Chapter 14

Unit Roots

14.1 The Random Walk Model

Chapter 6 gave a detailed analysis of the OLS estimator of the AR(1) model

$$x_t = \lambda x_{t-1} + u_t \tag{14.1.1}$$

subject to the stability condition $|\lambda| < 1$, under a range of assumptions about u_t. For the case $\lambda = 1$, the asymptotic theory is quite different from the stable case. This is a member of the class of integrated or I(1) processes, defined in §5.2.7. Borrowing the terminology from §5.1.2, they are also called unit root processes. In this case $u_t = \Delta x_t$, and the designation I(d) generally refers to the number of differencing operations needed to produce a process having the 'usual' stability properties, which is the I(0) case. To give a rigorous definition of I(0) is quite tricky. In §5.2.7 an I(0) process was characterized, in effect, as one having a stationary invertible ARMA representation, and this is the most usual definition, but is rather restrictive. It rules out stationary but nonlinear cases, such as GARCH processes, that could well arise in practice. Perhaps the best working definition to keep in mind, to avoid being too specific, is that an I(0) process should be amenable to the methods of statistical analysis applied up to this point in the book. In other words, it should obey the law of large numbers and central limit theorem, when suitably centred and normalized. As we now show, I(1) processes fail this test and have distinctive non-standard asymptotic properties.

If u_t is i.i.d. (14.1.1) is known as a *random walk*, and this simple case is a good one to start with. With $x_0 = 0$ the model $x_t = x_{t-1} + u_t$ solves as

$$x_t = \sum_{s=1}^{t} u_s. \tag{14.1.2}$$

Since the u_t are uncorrelated,

$$E(x_t x_{t+s}) = t\sigma^2 \text{ for } s = 0, 1, 2, 3, \dots, \ t \geq 1. \tag{14.1.3}$$

The variance of the process (the case $s = 0$) is therefore tending to infinity as $t \to \infty$, and the autocovariances do not tend to 0 as the time separation increases.

Next consider the sample mean $\bar{x}_n = n^{-1} \sum_{t=1}^{n} x_t$. Substitution yields

$$\bar{x}_n = \frac{1}{n} \sum_{t=1}^{n} \Big(\sum_{s=1}^{t} u_s \Big) = \sum_{t=1}^{n} \Big(\frac{n-t+1}{n} \Big) u_t. \tag{14.1.4}$$

All the weights in this sum of random terms lie between 0 and 1. These terms are martingale differences that satisfy the conditions of Theorem 6.2.3 if this is true of the u_t themselves, so consider the behaviour of the random variable $n^{-1/2}\bar{x}_n$ in the limit. Its mean is 0, and its variance is[1]

$$E\Big(\frac{\bar{x}_n}{\sqrt{n}} \Big)^2 = \frac{\sigma^2}{n} \sum_{t=1}^{n} \Big(\frac{n-t+1}{n} \Big)^2 \to \frac{\sigma^2}{3}. \tag{14.1.5}$$

Applying an appropriate CLT shows that

$$\frac{\bar{x}_n}{\sqrt{n}} \xrightarrow{\text{D}} \text{N}\Big(0, \frac{\sigma^2}{3} \Big). \tag{14.1.6}$$

The sample mean itself is therefore diverging.

Although this particular result yields an asymptotic normal distribution, this is the attribute of the partial sums of x_t scaled by $n^{-3/2}$, *not* $n^{-1/2}$. To extend the analysis further calls for a new kind of limit theory.

14.2 The Probability Background

14.2.1 Function Spaces

The most important departure is to transform the probability model from an infinite sequence of random variables to a random function on the closed interval $[0, 1]$. Let an I(1) time series, more formally called a partial-sum process, be $S_t = x_t - x_0 = \sum_{s=1}^{t} u_s$ for $t = 1, \ldots, n$. Define a function from $[0, 1]$ to $\mathbb{R}$ by

$$Y_n(r) = \begin{cases} \dfrac{S_{j-1} + (nr - j - 1)u_j}{\sqrt{n}\sigma}, & (j-1)/n \le r < j/n, \\ & j = 1, \ldots, n \\ \dfrac{S_n}{\sqrt{n}\sigma} & r = 1. \end{cases} \tag{14.2.1}$$

This may be written more compactly as

$$Y_n(r) = \frac{S_{[nr]} + (nr - [nr])u_{[nr]+1}}{\sqrt{n}\sigma} \quad 0 \le r < 1 \tag{14.2.2}$$

where $[nr]$ is the largest integer not exceeding nr. In graphical terms, this transformation just represents the joining up of dots with line segments. Figure 14.1

[1] See the second of the useful formulae in the footnote on page 149.

