}[\psi(L) - \psi(1)]$, and find that ΔZ_t satisfies $\Delta Z_t = [\psi(1) + (1 - L)\psi^*(L)]e_t$. Then by applying the operator $(1 - L)^{-1}$ we can write

$$Z_t = TS_t + C_t = \psi(1)S_t + \psi^*(L)e_t, \tag{2.4}$$

where $S_t = \Sigma_{j=1}^{t} e_j$ is a zero mean random walk. Here the trend function for the variable y_t contains not only the deterministic trend TD_t, but also a stochastic component $TS_t = \psi(1)S_t$, which affects the intercept of the trend in each period. This stochastic trend is obtained from the sum of the moving average coefficients for ΔZ_t, which is equivalent to the long-run effect of a unit shock e_t on the level of the noise Z_t. The noise or cyclical component is $C_t = \psi^*(L)e_t$, constructed to have no long-run effect on the level of Z_t.

The decomposition (2.4) can be used to develop measures of the importance of the stochastic trend TS_t for the behavior of the variable y_t. Campbell and Mankiw (1987) propose that the coefficient $\psi(1)$ is a natural measure of persistence in y_t, because it is the ratio of the long-run effect of an innovation e_t to the immediate effect. When $\psi(1) > 1$, the long-run impact of a univariate shock to y_t is greater than the immediate impact; when $\psi(1) < 1$, on the other hand, shocks tend to die out. The case where y_t is a random walk has $\psi(1) = 1$, while the trend-stationary model for y_t is the limiting case where $\psi(1) = 0$. Cochrane (1988) proposes a related measure of persistence, which is the ratio of the variance of innovations in TS_t to the variance of innovations in y_t. It is straightforward to show that this variance ratio can be written as $\psi(1)^2\sigma_e^2/\sigma_{\Delta y}^2$. The quantity $\psi(1)$ is also closely related to the spectral density of the change in y_t, evaluated at frequency zero. We use the notation $h_{\Delta y}(0)$ to denote this spectral density. Then we have $h_{\Delta y}(0) = \psi(1)^2\sigma_e^2$, the numerator of Cochrane's variance ratio.

The trend-stationary and difference-stationary processes described above can be thought of as reduced form models. It is possible to derive these processes as reduced forms of a structural unobserved components model (see Harvey, 1985; Clark, 1987,1989; Watson, 1986; among others). Consider, for example, an unobserved components model that represents y_t as the sum of a random walk with drift and an independent

stationary process. When the innovation variance of the random walk is zero, y_t is trend-stationary. More generally, the reduced form of this model is a difference-stationary process with constraints (see Clark, 1987; Watson, 1986). Of particular relevance is the constraint that $\psi(1)$ in (2.4) is less than 1, i.e., that the long-term effect of innovations is no greater than the immediate effect. (Of course this constraint can be relaxed in more general unobserved components models.)

More recently, various nonlinear structural models have been proposed. These yield nonlinear reduced forms rather than the linear trend-stationary or difference-stationary reduced forms discussed so far. They try to capture the idea that two fundamentally different types of shocks are present. Some, which might be called "big shocks," occur infrequently and affect the trend function of the series in a permanent way. The others, call them "regular shocks," occur every period and may or may not affect the level of the series permanently. The unit root issue, in this context, centers on whether the "regular shocks" have a permanent effect on the level of the series.

One such class of models has been proposed by Hamilton (1989). His structural model makes y_t the sum of a nonlinear trend function and a linear ARIMA process with a root on the unit circle. The trend function is a random walk with a drift that switches between low and high values according to a first-order Markov process. Lam (1990) has derived a computational algorithm for a slightly more general version of this model where a unit root is not imposed a priori on the linear part of the process. Unfortunately technical difficulties are such that no procedures are yet available to test whether the linear part of the process does have a unit root or not.

Perron (1989a) has suggested that a time series structure with very infrequent changes in slope can be a useful approximation in empirical applications, indeed a simple one-time change in slope can be enough to characterize many series of interest. By restricting the number of changes in slope a priori, one can circumvent the technical difficulties with unit root tests in the Hamilton–Lam framework and obtain asymptotically valid tests of the null hypothesis that the linear part of the process contains a unit root. In this restrictive, but empirically useful, framework the reduced form of the series is described by (2.1) with the deterministic component given by $TD_t = \kappa + \delta_0 t + \delta_1(t - T_B) \cdot 1(t > T_B)$, where $1(\cdot)$ is the indicator function and T_B is the time of the change in the slope of the trend function. If a unit root is present in Z_t the trend function also contains a stochastic component in a manner similar to the usual difference-stationary process.

A similar model can be derived for series with infrequent changes in

intercept. Again, it was argued in Perron (1989a,1990a) that such a model with a single change can be a useful approximation in practice. In this restricted framework, it becomes possible to test the unit root hypothesis for the linear part of the process by specifying the deterministic component of the trend function as

$$TD_t = \kappa_0 + \kappa_1 1(t > T_B) + \delta t. \tag{2.5}$$

Thus the reduced form models described in Perron (1989a) can be viewed as approximations to structural models where infrequent changes in the intercept and/or slope are modeled stochastically as in Hamilton (1989) or Chen and Tiao (1990). The implicit assumption is that, in the given data set of interest, there is only one such "big shock." Of course, with other types of series or a longer span of data, it may be necessary to allow for more than one change.

2.2 TESTING FOR A UNIT ROOT

We begin by considering the simplest case where the noise component Z_t (the series y_t less its deterministic trend) is an AR(1) process with no moving average component, i.e., $Z_t = \phi Z_{t-1} + e_t$. This process can be rewritten as

$$\Delta Z_t = \pi Z_{t-1} + e_t, \tag{2.6}$$

where $\pi = \phi - 1$. Here the null hypothesis of a unit root is given by $\pi = 0$, while trend-stationarity implies that $\pi < 0$. This simplified framework is not realistic for most empirical applications, but it makes many of the issues easier to discuss. Later we outline how the procedures are modified if allowance is made for additional serial correlation.

2.2.1 Basic Tests of the Null Hypothesis of a Unit Root In testing the unit root hypothesis, it is important to draw a clear distinction between the maintained DGP and the regression equations that are used to test the null hypothesis. An important issue that often causes confusion is the appropriate treatment of the deterministic trend in y_t in these regression equations.

We use the notation DV_t (*deterministic variables*) for the set of variables that appears in the deterministic trend under the maintained DGP. In most applications $DV_t = \{1\}$, a constant, or $DV_t = \{1, t\}$, allowing a first-order polynomial in t. However DV_t can be more complicated; for example, the nonlinear structural model with a deterministic change in the intercept at date T_B has $DV_t = \{1, t, 1(t > T_B)\}$. Since we are interested in

the properties of the noise function, a natural strategy is first to "detrend" the series and analyze the time series behavior of the estimated residuals. We use the notation $\tilde{y}_t$ for the residuals of a projection of y_t on a set of *deterministic regressors* DR_t. The unit root hypothesis can be tested by estimating the pair of regressions:

$$y_t = \tau' DR_t + \tilde{y}_t; \qquad \Delta\tilde{y}_t = \pi\tilde{y}_{t-1} + u_t, \tag{2.7}$$

and using the t-statistic for testing $\pi = 0$, denoted $t_{\hat{\pi}}$. The natural choice of regressors DR_t is just the set of variables DV_t that appears in the deterministic trend under the maintained data generating process; however we discuss below what happens when DR_t differs from DV_t.

When the deterministic trend is linear in time ($DV_t = \{1\}$ or $\{1,t\}$), this two-step procedure will be asymptotically equivalent to a conventional one-step procedure where deterministic regressors DR_t^* are included in the autoregression,

$$\Delta y_t = \tau' DR_t^* + \pi y_{t-1} + u_t, \tag{2.8}$$

and where $DR_t^* = DR_t$. The regressors DR_t^* must include all the elements of DR_t for this asymptotic equivalence to hold. In particular, consider the case where $DV_t = DR_t = \{1,t\}$. The one-step procedure will be asymptotically equivalent to the two-step procedure only if the regressors DR_t^* in (2.8) include the trend t. The coefficient on the trend is $-\delta\pi$, which is zero under the null hypothesis of a unit root but is nonzero under the alternative hypothesis that y_t is trend-stationary. Thus the trend t must be included to enable the regression equation (2.8) to nest both the null hypothesis and the alternative hypothesis.

When the deterministic trend function TD_t is nonlinear, the relationship between the one-step procedure (2.7) and the two-step procedure (2.8) is more complicated. In the case of a trend with a single change in intercept as described in (2.5), where $DV_t = \{1, t, 1(t > T_B)\}$, the two-step procedure with $DR_t = DV_t$ is equivalent to the one-step procedure with $DR_t^* = \{1, t, 1(t > T_B), D(T_B)_t\}$, where $D(T_B)_t$ is one for $t = T_B + 1$ and zero otherwise. The extra regressor $D(T_B)_t$ must be included in the one-step procedure to allow a proper nesting of the null and alternative hypotheses, but this is not necessary in the two-step procedure. In the case of a trend with a change in slope, the two-step and one-step procedures may not be equivalent even asymptotically. For simplicity, in what follows we discuss the properties of two-step procedures, which are also the properties of one-step procedures in the usual case of a linear deterministic trend.

Once deterministic regressors have been chosen, we can test the implications of the unit root hypothesis for the regressions (2.7) or (2.8). We shall concentrate on the behavior of the t-statistic for testing $\pi = 0$ in (2.7) or (2.8) even though it is not the only statistic of interest in this unit root context. For example, the "normalized bias" $T\hat{\pi}$ also provides a valid test statistic as it is independent of nuisance parameters.[5]

The first important point to note is that the asymptotic distribution of $t_{\hat{\pi}}$, under the null hypothesis of a unit root, depends on the deterministic terms included as regressors. Assume for the moment that the included deterministic regressors contain at least all the deterministic components in the data generating process for y_t.

Rule 1: Suppose that the deterministic regressors DR_t used to construct $\tilde{y}_t$ in (2.7) contain at least the deterministic variables DV_t included in the maintained data generating process. Then under the null hypothesis of a unit root, the asymptotic distribution of $t_{\hat{\pi}}$ is nonnormal and varies with the set DR_t. In the case where the maintained DGP has a linear trend, the same result holds for regression Equation (2.8) when the deterministic regressors DR_t^* include at least the variables DV_t.

Critical values for the asymptotic distribution of $t_{\hat{\pi}}$ can be found in the following sources for different sets of included deterministic regressors. For $DR_t = \{\emptyset\}$, $\{1\}$ or $\{1, t\}$, see Fuller (1976); for $DR_t = \{1, t, t^2, \ldots, t^p; p = 2, \ldots, 5\}$, see Ouliaris, Park, and Phillips (1989); for $DR_t = \{1, 1(t > T_B)\}$, see Perron (1990a); for $DR_t = \{1, t, 1(t > T_B)\}$, $\{1, t, (t - T_B)1(t > T_B)\}$ and $\{1, t, 1(t > T_B), t\, 1(t > T_B)\}$, see Perron (1989a).[6] The basic reason for the dependence of the null asymptotic distribution on the included deterministic regressors is the fact that the specified trend function needs to be estimated. If the true coefficients of the DGP were known, only a single set of critical values would be needed, namely that where $DR_t = \{\emptyset\}$, the null set. The tabulated critical values also have important implica-

5. Asymptotic critical values of the normalized bias can be found in the same sources given below for the critical values of $t_{\hat{\pi}}$. The normalized bias forms the basis for a transformed test statistic proposed by Phillips and Perron (1988) and discussed below. Dickey and Fuller (1981) consider individual t-statistics on the coefficients of the deterministic components; these are, however, of little practical use because their null distribution depends on nuisance parameters. More useful are likelihood ratio statistics considered by Dickey and Fuller such as a test for the joint hypotheses that $\pi = 0$ and $\delta = 0$ in (2.8) with $DR_t = \{1,t\}$ as in (2.2). However simulation experiments reported in Dickey and Fuller (1981) suggest that these statistics have lower power than $t_{\hat{\pi}}$.

6. Finite sample and asymptotic critical values are also available for tests of the unit root in models with a structural change in intercept and/or slope when the date of the change is assumed unknown; see Banerjee, Lumsdaine, and Stock (1990), Perron (1990b), Perron and Vogelsang (1990), and Zivot and Andrews (1990).

tions for the power of unit root tests, that is, the probability that the tests reject the null hypothesis of a unit root when a trend-stationary alternative hypothesis is true. We summarize these implications in the following rule.

Rule 2: Under the null hypothesis of a unit root, the left-tailed critical values of the asymptotic distribution of $t_{\hat{\pi}}$ increase in absolute value with the number of included deterministic regressors.

Things are different when the set of included deterministic regressors does not contain all the components of the deterministic trend. Of particular interest is the following.

Rule 3: Suppose that DR_t omits a variable in DV_t that is growing at a rate at least as fast as any of the elements of DR_t. Then under the null hypothesis of a unit root, the statistic $t_{\hat{\pi}}$ in (2.7) can be normalized in such a way that its asymptotic distribution is standard normal. In the case where the maintained DGP has a linear trend, a similar result describes the set of regressors DR_t^* and the distribution of $t_{\hat{\pi}}$ in the one-step regression (2.8).

Rule 3 applies most obviously to the case where a nonzero linear trend is present in the DGP but is omitted from the deterministic regressors DR_t (Perron and Phillips, 1987; West, 1988). It also applies when the DGP contains higher-order polynomial trends that are omitted from the regression (Sims, Stock, and Watson, 1990).

It is important not to misinterpret Rule 3. The rule seems at first to suggest that one could increase the power of unit root tests by omitting certain deterministic regressors that are present in the data generating process. Consider for instance using the t-statistic for testing $\pi = 0$ in a regression without a trend in the case where the DGP is a unit root process with drift. In this case the asymptotic distribution is normal and the critical values are smaller (in absolute value) than the nonnormal asymptotic critical values obtained when a trend is included as a regressor. However there are two reasons why this approach is misguided. First, the finite sample distribution of $t_{\hat{\pi}}$ is not invariant to the values of the parameters of the trend and for small values the normal approximation may be inadequate. Second, and more important, this procedure leads to tests whose power goes to zero as the sample size increases. This is an extreme form of inconsistency (an inconsistent test being defined as one whose power against fixed alternatives does not go to one as the sample size increases). This is stated in the following rule.

Rule 4: (1) Assume that DR_t omits a variable in DV_t that is growing at a rate at least as fast as any of the elements of DR_t. Then the power of the statistic $t_{\hat{\pi}}$ in (2.7) goes to zero as the sample size increases. (2) Suppose that DR_t fails to include a variable in DV_t that is nontrending (e.g., a mean or a change in mean). Then $t_{\hat{\pi}}$ in (2.7) is a consistent test but the finite sample power is adversely affected and decreases as the coefficient on the omitted component increases. Similar results apply to the set of regressors DR_t^* in the one-step procedure (2.8).

It is best to illustrate these results with a few examples. For part (1), consider first the case where the DGP is a stationary process around a deterministic trend function of the form $TD_t = \kappa + \delta t$ and only a constant is included as a deterministic regressor. This case is discussed in Perron (1988). Now consider applying the regression equation $\Delta y_t = c + \pi y_{t-1} + e_t$. If the DGP contains a trend component, the only way to fit this trend is to have $\pi = 0$, in which case c becomes the coefficient δ on the trend.[7] In a similar way, if the DGP specifies a trend function with a changing slope, a test of the unit root constructed using only a constant and a time trend as deterministic regressors will yield an inconsistent test. For an example of part (2), suppose no deterministic regressors are included but the DGP specifies that y_t has a nonzero mean, then the power of the test will decrease to zero as the mean increases (in absolute value). Similarly, if the DGP specifies a change in the intercept of the trend function at some date and no regressors are included to account for it, the power of the test will decrease as the magnitude of the change in mean increases.

Rule 4 shows the importance of including as many deterministic regressors as there are deterministic components in the trend function of the data-generating process. Otherwise the test will at best lose finite-sample power or at worst have power that goes to zero as the sample size increases. On the other hand, it is desirable not to include extraneous deterministic regressors. The following rule states the general behavior when extraneous regressors are included.

Rule 5: Suppose that $t_{\hat{\pi}}$ is constructed using a set of deterministic regressors, DR_t, that includes at least all the deterministic components under the relevant DGP. The power of a test of the unit root hypothesis against stationary alternatives decreases as additional deterministic regressors are included.

7. Kleidon (1986) runs unit root tests on aggregate earnings and dividends omitting a time trend. These tests are an example of this problem in practice.

The statement in rule 5 is partially justified by the statement in rule 2 that the critical values increase (in absolute value) with the number of extraneous deterministic regressors. However, this must be counterbalanced by the fact that, in finite samples, there is a downward bias away from zero in the estimate of π and this bias increases as the number of extraneous deterministic regressors increases. The justification for the statement in rule 5 comes from various published and unpublished simulation studies (see, e.g., Schwert, 1989; Dejong, Nankervis, Savin, and Whiteman, 1990a).

