

Department of Economics
University of California at Berkeley

M. Jansson

Testing the Order of Differencing in ARIMA Models

Suppose $\{y_t : 1 \leq t \leq T\}$ is an observed scalar time series generated by the model

$$y_t = \mu_t + u_t,$$

where u_t has mean zero and $\mu_t = E(y_t)$ is some parametric function of t such as (i) $\mu_t = 0$, (ii) $\mu_t = \mu$, or (iii) $\mu_t = \mu + \delta t$. Moreover, suppose u_t is generated by an $ARIMA(p, d, q)$ model of the form

$$\phi(L)(1-L)^d u_t = \theta(L)\varepsilon_t,$$

where

- $\phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$, $\phi_p \neq 0$, and $\phi(z) = 0 \Rightarrow |z| > 1$
- $\theta(L) = 1 + \theta_1 L + \dots + \theta_q L^q$, $\theta_q \neq 0$, and $\theta(z) = 0 \Rightarrow |z| > 1$
- $\phi(z)$ and $\theta(z)$ have no common zeros
- $\varepsilon_t \sim WN(0, \sigma^2)$; that is, $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, and $E(\varepsilon_t \varepsilon_s) = 0$ if $s \neq t$

We can fit the model to the data by proceeding as follows:

1. Select d
2. Given d , select (p, q) by minimizing an information criterion (e.g., AIC or BIC)
3. Given (p, d, q) , estimate $\phi(L)$, $\theta(L)$, and μ_t by minimizing a conditional sum of squares function

The differencing parameter d can be selected in a number of ways. The most well known and widely used approach is to base the choice between $d = 1$ and $d = 0$ on a test of $H_0 : \rho = 1$ vs. $H_1 : \rho < 1$ in a model of the form

$$\phi(L)(1 - \rho L)u_t = \theta(L)\varepsilon_t. \tag{1}$$

Under H_0 , $u_t \sim ARIMA(p, 1, q)$, while $u_t \sim ARMA(p + 1, q) = ARIMA(p + 1, 0, q)$ under H_1 . A test of $H_0 : \rho = 1$ vs. $H_1 : \rho < 1$ can therefore be viewed as a test of $d = 1$ against $d = 0$. Tests of this type are called *unit root tests*. Nelson and Plosser (1982, Journal of Monetary Economics, 10, 139-162) applied unit root tests to 14 U.S. macroeconomic time series (assuming $\mu_t = \mu + \delta \cdot t$). They were unable to reject the null hypothesis of *difference stationarity* in 13 cases. The only exception was the unemployment rate, which was found to be *trend stationary*.

An alternative approach is to select d based on a test of $H_0 : \theta^* = 1$ vs. $H_1 : \theta^* < 1$ in a model of the form

$$\phi(L)(1-L)u_t = (1-\theta^*L)\theta(L)\varepsilon_t. \quad (2)$$

When $\theta^* = 1$, there is a common factor in $\phi(L)(1-L)$ and $(1-\theta^*L)\theta(L)$, so $u_t \sim ARMA(p, q) = ARIMA(p, 0, q)$ under H_0 , whereas $u_t \sim ARIMA(p, 1, q+1)$ under H_1 . Tests of this type are called *stationarity tests* and are tests of $d = 0$ against $d = 1$. Kwiatkowski, Phillips, Schmidt and Shin (1992, *Journal of Econometrics*, 54, 159-178) applied stationarity tests to the Nelson-Plosser data set and rejected the null hypothesis of (trend) stationarity in 5 out of 14 cases. In practice, one often carries out both a unit root test and a stationarity test and hopes that they yield the same conclusion.

There are other ways to approach the problem of selecting d . One can consider a more general class of models, the *ARFIMA* class, in which the *fractional* differencing parameter d can be any real number (e.g., Hamilton, p. 448). Another approach is to act as a Bayesian and apply a decision theoretic classification scheme (see section 6.2 of Stock's (1994) handbook chapter).

Unit Root Tests

Suppose

$$y_t = \mu_t + u_t,$$

where μ_t is a deterministic component and u_t is a zero-mean scalar time series. Unit root tests are tests of $H_0 : \rho = 1$ vs. $H_1 : \rho < 1$ in a model of the form

$$(1 - \rho L) u_t = \psi(L) \varepsilon_t \quad (t = 1, \dots, T), \quad (3)$$

where $u_0 = 0$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$, and $\varepsilon_t \sim WN(0, \sigma^2)$.

The assumption $u_0 = 0$ is convenient and can be relaxed somewhat without changing the results. The model in (3) reduces to the model in (1) when $\psi(L) = \phi(L)^{-1} \theta(L)$. We study the more general case where only mild summability conditions are imposed on $\{\psi_i : i \geq 0\}$ partly because we wish to avoid making specific assumptions about the orders of $\phi(L)$ and $\theta(L)$ in the context of model selection.

Each of the following cases will be considered:

1. $\mu_t = 0$, $\psi(L) = 1$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.
2. $\mu_t = 0$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi(1) = \sum_{i=0}^{\infty} \psi_i \neq 0$, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.
3. $\mu_t = \mu$ or $\mu_t = \mu + \delta t$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi(1) \neq 0$, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.

Case 1 is the canonical case. It is useful for illustrating the fundamental differences between testing in the unit root ($\rho = 1$) case and the stationary case ($|\rho| < 1$), but it is too simple to be of empirical relevance. Case 2 introduces serial correlation, while case 3 accommodates a nonzero mean in y_t .

CASE 1: The canonical case

Suppose

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (t = 1, \dots, T), \quad (4)$$

where $y_0 = 0$ and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.

A natural test statistic is $t(1)$, where, for any ρ_0 ,

$$t(\rho_0) = \frac{\hat{\rho} - \rho_0}{\hat{\sigma} / \sqrt{\sum_{t=1}^T y_{t-1}^2}},$$

with

$$\hat{\rho} = \frac{\sum_{t=1}^T y_{t-1} y_t}{\sum_{t=1}^T y_{t-1}^2}, \quad \hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^T (y_t - \hat{\rho} y_{t-1})^2.$$

As the notation suggests, $t(\rho_0)$ is the OLS t -statistic used to test $H_0 : \rho = \rho_0$. The following results explain why and how $t(\rho_0)$ “should” be used when $|\rho_0| < 1$:

- (i) The test which rejects for small values of $t(\rho_0)$ is “the” asymptotically uniformly most powerful test of $\rho = \rho_0$ against $\rho < \rho_0$ when $\varepsilon_t \sim i.i.d. \mathcal{N}(0, \sigma^2)$.
- (ii) $t(\rho_0) \rightarrow_d \mathcal{N}(0, 1)$ when $\rho = \rho_0$.

It turns out that neither of these results hold in the unit root case where $\rho_0 = 1$. Stock (1994) discusses the failure of (i) and surveys recent research on optimal unit root tests. In spite of the fact that (i) is false when $\rho_0 = 1$, we will use $t(1)$ as the test statistic. In order to do so, we must address the failure of (ii) and find a way of characterizing the limiting null distribution of $t(1)$.

The standard case. To see why (ii) breaks down when $\rho_0 = 1$, it might be instructive to see why it holds when $|\rho_0| < 1$. When $|\rho| < 1$, a law of large numbers (LLN) can be used to show that

$$\hat{\sigma}^2 \rightarrow_p \sigma^2$$

and

$$\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 = E \left(\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right) + o_p(1) \rightarrow_p \frac{\sigma^2}{1 - \rho^2},$$

while a central limit theorem (CLT) can be used to show that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T y_{t-1} \varepsilon_t \rightarrow_d \mathcal{N} \left(0, \frac{\sigma^4}{1 - \rho^2} \right).$$

Therefore,

$$\sqrt{T}(\hat{\rho} - \rho) = \left(\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T y_{t-1} \varepsilon_t \right) \rightarrow_d \left(\frac{\sigma^2}{1 - \rho^2} \right)^{-1} \mathcal{N} \left(0, \frac{\sigma^4}{1 - \rho^2} \right) = \mathcal{N} (0, 1 - \rho^2)$$

and

$$\hat{\sigma} / \sqrt{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2} \rightarrow_p \sigma / \sqrt{\frac{\sigma^2}{1 - \rho^2}} = \sqrt{1 - \rho^2},$$

implying

$$t(\rho) = \frac{\sqrt{T}(\hat{\rho} - \rho)}{\hat{\sigma} / \sqrt{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2}} \rightarrow_d \frac{\mathcal{N}(0, 1 - \rho^2)}{\sqrt{1 - \rho^2}} = \mathcal{N}(0, 1).$$

The intermediate results

$$\sqrt{T}(\hat{\rho} - \rho) \rightarrow_d \mathcal{N}(0, 1 - \rho^2), \quad \hat{\sigma} / \sqrt{\frac{1}{T} \sum_{t=1}^T y_{t-1}^2} \rightarrow_p \sqrt{1 - \rho^2},$$

continue to hold when $\rho = 1$, but are not very useful because they merely show that $\sqrt{T}(\hat{\rho} - \rho) \rightarrow_p 0$ and $\hat{\sigma} / \sqrt{T^{-1} \sum_{t=1}^T y_{t-1}^2} \rightarrow_p 0$ in the unit root case.

The unit root case. In the unit root case, the result $\hat{\sigma}^2 \rightarrow_p \sigma^2$ still holds. Moreover, it turns out that $T^{-2} \sum_{t=1}^T y_{t-1}^2$ and $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$ both have nondegenerate limiting distributions. Therefore,

$$t(1) = \frac{T(\hat{\rho} - 1)}{\hat{\sigma} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T Y_{t-1}^2}} \stackrel{(\rho=1)}{=} \frac{\left(\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} \varepsilon_t \right)}{\hat{\sigma} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2}}$$

does have a limiting distribution when $\rho = 1$. To characterize that (nonstandard) limiting distribution, we must find a way of characterizing the joint limiting distribution of $T^{-2} \sum_{t=1}^T y_{t-1}^2$ and $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$.

