Ch. 21 Univariate Unit Root Process

1 Introduction

Much conventional asymptotic theory for least-squares estimation (e.g. the standard proofs of consistency and asymptotic normality of OLS estimators) assumes stationarity of the explanatory variables, possibly around a deterministic trend. Not all economic times series are stationary, as we saw in Chapter 19, and for many important ones, including aggregate consumption and national income, stationarity is not even a sensible approximation. This chapter describes methods of testing for a unit root in an observed series. Both parametric regression tests and non-parametric adjustments to theses test statistics are considered. To begin with, consider OLS estimation of a $AR(1)$ process,

$$Y_t = \rho Y_{t-1} + u_t,$$

where $u_t \sim i.i.d.(0, \sigma^2)$, and $Y_0 = 0$. The OLS estimator of ρ is given by

$$\hat{\rho}_T = \frac{\sum_{t=1}^T Y_{t-1}Y_t}{\sum_{t=1}^T Y_{t-1}^2} = \left(\sum_{t=1}^T Y_{t-1}^2\right)^{-1}\left(\sum_{t=1}^T Y_{t-1}Y_t\right)$$

and we also have

$$(\hat{\rho}_T - \rho) = \left(\sum_{t=1}^T Y_{t-1}^2\right)^{-1}\left(\sum_{t=1}^T Y_{t-1}u_t\right). \tag{1}$$

When the true value of ρ is less than 1 in absolute value, then Y_t (so does Y_t^2 ?) is a covariance-stationary process. Applying LLN for a covariance process (see 9.19 of Ch. 4) we have

$$\left(\sum_{t=1}^T Y_{t-1}^2\right)/T \xrightarrow{p} E\left[(\sum_{t=1}^T Y_{t-1}^2)/T\right] = \left[\frac{T\cdot\sigma^2}{1-\rho^2}\right]/T = \sigma^2/(1-\rho^2). \tag{2}$$

Since $Y_{t-1}u_t$ is a martingale difference sequence[1] with variance

$$E(Y_{t-1}u_t)^2 = E(Y_{t-1}^2)E(u_t^2) = \sigma^2\frac{\sigma^2}{1-\rho^2}\ (\textit{since } Y_{t-1} \textit{ and } u_t \textit{ are independent})$$

[1] $E(Y_{t-1}u_t|\mathcal{F}_{t-1}) = Y_{t-1}E(u_t|\mathcal{F}_{t-1}) = Y_{t-1}\cdot 0 = 0$. Therefore, $Y_{t-1}u_t$ is a martingale difference sequence.

and

$$\frac{1}{T}\sum_{t=1}^{T}\left[\sigma^2\frac{\sigma^2}{1-\rho^2}\right] \to \sigma^2\frac{\sigma^2}{1-\rho^2}.$$

Applying CLT for a martingale difference sequence to the second term in the righthand side of (1) we have

$$\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{T} Y_{t-1}u_t\right) \xrightarrow{L} N\left(0, \sigma^2\frac{\sigma^2}{1-\rho^2}\right). \tag{3}$$

Substituting (2) and (3) to (1) we have

$$\begin{aligned}
\sqrt{T}(\hat{\rho}_T - \rho) &= \left[(\sum_{t=1}^{T} Y_{t-1}^2)/T\right]^{-1} \cdot \sqrt{T}\left[(\sum_{t=1}^{T} Y_{t-1}u_t)/T\right] && (4)\\
&\xrightarrow{L} \left[\frac{\sigma^2}{1-\rho^2}\right]^{-1} N\left(0, \sigma^2\frac{\sigma^2}{1-\rho^2}\right) && (5)\\
&\equiv N(0, 1-\rho^2). && (6)
\end{aligned}$$

(6) is not valid for the case when $\rho = 1$, however. To see this, recall that the variance of Y_t when $\rho = 1$ is $t\sigma^2$, then the LLN as in (2) would not be valid since if we apply CLT, then it would incur that

$$(\sum_{t=1}^{T} Y_{t-1}^2)/T \xrightarrow{p} E[(\sum_{t=1}^{T} Y_{t-1}^2)/T] = \sigma^2\frac{\sum_{t=1}^{T} t}{T} \to \infty. \tag{7}$$

Similar reason would show that the CLT would not apply for $\frac{1}{\sqrt{T}}(\sum_{t=1}^{T} Y_{t-1}u_t)$. (In stead, $T^{-1}(\sum_{t=1}^{T} Y_{t-1}u_t)$ converges.) To obtain the limiting distribution, as we shall prove in the following, for $(\hat{\rho}_T - \rho)$ in the unit root case, it turn out that we have to multiply $(\hat{\rho}_T - \rho)$ by T rather than by $\sqrt{T}$:

$$T(\hat{\rho}_T - \rho) = \left[\left(\sum_{t=1}^{T} Y_{t-1}^2\right)/T^2\right]^{-1}\left[T^{-1}\left(\sum_{t=1}^{T} Y_{t-1}u_t\right)\right]. \tag{8}$$

Thus, the unit root coefficient converge at a faster rate (T) than a coefficient for stationary regression (which converges at $\sqrt{T}$).

2 Unit Root Asymptotic Theories

The fact that unit root process violate the restriction that certain moments be bounded, e.g. $E|X_t^2|^{1+\delta} < \Delta < \infty$ for some $\delta > 0$ and for all t had made inapplicable of traditional asymptotic theory to the unit root process. In this section, we develop tools to handle the asymptotics of unit root process.

2.1 Random Walks and Wiener Process

Consider a random walk,

$$Y_t = Y_{t-1} + \varepsilon_t,$$

where $Y_0 = 0$ and ε_t is $i.i.d.$ with mean zero and $Var(\varepsilon_t) = \sigma^2 < \infty$.

By repeated substitution we have

$$\begin{aligned} Y_t &= Y_{t-1} + \varepsilon_t = Y_{t-2} + \varepsilon_{t-1} + \varepsilon_t \\ &= Y_0 + \sum_{s=1}^{t} \varepsilon_s \\ &= \sum_{s=1}^{t} \varepsilon_s. \end{aligned}$$

Before we can study the behavior of estimators based on random walks, we must understand in more detail the behavior of the random walk process itself. Thus, consider the random walk $\{Y_t\}$, we can write

$$Y_T = \sum_{t=1}^{T} \varepsilon_t.$$

Rescaling, we have

$$T^{-1/2} Y_T/\sigma = T^{-1/2} \sum_{t=1}^{T} \varepsilon_t/\sigma.$$

(It is important to note here σ^2 should be read as $Var(T^{-1/2}\sum_{t=1}^{T}\varepsilon_t) = E[T^{-1}(\sum \varepsilon_t)^2] = \frac{T\cdot\sigma^2}{T} = \sigma^2$.) According to the Lindeberg-Lévy CLT, we have

$$T^{-1/2} Y_T/\sigma \xrightarrow{L} N(0,1).$$

More generally, we can construct a variable $Y_T(r)$ from the *partial sum* of ε_t

$$Y_T(r) = \sum_{t=s}^{[Tr]} \varepsilon_s,$$

where $0 \leq r \leq 1$ and $[Tr]$ denotes the largest integer that is less than or equal to Tr.

Applying the same rescaling, we define

$$\begin{aligned} W_T(r) &\equiv T^{-1/2} Y_T(r)/\sigma & (9) \\ &= T^{-1/2} \sum_{s=1}^{[Tr]} \varepsilon_s/\sigma. & (10) \end{aligned}$$

Now

$$W_T(r) = T^{-1/2}([Tr])^{1/2} \left\{ ([Tr])^{-1/2} \sum_{s=1}^{[Tr]} \varepsilon_s/\sigma \right\},$$

and for a given r, the term in the brackets $\{\cdot\}$ again obeys the CLT and converges in distribution to $N(0,1)$, whereas $T^{-1/2}([Tr])^{1/2}$ converges to $r^{1/2}$. It follows from standard arguments that $W_T(r)$ converges in distribution to $N(0,r)$.

Suppose that we first choose r_t so that for some integer t, $Tr_t = t$ (i.e., $r_t = t/T$). Then we will consider what happens as r increase to the value $(t+1)/T$. With $r_t = t/T$ we have $[Tr_t] = t$, so

$$W_T(r) = T^{-1/2} \sum_{s=1}^{t} \varepsilon_s/\sigma, \quad r = t/T.$$

For $t/T < r < (t+1)/T$, we still have $[Tr] = t$, so

$$W_T(r) = T^{-1/2} \sum_{s=1}^{t} \varepsilon_s/\sigma, \quad t/T < r < (t+1)/T.$$

That is, $W_T(r)$ is constant for $t/T \leq r < (t+1)/T$. When r hits $(t+1)/T$, we see that $W_T(r)$ jumps to

$$W_T(r) = T^{-1/2} \sum_{s=1}^{t+1} \varepsilon_s/\sigma, \quad r = (t+1)/T.$$

Thus, $W_T(r)$ is a piecewise constant function that jumps to a new values whenever $r = t/T$ for interger t. In this way we are able to concentrate the original horizontal axis of 1 to T to the close interval $[0, 1]$, indexing the observations by r. If, for example, $T = 100$, the original observation Y_{50} will be indexed by $r \in [0.50, 0.51)$, and so on.

We have written $W_T(r)$ so that it is clear that W_T can be considered to be a function of r. Also, because $W_T(r)$ depends on the $\varepsilon_t's$, it is random. Therefore, we can think of $W_T(r)$ as defining a *random function* of r, which we write $W_T(\cdot)$. Just as the CLT provides conditions ensuring that the rescaled random walk $T^{-1/2}Y_T/\sigma$ (which we can now write as $W_T(1)$) converges, as T become large, to a well-defined limiting random variables (the standard normal), the *function central limit theorem* (FCLT) provides conditions ensuring that the random function $W_T(\cdot)$ converge, as T become large, to a well-defined limit random function, say $W(\cdot)$. The word "Functional" in Functional Central Limit theorem appears because this limit is a function of r.

Some further properties of random walk, suitably rescaled, are in the following.

Proposition:

If Y_t is a random walk, then $Y_{t_4} - Y_{t_3}$ is independent of $Y_{t_2} - Y_{t_1}$ for all $t_1 < t_2 < t_3 < t_4$. Consequently, $W_t(r_4) - W_T(r_3)$ is independent of $W_t(r_2) - W_T(r_1)$ for all $[T \cdot r_i]^* = t_i, i = 1, .., 4$.

Proof:

Note that

$$Y_{t_4} - Y_{t_3} = \varepsilon_{t_4} + \varepsilon_{t_4-1} + ... + \varepsilon_{t_3+1},$$
$$Y_{t_2} - Y_{t_1} = \varepsilon_{t_2} + \varepsilon_{t_2-1} + ... + \varepsilon_{t_1+1}.$$

Since $(\varepsilon_{t_2}, \varepsilon_{t_2-1}, ..., \varepsilon_{t_1+1})$ is independent of $(\varepsilon_{t_4}, \varepsilon_{t_4-1}, ..., \varepsilon_{t_3+1})$ it follow that $Y_{t_4} - Y_{t_3}$ and $Y_{t_2} - Y_{t_1}$ are independent.

Consequently,

$$W_T(r_4) - W_T(r_3) = T^{-1/2}(\varepsilon_{t_4} + \varepsilon_{t_4-1} + ... + \varepsilon_{t_3+1})/\sigma$$

is independent of

$$W_T(r_2) - W_T(r_1) = T^{-1/2}(\varepsilon_{t_2} + \varepsilon_{t_2-1} + ... + \varepsilon_{t_1+1})/\sigma.$$

Proposition:
For given $0 \leq a < b \leq 1$, $W_T(b) - W_T(a) \xrightarrow{L} N(0, b-a)$ as $T \to \infty$.

Proof:
By definition

$$\begin{aligned} W_T(b) - W_T(a) &= T^{-1/2} \sum_{t=[Ta]+1}^{[Tb]} \varepsilon_t \\ &= T^{-1/2}([Tb]-[Ta])^{1/2} \times ([Tb]-[Ta])^{-1/2} \sum_{t=[Ta]+1}^{[Tb]} \varepsilon_t. \end{aligned}$$

The last term $([Tb]-[Ta])^{-1/2}\sum_{t=[Ta]+1}^{[Tb]} \varepsilon_t \xrightarrow{L} N(0,1)$ by the CLT, and $T^{-1/2}([Tb]-[Ta])^{1/2} = (([Tb]-[Ta])/T)^{1/2} \to (b-a)^{1/2}$ as $T \to \infty$. Hence $W_T(b)-W_T(a) \xrightarrow{L} N(0, b-a)$.

In words, the random walk has **independent increments** and those increments have a limiting normal distribution, with a variance reflecting the size of the interval $(b-a)$ over which the increment is taken.

It should not be surprising, therefore, that the limit of the sequence of function $W_T(\cdot)$ constructed from the random walk preserves these properties in the limit in an appropriate sense. In fact, these properties form the basis of the definition of the Wiener process.

Definition:
Let $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ be a complete probability space. Then $W: \mathcal{S} \times [0,1] \to \mathbb{R}^1$ is a standard Wiener process if each of $r \in [0,1]$, $W(\cdot, r)$ is $\mathcal{F}$-measurable, and in addition,
(a). The process starts at zero: $\mathcal{P}[W(\cdot, 0) = 0] = 1$.
(b). The increments are independent: if $0 \leq a_0 \leq a_1 ... \leq a_k \leq 1$, then

$W(\cdot, a_i) - W(\cdot, a_{i-1})$ is independent of $W(\cdot, a_j) - W(\cdot, a_{j-1})$, $j = 1, .., k$, $j \neq i$ for all $i = 1, ..., k$.
(c). The increments are normally distributed: for $0 \leq a \leq b \leq 1$, the increment $W(\cdot, b) - W(\cdot, a)$ is distributed as $N(0, b-a)$.

In the definition, we have written $W(\cdot, a)$ explicitness; whenever convenient, however, we will write $W(a)$ instead of $W(\cdot, a)$, analogous to our notation elsewhere. The Wiener process is also called a Brownian motion in honor of Norbert Wiener (1924), who provided the mathematical foundation for the theory of random motions observed and described by nineteenth century botanist Robert Brown in 1827.