Rules 4 and 5 suggest that care must be exercised in choosing the appropriate deterministic regressors to include to have tests with reasonable power properties. When it is not clear which set of deterministic regressors to include, a sequential testing procedure may be useful. Such a sequential testing strategy is described in Perron (1988) for the case where the class of trend functions under the DGP includes either no component, a constant, or a constant and a trend. Briefly, it was argued in that paper that a proper testing strategy should start from the most general trend specification (in that context, a first-order trend polynomial) and test down to more restricted specifications. In the more general case where the deterministic trend component is allowed to contain more than a simple first-order polynomial in time, such a sequential testing procedure cannot yet be applied given that the distribution theory for the relevant statistics has not been derived. Experimentation with various trend specifications should be guided by the following general rule, which summarizes our discussion of deterministic components.

Rule 6: A nonrejection of the unit root hypothesis may be due to misspecification of the deterministic components included as regressors.

2.2.2 Issues Concerning Power and Frequency of the Data Applied researchers are often faced with choices among different types of data set for a given time series. This can occur, in particular, when data are available at different sampling frequencies for different lengths of time. For instance, it is common to have quarterly observations for the period after World War II, while monthly observations may be available starting in the early 1960s. On the other hand, data covering longer horizons are often available only at an annual frequency. An annual data set might typically contain around 100 observations, while a quarterly data set might contain more than 160 and a monthly one over 300. It is then natural to ask which data set would allow the greatest discriminating power. Is a greater number of observations better in terms of power?

It turns out that for tests of the unit root hypothesis versus stationary alternatives the power depends very little on the number of observations per se but is rather influenced in an important way by the span of the data. For a given number of observations, the power is largest when the span is longest. For a given span, additional observations obtained using data sampled more frequently lead only to a marginal increase in power, the increase becoming negligible as the sampling interval is decreased (see Shiller and Perron, 1985; Perron, 1990c).[8] In most applications of interest, a data set containing fewer annual data over a long time period will lead to tests having higher power than if use was made of a data set containing more observations over a short time period. These results show that, whenever possible, tests of the unit root hypothesis should be performed using annual data over a long time period. This conclusion is reinforced by the fact that seasonal adjustment procedures often create a bias toward nonrejection of the unit root hypothesis (see Ghysels and Perron, 1990; Jaeger and Kunst, 1990).

On the other hand, long historical data series may pose other problems. First, it may be the case that the quality of historical data is questionable and that the early methods of construction spuriously induce a bias against one or the other hypothesis. For instance, Jaeger (1990) argues that before World War II the method of linear trend interpolation was common and may induce a bias in favor of rejecting the unit root hypothesis. Second, using a long sample of data increases the possibility that the series of interest is affected by a major structural change in the process characterizing either the trend function or the noise component. The presence of such a structural change would bias the test in favor of the unit root hypothesis. Hence, though using a data set over the longest period possible is desirable in terms of power properties, care must be taken in interpreting the results.

2.2.3 Extensions to Processes with Additional Correlation We now consider extensions that are necessary when allowance is made for possible additional serial correlation in the noise component of the DGP. We consider the case where the noise function Z_t obeys the ARMA(p,q) process (2.3), $A(L)Z_t = B(L)e_t$, rather than the AR(1) model (2.6). The points made above remain valid in this more general setting but a new issue arises,

8. Perron (1989b, 1990c) also considers testing the random walk hypothesis using a test of randomness applied to the first-differences of the data. Such a test is commonly used in finance. He shows that such a test has a power function that is dominated by unit root tests on levels. Also the power decreases to the size of the test as the number of observations increases with a fixed span of data. Hence in this case too many observations destroy the power of the test.

namely that the asymptotic distribution of the statistic $t_{\hat{\pi}}$ in first-order autoregressions such as (2.7) or (2.8) depends on the correlation structure of the data. Hence, modifications are necessary to get rid of this dependency on nuisance parameters. Two approaches seem natural, a parametric and a nonparametric one.

Dickey and Fuller (1979) and Said and Dickey (1984) consider a parametric correction motivated by the case of a pure AR(p) process, i.e., $A(L)Z_t = e_t$, where $A(L) = 1 - a_1L - \cdots - a_pL^p$. In this case, we can write $\Delta Z_t = \pi Z_{t-1} + \Sigma_{j=1}^{p-1}\gamma_j\Delta Z_{t-j}$, where $\pi = \Sigma_{i=1}^{p}a_i - 1$ is the difference between the sum of the autoregressive coefficients and one, while $\gamma_j = -\Sigma_{i=j+1}^{p}a_i$. As before, the noise component Z_t has a unit root if $\pi = 0$. The regression equation (estimated by OLS) then takes the form

$$y_t = \tau' DR_t + \tilde{y}_t; \qquad \Delta\tilde{y}_t = \pi\tilde{y}_{t-1} + \Sigma_{j=1}^{k}\gamma_j\Delta\tilde{y}_{t-j} + u_t, \tag{2.9}$$

or

$$\Delta y_t = \tau' DR_t^* + \pi y_{t-1} + \Sigma_{j=1}^{k}\gamma_j\Delta y_{t-j} + u_t, \tag{2.10}$$

where $k = p - 1$. Here DR_t and DR_t^* are vectors of deterministic regressors as discussed above. In the case of a pure AR(p), the asymptotic distribution of $t_{\hat{\pi}}$ obtained from (2.9) or (2.10) is the same as the asymptotic distribution of $t_{\hat{\pi}}$ obtained using a first-order autoregression with $\Delta Z_t = \pi Z_{t-1} + e_t$. In the more general case where the noise component is an ARMA(p,q), Said and Dickey (1984) suggest that the process can be approximated by a high-order autoregressive process, in which case the regression specifications (2.9) and (2.10) remain appropriate. The technical condition for such a procedure to remain asymptotically valid is that the order of the estimated autoregression, k, increases to infinity at a suitable rate as the sample size increases to infinity.

In practice, the choice of the truncation lag parameter k is an issue. First, even in the pure AR(p) case, the order p is usually an unknown variable. In the general ARMA(p,q) case, the theoretical conditions for the asymptotic validity of the procedure are not informative enough to guide any choice in finite samples. This problem is of importance because it is often the case that the outcome of the test depends on the particular choice of this truncation lag parameter. Several factors may explain such a sensitivity. First, too few lags may adversely affect the size of the test. Second, the introduction of too many lags may reduce power (because of more parameters being estimated and a reduced number of effective observations, given the need for additional initial conditions). Finally, as k changes, the initial conditions also change. This last factor

may be of importance given the noninvariance of the power function of the statistics to the initial conditions (see, e.g., DeJong, Nankervis, Savin, and Whiteman, 1990b). These factors point to the importance of choosing the truncation lag parameter judiciously. The choice of a fixed k, independent of the data, is likely to be inappropriate. The following data-dependent procedure is easy to implement and is likely to yield tests with better size and power properties.

SUGGESTED PROCEDURE TO SELECT k[9] Start with some upper bound on k, say k_{max}, chosen a priori. Estimate an autoregression of order k_{max}. If the last included lag is significant (using the standard normal asymptotic distribution), select $k = k_{max}$. If not reduce the order of the estimated autoregression by one until the coefficient on the last included lag is significant. If none is significant, select $k = 0$.

Such a procedure is studied in some detail by Hall (1990). It is in fact motivated by the pure AR(p) case. In the case of an AR(p), such a procedure will select k greater than or equal to the true order with probability one asymptotically and the distribution of $t_{\hat{\pi}}$ will be the same as in the fixed k case, provided the upper bound k_{max} is selected greater than the true order. In the general case where moving-average components are present no general consistency results are available yet. We conjecture, however, in analogy with the Said–Dickey (1984) extension, that the asymptotic distribution would remain unchanged provided the upper bound k_{max} increases at a suitable rate as the sample size increases to infinity. Simulation evidence presented in Hall (1990) suggests that such a data-based method induces little size distortion in finite samples. It is important, however, to note that the sequential method must proceed from a general model to more specific ones. An alternative procedure would be to select the order by starting from a parsimonious specification and including additional lags until the last one is significant, but this is not asymptotically valid and leads to more serious size distortions in finite samples.

An alternative way to handle additional serial correlation in the noise process Z_t has been proposed by Phillips (1987) and Phillips and Perron (1988). Their approach is to add to the original unit root test statistic a correction factor that eliminates the dependency of the asymptotic distribution on the serial correlation of Z_t. The correction uses a nonparamet-

9. Of course, this is not the only possible data dependent procedure for selecting k. Any procedure that selects k at least as large as the correct autoregressive order asymptotically, for example by using an information criterion or a joint F-test of significance on additional lags, will be adequate. The rule stated here has the advantage of simplicity, but additional work is needed on the finite-sample properties of alternative procedures.

ric estimate of the spectral density of ΔZ_t at frequency zero, measured relative to the sample variance of ΔZ_t. This nonparametric estimate is a weighted sum of the autocovariances of ΔZ_t, where the weights are chosen in such a way that the estimated spectral density is positive by construction. Phillips and Perron derive transformed versions of both the normalized bias $T\hat{\pi}$ and the t-statistic $t_{\hat{\pi}}$, but the former is preferable as it is more powerful.

These test statistics are easy to implement and asymptotically valid under quite general conditions. However, several simulation studies have shown that they have serious size distortions in finite samples when the data-generating process has a predominance of negative autocorrelations in first differences (see, e.g., Schwert, 1989; Phillips and Perron, 1988; DeJong, Nankervis, Savin, and Whiteman, 1990b). This suggests that the Phillips–Perron tests may be less reliable than the Dickey–Fuller methodology where a parametric correction is applied. An important fact that leaves some hope for this class of statistics is that simulation evidence suggests their size-adjusted power is substantially higher than the power of augmented Dickey–Fuller statistics. Therefore an important topic on the research agenda is to find a way to modify the Phillips–Perron procedure in such a way as to alleviate the size problem while retaining good power properties. Preliminary investigation by Stock (1990) seems to indicate that some improvements are possible on this front. We discuss some of this evidence in the next subsection.

2.2.4 Alternative Approaches to the Unit Root Issue So far we have followed the bulk of the existing literature by focusing on the properties of coefficients and t-statistics in autoregressions for the variable y_t. Recently some authors have explored the implications of the unit root model compared to those of a trend-stationary model by looking at the asymptotic behavior of the series $\{y_t\}$ itself. In many ways this is a simpler approach.

Suppose for the moment that the DGP contains no deterministic component so that $y_t = Z_t$, a zero mean ARMA(p,q) model. If y_t contains a unit root, we have, under general conditions, that $T^{-1/2}y_T$ converges in distribution to an appropriately scaled Brownian motion. Under the hypothesis that y_t does not contain a unit root, we have $T^{-1/2}y_T$ converging to zero. Stock (1990) has used this idea to develop a class of statistics to test the null hypothesis of a unit root. The statistics can easily be extended to allow for deterministic components in the trend function by running preliminary regressions of y_t on the deterministic variables. Just as before the asymptotic distribution of these statistics varies with the set of deterministic components included. Stock suggests, among other tests, modifications of the Sargan–Bhargava (1983) and Phillips–

Perron (1988) procedures based on an autoregressive spectral density estimator. Simulation evidence suggests that the size problem is alleviated while the power is greater than that of most available statistics. This is an interesting avenue for further research.

This idea of using the different behavior of sample moments of the data under the hypotheses of a unit root and of stationarity extends in a natural way to provide statistics for the null hypothesis of stationarity versus the alternative hypothesis of a unit root. Consider for instance the quantity $T^{-3/2}\Sigma_1^T t y_t$. Under the hypothesis that y_t is a zero mean stationary process, this converges to a nondegenerate normal distribution with a variance that is a function of the spectral density of y_t at frequency zero. Under the hypothesis that y_t follows a unit root process, this statistic explodes. Park and Choi (1988) suggest a test for the null hypothesis of stationarity that uses superfluous regressors. Their test can be seen as exploiting the behavior of the statistic discussed here. We give further details in a multivariate context below.

2.3 THE NEAR-OBSERVATIONAL EQUIVALENCE OF TREND- AND DIFFERENCE-STATIONARY PROCESSES

In the last section we discussed the possibility of testing both the null hypothesis of a unit root process and the null hypothesis of a trend-stationary process. This naturally leads us to ask what is the relation between these two classes of models and what is the importance of specifying one or the other hypothesis as the null.

We first recall from our discussion of the Beveridge–Nelson decomposition that a unit root process is one for which the spectral density of the first difference, $h_{\Delta y}(0)$, is nonzero. A trend-stationary process, by contrast, has $h_{\Delta y}(0) = 0$. This means that the unit root hypothesis is a composite null hypothesis, which has the following interesting implication.

Rule 7: In finite samples, any trend-stationary process can be approximated arbitrarily well by a unit root process (in the sense that the autocovariance structures will be arbitrarily close).

This point has been highlighted by Blough (1988) and Cochrane (1991). The idea is quite simple. For any trend-stationary process, we have $h_{\Delta y}(0) = 0$. A unit root process with $h_{\Delta y}(0) = \epsilon$, say, with $\epsilon > 0$ can arbitrarily approximate a trend-stationary process provided ϵ is chosen small enough relative to the sample size. The following example illustrates this point in a straightforward way. Consider an ARMA(1,1) process:

$$y_t = \phi y_{t-1} + u_t + \theta u_{t-1}. \tag{2.11}$$

This process is difference stationary when $\phi = 1$ and $-1 < \theta < 1$, but trend stationary (with a zero trend) when $-1 < \phi < 1$ and $-1 < \theta < 1$. Consider the case where the trend-stationary process has $\phi = \theta = 0$ (so the series is white noise), while the unit root process has $\phi = 1$, $-1 < \theta < 1$ [so the series is an IMA(1,1) with a negative moving average coefficient]. For any finite sample size, the trend-stationary process will be approximated arbitrarily well by the difference-stationary process (in the sense that they will have an arbitrarily close autocovariance structure) provided θ is close enough to but not equal to -1. This fact has the following interesting implication concerning the power of unit root tests in finite samples.

Rule 8: In finite samples, any test of the unit root hypothesis against trend-stationary alternatives must have power no greater than its size.

Rule 8 is simply an implication of the fact that the probability distributions of the statistics of interest are continuous in the parameters of the process for y_t. Therefore, given rule 7, the finite sample distribution of any statistic under a particular trend-stationary process can be arbitrarily close to the finite sample distribution of the statistic under a difference-stationary process that approximates the trend-stationary process. In terms of the example (2.11), the critical values of a unit root test must be chosen such that the probability of rejection is less than or equal to the size of the test for any value of the parameter θ in the interval $(-1 < \theta < 1)$. But when θ is arbitrarily close to -1 the unit root process is indistinguishable from a trend-stationary process, so the test must have power equal to its size against such a process. Using a rejection region based on the asymptotic distribution therefore means that any test will have an exact size greater than its nominal size for some part of the parameter space permitted under the null hypothesis. Schwert (1989) presents Monte Carlo results that illustrate this point.

Some have argued that this problem occurs because testing the null hypothesis of a unit root versus the alternative of a trend-stationary process implies testing a composite null hypothesis $[h_{\Delta y}(0) \neq 0]$ versus a point alternative $[h_{\Delta y}(0) = 0]$. The argument is then that the problem could be avoided by reversing the null and the alternative and testing the null hypothesis of trend-stationarity versus the alternative hypothesis of a unit root process. This argument is, however, incorrect as one can always express the trend-stationary hypothesis as a composite null and the unit root hypothesis as a point alternative. Consider, for example, the following measure. Let hl denote the half life of a shock e_t on the level of the series y_t. Consider now the quantity hl^{-1}. For any difference-

stationary process $hl^{-1} = 0$ while for any trend-stationary process $hl^{-1} > 0$. By analogy with rule 7 we have the following.

Rule 9: In finite samples, any unit root process can be approximated arbitrarily well by a trend-stationary process (in the sense that the autocovariance structures will be arbitrarily close).

This result follows because for any unit root process there will exist a trend-stationary process for which shocks have effects on the level of $\{y_t\}$ that are arbitrarily close to being infinite. For example in the simple first-order autoregressive model, $y_t = \phi y_{t-1} + e_t$, the random walk process can be arbitrarily well approximated, in any finite sample, by a stationary process with ϕ less than but close to one. Following the same logic as in the case of tests of the null hypothesis of a unit root, we have the following.

Rule 10: In finite samples, any test of the trend-stationarity hypothesis against unit root alternatives must have power no greater than its size.

The special feature of importance here is that for any trend-stationary process there is a difference-stationary process that approximates it arbitrarily well in finite samples *and vice versa*. It is this dual relationship stated in rules 7 and 9 that creates a problem beyond what one usually encounters in hypothesis testing. Given the statements in rules 8 and 10, should we altogether abandon the idea of trying to discriminate between a unit root process and a trend-stationary process? Some have argued that we should (e.g., Christiano and Eichenbaum, 1989). We favor a more pragmatic answer to this question, namely that we should still try to distinguish these two classes of processes while keeping in mind that strictly speaking we may reach incorrect conclusions if the DGP belongs to a particular subset of the parameter space.