We will proceed in two steps:

- I. Find an asymptotic approximation to the joint distribution of $(y_1, \dots, y_T)'$.
- II. Express $T^{-2} \sum_{t=1}^T y_{t-1}^2$ and $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$ as “smooth” functions of $(y_1, \dots, y_T)'$.

The first task is accomplished by using $(y_1, \dots, y_T)'$ to construct a random function, $Y_T(\cdot)$, which (i) contains all the information in $(y_1, \dots, y_T)'$ and (ii) converges in distribution to some random function, $Y(\cdot)$, with known distributional properties. Having done that, step II is carried out by showing that $T^{-2} \sum_{t=1}^T y_{t-1}^2$ and $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$ can be written as continuous functions of the random function $Y_T(\cdot)$ constructed in step I.

Step I. The basic idea is that the distribution of $(y_1, \dots, y_T)'$, the outcome of a discrete time random walk, can be approximated by that of a Gaussian (i.e., normally distributed) continuous time counterpart of a random walk. The key point is that the continuous time approximation to $(y_1, \dots, y_T)'$ is normally distributed no matter what the distribution of the underlying errors ε_t happens to be. That fact follows from *Donsker's theorem*, which can be viewed as a refined version of the Lindeberg-Lévy CLT.

Terminology. We say that a discrete time series $\{y_t : t \geq 0\}$ is a *random walk* if $y_0 = 0$ and the sequence $\{y_t - y_{t-1} : t \geq 1\}$ is *i.i.d.* $(0, \sigma^2)$. So, a random walk y_t has finite variance and

- starts at zero: $y_0 = 0$
- has *independent, stationary increments*: for all t and $k \geq 1$, $y_{t+k} - y_t$ is independent of (y_1, \dots, y_t) , has mean zero, and has a distribution which only depends on k

A time series $\{y_t : t \geq 0\}$ is random walk if and only if $y_t = \sum_{s=1}^t \varepsilon_s$ for some $\varepsilon_t \sim i.i.d. (0, \sigma^2)$. This representation and the CLT suggest that y_t is approximately Gaussian for large t . Any random walk is a *martingale*: for all t and $k \geq 1$, $E(y_{t+k} | y_1, \dots, y_t) = y_t$. A random walk $\{y_t\}$ is *Gaussian* if $y_{t+k} - y_t \sim \mathcal{N}(0, k \cdot \sigma^2)$ for any $k \geq 1$.

These concepts all have continuous time analogues. For our purposes, it suffices to define the continuous time process with which we will be dealing. A continuous time process $\{Y(r) : 0 \leq r \leq 1\}$ is a *Brownian motion with variance σ^2* , denoted $Y \sim BM(\sigma^2)$, if

- $Y(0) = 0$
- For $0 \leq r_1 < r_2 \leq 1$, $Y(r_2) - Y(r_1)$ is independent of $\{Y(r) : 0 \leq r \leq r_1\}$ and is distributed $\mathcal{N}[0, \sigma^2(r_2 - r_1)]$
- Any realization of $Y(\cdot)$ is continuous

A *standard Brownian motion*, $BM(1)$, is called a *Wiener process*. Clearly, $Y \sim BM(\sigma^2)$ if $Y(\cdot) = \sigma W(\cdot)$ for some Wiener process W (and vice versa). If $Y \sim BM(\sigma^2)$, the distribution of any finite-dimensional vector $[Y(r_1), \dots, Y(r_k)]$ is multivariate normal with $E[Y(r_i)] = 0$ and $E[Y(r_i)Y(r_j)] = \sigma^2 \min(r_i, r_j)$ for $1 \leq i, j \leq k$. In particular, $Y(r) \sim \mathcal{N}(0, \sigma^2 r)$ for any r .

Construction of $Y_T(\cdot)$. The aim is to construct a function $Y_T(\cdot)$ whose limiting distribution is (in a sense made precise later on) that of $Y(\cdot) \sim BM(0, \sigma^2)$. For $0 \leq r \leq 1$, let

$$Y_T(r) = \frac{y_{\lfloor Tr \rfloor}}{\sqrt{T}},$$

where $\lfloor \cdot \rfloor$ denotes the integer part of the argument; that is, let

$$Y_T(r) = \begin{cases} 0 & \text{for } 0 \leq r < 1/T \\ y_1/\sqrt{T} & \text{for } 1/T \leq r < 2/T \\ \vdots & \\ y_T/\sqrt{T} & \text{for } r = 1. \end{cases}$$

Starting with the points $(0, y_0 = 0)$, $(1, y_1)$, \dots , (T, y_T) , the function $Y_T(\cdot)$ is constructed in the following way. First, we rescale the x axis. Then a finite set of points is turned into a function by “filling in blanks”. Finally, the y axis is rescaled.

More precisely, the starting point is a plot of $\{y_t : 1 \leq t \leq T\}$ with x coordinates $0, 1, \dots, T$. Relabel these $0, 1/T, \dots, 1$ so that the domain is a subset of $[0, 1]$ for all T .

Then, to make the domain of the object under study independent of T (if the domain depends on T it does not make sense to talk about convergence), we use the dots $(0, 0), (1/T, y_1), \dots, (1, y_T)$ to construct a function on $[0, 1]$. To do so, we must define the values of the function in the intervals $0 < r < 1/T, 1/T < r < 2/T, \dots, (T-1)/T < r < 1$. One can either (i) interpolate between the dots or (ii) construct a right-continuous step function. Interpolation makes step I easier because the we are dealing with continuous functions in that case. On the other hand, step II is more tedious in the interpolation approach. Since we are only going to wave our hands in step I anyway, we choose the step function approach, as it makes step II easier.

Finally, we rescale the y axis by dividing all variables by \sqrt{T} . Heuristically, we must make sure that the variance of $Y_T(r)$ equals that of $Y(r) \sim \mathcal{N}(0, \sigma^2 r)$ in the limit. Since $\text{Var}(y_t) = \sigma^2 \cdot t$, it holds for any $r \in [0, 1]$ that as $T \rightarrow \infty$,

$$\text{Var}[Y_T(r)] = \text{Var}\left(y_{\lfloor Tr \rfloor} / \sqrt{T}\right) = \sigma^2 \frac{\lfloor Tr \rfloor}{T} \rightarrow \sigma^2 r = \text{Var}[Y(r)].$$

Functional limit theory. The function $Y_T(\cdot)$ is a member of $D[0, 1]$, the space of functions defined on the unit interval that are right continuous with (finite) left limits. Functions of this type are called *CADLAG* functions. It is possible to define probabilistic concepts such as convergence in distribution, convergence in probability etc. on abstract spaces of this form. To do so, one starts by defining a metric (a distance measure) on $D[0, 1]$. Once $D[0, 1]$ has been equipped with a metric, the usual (measure-theoretic) approach to probability theory can be applied: the metric induces a σ -algebra (the Borel σ -algebra) upon which a probability measure can be defined.

Defining the metric on $D[0, 1]$ in an “appropriate” way is far from trivial. For our purposes, it suffices to pretend that the distance between two functions in $D[0, 1]$ is measured by means of the uniform metric d_{sup} defined as

$$d_{\text{sup}}(f, g) = \sup\{|f(r) - g(r)| : 0 \leq r \leq 1\}.$$

Some technical difficulties arise when d_{sup} is used, but we will ignore those here (for details, see the book by Billingsley (1999)).

Remark. For obvious reasons, the function $\|\cdot\| = d_{\text{sup}}(\cdot, 0)$ is often called the *sup norm*, although some people call it the \mathcal{L}^∞ norm in recognition of certain relations between the sup norm and the \mathcal{L}^p norm defined as $\|f\|_p = \left(\int_0^1 |f(r)|^p dr\right)^{1/p}$ for $0 < p < \infty$. ■

A sequence $\{X_T\}$ of real-valued random variables *converges in probability* to the random variable X , denoted $X_T \rightarrow_p X$, if $\lim_{T \rightarrow \infty} \Pr(|X_T - X| > \varepsilon) = 0$ for all $\varepsilon > 0$. Convergence in probability generalizes in an obvious way to $D[0, 1]$: A sequence $\{Y_T(\cdot)\} \subseteq D[0, 1]$ *converges in probability* to $Y(\cdot) \in D[0, 1]$, denoted $Y_T \rightarrow_p Y$ or $Y_T(\cdot) \rightarrow_p Y(\cdot)$, if $\lim_{T \rightarrow \infty} \Pr(d_{\text{sup}}(Y_T, Y) > \varepsilon) = 0$ for all $\varepsilon > 0$. That is, $Y_T \rightarrow_p Y$ if the real-valued random variable $d_{\text{sup}}(Y_T, Y) = \sup_{0 \leq r \leq 1} |Y_T(r) - Y(r)|$ converges to zero in probability.

A sequence $\{X_T\}$ of real-valued random variables *converges in distribution* to the random variable X , denoted $X_T \rightarrow_d X$, if $\lim_{T \rightarrow \infty} \Pr(X_T \leq x) = \Pr(X \leq x)$ at all continuity points of $\Pr(X \leq \cdot)$. A general definition of convergence in distribution is suggested by the Helley-Bray theorem, according to which $X_T \rightarrow_d X$ if and only if $\lim_{T \rightarrow \infty} E[f(X_T)] = E[f(X)]$ for all bounded, continuous real-valued functions f defined on \mathbb{R} . We say that a sequence $\{Y_T(\cdot)\} \subseteq D[0, 1]$ *converges in distribution* to $Y(\cdot) \in D[0, 1]$,

denoted $Y_T \rightarrow_d Y$ or $Y_T(\cdot) \rightarrow_d Y(\cdot)$, if $\lim_{T \rightarrow \infty} E[F(X_T)] = E[F(X)]$ for all bounded, continuous real-valued functionals F defined on $D[0, 1]$. Here, F is *bounded* if $\sup_{f \in D[0, 1]} |F(f)| < \infty$, F is *continuous* if for any $f \in D[0, 1]$ and any $\varepsilon > 0$, there is $\delta > 0$ such that $|F(f) - F(g)| < \varepsilon$ for any $g \in D[0, 1]$ with $d_{\text{sup}}(f, g) < \delta$, and we use the term *functional* to point out that F is a function of a function.