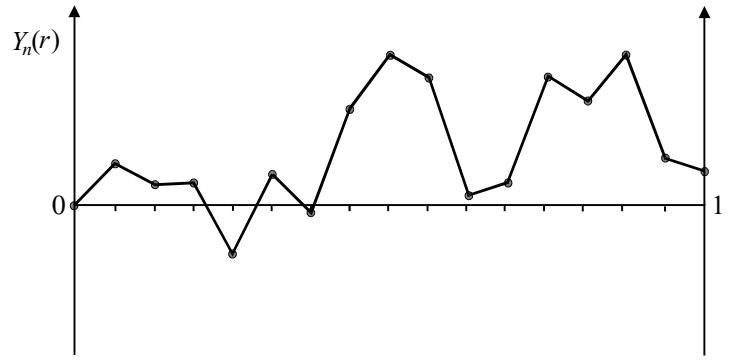

Figure 14.1: An Element of $C[0,1]$

provides a simple illustration, the vertices at points j/n on the horizontal axis representing the points $S_j/\sqrt{n}\sigma$. This resembles the usual sort of time series plot of points $S_1, \ldots, S_n$, and while the difference is crucial, it lies only in the axes. The horizontal axis is a continuum instead of representing n discrete points, and the vertical axis is re-scaled by the $1/\sqrt{n}\sigma$. Because the set $(u_1, \ldots, u_n)$ are random variables, Y_n can be thought of as a random drawing from the sample space $C[0,1]$ of *all possible* continuous functions on $[0,1]$. The interesting departure is to consider the distribution of these functions, and in particular, the limit of this distribution, if such can be shown to exist, as $n \to \infty$. Since the width of the interval is fixed, increasing n crams more and more points into the same space until, ultimately, their horizontal separation vanishes to give a curve that is 'all corners'. At the same time the variation of the partial sums is normalized so that the vertical scale can remain unchanged. This promises to be a limit of a rather interesting sort.

To handle this problem mathematically some structure must be imposed on the space $C[0,1]$, to be known simply as C when there is no risk of ambiguity. Let an element of C (that is, any continuous curve traversing the unit interval) be denoted x. The real numbers $x(r)$ representing the unique values of x at points $r \in [0,1]$ are called the *coordinates* of x. Given any two elements of the space, say $x \in C$ and $y \in C$, we need to be able to say how close together they are. Technically, C must be assigned a *metric*. For example, for any pair of real numbers x and y a distance between them is defined, the nonnegative number $d_E(x,y) = |x - y|$. d_E is known as the *Euclidean* metric, and the pair $(\mathbb{R}, d_E)$ (the set of points paired with the distance measure) defines a *metric space*, in this case Euclidean space. To be a metric, the function d must satisfy the properties

$$d(x,y) \geq 0, \text{and } d(x,y) = 0 \text{ if and only if } x = y \tag{14.2.3a}$$

$$d(y,x) = d(x,y) \tag{14.2.3b}$$

$$d(x,y) \leq d(x,z) + d(y,z). \tag{14.2.3c}$$

It can be verified that the Euclidean metric on $\mathbb{R}$ does so. In fact this is not the

only way to define a metric on $\mathbb{R}$. The *discrete* metric, taking the value 0 if $x = y$ and 1 otherwise, also satisfies conditions (14.2.3).

For elements of C there are also several ways to define a metric. A convenient choice is the *uniform* metric,

$$d_U(x, y) = \sup_{0 \le r \le 1} |x(r) - y(r)| \tag{14.2.4}$$

This is just the largest vertical separation between the pair of functions over the interval, and can be verified to satisfy the definition. (C, d_U) is a metric space. Intuitively, is helpful to think of metric spaces as sharing certain essential properties with the real numbers. In particular, there is the generalization of the notion of an open interval. Sets of the form $\{x : d_u(x, y) < r\}$ can be defined, called the *open spheres* with centre y and radius r. The σ-field generated from the open spheres of a metric space is called the Borel field of the space, the generalization of the concept defined in §B.3 for the real line. The Borel field of (C, d_U) is denoted $\mathcal{B}_C$, and probabilities can be assigned to the elements of $\mathcal{B}_C$ to define a probability space $(C, \mathcal{B}_C, \mu)$. In many ways this is analogous to the familiar distribution on the line, although unlike $\mathbb{R}$ there is no ordering of elements defined on C, and hence no natural counterpart of the c.d.f. to represent μ.

Now let us focus attention on elements of C such as the one illustrated in Figure 14.2. The key idea is that these have a distribution derived from the joint distribution of the elements $u_1, \ldots, u_n$. For example, this might be designated as i.i.d. with mean 0 and variance σ^2. Let μ_n denote the distribution in question, such that $\mu_n(A)$ represents the probability of the curve lying in the set A, for each $A \in \mathcal{B}_C$.

Now consider the sequence $\{\mu_n, n = 1, 2, 3, \ldots\}$, to correspond with increasing sample sizes, and a tighter packing of sample points into the unit interval. If there exists a limiting distribution μ such that $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{B}_C$, possibly excepting cases where $\mu(\partial A) > 0$ where ∂A denotes the boundary points of A, μ_n is said to *converge weakly* to μ, denoted $\mu_n \Longrightarrow \mu$. This is the natural generalization of Definition 3.1.1 for distributions on $\mathbb{R}$. The earlier stipulation of convergence only at 'continuity points of F' specializes the condition just given, allowing the convergence to fail for sets whose boundaries have positive measure.

14.2.2 Brownian Motion

If $u_t \sim$ i.i.d.$(0, \sigma^2)$, something is already known about the postulated convergence. Since the increments u_t are small relative to their sums from 0 to t according to (14.1.5), the limit curves should retain their continuity. The interesting thing is that, as a consequence of the central limit phenomenon, their evolution follows a normal law. Start with the case $Y_n(1)$. Looking at (14.2.1), the Lindeberg–Lévy CLT implies that $Y_n(1) = S_n/\sqrt{n}\sigma \xrightarrow{\text{D}} \text{N}(0, 1)$ on the stated assumptions. However if $n \to \infty$, so does nr for any $r > 0$. It is therefore also the case that

$$\frac{S_{[nr]}}{\sqrt{nr}\sigma} \xrightarrow{\text{D}} \text{N}(0, 1) \tag{14.2.5}$$

and hence

$$Y_n(r) = \sqrt{r}\frac{S_{[nr]}}{\sqrt{nr}\sigma} + \frac{(nr - [nr])u_{[nr]+1}}{\sqrt{n}\sigma} \xrightarrow{\text{D}} \text{N}(0, r) \tag{14.2.6}$$

noting that the second of the two terms is $O_p(n^{-1/2})$. Pursuing the same logic further still, one can say using the independence of the increments that

$$Y_n(r_2) - Y_n(r_1) \xrightarrow{\text{D}} \text{N}(0,\ r_2 - r_1) \tag{14.2.7}$$

for all $0 \leq r_1 < r_2 \leq 1$. To see this, let Y represent the limit of Y_n and note that, for example, $Y(\frac{1}{2})$ and $Y(1) - Y(\frac{1}{2})$ are a pair of independent normals each with variance of $\frac{1}{2}$, given that $Y(0) = 0$ with probability 1. Their sum $Y(1)$ clearly has a variance of 1, as required.