2.2 Functional Central Limit Theorems

We earlier defined convergence in law for random variables, and now we need to extend the definition to cover random functions. Let $S(\cdot)$ represent a continuous-time stochastic process with $S(r)$ representing its value at some date r for $r \in [0, 1]$. Suppose, further, that any given realization, $S(\cdot)$ is a continuous function of r with probability 1. For $\{S_T(\cdot)\}_{T=1}^{\infty}$ a sequence of such continuous function, we say that the sequence of probability measure induced by $\{S_T(\cdot)\}_{T=1}^{\infty}$ **weakly converge** to the probability measure induced by $S(\cdot)$, denoted by $S_T(\cdot) \Longrightarrow S(\cdot)$ if all of the following hold:
(a). For any finite collection of k particular dates,

$$0 \leq r_1 < r_2 < ... < r_k \leq 1,$$

the sequence of k-dimensional random vector $\{\mathbf{y}_T\}_{T=1}^{\infty}$ converges in distribution to the vector $\mathbf{y}$, where

$$\mathbf{y}_T \equiv \begin{bmatrix} S_T(r_1) \\ S_T(r_2) \\ . \\ . \\ . \\ S_T(r_k) \end{bmatrix} \quad \mathbf{y} \equiv \begin{bmatrix} S(r_1) \\ S(r_2) \\ . \\ . \\ . \\ S(r_k) \end{bmatrix};$$

(b). For each $\epsilon > 0$, the probability that $S_T(r_1)$ differs from $S_T(r_2)$ for any dates r_1 and r_2 within δ of each other goes to zero uniformly in T as $\delta \to 0$;

(c). $P\{|S_T(0)| > \lambda\} \to 0$ uniformly in T as $\lambda \to \infty$.

This definition applies to sequences of continuous functions, though the function in (9) is a discontinues step function. Fortunately, the discountinuities occur at a countable set of points. Formally, $S_T(\cdot)$ can be replaced with a similar continuous function, interpolating between the steps.

The Function Central Limit Theorem (FCLT) provides conditions under which W_T converges to the standard Wiener process, W. The simplest FCLT is a generalization of the Lindeberg-Lévy CLT, known as Donsker's theorem.

Theorem: (Donsker)
Let ε_t be a sequence of *i.i.d.* random scalars with mean zero. If $\sigma^2 \equiv Var(\varepsilon_t) < \infty$, $\sigma^2 \neq 0$, then $W_T \Longrightarrow W$.

Because pointwise convergence in distribution $W_T(\cdot, r) \xrightarrow{L} W(\cdot, r)$ for each $r \in [0, 1]$ is necessary (but not sufficient) for weak convergence $W_T \Longrightarrow W$, the Lindeberg-Lévy CLT ($W_T(\cdot, 1) \xrightarrow{L} W(\cdot, 1)$) follows immediately from Donsker's theorem. Donsker's theorem is strictly stronger than Lindeberg-Lévy however, as both use identical assumptions, but Donsker's theorem delivers a much stronger conclusion. Donsker called his result an *invariance principle*. Consequently, the FCLT is often referred as an invariance principle.

So far, we have assumed that the sequence ε_t used to construct W_T is *i.i.d.*. Nevertheless, just as we can obtain central limit theorems when ε_t is not necessary *i.i.d.*. In fact, versions of the FCLT hold for each CLT previous given in Chapter 4. See the following section.

In Chapter 4 we saw that if $\{X_T\}_{T=1}^{\infty}$ is a sequence of random variables with $X_T \xrightarrow{L} X$ and if $g : \mathbb{R}^1 \to \mathbb{R}^1$ is a continuous function, then $g(X_T) \xrightarrow{L} g(X)$. A similar result holds for sequences of random functions.

Theorem: (Continuous Mapping Theorem):
If $S_T(\cdot) \Longrightarrow S(\cdot)$ and $g(\cdot)$ is a continuous functional, then $g(S_T(\cdot)) \Longrightarrow g(S(\cdot))$.

In the above theorem, continuity of a functional $g(\cdot)$ means that for any $\varsigma > 0$, there exist a $\delta > 0$ such that if $h(r)$ and $k(r)$ are any continuous bounded functions on $[0, 1]$, $h : [0, 1] \rightarrow \mathbb{R}^1$ and $k : [0, 1] \rightarrow \mathbb{R}^1$, such that $|h(r) - k(r)| < \delta$ for all $r \in [0, 1]$, then $|g(h(\cdot)) - g(k(\cdot))| < \varsigma(\delta)$.

3 Regression with a Unit Root

3.1 Dickey-Fuller Test, Y_t is $AR(1)$ process

Consider the following simple $AR(1)$ process with a unit root,

$$Y_t = \beta Y_{t-1} + u_t, \tag{11}$$

$$\beta = 1 \tag{12}$$

where $Y_0 = 0$ and u_t is $i.i.d.$ with mean zero and variance σ^2.

We consider the three least square regressions:

$$Y_t = \breve{\beta} Y_{t-1} + \breve{u}_t, \tag{13}$$

$$Y_t = \hat{\alpha} + \hat{\beta} Y_{t-1} + \hat{u}_t, \tag{14}$$

and

$$Y_t = \tilde{\alpha} + \tilde{\beta} Y_{t-1} + \tilde{\delta} t + \tilde{u}_t, \tag{15}$$

where $\breve{\beta}, (\hat{\alpha}, \hat{\beta})$, and $(\tilde{\alpha}, \tilde{\beta}, \tilde{\delta})$ are the conventional least-squares regression coefficients. Dickey and Fuller (1979) were concerned with the limiting distribution of the regression in (13), (14), and (15) ($\breve{\beta}, (\hat{\alpha}, \hat{\beta})$, and $(\tilde{\alpha}, \tilde{\beta}, \tilde{\delta})$) under the null hypothesis that the data are generated by (11) and (12).

We first provide the following asymptotic results of the sample moments which are useful to derive the asymptotics of the OLS estimator.

Lemma:

Let u_t be a $i.i.d.$ sequence with mean zero and variance σ^2 and

$$y_t = u_1 + u_2 + ... + u_t \quad for\ t = 1, 2, ..., T, \tag{16}$$

with $y_0 = 0$. Then

(a) $T^{-\frac{1}{2}} \sum_{t=1}^{T} u_t \xrightarrow{L} \sigma W(1)$,

(b) $T^{-2} \sum_{t=1}^{T} Y_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 [W(r)]^2 dr$,

(c) $T^{-\frac{3}{2}}\sum_{t=1}^{T} Y_{t-1} \xrightarrow{L} \sigma \int_0^1 W(r)dr$,

(d) $T^{-1}\sum_{t}^{T} Y_{t-1}u_t \xrightarrow{L} \frac{1}{2}\sigma^2[W(1)^2-1]$,

(e) $T^{-\frac{3}{2}}\sum_{t=1}^{T} tu_t \xrightarrow{L} \sigma[W(1)-\int_0^1 W(r)dr]$,

(f) $T^{-\frac{5}{2}}\sum_{t=1}^{T} tY_{t-1} \xrightarrow{L} \sigma \int_0^1 rW(r)dr$,

(g) $T^{-3}\sum_{t=1}^{T} tY_{t-1}^2 \xrightarrow{L} \sigma^2 \int_0^1 r[W(r)]^2dr$.

A joint weak convergence for the sample moments given above to their respective limits is easily established and will be used below.

Proof:

(a) is a straightforward results of Donsker's Theorem with $r = 1$.

(b) First rewrite $T^{-2}\sum_{t=1}^{T} Y_{t-1}^2$ in term of $W_T(r_{t-1}) \equiv T^{-1/2}Y_{t-1}/\sigma = T^{-1/2}\sum_{s=1}^{t-1} u_s/\sigma$, where $r_{t-1} = (t-1)/T$, so that $T^{-2}\sum_{t=1}^{T} Y_{t-1}^2 = \sigma^2 T^{-1}\sum_{t=1}^{T} W_T(r_{t-1})^2$. Because $W_T(r)$ is constant for $(t-1)/T \leq r < t/T$, we have

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} W_T(r_{t-1})^2 &= \sum_{t=1}^{T}\int_{(t-1)/T}^{t/T} W_T(r)^2dr \\ &= \int_0^1 W_T(r)^2dr. \end{aligned}$$

The continuous mapping theorem applies to $h(W_T) = \int_0^1 W_T(r)^2dr$. It follows that $h(W_T) \Longrightarrow h(W)$, so that $T^{-2}\sum_{t=1}^{T} Y_{t-1}^2 \Longrightarrow \sigma^2\int_0^1 W(r)^2dr$, as claimed.

(c). The proof of item (c) is analogous to that of (b). First rewrite $T^{-3/2}\sum_{t=1}^{T} Y_{t-1}$ in term of $W_T(r_{t-1}) \equiv T^{-1/2}Y_{t-1}/\sigma = T^{-1/2}\sum_{s=1}^{t-1} u_s/\sigma$, where $r_{t-1} = (t-1)/T$, so that $T^{-3/2}\sum_{t=1}^{T} Y_{t-1} = \sigma T^{-1}\sum_{t=1}^{T} W_T(r_{t-1})$. Because $W_T(r)$ is constant for $(t-1)/T \leq r < t/T$, we have

$$\begin{aligned} T^{-1}\sum_{t=1}^{T} W_T(r_{t-1}) &= \sum_{t=1}^{T}\int_{(t-1)/T}^{t/T} W_T(r)dr \\ &= \int_0^1 W_T(r)dr. \end{aligned}$$

The continuous mapping theorem applies to $h(W_T) = \int_0^1 W_T(r)dr$. It follows that $h(W_T) \Longrightarrow h(W)$, so that $T^{-3/2}\sum_{t=1}^T Y_{t-1} \Longrightarrow \sigma \int_0^1 W(r)dr$, as claimed.

(d). For a random walk, $Y_t^2 = (Y_{t-1} + u_t)^2 = Y_{t-1}^2 + 2Y_{t-1}u_t + u_t^2$, implying that $Y_{t-1}u_t = (1/2)\{Y_t^2 - Y_{t-1}^2 - u_t^2\}$ and then $\sum_{t=1}^T Y_{t-1}u_t = (1/2)\{Y_T^2 - Y_0^2\} - (1/2)\sum_{t=1}^T u_t^2$. Recall that $Y_0 = 0$, and thus it is convenient to write $\sum_{t=1}^T Y_{t-1}u_t = \frac{1}{2}Y_T^2 - \frac{1}{2}\sum_{t=1}^T u_t^2$. From items (a)) we know that $T^{-1}Y_T^2$ $(= (T^{-1/2}\sum_{t=1}^T u_s)^2 \xrightarrow{L} \sigma^2 W^2(1)$ and $T^{-1}\sum_{t=1}^T u_t^2 \xrightarrow{p} \sigma^2$ by LLN (Kolmogorov); then, $T^{-1}\sum_{t=1}^T Y_{t-1}u_t \Longrightarrow \frac{1}{2}\sigma^2[W(1)^2 - 1]$.

(e). We first observe that $\sum_{t=1}^T Y_{t-1} = [u_1 + (u_1+u_2) + (u_1+u_2+u_3) + ... + (u_1 + u_2 + u_3 + ... + u_{T-1})] = [T-1)u_1 + (T-2)u_2 + (T-3)u_3 + ... + [T-(T-1)u_{T-1}] = \sum_{t=1}^T (T-t)u_t = \sum_{t=1}^T Tu_t - \sum_{t=1}^T tu_t$, or $\sum_{t=1}^T tu_t = T\sum_{t=1}^T u_t - \sum_{t=1}^T Y_{t-1}$. Therefore, $T^{-\frac{3}{2}}\sum_{t=1}^T tu_t = T^{-\frac{1}{2}}\sum_{t=1}^T u_t - T^{-\frac{3}{2}}\sum_{t=1}^T Y_{t-1}$. By applying the continuous mapping theorem to the joint convergence of items (a) and (c), we have

$$T^{-\frac{3}{2}}\sum_{t=1}^T tu_t \Rightarrow \sigma[W(1) - \int_0^1 W(r)dr].$$

The proofs of item (f) and (g) is analogous to those of (c) and (b). First rewrite $T^{-5/2}\sum_{t=1}^T tY_{t-1}$ in term of $W_T(r_{t-1}) \equiv T^{-1/2}Y_{t-1}/\sigma = T^{-1/2}\sum_{s=1}^{t-1} u_s/\sigma$, where $r_{t-1} = (t-1)/T$, so that $T^{-5/2}\sum_{t=1}^T tY_{t-1} = \sigma T^{-2}\sum_{t=1}^T tW_T(r_{t-1})$. Because $W_T(r)$ is constant for $(t-1)/T \le r < t/T$, we have

$$\begin{aligned} T^{-1}\sum_{t=1}^T (t/T)W_T(r_{t-1}) &= \sum_{t=1}^T \int_{(t-1)/T}^{t/T} rW_T(r)dr \quad for\ r = t/T \\ &= \int_0^1 rW_T(r)dr. \end{aligned}$$

The continuous mapping theorem applies to $h(W_T) = \int_0^1 rW_T(r)dr$. It follows that $h(W_T) \Longrightarrow h(W)$, so that $T^{-5/2}\sum_{t=1}^T tY_{t-1} \Longrightarrow \sigma\int_0^1 rW(r)dr$, as claimed.

We also write $T^{-3}\sum_{t=1}^T tY_{t-1}^2$ in term of $W_T(r_{t-1}) \equiv T^{-1/2}Y_{t-1}/\sigma = T^{-1/2}\sum_{s=1}^{t-1} u_s/\sigma$, where $r_{t-1} = (t-1)/T$, so that $T^{-3}\sum_{t=1}^T tY_{t-1}^2 = \sigma^2 T^{-2}\sum_{t=1}^T tW_T(r_{t-1})^2$. Because $W_T(r)$ is constant for $(t-1)/T \le r < t/T$, we have

$$\begin{aligned} T^{-1}\sum_{t=1}^T (t/T)W_T(r_{t-1})^2 &= \sum_{t=1}^T \int_{(t-1)/T}^{t/T} rW_T(r)^2 dr \quad for \\ &= \int_0^1 rW_T(r)^2 dr. \end{aligned}$$

The continuous mapping theorem applies to $h(W_T) = \int_0^1 rW_T(r)^2 dr$. It follows that $h(W_T) \Longrightarrow h(W)$, so that $T^{-3}\sum_{t=1}^T tY_{t-1}^2 \Longrightarrow \sigma^2 \int_0^1 rW(r)^2 dr$. This completes the proofs of this lemma.