For the argument that follows, consider the usual framework where the unit root is the null hypothesis. When applying any test of the unit root using asymptotic critical values, it must be the case that the test has an exact size greater than the nominal size for difference-stationary processes that are within some neighborhood region of the class of trend-stationary processes.[10] The magnitude of this region decreases as the

10. One might think that this neighborhood could be conveniently parameterized in terms of the quantity $h_{\Delta y}(0)$. Unfortunately this is not the case because one can take a unit root process with any value of $h_{\Delta y}(0)$ and find a trend-stationary process that approximates it arbitrarily well in a finite sample. It might be possible to characterize the neighborhood region in terms of the behavior of the spectral density function near the zero frequency, but more work is needed on this topic.

sample size increases. A given testing procedure is said to have better finite sample properties than another procedure if the region where the size becomes greater than its nominal counterpart is smaller for a given sample size T. In any event, unit root tests must be viewed in a context where the parameter space under the null hypothesis is restricted (the more so with smaller sample sizes). The same comments apply to tests of the null hypothesis of trend-stationarity.

Why should we be willing to use procedures that yield improper inference for some part of the parameter space? The answer is a pragmatic one. For practical purposes it does not really matter if we label a difference-stationary process with coefficient $h_{\Delta y}(0)$ close to zero as a trend-stationary process, or if we label a trend-stationary process with extremely persistent shocks as a difference-stationary process. Indeed these kinds of errors may even have practical advantages.

To illustrate this last point we conducted a small Monte Carlo experiment. We considered the family of ARMA(1,1) processes given in (2.11). We simulated both difference-stationary ARMA(1,1) processes that are close to being trend-stationary, and trend-stationary ARMA(1,1) processes that are close to being difference-stationary. The former processes have $\phi = 1$ and θ approaching -1 (for the Monte Carlo experiment we chose $\theta = -0.5, -0.8, -0.9, -0.95$, and -0.98). The latter processes have ϕ approaching 1 (we set $\theta = 0$ and chose $\phi = 0.5, 0.8, 0.9, 0.95$, and 0.98). For each data-generating process, we drew 5000 samples of length 100 and ran standard unit root tests with estimated linear trends. We calculated the augmented Dickey–Fuller $t_{\hat{\pi}}$ statistic and the Phillips–Perron transformation of the normalized bias $T\hat{\pi}$, denoted by $Z(\hat{\pi})$. For the former we used the lag length selection procedure described in the text, setting $k_{\max} = 6$; for the latter we set $k = k_{\max}$. Table 1 reports the fraction of 5000 runs in which the unit root test statistics exceeded their asymptotic 5% critical values.

Two points are very clear from this exercise. First, when the true DGP has a unit root but is close to being stationary, the unit root tests have severe size distortions: They reject the true null hypothesis too often. To take the most extreme case, when $\phi = 1$ and $\theta = -0.98$, the unit root hypothesis is falsely rejected at the 5% level at least 98% of the time. The reason for this is of course that in a finite sample the process looks very much like white noise; the unit root component, which dominates asymptotically, has only a small effect in a sample of length 100. Second, when the true DGP is stationary but has a root close to unity, then the unit root tests have very little power. If we compare the integrated case $\phi = 1$ and $\theta = 0$ with the stationary case $\phi = 0.98$ and $\theta = 0$, for example,

Table 1 UNIVARIATE MONTE CARLO RESULTS

Data generating process (ϕ,θ)	*Fraction of rejections at 5% level*		*Out-of-sample mean squared error*			
	Said–Dickey	Phillips–Perron	Levels	Differences	Said–Dickey	Phillips–Perron
1, −0.98	0.98	1.00	1.06	1.23	1.06	1.06
			1.06	1.24	1.07	1.06
1, −0.95	0.96	1.00	1.05	1.21	1.06	1.05
			1.12	1.26	1.13	1.12
1, −0.90	0.91	1.00	1.08	1.17	1.09	1.08
			1.38	1.39	1.37	1.38
1, −0.80	0.71	1.00	1.15	1.12	1.15	1.15
			2.48	2.11	2.31	2.48
1, −0.50	0.28	0.77	1.15	1.06	1.10	1.14
			9.87	7.07	7.54	8.96
1, 0	0.09	0.06	1.13	1.05	1.06	1.06
			34.3	25.2	25.6	25.4
0.98, 0	0.10	0.06	1.13	1.06	1.07	1.07
			28.0	21.6	21.8	21.8
0.95, 0	0.14	0.11	1.11	1.06	1.08	1.08
			17.4	16.0	16.1	16.1
0.90, 0	0.29	0.25	1.10	1.09	1.10	1.10
			9.54	10.4	10.2	10.3
0.80, 0	0.70	0.73	1.08	1.14	1.11	1.11
			4.23	5.78	4.98	4.85
0.50, 0	0.96	1.00	1.08	1.23	1.08	1.08
			1.60	2.16	1.61	1.60

Notes: This table reports the results of a Monte Carlo experiment with 5000 replications. Samples of length 100 were generated from the process $X_t = \phi X_{t-1} + u_t + \theta u_{t-1}$, with standard normal innovations u_t and values of ϕ and θ given in the first column. Said–Dickey t_π and Phillips–Perron $Z(\pi)$ unit root tests were performed on each sample, using estimated trends and selecting lag length by the procedure described in the text, with the maximum lag length $k_{max} = 6$. The second and third columns of the table report the empirical rejection probabilities of nominal 5% tests.

At the end of each sample, 1- and 20-period-ahead forecasts were formed using an autoregressive model in levels, and an autoregressive model in differences. For each model lag length was chosen using the selection procedure described in the text, with the maximum lag length $k_{max} = 6$. For each sample and forecast horizon, out-of-sample mean squared errors of forecast were calculated using 25 draws of the data-generating process. The table reports average out-of-sample mean squared errors across all replications, for the levels model, the differences model, and two mixed models. The mixed models use the levels model when the Said–Dickey or Phillips–Perron unit root test rejects, and the differences model otherwise. For each data generating process, the first row gives the results for one-period-ahead forecasts, and the second row gives the results for 20-period-ahead forecasts.

we find that the rejection rate is no more than 1% greater for the stationary case than for the unit root case.

So far these Monte Carlo results are quite standard (see, for example, Schwert, 1989). However we now show that the cases in which the unit root tests give false answers are also ones in which these false answers may have some practical utility. For each of our artificial data samples, we estimated an autoregressive forecasting model in levels and another autoregressive model in differences. In the former model a linear trend is estimated, while in the latter model the mean of the differenced data is estimated. We chose the lag length for each model using the lag length procedure described in the text, with $k_{max} = 6$. Then we used the models estimated up through period 100 to form out-of-sample forecasts one period ahead and 20 periods ahead, that is forecasts of y_{101} and y_{120}. We drew 25 realizations of y_{101} and y_{120} from the true DGP, and calculated out-of-sample mean squared errors of forecast for the simulation. Finally we averaged across all 5000 simulations to get average mean squared errors at horizons one and 20 for the levels and differences forecasting models. These average mean squared errors are reported in the fourth and fifth columns of Table 1. For each DGP, the mean squared errors for one-period-ahead forecasts appear above those for 20-period ahead forecasts.

The main point to note is that near-stationary unit root DGPs are better forecast using stationary forecasting models, while near-integrated stationary DGPs are better forecast using integrated forecasting models. Among the DGPs we consider, stationary forecasting models are superior for all processes with $\phi = 1$ and $\theta \leq -0.90$, while unit root forecasting models are superior for all processes with $\theta = 0$ and $\phi \geq 0.90$ (one period ahead) or 0.95 (20 periods ahead). The table also reports the average out-of-sample mean squared errors for mixed strategies. These use the levels forecasting model when the Said–Dickey or Phillips–Perron tests reject at the 5% level, and the differences forecasting model otherwise. For most DGPs, the mixed strategies have mean squared errors that are close to those of the best pure forecasting model.[11] These results illustrate that unit root test procedures can be practically useful for improving the quality of macroeconomic forecasts, even in small samples where they have only a limited ability to distinguish unit root processes from stationary processes. The example studied here is simple, but we believe it illustrates a fairly general principle.

11. There are some DGPs for which the unit root tests do not achieve the best possible mean squared errors. For example when $\phi = 1$ and $\theta = -0.8$, the unit root tests tend to reject the null even though the best forecasting model is a difference stationary model. However this phenomenon tends to occur in cases where the difference in forecasting performance between the unit root and trend stationary models is small.

Unit root tests have some other uses in finite samples. They can be helpful if one wishes to know whether stationary or integrated asymptotic distributions provide a better approximation in a particular application. Consider for example an AR(1) model. The limiting distribution of the least-squares estimator of the autoregressive parameter has a normal asymptotic distribution if the autoregressive parameter is less than one. If this parameter is close to one, however, the unit root asymptotic distribution actually provides a better finite-sample approximation than the asymptotically correct normal distribution (Evans and Savin, 1981). In more general contexts also, it may be better to use integrated asymptotic theory for near-integrated stationary models, and stationary asymptotic theory for near-stationary integrated models. In principle, of course, it would be better to have recourse to the exact finite sample distribution but in practice this can rarely be calculated analytically. Unit root tests are a simple alternative to extensive Monte Carlo simulations, which are usually needed to calculate finite sample distributions.

Unit root tests can also help researchers to impose plausible restrictions on more structural time series models. Unit root restrictions may help to increase the efficiency of estimates (i.e., reduce mean squared error) even if the variables in the model do not have true unit roots but are near-integrated. This is just a restatement of the general principle, familiar in the case of zero restrictions, that imposing false restrictions may help reduce the mean squared error of estimates. False restrictions increase the bias of forecasts, but they may reduce the variance by enough that the mean squared error is actually reduced. The Monte Carlo experiment described above illustrates this phenomenon in a univariate context, but the general principle is perhaps even more important in multivariate time series models.

3. *Review of Multivariate Procedures and Issues*

This section discusses issues related to unit roots in a multivariate context. Throughout we shall consider the properties of a vector y_t of n variables, for each of which a sample of size T is available. The discussion is organized into four main sections: (1) representation and characterization of the models describing the evolution of the vector y_t with particular emphasis on the issue of cointegration among their elements, (2) testing procedures related to cointegration, (3) estimation and inference in multivariate models with cointegration, and (4) a discussion about situations where these techniques are necessary and where they are not.

3.1 REPRESENTATION OF MULTIPLE TIME SERIES WITH SOME UNIT ROOTS

3.1.1 Basic Concepts of Cointegration We start with an $(n \times 1)$ vector of variables y_t. To keep the framework simple, we suppose that each element of this vector has a representation given by

$$y_{it} = TD_{it} + Z_{it}; \qquad A_i(L)Z_{it} = B_i(L)e_{it}, \qquad (i = 1, \ldots, n) \tag{3.1}$$

where TD_{it} is the deterministic component of variable i, Z_{it} is its noise function modeled as an ARMA process, and the innovation e_{it} is $N(0, \sigma_i^2)$. This is the same model we considered in the previous section. As before, we assume that y_{it} contains at most one autoregressive unit root and that the remaining roots are strictly outside the unit circle.[12] Note that the model (3.1) allows all variables to have nonzero deterministic trends. For simplicity of exposition we suppose that the deterministic component of each series can be modeled by a first-order trend polynomial, i.e., $TD_t = \kappa + \delta t$ where κ and δ are now $(n \times 1)$ vectors rather than scalars. This is the main case that has been studied in the literature.

A central concept in the analysis of a set of nonstationary variables is that of cointegration due to Granger (1981, 1983) and Granger and Weiss (1983) and discussed in more detail in Engle and Granger (1987). The idea is that even though each series may have a unit root, there may exist various linear combinations of the variables that are stationary. Stated more precisely, we have the following definition.

Definition 1: A vector of variables defined by (3.1) is said to be cointegrated if there exists at least one nonzero n-element vector β_i such that $\beta_i' y_t$ is trend-stationary. β_i is called a cointegrating vector. If there exist r such linearly independent vectors, $\beta_i (i = 1, \ldots, r)$, we say that $\{y_t\}$ is cointegrated with cointegrating rank r. We then define the $(n \times r)$ matrix of cointegrating vectors $\beta = (\beta_1, \ldots, \beta_r)$. The r elements of the vector $\beta' y_t$ are trend-stationary and β is called the cointegrating matrix.

An important fact to note about cointegrating vectors is the following:

Rule 11: The cointegrating vectors are identifiable at most up to a scale transformation. That is, if $\beta_i' y_t$ is $I(0)$, then $c\beta_i' y_t$ is also $I(0)$ for any constant $c \neq 0$.

12. The analysis could be made more comprehensive by allowing the possibility of multiple unit roots. We refrain from considering this more general case for two reasons. First, it would make the interpretation of the issues involved more difficult to convey without adding much insight. Second, the case of practical interest is that where each series is either integrated of order one or is trend-stationary.

Additionally, there are a few things to note about definition 1. First, this definition allows the linear combinations of the variables that eliminate the unit roots to have nonzero linear trends. This corresponds to the notion of "stochastic cointegration" in Ogaki and Park (1990). A stronger definition of cointegration, called "deterministic cointegration" by Ogaki and Park, would require that the same vectors β_i that eliminate the unit roots also eliminate the deterministic trends from the data. For deterministic cointegration the matrix β must be such that both $\beta' TD_t$ is a constant and also $\beta' Z_t$ is $I(0)$. When the deterministic trend is linear in time, this requires that $\beta' \delta = 0$.[13]

Second, definition 1 does not require that each of the individual series be integrated of order one; some or all series can be trend-stationary. In this respect definition 1 differs from the definition given in Engle and Granger (1987). The motivation for our more general definition is that in practice a researcher is often faced with a vector of series that can be quite different in nature incorporating some variables with $I(1)$ noise components and others with $I(0)$ noise components. Allowing the presence of trend-stationary variables has important implications. If y_t contains a trend-stationary variable it is trivially cointegrated, the cointegrating vector being the unit vector which selects the stationary variable. If all the series are trend-stationary, the system is again trivially cointegrated since any linear combination would yield a trend-stationary variable. An important point to note is the following.

Rule 12: In the case where at least one integrated variable is present, there cannot exist more than $n - 1$ linearly independent cointegrating vectors.

To see this, suppose first that there are two variables, one being $I(0)$ and the other $I(1)$. Since a nonstationary variable cannot be combined with a stationary variable to yield a stationary variable, the only cointegrating vector is the unit vector (or a scale transformation) that selects the $I(0)$ variable. Suppose now that there are two $I(1)$ variables and that the normalized linear combination $y_{1t} + ay_{2t}$ is $I(0)$. The cointegrating vector $(1\ a)$ is then unique (up to a scale transformation) since if another cointegrating vector existed, it could be combined with the first to imply that both the original variables were $I(0)$. This line of reasoning extends to systems of higher dimensions. This feature will prove of some impor-

13. An example of a system that is stochastically cointegrated but not deterministically cointegrated is a system where the individual variables are log real output levels of countries with different deterministic rates of population growth, and stationary log differences of per capita output.

tance when discussing the properties of models involving cointegrated systems.

3.1.2 Why Is Cointegration Interesting? Before turning to the detail of cointegrated systems it is useful to provide some motivation for studying them. At first sight the idea of cointegration among variables may seem to be an unlikely special case. If one has a set of integrated variables, it may seem highly restrictive to assume that some linear combination of them is stationary. In fact, however, the idea of cointegration has become extremely popular in macroeconomic analysis precisely because it arises naturally in multivariate macroeconomic models with unit root driving processes.

There are two main mechanisms that can give rise to cointegration in a macroeconomic model. To understand these, we first mention the Granger Representation Theorem that relates cointegration to the existence of an error-correction representation for the data (to be discussed more precisely below). In the error-correction representation, the stationary linear combination of the model variables Granger causes changes in at least some of the variables. As always, this Granger causality can arise either from true causality or because some variables in the model are forecasting others.

The first mechanism is one of true causality. This is emphasized in the work of the "LSE School" (Phillips, 1954; Sargan, 1964; Davidson, Hendry, Srba, and Yeo, 1978; Dávidson and Hendry, 1981; Hendry, 1986) and by Engle and Granger (1987). These authors envisage a sluggish adjustment to some long-run equilibrium described by economic theory. The short-run adjustment is a "servo-mechanism" that econometricians are free to model pragmatically. As Granger (1986, p. 213) puts it,

At the least sophisticated level of economic theory lies the belief that certain pairs of economic variables should not diverge from each other by too great an extent, at least in the long-run. Thus, such variables may drift apart in the short-run or according to seasonal factors, but if they continue to be too far apart in the long-run, then economic forces, such as a market mechanism or government intervention, will begin to bring them together again.

Any model that imposes a deterministic long-run relationship between two integrated macroeconomic variables, but that allows the variables to deviate in the short-run, will display cointegration.