A random function that possesses these properties of Y is known as a *Wiener process* or *Brownian motion process.* Formally:

Definition 14.2.1 A Brownian motion B is a real random function on the unit interval,[2] with the following properties:

(a) $B \in C$ with probability 1.

(b) $B(0) = 0$ with probability 1.

(c) for any set of subintervals defined by arbitrary $0 \leq r_1 < r_2 < \ldots < r_k \leq 1$, the increments $B(r_1)$, $B(r_2) - B(r_1)$, $\ldots$, $B(r_k) - B(r_{k-1})$ are totally independent.

(d) $B(t) - B(s) \sim \text{N}(0, t - s)$ for $0 \leq s < t \leq 1$. □

Given property (a), B can be thought of as a random drawing from the probability space $(C, \mathcal{B}_C, W)$, where W is the probability measure that assigns probabilities to B in accordance with the definition, called *Wiener measure.*

Although continuous, the sample paths of the Wiener process are extremely tortuous. They are curves of the class named *fractals* by Mandelbrot (1983). Even when traversing a finite interval they are curves of infinite length, for note that

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}|u_j| = O_p\big(\sqrt{n}\big). \tag{14.2.8}$$

Almost every point is a corner, and they are nondifferentiable at every point of $[0, 1]$ with probability 1. To see this, note that

$$\frac{B(r+h) - B(r)}{h} \sim \text{N}(0, h^{-1}) \quad \text{all } h > 0. \tag{14.2.9}$$

Letting h tend to 0 does not yield a well-defined limit in distribution.

[2] Brownian motion can be defined on any interval of the line, including $[0, \infty)$. The present definition is the relevant one for our purposes.

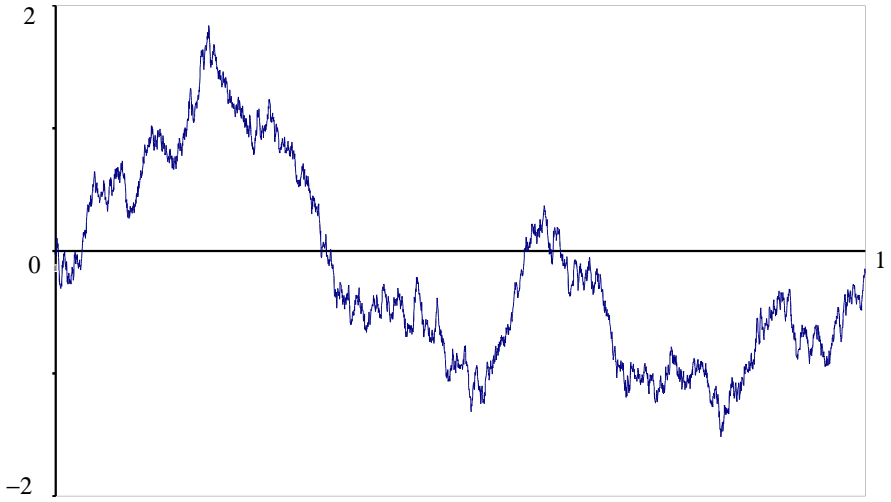

Figure 14.2: A Realization of B

Figure 14.2 gives an idea of the appearance of these sample paths,[3] and may be compared with the finite realization of an integrated process shown in Figure 5.3. Realizations of B have the interesting property of exhibiting purposeful-seeming local trends. At any point the curve rises or falls with the same probability, and S_n has zero mean for every n, but it can keep the same sign for very long periods.

14.2.3 The Functional CLT

The basic weak convergence result, to be thought of as the generalization to a function space of the Lindeberg–Lévy CLT, is *Donsker's Theorem* (Donsker, 1951).

Theorem 14.2.1 If $\{u_j\ ,\ j = 1, 2, 3, \dots\}$ is i.i.d.$(0, \sigma^2)$, then $Y_n \xrightarrow{\text{D}} B$. □

This is a case of the *functional central limit theorem* (FCLT). The term *invariance principle* is also used to refer to a result in which the weak limit of a stochastic process is invariant to the distribution of the underlying components.

The proof of the FCLT involves two fundamental steps, one of which is obvious in view of the earlier discussion. This is to show that the finite-dimensional distributions of the process (joint distributions of finite collections of coordinates) converge to those of the limit process. As remarked, the Lindeberg–Lévy theorem achieves this in the present case. However, the convergence of the finite dimensional distributions does *not* suffice by itself. These do not define all the

[3] The figure was constructed by plotting 3600 points of a partial sum of independent $\text{N}(0, 1)$ increments. Since this number approaches the maximum resolution of the graphics software, adding more points would not change the overall appearance of the curve.

characteristics of a continuous stochastic process. To cite just one example, they do not reveal what the distribution of $\max_{0\le r\le 1}|Y(r)|$ is.

