Introduction

Cointegration is a widely used tool in statistical based economic analysis. Cointegration between two non-stationary time series implies there exists some definite and constant relation between them which is stationary. An example will be explored here between the consumer price index (CPI) of Italy and the US, both of which, along with the exchange rate between the two countries, form non-stationary series. Yet the notion of purchase power parity would say that there exists a definite relation between these two price indices which operates so as to make them more or less equivalent to one another.

The above example is just one which has been coded for in the github repo. All examples follow those presented in chapters 17-19 of James D. Hamilton's *Time Series Analysis*[1]. In the code I make numerous references to specific equations and page numbers of Hamilton's text so the interested reader can deduce what is happening and why by referring to Hamilton's text.

Cointegration analysis is a non-trivial subject. Here I largely focus on all of the requisite details *not* covered chapters 17-19; all the fundamental things which build up to cointegration analysis, without an understanding of which one is ill prepared to reliably execute cointegration tests. Topics covered are mainly from chapters 1-3, 8, and 15, 17, and 19.

1 Stationary

A time series y_t is said to be stationary or covariance stationry if the expected value and covariance do not change over time (respectively). So we have,

$$\begin{aligned} E(y_t) &= \mu && for\ all\ t \\ E[(y_t - \mu)(y_{t-j} - \mu)] &= \gamma_j && for\ all\ t \end{aligned}$$

where μ is some constant. For covariance stationary processes the autocovariance function has the property that $\gamma_j = \gamma_{-j}$ for all integers j.

Now let ε_i be an i.i.d. sequence of random numbers with zero mean and variance σ^2. A *trend stationary* process has the general form

$$y_t = \alpha + \delta t + \psi(L)\varepsilon_t \tag{1}$$

where α and δ are constants, L is the backshift operator which has the property $Ly_t = y_{t-l}$, and $\psi(L) = 1 + \psi_1 L + \psi_2 L^2 ...$ where ψ_j are constant coefficients which obey the restriction $\sum_{j=0}^{\infty} |\psi_j| < \infty$, the meaning of which will become apparent in section 3. The terms $\psi(L)\varepsilon_t = \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2 L\varepsilon_{t-2} ...$ can be seen as unanticipated 'shocks' to the system which occurred at previous times. But these shocks are presumed to have zero mean, ergo, were we to subtract the deterministic terms $\alpha + \delta t$ ('drift' plus time trend), or even were we to just subtract the time trend, we'd be left with a stationary value on the right side. If the time trend is not subtracted then a model with a time trend is clearly not stationary as $E[y_t] \propto \delta t$ which changes in value with the time variable.

A *unit-root* process has the general form,

$$(1 - L)y_t = \Delta y_t = \delta + \psi(L)\varepsilon_t \tag{2}$$

Noting that $\psi(L)$ forms a univariate polynomial which by the fundamental theorem of algebra can be factored into monic terms, i.e. given $\psi(L) = 1 + \sum_{i=1}^{n} \psi_i L^i$ we have,

$$(1 - L)y_t = \delta + (L - \xi_1)(L - \xi_2)...(L - \xi_n)\varepsilon_t \tag{3}$$

where ξ_j are roots of the polynomial[1]. Suppose some $\xi_j = 1$, i.e. a unit root exists. Substituting $L = 1$ would cause both the left side and the polynomial on the right to go to zero and we'd get $\delta = 0$. This in turn implies y_t is stationary as can be seen by the following iterated relation,

$$\begin{aligned} y_t &= y_{t-1} + \psi(L)\varepsilon_t \\ &= y_{t-2} + \psi(L)(\varepsilon_t + \varepsilon_{t-1}) \\ &\vdots \\ &= y_0 + \psi(L)(\varepsilon_t + \varepsilon_{t-1} + \ldots) \end{aligned}$$

Thus, the stipulation of the unit root process that $\psi(1) \neq 0$ is to ensure that y_t is not stationary, which is to say we are instead interested in process for which y_t is non-stationary but $\Delta y_t = y_t - y_{t-1}$ may be stationary. If it is, such a process is referred to as integrated of order 1, or $I(1)$. Notice that equation 1 is also $I(1)$ because $(1 - L)\delta t = \delta$. The stipulation that $\psi(1) \neq 0$ also ensures invertibility of the $\psi(L)$ operator as we will soon see.

For infinite times the difference between trend stationary and unit root processes are negligible. For finite times, one of the most significant differences is in how a given shock ε_t changes the forecast y_{t+s}.

