MACROECONOMETRICS, Spring 2022.

Bent E. Sørensen

1 Unit Roots.

A good place to look after you have read this section is the survey in *Handbook of Econometrics* Vol. IV by Jim Stock although there are many good textbook treatments now, like Bruce Hansen's forthcoming book or (more complicated) James Hamilton's Time Series book.

In the statistical literature it has long been known that unit root processes behave differently from stable processes.

For example in the scalar AR(1) model, consider the distribution of the OLS estimator of the parameter a in the simple first order process,

$$(1)\ y_t \ = \ a\ y_{t-1} \ + \ e_t\ .$$

If e_t are independently identically normally distributed (niid) variables and a_N denotes the least squares estimator of b based on $y_0, y_1, .., y_N$, then Mann and Wald (1943) showed that $N^{1/2}(a_N - a)$ has a limiting normal distribution if $a < 1$. White (1958) showed that $|\,a\,|\ N(a^2-1)\ (a_N - a)$ has a limiting Cauchy distribution if $a > 1$, whereas $N(a_N - 1)$ has a limiting distribution that can be written in terms of a ratio of two functionals of a Wiener process, when $a = 1$. In the later years a lot of theoretical work has been

done on the distribution of least squares estimators in the presence of unit roots. Some notable early contributions are Fuller (1976), Dickey and Fuller (1979,1981), and Evans and Savin (1981,1984). The authoritative paper by Phillips (1987) sums up most of the theory.

It was the paper by Nelson and Plosser (1982) that sparked the huge surge in interest for unit root models among economists. They examined time series for some of the most important U.S. aggregate economic variables and concluded that almost all them were better described as being integrated of order one rather than stable. (They further went on to suggest that this favored real-business-cycle type of classical models in favor of monetarist and Keynesian models. In my opinion it is plain crazy to try and derive such sweeping conclusions from atheoretical time series modeling, and it is even crazier to do so on aggregate data; but maybe the huge interest that this paper generated is partly due to this provocative statement).

The reason why unit roots are so important is that the limiting distributions of estimates and test statistics are very different from the stationary case. Most importantly, one can not (in general) obtain limiting χ^2 (or t- or F-) distributions. Rather one obtains limiting distributions that can be expressed as functionals of Brownian motions. Also notice that in the case where you have a stable model that is "close" to an inte-

grated process, then the distributions of estimators will look more like the distributions from unit root models than it will look like the asymptotic (normal type) distribution in small samples. This phenomenon is treated in the literature under the heading of near-integrated (or near unit-root, or nearly non-stationary) models. We may not have time to go into this; but the reading list contains a quite detailed bibliography (since I happened to have written a paper in that area - I had the references typed up already). I personally doubt whether many series in economics are best thought of as genuinely non-stationary, and I don't think that one really can decide that on the basis of the statistical evidence (there has been written lots of papers on that question since the influential paper of Nelson and Plosser (1982)). My point of view is that it does not really matter. The models will have very different predictions for the very long run whether they truly have unit roots or not; but to cite Lord Keynes: "In the long run we are all dead"; or to say it less dramatically - I do not think that simple time series models are useful for forecasting 20 years ahead under any circumstances. What matters is that the small sample distributions look like the asymptotic unit root distributions, so if you do not use those you will make wrong statistical inferences.

1.1 Brownian Motions and Stochastic Integrals.

The easiest way to think of the Brownian motions is in the following way (which corresponds exactly to the way that you will simulate Brownian motions on the computer):

Let

$$B_N(t) \ = \frac{1}{\sqrt{N}} \ (\ e_1 \ + \ e_2 \ + \ ... \ + e_{[Nt]}) \ ; \ t \in [0,T] \ ,$$

where $e_1, ..., e_{[NT]}$ are iid $N(0,1)$. The notation $[Nt]$ means the integer part of Nt, i.e. the largest integer less than of equal to Nt. Note that $B_N(t)$ is a stochastic function from the closed interval $[0,T]$ to the real numbers. If N is large $B_N(.)$ is a good approximation to the Brownian motion $B(t); t \in [0,T]$ which is defined as

$$B(t) \ = \ \lim_{N \to \infty} B_N(t) \ .$$

For a fixed value of t it is obvious that $B_N(t)$ converges to a normally distributed random variable with mean zero and variance t. To show that $B_N(t)$ converges as a function to a continuous function $B(t)$ takes a large mathematical apparatus. You can find that in the classical text of Billingsley (1968); but be warned that this is a book written for mathematicians (but given that it is very well written).