3.1.1 No Constant Term or Time Trend in the Regression; True Process Is a Random Walk

We first consider the case that no constant term or time trend in the regression model, but true process is a random walk. The asymptotic distributions of OLS unit root coefficient estimator and t-ratio test statistics are in the following.

Theorem 1:
Let the data Y_t be generated by (11) and (12); then as $T \to \infty$, for the regression model (13),

$$T(\breve{\beta}_T - 1) \xrightarrow{L} \frac{1/2\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2 dr}$$

and

$$\breve{t} = \frac{(\breve{\beta}_T - 1)}{\breve{\sigma}_{\breve{\beta}_T}} \xrightarrow{L} \frac{1/2\{[W(1)]^2 - 1\}}{\{\int_0^1 [W(r)]^2 dr\}^{1/2}},$$

where $\breve{\sigma}^2_{\breve{\beta}_T} = [s_T^2 \div \sum_{t=1}^T Y_{t-1}^2]^{1/2}$ and s_T^2 denote the OLS estimate of the disturbance variance:

$$s_T^2 = \sum_{t=1}^T (Y_t - \breve{\beta}_T Y_{t-1})^2/(T-1).$$

Proof:
Since the deviation of the OLS estimate from the true value is characterized by

$$T(\breve{\beta}_T - 1) = \frac{T^{-1}\sum_{t=1}^T Y_{t-1}u_t}{T^{-2}\sum_{t=1}^T Y_{t-1}^2}, \tag{17}$$

which is a continuous function of Lemma 1(b) and 1(d), it follows that under the null hypothesis that $\beta = 1$, the OLS estimator $\breve{\beta}$ is characterized by

$$T(\breve{\beta}_T - 1) \Rightarrow \frac{1/2\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2 dr}. \tag{18}$$

To prove second part of this theorem, We first show the consistency of s_T^2. Notice that the population disturbance sum of squares can be written

$$\begin{aligned}
& (\mathbf{y}_T - \mathbf{y}_{T-1}\beta)'(\mathbf{y}_T - \mathbf{y}_{T-1}\beta) \\
= \; & (\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta} + \mathbf{y}_{T-1}\breve{\beta} - \mathbf{y}_{T-1}\beta)'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta} + \mathbf{y}_{T-1}\breve{\beta} - \mathbf{y}_{T-1}\beta) \\
= \; & (\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) + (\mathbf{y}_{T-1}\breve{\beta} - \mathbf{y}_{T-1}\beta)'(\mathbf{y}_{T-1}\breve{\beta} - \mathbf{y}_{T-1}\beta),
\end{aligned}$$

where $\mathbf{y}_T = [Y_1 \; Y_2 \; ... \; Y_T]'$ and the cross-product have vanished, since

$$(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'\mathbf{y}_{T-1}(\breve{\beta} - \beta) = 0,$$

by the OLS orthogonality condition ($\mathbf{X}'\mathbf{e} = \mathbf{0}$). Dividing last equation by T,

$$\begin{aligned}
& (1/T)(\mathbf{y}_T - \mathbf{y}_{T-1}\beta)'(\mathbf{y}_T - \mathbf{y}_{T-1}\beta) \\
= \; & (1/T)(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) + (\breve{\beta} - \beta)'[(\mathbf{y}'_{T-1}\mathbf{y}_{T-1})/T](\breve{\beta} - \beta)
\end{aligned}$$

or

$$(1/T)(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) \tag{19}$$

$$= (1/T)(\sum_{t=1}^T u_t^2) - T^{1/2}(\breve{\beta} - \beta)'[(\mathbf{y}'_{T-1}\mathbf{y}_{T-1})/T^2]T^{1/2}(\breve{\beta} - \beta). \tag{20}$$

Now $(1/T)(\sum_{t=1}^T u_t^2) \xrightarrow{p} E(u_t^2) \equiv \sigma^2$ by LLN for $i.i.d.$ sequence, $T^{1/2}(\breve{\beta} - \beta) \to 0$ and $(\mathbf{y}'_{T-1}\mathbf{y}_{T-1})/T^2 \Longrightarrow \sigma^2 \int_0^1 [W(r)]^2 dr$ from (18) and Lemma 1(b), respectively. We thus have

$$T^{1/2}(\breve{\beta} - \beta)'[(\mathbf{y}'_{T-1}\mathbf{y}_{T-1})/T^2]T^{1/2}(\breve{\beta} - \beta) \xrightarrow{p} 0'\sigma^2 \int_0^1 [W(r)]^2 dr \cdot 0 = 0.$$

Substituting these results into (20) we have

$$(1/T)(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) \xrightarrow{p} \sigma^2.$$

The OLS disturbance's variance estimator

$$s_T^2 = [1/(T-1)](\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) \tag{21}$$

$$= [T/(T-1)](1/T)(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta})'(\mathbf{y}_T - \mathbf{y}_{T-1}\breve{\beta}) \tag{22}$$

$$\xrightarrow{p} 1 \cdot \sigma^2 = \sigma^2, \tag{23}$$

therefore is a consistent estimator.

Finally, we can express the t statistics alternatively as

$$\breve{t}_T = T(\breve{\beta}_T - 1)\left\{T^{-2}\sum_{t=1}^{T} Y_{t-1}^2\right\}^{1/2} \div (s_T^2)^{1/2}$$

or

$$\breve{t}_T = \frac{T^{-1}\sum_{t=1}^{T} Y_{t-1}u_t}{\left\{T^{-2}\sum_{t=1}^{T} Y_{t-1}^2\right\}^{1/2}(s_T^2)^{1/2}},$$

which is a continuous function of Lemma 1(b) and 1(d), it follows that under the null hypothesis that $\beta = 1$, the asymptotic distribution of OLS t statistics is characterized by

$$\breve{t}_T \xrightarrow{L} \frac{1/2\sigma^2\{[W(1)]^2 - 1\}}{\left\{\sigma^2\int_0^1 [W(r)]^2 dr\right\}^{1/2}(\sigma^2)^{1/2}} = \frac{1/2\{[W(1)]^2 - 1\}}{\left\{\int_0^1 [W(r)]^2 dr\right\}^{1/2}}. \tag{24}$$

This complete the proof of this Theorem.

Statistical tables for the distributions of (18) and (24) for various sample size T are reported in the section labeled Case 1 in Table B.5 and B.6, respectively. This finite sample result assume Gaussian innovations.

3.1.2 Constant Term but No Time Trend included in the Regression; True Process Is a Random Walk

We next consider the case that a constant term is added in the regression model, but true process is a random walk. The asymptotic distributions of OLS unit root coefficient estimator and t-ratio test statistics are in the following.

Theorem 2:

Let the data Y_t be generated by (11) and (12); then as $T \to \infty$, for the regression model (14),

$$T(\hat{\beta}_T - 1) \xrightarrow{L} \frac{1/2\{[W(1)]^2 - 1\} - W(1)\cdot\int_0^1 W(r)dr}{\int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r)dr\right]^2} \tag{25}$$

and

$$\hat{t} = \frac{(\hat{\beta}_T - 1)}{\hat{\sigma}_{\hat{\beta}_T}} \xrightarrow{L} \frac{1/2\{[W(1)]^2 - 1\} - W(1) \cdot \int_0^1 W(r)dr}{\left\{ \int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r)dr \right]^2 \right\}^{1/2}}, \tag{26}$$

where $\hat{\sigma}^2_{\hat{\beta}_T} = s_T^2 \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} T & \sum Y_{t-1} \\ \sum Y_{t-1} & \sum Y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and s_T^2 denote the *OLS* estimate of the disturbance variance:

$$s_T^2 = \sum_{t=1}^{T} (Y_t - \hat{\alpha}_T - \hat{\beta}_T Y_{t-1})^2/(T-2).$$

Proof:
The proof of this theorem is analogous to that of Theorem 1 and is omitted here.

Statistical tables for the distributions of (25) and (26) for various sample size T are reported in the section labeled Case 2 in Table B.5 and B.6, respectively. This finite sample result assume Gaussian innovations.

These statistics test the null hypothesis that $\beta = 1$. However, a maintained assumption on which the derivation of Theorem 2 was based on is that the true value of α is zero. Thus, it might seem more natural to test for a unit root in this specification by testing the **joint hypothesis** that $\alpha = 0$ and $\beta = 1$. Dickey and Fuller (1981) derive the limiting distribution of the likelihood ratio test for the hypothesis that $(\alpha, \beta) = (0, 1)$ and used Monte Carlo to calculate the distribution of the *OLS F* test of this hypothesis. Their values are reported under the heading Case 2 in table B.7.

3.1.3 Constant Term and Time Trend Included in the Regression; True Process Is a Random Walk With or Without Drift

We finally in the section consider the case that a constant term and a linear trend are added in the regression model, but true process is a random walk with a drift. However, the true value of this drift **turns out not to matter** for the asymptotic distributions of *OLS* unit root coefficient estimator and t-ratio test statistics in this case.

Theorem 3:
Let the data Y_t be generated by (11) and (12); then as $T \to \infty$, for the regression model (15),

$$T(\tilde{\beta}-1) \Rightarrow \frac{1/2\{[W(1)-2\int_0^1 W(r)dr][W(1)+6\int_0^1 W(r)dr - 12\int_0^1 rW(r)dr]-1\}}{\int_0^1 [W(r)]^2 dr - 4[\int_0^1 W(r)dr]^2 + 12\int_0^1 W(r)dr\int_0^1 rW(r)dr - 12[\int_0^1 rW(r)dr]^2}$$

and

$$\tilde{t} = \frac{(\tilde{\beta}_T - 1)}{\tilde{\sigma}_{\tilde{\beta}_T}} \Rightarrow T(\tilde{\beta}-1) \div \sqrt{Q},$$

where $\tilde{\sigma}^2_{\tilde{\beta}_T} = s_T^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} T & \sum_{t=1}^T \xi_{t-1} & \sum_{t=1}^T t \\ \sum_{t=1}^T \xi_{t-1} & \sum_{t=1}^T \xi_{t-1}^2 & \sum_{t=1}^T t\xi_{t-1} \\ \sum_{t=1}^T t & \sum_{t=1}^T t\xi_{t-1} & \sum_{t=1}^T t^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},$

$\xi_t = Y_t - \alpha t$, s_T^2 denote the OLS estimate of the disturbance variance:

$$s_T^2 = \sum_{t=1}^T (Y_t - \tilde{\alpha} - \tilde{\beta}_T Y_{t-1} - \tilde{\delta} t)^2/(T-3),$$

and

$$Q \equiv \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr & 1/2 \\ \int W(r)dr & \int [W(r)]^2 dr & \int rW(r)dr \\ 1/2 & \int rW(r)dr & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Proof:
(a). Let the data generating process be

$$Y_t = \alpha + Y_{t-1} + u_t,$$

and the regression model be

$$Y_t = \alpha + \beta Y_{t-1} + \delta t + u_t. \tag{27}$$

Note that the regression model of (27) can be equivalently rewritten as

$$Y_t = (1-\beta)\alpha + \beta(Y_{t-1} - \alpha(t-1)) + (\delta + \beta\alpha)t + u_t,$$

$$\equiv \alpha^* + \beta^*\xi_{t-1} + \delta^* t + u_t, \tag{28}$$

where $\alpha^* = (1-\beta)\alpha$, $\beta^* = \beta$, $\delta^* = \delta + \beta\alpha$, and $\xi_{t-1} = Y_{t-1} - \alpha(t-1)$. Moreover, under the null hypothesis that $\beta = 1$ and $\delta = 0$,

$$\xi_t = Y_0 + u_1 + u_2 + ... + u_t;$$

that is, ξ_t is the random walk as described in Lemma 1. Under the maintained hypothesis, $\alpha = \alpha_0$, $\beta = 1$, and $\delta = 0$, which in (28) means that $\alpha^* = 0$, $\beta^* = 1$ and $\delta^* = \alpha_0$. The deviation of the OLS estimate from these true values is given by

$$\begin{bmatrix} \tilde{\alpha}^* \\ \tilde{\beta} - 1 \\ \tilde{\delta}^* - \alpha_0 \end{bmatrix} = \begin{bmatrix} T & \sum_{t=1}^T \xi_{t-1} & \sum_{t=1}^T t \\ \sum_{t=1}^T \xi_{t-1} & \sum_{t=1}^T \xi_{t-1}^2 & \sum_{t=1}^T t\xi_{t-1} \\ \sum_{t=1}^T t & \sum_{t=1}^T t\xi_{t-1} & \sum_{t=1}^T t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^T u_t \\ \sum_{t=1}^T \xi_{t-1}u_t \\ \sum_{t=1}^T tu_t \end{bmatrix}, \tag{29}$$

or in shorthand as

$$\mathbf{C} = \mathbf{A}^{-1}\mathbf{f}.$$

From Lemma 1, the order of probability of the individual terms in (29) is as follows,

$$\begin{bmatrix} \tilde{\alpha}^* \\ \tilde{\beta} - 1 \\ \tilde{\delta}^* - \alpha_0 \end{bmatrix} = \begin{bmatrix} O_p(T) & O_p(T^{\frac{3}{2}}) & O_p(T^2) \\ O_p(T^{\frac{3}{2}}) & O_p(T^2) & O_p(T^{\frac{5}{2}}) \\ O_p(T^2) & O_p(T^{\frac{5}{2}}) & O_p(T^3) \end{bmatrix}^{-1} \begin{bmatrix} O_p(T^{\frac{1}{2}}) \\ O_p(T^1) \\ O_p(T^{\frac{3}{2}}) \end{bmatrix}.$$

We define a rescaling matrices,

$$\Upsilon_T = \begin{bmatrix} T^{\frac{1}{2}} & 0 & 0 \\ 0 & T^1 & 0 \\ 0 & 0 & T^{\frac{3}{2}} \end{bmatrix}.$$