However, if the sequence of distributions $\{\mu_n,\ n=1,2,3,\dots\}$ does possess a well defined limit, this can only be W. The problem is to demonstrate that the limit exists, and it will do so only if the sequence of distributions is *uniformly tight*. Tightness of a distribution is the condition that prohibits the escape of probability mass to infinity. For example, consider the sequence of distributions on $\mathbb{R}$ defined by (14.2.9) with $h=1/n$. The distributions $\{\mathrm{N}(0,n),\ n=1,2,3,\dots\}$ are tight for each n, but the limit of the sequence would appear to assign positive probability to infinitely large values. In fact, it is not a well-defined distribution, and this sequence is not uniformly tight. In the context of a sequence of probability measures on C, a tight distribution, roughly speaking, is one which is both bounded at the origin and assigns arbitrarily small probability to functions with discontinuities. Fortunately, uniform tightness of the distributions of the Y_n is readily shown by a technical argument under the condition of i.i.d. increments, completing the proof of the Donsker FCLT.

The foregoing approach to the problem is neat in principle, but algebraically rather untidy. In (14.2.6), it might seem desirable to omit the asymptotically negligible term at the outset, and working directly with the simple partial-sum function X_n where

$$X_n(r)=\frac{S_{[nr]}}{\sqrt{n}\sigma}\quad 0\le r\le 1 \tag{14.2.10}$$

in place of Y_n. This function is drawn in Figure 14.3 for comparison with the continuous version in Figure 14.1. The substitution looks trivial but, unfortunately,

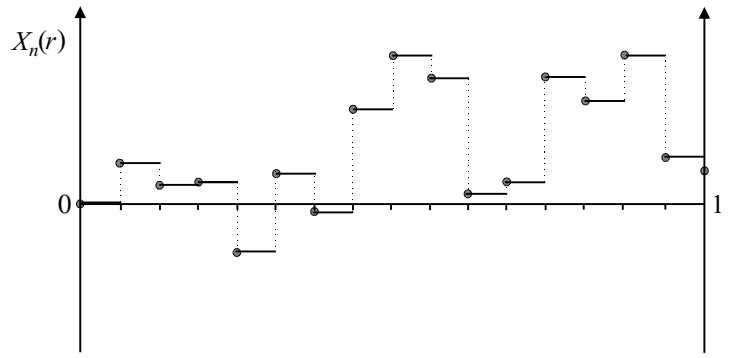

Figure 14.3: An Element of $D[0,1]$

is not as straightforward as it might appear. The problem is that X_n is not an element of $C[0,1]$. Technically, it belongs to the space $D[0,1]$ of *cadlag* functions, to be written simply as D when there is no risk of ambiguity.[4] Such functions may

[4] This is a French acronym, standing for 'continue à droite, limites à gauche'.

have discontinuities, although every *decreasing* sequence of points on a function segment has a limit point, and a small enough *increase* in r never need result in a jump in $X_n(r)$; whereas neither of these statements is true if the directions are interchanged. The space C forms a subset of D, and hence in particular a Brownian motion is an element of D. In this case condition (a) is an important component of Definition 14.2.1.

It is possible to make D a metric space by associating the uniform metric with it, but here is where the problems arise. The space (D, d_U) is *nonseparable*. A separable metric space is one that contains a countable,[5] dense subset, a subset being *dense* if every point of the space is arbitrarily close to one of its elements. For example, $\mathbb{R}$ is separable because the rational numbers are both countable and dense in $\mathbb{R}$. It is possible to show that (C, d_U) is also separable. However, if

$$x_\theta(r) = \begin{cases} 0, \ 0 \leq r < \theta \\ 1, \ \theta \leq r \leq 1 \end{cases} \tag{14.2.11}$$

the functions x_θ are elements of D for each $\theta \in [0, 1]$. There are an uncountable number of these, but $d_U(x_\theta, x_{\theta+\delta}) = 1$ for *every* $0 < \delta \leq 1 - \theta$. So no countable subset of (D, d_U) can exist that is arbitrarily close to all of these functions. The problem with nonseparability is that the Borel field of (D, d_U) contains 'too many sets'. It is not possible to associate probabilities with all these sets without running into contradictions.

The favoured solution to the nonseparability problem is to adopt a different metric for D. Common sense suggests that x_θ and $x_{\theta+\delta}$ should be thought of as close when δ is small. A suitable metric with this property has been constructed by Billingsley (1968).[6] Letting d_B denote Billingsley's metric, and $\mathcal{B}_D$ the associated Borel field, probability spaces of the form $((D, d_B), \mathcal{B}_D, \mu)$ are well-defined. With this set-up, the asymptotic analysis of the processes X_n can proceed just as for the continuous Y_n, with a worthwhile gain in simplicity.

For econometric problems we would also like to extend the FCLT beyond the case of i.i.d. increments. A route can be followed similar to the one pursued in Chapter 6. The following result has conditions nearly the same as those for the CLT, but with a crucial extra condition:

Theorem 14.2.2 Let $S_n = \sum_{t=1}^n u_t$, where the sequence $\{u_t, \, t = 1, \ldots, n\}$ has mean of zero, variances $\sigma_t^2 < \infty$, and meets the conditions specified for $\{X_t\}$ in at least one of Theorems 3.3.1, 3.3.2, 6.2.3 and 6.4.5. In addition, assume

$$\frac{E(S_n^2)}{n} \to \sigma^2 < \infty. \tag{14.2.12}$$

If $X_n(r) = S_{[nr]}/\sqrt{n}\sigma$, then $X_n \xrightarrow{\text{D}} B$. □

[5] An infinite set is countable if its elements can be labelled by the natural numbers $n = 1, 2, 3, \ldots$ The points of the line $0 \leq r \leq 1$ are uncountable, representing a higher order of infinity.