For the purpose of the present course this is all you need to know about the Brownian motion. One can show that any stationary continuous stochastic process $B(t)$ for which

$$B(t_4) - B(t_3) \ \text{ and } \ B(t_2) - B(t_1)$$

are independent for all $t_4 \geq t_3 \geq t_2 \geq t_1$ and with

$$E\{B(t_2) - B(t_1)\} \ = \ 0 \ , \text{ and } \ Var\{B(t_2) - B(t_1)\} \ = \ t_2 - t_1 \ ;$$

has to be a Brownian Motion. The Brownian motion is an example of a *process with identical independent increments*. You can convince yourself from the definition I gave

of Brownian motion, that this formula for the variance is true. Notice that if you start the process at time 0 then

$$Var(B(t)) \;=\; t \;.$$

So it is obvious that the unconditional variance of $B(t)$ tends to infinity as t tends to infinity. This corresponds to the behavior of the discrete time random walk, which again is just an AR(1) with an autoregressive coefficient of 1. So it is not surprising that Brownian motions show up in the (properly normalized) asymptotic distribution of estimators of AR models. Brownian motions are quite complicated if you look into some of the finer details of their sample paths. One can show that Brownian motions with probability 1 are only differentiable on a set of measure zero. You can also show that a Brownian motion that you start at zero at time zero will cross the x-axis infinitely many times in any finite interval that includes 0. Properties like those are very important in continuous time finance, so if you want to specialize in that field you should go deeper into the theory of Brownian motions. I have supplied a list of references in the reading list, that will be useful for that purpose; but for unit root theory this is not absolutely necessary.

You will also need to be able to integrate with respect to Brownian motions. We want to give meaning to the symbol $\int_0^1 f(s)dB(s)$, where f is a function that will often be stochastic itself. In many asymptotic formulae $f(s)$ is actually the same as $B(s)$. We will define the so-called Ito-integral (named after the Japanese mathematician Kiyosi

Ito):

$$\int_0^1 f(s)dB(s) = \lim_{K\to\infty} \sum_{k=0}^{K} f(\frac{k-1}{K})\Delta B(\frac{k}{K}) \ ,$$

where $\Delta B(\frac{k}{K}) = B(\frac{k}{K}) - B(\frac{k-1}{K})$, and where the limit is *in probability*. You can *not* obtain convergence sample path by sample path almost surely (with probability one). If we temporarily call the stochastic integral on the left hand side for I and the approximating sum on the right hand side for I_K then the convergence in probability means that there exists a stochastic variable I such that

$$\lim_{K\to\infty} P\{|I - I_K| > \epsilon\} = 0$$

for any given $\epsilon > 0$. This is however a probability statement and it does not preclude that I_K for a given sample path can be found arbitrarily far from I for arbitrarily large K. For our purpose that does not really matter, since convergence in probability is all we need. If you want to go through the mathematics (which is detailed in the probability books in the reading list, with the book by Øksendal as the most accessible), then the hard part is to show the existence of the limit I to which I_K converges, and if you don't want to go through the mathematics you should just take it for a fact that the limit exists in a well defined sense.

Notice that the sum in the approximating sum is over values of the function multiplied by a "forward looking" increment in the integrator B. This is essential and you

will not get the right answer if you do not do it like that. This is in contrast to the standard Riemann-Stieltjes integral where this does not matter. The definition of the stochastic integral is given in the way that you can actually simulate the distribution, and this is the way it is done in many Monte Carlo studies in the literature.

Ito's Lemma

The main tool of modern stochastic calculus is Ito's lemma. It is usually formulated as

$$df(X,t) \;=\; \frac{\partial f}{\partial t}dt \;+\; \frac{\partial f}{\partial X}dX \;+\; \frac{1}{2}\frac{\partial^2 f}{\partial X^2}(dX)^2 \;\text{ where }\; dt^2 \;=\; 0 \text{ , and } \; dB^2 \;=\; dt \;.$$

In this course Ito's lemma is not so central; but you may meet the following equality

$$(*) \quad \int_0^1 B(s)dB(s) \;=\; \frac{1}{2}(B(1)^2 - 1) \;,$$

which is a simple consequence of Ito's lemma. Ito's lemma is essential in continuous time finance and in more advanced examinations of unit root theory.

I leave the proof of (*) for the homework.