Multiplying the rescaling matrices on (29), we get

$$\Upsilon_T\mathbf{C} = \Upsilon_T\mathbf{A}^{-1}\Upsilon_T\Upsilon_T^{-1}\mathbf{f} = \left[\Upsilon_T^{-1}\mathbf{A}\Upsilon_T^{-1}\right]^{-1}\Upsilon_T^{-1}\mathbf{f} \tag{30}$$

Substituting the results of Lemma A.1 to (30), we establish that

$$\tilde{\mathbf{b}}_{\mathbf{1}} \Rightarrow \mathbf{Q}^{-1}\mathbf{h}_{\mathbf{1}}, \tag{31}$$

where

$$\begin{aligned}
\tilde{\mathbf{b}}_{\mathbf{1}} &\equiv \begin{bmatrix} T^{1/2}\tilde{\alpha}^* \\ T(\tilde{\beta}-1) \\ T^{3/2}(\tilde{\delta}^*-\alpha_0) \end{bmatrix}, \\
\mathbf{Q} &\equiv \begin{bmatrix} 1 & \int W(r)dr & 1/2 \\ \int W(r)dr & \int [W(r)]^2 dr & \int rW(r)dr \\ 1/2 & \int rW(r)dr & 1/3 \end{bmatrix} \\
\mathbf{h}_{\mathbf{1}} &\equiv \begin{bmatrix} \sigma W(1) \\ \frac{1}{2}\sigma^2\{[W(1)]^2-1\} \\ \sigma[W(1)-\int W(r)dr] \end{bmatrix}.
\end{aligned}$$

Thus, the asymptotic distribution of $T(\tilde{\beta}-1)$ is given by the middle row of (31), that is,

$$T(\tilde{\beta}-1) \Rightarrow \frac{1/2\{[W(1)-2\int_0^1 W(r)dr][W(1)+6\int_0^1 W(r)dr-12\int_0^1 rW(r)dr]-1\}}{\int_0^1 [W(r)]^2 dr - 4[\int_0^1 W(r)dr]^2 + 12\int_0^1 W(r)dr\int_0^1 rW(r)dr - 12[\int_0^1 rW(r)dr]^2}.$$

Note that this distribution does not depend on either α or σ; in particular, it doesn't matter whether or not the true value of α is zero. .

(b). The asymptotic distribution of the *OLS* t statistics can be founded using similar calculation as those in (23). Notice that

$$
\begin{aligned}
T^2 \cdot \tilde{\sigma}^2_{\tilde{\beta}_T} &= T^2 \cdot s_T^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} T & \sum_{t=1}^{T} \xi_{t-1} & \sum_{t=1}^{T} t \\ \sum_{t=1}^{T} \xi_{t-1} & \sum_{t=1}^{T} \xi_{t-1}^2 & \sum_{t=1}^{T} t\xi_{t-1} \\ \sum_{t=1}^{T} t & \sum_{t=1}^{T} t\xi_{t-1} & \sum_{t=1}^{T} t^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= s_T^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix} \\
&\quad \times \begin{bmatrix} T & \sum_{t=1}^{T} \xi_{t-1} & \sum_{t=1}^{T} t \\ \sum_{t=1}^{T} \xi_{t-1} & \sum_{t=1}^{T} \xi_{t-1}^2 & \sum_{t=1}^{T} t\xi_{t-1} \\ \sum_{t=1}^{T} t & \sum_{t=1}^{T} t\xi_{t-1} & \sum_{t=1}^{T} t^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= s_T^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & T^{-3/2} \sum_{t=1}^{T} \xi_{t-1} & T^{-2} \sum_{t=1}^{T} t \\ \sum_{t=1}^{T} T^{-3/2} \xi_{t-1} & \sum_{t=1}^{T} T^{-2} \xi_{t-1}^2 & \sum_{t=1}^{T} T^{-5/2} t\xi_{t-1} \\ \sum_{t=1}^{T} T^{-2} t & \sum_{t=1}^{T} T^{-5/2} t\xi_{t-1} & \sum_{t=1}^{T} T^{-3} t^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&\xrightarrow{L} \sigma^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \sigma \int W(r)dr & 1/2 \\ \sigma \int W(r)dr & \sigma^2 \int [W(r)]^2 dr & \sigma \int rW(r)dr \\ 1/2 & \sigma \int rW(r)dr & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \\
&\quad \times \begin{bmatrix} 1 & \int W(r)dr & 1/2 \\ \int W(r)dr & \int [W(r)]^2 dr & \int rW(r)dr \\ 1/2 & \int rW(r)dr & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)dr & 1/2 \\ \int W(r)dr & \int [W(r)]^2 dr & \int rW(r)dr \\ 1/2 & \int rW(r)dr & 1/3 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&\equiv Q.
\end{aligned}
$$

From this result it follows that the asymptotic distribution of the OLS t test of the hypothesis that $\beta = 1$ is given by

$$
\tilde{t}_T = T(\tilde{\beta}_T - 1) \div (T^2 \cdot \tilde{\sigma}^2_{\tilde{\beta}_T})^{1/2} \xrightarrow{p} T(\tilde{\beta}_T - 1) \div \sqrt{Q}.
$$

This completes the proofs of Theorem 3.

Again, this distribution does not dependent on α or σ. The small-sample distribution of the OLS t statistics under the assumption of Gaussian disturbance is presented under case 4 in Table B.6. If this distribution were truly t, then a value below -2.0 would be sufficient to reject the null hypothesis. However, Table B.6 reveals that, because of the nonstandard distribution, the t statistic must below -3.4 before the null hypothesis of a unit root could be rejected.

The assumption that the true value of δ is equal to zero is again an auxiliary hypothesis upon which the asymptotic properties of the test depend. Thus, as in section 3.1.2, it is natural to consider the OLS F test of the joint null hypothesis that $\delta = 0$ and $\beta = 1$. Though this F test statistic is calculated in the usual way, its asymptotic distribution is nonstandard, and the calculated F statistic should be compared with the value under case 4 in Table B.7.

Remark:
To derive the asymptotic distribution in this chapter, it is useful to use the following result:

$$\begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} = \begin{bmatrix} \alpha^2 a & \alpha\beta b \\ \alpha\beta c & \beta^2 d \end{bmatrix}.$$

The unit root tests are not confining themselves to the simple $AR(1)$ process as the original Dickey and Fuller (1979). There are two ways to generalize the unit root process. The first one is the parametric model that assume Y_t in (11) is a $AR(p)$ process or u_t is a $ARMA(p,q)$ (Said-Dickey, 1984) process. The second one is a non-parametric model that assume u_t satisfy certain memory and moment constrains (Phillips-Perron, 1987). These modifications of unit root models are discussed in the following.

3.2 Augmented Dickey-Fuller Test, Y_t is $AR(p)$ process

Instead of (11) and (12), suppose that the data were generated from an $AR(p)$ process,

$$(1-\phi_1 L-\phi_2 L^2-...-\phi_p L^p)Y_t=\varepsilon_t, \tag{32}$$

where ε_t is $i.i.d.$ sequence with zero mean and variance σ^2 and finite fourth moment. It is helpful to write the autoregression (32) in a slightly different form. To do so, define

$$\begin{aligned}\beta &\equiv \phi_1+\phi_2+...+\phi_p\\ \zeta_j &\equiv -[\phi_{j+1}+\phi_{j+2}+...+\phi_p] \quad for\ j=1,2,...,p-1.\end{aligned}$$

Notice that for any value of $\phi_1,\phi_2,...,\phi_p$, the following polynomials in L are equivalent:

$$\begin{aligned}& (1-\beta L)-(\zeta_1 L+\zeta_2 L^2+...+\zeta_{p-1}L^{p-1})(1-L)\\ =\ & 1-\beta L-\zeta_1 L+\zeta_1 L^2-\zeta_2 L+\zeta_2 L^3-...-\zeta_{p-1}L^{p-1}+\zeta_{p-1}L^p\\ =\ & 1-(\beta+\zeta_1)L-(\zeta_2-\zeta_1)L^2-(\zeta_3-\zeta_2)L^3-...-(\zeta_{p-1}-\zeta_{p-2})L^{p-1}-(-\zeta_{p-1})L^p\\ =\ & 1-[(\phi_1+\phi_2+...+\phi_p)-(\phi_2+...+\phi_p)]L-[-(\phi_3+\phi_4+...+\phi_p)+(\phi_2+...+\phi_p)]L^2\\ & -...-[-(\phi_p)+(\phi_{p-1}+\phi_p)]L^{p-1}-(\phi_p)L^p\\ =\ & 1-\phi_1 L-\phi_2 L^2-...-\phi_p L^p.\end{aligned}$$

Thus, the autoregression (32) can be equivalently be written

$$\{(1-\beta L)-(\zeta_1 L+\zeta_2 L^2+...+\zeta_{p-1}L^{p-1})(1-L)\}Y_t=\varepsilon_t \tag{33}$$

or

$$Y_t=\beta Y_{t-1}+\zeta_1\triangle Y_{t-1}+\zeta_2\triangle Y_{t-2}+...+\zeta_{p-1}\triangle Y_{t-p+1}+\varepsilon_t. \tag{34}$$

Example:

In the case of $p=3$, (34) is

$$\begin{aligned}Y_t &= \phi_1 Y_{t-1}+\phi_2 Y_{t-2}+\phi_3 Y_{t-3}+\varepsilon_t\\ &= (\phi_1+\phi_2+\phi_3)Y_{t-1}-(\phi_2+\phi_3)[Y_{t-1}-Y_{t-2}]-(\phi_3)[Y_{t-2}-Y_{t-3}]\\ &= \beta Y_{t-1}+\zeta_1\triangle Y_{t-1}+\zeta_2\triangle Y_{t-2}.\end{aligned}$$

Suppose that the process that generated Y_t contain a single unit root; that is suppose one root of

$$(1 - \phi_1 z - \phi_2 z^2 - ... - \phi_p z^p) = 0 \tag{35}$$

is unity, then

$$1 - \phi_1 - \phi_2 - ... - \phi_p = 0, \tag{36}$$

and all other root of (35) are outside the unit circle. Notice that (36) implies that $\beta = 1$ in (34). Moreover, when $\beta = 1$, it would imply that

$$1 - \phi_1 L - \phi_2 L^2 - ... - \phi_p L^p = (1 - L)(1 - \zeta_1 L - \zeta_2 L^2 - ... - \zeta_{p-1} L^{p-1})$$

and all the roots of $(1 - \zeta_1 L - \zeta_2 L^2 - ... - \zeta_{p-1} L^{p-1}) = 0$ would lie outside the unit circle. Under the null hypothesis that $\beta = 1$, expression (34) could then be written as

$$\triangle Y_t = \zeta_1 \triangle Y_{t-1} + \zeta_2 \triangle Y_{t-2} + ... + \zeta_{p-1} \triangle Y_{t-p+1} + \varepsilon_t. \tag{37}$$

or

$$\triangle Y_t = u_t \tag{38}$$

where

$$u_t = \psi(L)\varepsilon_t = (1 - \zeta_1 L - \zeta_2 L^2 - ... - \zeta_{p-1} L^{p-1})^{-1} \varepsilon_t.$$

That is, we may express this $AR(p)$ with an unit root process as the $AR(1)$ with an unit root process as (38) but with serially correlated u_t.

Dickey and Fuller (1979) propose a test of unit root in this $AR(p)$ model and is known as the Augmented Dickey-Fuller (ADF) test.

3.2.1 Constant Term but No Time Trend included in the Regression; True Process Is a Autoregressive with no Drift

Assume that the initial sample is of size $T + p$, with observation numbered $\{Y_{-p+1}, Y_{-p+2}, ..., Y_T\}$, and conditional on the first p observations. We are interested in the properties of OLS estimation of

$$\begin{aligned} Y_t &= \zeta_1 \triangle Y_{t-1} + \zeta_2 \triangle Y_{t-2} + ... + \zeta_{p-1} \triangle Y_{t-p+1} + \alpha + \beta Y_{t-1} + \varepsilon_t & (39) \\ &= \mathbf{x}_t' \boldsymbol{\beta} + \varepsilon_t & (40) \end{aligned}$$

under the null hypothesis that $\alpha = 0$ and $\beta = 1$, i.e. the DGP is

$$\triangle Y_t = \zeta_1 \triangle Y_{t-1} + \zeta_2 \triangle Y_{t-2} + ... + \zeta_{p-1} \triangle Y_{t-p+1} + \varepsilon_t, \qquad (41)$$

where $\boldsymbol{\beta} \equiv (\zeta_1, \zeta_2, ..., \zeta_{p-1}, \alpha, \beta)'$ and $\mathbf{x}_t \equiv (\triangle Y_{t-1}, \triangle Y_{t-2}, ..., \triangle Y_{t-p+1}, 1, Y_{t-1})'$. The asymptotic distribution of the *OLS* coefficients estimators, $\hat{\boldsymbol{\beta}}_{\boldsymbol{T}}$ are in the following.