For a trend stationary process it is easy to show that,

$$\hat{y}_{t+s|t} = \alpha + \delta(t + s) + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \ldots \tag{4}$$

where the bar symbol "|" in the subscript of y indicates a conditional statement; given the values of $\hat{y}_t, \hat{y}_{t-1}, \hat{y}_{t-2}, ...$ then the value of $\hat{y}_{t+s}$ is the expression on the right. One should still wonder what happened to the variables $\varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + ... + \psi_{s-1}\varepsilon_{t+1}$? This is explained by the hat symbol on $\hat{y}_{t+s}$ which represents an estimate of y_{t+s} using the projection theorem which is discussed in section[2] 4.

For a unit root process it can be shown that[3],

$$\hat{y}_{t+s|t} = s\delta + (\psi_1 + \psi_2 + ... + \psi_s)\varepsilon_t + (\psi_2 + \psi_3 + ... + \psi_{s+1})\varepsilon_{t-1} + ... \tag{5}$$

Taking derivatives with respect to ε_t in equations 4 and 5 gives the effect which a change in ε_t has on y_{t+s}

$$\begin{aligned} \frac{\partial \hat{y}_{t+s}}{\partial \varepsilon_t} &= \psi_s && \longrightarrow && \textit{trend stationary} \\ \frac{\partial \hat{y}_{t+s}}{\partial \varepsilon_t} &= 1 + \psi_1 + \psi_2 + ... + \psi_s && \longrightarrow && \textit{unit root} \end{aligned} \tag{6}$$

Taking the limit $s \to \infty$ we get that for the trend stationary process $\psi_s \to 0$ due to the restriction $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Thus, the effects of a given shock eventually wear off for a trend stationary process, but not for a unit root process. Again, the difference between trend stationary and unit root process are negligible for infinite times, so we are making the presumption that samples are finite yet large enough for applications of the central limit theorem (typically 100-500 time samples).

2 Generalized Method for Impulse Response Coefficient

We now discuss a more general method of calculating the term $\frac{\partial y_{t+s}}{\partial \varepsilon_t}$. For the generalized $AR[p]$ process,

[1] we find the roots by replacing the operator L with a dummy variable z.
[2] also see $ACF\_Innovations\_READ\_ME$ file contained in the $Time - Series - Analysis$ folder of the github repo.
[3] see pg 439 of Hamilton.

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t \tag{7}$$

Iterating the above relation for times $t = 0, 1, 2, ..., t-1, t$ produces the matrix equation,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p+2} \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_{p-1} & \phi_p \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \tag{8}$$

Where the ones just below the diagonal entry are merely a clever way of constructing the matrix equations such that equation 7 holds at time t but for times less than t, e.g. $t-1$ we'd get $y_{t-1} = y_{t-1}$. Notice however that this in no way precludes equation 7 from holding at time $t-1$; it is simply a different way of making a true statement. The above matrix equation is abbreviated as,

$$\mathbf{Y}_t = \mathbf{F}\ \mathbf{Y}_{t-1} + \mathbf{e}_t$$

Iterating the above relation we get,

$$\mathbf{Y}_{t+j} = \mathbf{F}^{j+1}\mathbf{Y}_{t-1} + \mathbf{F}^j\mathbf{e}_t + \mathbf{F}^{j-1}\mathbf{e}_{t+1} + \mathbf{F}^{j-2}\mathbf{e}_{t+2} + ... + \mathbf{F}\mathbf{e}_{t+j-1} + \mathbf{e}_{t+j} \tag{9}$$

differentiating with respect to ε_t gives,

$$\frac{\partial y_{t+s}}{\partial e_t} = f_{11}^j \tag{10}$$

where f_{11}^j represents the $[1,1]$ entry of the matrix $\mathbf{F}^j$. The fact that only $[1,1]$ element is needed is explained by the fact that A) $\mathbf{e}_t$ picks the first column of matrix $\mathbf{F}^j$ and B) we are concerned with the derivative taken on y_{t+j} which is the first [top] element in the column vector $\mathbf{Y}_{t+j}$.