Example:

Find $d\log(\log B)$. We do this is 2 stages. First we set $X = \log(B)$. Then (since $(dB)^2 = dt$)

$$dX \;=\; \frac{1}{B}dB \;-\; \frac{1}{2}\frac{1}{B^2}dt \;.$$

Then

$$\begin{aligned} d\log(\log B) &= \frac{1}{B}dX - \frac{1}{2}\frac{1}{B^2}(dX)^2 \\ &= \frac{1}{B}(\frac{1}{B}dB - \frac{1}{2B^2}dt) - \frac{1}{2B^2}\frac{1}{B^2}dt \\ &= \frac{1}{B^2}dB - (\frac{1}{2B^3} + \frac{1}{2B^4})dt \end{aligned}$$

(I cannot think of an application of this particular result but it illustrates the method clearly.)

OLS estimation of AR(1) model Now consider the process (1) again. I will give you the intuition (I hope) for why the coefficent in the AR(1) model converges at rate T when it is a random walt and at rate $\sqrt{T}$ when it is stable. Notice that if $a = 1$ then

$$y_t = y_{t-1} + e_t = \sum_{k=1}^{t} e_k \, .$$

For the absolute value of a less than one, notice that

$$y_t = e_t + ae_{t-1} + a^2 e_{t-2} + ... + a^t e_0 \, ,$$

if the process start at 0. This converges to a random variable with mean 0 and variance $\sigma_e^2 \frac{1}{1-a^2}$ for T going to infinity. The least squares estimator is

$$\hat{a} = \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2} = \frac{\sum_{t=1}^{T} (ay_{t-1} + e_t) y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2} = a + \frac{\sum_{t=1}^{T} y_{t-1} e_t}{\sum_{t=1}^{T} y_{t-1}^2}$$

Now notice the numerator in the last term is mean 0 (because e_t is) and the variance of $y_{t-1}e_t$ is limited, so it satisfies a CLT:

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_{t-1}e_t \to N(0,\sigma^2)\,,$$

for some variance σ^2 (which isn't to hard to find). And the denominator satisfies a LLN:

$$\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2 \to \kappa\,,$$

for some constant (which also isn't hard to find). So

$$\sqrt{T}(\hat{a}-a) \;=\; \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T} y_{t-1}e_t}{\frac{1}{T}\sum_{t=1}^{T} y_{t-1}^2} \to X\,,$$

where X is a normal random variable.

When $a = 1$, none of this goes through and we have to normalize differently:

$$T(\hat{a}\;-\;1) \;=\; T\frac{\sum_{t=1}^{T} y_{t-1}\Delta y_t}{\sum_{t=1}^{T} y_{t-1}^2} \;=\; \frac{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(y_{t-1}/\sqrt{T})e_t}{\frac{1}{T}\sum_{t=1}^{T}(y_{t-1}/\sqrt{T})^2} \;=\; \frac{\sum_{t=1}^{T}(y_{t-1}/\sqrt{T})(e_t/\sqrt{T})}{\sum_{t=1}^{T}(y_{t-1}/\sqrt{T})^2\frac{1}{T}}\,.$$

The next to last term shows how the normalization with $\frac{1}{\sqrt{T}}$ keeps the variance of y terms from going to infinity and the last term involves the terms that converges to functions of Brownian motion $e_t/\sqrt{T}$ is the dB in the discrete approximation, and $1/T$ is dt in a standard intergral. So it is not so mysterious (well, maybe less mysterious at least), because for large T, y_T has variance proportional to T, we need do divide y_T by $\sqrt{T}$ to have any hope of converging to something finite. And one can see from the way that we

defined the Brownian motion that $y_{t-1}/\sqrt{T}$ converges to a Brownian motion, and from the way that we defined the stochastic integral, one can see (at least one would guess) that

$$T(\hat{a} \ - \ 1) \ \Rightarrow \ \frac{\int_0^1 B(s)dB(s)}{\int_0^1 B(s)^2 ds} \ .$$

Intuitively you should always think of e_t as $dB(t)$ and y_t which under the null of a unit root is equal to $\sum_{k=0}^{t} e_t$ corresponds then to $\int_0^{t/T} dB(s) = B(s/T)$ (for $B(0) = 0$). From our application of Ito's lemma one can see that another expression would be

$$T(\hat{a} \ - \ 1) \ \Rightarrow \ \frac{1}{2}\frac{B(1)^2 - 1}{\int_0^1 B(s)^2 ds} \ .$$

Notice that $B(1)$ is just a standard normal distribution, so that $B(1)^2$ is just a standard $\chi^2(1)$ distributed random variable. Contrary to the stable model the denominator of the expression for the least squares estimator does not converge to a constant almost surely; but rather to a stochastic variable that is strongly correlated with the numerator. For these reasons the asymptotic distribution does not look like a normal, and it turns our that the limiting distribution of the least squares estimator is highly skewed, with a long tail to the left - look at the graph of the distribution in e.g. Evans and Savin (1981).