Theorem 4:

Let the data Y_t be generated by (41); then as $T \to \infty$, for the regression model (39),

$$\boldsymbol{\Upsilon}_T(\hat{\boldsymbol{\beta}}_{\boldsymbol{T}} - \boldsymbol{\beta}) \xrightarrow{L} \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V}^{-1}\mathbf{h}_1 \\ \mathbf{Q}^{-1}\mathbf{h}_2 \end{bmatrix}, \qquad (42)$$

where

$$\boldsymbol{\Upsilon}_T = \begin{bmatrix} \sqrt{T} & 0 & . & . & . & 0 & 0 \\ 0 & \sqrt{T} & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sqrt{T} & 0 \\ 0 & 0 & . & . & . & 0 & T \end{bmatrix}, \quad \mathbf{V} = \begin{bmatrix} \gamma_0 & \gamma_1 & . & . & . & \gamma_{p-2} \\ \gamma_1 & \gamma_0 & . & . & . & \gamma_{p-3} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \gamma_{p-2} & \gamma_{p-3} & . & . & . & \gamma_0 \end{bmatrix},$$

$$\mathbf{Q} = \begin{bmatrix} 1 & \lambda \int W(r)dr \\ \lambda \int W(r)dr & \lambda^2 \int [W(r)]^2 dr \end{bmatrix},$$

$$\mathbf{h}_1 \sim N_{p-1}(\mathbf{0}, \mathbf{V}), \;\; \mathbf{h}_2 \sim \begin{bmatrix} \sigma W(1) \\ 1/2\sigma\lambda\{[W(1)]^2 - 1\} \end{bmatrix}, \gamma_j = E[(\triangle Y_t)(\triangle Y_{t-j})]$$

and

$$\lambda = \sigma \cdot \psi(1) = \sigma/(1 - \zeta_1 - \zeta_2 - ... - \zeta_{p-1}). \qquad (43)$$

The results reveals that in a regression of $I(1)$ variables on $I(1)$ and $I(0)$ variables, **the asymptotic distribution of the coefficient of $I(1)$ and $I(0)$ variables are independent**. Thus, the asymptotic distribution of $\sqrt{T}(\hat{\zeta}_j - \zeta_j)$, $j = 1, 2, .., p-1$ and $T(\hat{\beta} - \beta)$ are independent. This results can be used for showing that the distribution of $\hat{\beta}_T$ in the *ADF* regression is the Dickey-Fuller distribution (taking into account of serially correlated in the u_t, see (44)). Also the asymptotic distribution of $\sqrt{T}(\hat{\zeta}_j - \zeta_j)$ is normal.

Therefore, the limiting distribution of $T(\hat{\beta}-\beta)$ is given by the second element of

$$\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\begin{bmatrix} \hat{\alpha}-0 \\ \hat{\beta}_T - 1 \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & \lambda\int W(r)dr \\ \lambda\int W(r)dr & \lambda^2\int[W(r)]^2dr \end{bmatrix}^{-1}\begin{bmatrix} \sigma W(1) \\ 1/2\sigma\lambda\{[W(1)]^2-1\} \end{bmatrix}$$
$$\equiv \begin{bmatrix} \sigma & 0 \\ 0 & \sigma/\lambda \end{bmatrix}\begin{bmatrix} 1 & \int W(r)dr \\ \int W(r)dr & \int[W(r)]^2dr \end{bmatrix}^{-1}\begin{bmatrix} W(1) \\ 1/2\{[W(1)]^2-1\} \end{bmatrix},$$

or

$$T(\hat{\beta}_T - 1) \Rightarrow (\sigma/\lambda)\cdot\frac{1/2\{[W(1)]^2-1\} - W(1)\cdot\int_0^1 W(r)dr}{\int_0^1[W(r)]^2dr - \left[\int_0^1 W(r)dr\right]^2}. \quad (44)$$

The parameter (σ/λ) is the factor to correct the serial correlation in u_t. When u_t is *i.i.d.*, form (38) we have $\zeta_i = 0$ and $\lambda = \sigma$, that is $(\sigma/\lambda) = 1$. This distribution is back to simple Dickey-Fuller distribution. We are now in a position to propose the *ADF* test statistics which correct for (σ/λ) and have the same distribution as *DF*.

Theorem 5 (*ADF*):
Let the data Y_t be generated by (41); then as $T \to \infty$, for the regression model (39),
(a).

$$\frac{T(\hat{\beta}_T - 1)}{1-\hat{\zeta}_1-\hat{\zeta}_2-...-\hat{\zeta}_{p-1}} \Rightarrow \frac{1/2\{[W(1)]^2-1\} - W(1)\cdot\int_0^1 W(r)dr}{\int_0^1[W(r)]^2dr - \left[\int_0^1 W(r)dr\right]^2},$$

(b).

$$t_T = \frac{(\hat{\beta}_T - 1)}{\{s_T^2\mathbf{e}'_{p+1}(\sum\mathbf{x}_t\mathbf{x}'_t)^{-1}\mathbf{e}_{p+1}\}^{1/2}} \Rightarrow \frac{1/2\{[W(1)]^2-1\} - W(1)\cdot\int_0^1 W(r)dr}{\left\{\int_0^1[W(r)]^2dr - \left[\int_0^1 W(r)dr\right]^2\right\}^{1/2}},$$

where $\mathbf{e}_{p+1} = [0\ 0\ ...0\ 1]'$.

Proof:
(a). From (44) we have

$$T\cdot(\lambda/\sigma)(\hat{\beta}_T - 1) \Rightarrow \frac{1/2\{[W(1)]^2-1\} - W(1)\cdot\int_0^1 W(r)dr}{\int_0^1[W(r)]^2dr - \left[\int_0^1 W(r)dr\right]^2}. \quad (45)$$

Recall from (43) that

$$\lambda/\sigma = (1 - \zeta_1 - \zeta_2 - ... - \zeta_{p-1})^{-1}.$$

Since $\hat{\zeta}_j$ is consistent from (42), this magnitude is clearly consistently estimated by

$$\widehat{\lambda/\sigma} = (1 - \hat{\zeta}_1 - \hat{\zeta}_2 - ... - \hat{\zeta}_{p-1})^{-1}. \quad (46)$$

It follows that

$$\begin{aligned} & [T(\hat{\beta}_T - 1) \cdot (\lambda/\sigma)] - [T(\hat{\beta}_T - 1) \cdot (\widehat{\lambda/\sigma})] \\ = \; & T(\hat{\beta}_T - 1) \cdot [(\lambda/\sigma) - (\widehat{\lambda/\sigma})] \\ = \; & O_p(1) \cdot o_p(1) = o_p(1). \end{aligned}$$

Thus $[T(\hat{\beta}_T - 1) \cdot (\widehat{\lambda/\sigma})]$ and $[T(\hat{\beta}_T - 1) \cdot (\lambda/\sigma)]$ have the same asymptotic distribution. This complete the proof of part (a).

To prove art (b), we first multiply the numerator and denominator of t_T by T results in

$$t_T = \frac{T(\hat{\beta}_T - 1)}{\{s_T^2 \mathbf{e}'_{p+1} \boldsymbol{\Upsilon}_T (\sum \mathbf{x}_t \mathbf{x}'_t)^{-1} \boldsymbol{\Upsilon}_T \mathbf{e}_{p+1}\}^{1/2}}. \quad (47)$$

But

$$\begin{aligned} \mathbf{e}'_{p+1} \boldsymbol{\Upsilon}_T (\sum \mathbf{x}_t \mathbf{x}'_t)^{-1} \boldsymbol{\Upsilon}_T \mathbf{e}_{p+1} &= \mathbf{e}'_{p+1} \left[\boldsymbol{\Upsilon}_T^{-1} (\sum \mathbf{x}_t \mathbf{x}'_t) \boldsymbol{\Upsilon}_T^{-1} \right]^{-1} \mathbf{e}_{p+1} \\ &\xrightarrow{L} \mathbf{e}'_{p+1} \begin{bmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^{-1} \end{bmatrix} \mathbf{e}_{p+1} \\ &= \frac{1}{\lambda^2 \cdot \left\{ \int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r) dr \right]^2 \right\}}. \end{aligned}$$

Hence from (45) and (47),

$$\begin{aligned} t_T \; & \xrightarrow{L} \; (\sigma/\lambda) \frac{1/2\{[W(1)]^2 - 1\} - W(1) \cdot \int_0^1 W(r) dr}{\int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r) dr \right]^2} \\ & \quad \div \left\{ \frac{\sigma^2}{\lambda^2 \cdot \left\{ \int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r) dr \right]^2 \right\}} \right\}^{1/2} \\ & = \frac{1/2\{[W(1)]^2 - 1\} - W(1) \cdot \int_0^1 W(r) dr}{\left\{ \int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r) dr \right]^2 \right\}^{1/2}}. \end{aligned}$$

This is the same distribution as in (26). Thus, the usual t test of $\beta = 1$ for OLS estimation of $AR(p)$ can be compared with Theorem 2 and use case 2 of Table B.6 without any correction for the fact u_t is serially correlated (or $\triangle Y$ are included in the regression).

3.3 Augmented Dickey-Fuller Test, Y_t is a $ARMA(p,q)$ process

The fact that the distribution of $\hat{\beta}_T$ in the ADF regression is the Dickey-Fuller distribution has been extended by Said and Dickey (1984) to the more general case in which, under the null hypothesis, the series of first difference are of the general $ARMA(p,q)$ form with unknown p and q. They showed that a regression model, such as (39), is still valid for testing the unit root null under the presence of the serial correlations of error, if the number of lags $\triangle Y$ included as regressor increase with the sample size at a controlled rate $T^{1/3}$. Essentially the moving terms are being approximated by including enough autoregressive terms.

Consider the general $ARIMA(p,1,q)$ model is defined by

$$\begin{aligned} Y_t &= \beta Y_{t-1} + u_t, & (48)\\ (1-\phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)u_t &= (1+\theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)\varepsilon_t & (49) \end{aligned}$$

with $Y_0 = 0$, ε_t is $i.i.d.$ and $\beta = 1$.

(48) and (49) is equivalently written as

$$(1-\phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)\triangle Y_t = (1+\theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)\varepsilon_t. \quad (50)$$

Therefore, (41) as a special case of (50) in which $\theta_j = 0, j = 1, 2, ..., q$.

Rewrite (50) as

$$\eta(L)\triangle Y_t = \varepsilon_t \quad (51)$$

where $\eta(L) = (1-\eta_1 L - \eta_2 L^2 - ....) = (1+\theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)^{-1}(1-\phi_1 L - \phi_2 L^2 - ... - \phi_p L^p)$. That is,

$$\triangle Y_t = \eta_1 \triangle Y_{t-1} + \eta_2 \triangle Y_{t-2} + \eta_3 \triangle Y_{t-3} + ... + \varepsilon_t$$

or

$$Y_t = Y_{t-1} + \eta_1 \triangle Y_{t-1} + \eta_2 \triangle Y_{t-2} + \eta_3 \triangle Y_{t-3} + ... + \varepsilon_t. \quad (52)$$

This motives us to estimate the coefficient in (52) by regression Y_t on $Y_{t-1}, \triangle Y_{t-1}, \triangle Y_{t-2}, ...,$ $\triangle Y_{t-k}$ where k is a suitably chosen integer. To get consistent estimator of the coefficient in (52) it is necessary to let k as a function of T.

Consider a truncated version of (52)

$$\begin{aligned} Y_t &= \alpha + \beta Y_{t-1} + \eta_1 \triangle Y_{t-1} + \eta_2 \triangle Y_{t-2} + ... + \eta_k \triangle Y_{t-k} + e_{tk} \\ &= \mathbf{x}_t' \boldsymbol{\beta} + e_{tk}, \end{aligned} \tag{53}$$

where $\boldsymbol{\beta} \equiv (\alpha, \beta, \zeta_1, \zeta_2, ..., \zeta_k)'$ and $\mathbf{x}_t \equiv (1, Y_{t-1}, \triangle Y_{t-1}, \triangle Y_{t-2}, ..., \triangle Y_{t-k})'$. Notice that e_{tk} is not a white noise. In this case, the limiting distribution of t statistics of the coefficient on the lagged Y_{t-1} (i.e., $\hat{\beta}_T$) from OLS estimation of has the same Dickey-Fuller t-distribution as when u_t is $i.i.d.$.

Theorem 6 (Said-Dickey ADF):
Let the data Y_t be generated by (48) and (49) and the regression model be (53). We assume that $T^{-1/3}k \rightarrow 0$ and there exist $c > 0, r > 0$ such that $ck > T^{1/r}$, then as $T \rightarrow \infty$,

$$t_T = \frac{(\hat{\beta}_T - 1)}{\{s_T^2 \mathbf{e}'_{k+1} (\sum \mathbf{x}_t \mathbf{x}'_t)^{-1} \mathbf{e}_{k+1}\}^{1/2}} \xrightarrow{L} \frac{1/2\{[W(1)]^2 - 1\} - W(1) \cdot \int_0^1 W(r)dr}{\left\{ \int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r)dr \right]^2 \right\}^{1/2}},$$

where $\mathbf{e}_{k+1} = [0\ 1\ ...0\ 0]'$.

The intuition behind that k should be a function of T is clear from the fact

$$e_{tk} = \zeta_{k+1} \triangle Y_{t-k-1} + \zeta_{k+2} \triangle Y_{t-k-2} + ... + \varepsilon_t.$$

Then as $k \rightarrow \infty$

$$\begin{aligned} e_{tk} - \varepsilon_t &= \zeta_{k+1} \triangle Y_{t-k-1} + \zeta_{k+2} \triangle Y_{t-k-2} + ... \\ &\xrightarrow{p} 0 \end{aligned}$$

from the absolute summability of ζ_j, i.e $\sum_{j=0}^{\infty} |\zeta_j| < \infty$ which imply $\zeta_j \rightarrow 0$ and $\zeta_{k+1} \triangle Y_{t-k-1} + \zeta_{k+2} \triangle Y_{t-k-2} + ... \longrightarrow 0$. Then, it is expected that a $ARIMA(p, 1, q)$ process here would have the same asymptotic result with the $ARIMA(p, 1, 0)$ process. Therefore, the asymptotic result should be derived under the condition that $k \rightarrow \infty$. However, it can not increase quickly then T, i.e. we need the condition that $T^{-1/3}k \rightarrow 0$.

3.3.1 Choice of Lag-Length in the *ADF* Test

It has been observed that the size and power properties of the *ADF* test are sensitive to the number of lagged terms (k) used. Several guidelines have been suggested for the choice of k. Ng and Perron (1995) examine these in details.

The guidelines are:

(a). **Rule for fixing k at an arbitrary level independent of T**. Overall, choosing a fixed k is not desirable from their detailed simulation.

(b). **Rule for fixing k as a function of T**. A rule commonly used is the one suggested by Schwert (1989) which is to choose

$$k = Int\{c(T/100)^{1/d}\}.$$

Schwert suggest $c = 12$ and $d = 4$. The problem with such a rule is that it need not be optional for all p and q in the $ARMA(p, q)$.