For large j values taking matrix powers can lead to computational problems. Thankfully, Jordan decomposition allows the matrix $\mathbf{F}$ to be written as,

$$\mathbf{F} = \mathbf{M}^{-1}\mathbf{J}\mathbf{M}$$

where $\mathbf{M}$ is the matrix which trasforms $\mathbf{F}$ into its Jordan form $\mathbf{J}$. The property we wish to exploit is, because the matrix $\mathbf{M}^{-1}\mathbf{M} = \mathbf{I}$, we can write the j^{th} power of $\mathbf{F}$ as $\mathbf{F}^j = \mathbf{M}^{-1}\mathbf{J}^j\mathbf{M}$. The Jordan matrix is guaranteed to have only two diagonal rows (diagonal and supradiagonal) which makes computing higher orders of the matrix more stable. This is especially true when all of the eigen-values are less than unity and are distinct in which case $\mathbf{J}$ is a diagonal matrix and f_{11}^j will tend to zero as $j \to \infty$.

For the more general case of repeated eigenvalues the matrix $\mathbf{J}^j$ takes the form,

$$\mathbf{J}^j = \begin{bmatrix} \mathbf{J}_1^j & & 0 & 0 & \dots & 0 \\ 0 & \mathbf{J}_2^j & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \mathbf{J}_s^j \end{bmatrix} \tag{11}$$

where the $n_i \times n_i$ Jordan block matrices $\mathbf{J}_i^j$ have the form[4],

$$\mathbf{J}_i^j = \begin{bmatrix} \lambda_i^j & \binom{j}{1}\lambda_i^{j-1} & \binom{j}{2}\lambda_i^{j-2} & \dots & \binom{j}{n_i-1}\lambda_i^{j-n_i+1} \\ 0 & \lambda_i^j & \binom{j}{1}\lambda_1^{j-1} & \dots & \binom{j}{n_i-2}\lambda_i^{j-n_i+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \lambda_i^j \end{bmatrix} \tag{12}$$

where

$$\binom{j}{n} = \begin{cases} \frac{j(j-1)\dots(j-n+1))}{n!}, & \text{if } j \geq n \\ 0, & \text{otherwise} \end{cases}$$

For the above case of repeated eigenvalues, and assuming all eigenvalues are less than 1, to my knowledge there is no assurance that f_{11}^j converges so that the effects of a shock ε_t eventually wear off or at least converge to a finite values as $j \to \infty$. Attempting convergence test for $\binom{j}{n}\lambda_i^{j-n}$ are unfruitful. Though it can be shown[5] that for all eigenvalues less than 1 we have $\sum_{j=0}^{\infty} \frac{\partial y_{t+j}}{\partial \varepsilon_t} = 1/(1-\phi_1-\phi_2-\dots-\phi_p)$ the triangle inequality would point in the wrong direction for us to use this as a means of proving that $\frac{\partial y_{t+j}}{\partial \varepsilon_t}$ converges to a finite number or to zero. This makes it difficult to establish a relation between equations 6 and 10. But this is not a cause for alarm as we've yet to attempt to put equation 7 on which the above analysis was based in the form of a unit root or trend stationary process.

For the moment we can calculate the impulse response coefficients and deduce from the graph whether y_t is stable in response to an impulse. By 'stable' it is here meant that the effects of ε_t on y_{t+j} are finite as $j \to \infty$.

Example:

On pages 583-586 Hamilton[1] tests for a cointegrating relation between the consumer price index of Italy vs. the U.S. between 1973:1 (january, 1973) and 1989:10 (October, 1989). Under purchasing power parity it is hypothesized that goods would sell for the same effective cost in both countries. If this hypothesis holds then we'd have $P_t = s_t P_t^*$ where P_t is the U.S. consumer price index, P_t^* is the Italian consumer price index, and S_t is the exchange rate between the two countries. Taking 100 times the natural log of this relation then subtracting the initial values gives $p_t = s_t + p_t^*$ where $p_t = 100(ln(p_t) - ln(p_0))$ with similar expressions for s_t and p_t^*. We subtracted the initial values only to make the relative trends between the three quantities more apparent as is seen in figure 1.a. So too is it apparent that all three processes p_t, s_t and p_t^* are non-stationary. However, if purchase power parity holds then we could construct a variable z_t,

$$z_t = p_t - s_t - p_t^*$$

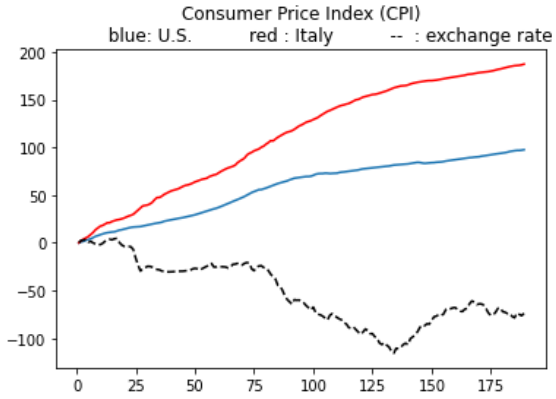
and we would expect z_t to only fluctuate randomly about a value of zero, i.e. z_t would be stationary. Note that p_t, s_t and p_t^* are presumed column vectors. We could then stack them into an array $y_t = [p_t, s_t, p_t^*]$ which would be an $I(1)$ process, i.e. z_t is stationary without differencing. We would then have $z_t = ay_t'$ where $a = [1, -1, -1]$.

definition: *a vector a which acts on an $I[1]$ process to produce an $I[0]$ process is called a* cointegrating vector.