The second mechanism by which cointegration can arise involves forecasting rather than true causality. As described in Campbell and Shiller (1987, 1988a), if one variable is first-order integrated and another variable

is a rational forecast of future values of the first variable, then the two variables will be cointegrated. In general, when the forecasts are based on a multivariate information set, the stationary linear combination of the two variables will Granger cause at least the variable being forecast, and possibly the variable that embodies agents' forecasts as well. The term structure of interest rates provides an example. If short-term interest rates are $I(1)$ and term premia are stationary, then long-term and short-term interest rates will be cointegrated, with Granger causality from the yield spread to changes in both short-term rates and long-term rates. This Granger causality need not reflect any causal mechanism relating short rate changes to past slopes of the term structure.

Cointegration can also arise in models with forward-looking, optimizing agents and $I(1)$ forcing variables. For example, a real business cycle model with a Cobb–Douglas production function and a random walk technology shock implies that log consumption, log investment, and log output are cointegrated (King, Plosser, and Rebelo, 1988). Here both the two mechanisms discussed above are at work.

3.1.3 Some Useful Representations We now discuss a number of alternative representations of multivariate systems with unit roots and possible cointegration. These are necessary background for the methods of testing and estimation to be discussed subsequently.

THE AUTOREGRESSIVE AND ERROR-CORRECTION REPRESENTATIONS Following the work of Sims (1980a), vector autoregressive (VAR) systems have become an increasingly popular device to model the stochastic properties of a multivariate system. It is therefore useful to analyze how, if at all, the possibility of unit roots and cointegration affects the estimation and interpretation of VAR models. In keeping with our practice of separating the trend function from the noise function, we start with the following reduced form representation:

$$y_t = \kappa + \delta t + Z_t; \qquad A(L)Z_t = e_t, \tag{3.2}$$

where κ and δ are n-element vectors of fixed coefficients. $A(L)$ is a matrix lag polynomial of order p such that $A(L) = I - A_1L - A_2L^2 - \cdots - A_pL^p$. The A_i are $(n \times n)$ matrices of fixed coefficients and the disturbances are assumed to be distributed normally with mean 0 and covariance matrix Σ. Just as in the univariate case, this vector autoregressive system of order p can be viewed as an approximation to a more general multivariate ARMA process. For the issues to be discussed, there is little loss

of generality in considering VAR models. Following Dickey and Fuller (1979) in the univariate case, we can write (3.2) as follows:

$$\Delta y_t = \mu + \Pi[y_{t-1} - \delta(t-1)] + \Sigma_{j=1}^{k}\Gamma_j \Delta y_{t-j} + e_t, \tag{3.3}$$

where $k = p - 1$, $\Pi = \Sigma_{i=1}^{p} A_i - I$, and $\Gamma_j = -\Sigma_{i=j+1}^{p} A_i$ for $j = 1, \ldots, k$. The constant vector μ is related to the drift vector δ and the other parameters of the model, as discussed further below. The matrix Π is the multivariate analog of the coefficient π in the univariate case (2.10), which measured the sum of the autoregressive coefficients less one. In the univariate case, a unit root is present if the sum of the AR coefficients is one, equivalently if $\pi = 0$. In the present multivariate case, the correspondence between unit roots and the nature of the matrix Π is not as straightforward. Many of the issues concerning cointegration can be analyzed by simply using (3.3) and searching for conditions on the nature of the variables y_t and the matrix Π such that both sides of (3.3) are stationary. Since we do not consider processes with more than one unit root, the left-hand side of (3.3) is stationary. The right-hand side will also be stationary if and only if the components of $\Pi[y_{t-1} - \delta(t-1)]$ are stationary. There are three cases of interest and they relate to the rank of the matrix Π. In the following discussion, we use the notation y_t^* to denote $y_t - \delta t$, the deviation of the series from their deterministic trends.

Consider first the case where Π is of full rank n. For all the elements of $\Pi[y_{t-1} - \delta(t-1)] = \Pi y_{t-1}^*$ to be stationary we need all n linearly independent combinations of y_{t-1}^* formed by the rows of Π to be stationary. Given rule 12 it must then be the case that all the elements of y_t are stationary around the trend vector δt. This case corresponds to the standard VAR model where no restrictions are imposed on the reduced form representation. Here standard VAR analysis applied to the level of the series is the appropriate estimation strategy. Consider now the case where the only way to make Πy_{t-1}^* stationary is to have the rank of Π be zero. This implies that $\Pi = 0$, an $(n \times n)$ matrix of zeros, and that there are no linear combinations of the variables y_t that are trend-stationary. In this case (3.3) becomes a VAR in first-differences.

The case of particular interest is when Π is neither of rank zero nor of full rank. Denote the rank of Π by r. Then there exist $(n \times r)$ matrices α and β such that

$$\Pi = \alpha\beta'. \tag{3.4}$$

To have Πy_{t-1}^* stationary, it must be the case that $\beta' y_{t-1}^*$ is stationary. Hence β is the matrix whose columns are the linearly independent

cointegrating vectors and the rank of the matrix Π corresponds to the cointegrating rank of the system of variables y_t (using definition 1). In the stronger case of deterministic cointegration, where we require that Πy_{t-1} is stationary, $\beta' y_{t-1}$ must be stationary so we require $\beta'\delta = 0$. In analogy with rule 11, which applied to a single cointegrating vector, we have the following important fact.

Rule 13: The parameters of α and β are not identified since for any nonsingular matrix F the matrices αF and $\beta(F')^{-1}$ yield the same matrix Π. This implies that the data can give information only about the space spanned by the columns of β (the cointegrating space) and the space spanned by α.

One solution to this identification problem is to normalize one element of each column of β, for example by imposing a unit coefficient on one variable in each equation. Johansen (1989a) points out, however, that this is valid only if we have a priori knowledge that the coefficient associated with that variable is nonzero. In practice this may be unappealing. Fortunately, it is often unnecessary to separately identify the parameters of α and β; but care must be taken to avoid testing hypotheses about these parameters when they have not been identified by normalization or other prior restrictions (Park, 1990b).

Before describing how one can interpret the matrix α, it is useful to give an interpretation of the vector $\beta' y_t^*$. Each column of the matrix β can be viewed as describing some long-run relationship between the integrated series in the vector y_t^*. The fact that we require $\beta' y_t^*$ only to be stationary (as opposed to constant or white noise) means that allowance is made for the possibility of serially correlated but temporary divergences from this relationship. Accordingly, the elements of $\beta' y_t^*$ are sometimes called "equilibrium errors," the equilibrium relation being described by the cointegrating vectors. Of course, as discussed in Section 3.1.2, this use of the term equilibrium should not be confused with the common use in macroeconomics to refer to the outcome of a market-clearing economic model. Equilibrium errors can arise in the most classical of macroeconomic models, as discussed by King, Plosser, and Rebelo (1988).

We now introduce the notation $z_{t-1} = \beta' y_{t-1}^* = \beta'[y_{t-1} - \delta(t-1)]$, and rewrite (3.3) as

$$\Delta y_t = \mu + \alpha z_{t-1} + \Sigma_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t. \tag{3.5}$$

Equation (3.5) describes what we referred to in Section 3.1.2 as an error-correction model. It specifies that the change in y_t depends not only on

the lagged values of these changes but also on the equilibrium error that occurred in the previous period. Viewed in this error correction framework the matrix α can be interpreted as a measure of the speed by which the system corrects last period's equilibrium error. The matrix α is often called the adjustment matrix, although of course it need not arise from costs of adjustment.

The error-correction model (3.5) implies that when both unit roots and cointegration are present, an unrestricted VAR in the first-differences of y_t is misspecified because the lagged equilibrium errors z_{t-1} are omitted as regressors. An unrestricted VAR in the levels of y_t is not misspecified but involves a loss of efficiency because some restrictions are omitted, namely the reduced rank of the matrix Π in (3.3). We note also that it is possible to rewrite the error-correction model (3.5) as a VAR, not in first differences or levels of the original series, but in the levels of r equilibrium errors z_t and the differences of $n - r$ of the original series. This representation is used in Campbell and Shiller (1987, 1988b). We summarize the major points discussed so far in the following rule.

Rule 14: In the general VAR process (3.3) three cases are possible: (1) rank(Π) = n in which case all variables are trend-stationary and the application of an unrestricted VAR in levels is appropriate; (2) rank(Π) = 0 (or Π = 0), in which case no cointegrating relation exists and the application of an unrestricted VAR in first-differences is appropriate; and (3) $0 <$ rank(Π) = $r < n$, in which case at least one integrated variable and one cointegrating relation are present. In the latter case an unrestricted VAR (either in levels or in differences) is inappropriate but the data can be described by an error-correction model of the form (3.5) or by a VAR in r stationary combinations and $n - r$ differences of the original variables.

One final point about the error-correction representation is emphasized by Johansen (1989a,b). In general the error-correction model (3.5) contains an unrestricted constant vector μ. When the series y_t are not trending ($\delta = 0$), however, one can show that $\mu = -\Pi\kappa = -\alpha\beta'\kappa$, where κ is the vector of intercepts in (3.2). In this case (3.5) can be rewritten as

$$\Delta y_t = \alpha(\beta' y_{t-1} - \beta'\kappa) + \Sigma_{j=1}^{k}\Gamma_j \Delta y_{t-j} + e_t. \tag{3.6}$$

In the representation (3.6), constant terms appear only because the equilibrium errors have nonzero means $\beta'\kappa$. These need to be subtracted from the equilibrium errors on the right-hand side of (3.6) in order to satisfy the condition that the unconditional mean of Δy_t is zero when

there are no trends in y_t. Johansen (1989a,b) emphasizes that the restriction in (3.6), as compared with (3.5), can affect estimation and testing.

THE MOVING AVERAGE REPRESENTATION It is useful at this point to consider how cointegration restricts the moving-average representation of the first-differences of the data. By analogy with the earlier univariate analysis, we use the notation $\Psi(L)$ to denote the moving average representation of the first difference of the multivariate noise process Z_t: $\Delta Z_t = \Psi(L)e_t$. In general $\Psi(L)$ is a matrix polynomial in L of infinite order and can be written in the form $\Psi(L) = \Sigma_{j=0}^{\infty}\Psi_j L^j$, where $\Psi_0 = I$. The first difference of the data can be written as

$$(1 - L)y_t = \delta + \Psi(L)e_t. \tag{3.7}$$

A multivariate version of the Beveridge–Nelson decomposition (2.4) enables us to rewrite (3.7) as

$$y_t = \mu^* + \delta t + \Psi(1)S_t + \Psi^*(L)e_t, \tag{3.8}$$

where as before S_t is defined by $S_t = \Sigma_{j=1}^{t}e_t$ (this is now a vector of n random walks), and $\Psi^*(L)$ is defined to equal $(1 - L)^{-1}[\Psi(L) - \Psi(1)]$. The vector μ^* is a vector of constants. Multiplying both sides of (3.8) by the matrix β' and using the definition of z_t, the vector of equilibrium errors, we have

$$z_t = \beta'\mu^* + \beta'\Psi(1)S_t + \beta'\Psi^*(L)e_t. \tag{3.9}$$

Given that the left-hand side of (3.9) is stationary we require its right-hand side to be likewise. The first term is a constant and the third term is stationary given the properties of $\Psi^*(L)$, but the second term involves the random walk component S_t. Therefore we need $\beta'\Psi(1) = 0$. Since β is an $(n \times r)$ matrix, this implies that the rank of $\Psi(1)$ is $n - r$. It is also easy to verify (see Engle and Granger, 1987) that $\Psi(1)\alpha = 0$, expressing a relationship between the matrix sum of the moving-average coefficients and the adjustment matrix. These results, which we summarize in the following rule, are helpful in discussing the nature of some testing procedures.

Rule 15: In the moving-average representation of the first-differences of the data (3.7), the presence of a cointegrating relation among the components of the vector y_t implies that $\Psi(1)$, the matrix of the sum of the moving average coefficients, is singular. If there are r linearly indepen-

dent cointegrating vectors, $\Psi(1)$ is of rank $n - r$ and the following relation holds: $\beta'\Psi(1) = 0$ and $\Psi(1)\alpha = 0$.

THE LONG-RUN COVARIANCE MATRIX Consider again the first-differenced representation of the data given by (3.7). The existence of cointegrating relationships implies restrictions on the long-run covariance matrix of this vector of first-differences. To be more precise, we mean by the long-run covariance matrix the spectral density matrix evaluated at frequency zero which, following the notation used for the univariate case, we denote as $h_{\Delta y}(0)$. This long-run covariance matrix is related to Σ, the covariance matrix of the vector e_t, and to $\Psi(1)$, the matrix sum of the moving-average coefficients, as follows:

$$h_{\Delta y}(0) = \Psi(1)\Sigma\Psi(1)'. \tag{3.10}$$

We stated this relationship for the scalar case when discussing the Beveridge–Nelson decomposition in Section 2. Since $\beta'\Psi(1) = 0$, as discussed in the last subsection, we have $\beta' h_{\Delta y}(0) = 0$, which means that the rank of the matrix $h_{\Delta y}(0)$ is $n - r$. This fact has been used by Phillips and Ouliaris (1988, 1990) to derive testable implications associated with cointegration. We summarize the facts of practical interest as follows.

Rule 16: The long-run covariance matrix of the first-differenced data, $h_{\Delta y}(0)$, will be of full rank only if no cointegrating relation exists among the variables y_t. If there exist r cointegrating relations $h_{\Delta y}(0)$ is singular and has rank $n - r$.

This singularity of the long-run covariance matrix when cointegration is present corresponds to the fact that the first difference of a stationary univariate series has zero long-run variance. In the multivariate case considered here, the long-run covariance matrix will also be singular if one of the series is stationary since this implies the presence of a cointegrating relation (though a trivial one). The relation between the univariate and multivariate case is probably best understood by noting that if all the series are stationary then $h_{\Delta y}(0)$ is the null matrix. This holds because in this case there are n cointegrating vectors and hence the rank of $h_{\Delta y}(0)$ is zero.

THE COMMON TREND REPRESENTATION An interesting specification put forward and used to derive test statistics by Stock and Watson (1988a) is the so-called common trend representation. It highlights the fact that if r cointegrating relations exist among a set of n variables [all of which could

be $I(1)$], then the variables can be written as linear combinations of n linear trends, $n - r$ pure random walks, and n stationary random terms. To see this, recall that Equation (3.8) gives the stochastic trend in the vector y_t as $\Psi(1)S_t$, where S_t is a vector of n random walks. Stock and Watson show that since $\Psi(1)$ has rank $n - r$, the stochastic trend in (3.8) can be rewritten in terms of $n - r$ random walks. This is the common trend representation.

In the stronger case of deterministic cointegration, we have the additional restriction that $\beta'\delta = 0$, where δ is the vector of linear trends in the series y_t. Since also $\beta'\Psi(1) = 0$, δ lies in the column space of $\Psi(1)$ and can be written as a linear combination of $\Psi(1)$, i.e., $\delta = \Psi(1)\delta^*$. (3.8) now becomes

$$y_t = \mu^* + \Psi(1)S_t^* + \Psi^*(L)e_t, \tag{3.11}$$

where $S_t^* = \delta^* t + \Sigma_{j=1}^{t} e_t$ is a vector of n random walks with drift δ^*. Again, Stock and Watson show that this can be reduced to a representation with $n - r$ random walks. When the original series y_t are trending, then the common random walks have nonzero drifts; they have zero drifts when the original series are not trending ($\delta = 0$). We summarize the common trend representation in the following.

Rule 17: A multivariate system with r cointegrating vectors represented by the matrix β with corresponding adjustment matrix α can be written as n trends, plus $n - r$ random walks ("common trends"), plus n stationary components. When the cointegration is deterministic ($\beta'\delta = 0$), the system can be represented as $n - r$ random walks, plus n stationary components. These random walks generally have nonzero drifts, but they have zero drifts when the series y_t are not trending ($\delta = 0$).

3.2 TESTING FOR COINTEGRATION

In this section we review some of the statistical procedures that can be used to test hypotheses about cointegration. We show how each procedure uses one of the different representations outlined in the previous section. In Sections 3.2.1 and 3.2.2 we discuss procedures that are designed to distinguish a system without cointegration from a system with at least one cointegrating relationship. These procedures do not try to estimate the number of cointegrating vectors. Then in Section 3.2.3 we consider procedures that test for a particular number of cointegrating relationships. A last subsection briefly discusses issues related to the size and power of the tests. In testing for cointegration, just as in testing for a

unit root in a univariate context, the treatment of deterministic regressors is important. Throughout our discussion we emphasize this point.

3.2.1 Tests Based on a Static Regression THE FRAMEWORK OF THE STATIC REGRESSION Tests based on a static regression are probably the most popular class of tests for cointegration. These tests give a special role to one of the variables in a vector y_t, which is chosen to be the dependent variable in the regression. The tests try to distinguish the hypothesis that there is no cointegration between the dependent variable and the regressors, from the hypothesis that there is at least one cointegrating relationship between the dependent variable and the regressors. The procedures presume that any $I(0)$ variables in the original vector y_t have been removed from consideration before the procedure begins, so that all remaining variables are $I(1)$. Hence, each of the individual variables must pass a unit root test before being included in the static regression. The cointegration tests are conditional on this pretesting procedure. In principle, their critical values should be adjusted to account for pretesting, but no theory is currently available that would allow us to do this.