[6] Billingsley's metric has the technical property of inducing the so-called *Skorokhod topology* on the space.

A proof is given in Davidson (1994a), Theorems 27.14 and 29.6.[7] Condition (14.2.12) is *global* wide-sense stationarity. Notice that σ^2 is not the same thing as $\sigma_u^2 = \lim_{n\to\infty} n^{-1}\sum \sigma_t^2$ unless the increments are uncorrelated. Otherwise, all the terms $E(u_t u_{t-j})$ for $j > 0$ and $1 \leq t \leq n$ are included in the sum. This extra condition is sufficient for uniform tightness of the distributions, and ensures that

$$\frac{S_{[nr]}}{\sqrt{n}\sigma} \xrightarrow{\text{D}} \text{N}(0, r) \tag{14.2.13}$$

for each r in $(0, 1]$. For the ordinary CLT this need hold only for the case $r = 1$, and technically (14.2.12) need not hold, though it is a fairly innocuous condition and would be expected to hold in most applications of the CLT in any case.

Always paired with the FCLT is the function-space version of the continuous mapping theorem (Theorem 3.1.3). This following is a special case of Davidson (1994a) Theorem 26.13.

Theorem 14.2.3 Let $h : D \mapsto \mathbb{R}$ be a measurable functional[8] that is continuous except on a set $D_h \subset D$ with $P(B \in D_h) = 0$. If $X_n \xrightarrow{\text{D}} B$ then $h(X_n) \xrightarrow{\text{D}} h(B)$. □

Examples of continuous functionals of X_n include the coordinate functions $X_n(r)$ for any fixed r, and also integrals over $[0, 1]$, of X_n and of transformations such as X_n^2.

14.3 The Unit Root Autoregression

14.3.1 Basic Convergence Results

Consider the asymptotic properties of the autoregression in the model

$$x_t = \lambda x_{t-1} + u_t \quad \lambda = 1. \tag{14.3.1}$$

For the main development, it is not necessary to make any assumptions about u_t beyond what is specified in Theorem 14.2.2. The initial value x_0 is assumed to be observed, so that the error-of-estimate takes the form

$$\hat{\lambda} - 1 = \frac{\sum_{t=1}^n u_t x_{t-1}}{\sum_{t=1}^n x_{t-1}^2} = \frac{\sum_{t=1}^n (x_t - x_{t-1}) x_{t-1}}{\sum_{t=1}^n x_{t-1}^2}. \tag{14.3.2}$$

If in fact $x_0 = 0$, this is equivalent to summing from 2 to n. Recall that $S_t = x_t - x_0$ and $X_n(r) = n^{-1/2} S_{[nr]}/\sigma$. Note that $\int_{(t-1)/n}^{t/n} dr = n^{-1}$, and therefore

$$S_{t-1} = nS_{t-1}\int_{(t-1)/n}^{t/n} dr = n\int_{(t-1)/n}^{t/n} S_{[nr]} dr = n^{3/2}\sigma \int_{(t-1)/n}^{t/n} X_n(r) dr \tag{14.3.3}$$

[7] De Jong and Davidson (2000b) prove the FCLT under even weaker conditions than these.

[8] A functional is function that takes another function as its argument. The integral of a function is a familiar example.

and

$$S_{t-1}^2 = nS_{t-1}^2 \int_{(t-1)/n}^{t/n} dr = n \int_{(t-1)/n}^{t/n} S_{[nr]}^2 dr = n^2\sigma \int_{(t-1)/n}^{t/n} X_n(r)^2 dr. \quad (14.3.4)$$

Put $S_0 = 0$, and then

$$\frac{1}{n^{3/2}} \sum_{t=1}^{n} S_{t-1} = \sigma \sum_{t=1}^{n} \int_{(t-1)/n}^{t/n} X_n(r)dr = \sigma \int_0^1 X_n(r)dr$$

$$\xrightarrow{\text{D}} \sigma \int_0^1 B(r)dr \quad (14.3.5)$$

and

$$\frac{1}{n^2} \sum_{t=1}^{n} S_{t-1}^2 = \sigma \sum_{t=1}^{n} \int_{(t-1)/n}^{t/n} X_n(r)^2 dr = \sigma \int_0^1 X_n(r)^2 dr$$

$$\xrightarrow{\text{D}} \sigma^2 \int_0^1 B(r)^2 dr \quad (14.3.6)$$

where σ^2 is defined in (14.2.12). The indicated convergence follows from Theorems 14.2.2 and 14.2.3, noting that the expressions are both continuous functionals of X_n.

The random variables $\int_0^1 Bdr$ and $\int_0^1 B^2 dr$ are called *functionals of Brownian motion.* The convergence in (14.3.5) has already been encountered in a different form as (14.1.6); in other words, it has been shown that $\int_0^1 Bdr \sim \text{N}(0, \frac{1}{3})$.[9] However, no such simple closed form exists for the density function of $\int_0^1 B^2 dr$. Although analytical formulae for the distributions of these statistics can be obtained,[10] these methods are difficult and inflexible. The preferred method of actually tabulating such distributions is by Monte Carlo simulation. To obtain a good approximation to the limiting distribution of $\int_0^1 B^2 dr$, for example, the integral might be approximated by the formula on the left-hand side of (14.3.6), with the S_t obtained as the partial sums of N(0, 1) random numbers, and n chosen suitably large (say 1000). The distribution is built up experimentally from a still larger number of replications of the variable. Percentiles can be estimated quite accurately with 10,000 or more replications. This kind of computation can be done routinely on a fast PC.