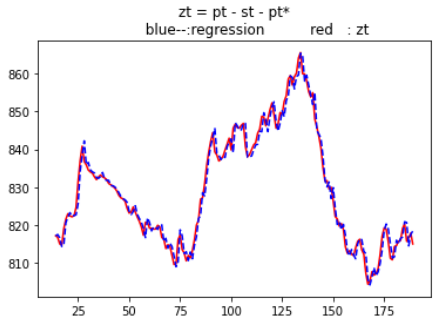
Specific tests for cointegration are executed in the code and are discussed at length in chapters 17-19 of Hamilton. Here we only wish to examine the impulse response coefficient, and in sections to come we will

[4] see pg. 444 of Chiang, Chin Long. 1980. *An Introduction to Stochastic Processes and Their Applications.* Huntington, N.Y. Krieger

[5] see page 20 of Hamilton



(a) figure 1.a: Consumer price index of US vs. Italy



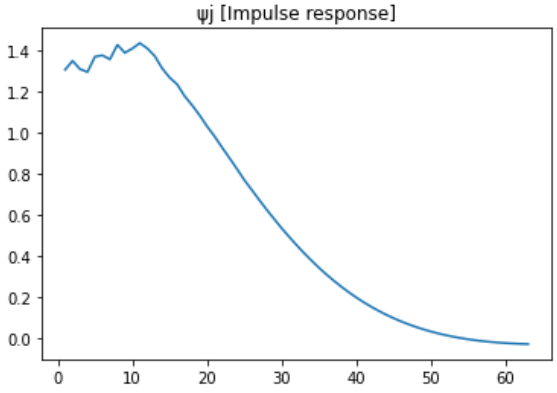
(b) figure 1.b: $z_t = p_t - s_t - p_t^*$

examine the fundamentals used in the aforementioned chapters to test for cointegration. If one understands the underlying fundamental principles then the tools presented in chapters 17-19 (also 20) are merely clever extensions of these principles.

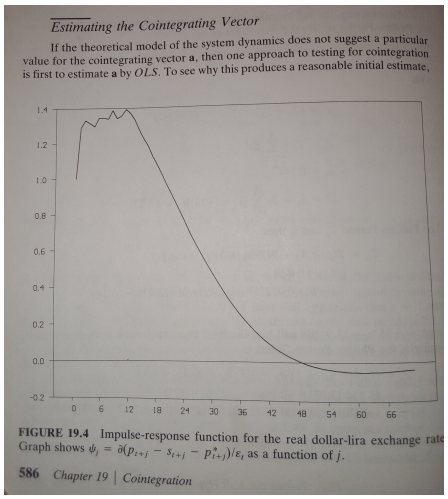
The file *Impulse_Response_CPI_US_vs_Italy.py* performs a regression of the form of equation 7, calculates the value of the impulse response coefficient(s),

$$\frac{\partial z_{t+j}}{\partial \varepsilon_t} = f_{11}^j$$

then plots the values of f_{11}^j for $j = 0, 1, 2, ..., 70$. The resulting graph in figure 2.a is reassuringly in line with the results given on page 586 by Hamilton (figure 1.b).



(c) figure 2.a: Calculated Impulse response coefficients (see github repo)


Estimating the Cointegrating Vector

If the theoretical model of the system dynamics does not suggest a particular value for the cointegrating vector **a**, then one approach to testing for cointegration is first to estimate **a** by *OLS*. To see why this produces a reasonable initial estimate,

FIGURE 19.4 Impulse-response function for the real dollar-lira exchange rate. Graph shows $\psi_j = \partial(p_{t+j} - s_{t+j} - p^*_{t+j})/\varepsilon_t$ as a function of j.