The important thing to realize is that it makes sense to talk about convergence in probability and convergence in distribution in function spaces such as $D[0, 1]$. The definition of convergence in probability is fairly operational, whereas the definition of convergence in distribution is somewhat abstract. It is possible to give more primitive sufficient (and necessary) conditions for convergence in distribution, but fortunately we will never have to establish convergence in distribution directly. Instead, we will always be able to write the functions of interest as the sum of two terms, $Y_T = Y_T^L + Y_T^R$ say, such that Y_T^L (the “leading” term) can be analyzed using existing convergence results and Y_T^R (the “remainder” term) converges in probability to a nonrandom function.

The result that will enable us to complete step I is Donsker’s theorem:

Theorem. *If $\varepsilon_t \sim i.i.d. (0, \sigma^2)$, then*

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor T \cdot \rfloor} \varepsilon_t \rightarrow_d \sigma W(\cdot),$$

where W is a Wiener process.

It follows from Donsker’s theorem that as $T \rightarrow \infty$, the function $Y_T(\cdot)$ defined above converges in distribution to $Y(\cdot) \sim BM(\sigma^2)$. In large samples it therefore holds that the distribution of the function $Y_T(\cdot)$, which contains all the information in $(y_1, \dots, y_T)'$, can be approximated by that of a Brownian motion with variance σ^2 .

Remarks. (i) Results such as Donsker’s theorem are often called functional central limit theorems (FCLTs). As the terminology suggests, it is a central limit theorem for functions. Functional central limit theorems are refinements of conventional central limit theorems. This is so because a necessary condition for $Y_T \rightarrow_d Y$ is that for any $k \geq 1$ and any $0 \leq r_1, \dots, r_k \leq 1$, the finite-dimensional vector $[Y_T(r_1), \dots, Y_T(r_k)]$ converges in distribution to $[Y(r_1), \dots, Y(r_k)]$. In particular, it must be the case that $Y_T(1) \rightarrow_d Y(1)$. Now, $Y_T(1) = y_T/\sqrt{T} = T^{-1/2} \sum_{t=1}^T \varepsilon_t$ and $Y(1) \sim \mathcal{N}(0, \sigma^2)$, so Donsker’s theorem contains the conventional Lindeberg-Lévy CLT as a special case.

(ii) Necessary and sufficient conditions for convergence in distribution in $D[0, 1]$ are finite-dimensional convergence in distribution and a technical condition called *tightness*. Essentially, tightness is a function space analogue stochastic boundedness (a sequence $\{X_T\}$ of real-valued random variables is *stochastically bounded*, denoted $X_T = O_p(1)$, if for any $\delta > 0$ there is an M such that $\Pr(|X_T| < M) < \delta$ for all T). Because finite-dimensional convergence in distribution usually follows from a conventional CLT, the hard part of the proof of an FCLT is to establish tightness. For details, consult Billingsley (1999). ■

Step II. Armed with Donsker's theorem, we can derive the limiting distribution of $t(1)$ by showing that apart from additive terms with degenerate limiting distributions, $T^{-2} \sum_{t=1}^T y_{t-1}^2$ and $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$ can be written as continuous functions of the random function $Y_T(\cdot)$ constructed in step I. The result that justifies this claim is the *continuous mapping theorem* (CMT), which states that if $Y_T \rightarrow_d Y$ in $D[0, 1]$, then $F(Y_T) \rightarrow_d F(Y)$ in \mathbb{R}^k whenever F is a continuous \mathbb{R}^k -valued functional defined on $D[0, 1]$.

First, consider $T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t$. When $\rho = 1$, $y_t = y_{t-1} + \varepsilon_t$. Squaring and rearranging, we have:

$$y_t^2 = (y_{t-1} + \varepsilon_t)^2 = y_{t-1}^2 + \varepsilon_t^2 + 2y_{t-1}\varepsilon_t \quad \Leftrightarrow \quad y_{t-1}\varepsilon_t = \frac{1}{2} (y_t^2 - y_{t-1}^2 - \varepsilon_t^2).$$

Therefore, using $y_0 = 0$,

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^T y_{t-1} \varepsilon_t &= \frac{1}{2} \frac{1}{T} \sum_{t=1}^T (y_t^2 - y_{t-1}^2 - \varepsilon_t^2) = \frac{1}{2} \left(\frac{y_T^2}{T} - \frac{y_0^2}{T} - \frac{1}{T} \sum_{t=1}^T \varepsilon_t^2 \right) \\ &= \frac{1}{2} \left(\frac{y_T^2}{T} - \frac{1}{T} \sum_{t=1}^T \varepsilon_t^2 \right). \end{aligned}$$

Now, $T^{-1} \sum_{t=1}^T \varepsilon_t^2 \rightarrow_p \sigma^2$ by LLN. Moreover, $Y_T(1)^2$ is a continuous function of $Y_T(\cdot)$, so

$$\frac{y_T^2}{T} = Y_T(1)^2 \rightarrow_d Y(1)^2 = \sigma^2 W(1)^2$$

by Donsker's theorem and CMT, where $Y \sim BM(\sigma^2)$ and W is a Wiener process. As a consequence,

$$T^{-1} \sum_{t=1}^T y_{t-1} \varepsilon_t \rightarrow_d \frac{1}{2} [\sigma^2 W(1)^2 - \sigma^2] = \frac{1}{2} \sigma^2 [W(1)^2 - 1].$$

Remark. (Lipschitz) continuity of the functional $F : D[0, 1] \rightarrow \mathbb{R}$ defined as $F(f(\cdot)) = f(1)$ is a consequence of the inequality

$$|F(f(\cdot)) - F(g(\cdot))| = |f(1) - g(1)| \leq \sup_{0 \leq r \leq 1} |f(r) - g(r)| = d_{\text{sup}}(f, g). \quad \blacksquare$$

Next,

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 = \frac{1}{T} \sum_{t=1}^T Y_T \left(\frac{t-1}{T} \right)^2 = \int_0^1 Y_T(r)^2 dr \rightarrow_d \int_0^1 Y(r)^2 dr = \sigma^2 \int_0^1 W(r)^2 dr,$$

where the first equality follows from the definition of $Y_T(\cdot)$, the convergence result follows from Donsker's theorem and CMT, and the random functions Y and W are the same as before. The proof of the second equality is left as an exercise, as is the proof of the fact that $\int_0^1 Y_T(r)^2 dr$ is a continuous function of $Y_T(\cdot)$.

Combining the results in the preceding displays, we have that if $\rho = 1$, then

$$T(\hat{\rho} - 1) = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-1} \varepsilon_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \rightarrow_d \frac{\frac{1}{2} \sigma^2 [W(1)^2 - 1]}{\sigma^2 \int_0^1 W(r)^2 dr} = \frac{\frac{1}{2} [W(1)^2 - 1]}{\int_0^1 W(r)^2 dr}.$$

Notice that $\hat{\rho} - \rho$ must be multiplied by T rather than \sqrt{T} to obtain a nondegenerate limiting distribution when $\rho = 1$. This property, not enjoyed when $|\rho| < 1$, is referred to as *superconsistency*.

Using the (readily verified) fact that $\hat{\sigma}^2 \rightarrow_p \sigma^2$,

$$\hat{\sigma} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \rightarrow_d \sigma / \sqrt{\sigma^2 \int_0^1 W(r)^2 dr} = 1 / \sqrt{\int_0^1 W(r)^2 dr},$$

so

$$t(1) = \frac{T(\hat{\rho} - 1)}{\hat{\sigma} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2}} \rightarrow_d \frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}}$$

when $\rho = 1$.

The preceding display characterizes the limiting null distribution of $t(1)$, but it should be emphasized that the quantity on the right hand side is *not* a distribution. Rather, it is a random variable whose distribution is the limiting distribution of $t(1)$ when $\rho = 1$ in (4). That distribution is sometimes called the *Dickey-Fuller distribution*. Relative to the standard normal distribution, the Dickey-Fuller distribution is skewed to the left. Percentiles can be obtained by simulation and are reported in Hamilton's (1994) Table B.6. For instance,

$$\Pr \left(\frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}} \leq -1.95 \right) = 0.05,$$

so a 5% *Dickey-Fuller test* rejects the unit root null hypothesis if $t(1) < -1.95$.

Remark. An alternative unit root test, the *Dickey-Fuller coefficient test*, can be based on $T(\hat{\rho} - 1)$. Like $t(1)$, that statistic is asymptotically pivotal. Percentiles can be found in Hamilton's (1994) Table B.5. ■

CASE 2: Serial correlation

Suppose

$$y_t = \rho y_{t-1} + v_t \quad (t = 1, \dots, T), \quad (5)$$

where $y_0 = 0$, $v_t = \psi(L)\varepsilon_t$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi(1) \neq 0$, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.

The condition $\psi(1) \neq 0$ rules out a common factor in $(1 - \rho L)$ and $\psi(L)$ when $\rho = 1$. If v_t is generated by an *ARMA* model, say $\phi(L)v_t = \theta(L)\varepsilon_t$, then $\psi(L) = \phi(L)^{-1}\theta(L)$ and the condition $\psi(1) \neq 0$ holds provided $\theta(L)$ is invertible and $\phi(z) = 0$ implies $|z| > 1$. The summability condition $\sum_{i=1}^{\infty} i |\psi_i| < \infty$ is known as *1-summability* and holds whenever $v_t \sim ARMA(p, q)$, because the coefficients ψ_i exhibit exponential decay in that case (that is, there is a $0 < \lambda < 1$ and constant $M < \infty$ such that $|\psi_i| < M\lambda^i$ for all i).

We start by deriving the limiting distribution of $t(1)$ in the serially correlated case. It turns out that this limiting distribution depends on nuisance parameters, so $t(1)$ is not asymptotically pivotal and cannot be used as a test statistic. In spite of this, the derivation of the limiting distribution of $t(1)$ is of interest because (i) it illustrates a very general approach to deriving limiting distributions in models with serial correlation and (ii) the form of the limiting distribution suggests how an asymptotically similar test can be constructed.