OK, if you couldn't quite follow the "derivation" of the limiting distribution, don't despair. One needs quite a bit more machinery and notation to make sure that all the limit operations are legal; but all you need is the basic intuition, so try and get that.

Most of the part of the unit root literature that is concerned with asymptotic theory contains the limiting distribution given here. We will refer to the distribution of

$$\frac{\int_0^1 B(s)dB(s)}{\int_0^1 B(s)^2 ds}$$

as *the unit root distribution*. It is not possible to find any simple expression for the density of this distribution but one can find the characteristic function, which can be inverted in order to tabulate the distribution function—(see Evans and Savin (1981)). You can also evaluate the distribution by Monte Carlo simulation, which is performed by choosing a large value of T, and then drawing the innovation terms from a pseudo random number generator and the generating the series y_t from the defining equation (1). For large T the distribution of the LS estimator is close to the limiting distribution, which can be graphed by repeating this exercise like 10– or 20,000 times and plotting the result.

TS and DS models

If you look at a plot of a typical macro economic time series, like real GNP, it is obvious that it displays are very pronounced *trend*. What is a trend? Well, for many years that question was not considered for many seconds—a trend was simply assumed to be linear function of time, and econometricians would routinely "detrend" their series by using the residuals from a regression on time (and a constant) rather than the original series.

This practice was challenged by the Box-Jenkins methodology, which became somewhat popular in economics in the seventies, although it originated from engineering. The Box-Jenkins methodology had as one of its major steps the "detrending" of variables by the taking of differences. In the 80ies a major battle between thes two approaches raged, with the difference-detrenders seemingly having the upper hand in the late 80ies, although challenged from many sides—the Bayesians often being the most opinionated (note, that Bayesians mostly do not rely on asymptotic theory and in finite samples there is not discontinuous change in the distribution at the unit root).

During that period the following terminology took hold. The model

$$(DS)\ y_t \ = \ \mu \ + \ \ y_{t-1} \ + \ e_t$$

is called Difference Stationary (DS) since it is stationary after the application of the differencing operation, and the model and

$$(TS)\ y_t \ = \ \mu \ + \ \beta t \ + \ a\ y_{t-1} \ + \ e_t \ ; \ \ a < 1 \ ,$$

is called Trend Stationary (TS) since it is stationary after good old-fashioned detrending by regressing on a time-trend. Most tests for unit roots are formulated as testing TS versus DS.

1.2 Unit Root tests

1.2.1 Dickey-Fuller tests

The most famous of the unit root tests are the ones derived by Dickey and Fuller and described in Fuller (1976).

Dickey and Fuller considered the estimation of the parameter a from the models

$$(1) \quad y_t \;=\; \rho\, y_{t-1} \;+\; e_t \;,$$

$$(2) \quad y_t \;=\; \mu \;+\; \rho\, y_{t-1} \;+\; e_t \;.$$

and

$$(3) \quad y_t \;=\; \mu \;+\; \beta t \;+\; \rho\, y_{t-1} \;+\; e_t \;.$$

The parameter ρ is just the AR-parameter that we have denoted by a so far; but here I use the notation of Fuller (1976). It is assumed that $y_0 = 0$.

The simplest Dickey-Fuller test is simply to estimate (1) by least squares and compare $T(\hat{\rho} - 1)$ to a table of the distribution derived from a Monte Carlo study (or, as shown in Evans and Savin (1981), one can find the characteristic function and invert it). This test is sometimes known as the Dickey-Fuller ρ-test. The critical values for this and the following Dickey-Fuller (DF) tests can be found in Fuller (1976), p. 373. Simplified versions of the tables from Fuller can be found many places, (The Monte Carlo simulations were actually done as part of David Dickey's Ph.D. thesis). In practice the model

(1) is often too simple and one would like to allow for a mean term in the estimation. Dickey and Fuller suggested that one estimates (2) by first calculating the average $\bar{y}$ of the observations $y_2, ..., y_T$ and the average $\bar{y}_0$ of $y_1, ..., y_{T-1}$ and then calculates the least squares estimator $\hat{\rho}_\mu$ by regressing $y_t - \bar{y}$ on $y_{t-1} - \bar{y}_0$. Comparing $T(\hat{\rho}_\mu - 1)$ to the critical values in Fuller is known as the Dickey-Fuller ρ_μ-test. Dickey and Fuller also suggested estimating $\hat{\rho}_\tau$ from model (3) by a standard regression and they tabulated the critical values of $T(\hat{\rho}_\tau - 1)$ under the composite null hypothesis $\rho = 1$ and $\beta = 0$. Note that this last test is a test for the DS model against the TS model, and it is known as Dickey-Fuller ρ_τ-test.