14.3.2 Tests of the I(1) Hypothesis

Now consider the limiting distribution of the ratio in (14.3.2) multiplied by a suitable scale factor. First, the denominator. Since $x_t = S_t + x_0$,

$$\frac{1}{n^2} \sum_{t=1}^{n} x_{t-1}^2 = \frac{1}{n^2} \sum_{t=1}^{n} S_{t-1}^2 + \frac{2x_0}{n^2} \sum_{t=1}^{n} S_{t-1} + \frac{x_0^2}{n}. \quad (14.3.7)$$

[9] (14.1.6) was shown for the uncorrelated case, but since (14.3.5) holds generally, the two limits must coincide.

[10] See Evans and Savin (1981), also Tanaka (1996).

The fact that $x_0 n^{-2} \sum_{t=1}^{n} S_{t-1} = O_p(n^{-1/2})$ follows from (14.3.5), and hence the last two terms converge in probability to zero. Invoking Theorem 3.1.2(i) gives the result

$$\frac{1}{n^2} \sum_{t=1}^{n} x_{t-1}^2 \xrightarrow{\text{D}} \sigma^2 \int_0^1 B^2 dr. \tag{14.3.8}$$

Thus, the distribution does not depend on x_0, which may be either a constant or a random variable.

The next task is to consider the numerator in (14.3.2). Since $S_t = S_{t-1} + u_t$,

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^{n} S_{t-1} u_t &= \frac{1}{2n} \sum_{t=1}^{n} (S_t^2 - S_{t-1}^2 - u_t^2) \\ &= \frac{1}{2n} \left(S_n^2 - \sum_{t=1}^{n} u_t^2 \right) \\ &= \frac{\sigma^2}{2} \left(X_n(1)^2 - \frac{1}{n\sigma^2} \sum_{t=1}^{n} u_t^2 \right) \\ &\xrightarrow{\text{D}} \frac{\sigma^2}{2} \left(B(1)^2 - \frac{\sigma_u^2}{\sigma^2} \right) \end{aligned} \tag{14.3.9}$$

where Theorem 6.4.4, for example, is used to show that $\operatorname{plim} n^{-1} \sum_{t=1}^{n} u_t^2 = \sigma_u^2$, and Theorem 3.1.2(i) is applied in the final step. It follows from the same sort of reasoning that $\operatorname{plim} n^{-1} \sum_{t=1}^{n} u_t = 0$, and hence that

$$\frac{1}{n} \sum_{t=1}^{n} x_{t-1} u_t = \frac{1}{n} \sum_{t=1}^{n} S_{t-1} u_t + x_0 \frac{1}{n} \sum_{t=1}^{n} u_t \xrightarrow{\text{D}} \frac{\sigma^2}{2} \left(B(1)^2 - \frac{\sigma_u^2}{\sigma^2} \right). \tag{14.3.10}$$

Note that $B(1) \sim \text{N}(0,1)$ by the definition, and hence $B(1)^2 \sim \chi^2(1)$. Also, if u_t is serially uncorrelated then $\sigma^2 = \sigma_u^2$, as noted following Theorem 14.2.2.

It is worth digressing to note that this result provides, in effect, a proof of a famous result in stochastic calculus known as the Itô integral formula. This supplies the generalization to stochastic integrals of the formula for integration by parts. If B is standard Brownian motion, the Itô rule says that

$$\int_0^1 B(r) dB(r) = \tfrac{1}{2}[B(1)^2 - 1] \text{ a.s.} \tag{14.3.11}$$

where the left-hand side of (14.3.11) can be viewed (assuming $\sigma^2 = 1$) as the limit in distribution as $n \to \infty$ of the left hand side of (14.3.10), assuming $\sigma^2 = 1$. In contrast to (14.3.11), the formula for integration by parts applied to the ordinary Stieltjes integral with respect to a function of bounded variation $h(r) : [0,1] \mapsto \mathbb{R}$ (say) would yield

$$\int_0^1 h(r) dh(r) = \tfrac{1}{2} h(1)^2. \tag{14.3.12}$$

The appearance of the extra term of $-\frac{1}{2}$ in the Itô formula is linked to the fact, noted in §14.2.2, that almost all realizations of B are functions of unbounded variation.

Now, returning to equation (14.3.2), a further application of the continuous mapping theorem delivers the required limit result. From (14.3.8) and (14.3.10),

$$n(\hat{\lambda}-1) \xrightarrow{\text{D}} \frac{B(1)^2 - \frac{\sigma_u^2}{\sigma^2}}{2\int_0^1 B^2 dr}. \tag{14.3.13}$$

Contrast (14.3.13) with (6.3.20). In particular, observe the scaling factor of n in the former case, instead of $\sqrt{n}$. $\hat{\lambda}$ is not merely consistent but 'super-consistent'. If the data are serially uncorrelated the limit in (14.3.13) reduces to

$$\frac{B(1)^2 - 1}{2\int_0^1 B^2 dr}. \tag{14.3.14}$$

This is a random variable whose distribution, although nonstandard, does not depend on nuisance parameters and can be tabulated by Monte Carlo simulation. This was done originally by Dickey and Fuller, see Fuller (1976). The statistic $n(\hat{\lambda}-1)$ accordingly provides the basis for a test of $\lambda = 1$. Under the alternative hypothesis of $|\lambda| < 1$, the standard asymptotics for stationary processes apply, $\hat{\lambda}$ is $\sqrt{n}$-consistent for λ, and hence

$$n|\hat{\lambda}-1| = O_p(n). \tag{14.3.15}$$

The test is accordingly consistent.