When $\rho = 1$,

$$T(\hat{\rho} - 1) = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-1} v_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2}$$

and it follows as in case I that

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 = \int_0^1 Y_T(r)^2 dr$$

and

$$\frac{1}{T} \sum_{t=1}^T y_{t-1} v_t = \frac{1}{2} \left[Y_T(1)^2 - \frac{1}{T} \sum_{t=1}^T v_t^2 \right],$$

where $Y_T(r) = y_{\lfloor Tr \rfloor} / \sqrt{T}$.

The limiting null distribution of

$$t(1) = \frac{\hat{\rho} - 1}{\hat{\sigma} / \sqrt{\sum_{t=1}^T y_{t-1}^2}} \stackrel{(\rho=1)}{=} \frac{\frac{1}{2} \left[Y_T(1)^2 - \frac{1}{T} \sum_{t=1}^T v_t^2 \right]}{\hat{\sigma} \sqrt{\int_0^1 Y_T(r)^2 dr}}$$

can therefore be derived by finding the limiting distribution of $Y_T(\cdot)$ and the probability limits of $\hat{\sigma}^2$ and $T^{-1} \sum_{t=1}^T v_t^2$.

Asymptotics for linear processes. The following algebraic result, adapted from Phillips and Solo (1992, *Annals of Statistics*, 20, 971-1001), is known as the *Beveridge-Nelson (BN) decomposition* of $\psi(L)$.

Lemma. If $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$, then

$$\psi(L) = \psi(1) + (1-L)\tilde{\psi}(L),$$

where $\tilde{\psi}(L) = \sum_{i=0}^{\infty} \tilde{\psi}_i L^i$ with $\tilde{\psi}_i = -\sum_{j=i+1}^{\infty} \psi_j$. If $\sum_{i=1}^{\infty} i^{1/2} |\psi_i| < \infty$ or $\sum_{i=1}^{\infty} i^2 \psi_i^2 < \infty$, then $\sum_{i=0}^{\infty} \tilde{\psi}_i^2 < \infty$.

Remark. Both summability conditions $\sum_{i=1}^{\infty} i^{1/2} |\psi_i| < \infty$ and $\sum_{i=1}^{\infty} i^2 \psi_i^2 < \infty$ are weaker than 1-summability, so $\sum_{i=1}^{\infty} i |\psi_i| < \infty$ implies $\sum_{i=0}^{\infty} \tilde{\psi}_i^2 < \infty$ in view of the lemma. In fact, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$ implies $\sum_{i=0}^{\infty} |\tilde{\psi}_i| < \infty$ because

$$\sum_{i=0}^{\infty} |\tilde{\psi}_i| = \sum_{i=0}^{\infty} \left| -\sum_{j=i+1}^{\infty} \psi_j \right| \leq \sum_{i=0}^{\infty} \sum_{j=i+1}^{\infty} |\psi_j| = \sum_{i=1}^{\infty} i |\psi_i|,$$

where the inequality uses the triangle inequality. ■

Applying the BN decomposition to $\psi(L)$, we can write v_t as

$$v_t = \psi(L)\varepsilon_t = \psi(1)\varepsilon_t + \tilde{\psi}(L)\varepsilon_t - \tilde{\psi}(L)\varepsilon_{t-1}.$$

When $\rho = 1$ in (5), the corresponding decomposition of y_t is

$$y_t = \sum_{s=1}^t v_s = \psi(1) \sum_{s=1}^t \varepsilon_s + \tilde{\psi}(L)\varepsilon_t - \tilde{\psi}(L)\varepsilon_0. \quad (6)$$

The BN decomposition facilitates the derivation of the limiting distribution of $Y_T(\cdot)$ because it delivers a decomposition of y_t in which (i) the leading term, $\psi(1) \sum_{s=1}^t \varepsilon_s$, is easy to analyze and (ii) the remainder term, $\tilde{\psi}(L)\varepsilon_t - \tilde{\psi}(L)\varepsilon_0$, is asymptotically negligible.

Remark. Beveridge and Nelson (1981, *Journal of Monetary Economics*, 7, 151-174) interpreted (6) as a *permanent-transitory decomposition* of y_t . This terminology reflects the fact that we can view (6) as a decomposition of y_t into a permanent component, $\psi(1) \sum_{s=1}^t \varepsilon_s$, a transitory component, $\tilde{\psi}(L)\varepsilon_t$, and an initial condition, $-\tilde{\psi}(L)\varepsilon_0$. ■

Let $Y_T(r) = y_{[Tr]}/\sqrt{T} = Y_T^L(r) + Y_T^R(r)$, where

$$Y_T^L(r) = \frac{1}{\sqrt{T}} \sum_{s=1}^{[Tr]} \psi(1)\varepsilon_s$$

and

$$Y_T^R(r) = \frac{1}{\sqrt{T}} \left[\tilde{\psi}(L)\varepsilon_{[Tr]} - \tilde{\psi}(L)\varepsilon_0 \right].$$

Because $\psi(1)\varepsilon_t \sim i.i.d. [0, \psi(1)^2\sigma^2]$, it follows from Donsker's theorem that $Y_T^L(\cdot) \rightarrow_d \omega W(\cdot)$, where

$$\omega^2 = \psi(1)^2\sigma^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \text{Var} \left(\sum_{t=1}^T v_t \right)$$

is the *long-run variance* of v_t and W is a Wiener process.

As a consequence, $Y_T(\cdot) = Y_T^L(\cdot) + Y_T^R(\cdot) \rightarrow_d \omega W(\cdot)$ if $Y_T^R(\cdot)$ is asymptotically negligible in the sense that $Y_T^R(\cdot) \rightarrow_p 0$. Recall that $Y_T^R(\cdot) \rightarrow_p 0$ is shorthand for

$$\sup_{0 \leq r \leq 1} |Y_T^R(r)| = \sup_{0 \leq r \leq 1} \left| \frac{1}{\sqrt{T}} \left[\tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} - \tilde{\psi}(L)\varepsilon_0 \right] \right| \rightarrow_p 0.$$

By the triangle inequality,

$$\left| \frac{1}{\sqrt{T}} \left[\tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} - \tilde{\psi}(L)\varepsilon_0 \right] \right| \leq \left| \frac{1}{\sqrt{T}} \tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} \right| + \frac{1}{\sqrt{T}} \left| \tilde{\psi}(L)\varepsilon_0 \right|,$$

so

$$\sup_{0 \leq r \leq 1} \left| \frac{1}{\sqrt{T}} \left[\tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} - \tilde{\psi}(L)\varepsilon_0 \right] \right| \leq \sup_{0 \leq r \leq 1} \left| \frac{1}{\sqrt{T}} \tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} \right| + \frac{1}{\sqrt{T}} \left| \tilde{\psi}(L)\varepsilon_0 \right|.$$

Because $\sum_{i=0}^{\infty} \tilde{\psi}_i^2 < \infty$, $\tilde{\psi}(L)\varepsilon_0$ is well defined as satisfies $T^{-1/2} \left| \tilde{\psi}(L)\varepsilon_0 \right| \rightarrow_p 0$. It can also be shown that

$$\sup_{0 \leq r \leq 1} \left| \frac{1}{\sqrt{T}} \tilde{\psi}(L)\varepsilon_{\lfloor Tr \rfloor} \right| = \max_{1 \leq t \leq T} \left| \frac{1}{\sqrt{T}} \tilde{\psi}(L)\varepsilon_t \right| \rightarrow_p 0.$$

(For a proof, see Phillips and Solo (1992).). Therefore, $Y_T^R(\cdot) \rightarrow_p 0$ and we have completed the derivation of the limiting distribution of $Y_T(\cdot)$:

$$Y_T(\cdot) \rightarrow_d \omega W(\cdot).$$

By analogy with case I,

$$\frac{y_T^2}{T} = Y_T(1)^2 \rightarrow_d \omega^2 W(1)^2$$

and

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 = \int_0^1 Y_T(r)^2 dr \rightarrow_d \omega^2 \int_0^1 W(r)^2 dr.$$

To complete the derivation of the limiting distribution of $t(1)$, we still need to find the probability limits of $\hat{\sigma}^2$ and $T^{-1} \sum_{t=1}^T v_t^2$. It can be shown that $T^{-1} \sum_{t=1}^T v_t^2 \rightarrow_p \gamma_v(0)$, where

$$\gamma_v(0) = \sigma^2 \sum_{i=0}^{\infty} \psi_i^2 = \text{Var}(v_t)$$

is the *short-run variance* of v_t . In turn, this result can be used to show that $\hat{\sigma}^2 \rightarrow_p \gamma_v(0)$.

Limiting distributions. Combining these intermediate results, we have

$$T(\hat{\rho} - 1) = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-1} v_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \rightarrow_d \frac{\frac{1}{2} [\omega^2 W(1)^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W(r)^2 dr} = \frac{\frac{1}{2} [W(1)^2 - 1]}{\int_0^1 W(r)^2 dr} + \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W(r)^2 dr}$$

and

$$t(1) = \frac{T(\hat{\rho} - 1)}{\hat{\sigma} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2}} \rightarrow_d \frac{\frac{1}{2} [\omega^2 W(1)^2 - \gamma_v(0)]}{\sqrt{\gamma_v(0)} \sqrt{\omega^2 \int_0^1 W(r)^2 dr}} = \sqrt{\frac{\omega^2}{\gamma_v(0)}} \left(\frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}} + \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \sqrt{\int_0^1 W(r)^2 dr}} \right).$$

Superconsistency. Even in the serially correlated case, the OLS estimator $\hat{\rho}$ is superconsistent when $\rho = 1$ in (5). This result is quite remarkable because conventional wisdom and results for the case where $|\rho| < 1$ suggest that $\hat{\rho}$ should be inconsistent when $E(y_{t-1}v_t) \neq 0$ in (5). The following heuristic explanation is due to Phillips (1987, *Econometrica*, 55, 277-301): *Intuitively, when the model (5) has a unit root, the strength of the signal (as measured by the sample variation of the regressor y_{t-1}) dominates the noise by a factor of $O(T)$, so that the effects of any regressor-error correlation are annihilated in the regression as $T \rightarrow \infty$.* As noted by Stock (1994, pp. 2759-2760), the fact that $\hat{\rho}$ is an extremum estimator ($\hat{\rho} = \arg \min_{\rho} T^{-1} \sum (y_t - \rho y_{t-1})^2$) can also be used to motivate the super consistency result.