It is often not realistic that a data series should follow as simple a model as (1), (2), or (3) with iid error terms. In most economic data series there will also be substantial short term autocorrelations, so it may be more reasonable to assume the model

$$(4) \quad y_t \;=\; a_1 y_{t-1} \;+ ... +\; a_p y_{t-p} \;+\; e_t \;,$$

where e_t is iid normal, rather than model (1) (and equivalent for models (2) and (3)). An equivalent way of writing model (4) is

$$(4') \quad y_t \;=\; \theta_1 y_{t-1} \;+\; \theta_2 \Delta y_{t-1} + ... +\; \theta_p \Delta y_{t-p+1} \;+\; e_t \;.$$

It is simple to show the equivalence of (4) and (4') and it is left as an exercise. A unit root in (4) will show up as $\theta_1 = 1$ in (4'). So to test for a unit root in (4) you might

want to use $T(\hat{\theta}_1 - 1)$. One would expect the distribution of this statistic to depend on nuisance parameters as e.g. the minimum order p of the autoregression, and it does. One can show (see Fuller (1976) that $cT(\hat{\theta}_1 - 1)$ has the same limiting distribution as $T(\hat{\rho} - 1)$ has, for some constant c, where c is the sum of the terms in the MA representation for e_t. It is somewhat complicated to find c which makes this test less appealing; but it turns out that if you instead of using the distribution of the coefficient you use the distribution of the t-statistic, then the test will *not* depend on the form of the autoregression. The t-statistic has the usual form, that is calculated by any regression package, namely

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

in the simplest model with zero mean and no trend.

For model (1) Dickey and Fuller simulated the critical values under the null of a unit root and the critical values of the test (called the $\hat{\tau}$-test) is tabulated in Fuller (1976) p. 373. Dickey and Fuller also simulated the corresponding t-test for model (2) and (3), which they denoted $\hat{\tau}_\mu$ and $\hat{\tau}_\tau$.

Testing for unit roots using one of the models (1)-(3) is known as Dickey-Fuller tests, whether the $\hat{\rho}$ or the $\hat{\tau}$-tests are used. Estimating a higher order autoregressive process (maybe with non-zero mean and trend) and using the critical $\hat{\tau}$ tables are known as Augmented Dickey-Fuller (ADF) tests. ADF-tests are the most widely applied tests for unit roots. In the influential paper by Engle and Granger (1987) on co-integration, they rec-

ommended ADF tests after examining ADF and several of its competitors. (In the case of co-integration test the critical values are different, see the section on co-integration).

Since the publication of Engle and Granger's article scores of new tests have been developed and we will look at the most popular contenders below. But let us consider how to choose between the different DF tests. Experience from Monte Carlo studies, see for example Dickey, Bell, and Miller (1986), shows clearly that the $\hat{\rho}$-tests have the highest power, as was to be expected. Which $\hat{\rho}$-test should you choose then? That depends, but in any raw series you typically must assume a non-zero mean, which means that you should "at least" use the $\hat{\rho}_{\mu}$- test. If you think that it is likely that there may be a linear trend, then you should use the $\hat{\rho}_{\tau}$-test; but as a general rule this is not "for free". In small samples (as we quite commonlu use in economics) there is a loss in power that increases with the number of parameters that you estimate. On the other hand it you allow for too few parameters, then your model is misspecified and your tests will be wrong. In the present case this means that you will prefer not to include a time-trend if you are reasonably sure that it can safely be left out. Note, however, that the model 2 under the null hypothesis has

$$(2') \quad y_t \;=\; \mu t \;+\; \sum e_t \;,$$

which is *not* invariant to the μ- term after subtracting the mean y_t. In the stable process on the contrary, the de-meaned value of y_t will be independent of the mean term μ (since μ here shows up as a mean term $\mu/[1-\rho]$). Therefore you have to use the $\hat{\tau}_\tau$ test if you think that model (2) might be true with a unit root and a non-zero μ.

What do phrases like "reasonably sure" mean? That is hard to tell - it depends strongly on the number of observations that you have, if you have lots of data: take the most general model. If you have few data points, you will have to rely partly on the knowledge that you have about the particular area of economics, that you are researching. You may even feel that you need higher order deterministic terms than Fuller's tables allow for, in which case you may have to simulate the distribution under the null yourself (although you should consult a time series specialist first to see if they are already available). Be aware that it is not very hard to simulate the null distribution in GAUSS (or most other modern computer languages).