The distribution of the conventional t statistic for the test of $\lambda = 1$ is found similarly. Writing

$$t_\lambda = \frac{\hat{\lambda}-1}{s_u}\sqrt{\sum_{t=1}^{n} x_{t-1}^2} \tag{14.3.16}$$

where s_u^2 is the least squares residual variance, note that

$$\begin{aligned} s_u^2 &= \frac{1}{n}\sum_{t=1}^{n}(u_t - (\hat{\lambda}-1)x_{t-1})^2 \\ &= \frac{1}{n}\sum_{t=1}^{n} u_t^2 - 2(1-\hat{\lambda})\frac{1}{n}\sum_{t=1}^{n} x_{t-1}u_t + (1-\hat{\lambda})^2\frac{1}{n}\sum_{t=1}^{n} x_{t-1}^2 \\ &\xrightarrow{\text{pr}} \sigma_u^2 \end{aligned} \tag{14.3.17}$$

since the last two terms in the third member of (14.3.17) are $O_p(n^{-1})$ by (14.3.10) and (14.3.8), respectively. It follows that, under the null hypothesis $\lambda = 1$,

$$t_\lambda \xrightarrow{\text{D}} \frac{\frac{\sigma}{2\sigma_u}\left(B(1)^2 - \frac{\sigma_u^2}{\sigma^2}\right)}{\sqrt{\int_0^1 B(r)^2 dr}}. \tag{14.3.18}$$

Again, if the time series is uncorrelated then the limiting distribution reduces to

$$\frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B^2 dr}}. \tag{14.3.19}$$

This is free of nuisance parameters, but as with (14.3.14) is not a standard distribution. As in the case of $n(\hat{\lambda} - 1)$, it must be tabulated by Monte Carlo, and this was also done by Dickey and Fuller.

Thus, one of the basic intuitions of standard regression theory, that the t ratio has a variance standardised to 1 and a normal distribution in the limit, proves inappropriate to this case. The sample standard error that converges to a constant limit in the stable case here converges to a random variable, a functional of Brownian motion. However, it can be verified that under the alternative hypothesis $|\lambda| < 1$,

$$|t_\lambda| = O_p\big(\sqrt{n}\big) \tag{14.3.20}$$

and hence this test is also consistent.

Since the explosive case $\lambda > 1$ is ruled out (in large samples such data must approach infinity at an exponential rate), the $n(\hat{\lambda}-1)$ and t tests of a unit root are always performed as one-tail tests, with the rejection region lying under the left-hand tail. In other words, the null is rejected when the statistics take absolutely large negative values. In practice, the test is often performed by reparameterizing (14.1.1) as

$$\Delta x_t = (\lambda - 1)x_{t-1} + u_t \tag{14.3.21}$$

where $\Delta x_t = u_t$ by hypothesis. The OLS estimate and t ratio of the regression coefficient in (14.3.21) are identical to $\hat{\lambda} - 1$ and t_λ respectively. The tests based on these two statistics are the alternative versions of the so-called Dickey–Fuller test for a unit root (Dickey and Fuller 1979, 1981).

14.3.3 Serial Correlation

For the more usual case when u_t is serially correlated, two approaches have been developed. The most popular procedure is to try to model the correlation in u_t by fitting an autoregression. Instead of (14.3.21), run the regression

$$\Delta x_t = (\lambda - 1)x_{t-1} + \sum_{j=1}^{k} \beta_j \Delta x_{t-j} + u_t \tag{14.3.22}$$

assuming that for a suitable value of k, u_t so defined is an uncorrelated sequence. This is the augmented Dickey–Fuller (ADF) test. The following is the basic result needed to justify it.

Theorem 14.3.1 If (14.3.22) holds where u_t meets the conditions specified in Theorem 14.2.2, and is also serially uncorrelated, (14.3.14) and (14.3.19) are the limiting distributions of the ADF statistics $n(\hat{\lambda} - 1)$ and t_λ respectively. □

See §14.6 for the proof. The problem with this approach is to choose k appropriately. If the process generating Δx_t has moving average as well as autoregressive components, the true k is actually infinite and Theorem 14.3.1 does not apply. However, if k is made a suitable increasing function of n, it can be shown that t_λ does converge to the DF distribution under H_0. Said and Dickey (1984) show that if Δx_t is ARMA(1,1), $k = o(n^{1/3})$ is appropriate. In practice, *ad hoc* methods such as choosing k on the basis of model selection criteria (see §9.4.1) are often adopted. It should be borne in mind that over-parameterization as well as neglect of autocorrelation can affect the performance of the test in finite samples.

An alternative approach due to Phillips and Perron (see Phillips 1987, Phillips and Perron 1988) is to treat the autocorrelation nonparametrically. The nuisance parameters in (14.3.13) and (14.3.18) are estimated consistently using the methods of §9.4.3. Applying the Newey and West (1987) suggestion to use the Bartlett kernel leads to the estimator

$$s_{nl}^2 = \frac{1}{n}\left(\sum_{t=1}^{n} \hat{u}_t^2 + 2\sum_{\tau=1}^{l(n)} w_{\tau l} \sum_{t=\tau+1}^{n} \hat{u}_t \hat{u}_{t-\tau}\right) \tag{14.3.23}$$

where $\hat{u}_t = x_t - \hat{\lambda} x_{t-1}$, $w_{\tau l} = 1 - \tau/(l(n)+1)$ and $l(n) = O(n^{1/3})$ is recommended, see §9.4.3 for details.