We can write $\hat{\rho} - \rho$ as

$$\hat{\rho} - \rho = \left(\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T y_{t-1} v_t \right).$$

When $|\rho| < 1$,

$$\hat{\rho} - \rho = \underbrace{\left(\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right)^{-1}}_{\rightarrow_p \bar{E}(y_{t-1}^2) \neq 0} \underbrace{\left(\frac{1}{T} \sum_{t=1}^T y_{t-1} v_t \right)}_{\rightarrow_p \bar{E}(y_{t-1} v_t)} \rightarrow_p \bar{E}(y_{t-1}^2)^{-1} \bar{E}(y_{t-1} v_t),$$

where $\bar{E}(y_{t-1}^2) = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T E(y_{t-1}^2) \neq 0$ and $\bar{E}(y_{t-1} v_t) = \lim_{T \rightarrow \infty} T^{-1} \sum_{t=1}^T E(y_{t-1} v_t)$. Therefore, $\hat{\rho}$ is inconsistent when $\bar{E}(y_{t-1} v_t) \neq 0$. When $\rho = 1$, in contrast,

$$\left(\frac{1}{T} \sum_{t=1}^T y_{t-1}^2 \right)^{-1} = \frac{1}{T} \cdot \left(\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2 \right)^{-1} = \frac{1}{T} \cdot O_p(1) \rightarrow_p 0$$

and $\hat{\rho}$ is consistent whenever $T^{-1} \sum_{t=1}^T y_{t-1} v_t$ is stochastically bounded. Indeed, $\left(T^{-2} \sum_{t=1}^T y_{t-1}^2 \right)^{-1}$ is stochastically bounded, so $\hat{\rho}$ is consistent whenever $T^{-2} \sum_{t=1}^T y_{t-1} v_t \rightarrow_p 0$.

“Bias”. Although the estimator remains consistent, its limiting distribution is affected when v_t is serially correlated. The long-run variance is

$$\omega^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \text{Var} \left(\sum_{t=1}^T v_t \right),$$

whereas the short-run variance can be written as

$$\gamma_v(0) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \text{Var}(v_t).$$

Therefore, the terms involving $\omega^2 - \gamma_v(0)$ generally do not vanish unless v_t is white noise (in that case, $\omega^2 = \gamma_v(0) = \sigma^2$ and we are back in case I). It can be shown that

$$\frac{1}{2} [\omega^2 - \gamma_v(0)] = \bar{E}(y_{t-1}v_t) = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T E(y_{t-1}v_t),$$

so the second term in the limiting distributions of $T(\hat{\rho} - 1)$ and $t(1)$ reflects the “endogeneity bias” caused by nonzero values of $E(y_{t-1}v_t)$.

“Bias-correction”. The OLS variance estimator $\hat{\sigma}^2$ is a consistent estimator of $\gamma_v(0)$. Moreover, it turns out that it is possible to construct a consistent (nonparametric) estimator of ω^2 . For now, let $\hat{\omega}^2$ denote any such estimator. Consider the following “bias-corrected” version of the OLS estimator $\hat{\rho}$:

$$\hat{\rho}^+ = \hat{\rho} - T \cdot \frac{\frac{1}{2} [\hat{\omega}^2 - \hat{\sigma}^2]}{\sum_{t=1}^T y_{t-1}^2}.$$

Since

$$\frac{\frac{1}{2} [\hat{\omega}^2 - \hat{\sigma}^2]}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \rightarrow_d \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W(r)^2 dr},$$

we have:

$$\begin{aligned} T(\hat{\rho}^+ - 1) &= T(\hat{\rho} - 1) - \frac{\frac{1}{2} [\hat{\omega}^2 - \hat{\sigma}^2]}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2} \\ &\rightarrow_d \frac{\frac{1}{2} [W(1)^2 - 1]}{\int_0^1 W(r)^2 dr} + \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W(r)^2 dr} - \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W(r)^2 dr} = \frac{\frac{1}{2} [W(1)^2 - 1]}{\int_0^1 W(r)^2 dr}. \end{aligned}$$

As desired, the term involving $\omega^2 - \gamma_v(0)$ does not appear in the limiting representation of $T(\hat{\rho}^+ - 1)$.

Consider the test statistic

$$t_P(1) = \frac{\hat{\rho}^+ - 1}{\hat{\omega} / \sqrt{\sum_{t=1}^T y_{t-1}^2}}.$$

The statistic $t_P(1)$, known as *Phillips' Z_t statistic*, is obtained by making two adjustments to $t(1)$. First, the coefficient estimator in the numerator is $\hat{\rho}^+$ rather than $\hat{\rho}$. Second, the variance estimator in the denominator is an estimator of ω^2 rather than σ_v^2 . The former adjustment removes “endogeneity bias”, while the latter adjustment reflects the fact that with serially correlated error terms we should use “robust” standard errors when doing inference.

The limiting distribution of $t_P(1)$ is the Dickey-Fuller distribution when $\rho = 1$:

$$t_P(1) = \frac{T(\hat{\rho}^+ - 1)}{\hat{\omega} / \sqrt{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^2}} \rightarrow_d \frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}}.$$

Since Phillips' Z_t test handles serial correlation by employing a nonparametric estimator $\hat{\omega}^2$, it is sometimes referred to as a *semiparametric unit root test*.

Robust standard errors. The long-run variance ω^2 can be written as $\gamma_v(0) + 2 \sum_{i=1}^{\infty} \gamma_v(i)$, where $\gamma_v(i) = E(v_t v_{t-i})$ for any $t > i \geq 0$. Equivalently, ω^2 equals 2π times the zero-frequency spectral density of v_t . Consistent estimation of a spectral density was discussed in the first half. Any consistent spectral density estimator can be used to estimate ω^2 and a popular nonparametric choice is the *Newey-West estimator*

$$\hat{\omega}^2 = \hat{\gamma}_v(0) + 2 \sum_{i=1}^b \left(1 - \frac{i}{b}\right) \hat{\gamma}_v(i),$$

where $\hat{\gamma}_v(i) = T^{-1} \sum_{t=i}^T \hat{v}_t \hat{v}_{t-i}$, $\hat{v}_t = y_t - \hat{\rho} y_{t-1}$ and b is a *bandwidth parameter*.

The Newey-West estimator $\hat{\omega}^2$ is a weighted sum of the estimated autocovariances $\{\hat{\gamma}_v(i) : 0 \leq i \leq b\}$. It is consistent if $b \rightarrow \infty$ but $b/\sqrt{T} \rightarrow 0$ as $T \rightarrow \infty$. In small samples, the choice of the bandwidth parameter b is very important. Specific recommendations with respect to (data-dependent) bandwidth selection have been provided by Andrews (1991, *Econometrica*, 59, 817-858).

Augmented Dickey-Fuller tests. A popular alternative to Phillips' Z_t test is the *augmented Dickey-Fuller (ADF) test*. Suppose $v_t \sim AR(p)$:

$$v_t = \phi_1 v_{t-1} + \dots + \phi_p v_{t-p} + \varepsilon_t,$$

where p is known, $\phi(z) = 1 - \sum_{i=1}^p \phi_i z^i = 0$ has roots outside the unit circle, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$. In other words, consider the model:

$$(1 - \rho L) \phi(L) y_t = \varepsilon_t, \quad \varepsilon_t \sim i.i.d. (0, \sigma^2). \quad (7)$$

The model can be written as

$$y_t = \alpha y_{t-1} + \beta_1 \Delta y_{t-1} + \dots + \beta_p \Delta y_{t-p} + \varepsilon_t,$$

where $\alpha = 1 - (1 - \rho) \phi(1)$ and $\{\beta_i : 1 \leq i \leq p\}$ are functions of ρ and $\{\phi_i : 1 \leq i \leq p\}$. Notice that $\alpha = 1$ if and only if $\rho = 1$. The ADF test is the OLS t -test of $\alpha = 1$ from the OLS regression

$$y_t = \hat{\alpha}y_{t-1} + \hat{\beta}_1\Delta y_{t-1} + \dots + \hat{\beta}_p\Delta y_{t-p} + \hat{\varepsilon}_t \quad (t = p+1, \dots, T).$$

It can be shown that the ADF test statistic, denoted $t_{ADF}(1)$, satisfies

$$t_{ADF}(1) \rightarrow_d \frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}} \quad (8)$$

when $\rho = 1$.

Said and Dickey (1984, *Biometrika*, 71, 599-607) considered the case where v_t is generated by an *ARMA* model of unknown order. They showed that (8) holds even in that case provided the lag length p satisfies $p \rightarrow \infty$ and $p/\sqrt[3]{T} \rightarrow 0$ as $T \rightarrow \infty$.

CASE 3: Deterministic components

Suppose

$$y_t = \mu_t + u_t, \quad u_t = \rho u_{t-1} + v_t \quad (t = 1, \dots, T), \quad (9)$$

where μ_t is a deterministic component, $y_0 = \mu_0$, $v_t = \psi(L)\varepsilon_t$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi(1) \neq 0$, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$. Two cases will be considered, the *constant mean case* with $\mu_t = \mu$ and the *linear trend case* with $\mu_t = \mu + \delta t$. In both cases, the limiting distributions of interest are similar to, but different from, those encountered in case 2.

Constant mean case. In the *constant mean case* with $\mu_t = \mu$, the model can be written as

$$y_t = \rho y_{t-1} + \mu^* + v_t, \quad (10)$$

where $\mu^* = (1 - \rho)\mu$. Suppose (10) is estimated by OLS and let $(\hat{\rho}_\mu, \hat{\mu}^*)$ denote the OLS estimator of (ρ, μ^*) . Since a constant term is included in the regression, no generality is lost by assuming $\mu = 0$. Moreover, by the properties of OLS,

$$T(\hat{\rho}_\mu - 1) = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-1}^\mu v_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^{\mu 2}},$$

where $y_{t-1}^\mu = y_{t-1} - T^{-1} \sum_{s=1}^T y_{s-1}$ for $t = 1, \dots, T+1$.