We partition the variables in y_t as (y_{1t}, y_{2t}), where y_{1t} is a scalar $I(1)$ variable and y_{2t} is an m-element vector of $I(1)$ variables. We also assume that if there exists a cointegrating relationship, the variable y_{1t} has a nonzero coefficient in the cointegrating vector. Since cointegrating vectors are identifiable only up to scale, we may without loss of generality set this coefficient to 1. The hypothesis of cointegration then asserts that there exists an m-element vector of coefficients θ such that $y_{1t} - \theta' y_{2t}$ is $I(0)$, i.e., the cointegrating vector is $\beta' = (1 \; -\theta')$. Assume for the moment that it is known a priori that the mean of the linear combination $y_{1t} - \theta' y_{2t}$ is 0 and consider the following regression equation:

$$y_{1t} = \theta' y_{2t} + u_t. \tag{3.12}$$

The hypothesis that y_{1t} and y_{2t} are not cointegrated can now be stated as the hypothesis that there does not exist any vector of coefficients θ such that $u_t = y_{1t} - \theta' y_{2t}$ is $I(0)$. The hypothesis that y_{1t} and y_{2t} are cointegrated is that a vector of coefficients can be found such that u_t is $I(0)$. Note that even if no such vector of coefficients exists, there still could be cointegration among the variables y_{2t} on the right-hand side of (3.12).

RESIDUAL-BASED TESTS In (3.12), a test of the null hypothesis versus the alternative hypothesis amounts to a unit root test on the equation errors. Since we do not observe the errors u_t, we must use some estimates of

their values. A straightforward approach is to apply OLS to (3.12) and conduct a unit root test on the estimated residuals, $\hat{u}_t$, as a proxy for the true residuals.

In principle, any test for a unit root versus stationarity applicable in a univariate context can be used as a test for no cointegration versus cointegration when applied to the series $\hat{u}_t$. These include the augmented Dickey–Fuller method, the tests proposed by Phillips and Perron (1988) or the tests suggested by Stock (1990) that were discussed in Section 2. Just as in the univariate analysis, it is important to include a constant in the static regression if the alternative hypothesis of cointegration allows a nonzero mean for $\beta' y_t$, and a trend in the static regression if the alternative hypothesis is "stochastic cointegration," allowing a nonzero trend for $\beta' y_t$. The critical values of the unit root tests on $\hat{u}_t$ depend on whether a constant and/or a time trend are included in the static regression. In addition it is important to note that the critical values are not the same as those for unit root tests applied to raw data. They depend on the number of integrated regressors in (3.12) and whether or not these regressors are trending. We give some intuition for this when we discuss "spurious regression" below. These points are summarized in the following rule.

Rule 18: When unit root tests are applied to the estimated residuals from regression (3.12), their asymptotic distribution under the null hypothesis depends on whether a constant and/or a time trend are included in (3.12). In addition they depend on the number of integrated regressors, that is the dimension of the vector y_{2t}, and the nature of the deterministic trends in y_{2t}. The asymptotic distribution is never identical to the case where the unit root tests are applied to raw data.

Phillips and Ouliaris (1990) and Engle and Yoo (1987) have tabulated critical values for the augmented Dickey–Fuller t test and the Phillips–Perron $Z(\hat{\pi})$ and $Z(t_{\hat{\pi}})$ tests applied to residuals from (3.12), where the dimension of the vector y_{2t} ranges from 1 to 5. In the case where the regressors may contain deterministic trends, care must be exercised in using these tabulated critical values. This issue has been investigated by Hansen (1990b), and we summarize his results as follows. If the static regression includes a time trend, parts (c) of Phillips and Ouliaris' tables are appropriate and the tests are for stochastic cointegration. If the static regression includes only a constant and the regressors do not contain deterministic trends, parts (b) of Phillips and Ouliaris' tables are appropriate. Finally consider the case where the static regression includes only a constant but the regressors contain deterministic trends so that one is

testing for deterministic cointegration among trending variables. Here parts (c) of Phillips and Ouliaris' tables are appropriate but one should use the critical values for $m - 1$ regressors instead of m regressors.[14]

A DIGRESSION ON SPURIOUS REGRESSION If there is no cointegration between the dependent variable and the regressors of (3.12), then the regression is described as "spurious." This term was introduced by Granger and Newbold (1974) who gave special attention to the case where a random walk (y_{1t}) is regressed on an independent random walk (y_{2t}). Since these variables are not cointegrated, the residuals are $I(1)$ and this results in a low value of the Durbin–Watson statistic. Granger and Newbold (1974) also documented the fact that in such a regression the R^2 statistic will typically be high giving the impression of a good fit.

This phenomenon has been theoretically investigated by Phillips (1986) who shows that in a spurious regression the following results hold as the sample size increases to infinity: (1) the Durbin–Watson statistic converges to 0, (2) the R^2 of the regression converges to a random variable, and (3) the t-statistics on the coefficients of the vector θ diverge. To understand these results intuitively, consider two independent random walks. If one could observe an increasing number of samples drawn from these stochastic processes with a fixed initial condition, then the correlation of the samples would be zero. But instead one observes a single sample of increasing length from each process. In this single sample random shocks have effects that never die out, so the regression coefficient of one time series on the other and the R^2 converge to random variables rather than to zero.

The characteristics of spurious regression help one to understand why unit root tests on static regression residuals have critical values which depend on the number of regressors (rule 18). Under the null hypothesis of no cointegration, the static regression contains m spurious regressors whose coefficients have random limits. This affects the asymptotic distribution of the static regression residuals, and hence the null distribution of the unit root test statistics.

Granger and Newbold (1974) proposed a rule of thumb for detecting spurious regression. It states that, in an estimated regression, an R^2 higher than the Durbin–Watson statistic should be viewed as a warning of a spurious relationship (see Hendry, 1980, for an interesting illustration). In light of the theoretical apparatus that is now available, this rule

14. In an interesting recent contribution, Hansen (1990a) has suggested working with the estimated residuals from a Cochrane–Orcutt version of the static regression that allows for AR(1) errors. Hansen's test statistics have the advantage that their limiting distributions are always the same as those of univariate unit root tests.

of thumb should be viewed as a preliminary sign that a more formal cointegration analysis is warranted. It also highlights the importance of residual diagnostic statistics such as the Durbin–Watson as indicators of possible misspecification. Nevertheless, though the Durbin–Watson statistic is a useful preliminary indicator it cannot be used as the basis of a formal test statistic for cointegration as suggested by Engle and Granger (1987). The reason is that under the null hypothesis of no cointegration, the asymptotic distribution of the Durbin–Watson statistic depends on nuisance parameters such as the correlations among the first differences of the variables included in the regression. This important practical result is stated in the following rule.

Rule 19: The Durbin–Watson statistic should not be used as the basis of a test of the null hypothesis of no cointegration versus the alternative hypothesis of cointegration.

TESTS BASED ON THE SIGNIFICANCE OF SPURIOUS REGRESSORS The idea of spurious regression has been used in several recent papers by Park and Choi (1988), Park, Ouliaris, and Choi (1988), and Park (1990a) to derive tests for cointegration. These tests have the advantage that they can be formulated either with a null of no cointegration or with a null of cointegration. Consider the following version of (3.12) with added regressors:

$$y_{1t} = \tau' DR_t + \theta' y_{2t} + \varphi_1' s_{1t} + \varphi_2' s_{2t} + u_t. \tag{3.13}$$

Here DR_t is a vector of deterministic regressors that capture the deterministic trend of the variables y_{1t} and y_{2t}. s_{1t} is a vector of q nonstationary deterministic functions that are of a higher order than the variables in DR_t. For example, if (y_{1t}, y_{2t}) are characterized by a first-order linear trend, $DR_t = \{1,t\}$ and s_{1t} could include the regressors $\{t^2, t^3, \ldots, t^{q+1}\}$. The vector s_{2t} is a p-element vector containing variables that are integrated of order one. The specific assumptions needed about these regressors depend on whether the null hypothesis being tested is that of no cointegration or that of cointegration.

Consider first the case where the null hypothesis is the familiar one of no cointegration. Here we assume that $\Delta s_{2t} = w_t$, where w_t is asymptotically uncorrelated with the errors u_t of regression (3.13). An obvious choice for such a regressor is a computer generated random walk. Denote by $F(\varphi_1,\varphi_2)$ the Wald statistic for the joint hypothesis that $\varphi_1 = 0$ and $\varphi_2 = 0$. Park et al. show that the Wald statistic normalized by the sample size, $F(\varphi_1,\varphi_2)/T$, has a nondegenerate asymptotic distribution under the null hypothesis, but converges to zero under the alternative hypothesis;

thus it is a consistent test of the null against a cointegrated alternative. The critical values depend on (1) the number of nonstationary regressors in y_{2t} (m), (2) the number of nonstationary regressors in s_{2t} (p), and (3) the number and specific form of the nonstationary deterministic regressors included in the vector s_{1t}. Asymptotic critical values can be found in the papers mentioned above.

Consider now testing the null hypothesis of cointegration versus the alternative hypothesis of no cointegration. The same regression Equation (3.13) can be used if it is assumed that the superfluous stochastic regressors included in the vector s_{2t} are not cointegrated with the variables in y_{2t}. The same Wald statistic for testing the joint hypothesis $\varphi_1 = 0$ and $\varphi_2 = 0$ can be used as the basis of the test. Indeed, it can be shown that $F(\varphi_1, \varphi_2)$ (not divided by T) has a nondegenerate asymptotic distribution under the hypothesis of cointegration but diverges to infinity under the hypothesis of no cointegration. Unfortunately the limiting distribution under cointegration depends on nuisance parameters involving the serial correlation of the variables, but Park (1990a) shows how to transform the Wald test in such a way that its limiting distribution converges to a χ^2 with $(p + q)$ degrees of freedom.

The intuition behind these tests is that under the null hypothesis of no cointegration neither the variables of interest (y_{2t}) nor the superfluous regressors (s_{1t} and s_{2t}) have a long-run relation with the dependent variable y_{1t}. Hence the regression is spurious and from our earlier discussion the t or F statistics associated with all the regressors, including the superfluous ones, will diverge. Under the hypothesis that the dependent variable y_{1t} is cointegrated with the variables y_{2t}, the regression is not spurious, and the t or F statistics for the superfluous regressors will asymptotically reflect their insignificance as in a standard regression framework. However, Park's corrections are necessary to account for the presence of possible correlation between the errors and the regressors.

3.2.2 Cointegration Tests on the Long-Run Covariance Matrix We now describe some tests that exploit the fact that any cointegrating relationship among the variables y_t implies a singular long-run covariance matrix $h_{\Delta y}(0)$, as defined in Section 3.1.3. These tests still do not attempt to estimate the number of cointegrating relationships, but some tests of this type avoid the assumption of the previous section that a particular variable y_{1t} has a nonzero coefficient in the cointegrating vector.

For notational convenience we partition $h_{\Delta y}(0)$ as follows in accordance with the partition on y_t such that $y_t = (y_{1t}\ y_{2t})$ with y_{1t} a scalar and y_{2t} an m vector:

$$h_{\Delta y}(0) = \Omega = \begin{Bmatrix} \omega_{11} & \omega_{21}' \\ \omega_{21} & \Omega_{22} \end{Bmatrix}, \tag{3.14}$$

and define $\omega_{11 \cdot 2} \equiv \omega_{11} - \omega_{21}'\Omega_{22}^{-1}\omega_{21}$, a measure of the long-run variance of Δy_{1t}, conditional on the elements of Δy_{2t}. (If the elements of Ω were variances and covariances rather than spectral densities at frequency zero, then $\omega_{11 \cdot 2}$ would be the variance of the error in a regression of Δy_{1t} on Δy_{2t}). We have that $\omega_{11 \cdot 2} = 0$ if y_{1t} and y_{2t} are cointegrated. Letting det(·) denote the determinant, it can be shown that

$$\det[h_{\Delta y}(0)] = \omega_{11 \cdot 2} \det(\Omega_{22}). \tag{3.15}$$

When there is no cointegration among the elements of y_{2t}, $\det(\Omega_{22}) \neq 0$. In this case $h_{\Delta y}(0)$ is singular if and only if $\omega_{11 \cdot 2} = 0$. Phillips and Ouliaris (1990) use this framework to propose what they call the "variance ratio test." It is an appropriately scaled version of an estimate $\hat{\omega}_{11 \cdot 2}$ of the quantity $\omega_{11 \cdot 2}$. This test statistic by itself does not offer any generalization over the statistics discussed earlier, but the same framework can be used to construct statistics that test the null of no cointegration against the alternative of at least one cointegrating vector without specifying a priori any specific element of y_t as having a nonzero coefficient. The idea is to directly test whether $h_{\Delta y}(0) \equiv \Omega$ is singular. Phillips and Ouliaris propose a "multivariate trace statistic" for this purpose.

In constructing these tests, constants and/or linear trends should be included in the system if they appear in the data-generating process. As always the presence of these deterministic regressors affects the critical values for the tests. In addition, the critical values depend on the number of variables in the system. Phillips and Ouliaris present tabulated critical values for any number of regressors in y_{2t} between 1 and 5. Note also that the estimates of the long-run covariance matrix in these procedures must be based on the residuals from a regression of y_t on y_{t-1}. Under the null hypothesis of no cointegration this is equivalent to using the first-differences of the data since the least-squares estimator of the matrix of coefficients converges to the identity matrix. But under the alternative hypothesis of cointegration this equivalence breaks down, and the regression residuals must be used to ensure the consistency of the test statistic.

3.2.3 Tests for Cointegrating Rank We now consider testing procedures that allow the investigator to estimate the number of cointegrating vectors. We start with a procedure suggested by Johansen (1988, 1989a), and

then describe the approach of Stock and Watson (1988a). Both procedures allow one to specify as the null hypothesis an arbitrary number of cointegrating vectors provided that this number is less than the number of cointegrating vectors allowed under the alternative hypothesis. For example, in a three-variable system one can test the null hypothesis of one cointegrating vector against the alternative hypothesis that there are two or three cointegrating vectors, the latter corresponding to the hypothesis that all series are trend-stationary. These procedures also do not impose any prior assumption that some or all of the series investigated are $I(1)$. Therefore they contain as a special case a univariate test for a unit root versus stationary alternatives.

Johansen's procedure applies maximum likelihood to the autoregressive model discussed in Section 3.1.3, assuming that the errors are Gaussian. More specifically the estimated model is given by

$$\Delta y_t = \mu + \Pi y_{t-1} + \Sigma_{j=1}^{k} \Gamma_j \Delta y_{t-j} + e_t, \tag{3.16}$$

where $e_t \sim N(0, \Sigma)$. This is the autoregressive representation (3.3) with the added assumption that cointegration is deterministic, i.e., $\Pi\delta = 0$, or that the series are driftless, i.e., $\delta = 0$. This assumption allows the system to contain $I(1)$ or stationary variables (or linear combinations of variables), but it rules out trend-stationary variables (or linear combinations of variables) with nonzero trends. Johansen's method tries to estimate the rank of the matrix Π, the matrix of coefficients on the lagged levels in (3.16). Recall that when there are r cointegrating relationships, the rank of Π is equal to r.

We first consider how to construct a likelihood ratio test of the null hypothesis of r cointegrating vectors versus the alternative hypothesis of n cointegrating vectors. Under the alternative hypothesis Π is unrestricted and the maximum likelihood estimates of the coefficients of (3.16) are obtained by OLS. Under the null hypothesis the matrix Π is restricted by the relationship $\Pi = \alpha\beta'$. Maximum likelihood estimates of the matrices Γ_i ($i = 2, \ldots, k$) and the vector μ can again be obtained by OLS. However, the maximum likelihood estimates of α, β, and Σ are different and are obtained by solving an eigenvalue problem. Here we simply sketch the basic idea of the procedure, as details are given in Johansen (1988, 1989a). If the matrix β were known, then α could be obtained by first regressing Δy_t and y_{t-1} on lagged changes Δy_{t-j}, then taking the residuals (indicated by tildes) and regressing $\Delta\tilde{y}_t$ on $\beta'\tilde{y}_{t-1}$. Maximum likelihood estimates of β are obtained by minimizing the determinant of the covariance matrix of the residuals of this second stage regression; this is equivalent to choosing r eigenvectors corresponding to

the r largest eigenvalues of a particular matrix. The maximized likelihood function is multiplicative in these eigenvalues, so that a likelihood ratio test statistic can be computed from them. Johansen calls this the "trace statistic." Its asymptotic distribution is not given by standard theory, because under the null hypothesis the calculated eigenvalues correspond to $n - r$ nonstationary common trends rather than stationary linear combinations of the data. The asymptotic distribution depends on $n - r$, and is tabulated in Johansen (1989b) and Johansen and Juselius (1990) for values of $n - r$ between 1 and 5. Johansen also considers a likelihood ratio statistic for the null hypothesis of r cointegrating vectors versus the alternative of $r + 1$ cointegrating vectors. This is called the "maximum eigenvalue statistic." The asymptotic critical values again are nonstandard, depend on the number of nonstationary components $(n - r)$, and have been tabulated for $n - r$ ranging from 1 to 5.