(c). **Information based rules**. The information criteria suggest choosing k to minimize an objective function that trades off parsimony against reduction in sum of squares. The objective function is of the form (see also p.5 of Ch. 16)

$$I_k = \log \hat{\sigma}_k^2 + k\frac{C_T}{T}.$$

The Akaike information criterion (AIC) choose $C_T = 2$. The Schwarz Bayesian information criterion (BIC) chooses $C_T = \log T$. Ng and Perron argue that both AIC and BIC are asymptotically the same with $ARMA(p, q)$ models and that both of them choose k proportional to $\log T$.

(d). **Sequential rules**. Hall (1994) discusses a *general to specific rule* which is to start with a large value of k (k_{max}). We test the significance of the last coefficient and reduce k iteratively until a significant statistic is encountered.

Ng and Perron (1995) compare AIC, BIC, and Hall's general to specific approach through a Monte Carlo study. The major conclusions are:
(a). Both AIC and BIC choose very small value of k (e.g. $k = 3$). This results in high size distortions, especially with MA errors (Remark: The intuition behind this conclusion is that with MA error, this model is $AR(\infty)$, if you use too small lagged value, your model look not very much like a $AR(\infty)$, and since this

asymptotic results is derived from $k \to \infty$, your finite sample distribution is far away from the asymptotic distribution and results in size distortion.)
(b). Hall's criterion tends to choose higher values of k. The higher the k_{max} is, the higher is the chosen value of k. This results in the size being at the nominal level, but of course with a loss of power. (Remark: Unit root test statistics sa far we have derived is the distribution under the null. On the other side, we can derive the test statistics under the alternative hypothesis of stationary or fractional difference process. Theses test statistics should go to $-\infty$, say, against the left-tailed hypothesis, when the alternative hypothesis is true to be able to consistently reject the null hypothesis, or what is called power is one. A common results is that the asymptotic distribution of the unit root test statistics under the alternative hypothesis is function of k, and with precisely, (T/k), say. For a fixed T, (T/K) is smaller with a larger k. This cause the unit roots distribution tend to $-\infty$ more slower under the alternative hypothesis and has a lower power. See Lee and Schmidt (1996) and Lee and Shie (2004).)

What this study suggests is that Hall's general to specific methods is preferable to the others. DeJong et al. (1992) show that increasing k typically results in a modest decrease in power but a substantial decrease in size distortion. If this is the case the information criteria are at a disadvantage because they result in a choice of very small value of k. However, Stock (1994) propose opposite evidence arguing in favor of BIC compared with Hall's method.

3.4 Phillips-Perron Test, u_t is mixing process

In contrast to ADF where a $AR(1)$ unit root process has been extend to $AR(p)$ and $ARMA(p,q)$ with a unit root, Phillips (1987) and Phillips and Perron (1988) extend the random walk ($AR(1)$ process with a unit root) to a general setting that allows for weakly dependent and heterogeneously distributed innovations. The model Phillips (1987) consider are

$$Y_t = \beta Y_{t-1} + u_t, \tag{54}$$

$$\beta = 1 \tag{55}$$

where $Y_0 = 0$ (not necessary in Phillips's original paper), $\beta = 1$ and u_t is a weakly dependent and heterogeneously distributed innovations sequence to be specified below.

We consider the three least square regressions

$$Y_t = \breve{\beta} Y_{t-1} + \breve{u}_t, \tag{56}$$

$$Y_t = \hat{\alpha} + \hat{\beta} Y_{t-1} + \hat{u}_t, \tag{57}$$

and

$$Y_t = \tilde{\alpha} + \tilde{\beta} Y_{t-1} + \tilde{\delta}(t - \frac{1}{2}T) + \tilde{u}_t, \tag{58}$$

where $\breve{\beta}, (\hat{\alpha}, \hat{\beta})$, and $(\tilde{\alpha}, \tilde{\delta}, \tilde{\beta})$ are the conventional least-squares regression coefficients. Phillips (1987) and Phillips and Perron (1978) were concerned with the limiting distribution of the regression in (56), (57), and (58) ($\breve{\beta}, (\hat{\alpha}, \hat{\beta})$, and $(\tilde{\alpha}, \tilde{\delta}, \tilde{\beta})$) under the null hypothesis that the data are generated by (54) and (55).

So far, we have assumed that the sequence u_t used to construct W_T is $i.i.d.$. Nevertheless, just as we can obtain central limit theorems when u_t is not necessary $i.i.d.$, so also can we obtain FCLT when u_t is not necessary $i.i.d.$. Here we present a version of FCLT, due to McLeish (1975), under a very weak assumption on u_t.

Theorem 7 (McLeish): Let u_t satisfies that
(a). $E(u_t) = 0$,
(b). $sup_t\ E|u_t|^\gamma < \infty$ for some $\gamma > 2$,
(c). $\lambda^2 = \lim_{T\to\infty} E[T^{-1}(\sum u_t)^2]$ exists and $\lambda^2 > 0$, and

(d). u_t is strong mixing with mixing coefficients α_m that satisfy $\sum_1^{\infty} \alpha_m^{1-2/\gamma} < \infty$, then $W_T \Longrightarrow W$, where $W_T(r) \equiv T^{-1/2} \sum_{t=1}^{[Tr]} u_t/\lambda$.

These conditions (a)–(d) allow for both temporal dependence (by mixing) and heteroscedasticity (as long as $\lambda^2 = \lim_{T\to\infty} E[T^{-1}(\sum u_t)^2]$ exist) in the process u_t. (Hint: When u_t is an *i.i.d.* process, then $\lambda^2 = \lim_{T\to\infty} E[T^{-1}(\sum u_t)^2] = \sigma^2$, and this result is back to (10).)

We first provide the following asymptotic results of the sample moments which are useful to derive the asymptotics of the OLS estimator.

Lemma 2:
Let u_t be a random sequence that satisfies the assumptions in Theorem 7, and if $sup_t E|u_t|^{\gamma+\eta} < \infty$ for some $\eta > 0$,

$$y_t = u_1 + u_2 + ... + u_t \quad for\ t = 1, 2, ..., T, \tag{59}$$

with $y_0 = 0$. Then

(a) $T^{-\frac{1}{2}} \sum_{t=1}^{T} u_t \xrightarrow{L} \lambda W(1)$,
(b) $T^{-2} \sum_{t=1}^{T} Y_{t-1}^2 \Longrightarrow \lambda^2 \int_0^1 [W(r)]^2 dr$,
(c) $T^{-\frac{3}{2}} \sum_{t=1}^{T} Y_{t-1} \Longrightarrow \lambda \int_0^1 W(r) dr$,
(d) $T^{-1} \sum_{t=1}^{T} u_t^2 \xrightarrow{p} \sigma_u^2 = T^{-1} \sum_{t=1}^{T} E(u_t^2)$.
(e) $T^{-1} \sum_{t} Y_{t-1} u_t \Longrightarrow \frac{1}{2}\{\lambda^2 [W(1)^2] - \sigma_u^2\}$,
(f) $T^{-\frac{3}{2}} \sum_{t=1}^{T} t u_t \Longrightarrow \lambda [W(1) - \int_0^1 W(r) dr]$,
(g) $T^{-\frac{5}{2}} \sum_{t=1}^{T} t Y_{t-1} \Longrightarrow \lambda \int_0^1 r W(r) dr$,
(h) $T^{-3} \sum_{t=1}^{T} t Y_{t-1}^2 \Longrightarrow \lambda^2 \int_0^1 r [W(r)]^2 dr$.

A joint weak convergence for the sample moments given above to their respective limits is easily established and will be used below.

Proof:

Proofs of items (a), (b), (c), (f), (g) and (h) are analogous to those proofs at Lemma 1. Item (d) is the results of LLN for a mixing process.

(e). For a random walk, $Y_t^2 = (Y_{t-1} + u_t)^2 = Y_{t-1}^2 + 2Y_{t-1}u_t + u_t^2$, implying that $Y_{t-1}u_t = (1/2)\{Y_t^2 - Y_{t-1}^2 - u_t^2\}$ and then $\sum_{t=1}^T Y_{t-1}u_t = (1/2)\{Y_T^2 - Y_0^2\} - (1/2)\sum_{t=1}^T u_t^2$. Recall that $Y_0 = 0$, and thus it is convenient to write $\sum_{t=1}^T Y_{t-1}u_t = \frac{1}{2}Y_T^2 - \frac{1}{2}\sum_{t=1}^T u_t^2$. From items (a)) we know that $T^{-1}Y_T^2$ $(= (T^{-1/2}\sum_{t=1}^T u_s)^2 \xrightarrow{L} \lambda^2 W^2(1)$ and $\sum_{t=1}^T u_t^2 \xrightarrow{p} \sigma_u^2$ by LLN (MacLeish); then, $\sum_{t=1}^T Y_{t-1}u_t \Longrightarrow \frac{1}{2}\{\lambda^2[W(1)^2] - \sigma_u^2\}$.

3.4.1 No Constant Term or Time Trend in the Regression; True Process Is a Random Walk

We first consider the case that no constant term or time trend in the regression model, but true process is a random walk. The asymptotic distributions of OLS unit root coefficients estimator and t-ratio test statistics are in the following.

Theorem 8:

Let the data Y_t be generated by (54) and (55); and u_t be a random sequence that satisfies the assumptions in Theorem 7, and if $sup_t E|u_t|^{\gamma+\eta} < \infty$ for some $\eta > 0$, then as $T \to \infty$, for the regression model (56),

$$T(\breve{\beta}_T - 1) \Rightarrow \frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr}$$

and

$$t = \frac{(\breve{\beta}_T - 1)}{\breve{\sigma}_{\breve{\beta}_T}} \Rightarrow \frac{(\lambda/2\sigma_u)\{[W(1)]^2 - \sigma_u^2/\lambda^2\}}{\{\int_0^1 [W(r)]^2 dr\}^{1/2}},$$

where $\breve{\sigma}_{\breve{\beta}_T}^2 = [s_T^2 \div \sum_{t=1}^T Y_{t-1}^2]^{1/2}$ and s_T^2 denote the OLS estimate of the disturbance variance:

$$s_T^2 = \sum_{t=1}^T (Y_t - \breve{\beta}_T Y_{t-1})^2/(T-1).$$

Proof:
Since the deviation of the OLS estimate from the true value is characterized by

$$T(\breve{\beta}_T - 1) = \frac{T^{-1}\sum_{t=1}^{T} Y_{t-1}u_t}{T^{-2}\sum_{t=1}^{T} Y_{t-1}^2}, \quad (60)$$

which is a continuous function function of Lemma 2's (b) and (e), it follows that under the null hypothesis that $\beta = 1$, the OLS estimator $\breve{\beta}$ is characterized by

$$T(\breve{\beta}_T - 1) \Rightarrow \frac{1/2\{\lambda^2[W(1)^2] - \sigma_u^2\}}{\lambda^2 \int_0^1 [W(r)]^2 dr} = \frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr}. \quad (61)$$

To prove second part of this theorem, We first note since $\breve{\beta}$ is a consistent estimator of β from (61), s_T^2 is a consistent estimator of σ_u^2, from the analogous proofs as in Theorem 2. Then, we can express the t statistics alternatively as

$$t_T = T(\breve{\beta}_T - 1)\left\{T^{-2}\sum_{t=1}^{T} Y_{t-1}^2\right\}^{1/2} \div (s_T^2)^{1/2}$$

or

$$t_T = \frac{T^{-1}\sum_{t=1}^{T} Y_{t-1}u_t}{\left\{T^{-2}\sum_{t=1}^{T} Y_{t-1}^2\right\}^{1/2}(s_T^2)^{1/2}},$$

which is a continuous function function of Lemma 2's (b) and (e), it follows that under the null hypothesis that $\beta = 1$, the asymptotic distribution of OLS t statistics is characterized by

$$t_T \xrightarrow{L} \frac{(1/2)\{[\lambda^2 W(1)]^2 - \sigma_u^2\}}{\left\{\lambda^2 \int_0^1 [W(r)]^2 dr\right\}^{1/2}(\sigma_u^2)^{1/2}} = \frac{(\lambda/2\sigma_u)\{[W(1)]^2 - \sigma_u^2/\lambda^2\}}{\{\int_0^1 [W(r)]^2 dr\}^{1/2}}. \quad (62)$$

This complete the proof of this Theorem.

Theorem 8 extends (18) and (24) to the very general case of weakly dependent and heterogeneously distributed data. When u_t is a $i.i.d.$ sequence, $\lambda^2 = T^{-1}E[(\sum u_t)^2] = T^{-1}\sum E(u_t^2) = \sigma_u^2$, and we see that the results of Theorem 8 reduces to those of Theorem 2.

Several interesting things are noteworthy to the asymptotic distribution results for the least squares estimators, (61). First, note that the scale factor here is T, not $\sqrt{T}$ as it previously has been. Thus, $\breve{\beta}$ is "collapsing " to its limits at a much faster rate than before. This is sometimes called *superconsistency*. Next, note that the limiting distribution is no longer normal; instead, we have a distribution that is somewhat complicated function of a Wiener process. When $\sigma_u^2 = \lambda^2$ (independence) we have the distribution of J.S. White (1958, p.1196), apart from an incorrect scaling there, as noted by Phillips (1987). For $\sigma_u^2 = \lambda^2$, this distribution is also that tabulated by Dickey and Fuller (1979) in their famous work on testing for unit root.

In the regression setting studied in previous chapters the existence of serial correlation in u_t in the presence of a lagged dependent variable regressor leads to the inconsistency of $\breve{\beta}$ for β_0 as discussed in Chapter 10. Here, however, the situation is quite different. Even though the regressor is a lagged dependent variables, $\breve{\beta}$ is consistent for $\beta_0 = 1$ despite the fact that condition (c) and (d) of Theorem 7 permit u_t to display considerable correlation.

The effect of the serial correlation is that $\sigma_u^2 \neq \lambda^2$. This results is a shift of the location of the asymptotic distribution away from zero (since $E(W(1)^2 - \sigma_u^2/\lambda^2) \neq 0$ where $E(W(1)^2) = E(\chi^2_{(1)}) = 1.$) relative to the $\sigma_u^2 = \lambda^2$ case of no serial correlation). Despite this effect of the serial correlation in u_t, we no longer have the serious adverse consequence of the inconsistency of $\breve{\beta}$.