Proceeding in cases I and II, we have:

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^{\mu 2} = \int_0^1 Y_T^\mu(r)^2 dr$$

and

$$\frac{1}{T} \sum_{t=1}^T y_{t-1}^\mu v_t = \frac{1}{2} \left[Y_T^\mu(1)^2 - Y_T^\mu(0)^2 - \frac{1}{T} \sum_{t=1}^T v_t^2 \right],$$

where $Y_T^\mu(r) = y_{[Tr]}^\mu / \sqrt{T}$ and the inclusion of $Y_T^\mu(0)^2$ reflects the fact that $y_0^\mu \neq 0$ in general. Now,

$$Y_T^\mu(r) = \frac{1}{\sqrt{T}} y_{[Tr]} - \frac{1}{T^{3/2}} \sum_{s=1}^T y_s = Y_T(r) - \frac{1}{T} \sum_{s=1}^T Y_T \left(\frac{s-1}{T} \right) = Y_T(r) - \int_0^1 Y_T(s) ds,$$

where $Y_T(\cdot) = y_{[T\cdot]} / \sqrt{T}$ is identical to $Y_T(\cdot)$ from case II. We therefore know that $Y_T(\cdot) \rightarrow_d \omega W(\cdot)$, where ω and $W(\cdot)$ are defined as in case II. Moreover, $\int_0^1 Y_T(s) ds$ is a continuous function of $Y_T(\cdot)$, so

$$Y_T^\mu(\cdot) \rightarrow_d \omega W^\mu(\cdot),$$

where W^μ is a *demeaned Wiener process* given by $W^\mu(r) = W(r) - \int_0^1 W(s) ds$.

Remark. The demeaned series $\{y_{t-1}^\mu : 1 \leq t \leq T\}$ contains the residuals from a (discrete time) OLS regression of $\{y_{t-1} : 1 \leq t \leq T\}$ on a constant:

$$y_{t-1}^\mu = y_{t-1} - 1 \cdot \tilde{\mu}_y, \quad \tilde{\mu}_y = \frac{1}{T} \sum_{s=1}^T y_{s-1} = \arg \min_{\mu} \sum_{t=1}^T (y_{t-1} - \mu)^2.$$

Analogously, the demeaned Wiener process $\{W^\mu(r) : 0 \leq r \leq 1\}$ can be interpreted as the residual process from a continuous time least squares regression of $\{W(r) : 0 \leq r \leq 1\}$ on a constant:

$$W^\mu(r) = W(r) - 1 \cdot \tilde{\mu}_W, \quad \tilde{\mu}_W = \int_0^1 W(s) ds = \arg \min_{\mu_W} \int_0^1 [W(s) - \mu_W]^2 ds. \quad \blacksquare$$

Since $Y_T^\mu(\cdot) \rightarrow_d \omega W^\mu(\cdot)$ and $T^{-1} \sum_{t=1}^T v_t^2 \rightarrow_p \gamma_v(0)$,

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^{\mu 2} \rightarrow_d \omega^2 \int_0^1 W^\mu(r)^2 dr$$

and

$$\frac{1}{T} \sum_{t=1}^T y_{t-1}^\mu v_t \rightarrow_d \frac{1}{2} \left[\omega^2 W^\mu(1)^2 - \omega^2 W^\mu(0)^2 - \gamma_v(0) \right],$$

implying in particular that if $\rho = 1$, then

$$T(\hat{\rho}_\mu - 1) \rightarrow_d \frac{\frac{1}{2} [W^\mu(1)^2 - W^\mu(0)^2 - 1]}{\int_0^1 W^\mu(r)^2 dr} + \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W^\mu(r)^2 dr}.$$

Although similar in form to the limiting distribution encountered in case II, the limiting distribution of $T(\hat{\rho}_\mu - 1)$ is different from that of $T(\hat{\rho} - 1)$. By example, we have demonstrated that distributional results can change when we change the nature of the deterministic component μ_t . This phenomenon is the rule rather than the exception in models with integrated processes.

By analogy with case II, define

$$\hat{\rho}_\mu^+ = \hat{\rho}_\mu - T \cdot \frac{\frac{1}{2} [\hat{\omega}^2 - \hat{\sigma}^2]}{\sum_{t=1}^T y_{t-1}^{\mu 2}}$$

and

$$t_P^\mu(1) = \frac{\hat{\rho}_\mu^+ - 1}{\hat{\omega} / \sqrt{\sum_{t=1}^T y_{t-1}^{\mu 2}}},$$

where $\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (y_t - \hat{\rho}_\mu y_{t-1} - \hat{\mu}^*)^2$ and $\hat{\omega}^2$ is a consistent estimator of ω^2 (e.g., the Newey-West estimator).

As might be expected, we have

$$T(\hat{\rho}_\mu^+ - 1) \rightarrow_d \frac{\frac{1}{2} [W^\mu(1)^2 - W^\mu(0)^2 - 1]}{\int_0^1 W^\mu(r)^2 dr}$$

and

$$t_P^\mu(1) \rightarrow_d \frac{\frac{1}{2} \left[W^\mu(1)^2 - W^\mu(0)^2 - 1 \right]}{\sqrt{\int_0^1 W^\mu(r)^2 dr}}$$

when $\rho = 1$. The test which rejects for small values of $t_P^\mu(1)$ is called the *Phillips-Perron test*.

The demeaned version of the ADF test, denoted $t_{ADF}^\mu(1)$, is the OLS t -test of $\alpha = 1$ in the regression

$$y_t = \hat{\alpha}y_{t-1} + \hat{\beta}_1\Delta y_{t-1} + \dots + \hat{\beta}_p\Delta y_{t-p} + \hat{\beta}_\mu + \hat{\varepsilon}_t \quad (t = p+1, \dots, T).$$

If $v_t \sim AR(p)$ or if $v_t \sim ARMA$ and the lag length p satisfies $p \rightarrow \infty$ and $p/\sqrt[3]{T} \rightarrow 0$ as $T \rightarrow \infty$,

$$t_{ADF}^\mu(1) \rightarrow_d \frac{\frac{1}{2} \left[W^\mu(1)^2 - W^\mu(0)^2 - 1 \right]}{\sqrt{\int_0^1 W^\mu(r)^2 dr}}$$

when $\rho = 1$.

Percentiles of the limiting distribution of $t_P^\mu(1)$ and $t_{ADF}^\mu(1)$ are reported in Hamilton's (1994) Table B.6. A 5% *ADF test* rejects the unit root null hypothesis if $t_{ADF}^\mu(1) < -2.86$.

Linear trend case. In the *linear trend case* with $\mu_t = \mu + \delta t$, the model can be written as

$$y_t = \rho y_{t-1} + \mu^* + \delta^* t + v_t, \quad (11)$$

where $\mu^* = (1 - \rho)\mu + \rho\delta$ and $\delta^* = (1 - \rho)\delta$. Suppose (11) is estimated by OLS and let $(\hat{\rho}_\tau, \hat{\mu}^*, \hat{\delta}^*)$ denote the OLS estimator of (ρ, μ^*, δ^*) . By the properties of OLS,

$$T(\hat{\rho}_\tau - 1) = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-1}^\tau v_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^{\tau 2}},$$

where y_{t-1}^τ are the OLS residuals from a regression of y_{t-1} on a constant and a trend:

$$y_{t-1} = \tilde{\mu} + \tilde{\delta}t + y_{t-1}^\tau \quad (t = 1, \dots, T).$$

Defining $Y_T^\tau(r) = T^{-1/2}y_{[Tr]}^\tau$, where $y_T^\tau = y_T - \tilde{\mu} - \tilde{\delta}T$, we have

$$\frac{1}{T^2} \sum_{t=1}^T y_{t-1}^{\tau 2} = \int_0^1 Y_T^\tau(r)^2 dr$$

and

$$\frac{1}{T} \sum_{t=1}^T y_{t-1}^\tau v_t = \frac{1}{2} \left[Y_T^\tau(1)^2 - Y_T^\tau(0)^2 - \frac{1}{T} \sum_{t=1}^T v_t^2 \right].$$

It can be shown that

$$Y_T^\tau(\cdot) \rightarrow_d \omega W^\tau(\cdot),$$

where W^τ is a *detrended Wiener process*,

$$W^\tau(r) = W(r) - (4 - 6r) \int_0^1 W(s) ds - (12r - 6) \int_0^1 sW(s) ds,$$

which can be interpreted as the residual process from a continuous time least squares regression of $\{W(r) : 0 \leq r \leq 1\}$ on $\{(1, r)^\tau : 0 \leq r \leq 1\}$.

The limiting distribution of $T(\hat{\rho}_\tau - 1)$ is different from those of $T(\hat{\rho}_\mu - 1)$ and $T(\hat{\rho} - 1)$:

$$T(\hat{\rho}_\tau - 1) \rightarrow_d \frac{\frac{1}{2} [W^\tau(1)^2 - W^\tau(0)^2 - 1]}{\int_0^1 W^\tau(r)^2 dr} + \frac{\frac{1}{2} [\omega^2 - \gamma_v(0)]}{\omega^2 \int_0^1 W^\tau(r)^2 dr}$$

when $\rho = 1$ in (9). Let

$$\hat{\rho}_\tau^+ = \hat{\rho}_\tau - T \cdot \frac{\frac{1}{2} [\hat{\omega}^2 - \hat{\sigma}^2]}{\sum_{t=1}^T y_{t-1}^{\tau 2}}, \quad t_P^\tau(1) = \frac{\hat{\rho}_\tau^+ - 1}{\hat{\omega} / \sqrt{\sum_{t=1}^T y_{t-1}^{\tau 2}}},$$

where $\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T (y_t - \hat{\rho}_\tau y_{t-1} - \hat{\mu}^* - \hat{\delta}^* t)^2$ and $\hat{\omega}^2$ is a consistent estimator of ω^2 . When $\rho = 1$,

$$T(\hat{\rho}_\tau^+ - 1) \rightarrow_d \frac{\frac{1}{2} [W^\tau(1)^2 - W^\tau(0)^2 - 1]}{\int_0^1 W^\tau(r)^2 dr},$$

while the Phillips-Perron test statistic $t_P^\tau(1)$ satisfies

$$t_P^\tau(1) \rightarrow_d \frac{\frac{1}{2} [W^\tau(1)^2 - W^\tau(0)^2 - 1]}{\sqrt{\int_0^1 W^\tau(r)^2 dr}}.$$

The detrended version of the ADF test, denoted $t_{ADF}^\tau(1)$, is the OLS t -test of $\alpha = 1$ in the regression

$$y_t = \hat{\alpha} y_{t-1} + \hat{\beta}_1 \Delta y_{t-1} + \dots + \hat{\beta}_p \Delta y_{t-p} + \hat{\beta}_\mu + \hat{\beta}_\tau \cdot t + \hat{\varepsilon}_t \quad (t = p+1, \dots, T).$$

If $v_t \sim AR(p)$ or if $v_t \sim ARMA$ and the lag length p satisfies $p \rightarrow \infty$ and $T^{-1/3}p \rightarrow 0$ as $T \rightarrow \infty$,

$$t_{ADF}^\tau(1) \rightarrow_d \frac{\frac{1}{2} [W^\tau(1)^2 - W^\tau(0)^2 - 1]}{\sqrt{\int_0^1 W^\tau(r)^2 dr}}$$

when $\rho = 1$.