In applying the Johansen approach it is essential to handle deterministic trends with some care, because the critical values of the test depend on the trend characteristics of the data. Note first that the system (3.16) excludes time as a regressor so it can be used only to test for deterministic cointegration. Furthermore, the Johansen test calculates eigenvalues that under the null hypothesis are associated with common trends rather than stationary linear combinations of the data. To understand the importance of this, recall that an omitted nonzero trend in a univariate unit root test causes the t-statistic on the lagged level variable to have a standard normal distribution rather than a Dickey–Fuller distribution (rule 3). Analogously, the critical values of the Johansen test depend on whether the omitted drifts in the common trends are zero or not. If they are zero, this implies non-trending data which in turn implies the restriction $\mu = -\alpha\beta'\kappa$ as shown in Equation (3.6). In summary, there are three possibilities: (1) trending data ($\delta \neq 0$), (2) nontrending data ($\delta = 0$, $\mu = -\alpha\beta'\kappa$) with the statistic calculated using an autoregression with an unrestricted intercept term, and (3) nontrending data with the statistic calculated using an autoregression imposing the restriction $\mu = -\alpha\beta'\kappa$. A conservative procedure for determining the cointegrating rank is to use the maximal critical values over all the cases. A sequential procedure to estimate both r and the presence or absence of trends is also possible (Johansen, 1991). The Johansen approach could be generalized to allow stochastic cointegration in the null hypothesis by including linear trends in (3.16), but no tables of critical values have yet been published.

An alternative test for cointegrating rank has been proposed by Stock and Watson (1988a). Their approach is implicitly based on trying to estimate the rank of the matrix $\Psi(1)$, the sum of the matrix coefficients in the moving-average representation of the series in first differences. As

we discussed in Section 3.1.3, the rank of $\Psi(1)$ is equal to $n - r$ under the hypothesis that there are r cointegrating vectors. This is also the number of common trends in the representation discussed in Section 3.1.3. Hence estimating the rank of the matrix $\Psi(1)$ is equivalent to estimating the number of common trends present in the set of variables under study. Stock and Watson estimate the common trends under the null hypothesis that $n - r$ of them are present and test whether these constructed series show further evidence of cointegration. To estimate the common trends, Stock and Watson suggest using principal components analysis. The idea is to choose the $n - r$ linear combinations of y_t with the largest variance as the estimated common trends. The intuition for this procedure is that the common trends are $I(1)$ so their variance increases with the sample size and will asymptotically dominate the variance of the stationary linear combinations of the data.

Once the vector of common trends has been estimated, Stock and Watson regress the vector on its own first lag to obtain a coefficient matrix P. The limiting distribution of the coefficient estimates depends on the serial correlation of the data. Stock and Watson consider both a nonparametric correction similar to that used by Phillips and Perron (1988) and a vector autoregressive parametric correction similar to that used by Dickey and Fuller (1979).[15] To test the hypothesis of $n - r$ common trends versus m common trends, Stock and Watson calculate the $m + 1$ smallest eigenvalue of the corrected least squares estimator $\hat{P}_c$ and test whether its real part is different from one. Stock and Watson have tabulated the asymptotic critical values of the normalized eigenvalue under the null hypothesis. These critical values depend on two parameters, the number of common trends under the null hypothesis, $n - r$, and the number of common trends under the alternative hypothesis, m. Critical values are tabulated for $n - r$ and $n - m$ ranging from 1 to 6. Note that a test of the null hypothesis of no-cointegration versus the alternative of cointegration can be obtained by specifying $r = 0$ and $m = n - 1$. Similarly a univariate test for a unit root can be obtained by specifying $n = r = 1$ (in which case there is no need to estimate the vector of common trends since it is the univariate series itself) and $m = 0$. In the latter case the nonparametrically corrected statistic reduces to that proposed by Phillips and Perron (1988). The Stock–Watson approach allows the investigator to remove constants and/or linear trends from the data. As always, this affects the asymptotic critical values and should be done whenever constants and/or trends are present in the

15. They recommend the parametrically corrected version for reasons of size robustness similar to the univariate case.

data-generating process. The way the detrending is performed depends on the hypothesis being tested. For testing stochastic cointegration, the original data should be detrended, while to test for deterministic cointegration, the estimated common trends should be detrended.[16]

3.2.4 Some Comments about Size and Power It is useful, in closing this section, to return to the near-observational equivalence issue discussed in Section 2.3. Recall that in the univariate case we argued that the trend-stationary and integrated processes are classes for which each member of one class can be arbitrarily well approximated, in finite samples, by a member of the other class. The same principle applies when comparing the classes of cointegrated and noncointegrated processes. Every cointegrated process can be arbitrarily well approximated, in finite samples, by a noncointegrated process and vice versa. To make this point more precise, consider a set of variables that displays no cointegrating relationship. Then the matrix of the sum of the autoregressive coefficients Π has rank zero. Any such process can be arbitrarily well approximated by a cointegrated process for which the matrix Π has "nearly zero rank." Conversely, consider a process that is cointegrated in which case the matrix of the sum of the moving-average coefficients, $\Psi(1)$, is singular. Any such process can be arbitrarily well approximated by a system with no cointegrating relationship with a matrix $\Psi(1)$ "nearly singular."

In the case of multiple time series, this argument can be taken one step further. A system with r cointegrating vectors can be arbitrarily well approximated, in finite samples, by a system with *any* number of cointegrating vectors. Consider approximating a process with r cointegrating vectors by a process with a smaller number of cointegrating vectors, say m ($m < r$). With r cointegrating vectors, the matrix $\Psi(1)$ has rank $n - r$. In finite samples there will exist a process with $\Psi(1)$ of rank $n - m$ that is a close approximation if its matrix $\Psi(1)$ is nearly of rank $n - r$. Conversely, a process with r cointegrating vectors can be well approximated by a process with a greater number of cointegrating vectors, say $m^* > r$. The Π matrix for such a process has rank m^* but is "nearly of rank r."

This discussion suggests that it may be difficult to distinguish processes that exhibit cointegration from those that do not, and more so to estimate precisely the exact number of cointegrating relationships. If in

16. There is as yet little work comprehensively comparing the finite sample behavior of alternative tests for cointegration. An exception is Gregory (1991), who analyzes a wide variety of procedures within a linear-quadratic model. He recommends the Dickey–Fuller ADF and Phillips $Z(\pi)$ tests applied to static regression residuals as analyzed by Phillips and Ouliaris (1990).

fact the goal of cointegration tests is to uncover the "true relation among the variables," these issues are disconcerting. As in the univariate case, our response is a pragmatic one. In many applications the goal is not to uncover the "true number" of cointegrating relationships per se but rather to have a useful guide in imposing restrictions that may lead to more efficient estimation. Viewed in this light, it is inconsequential if we label a process that really has m cointegrating vectors but is "nearly" cointegrated of order r as one having r cointegrating vectors. The testing procedures described can play a useful role in identifying these (possibly approximate) restrictions and may permit more precise estimates of the coefficients governing short-run dynamics.

3.3 ESTIMATION OF MULTIVARIATE MODELS WITH COINTEGRATION

We now consider methods for optimal estimation of the parameters in a multivariate model with cointegration. First recall that in the error-correction representation (3.5), the changes in a given variable are functions of lagged changes in all the variables and the r lagged equilibrium errors z_{t-1}. Suppose first that both the number of cointegrating vectors r and the true coefficients in the matrix β are known a priori. The equilibrium errors z_t are then known quantities (when the drifts δ are either known or zero), and optimal estimation of (3.5) is simply OLS applied equation by equation. In practice, however, both r and β are unknown quantities. Methods for estimating r, the number of cointegrating relationships, were discussed in the last subsection. In all the methods that we shall discuss concerning the estimation of the other parameters, it is important to note that the value r is chosen as the estimated value from one of these procedures. Methods for estimating β and the other parameters of the model treat the estimated value of r as if it were the true value, so inference is conditioned on that value of r. This point is summarized in the following.

Rule 20: In all the procedures that follow the number of cointegrating vectors used to specify the estimation procedure is treated as known even though, in practice, it is obtained from a testing procedure and hence is a random variable. In principle this pretesting affects the appropriate distribution theory underlying the estimates, even asymptotically.

It is not known whether this pretesting problem is of any practical importance. We believe that it is not likely to be serious. In any event, no theoretical results are currently available that might shed light on this problem and we shall accordingly ignore it. Conditioning on the number

of cointegrating vectors amounts to imposing some long-run restrictions. In a sense even if the true value of r is different from the estimated value, it may still be the case that the restrictions imposed allow estimates with greater precision. Once one conditions on the number of cointegrating vectors there remains the issue of estimating (1) the cointegrating vector β, and (2) the coefficients related to the short-run dynamics (α, μ, and the Γ_j). Two main approaches are available in the literature: (1) estimating the cointegrating vector and the short-run dynamics jointly, and (2) estimating them separately. Within each approach there is also the issue of estimating the system as a whole versus estimating the system equation by equation. It turns out that each of these methods yields consistent estimates, but these can have quite different finite sample and asymptotic properties. To understand the issues involved it is useful to review a simple method that was suggested by Engle and Granger (1987). For simplicity we consider the case where there is a single cointegrating vector that has a nonzero coefficient on a variable y_{1t}; this coefficient can then be normalized to equal one. All the other variables in the system, y_{2t}, are assumed to be $I(1)$.[17]

Engle and Granger (1987) suggest a two-step procedure where the cointegrating vector is estimated in the first step. The method is simple and appealing. To estimate the cointegrating vector, just apply OLS to the static regression:

$$y_{1t} = \tau' DR_t + \theta' y_{2t} + u_t. \tag{3.17}$$

Once OLS estimates $\hat{\theta}$ have been obtained from (3.17), one can construct an estimate of the vector $\hat{\beta}' y_t$, where $\hat{\beta} = [1 \quad -\hat{\theta}']$. The parameters associated with the short-run dynamics can then be estimated by OLS on each equation of the error-correction representation (3.5), with $\hat{\beta}' y_t$ substituted for $\beta' y_t$ and trends added if necessary. The second step of this method involves OLS estimation of a model with generated regressors. Following the work of Pagan (1984), one might suspect that standard errors would need to be adjusted to account for the use of generated regressors. However, this is not the case here as shown by Stock (1987) and Engle and Granger (1987). The relevant facts are stated as follows.

Rule 21: OLS estimates of the parameters of the cointegrating vector obtained using (3.17) are consistent and converge at rate T to the true values. Furthermore, the parameter estimates obtained from OLS on

17. If some of the variables in the system are $I(0)$, they can be ignored in the first step, which estimates the cointegrating vector and reintroduced in the second step, which estimates the short-term dynamics.

(3.5) using the estimated $\hat{\beta}'y_t$ are consistent, asymptotically normal, and have the same asymptotic distribution that would prevail if one applied OLS to (3.5) using the true values $\beta'y_t$.

The important point in rule 21 is that the estimate of the cointegrating vector converges to its true value at rate T instead of the usual rate $T^{1/2}$. This is the reason why using estimated values in the second step regression leads to the same outcome (asymptotically) as if one were able to use the true values. The reason why the estimates converge at rate T, even in the presence of substantial serial correlation in the errors and correlation across variables, is that the residuals are $I(1)$ for all parameter values except those corresponding to the cointegrating vector. Hence as the sample size increases, the variance of the residuals grows without bound for all combinations of parameters other than those in the cointegrating vector.

These results are appealing, but simulation studies show that in finite samples the OLS procedure can lead to severe biases which often decrease only slowly with the sample size (see, in particular, Banerjee et al., 1986).[18] Hence, it appears that the rate T convergence result is not sufficient to ensure parameter estimates with good finite sample properties. As we will discuss below this is due to the fact that the least-squares estimate of the cointegrating vector obtained from (3.17) is not asymptotically optimal. Another disadvantage of the OLS procedure is the following:

Rule 22: OLS estimates of the parameters of the cointegrating vector in (3.17) have an asymptotic distribution that depends on nuisance parameters. Therefore, one should not attempt to test hypotheses about the cointegrating vector using these estimates unless the effect of the nuisance parameters is taken into account.

The nuisance parameters in rule 22 are of two types.[19] First, there is the serial correlation of the errors u_t in (3.17). This can be dealt with fairly easily using a nonparametric correction like the ones discussed above for the Phillips–Perron univariate unit root test or the Stock–Watson cointegration test. Second and more important, the asymptotic distribution of $T(\hat{\theta} - \theta)$ is affected by the endogeneity of the regressors y_{2t}. If the innova-

18. This study concentrates on the properties of the estimates of the cointegrating vectors and not the estimates of the parameters in (3.5). The estimates of the cointegrating vectors are indeed those of primary interest. If one has good estimates of these quantities the estimates of the parameters of the short-run dynamics should be well behaved.
19. For an extensive treatment of asymptotic distribution in models with unit roots and cointegration, see Park and Phillips (1988, 1989).

tions in y_{1t} do not Granger cause the innovations in y_{2t}, this problem disappears. Thus we have the following.

Rule 23: If the error u_t in (3.17) is serially uncorrelated and the innovations in y_{1t} do not Granger cause the innovations in y_{2t}, then the asymptotic distribution of $T(\hat{\theta} - \theta)$ is free of nuisance parameters. Furthermore, the OLS estimate $\hat{\theta}$ is asymptotically optimal and standard hypothesis tests such as Wald tests on the parameters of θ are asymptotically distributed as chi-squared random variables.

These conditions under which OLS on (3.17) yields optimal estimates are highly restrictive. Accordingly, there is a need to derive alternative estimation procedures that are asymptotically optimal in the general case. Phillips (1991b) discusses optimal inference in cointegrated systems and shows that Full Information Maximum Likelihood (FIML) methods yield asymptotically optimal estimates if they incorporate the correct prior restriction that $n - r$ unit roots are present in the system.[20] In this case an optimal asymptotic theory of inference applies and hypotheses can be tested using the usual chi-squared distribution.[21]

The intuition behind the optimality of FIML is first that it accounts parametrically for serial correlation in the static regression errors. More importantly, it is a system estimation method that accounts for the presence of endogeneity and feedback among the variables. It is important to note that the unit roots in the model need to be imposed a priori and not estimated. FIML is not optimal when it is applied to a system that does not impose unit roots, such as an unrestricted VAR. An example of FIML fully restricted by the a priori imposition of unit roots is the method proposed by Johansen (1988, 1989a) and Ahn and Reinsel (1990) for the case of a Gaussian multivariate autoregressive system. We showed in Section 3.2 how one can implement this procedure to estimate the parameters of the model, in particular those of the cointegrating matrix.[22] Johansen also discusses the algorithm implied by FIML in the case where restrictions are imposed on the cointegrating matrix β, the adjustment matrix α, or both. It is then possible to form likelihood ratio tests that are asymptotically distributed chi-squared. Wald tests of restrictions on either α or β are also asymptotically chi-squared. Ahn and Reinsel (1990)

20. For a discussion of asymptotic optimality for inference in time series models, see Jeganathan (1988).
21. Phillips (1991a) also considers system estimation procedures in the frequency domain. These also share the property of being asymptotically optimal. We omit their discussion here. For an application, see Corbae, Ouliaris, and Phillips (1990).
22. Simulation evidence on the finite sample performance of Johansen's (1988) maximum likelihood procedure is presented in Gonzalo (1989).

consider a two-step reduced rank procedure that is asymptotically equivalent to FIML.

Despite the availability of the Johansen procedure, there are several reasons to consider single equation estimation methods. First, such methods are often easier to apply. Second, knowledge of the presence of unit roots is usually obtained via a pretesting procedure that, if properly taken into account, could affect the asymptotic properties of subsequent estimates. Such pretesting issues are likely to be more severe in systems estimation than in single equation estimation procedures. Third, in the Johansen approach as in any system estimation procedure, the estimates of one equation are sensitive to possible misspecification in another equation.

In discussing single equation methods, for simplicity we first return to the assumption that there is only one cointegrating vector. The aim of the single equation methods is to find an estimator of the coefficients of this vector that has optimal properties and for which hypotheses can be tested using the normal or chi-square asymptotic distribution. Once the estimates of the cointegrating vector are available one can construct an estimate of the equilibrium error z_t, substitute it into the error-correction regression (3.5), and then estimate that regression by OLS. To our knowledge, there exist three single equation estimation methods for the cointegrating vectors that have the same asymptotic distribution as the FIML estimates, and hence that are asymptotically optimal. These are due to Phillips and Hansen (1990);[23] Saikkonen (1990); Stock and Watson (1989b); and Phillips and Loretan (1989). They vary according to whether the corrections for serial correlation in the residuals of (3.17) and the presence of endogeneity are of a parametric or nonparametric nature. Phillips and Hansen's procedure is fully nonparametric; Saikkonen and Stock and Watson correct for endogeneity in a parametric way while the correction for serial correlation is nonparametric; and Phillips and Loretan's procedure is fully parametric. These methods are asymptotically equivalent.