One way of understanding why this is so is succinctly expressed by Phillips (1987, p.283):

Intuitively, when the data generating process has a unit root, the strength of the single (as measured by the sample variation of the regressor Y_{t-1} dominates the noise by a factor of $O(T)$, so that the effect of any regressor-error correlation are annihilated in the regression as $T \to \infty$.

Note that , however, even when $\sigma_u^2 = \lambda^2$ the asymptotic distribution given in (61) is not centered about zero, so an asymptotic bias is still present. The reason for this is that there generally exist a strong (negative) correlation between $W(1)^2$ and $(\int_0^1 [W(r)^2]dr)^{-1}$, resulting from the fact that $W(1)^2$ and $W(r)^2$ are highly correlated for each r. Thus, even though $E(W(1)^2 - \sigma_u^2/\lambda^2) = 0$ with $\sigma_u^2 = \lambda^2$, we do **not** have

$$E\left[\frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr}\right] = 0.$$

See Abadir (1995) for further details.

3.4.2 Estimation of λ^2 and σ_u^2

The limiting distribution given in Theorem 8 depend on unknown parameters σ_u^2 and λ^2. Theses distributions are therefore not directly useable for statistical testing. However, both theses parameters may be consistently estimated and the estimates may be used to construct modified statistics whose limiting distribution are independent of (λ^2, σ_u^2). As we shall see, these new statistics provide very general tests for the presence of a unit root in (54).

As shown in Lemma 2 (d), $T^{-1}\sum_{t=1}^{T} u_t^2 \to \sigma_u^2$. This provides us with the simple estimator

$$s_u^2 = T^{-1}\sum_{t=1}^{T}(Y_t - Y_{t-1})^2 = T^{-1}\sum_{t=1}^{T} u_t^2,$$

which is consistent for σ_u^2 under the null hypothesis $\beta = 1$. Since $\breve{\beta} \to 1$ by Theorem 8 we may also use $\breve{s}_u^2 = T^{-1}\sum_1^T (Y_t - \breve{\beta}Y_{t-1})^2$ as a consistent estimator of σ_u^2.

Consistent estimation of $\lambda^2 = \lim_{T\to\infty} T^{-1}E(\sum_1^T u_t)^2$ is more difficult. We start by defining

$$\begin{aligned}\lambda_T^2 &= T^{-1}E(\sum_1^T u_t)^2 \\ &= T^{-1}\sum_1^T E(u_t^2) + 2T^{-1}\sum_{\tau=1}^{T-1}\sum_{t=\tau+1}^{T} E(u_t u_{t-\tau})\end{aligned}$$

and by introducing the approximate

$$\lambda_{Tl}^2 = T^{-1}\sum_1^T E(u_t^2) + 2T^{-1}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T} E(u_t u_{t-\tau}).$$

We shall call l the lag truncation number. For large T and large $l < T$, λ_{Tl}^2 may be expected to very close to λ_T^2 if the total contribution in λ_T^2 of covariance such as $E(u_t u_{t-\tau})$ with long lags $\tau > l$ is small. This will be true if u_t satisfies the assumption in Theorem 7. Formally, we have the following lemma.

Lemma 3:

If the sequence u_t satisfies the assumption in Theorem 7 and if $l \to \infty$ as $T \to \infty$, then $\lambda_T^2 - \lambda_{Tl}^2 \to 0$ as $T \to \infty$.

This lemma suggests that under suitable conditions on the rate at which $l \to \infty$ as $T \to \infty$ we may proceed to estimate λ^2 from finite sample of data by sequentially estimating λ_{Tl}^2. We define

$$s_{Tl}^2 = T^{-1}\sum_1^T u_t^2 + 2T^{-1}\sum_{\tau=1}^{l}\sum_{t=\tau+1}^{T} u_t u_{t-\tau}. \tag{63}$$

The following result establishes that s_{Tl}^2 is a consistent estimator of λ^2.

Theorem 9:
If
(a). u_t satisfies all the assumptions in Theorem 7 except that part (b) is replaced by the stronger moment condition: $sup_t E|u_t|^{2\gamma} < \infty$, for some $\gamma > 2$;
(b). $l \to \infty$ as $T \to \infty$ such that $l = o(T^{1/4})$;
then $s_{Tl}^2 \to \lambda^2$ as $T \to \infty$.

According to this result, if we allow the number of estimated autocovariances to increase as $T \to \infty$ but control the rate of increase so that $l = o(T^{1/4})$, then s_{Tl}^2 yields a consistent estimator of λ^2. Inevitably the choice of l will be an empirical matter.

Rather than using the first difference $u_t = Y_t - Y_{t-1}$ in the construction of s_{Tl}^2, we could have used the residuals $\breve{u}_t = Y_t - \breve{\beta} Y_{t-1}$ from the least squares regression. Since $\breve{\beta} \to 1$, this estimator is also consistent for λ^2 under the null hypothesis $\beta = 1$.

We remark that s_{Tl}^2 is not constrained to be nongative as it presently defined in (63). When there are large negative sample serial covariance, s_{Tl}^2 can take on negative values. Newey and West (1987) have suggested a modification to variance estimator such as s_{Tl}^2 which ensures that they are nonnegative. In the present case, the modification yield:

$$\breve{s}_{Tl}^2 = T^{-1}\sum_1^T \breve{u}_t^2 + 2T^{-1}\sum_{\tau=1}^{l} w_{\tau l}\sum_{t=\tau+1}^{T} \breve{u}_t \breve{u}_{t-\tau}, \tag{64}$$

where

$$w_{Tl} = 1 - \tau/(l+1), \tag{65}$$

which put a higher weight on more recent autocovariances.

3.4.3 New Tests for a Unit Root

The consistent estimator $\breve{s}_u^2$ and $\breve{s}_{Tl}^2$ may be used to develop new tests for unit roots that apply under very general conditions. We define the statistics:

$$Z_{\breve{\beta}} = T(\breve{\beta} - 1) - \frac{1/2(\breve{s}_{Tl}^2 - \breve{s}_u^2)}{\left(T^{-2}\sum_1^T Y_{t-1}^2\right)} \tag{66}$$

and

$$Z_{\breve{t}} = t_T \cdot (\breve{s}_u^2/\breve{s}_{Tl}^2)^{1/2} - 1/2(\breve{s}_{Tl}^2 - \breve{s}_u^2)\left[\breve{s}_{Tl}\left(T^{-2}\sum_1^T Y_{t-1}^2\right)^{1/2}\right]^{-1}. \tag{67}$$

$Z_{\breve{\beta}}$ is a transformation of the standardized estimator $T(\breve{\beta} - 1)$ and $Z_{\breve{t}}$ is a transformation of the regression t statistics as in (62). The limiting distribution of $Z_{\breve{\beta}}$ and $Z_{\breve{t}}$ are given by the following results.

Theorem 10 (Phillips 1987):
If the condition of Lemma 3 are satisfied, then as $T \to \infty$,
(a).

$$Z_{\breve{\beta}} \Rightarrow \frac{1/2\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2 dr}$$

and
(b).

$$Z_{\breve{t}} \Rightarrow \frac{1/2\{[W(1)]^2 - 1\}}{\{\int_0^1 [W(r)]^2 dr\}^{1/2}}$$

under the null hypothesis that data is generated by (54) and (55).

Proof:
(a). From (61) we have

$$T(\breve{\beta}_T - 1) \Rightarrow \frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr}$$

and from Lemma 2 (b) and Theorem 9 we have

$$\frac{1/2(s_{Tl}^2 - s_u^2)}{T^{-2}\sum_1^T Y_{t-1}^2} \Rightarrow \frac{1/2(\lambda^2 - \sigma_u^2)}{\lambda^2 \int_0^1 [W(r)]^2 dr} \equiv \frac{1/2(1 - \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr}.$$

Therefore, the test statistics $Z_{\breve{\beta}}$ is distributed as

$$\begin{aligned} Z_{\breve{\beta}} = T(\breve{\beta}_T - 1) - \frac{1/2(\breve{s}_{Tl}^2 - \breve{s}_u^2)}{T^{-2}\sum_1^T Y_{t-1}^2} &\Rightarrow \frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2 - 1 + \sigma_u^2/\lambda^2)}{\int_0^1 [W(r)]^2 dr} \\ &\equiv \frac{1/2([W(1)^2] - 1)}{\int_0^1 [W(r)]^2 dr}. \end{aligned}$$

(b). From (62) and the consistency $\breve{s}_u^2$ and $\breve{s}_{Tl}^2$ we have

$$t_T \cdot (\breve{s}_u^2/\breve{s}_{Tl}^2)^{1/2} \Rightarrow \frac{1/2\{[W(1)]^2 - \sigma_u^2/\lambda^2\}}{\{\int_0^1 [W(r)]^2 dr\}^{1/2}}. \tag{68}$$

Consider the following statistics:

$$\frac{1/2(\breve{s}_{Tl}^2 - \breve{s}_u^2)}{\breve{s}_{Tl}\left[T^{-2}\sum_1^T Y_{t-1}^2\right]^{1/2}} \Rightarrow \frac{1/2(\lambda^2 - \sigma_u^2)}{\lambda^2\left\{\int_0^1 [W(r)]^2 dr\right\}^{1/2}} \equiv \frac{1/2(1 - \sigma_u^2/\lambda^2)}{\left\{\int_0^1 [W(r)]^2 dr\right\}^{1/2}}. \tag{69}$$

Combining (68) and (69) we have

$$\begin{aligned} Z_{\breve{t}} = t_T \cdot (\breve{s}_u^2/\breve{s}_{Tl}^2)^{1/2} - \frac{1/2(\breve{s}_{Tl}^2 - \breve{s}_u^2)}{\breve{s}_{Tl}\left[T^{-2}\sum_1^T Y_{t-1}^2\right]^{1/2}} &\Rightarrow \frac{1/2([W(1)^2] - \sigma_u^2/\lambda^2 - 1 + \sigma_u^2/\lambda^2)}{\left\{\int_0^1 [W(r)]^2 dr\right\}^{1/2}} \\ &\equiv \frac{1/2([W(1)^2] - 1)}{\left\{\int_0^1 [W(r)]^2 dr\right\}^{1/2}}. \end{aligned}$$

Theorem 10 demonstrate that the limiting distribution of the two statistics $Z_{\breve{\beta}}$ and $Z_{\breve{t}}$ are invariant within a very wide class of weakly dependent and possibly heterogeneously distributed innovation u_t. More, the limiting distribution of $Z_{\breve{\beta}}$ is identical to that of $T(\breve{\beta}-1)$ when $\lambda^2=\sigma_u^2$, so that the statistical tables reported in the section labeled Case 1 in Table B.5 are still useable.

The limiting distribution of $Z_{\breve{t}}$ given in Theorem (11) is identical to that of regression t_T statistics when $\lambda^2=\sigma_u^2$. This is, in fact, the limiting distribution of the t statistics when the innovation u_t is $i.i.d.(0,\sigma^2)$. Therefore, the statistical tables reported in the section labeled Case 1 in Table B.6 are still useable.

Phillips and Perron (1988) analysis the asymptotic results of estimator OLS when the regression contains a constant ($\hat{\beta}$) or a constant and a time trend ($\tilde{\beta}$) under the assumption that true data generating process is (54) and (55).

Theorem 11 (Phillips and Perron 1988):
If the condition of Lemma 3 are satisfied, then as $T\to\infty$,

$$Z_{\hat{\beta}}=T(\hat{\beta}-1)-\frac{1/2(\hat{s}_{Tl}^2-\hat{s}_u^2)}{T^{-2}\sum_1^T(Y_{t-1}-\bar{Y}_{-1})^2},$$

$$Z_{\hat{t}}=\hat{t}_T\cdot(\hat{s}_u^2/\hat{s}_{Tl}^2)^{1/2}-1/2(\hat{s}_{Tl}^2-\hat{s}_u^2)\left[\hat{s}_{Tl}\left(T^{-2}\sum_1^T(Y_{t-1}-\bar{Y}_{-1})^2\right)^{1/2}\right]^{-1},$$

$$Z_{\tilde{\beta}}=T(\tilde{\beta}-1)-\frac{T^6}{24D_X}(\tilde{s}_{Tl}^2-\tilde{s}_u^2),$$

and

$$Z_{\tilde{t}}=\tilde{t}_T\cdot(\tilde{s}_u^2/\tilde{s}_{Tl}^2)^{1/2}-\frac{T^3(s_{Tl}^2-s_u^2)}{4\sqrt{3}D_x^{1/2}s_{Tl}},$$

the limiting distribution of Z are identical to those of distribution when $\lambda^2=\sigma_u^2$, where $\bar{Y}_{-1}=\sum_1^{T-1}Y_t/(T-1)$ and $D_x=det(X'X)$ and the regressors are $X=(1,t,Y_{t-1})$, $\hat{s}_u^2=T^{-1}\sum_1^T(Y_t-\hat{\alpha}-\hat{\beta}Y_{t-1})^2$, $\hat{s}_{Tl}^2=T^{-1}\sum_1^T\hat{u}_t^2+2T^{-1}\sum_{\tau=1}^l w_{\tau l}\sum_{t=\tau+1}^T\hat{u}_t\hat{u}_{t-\tau}$, $\tilde{s}_u^2=T^{-1}\sum_1^T[Y_t-\tilde{\alpha}-\tilde{\beta}Y_{t-1}-\tilde{\delta}(t-\frac{1}{2}T)]^2$, and $\tilde{s}_{Tl}^2=T^{-1}\sum_1^T\tilde{u}_t^2+2T^{-1}\sum_{\tau=1}^l w_{\tau l}\sum_{t=\tau+1}^T\tilde{u}_t\tilde{u}_{t-\tau}$.

Exercise:
Reproduce Case 4 at Table B.5 and B.6 on Hamilton (1994)'s p.762 and 763, respectively. Two things to be noted:

(1). Confirms that the same results is obtained from non-gaussian *i.i.d.*
(2). The constant α will not affect this distribution. (so using $\alpha = 0$ and $\alpha = 10000$ will get identical results)

Exercise:
Reproduce Table 1 of Phillips and Perron (1988, p.344) from which you will have the chance to have your own unit root test's (ADF and PP) program in Gauss and the chance to see what are size distortion and the problem arrising from the lake of power of unit root tests.