Percentiles of the limiting distribution of $t_P^\tau(1)$ and $t_{ADF}^\tau(1)$ can be found in Hamilton's (1994) Table B.6. A 5% ADF test rejects the unit root null hypothesis if $t_{ADF}^\tau(1) < -3.41$.

Size and power: Asymptotic theory

The asymptotic properties of the ADF tests coincide with those of the Phillips(-Perron) tests, so it suffices to consider the ADF tests. Suppose y_t is generated by (5). An ADF test with nominal significance level α rejects H_0 if $t_{ADF}(1) < c_\alpha$, where c_α satisfies

$$\Pr \left(\frac{\frac{1}{2} [W(1)^2 - 1]}{\sqrt{\int_0^1 W(r)^2 dr}} < c_\alpha \right) = \alpha.$$

For any T and any ρ , let

$$\pi_T(\rho; \alpha) = \Pr_\rho [t_{ADF}(1) < c_\alpha]$$

denote the probability of rejecting H_0 in a sample of size T as function of ρ . The size of the ADF test is $\pi_T(1; \alpha)$, while the power against the alternative $\rho < 1$ is $\pi_T(\rho; \alpha)$.

Analogously, let $\pi_T^\mu(\rho; \alpha) = \Pr_\rho [t_{ADF}^\mu(1) < c_\alpha^\mu]$ and $\pi_T^\tau(\rho; \alpha) = \Pr_\rho [t_{ADF}^\tau(1) < c_\alpha^\tau]$, where c_α^μ and c_α^τ satisfy

$$\Pr \left(\frac{\frac{1}{2} [W^\mu(1)^2 - W^\mu(0)^2 - 1]}{\sqrt{\int_0^1 W^\mu(r)^2 dr}} < c_\alpha^\mu \right) = \alpha$$

and

$$\Pr \left(\frac{\frac{1}{2} [W^\tau(1)^2 - W^\tau(0)^2 - 1]}{\sqrt{\int_0^1 W^\tau(r)^2 dr}} < c_\alpha^\tau \right) = \alpha,$$

respectively.

For any given T (and α), the power functions $\pi_T(\cdot; \alpha)$, $\pi_T^\mu(\cdot; \alpha)$, and $\pi_T^\tau(\cdot; \alpha)$ depend on $\{\psi_i : i \geq 0\}$ and the distribution of ε_t . By taking limits (as $T \rightarrow \infty$), we obtain approximations to $\pi_T(\cdot; \alpha)$, $\pi_T^\mu(\cdot; \alpha)$ and $\pi_T^\tau(\cdot; \alpha)$ that do not depend on $\{\psi_i : i \geq 0\}$ and/or the distribution of ε_t .

Size. The statistics $t_{ADF}(1)$, $t_{ADF}^\mu(1)$, and $t_{ADF}^\tau(1)$ accommodate serial correlation and are asymptotically pivotal when $\mu_t = 0$, $\mu_t = \mu$ and $\mu_t = \mu + \delta t$, respectively. Therefore,

$$\lim_{T \rightarrow \infty} \pi_T(1; \alpha) = \lim_{T \rightarrow \infty} \pi_T^\mu(1; \alpha) = \lim_{T \rightarrow \infty} \pi_T^\tau(1; \alpha) = \alpha.$$

Consistency. The unit root tests are consistent in the sense that the probability of rejecting H_0 tends to unity as T increases if, in fact, $|\rho|$ is some fixed number less than one:

$$\lim_{T \rightarrow \infty} \pi_T(\rho; \alpha) = \lim_{T \rightarrow \infty} \pi_T^\mu(\rho; \alpha) = \lim_{T \rightarrow \infty} \pi_T^\tau(\rho; \alpha) = 1 \quad \forall |\rho| < 1.$$

Local asymptotic power. A nondegenerate approximation to the power function of a unit root test can be obtained by considering a sequence of local alternatives of the form $\rho = 1 + c/T$, where $c \leq 0$ is fixed as T increases. It turns out that for any $c < 0$,

$$\pi_{\infty}(c; \alpha) = \lim_{T \rightarrow \infty} \pi_T(1 + c/T; \alpha),$$

$$\pi_{\infty}^{\mu}(c; \alpha) = \lim_{T \rightarrow \infty} \pi_T^{\mu}(1 + c/T; \alpha),$$

and

$$\pi_{\infty}^{\tau}(c; \alpha) = \lim_{T \rightarrow \infty} \pi_T^{\tau}(1 + c/T; \alpha)$$

all exist and lie (strictly) between α and one.

For simplicity, consider $\pi_{\infty}(c; \alpha)$. That function is obtained by (i) characterizing the limiting behavior of $t_{ADF}(1)$ under a sequence of local alternatives of the form $\rho = 1 + c/T$, $c < 0$ and (ii) evaluating the distribution function of the limiting random variable at c_{α} . When y_t is generated by (5) with $\rho = 1 + c/T$, it can be shown that

$$Y_T(r) = c \int_0^r \exp(c(r-s)) \xi_T(s) ds + \xi_T(r) + Y_T^R(r),$$

where $Y_T(r) = y_{\lfloor Tr \rfloor} / \sqrt{T}$, $\xi_T(r) = T^{-1/2} \sum_{t=1}^{\lfloor Tr \rfloor} v_t$, and $Y_T^R(r)$ is an asymptotically negligible remainder term:

$$\sup_{0 \leq r \leq 1} |Y_T^R(r)| \rightarrow_p 0.$$

Since $\xi_T(\cdot) \rightarrow_d \omega W(\cdot)$, it follows from CMT that

$$Y_T(\cdot) \rightarrow_d \omega W_c(\cdot),$$

where $W_c(\cdot)$ is an *Ornstein-Uhlenbeck (O-U) process* given by

$$W_c(r) = c \int_0^r \exp(c(r-s)) W(s) ds + W(r).$$

Remark. The O-U process can be interpreted as an AR(1) process in continuous time. Specifically, W_c satisfies the stochastic differential equation

$$dW_c(r) = cW_c(r) dr + dW(r)$$

with initial condition $W_c(0) = 0$. Processes of this type are also used in continuous time finance. ■

As a consequence,

$$t_{ADF}(1) \rightarrow_d \frac{\frac{1}{2} [W_c(1)^2 - 1]}{\sqrt{\int_0^1 W_c(r)^2 dr}}$$

and

$$\pi_\infty(c; \alpha) = \Pr \left(\frac{\frac{1}{2} [W_c(1)^2 - 1]}{\sqrt{\int_0^1 W_c(r)^2 dr}} < c_\alpha \right),$$

which can be evaluated numerically.

Analogously, one can show that

$$\pi_\infty^\mu(c; \alpha) = \Pr \left(\frac{\frac{1}{2} [W_c^\mu(1)^2 - W_c^\mu(0)^2 - 1]}{\sqrt{\int_0^1 W_c^\mu(r)^2 dr}} < c_\alpha^\mu \right)$$

and

$$\pi_\infty^\tau(c; \alpha) = \Pr \left(\frac{\frac{1}{2} [W_c^\tau(1)^2 - W_c^\tau(0)^2 - 1]}{\sqrt{\int_0^1 W_c^\tau(r)^2 dr}} < c_\alpha^\tau \right),$$

where W_c^μ (W_c^τ) is a demeaned (detrended) O-U process.

The local asymptotic power functions such as π_∞ , π_∞^μ , and π_∞^τ are plotted in figures 1-3 on pp. 2774-2775 in Stock (1994). Among other things, these power functions can be used to evaluate the effect of detrending. As might be expected, it turns out that power is decreasing as more nuisance parameters are estimated. That is, $\pi_\infty^\tau(c; \alpha) < \pi_\infty^\mu(c; \alpha) < \pi_\infty(c; \alpha)$.

Remarks. (a) Local asymptotic power results can also be used to (i) compare competing unit root tests in terms of power and (ii) evaluate the efficiency properties of a given unit root test. Figures 1-3 on pp. 2774-2775 in Stock (1994) illustrate (i), while an illustration of (ii) can be found in Elliott, Rothenberg and Stock (1996, *Econometrica*, 64, 813-836). It turns out that the ADF tests are inefficient in case 3 (i.e., in the demeaned and detrended cases), but (for all practical purposes) efficient in case 2. Efficient unit root tests have been proposed by Elliott, Rothenberg and Stock (1996). One of these tests, the *DF-GLS test*, is a modified version of the ADF test.

(b) The local asymptotic power calculations proceed under the assumption that $\rho = 1 + c/T$, where $c \leq 0$ is fixed as T increases. This assumption formalizes the idea that a useful asymptotic approximation should deliver the prediction that we cannot reject the unit root hypothesis with certainty even if it is wrong. The parameterization $\rho = 1 + c/T$ is an approximation device which is hoped to deliver good asymptotic approximations to the finite sample distributions of test statistics such as $t_{ADF}(1)$. Often distributional approximations derived in this way are quite precise.