The fully nonparametric procedure of Phillips and Hansen (1990) starts with the OLS estimates obtained from (3.17). Two nonparametric corrections are applied to the OLS estimator to give it the same asymptotic distribution as the FIML estimator. The first correction deals with the presence of serial correlation in the residuals of (3.17) and is akin to the Phillips–Perron (1988) correction. The second correction uses a normalized nonparametric estimate of the long-run covariance between the

23. Park (1988) considers a closely related estimation method called "canonical cointegration regression" that eliminates nuisance parameter dependencies nonparametrically. His procedure also generalizes easily to a multivariate context.

innovations in y_{1t} and the innovations in y_{2t} to deal with the presence of Granger causality from y_{1t} to y_{2t}. The t-statistics for (3.17) can also be corrected in this way. Phillips and Hansen (1990) and Phillips and Loretan (1989) present simulation results showing that this estimator has rather poor finite sample properties (greater bias and mean squared error than simple OLS) when the model contains negative serial correlation. This is analogous to the behavior of the Phillips–Perron (1988) tests for a unit root as documented by Schwert (1989) and others. Otherwise, the estimator performs relatively well compared to OLS but less well than the following two procedures which incorporate some parametric structure.

The methods of Saikkonen (1990) and Stock and Watson (1989b) share with the approach of Phillips and Loretan (1989) a common parametric correction for the effect caused by the endogeneity of the regressors. Recall that the asymptotic distribution of the least-squares estimator in (3.17) is affected by the presence, in the general case, of Granger causality from the innovations in y_{1t} to the innovations in y_{2t}. From the work of Sims (1972) on testing for causality we recall the following fact. If a variable w_{1t} Granger causes another variable w_{2t}, then w_{1t} can be expressed as a linear combination of past, present, and future values of w_{2t}. The idea behind this result is that if w_{1t} Granger causes w_{2t}, future values of w_{2t} will contain information that is useful in predicting w_{1t}. Saikkonen, Stock and Watson, and Phillips and Loretan exploit this idea to propose a parametric correction to the least-squares regression (3.17) that asymptotically eliminates the effect of this endogeneity on the distribution of the least-squares estimator of θ. The idea is simply to add to the regression leads and lags of the first-differences of the regressors y_{2t}. This yields the regression

$$y_{1t} = \tau' DR_t + \theta' y_{2t} + d_1(L)\,\Delta y_{2t} + d_2(L^{-1})\,\Delta y_{2t} + v_t, \tag{3.18}$$

where $d_1(L) = \Sigma_{k=1}^{\infty} d_{1k} L^k$ and $d_2(L^{-1}) = \Sigma_{k=1}^{\infty} d_{2k} L^{-k}$. In principle the polynomials $d_1(L)$ and $d_2(L^{-1})$ have infinite order, but in practice one needs to truncate the infinite sum. This can be done using standard asymptotic distribution theory to eliminate insignificant additional lags to arrive at a parsimonious representation.

Equation (3.18) still has the problem that the residuals v_t are serially correlated, which affects the asymptotic distribution of the least-squares estimate of θ. Stock and Watson (1989b) propose two different methods for dealing with this. The first one uses OLS estimates of θ in (3.18), but corrects standard errors and Wald test statistics nonparametrically for the effect of serial correlation. The second method uses generalized

least-squares on (3.18). A preliminary OLS regression estimates the correlation structure of the residuals v_t parametrically. This is then used to construct the GLS estimator. In this case hypothesis tests on the coefficients of the cointegrating vector can be performed using the standard (unmodified) Wald test whose asymptotic distribution is then chi-squared.[24]

Phillips and Loretan (1989) instead use a parametric correction to obtain a regression equation with uncorrelated residuals. They include lags of the equilibrium error in the regression (3.18) leading to the specification:

$$y_{1t} = \tau' DR_t + \theta' y_{2t} + d_3(L)(y_{1t} - \theta' y_{2t}) + d_1(L)\Delta y_{2t} + d_2(L^{-1})\Delta y_{2t} + e_t, \quad (3.19)$$

where e_t is now a serially uncorrelated sequence and $d_3(L) = \sum_{k=1}^{\infty} d_{3k} L^k$ is an infinite lag polynomial in L, which as before is truncated in practice. Equation (3.19) differs from a single equation of an error-correction representation in that leads of Δy_{2t} are included in the regression. Also in (3.19) the coefficients θ enter nonlinearly so the equation must be estimated by nonlinear least-squares. The nonlinear least-squares estimator of θ in (3.19) has the same asymptotic distribution as the FIML estimator so that hypotheses can be tested using the standard chi-squared distribution. Phillips and Loretan present some preliminary simulation evidence about the performance of this single-equation estimator. They remark that hypothesis tests on the coefficients of the cointegrating vector appear to be sensitive to overfitting the lag length in (3.19). They suggest successively eliminating insignificant regressors in the spirit of Hendry's methodology (see Hendry, 1987; Hendry and Richard, 1982).

Our discussion of single-equation methods has considered the case where there is a single cointegrating vector in the system. In general there may be r cointegrating vectors. In this case any of the single-equation methods can be applied to a stacked system of r equations, each with a different dependent variable. The choice of r dependent variables, out of the n available, represents a normalization of the cointegrating vectors. This generalization is of course more complicated than estimating a single regression equation, but it may still be easier to apply than the Johansen system estimation method, as it requires only multivariate (linear or nonlinear) least-squares procedures. Of course, to use this approach one must know or estimate r, the cointegrating rank. Hypotheses about r can be tested following Stock and Watson (1988a) or Johansen (1989a), but if the Johansen approach is used for this purpose

24. Note that Stock and Watson (1989b) also consider the more general case where the variable can be integrated of any order with or without deterministic trends.

then it is easy to use it to estimate the cointegrating vectors as well. In this case the only remaining advantage of the single-equation methods is that they may be more robust to misspecification of the system. No comprehensive simulation study of the finite sample properties of alternative estimation procedures is yet available. It does seem that any of the optimal methods are better than using static OLS to estimate cointegrating vectors, but it is not at all clear how one should choose among the available optimal methods. More work is needed on this topic, especially on the robustness of each procedure to misspecification.

3.4 WHEN ARE COINTEGRATION METHODS NECESSARY?

In the previous sections we have developed in some detail the theory of representation, testing, and estimation for cointegrated systems. It is important to acknowledge, however, that there may be circumstances where macroeconomists can avoid using the cointegration methods we have described.

First, economic theory sometimes determines not only the existence, but also the parameters of cointegrating vectors. For example, when two variables in a model are measured in logs it is common to find that the log ratio of the variables is stationary, so that the variables are cointegrated with cointegrating vector $\beta = [1 \ -1]$. This occurs for example in real business cycle models with unit root shocks (King, Plosser, and Rebelo, 1988). Campbell and Shiller (1987, 1988b) have modeled asset price determination with unit root processes for dividends; when the model is formulated in levels, the cointegrating vector between prices and dividends is a function of the unknown discount factor, but an approximate log-linear model gives a known cointegrating vector equal to $[1 \ -1]$. When cointegrating vectors are known, the estimation and inference problem becomes fairly trivial. One can form the equilibrium errors $z_t = \beta'(y_t - \delta t)$, substitute them into the error-correction model (3.5), and estimate the model using OLS equation by equation. No nonstandard asymptotic theory is needed for testing hypotheses about the other parameters of the model. The a priori restrictions on the cointegrating vectors can be tested using univariate unit root test statistics on the equilibrium errors.

Even when the cointegrating vectors have unknown parameters, one can often avoid using cointegration methods if one is not directly interested in these parameters. Consider estimating an unrestricted VAR in levels and testing hypotheses about the VAR coefficients. Sims, Stock, and Watson (1990) point out that the asymptotic distribution of the test statistics is standard whenever the hypotheses can be expressed as restrictions on $I(0)$ regressors. This result is given in the following.

Rule 24: In an unrestricted VAR system in levels, parameter estimates have standard asymptotic distributions whenever the system can be written in such a way that the parameters appear on $I(0)$ variables. Hypothesis test statistics have standard distributions whenever the hypotheses can be expressed as restrictions on $I(0)$ variables.

At an intuitive level, this result should not be surprising. We have already seen that cointegrating vectors can be estimated in a preliminary regression, and can then be treated as known in subsequent OLS estimation of the error-correction representation. Because the estimates of cointegrating vectors converge rapidly to their true values, uncertainty about the cointegrating vectors does not affect the asymptotic distribution of the other parameters of the model (rule 21). The Sims, Stock, and Watson result extends this to the case where the cointegrating vectors are estimated simultaneously with the other parameters of the model, rather than in a first step, and where the unit roots are estimated rather than imposed on the system.

The practical usefulness of this result will depend very much on the circumstances of a particular macroeconomic investigation. Sims, Stock, and Watson discuss some leading examples. If one is testing for the significance of additional lags in a VAR, the final lag coefficients can always be written as coefficients on differences. This means that tests for lag length do not suffer from unit root problems even in a VAR in levels. The same holds more generally for restrictions that involve only a subset of the lagged levels that appear in the VAR. Tests for Granger causality from a variable y_{2t} to another variable y_{1t} are more problematic because they involve all the lagged levels of y_{2t} in a regression of y_{1t} on lags of itself and y_{2t}. Thus Granger causality test statistics have unit root distributions unless y_{2t} is cointegrated with y_{1t}.

The macroeconomic literature on the permanent income hypothesis also offers some examples. Hall (1978) and Flavin (1981) formulated a version of the permanent income hypothesis in which consumption follows a martingale. The model also implies that consumption and income are cointegrated (Campbell, 1987). Hall tested the model by regressing consumption on lagged levels of consumption and income, and testing whether the coefficients on variables other than the first lag of consumption were jointly significant. This can be expressed as a test on coefficients of changes in consumption and stationary combinations of consumption and income; therefore Hall's test is valid even when income has a unit root (Stock and West, 1988). Flavin (1981), on the other hand, tested the model by regressing the change in consumption on lagged levels of detrended income and testing the joint significance of the coefficients. This test

rejects too often asymptotically when income has a unit root, and rejects too often in finite samples even when income is stationary but has a root close to unity (Mankiw and Shapiro, 1985).

In some cases it is possible to learn about the questions at hand, while still benefitting from rule 24, by adapting the hypotheses to be tested. For example one can test the hypothesis that $K - 1$ lagged levels of y_{2t} do not help to forecast y_{1t} in a regression that includes K lagged levels of y_{1t} and y_{2t}. If this hypothesis is rejected, one has found evidence of Granger causality from y_{2t} to y_{1t} without running the standard test that is subject to unit root problems. Issues related to unit roots need be confronted only if this hypothesis is not rejected. This is an example of the general principle that one should try to test hypotheses of direct interest using procedures that are unaffected by extraneous characteristics of the problem such as the presence or absence of unit roots.

The above example involves estimation of a levels model, but one can sometimes avoid cointegration methods by working in differences. In the Hall (1978) version of the permanent income hypothesis, the joint process for consumption and income is an error-correction model, with a cointegrating vector that is unknown if one observes only a subset of total consumption (Campbell, 1987). Nevertheless one can test the model by regressing changes in consumption just on lagged changes in income and lagged changes in consumption. The omission of the error-correction term may affect the power of the test, but will not affect its size.

There remain many cases where cointegration methods have an important role to play in applied macroeconomics. First, researchers are often interested in testing for the presence of unit roots in a system of variables related by identities or behavioral models. Univariate unit root tests can yield different results, depending on which variables are chosen for the tests. In this situation a system approach such as that of Ahn and Reinsel (1990), Johansen (1988, 1989a,b), Park (1988), or Stock and Watson (1988a) can be useful. Second, economic models sometimes have underlying parameters that appear both in the cointegrating vectors and in the coefficients governing the short-run dynamics of the model. Kashyap and Wilcox (1990), for example, estimate an inventory model in which the parameters of firms' cost functions determine both the cointegrating vector between inventories and sales and the short-run dynamics of these variables.

4. *Conclusion*

We now return to the example with which we started this paper, and briefly discuss some of the implications of our analysis for estimation

and hypothesis testing of the five-variable system $X_t = (m_t, y_t, p_t, i_t, b_t)$. We note first that there is some evidence that the first three variables in this system need to be differenced twice to render them stationary, that is, they are $I(2)$ rather than $I(1)$.[25] One might suspect that this is due to a unit root in the inflation rate Δp_t rather than $I(2)$ behavior of real money and output, and indeed King, Plosser, Stock, and Watson (1991) argue that a transformed system $X_t^* = (m_t - p_t, y_t - p_t, \Delta p_t, i_t, b_t)$ contains variables all of which are $I(1)$. Here is a case where macroeconomic theory suggests certain cointegrating vectors that can be applied to the data; standard univariate unit root test statistics can be applied to the elements of X_t^* to test King, Plosser, Stock, and Watson's hypothesis.

In the introduction we mentioned three macroeconometric exercises that could be undertaken on the system X_t^*. First, one might want to estimate a money demand function relating the real money stock $m_t - p_t$ to real output $y_t - p_t$ and nominal interest rates i_t and b_t. Our review of multivariate systems with unit roots has shown that a critical issue is whether the real money stock is cointegrated with the other $I(1)$ variables in the vector X_t^*. If there is no cointegration, then the money demand regression is spurious and standard t and F tests on the estimated coefficients are meaningless. If there is cointegration, on the other hand, the parameters of the money demand regression can be estimated superconsistently by any of the methods discussed in Section 3.3. The estimated coefficients will be robust to the presence of measurement error and endogeneity of the regressors; this circumvents many of the standard problems in the money demand literature, such as which concepts of the money stock, real economic activity, and the interest rate to use, and how to adjust for endogenous responses of activity and nominal interest rates to the money supply process. It is also important to note that economic theory tells us which variables enter the cointegrated money demand relationship but does not deliver strong prior restrictions on the parameters of this cointegrating vector (the income elasticity and interest semielasticity of money demand); thus this is a case where the superconsistency result may have some practical benefits for macroeconomics.

Unfortunately the empirical evidence on cointegration of real money, real output, and nominal interest rates is mixed. King, Plosser, Stock, and Watson (1991) find evidence that these variables are cointegrated,

25. As always, this evidence is somewhat sensitive to assumptions made about trends and to the exact data series and sample period used. Stock and Watson (1989a), for example, argue that the growth rate of log M1 is trend-stationary with a positive trend in postwar U.S. data, whereas King, Plosser, Stock, and Watson (1991) argue that the growth rate of log M2 is $I(1)$ with no trend in a similar data set.

but Friedman and Kuttner (1990) argue that this evidence largely disappears in recent data as a result of the well-known change in the behavior of velocity in the early 1980s. Both these papers use log-linear deterministic trends; it is also possible that real money, output, and interest rates are cointegrated with a broken deterministic velocity trend. This illustrates another theme of our survey, that the treatment of deterministic trends is inseparable from the treatment of unit roots in macroeconomic data.

A second macroeconometric exercise follows Sims (1972, 1980a,b) to ask whether the real money stock Granger causes real output. It turns out that Granger causality tests from money to output are sensitive to whether output and money are used in raw form, or are detrended or differenced before the tests are conducted (Bernanke, 1986; Eichenbaum and Singleton, 1986; Christiano and Ljungqvist, 1988; Stock and Watson, 1989a). They are also sensitive to the inclusion of nominal interest rates in the system (Sims, 1980b). These findings should not be surprising given the result of Sims, Stock, and Watson (1990) summarized in our rule 24. If real money and real output are $I(1)$ variables, then Granger causality tests have a nonstandard distribution if the series are not cointegrated, but a standard distribution if they are cointegrated. If there is a cointegrating vector relating real money, real output, and nominal interest rates, but no cointegrating vector between real money and real output alone, then the distribution of Granger causality tests on levels will depend on whether nominal interest rates are included in the system.

A final exercise is to test the expectations theory of the term structure. This states that the long-term interest rate can be written as a constant, plus the expected discounted value of future short-term interest rates. When the short rate is $I(1)$, the expectations theory implies that the long rate is also $I(1)$. One might think that in this case the theory could be tested as a set of restrictions on a VAR in differences (Sargent, 1979). Unfortunately this strategy runs into problems because the expectations theory also implies that the spread between long rates and short rates is stationary, so long rates and short rates are cointegrated with cointegrating vector $[1 \; -1]$. This means, first, that no well-behaved VAR representation exists for differenced long and short rates, and second, that the expectations theory puts restrictions on levels that cannot be tested by looking only at differences (Campbell and Shiller, 1987). The theory can be tested using an error-correction model, which is conveniently transformed into a VAR for the yield spread and the change in the short-term interest rate.