Now then should you use DF or the ADF test? For example for U.S. macro economic time series it is not safe to assume that there is no autocorrelation apart from the potential unit root, which means that you should use the ADF test. How many lags should you include? Again, it depends on the number of data points that you have and on your a priori expectations - the considerations are totally parallel to those about mean and trend terms. What if the process really is an ARIMA(k,1,1)? The MA term corresponds

to an infinite AR process, so you may hope that including enough AR-terms and letting the number of included terms go to infinity with T, then you will obtain a consistent test. Said and Dickey (1984) show that this is exactly the case, given that on chooses $k = k(T) = k_0 T^{1/3}$. This is nice to know; but since k_0 is unknown it is pretty much useless as a practical guideline. You may sometimes see the ADF test referred to as the Said-Dickey test - that is a way of indicating that the author does not consider the order of the AR model as more than an approximation.

1.2.2 Phillips-Perron tests

The tests that has been most popular next to the DF-tests are the Phillips-Perron (PP) tests suggested in Phillips and Perron (1988).

In the case where you are convinced that there is no autocorrelation apart from the unit root, the PP test is identical to the DF test. The difference between the PP and the DF tests lies in the treatment of the autocorrelation. If there is autocorrelation that is not accounted for you will get bias. The hard part is to rid of that bias. The ADF does that by choosing an AR(p) process, but the weak point in that procedure is exactly the choice of p. Not only does a researcher usually only have vague ideas about the appropriate chice of p; but for many applied researches it is just too tempting to try

out a bunch of different values for p until the results corresponds to what the researcher wanted a priori. If results are derived by data-mining of that sort the asymptotic test statistics reported will be meaningless.

The idea of the Phillips-Perron test is to run a non-augmented Dickey Fuller regression, and then to adjust for the bias that might occur due to correlation in the innovation term so that the Dickey-Fuller tables can be used anyway.

Phillips-Perron suggest estimating $\hat{\rho}$ from model (2), then estimate the innovation error variance $s^2 = \sum_{t=2}^{T} \hat{e}_t^2$ and 2π times the spectral density at frequency zero by, $2\pi \hat{f}_e(0)$, (we will cover techniques for doing this later on). The Phillips-Perron $Z(\hat{\alpha})$ test is then

$$Z(\hat{\alpha}) \;=\; T(\hat{\rho}_\mu - 1) \;-\; \frac{T^2(2\pi \hat{f}(0) - s^2)}{2\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{-1})^2} \;,$$

where $\bar{y}_{-1}$ is the mean of y_{t-1}. (It has its name because Phillips-Perron uses α rather that ρ for the AR-parameter; but here we will not change notation again). The corresponding t-statistic is

$$Z(t_{\hat{\alpha}}) \;=\; \tau_\mu \sqrt{\frac{s^2}{2\pi \hat{f}(0)}} \;-\; \frac{T(2\pi \hat{f}(0) - s^2)}{2\sqrt{2\pi \hat{f}(0)\;\sum_{t=2}^{T}(y_{t-1} - \bar{y}_{-1})^2}} \;.$$

Correspondingly for the model that allows for a time trend under the alternative Phillips and Perron suggest the adjusted DF tests

$$Z(\tilde{\alpha}) \;=\; T(\hat{\rho}_\tau - 1) \;-\; \frac{T^2(2\pi \hat{f}(0) - s^2)}{2\sum_{t=2}^{T}(y_{t-1} - \hat{y})^2} \;,$$

and

$$Z(t_{\tilde{\alpha}}) \;=\; \tau_\tau \sqrt{\frac{s^2}{2\pi \hat{f}(0)}} \;-\; \frac{T(2\pi \hat{f}(0) - s^2)}{2\sqrt{2\pi \hat{f}(0) \sum_{t=2}^{T} (y_{t-1} - \hat{y})^2}} \;,$$

where $\hat{y}$ is the projection of y_{t-1} on $1, t$ (i.e. the fitted value from a regression of lagged y on a constant and a time trend) and s^2 is defined as before but now from the residuals from the regression of y_t on $y_{t-1}, 1, t$. The PP-tests now consist of comparing those adjusted DF-values with the corresponding tables used for the basic Dickey-Fuller test. You will hear those tests referred to as the Phillips- Perron tests or as the $Z\alpha$ and Zt tests.