(c) A time series y_t generated by (4), (5), or (9) with $\rho \approx 1$ is said to *nearly integrated*. In some applications, the value of ρ is of interest. A failure to reject the unit root hypothesis is nothing more than that and does not imply $\rho = 1$. Stock (1991, *Journal of Monetary Economics*, 28, 435-460) interprets the failure to reject the unit root hypothesis as evidence in favor of near integration and explains how to construct confidence intervals for ρ using the asymptotic distribution implied by the nesting $\rho = 1 + c/T$.

These confidence intervals do *not* have end points $\hat{\rho} \pm 1.96$ times a standard error, but are nonetheless relatively easy to construct. Using the Nelson-Plosser data set, he finds that confidence intervals are rather wide for most U.S. macroeconomic time series. ■

In order to conduct a unit root test in practice, we will typically have to choose between $t_{ADF}^{\mu}(1)$ and $t_{ADF}^{\tau}(1)$. When doing so we face a conventional efficiency vs. robustness trade-off: the local asymptotic power of $t_{ADF}^{\mu}(1)$ is higher than that of $t_{ADF}^{\tau}(1)$ in the demeaned case, but only the latter test is applicable in the detrended case. The validity of the latter claim is not as obvious as one might expect. Since $\delta^* = 0$ in (11) when $\rho = 1$ in (9), the limiting null distribution of $t_{ADF}^{\mu}(1)$ is

$$\frac{\frac{1}{2} \left[W^{\mu}(1)^2 - W^{\mu}(0)^2 - 1 \right]}{\sqrt{\int_0^1 W^{\mu}(r)^2 dr}}$$

even in the linear trend case. On the other hand, the test based on $t_{ADF}^{\mu}(1)$ is inconsistent when $\mu_t = \mu + \delta t$ with $\delta \neq 0$ (for details, see section 3.2.5 of Stock (1994)).

Size and power: Finite-sample evidence

In finite samples, the size properties of the unit root tests depend on the way in which autocorrelation is accommodated. Specifically, the size of the ADF tests can depend rather crucially on the value of the lag length p . Similarly, the performance of the Phillips(-Perron) tests depends on the value of the bandwidth parameter b used in the Newey-West estimator. Numerous Monte Carlo experiments have been performed and it seems to be the case that data-dependent selection rules for p (by means of BIC or some variant thereof) and b (by means of the “plug-in” bandwidth formulas provided in Andrews (1991) are better than deterministic rules. In addition, tests of the ADF variety often do better than the Phillips(-Perron) tests in terms of size. Finite-sample results concerning (size-adjusted) power often bear out the predictions from local asymptotic power studies. In particular, tests such as the DF-GLS test also tend to outperform their competitors in small samples.

In practice, it turns out to be very hard to discriminate between series with unit roots highly persistent series with values of ρ close to, but less than, unity. Indeed, a common complaint against unit root tests is that in samples of the size encountered in practice they have rather modest power against alternatives that may be plausible from an economic point of view. There is no easy solution to this problem. While increasing the “sample size” certainly helps, it is important to realize that in terms of power, it is the sample size measured in calendar units (as opposed to the number of observations in the sample) that matters so that collecting data at higher frequency does not solve the problem (although by doing that one might be able to get more precise estimates of the nuisance parameters characterizing the short-run behavior of the series). One way to try to control the type II error is to flip the hypotheses and complement the unit root test by conducting a test of $H_0 : d = 0$ vs. $H_1 : d = 1$. Such tests are called stationarity tests and will be considered next.

Stationarity Tests

Suppose y_t is generated by

$$y_t = \mu_t + u_t,$$

where μ_t is a deterministic component and u_t is a zero-mean scalar time series. Stationarity tests are tests of $H_0 : \theta = 1$ vs. $H_1 : \theta < 1$ in a model of the form

$$\begin{aligned} u_1 &= v_1, \\ \Delta u_t &= (1 - \theta L) v_t \quad (t = 2, \dots, T), \end{aligned} \tag{12}$$

where $v_t = \psi(L) \varepsilon_t$, $\psi(L) = \sum_{i=0}^{\infty} \psi_i L^i$ with $\psi(1) \neq 0$, $\sum_{i=1}^{\infty} i |\psi_i| < \infty$, and $\varepsilon_t \sim i.i.d. (0, \sigma^2)$.

Results for the (empirically uninteresting) the zero-mean case are fundamentally different from the results for the constant mean and linear trend cases, so only these latter cases will be considered.

Constant mean case.

When $\mu_t = \mu$ and $\psi(L) = 1$, a test $H_0 : \theta = 1$ vs. $H_1 : \theta < 1$ can be based on

$$L_{\varepsilon}^{\mu} = \frac{\frac{1}{T^2} \sum_{t=1}^T \left(\sum_{s=1}^t y_s^{\mu} \right)^2}{\hat{\sigma}^2},$$

where $y_t^{\mu} = y_t - T^{-1} \sum_{s=1}^T y_s$ and $\hat{\sigma}^2 = T^{-1} \sum_{t=1}^T y_t^{\mu 2}$.

Remark. The test which rejects for large values of L_{ε}^{μ} enjoys certain optimality properties when $\varepsilon_t \sim i.i.d. \mathcal{N}(0, \sigma^2)$. Specifically, it is the locally best invariant test, where invariance is respect to transformations of the form

$$y_t \rightarrow s y_t + m, \quad s > 0, m \in \mathbb{R}.$$

For our purposes, it suffices to realize that the test based on L_{ε}^{μ} is not completely ad hoc. ■

Serial correlation can be accommodated by replacing $\hat{\sigma}^2$, a consistent estimator of

$$\sigma^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \text{Var}(v_t),$$

with a consistent estimator of

$$\omega^2 = \lim_{T \rightarrow \infty} \frac{1}{T} \text{Var} \left(\sum_{t=1}^T v_t \right),$$

the long-run variance of v_t . The resulting test statistic is

$$L^\mu = \frac{\frac{1}{T^2} \sum_{t=1}^T \left(\sum_{s=1}^t y_s^\mu \right)^2}{\hat{\omega}^2},$$

where $\hat{\omega}$ is a consistent estimator of ω^2 such as the Newey-West estimator

$$\hat{\omega}^2 = \hat{\gamma}_v(0) + 2 \sum_{i=1}^b \left(1 - \frac{i}{b}\right) \hat{\gamma}_v(i), \quad \hat{\gamma}_v(i) = \frac{1}{T} \sum_{t=i}^T y_t^\mu y_{t-i}^\mu.$$

The test which rejects for large values of L^μ is known as the *KPSS test*.

The derivation of the limiting distribution of L^μ under H_0 is straightforward. We have:

$$L^\mu = \frac{\frac{1}{T^2} \sum_{t=1}^T \left(\sum_{s=1}^t y_s^\mu \right)^2}{\hat{\omega}^2} = \frac{\int_0^1 Y_T^\mu(r)^2 dr + \frac{1}{T} Y_T^\mu(1)^2}{\hat{\omega}^2},$$

where $Y_T^\mu(r) = T^{-1/2} \sum_{t=1}^{\lfloor Tr \rfloor} y_t^\mu$ for $0 \leq r \leq 1$. Since $u_t = v_t$ under H_0 ,

$$Y_T^\mu(r) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} y_t - \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} \left(\frac{1}{T} \sum_{s=1}^T y_s \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor Tr \rfloor} v_t - \frac{\lfloor Tr \rfloor}{T} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T v_t \right)$$

when $\theta = 1$ in (12). We know from the unit root case that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{\lfloor T \cdot \rfloor} v_t \rightarrow_d \omega W(\cdot),$$

so

$$Y_T^\mu(\cdot) \rightarrow_d \omega B^\mu(\cdot),$$

where B^μ is a *Brownian bridge*:

$$B^\mu(r) = W(r) - r \cdot W(1).$$

As a consequence,

$$L^\mu = \frac{\int_0^1 Y_T^\mu(r)^2 dr + T^{-1} Y_T^\mu(1)^2}{\hat{\omega}^2} \rightarrow_d \int_0^1 B^\mu(r)^2 dr$$

under H_0 . A 5% stationarity test rejects H_0 is $L^\mu > 0.463$.

Linear trend case.

When $\mu_t = \mu + \delta t$, a test $H_0 : \theta = 1$ vs. $H_1 : \theta < 1$ can be based on

$$L^\tau = \frac{\frac{1}{T^2} \sum_{t=1}^T \left(\sum_{s=1}^t y_s^\tau \right)^2}{\hat{\omega}^2},$$

where $\hat{\omega}^2$ is a consistent estimator of $\omega^2 = \lim_{T \rightarrow \infty} T^{-1} \text{Var} \left(\sum_{t=1}^T v_t \right)$ and y_t^τ are the OLS residuals from a regression of y_t on a constant and a trend:

$$y_t = \tilde{\mu} + \tilde{\delta}t + y_t^\tau \quad (t = 1, \dots, T).$$

Using the fact that $T^{-1/2} \sum_{t=1}^{\lfloor T \cdot \rfloor} v_t \rightarrow_d \omega W(\cdot)$, it can be shown that

$$Y_T^\tau(\cdot) \rightarrow_d \omega B^\tau(\cdot)$$

when $\theta = 1$ in (12), where $Y_T^\tau(r) = T^{-1/2} \sum_{t=1}^{\lfloor Tr \rfloor} y_t^\tau$ and B^τ is a *second-level Brownian bridge*:

$$B^\tau(r) = W(r) - rW(1) + 6r(1-r) \left[\frac{1}{2}W(1) - \int_0^1 W(s) ds \right].$$

As a consequence,

$$L^\tau = \frac{\int_0^1 Y_T^\tau(r)^2 dr + T^{-1} Y_T^\tau(1)^2}{\hat{\omega}^2} \rightarrow_d \int_0^1 B^\tau(r)^2 dr$$

under H_0 . A 5% stationarity test rejects H_0 is $L^\tau > 0.146$.