These examples illustrate some of the possible applications of multivariate unit root methods in macroeconometrics. Some of these methods

are still relatively new and have not yet found wide application in macroeconomics, but we expect that over the next few years they will become as well established as the now familiar test procedures for unit roots in univariate time series.

REFERENCES

Ahn, S. K., and G. C. Reinsel. (1990). Estimation for partially nonstationary multivariate autoregressive models. *Journal of the American Statistical Association* 85:813–823.

Banerjee, A., J. J. Dolado, D. F. Hendry, and G. W. Smith. (1986). Exploring equilibrium relationships in econometrics through static models: Some Monte Carlo evidence. *Oxford Bulletin of Economics and Statistics* 48:253–277.

Banerjee, A., R. L. Lumsdaine, and J. H. Stock. (1990). Recursive and sequential tests of the unit root and trend break hypothesis. Mimeo, Harvard University.

Bernanke, B. S. (1986). Alternative explanations of the money-income correlation. *Carnegie–Rochester Conference Series on Public Policy* 25:49–100.

Beveridge, S., and C. R. Nelson. (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the 'business cycle.' *Journal of Monetary Economics* 7:151–174.

Blough, S. R. (1988). On the impossibility of testing for unit roots and cointegration in finite samples. Working Paper No. 211, Johns Hopkins University.

Campbell, J. Y. (1987). Does saving anticipate declining labor income? An alternative test of the permanent income hypothesis. *Econometrica* 55:1249–1273.

Campbell, J. Y., and N. G. Mankiw. (1987). Are output fluctuations transitory? *Quarterly Journal of Economics* 102:857–880.

Campbell, J. Y., and R. J. Shiller. (1987). Cointegration and tests of present value models. *Journal of Political Economy* 95:1062–1088.

———. (1988a). Interpreting cointegrated models. *Journal of Economic Dynamics and Control* 12:505–522.

———. (1988b). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies* 1:195–228.

Chen, C., and G. C. Tiao. (1990). Random level-shift time series models, ARIMA approximations, and level-shift detection. *Journal of Business and Economic Statistics* 8:83–98.

Christiano, L. J., and M. Eichenbaum. (1989). Unit roots in real GNP: Do we know, and do we care? NBER Working Paper No. 3130.

Christiano, L. J., and L. Ljungqvist. (1988). Money does Granger-cause output in the bivariate money-output relation. *Journal of Monetary Economics* 22:217–235.

Clark, P. K. (1987). The cyclical component of U.S. economic activity. *Quarterly Journal of Economics* 102:798–814.

Clark, P. K. (1989). Trend reversion in real output and unemployment. *Journal of Econometrics* 40:15–32.

Cochrane, J. H. (1988). How big is the random walk in GNP? *Journal of Political Economy* 96:893–920.

———. (1991). A critique of the application of unit root tests. *Journal of Economic Dynamics and Control* 15:275–284.

Corbae, D., S. Ouliaris, and P. C. B. Phillips. (1990). A reexamination of the

consumption function using frequency domain regressions. Mimeo, University of Iowa.

Davidson, J. E. H., and D. F. Hendry. (1981). Interpreting econometric evidence: The behavior of consumer's expenditure in the U.K. *European Economic Review* 16:177–192.

Davidson, J. E. H., D. F. Hendry, F. Srba, and S. Yeo. (1978). Econometric modelling of the aggregate time-series relationship between consumer's expenditure and income in the United Kingdom. *Economic Journal* 88:661–692.

DeJong, D. N., J. C. Nankervis, N. E. Savin, and C. H. Whiteman. (1990a). Integration versus trend-stationarity in time series. Mimeo, University of Iowa.

———. (1990b). The power problems of unit root tests in time series with autoregressive errors. Mimeo, University of Iowa.

Dickey, D. A., and W. A. Fuller. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74:427–431.

———. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49:1057–1072.

Dickey, D. A. and S. G. Pantula. (1987). Determining the order of differencing in autoregressive processes. *Journal of Business and Economic Statistics* 5: 455–462.

Dickey, D. A., W. R. Bell, and R. B. Miller. (1986). Unit roots in time series models: Tests and implications. *American Statistician* 40:12–26.

Diebold, F. X., and M. Nerlove. (1990). Unit roots in economic time series: A selective survey. In *Advances in Econometrics: Cointegration, Spurious Regressions, and Unit Roots*. T.B. Fomby and G.F. Rhodes (eds.), pp. 3–70. Greenwich, CT: JAI Press.

Dolado, J. J., T. Jenkinson, and S. Sosvilla-Rivero. (1990). Cointegration and unit roots. *Journal of Economic Surveys* 4(3):249–273.

Eichenbaum, M., and K. J. Singleton. 1986. Do equilibrium real business cycle theories explain postwar U.S. business cycles? *NBER Macroeconomics Annual 1986*, 91–135.

Engle, R. F., and C. W. J. Granger. (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55:251–276.

Engle, R. F., and B. S. Yoo. (1987). Forecasting and testing in cointegrated systems. *Journal of Econometrics* 35:143–159.

Evans, G. B. A., and N. E. Savin. (1981). Testing for unit roots: 1. *Econometrica* 49:753–779.

Flavin, M. A. (1981). The adjustment of consumption to changing expectations about future income. *Journal of Political Economy* 89:974–1009.

Friedman, B. M., and K. N. Kuttner. (1990). Money, income, prices and interest rates after the 1980's. Mimeo, Harvard University.

Fuller, W. A. (1976). *Introduction to Statistical Time Series*. New York: John Wiley.

Ghysels, E., and P. Perron. (1990). The effect of seasonal adjustment filters on tests for a unit root. Econometric Research Program Memorandum No. 355, Princeton University. *Journal of Econometrics* (in press).

Gonzalo, J. (1989). Comparison of five alternative methods of estimating cointegrating vectors. Mimeo, University of California at San Diego.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16:121–130.

———. (1983). Co-integrated variables and error correcting models. Unpublished Discussion Paper No. 83-13, University of California at San Diego.
———. (1986). Developments in the study of cointegrated variables. *Oxford Bulletin of Economics and Statistics* 48:213–228.
Granger, C. W. J., and P. Newbold. (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2:111–120.
Granger, C. W. J., and A. A. Weiss. (1983). Time series analysis of error-correcting models. In *Studies in Econometrics, Time Series, and Multivariate Statistics*, pp. 255–278. New York: Academic Press.
Gregory, A. W. (1991). Testing for cointegration in linear quadratic models. Mimeo, Queen's University.
Hall, A. (1990). Testing for a unit root in time series with pretest data based model selection. Mimeo, North Carolina State University.
Hall, R. E. (1978). Stochastic implications of the life cycle-permanent income hypothesis: Theory and evidence. *Journal of Political Economy* 86:971–987.
Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57:357–384.
Hansen, B. E. (1990a). A powerful, simple test for cointegration using Cochrane–Orcutt. Mimeo, University of Rochester.
———. (1990b). Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends. *Journal of Econometrics* (in press).
Harvey, A. C. (1985). Trends and cycles in macroeconomic time series. *Journal of Business and Economic Statistics* 3:216–227.
Hendry, D. F. (1980). Econometrics: Alchemy or science. *Economica* 47:387–406.
———. (1986). Econometric modelling with cointegrated variables: An overview. *Oxford Bulletin of Economics and Statistics* 48:201–212.
———. (1987). Econometric methodology: A personal perspective. In *Advances In Econometrics*, Vol. 2, pp. 29–48. T. Bewley (ed.). Cambridge: Cambridge University Press.
Hendry, D. F., and J. F. Richard. (1982). On the formulation of empirical models in dynamic econometrics. *Journal of Econometrics* 20:3–33.
Jaeger, A. (1990). Shock persistence and the measurement of prewar output series. *Economics Letters* 34:333–337.
Jaeger, A., and R. M. Kunst. (1990). Seasonal adjustment and measuring persistence in output. *Journal of Applied Econometrics* 5:47–58.
Jeganathan, P. (1988). Some aspects of asymptotic theory with applications to time series models. Mimeo, University of Michigan.
Johansen, S. (1988). Statistical analysis of cointegrating vectors. *Journal of Economic Dynamics and Control* 12:231–254.
———. (1989a). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica* (in press).
———. (1989b). *Likelihood Based Inference on Cointegration Theory and Applications*. Centro Interuniversitario di Econometria, Bologna.
———. (1991). Determination of cointegrating rank in the presence of a linear trend. Mimeo, University of Copenhagen.
Johansen, S., and K. Juselius. (1990). Maximum likelihood estimation and inference on cointegration—with applications to the demand for money. *Oxford Bulletin of Economics and Statistics* 52:169–210.
Kashyap, A. K., and D. W. Wilcox. (1990). Production and inventory control at

the General Motors Corporation during the 1920's and 1930's. Federal Reserve Board Finance and Economics Discussion Paper 135.

King, R. G., C. I. Plosser, and S. T. Rebelo. (1988). Production, growth and business cycles II. New directions. *Journal of Monetary Economics* 21:309–341.

King, R. G., C. I. Plosser, J. H. Stock, and M. W. Watson. (1991). Stochastic Trends and Economic Fluctuations. *American Economic Review* (in press).

Kleidon, A. W. (1986). Variance bounds tests and stock price valuation models. *Journal of Political Economy* 94:953–1001.

Lam, P. S. (1990). The Hamilton model with a general autoregressive component: Estimation and comparison with other models of economic time series. Mimeo, The Ohio State University.

Mankiw, N. G., and M. D. Shapiro. (1985). Trends, random walks, and tests of the permanent income hypothesis. *Journal of Monetary Economics* 16:165–174.

Nelson, C. R., and C. I. Plosser. (1982). Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10:139–162.

Ogaki, M., and J. Y. Park. (1990). A cointegration approach to estimating preference parameters. Mimeo, University of Rochester.

Ouliaris, S., J. Y. Park, and P. C. B. Phillips. (1989). Testing for a unit root in the presence of a maintained trend. In *Advances in Econometrics and Modelling*. B. Raj (ed.), pp. 7–28. Dordrecht: Kluwer Academic Publishers.

Pagan, A. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review* 25:221–247.

Pantula, S. G. (1989). Testing for unit roots in time series data. *Econometric Theory* 5:256–271.

Park, J. Y. (1988). Canonical cointegrating regressions. CAE Working Paper No. 88-29, Cornell University.

———. (1990a). Testing for unit roots and cointegration by variable addition. In *Advances in Econometrics: Co-integration, Spurious Regressions, and Unit Roots*, pp. 107–133. T. B. Fomby and G. F. Rhodes (eds.). Greenwich, CT: JAI Press.

———. (1990b). Maximum likelihood estimation of simultaneous cointegrated models. Memo 1990-18, Institute of Economics, University of Aarhus.

Park, J. Y., and B. Choi. (1988). A new approach to testing for a unit root. Mimeo, Cornell University.

Park, J. Y., and P. C. B. Phillips. (1988). Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory* 4:468–497.

———. (1989). Statistical inference in regressions with integrated processes: Part 2. *Econometric Theory* 5:95–131.

Park, J. Y., S. Ouliaris, and B. Choi. (1988). Spurious regressions and tests for cointegration. Mimeo, Cornell University.

Perron, P. (1988). Trends and random walks in macroeconomic time series: Further evidence from a new approach. *Journal of Economic Dynamics and Control* 12:297–332.

———. (1989a). The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 57:1361–1401.

———. (1989b). Testing for a random walk: A simulation experiment of power when the sampling interval is varied. In *Advances in Econometrics and Modelling*. B. Raj (ed.), pp. 47–68. Dordrecht: Kluwer Academic Publishers.

———. (1990a). Testing for a unit root in a time series with a changing mean. *Journal of Business and Economic Statistics* 8:153–162.

Perron, P. (1990b). Further evidence on breaking trend functions in macroeconomic variables. Econometric Research Program Memorandum No. 350, Princeton University.

———. (1990c). Test consistency with varying sampling frequency. *Econometric Theory* (in press).

Perron, P., and P. C. B. Phillips. (1987). Does GNP have a unit root? A reevaluation. *Economics Letters* 23:139–145.

Perron, P., and T. J. Vogelsang. (1990). Nonstationarity and level shifts with an application to purchasing power parity. Unpublished manuscript, Princeton University.

Phillips, A. W. (1954). Stabilization policy in a closed economy. *Economic Journal* 64:290–323.

Phillips, P. C. B. (1986). Understanding spurious regression. *Journal of Econometrics* 33:311–340.

———. (1987). Time series regression with unit roots. *Econometrica* 55:277–302.

———. (1991a). Spectral regression for cointegrated time series. In *Nonparametric and Semiparametric Methods in Economics and Statistics*. pp. 413–435. W. Barnett, J. Powell, and G. Tauchen (eds.). Cambridge: Cambridge University Press.

———. (1991b). Optimal inference in cointegrated systems. *Econometrica* 59:283–306.

Phillips, P. C. B., and B. E. Hansen. (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* 57:99–125.

Phillips, P. C. B., and M. Loretan. (1991). Estimating long run economic equilibria. *Review of Economic Studies* 58:407–436.

Phillips, P. C. B., and S. Ouliaris. (1988). Testing for cointegration using principal components methods. *Journal of Economic Dynamics and Control* 12:205–230.

———. (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica* 58:165–193.

Phillips, P. C. B., and P. Perron. (1988). Testing for a unit root in time series regression. *Biometrika* 75:335–346.

Said, S. E., and D. A. Dickey. (1984). Testing for unit roots in autoregressive—moving average models of unknown order. *Biometrika* 71:599–608.

Saikkonen, P. (1990). Asymptotically efficient estimation of cointegration regressions. *Econometric Theory* (in press).

Sargan, J. D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. In *Econometric Analysis for National Economic Planning*. P. E. Hart, G. Mills, and J. N. Whittaker (eds.). London: Butterworths.

Sargan, J. D., and A. Bhargava. (1983). Testing the residuals from least squares regression for being generated by the Gaussian random walk. *Econometrica* 51:153–174.

Sargent, T. J. (1979). A note on maximum likelihood estimation of the rational expectations model of the term structure. *Journal of Monetary Economics* 5:133–143.

Schwert, G. W. (1989). Tests for unit roots: A Monte Carlo investigation. *Journal of Business and Economic Statistics* 7:147–160.

Shiller, R. J., and P. Perron. (1985). Testing the random walk hypothesis: Power versus frequency of observation. *Economics Letters* 18:381–386.

Sims, C. A. (1972). Money, income, and causality. *American Economic Review* 62:540–552.

———. (1980a). Macroeconomics and reality. *Econometrica* 48:1–48.

———. (1980b). Comparison of interwar and postwar business cycles: Monetarism reconsidered. *American Economic Review* 70:250–257.

Sims, C. A., J. H. Stock, and M. W. Watson. (1990). Inference in linear time series models with some unit roots. *Econometrica* 58:113–144.
Stock, J. H. (1987). Asymptotic properties of least-squares estimators of cointegrating vectors. *Econometrica* 55:1035–1056.
———. (1990). A class of tests for integration and cointegration. Mimeo, Harvard University.
Stock, J. H., and M. W. Watson. (1988a). Testing for common trends. *Journal of the American Statistical Association* 83:1097–1107.
———. (1988b). Variable trends in economic time series. *Journal of Economic Perspectives* 2(3):147–174.
———. (1989a). Interpreting the evidence on money-income causality. *Journal of Econometrics* 40:161–182.
———. (1989b). A simple MLE of cointegrating vectors in higher order integrated systems. National Bureau of Economic Research Technical Working Paper No. 83, Cambridge.
Stock, J. H., and K. D. West. (1988). Integrated regressors and tests of the permanent income hypothesis. *Journal of Monetary Economics* 21:85–96.
Watson, M. W. (1986). Univariate detrending methods with stochastic trends. *Journal of Monetary Economics* 18:1–27.
West, K. D. (1988). Asymptotic normality when regressors have a unit root. *Econometrica* 56:1397–1418.
Zivot, E., and D. W. K. Andrews. (1990). Further evidence on the great crash, the oil price shock and the unit root hypothesis. Cowles Foundation Discussion Paper No. 944, Yale University.

Comment

JOHN H. COCHRANE[1]
University of Chicago and NBER

This paper is an outstanding survey of unit root econometrics. It is an enormous and confusing literature, and Campbell and Perron's 24 rules are a tremendous and very practical condensation. If you decide to run unit root tests, this is a good place to start.

Rather than pick on rule 22, or survey some fields that the authors left out of this already massive paper (such as the Bayesian view or fractional unit roots), I will devote my comments to some reservations on practical usefulness. The bottom line is that, as much as I admire this paper as a survey of what *econometricians* know about unit roots, I am not yet convinced that this is what *macroeconomists* should know about unit roots.

For the moment, there are two broad uses of unit root econometrics, and I think it is best to organize my thoughts about what macroeconomists need to know about unit root econometrics by how they use it.

1. University of Chicago and NBER. I thank Jim Stock and Mark Watson for helpful discussions in preparing these comments.