An interesting (and extensive) Monte Carlo study is Schwert (1987). Schwert examines what happens to the size of the above mentioned tests if the true process is an ARIMA(1,1,1) and the Phillips-Perron or the ADF tests are used. A very clear conclusion from Schwert's study is that the ADF test, used with "many" lags, have the best size-properties in the sense that the size of the small sample distribution is close to the true 5% size if the Dickey-Fuller tables are used. The other tests can be *very* far of in terms of size, especially if the ARIMA-process has a large negative MA-coefficient, in which case the PP tests (and also the ADF test when the MA-coefficient is very close to -1), almost always reject the unit root model (even if it is true). Note that the happens

because the model has the form

$$(1 - L)y_t \;=\; (1 + bL)e_t \;.$$

One can see that the two lag polynomials "nearly cancel" when b is close to -1. Some authors call processes like that "nearly white noise processes".

Does this mean that one should always use the ADF-test with a high number of AR-terms? Not necessarily, since there is a cost in terms of power. It seems that this cost is not nearly as high as the cost incurred by falsely allowing for a trend-term (see Dickey, Bell and Miller (1986)), but there will presumably still be some loss in power. In an interesting Monte Carlo study, Campbell and Perron (1991) show that both the Said-Dickey test (with 6 lags) and the Phillips-Perron test falsely reject the unit root almost every time in the case of nearly white noise processes. They also show, however, that for simple short term forecasting, this will not necessarily result in a decline in the ability to make simple forecasts, since the trend stationary model forecasts just as well in the cases where the unit root model tends to be falsely rejected.

1.2.3 Approximate POI-tests

In a recent interesting paper Elliott, Rothenberg and Stock (1992) (ERS) examine "point optimal invariant tests" (POI) for unit roots. An invariant test is a test – like the augmented Dickey-Fuller test – that is invariant to nuisance parameters. In the unit root case that will be tests that are invariant to the parameters that capture the stationary movements around the unit roots (i.e., the parameters to $\Delta X_{t-1}, \Delta X_{t-2}, \cdots, \Delta X_{t-k}$ in the ADF-case).

The most commonly used test are the "τ-tests" that allows for a deterministic time trend (in the case where you do not allow for a time trend most tests are pretty well behaved), so I will only consider this type of test in this subsection.

When

$$y_t = \mu + \sigma t + u_t \ ; \quad u_t = \rho u_{t-1} + \epsilon_t \ ,$$

ERS show that the POI test for a unit root against $\rho = \overline{\rho}$ has the form

$$M_T = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \ ,$$

where

$$\tilde{\sigma}^2 \ = \ T^{-1} \sum_{t=1}^{T} \tilde{e}_t^2 \ ; \qquad \hat{\sigma}^2 \ = \ T^{-1} \sum_{t=1}^{T} \hat{e}_t^2 \ ,$$

and $\hat{e}_t$ and $\tilde{e}_b$ are the residuals from the GLS estimation of y_t on $z_t = (1, t)$ under

$\rho = 1$ and $\rho = \overline{\rho}$ respectively:

$$\hat{e} \;=\; \hat{y}_t - \hat{\beta}'\hat{z}_t \;; \qquad (\hat{\beta} \;=\; (\sum \hat{z}_t\hat{z}_t')^{-1}(\sum \hat{z}_t y_t) \;,$$

where

$$(\hat{y}_1, \cdots, \hat{y}_T) \;=\; (y_1, \Delta y_2, \cdots, \Delta y_T) \;,$$

and

$$(\hat{z}_1, \cdots, \hat{z}_T) \;=\; (z_1, \Delta z_2, \cdots, \Delta z_T) \;,$$

and correspondingly for $\tilde{e}$ except

$$(\tilde{y}_1, \cdots, \tilde{y}_T) \;=\; (y_1, (1-\overline{\rho}L)y_2, \cdots, (1-\overline{\rho}L)y_T) \;,$$

and

$$(\tilde{z}_1, \cdots, \tilde{z}_T) \;=\; (z_1, (1-\overline{\rho}L)z_2, \cdots, (1-\overline{\rho}L)z_T) \;.$$

This last transformation is the GLS-detrending (if you are not quite sure what I mean by that, check your first year econometrics text (look under AR(1) disturbances) - most often the first observation will also be rescaled, but asymptotically that does not make a difference).

The critical value for the test will depend on $\overline{c}$ where

$$\overline{\rho} \;=\; 1 - \frac{\overline{c}}{T} \;.$$

When

$$u_t = \rho u_{t-1} + \beta_2 \Delta u_{t-2} + \cdots + \beta_k \Delta u_{t-k} + \epsilon_t \;,$$

the critical value of the test statistic has to be adjusted in the same way as for the Phillips-Perron (PP) test (in practice you adjust the test-statistic rather than the critical value, (and you don't have to assume a finite AR-representation), exactly as in the PP case).