This estimator is always positive, and $s_{nl}^2 \xrightarrow{\text{pr}} \sigma^2$ under H_0. Thus, consider the Dickey–Fuller t statistic. It can be verified that inverting the formula in (14.3.18) gives

$$\frac{\sigma_u}{\sigma} t_\lambda + \frac{\sigma_u^2 - \sigma^2}{2\sigma^2 \sqrt{\int_0^1 B^2 dr}} \xrightarrow{\text{D}} \frac{B(1)^2 - 1}{2\sqrt{\int_0^1 B^2 dr}} \tag{14.3.24}$$

The sample counterpart is the statistic

$$\hat{Z}_t = \frac{s_u}{s_{nl}} t_\lambda + \frac{s_u^2 - s_{nl}^2}{2 s_{nl} \sqrt{n^{-2} \sum_{t=1}^{n} x_{t-1}^2}} \tag{14.3.25}$$

which has the same asymptotic distribution as (14.3.24).

While the formula in (14.3.25) shows how to modify the t statistic, it might be simpler to see how it is constructed directly from its constituent elements. Since

$$t_\lambda = \frac{\sum_{t=1}^{n} x_{t-1} \hat{u}_t}{s_u \sqrt{\sum_{t=1}^{n} x_{t-1}^2}}, \tag{14.3.26}$$

the modified statistic is

$$\hat{Z}_t = \frac{\sum_{t=1}^{n} (x_{t-1} \hat{u}_t - \hat{c})}{s_{nl} \sqrt{\sum_{t=1}^{n} x_{t-1}^2}} \tag{14.3.27}$$

where $\hat{c} = n^{-1} \sum_{\tau=1}^{l(n)} w_{\tau l} \sum_{t=\tau+1}^{n} \hat{u}_t \hat{u}_{t-\tau}$ from (14.3.23). Notice that $\hat{c}$ is actually the estimate of $E(x_{t-1} u_t)$, and so can be seen as a centring correction. The modifications to t_λ represented by $\hat{Z}_t$ are therefore two, the mean correction and the replacement of s_u by the consistent estimator s_{nl}. Similar corrections can be applied to the $n(\hat{\lambda} - 1)$ statistic, and it is left to the interested reader to construct the modifications for this case.

14.3.4 Including an Intercept

It is usual to run a regression such as (14.3.21) including an intercept. This was not assumed in §14.3.2 and including an intercept changes the asymptotic distributions, although both estimators are consistent. The OLS estimator of $\lambda - 1$ becomes

$$\hat{\lambda} - 1 = \frac{\sum_{t=1}^{n}(x_{t-1} - \bar{x}_{-1})u_t}{\sum_{t=1}^{n} x_{t-1}^2 - n\bar{x}_{-1}^2} \tag{14.3.28}$$

where $\bar{x}_{-1} = n^{-1}\sum_{t=1}^{n} x_{t-1}$. Note that the mean deviations do not depend on x_0.

Starting with the denominator, expanding x_{t-1} yields

$$\begin{aligned}\frac{1}{n^2}\Big(\sum_{t=1}^{n} x_{t-1}^2 - n\bar{x}_{-1}^2\Big) &= \frac{1}{n^2}\sum_{t=1}^{n} S_{t-1}^2 - \Big(\frac{1}{n^{3/2}}\sum_{t=1}^{n} S_{t-1}\Big)^2 \\ &\xrightarrow{\text{D}} \sigma^2\int_0^1 B^2 dr - \Big(\sigma\int_0^1 B dr\Big)^2 .\end{aligned} \tag{14.3.29}$$

Also, extending (14.3.10),

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}(x_{t-1} - \bar{x}_{-1})u_t &= \frac{1}{n}\sum_{t=1}^{n} S_{t-1}u_t - \frac{1}{n^{3/2}}\sum_{t=1}^{n} S_{t-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} u_t \\ &\xrightarrow{\text{D}} \frac{\sigma^2}{2}\Big(B(1)^2 - \frac{\sigma_u^2}{\sigma^2} - 2\int_0^1 B dr \cdot B(1)\Big).\end{aligned} \tag{14.3.30}$$

Thus, with the data expressed in sample mean deviations the null distribution of $n(\hat{\lambda}-1)$ is given by the ratio of (14.3.30) to (14.3.29), which differs from (14.3.13). Similarly,

$$t_\lambda = \frac{\hat{\lambda} - 1}{s}\sqrt{\sum_{t=1}^{n} x_{t-1}^2 - n\bar{x}_{-1}^2} \xrightarrow{\text{D}} \frac{B(1)^2 - \frac{\sigma_u^2}{\sigma^2} - 2B(1)\int_0^1 B dr}{2\frac{\sigma_u}{\sigma}\sqrt{\int_0^1 B^2 dr - \Big(\int_0^1 B dr\Big)^2}} \tag{14.3.31}$$

which is different from (14.3.18). Some simplification of these expressions is possible by noting that

$$\int_0^1 B^2 dr - \Big(\int_0^1 B dr\Big)^2 = \int_0^1 B^{*2} dr \tag{14.3.32}$$

where

$$B^* = B - \int_0^1 B dr \tag{14.3.33}$$

which is known as a *de-meaned* Brownian motion.

These distributions have also been tabulated by Dickey and Fuller, so that one has the choice of estimating either with or without an intercept provided the correct tabulation is used in each case. The corrections for serial correlation, either parametric or nonparametric, can be carried out as usual, and the data in mean deviation form can be substituted for the raw data to compute the variance in the latter case. The limiting distributions apply as before.