3.5 Phillips-Perron Test, u_t is a $MA(\infty)$ process

Hamilton (1994, Ch. 17) has parameterized the Phillips-Perron test by assuming that the innovation in (54) to be

$$u_t = \varphi(L)\varepsilon_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j}, \tag{70}$$

where ε_t is a white noise process $(0, \sigma_\varepsilon^2)$ and) $\sum_{j=0}^{\infty} j \cdot |\varphi_j| < \infty$.

3.5.1 Beveridge-Nelson Decomposition

Since (70) is a subcase of the assumption in Theorem 7 (McLeish), all we have to shown is that the "long-run" variance in Theorem 7, λ^2, here is equal to $\sigma_\varepsilon^2 \cdot \varphi^2(1)$ (Hamilton, p.505, eq. 17.5.10.). To do this, we need the Beveridge-Nelson Decomposition.

Theorem 12 (Beveridge-Nelson (B-N) Decomposition):
Let

$$u_t = \varphi(L)\varepsilon_t = \sum_{j=0}^{\infty} \varphi_j \varepsilon_{t-j}, \tag{71}$$

where ε_t is a white noise process $(0, \sigma_\varepsilon^2)$ and $\sum_{j=0}^{\infty} j \cdot |\varphi_j| < \infty$.

Then

$$u_1 + u_2 + ... + u_t = \varphi(1)(\varepsilon_1 + \varepsilon_2 + ... + \varepsilon_t) + \xi_t - \xi_0,$$

where $\varphi(1) \equiv \sum_{j=0}^{\infty} \varphi_j, \xi_t = \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}, \alpha_j = -(\varphi_{j+1} + \varphi_{j+2} + \varphi_{j+3} + ...)$ and $\sum_{j=0}^{\infty} |\alpha_j| < \infty$. (Therefore, ξ_t is a stationary process since it is a $MA(\infty)$ with absolute summable coefficients.)

Proof:

Observe that

$$\begin{aligned}
\sum_{s=1}^{t} u_s &= \sum_{s=1}^{t}\sum_{j=0}^{\infty}\varphi_j\varepsilon_{s-j} \\
&= \{\varphi_0\varepsilon_t+\varphi_1\varepsilon_{t-1}+\varphi_2\varepsilon_{t-2}+...+\varphi_t\varepsilon_0+\varphi_{t+1}\varepsilon_{-1}+...\} \\
&\quad +\{\varphi_0\varepsilon_{t-1}+\varphi_1\varepsilon_{t-2}+\varphi_2\varepsilon_{t-3}+...+\varphi_{t-1}\varepsilon_0+\varphi_t\varepsilon_{-1}+...\} \\
&\quad +\{\varphi_0\varepsilon_{t-2}+\varphi_1\varepsilon_{t-3}+\varphi_2\varepsilon_{t-4}+...+\varphi_{t-2}\varepsilon_0+\varphi_{t-1}\varepsilon_{-1}+...\} \\
&\quad +...+\{\varphi_0\varepsilon_1+\varphi_1\varepsilon_0+\varphi_2\varepsilon_{-1}+...\} \\
&= \varphi_0\varepsilon_t+(\varphi_0+\varphi_1)\varepsilon_{t-1}+(\varphi_0+\varphi_1+\varphi_2)\varepsilon_{t-2}+... \\
&\quad +(\varphi_0+\varphi_1+\varphi_2+...+\varphi_{t-1})\varepsilon_1+(\varphi_1+\varphi_2+...+\varphi_t)\varepsilon_0 \\
&\quad +(\varphi_2+\varphi_3+...+\varphi_{t+1})\varepsilon_{-1}+.... \\
&= (\varphi_0+\varphi_1+\varphi_2+...)\varepsilon_t-(\varphi_1+\varphi_2+\varphi_3+...)\varepsilon_t \\
&\quad +(\varphi_0+\varphi_1+\varphi_2+...)\varepsilon_{t-1}-(\varphi_2+\varphi_3+\varphi_4+...)\varepsilon_{t-1} \\
&\quad +(\varphi_0+\varphi_1+\varphi_2+...)\varepsilon_{t-2}-(\varphi_3+\varphi_4+\varphi_5+...)\varepsilon_{t-2}+... \\
&\quad +(\varphi_0+\varphi_1+\varphi_2+...)\varepsilon_1-(\varphi_t+\varphi_{t+1}+\varphi_{t+2}+...)\varepsilon_1 \\
&\quad +(\varphi_1+\varphi_2+\varphi_3+...)\varepsilon_0-(\varphi_{t+1}+\varphi_{t+2}+\varphi_{t+3}+...)\varepsilon_0 \\
&\quad +(\varphi_2+\varphi_3+\varphi_4+...)\varepsilon_{-1}-(\varphi_{t+2}+\varphi_{t+3}+\varphi_{t+4}+...)\varepsilon_{-1}+...
\end{aligned}$$

or

$$\sum_{s=1}^{t} u_s = \varphi(1)\cdot\sum_{s=1}^{t}\varepsilon_s+\xi_t-\xi_0,$$

where

$$\begin{aligned}
\xi_t &= -(\varphi_1+\varphi_2+\varphi_3+...)\varepsilon_t-(\varphi_2+\varphi_3+\varphi_4+...)\varepsilon_{t-1} \\
&\quad -(\varphi_3+\varphi_4+\varphi_5+...)\varepsilon_{t-3}-... \\
\xi_0 &= -(\varphi_1+\varphi_2+\varphi_3+...)\varepsilon_0-(\varphi_2+\varphi_3+\varphi_4+...)\varepsilon_{-1} \\
&\quad -(\varphi_3+\varphi_4+\varphi_5+...)\varepsilon_{-2}...
\end{aligned}$$

This theorem states that for any serial correlated process u_t that satisfy (71), its partial sum ($\sum u_t$) can be write as the sum of a random walk ($\varphi(1)\sum\varepsilon_t$) and a stationary process, ξ_t and a initial condition ξ_0. Notice that ξ_t is stationary from the fact that $\xi_t=\sum_{j=0}^{\infty}\alpha_j\varepsilon_{t-j}$, where $\alpha_j=-(\varphi_{j+1}+\varphi_{j+2}+...)$ and $\{\alpha_j\}_{j=0}^{\infty}$

is absolutely summable:

$$\begin{aligned}
\sum_{j=0}^{\infty}|\alpha_j| &= |\varphi_1+\varphi_2+\varphi_3+...|+|\varphi_2+\varphi_3+\varphi_4+...|+|\varphi_3+\varphi_4+\varphi_5+...|+...\\
&\leq \{|\varphi_1|+|\varphi_2|+|\varphi_3|+...\}+\{|\varphi_2|+|\varphi_3|+|\varphi_4|+...\}+\{|\varphi_3|+|\varphi_4|+|\varphi_5|+...\}+...\\
&= |\varphi_1|+2|\varphi_2|+3|\varphi_3|+...\\
&= \sum_{j=0}^{\infty} j\cdot|\varphi_j|,
\end{aligned}$$

which is bounded by the assumptions in Theorem 12.

3.5.2 The Equality of Long-Run Variance in Phillips's and Hamilton's Assumption

We now show that the long run variance $E(T^{-1}(\sum u_t)^2)$ in Theorem 7 is equal to $\varphi(1)^2\sigma_\varepsilon^2$ (Hamilton's p.505, (17.5.10). From B-N Decomposition we see that

$$u_1+u_2+...+u_T=\varphi(1)(\varepsilon_1+\varepsilon_2+...+\varepsilon_T)+\xi_T-\xi_0,$$

therefore,

$$\begin{aligned}
\lambda^2 &= T^{-1}E[(u_1+u_2+...+u_T)^2]\\
&= T^{-1}E\{[\varphi(1)^2(\varepsilon_1+\varepsilon_2+...+\varepsilon_T)^2]+\xi_T^2+\xi_0^2+2[\varphi(1)(\varepsilon_1+\varepsilon_2+...+\varepsilon_T)\cdot\xi_T]\\
&\quad +2[\varphi(1)(\varepsilon_1+\varepsilon_2+...+\varepsilon_T)\cdot\xi_0]+2[\xi_T\xi_0]\}\\
&= T^{-1}\left[(\varphi(1)^2T\sigma_\varepsilon^2)+E(\xi_T^2)+E(\xi_0^2)+2\left(\varphi(1)\sigma_\varepsilon^2\sum_{j=0}^{T-1}\alpha_j\right)+0+\sigma_\varepsilon^2\sum_{j=0}^{\infty}\alpha_j\alpha_{T+j}\right]\\
&\to \varphi(1)^2\sigma_\varepsilon^2
\end{aligned}$$

from the stationarity of ξ_T and absolute summability of α_j. (We have to show that $\sum_{j=0}^{\infty}|\alpha_j|<0$ imply $\sum_{j=0}^{\infty}\alpha_j<0$ and $\sum_{j=0}^{\infty}\alpha_j\alpha_{T+j}<0$.)

Therefore, the results of Hamilton is all the same with those of Phillips as long as we replace λ^2 with $\varphi(1)^2\sigma_\varepsilon^2$.

4 Issues in Unit Root Testing

4.1 Size Distortion and Low Power of Unit Root Tests

Schwart (1989) first presented Monte-Carlo evidence to point out the size distortion problem of the commonly used unit root test. He argued that the distribution of the Dickey-Fuller tests is far different from the distribution (this is the meaning of size-distortion, the distribution under the null hypothesis is not what you have expected, and therefore the 5% critical value is misleading) reported by Dickey and Fuller if the underlying distribution contains a moving-average component. He also suggests that the Phillips and Perron (PP) test suffer from size distortions when the MA parameters is large, which is the case with many economic time series as noted by Schwert (1989). The test with the least size distortion is the Said-Dickey (1984) high-order autoregression t-test. Whereas Schwert complained the size distortion of unit root tests, DeJong et al. complained about the low power of unit root tests, DeJong et al. (1992) argued about that the unit root tests have low power against possible trend-stationary alternatives. Similar problems about size distortions and low power were noticed by Agiakoglou and Newbold (1992).

The poor power problem is not unique to the unit root tests. Cochrane argue that any test of the hypothesis $\theta = \theta_0$ has arbitrarily low power against alternative $\theta_0 - \epsilon$ in small sample, **but in many cases the difference between θ_0 and $\theta_0 - \epsilon$ would not be considered important from the statistical or economic perspective**. (For example, the expected value of a population height is 170 or 171.) But the low power problem is particular disturbing in the unit root case because of the discontinuity of the distribution theory near unit root. (unit root test statistics has different asymptotic distribution under the null and the alternative.)

Mention must be made of a paper by Gonzalo and Lee (1996) who complain about the repetition of the phrase "lack of power unit root test". They show numerically that the lack of power and size distortion of the Dickey-Fuller tests for unit roots are similar to and in many situations even smaller than the lack of power and size distortions of the standard student t-tests for stationary roots in an autoregressive model. But arguments like this miss the important point. There is no discontinuity of inference in the latter case but there is in the case of unit root tests. Thus, the consequences of lack of power are vastly different in the two cases.

There have been several solutions to the problems of size distortion and low power of the ADF and PP tests. Some of these are modifications of the ADF and PP tests and others are new tests. See Maddala and Kim (1999) p.103 for a good survey.

4.2 Tests With Stationarity as Null, the KPSS

Kwiatkowski, Phillips, Schmidt, and Shin (1992), which is often referred to as $KPSS$, start with the model

$$Y_t = \psi + \delta t + \zeta_t + \varepsilon_t,$$

where ε_t is a stationary process and ζ_t is a random walk given by

$$\zeta_t = \zeta_{t-1} + u_t, \quad u_t \sim i.i.d.(0, \sigma_u^2).$$

The null hypothesis of stationarity is formulated as

$$H_0 : \ \sigma_u^2 = 0 \quad or \ \zeta_t \ \ is \ a \ constant.$$

The $KPSS$ test statistics for this hypothesis is given by

$$LM = \frac{\sum_{t=1}^{T} S_t^2}{\tilde{s}_{Tl}^2},$$

where

$$\tilde{s}_{Tl}^2 = T^{-1}\sum_{t=1}^{T} e_t + 2T^{-1}\sum_{\tau=1}^{l} w_{\tau l}\sum_{t=\tau+1}^{T} e_t e_{t-\tau}$$

is a consistent estimator of long run variance $\lim_{T\to\infty} T^{-1}E(S_T^2)$. Here $W_{\tau l}$ is an optimal weighting function that corresponds to the choice of a spectral window. $KPSS$ use the Bartlett window, as suggested by Newey and West (1987),

$$w_{\tau l} = 1 - \frac{\tau}{l+1},$$

and e_t are the residuals from the regression of Y_t on a constant and a time trend (remember that LM test statistics is constructed under the null hypothesis), and S_t is the partial sum of e_t defined by

$$S_t = \sum_{i=1}^{T} e_i, \quad t = 1, 2, ..., T.$$

For consistency of $\tilde{s}^2_{Tl}$, it is necessary that $l \to \infty$ as $T \to \infty$. The rate $l = o(T^{1/2})$ is usually satisfactory. $KPSS$ derive the asymptotic distribution of the LM statistic and tabulate the critical values by simulation.

For testing the null of level stationarity instead of trend stationarity the test is constructed the same way except that e_t is obtained as the residual from a regression of Y_t on an intercept only. The test is an upper tail test.

It has been suggested (see, e.g., $KPSS$, p.176 and Choi, 1994, p.721) that the tests using stationarity as null can be used for $confirmatory$ analysis, i.e., to confirm our conclusions about unit roots. However, if both tests fail to reject the respective nulls or both reject the respective nulls, **we do not have a confirmation**.

4.3 Panel Data Unit Root Tests

The principle motivation behind panel data unit root tests is to increase the power of unit root tests by increasing the sample size. An alternative route of increasing the sample size by using long time series data, it is argued, causes problems arising from structural changes. However, it is not clear whether this is more of a problem than cross-sectional heterogeneity, a problem with the use of panel data.

It is often argued that the commonly used unit root tests such as ADF and PP are not very powerful, and that using panel data you get a more power test. See Maddala and Kim (1999), p.134 for a good survey.