In practice, $\overline{\rho}$ is not known, so ERS suggest choosing an approximate value of $\overline{c}$. They suggest using the value of $\overline{c}$ that asymptotically gives a power of 50%. These values are $\overline{c} = -7$ in case of a mean and $\overline{c} = -13.5$ in the case of a trend). ERS show that this way we get approximately the POI test asymptotically.

Also, it turns out that if we instead do the GLS-adjustment and then perform the ADF-test (now without allowing for a mean or trend) we get approximately the POI-test. Elliott et al. call this test the DF-GLS$^{\tau}$ test. They argue that the theoretical examinations above point to the GLS-adjustment as the critical factor, and they then show by Monte Carlo simulation that the DF-GLS$^{\tau}$ seems to behave as well as the approximate POI test (using the values for $\overline{c}$ that I listed above).

The critical values depend on T. The critical values are

T	1%	5%
50	-3.77	-3.19
100	-3.58	-3.03
200	-3.46	-2.93
500	-3.47	-2.89
∞	-3.48	-2.89

If you need critical values for the variant of the test with mean but not trend, see the article by ERS. (They call that test the DF-GLS$^{\mu}$ test.)

1.3 The importance of unit roots

The discussion above indicates that there is not necessarily a sharp distinction between unit root processes and stable processes, which has led some researchers (Cochrane (1991), Christiano and Eichenbaum (1989) - the later paper titled "Unit roots in real GNP: Do we know and do we care?") to question the importance of unit root tests.

I do not think that arguments that stresses that some stable processes look a lot like unit root processes in finite time, makes it less important to test for unit roots. But

it is true that there are models, like the ARIMA models where the MA-process nearly cancels with the unit root, in which it really isn't interesting whether the model "truly" is a unit root model. In my opinion no time series model is true in econometrics anyway. If one accepts that these models are just approximations then there isn't really a problem - if the stable process does a good enough job at modeling the data, then let it!

It turns out that one can always decompose a unit root process into the sum of a random walk and a stable process. This is known as the Beveridge-Nelson (1981) composition. The Beveridge-Nelson (BN) decomposition states that any I(1) time series can be decomposed into the sum of a random walk an I(0) time series. In the vector case we can state the results as

Theorem: Beveridge-Nelson decomposition

Any I(1) process y_t can be written as the sum of a random walk s_t and a stable process c_t (s_t and c_t will *not* be independently distributed):

$$y_t \;=\; s_t \;+\; c_t$$

Proof: Since $(1-L)y_t$ is a stable process it has a Wold-decomposition

$$(1-L)y_t \;=\; \psi(L)e_t \;,$$

now write

$$(1-L)y_t \;=\; \psi(L)e_t \;=\; \psi(1)e_t \;+\; (\psi(L)-\psi(1))e_t \;,$$

which is equivalent to

$$(1-L)y_t \;=\; \psi(L)e_t \;=\; \psi(1)e_t \;+\; \psi^{**}(L)e_t \;,$$

where $\psi^{**}(1)=0$. This implies that

$$y_t \;=\; \psi(1)(1-L)^{-1}e_t \;+\; (1-L)^{-1}\psi^{**}(L)e_t \;,$$

which has the form

$$y_t \;=\; s_t \;+\; \psi^{*}(L)e_t \;,$$

where $\psi^*(L)e_t$ is a stable process and s_t is the random walk $\psi(1)(1-L)^{-1}e_t = \psi(1)\sum_s^t e_s$. QED.

Note that $\psi(1)e_t$ is the long run effect of e_t. For example if $y_0 = 0$ then $y_T \rightarrow \psi(1)e_1 + h_T(e_2, e_3, ..)$ for some function h. Also note that the derivation of the Campbell and Mankiw (1987) suggested the use of $\psi(1)$ as a measure of the persistence in y_t; whereas Cochrane (1988) suggests the use of

$$\psi(1)^2\sigma_e^2/\sigma_{\Delta y}^2 \;,$$

where $\sigma_{\Delta y}^2$ and σ_e^2 is the variance of Δy_t and e_t, respectively. (Note the importance of the spectral density for Δy at frequency zero). Cochrane refers to his measure as the

size of the unit root. This may be a slightly unfortunate term, and some researchers prefer not to use it; but you will meet it in the literature. Cochrane's measure is best if you want to measure the importance of the random walk component of a series relative to the stationary part of the series (if Cochrane's measure is very small an ADF test would most likely conclude falsely that the process was stable) whereas Campbell and Mankiw's measure is better if you are a macroeconomist who wants to know the impact in the infinite future of a shock happening today.