(d) figure 2.b: Impulse response coefficients given by Hamilton, pg. 586

In the calculated regression a value of $p = 13$ lags were taken, so $\mathbf{F}$ is a 13×13 matrix which in this case is diagonal because the eigenvalues are unique. Though the calculated regression coefficients are not exactly the same as those given by Hamilton, this might be due to a slightly different source of data[6] But the similarity

[6]The data utilized in the code pulls from fred.stlouisfed.org. There are a number of variations on the consumer price index. The data pulled for the U.S. was entitled *Consumer Price Index for All Urban Consumers: All Items in U.S. City Average.*

of the impulse response graphs is reassuring. Being as **F** is as a fairly large matrix these results are also a reassuring test of the Jordan decompositon algorithm contained in the *Jordan_Decompose.py* file within the $funcs\_$ folder (it would however, be an even more reassuring test of the Jordan decomposition algorithm were the eigenvalues not distinct which would imply the Jordan matrix is not diagonal).

Notice that the effects of an impulse wear off gradually. At first glance we might say this is indicative of a time trend. However, noting that the data is taken in monthly intervals, as far as three years into the future the effects of a unit increase would affect our prediction of z_t by about %25. This is not a negligible increase, and three years, while not a true measure of 'permanent' is not a negligible time-span! Ergo, we might suspect z_t to be more appropriately modeled by a unit-root process. This is affirmed by a plot of z_t in figure 1.b which shows no apparent deterministic trend in time of z_t. Figure 1.a plots $p_t - p_0$, $s_t - s_0$, and $p_t^* - p_0^*$.

3 Transforming the General Regression into Unit Root processes

Adding a constant α to equation 7 then moving all $\phi_j y_{t-j} = \phi_j L^j y_t$ terms to the right side gives,

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t = \alpha + \varepsilon_t \tag{13}$$

calling the polynomial on the left $(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) = 1 - \phi(L)$ then dividing by $1 - \phi(L)$ gives

$$y_t = \delta_\phi + u_t \tag{14}$$

where $u_t = \varepsilon_t/(1-\phi(L))$ and $\delta_\phi = \alpha/(1-\phi(1))$. Why the lag operator is taken to be $L = 1$ in the denominator of δ_ϕ can be understodd in two ways. The more intuitive way is to multiply the above equation by $1 - \phi(L)$ and note that when this operator acts on the constant α there is nothing for the lag operator to act on, so we get

$$\frac{(1 - \phi(L))\alpha}{1 - \phi(1)} = \frac{\alpha(1 - \phi(1))}{1 - \phi(1)} = \alpha$$

We'd then be left with our original equation (equation 13). A second and more common way to see this is by looking at the roots of the polynomial $(1 - \phi(L)) = 1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p$. Dividing the polynomial by L^p then calling $\lambda = 1/L$ this polynomial can be factored into monic terms as,

$$\begin{aligned}(1 - \phi(L)) &= \lambda^p - \phi_1 \lambda^{p-1} - \ldots - \phi_p \\ &= (x - \lambda_1)(x - \lambda_2) \ldots (x - \lambda_p) = 0\end{aligned}$$

where λ_j are the roots [eigenvalues] of the polynomial. In terms of the lag operator this is written as,

$$(1 - \phi(L)) = (1 - \lambda_1 L)(1 - \lambda_2 L) \ldots (1 - \lambda_p L) = 0$$

In terms of the invertibility of $(1 - \phi(L))$, if all of the roots λ_i are less than one then each of the terms $(1 - \lambda_1 L)$ form a bounded sequence when acting on y_t. Hamilton gives a detailed analysis of this on page 28 where it is shown that for $|\lambda_i| < 1$ the term $(1 - \lambda_1 L)^{-1}$ converges to an infinite series in a fashion which is reminiscent of a geometric series expansion. A collection of multiplied $(1 - \lambda_j L)^{-1}$ terms then can all be expanded and the coefficients on powers of L can be grouped together to form one single infinite series expansion for the polynomial. Explicitly we have,

$$\begin{aligned}(1-\phi(L))^{-1} &= [(1-\lambda_1 L)(1-\lambda_2 L)\dots(1-\lambda_p L)] \\ &= \psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \dots \\ &= \psi(L)\end{aligned}$$

recalling that the substitution that $\lambda = 1/L$ was used to find the roots of the polynomial $\lambda^p - \phi_1\lambda^{p-1} - \dots - \phi_p$, the requirement that λ_j values be less than one now becomes that the roots of the polynomial $1 - \phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$ be *greater* than one, or 'outside the unit circle'.

The relation $(1-\phi(L))^{-1} = \psi(L)$ would of course require that,

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)(\psi_0 + \psi_1 L + \psi_2 L^2 + \dots) = 1$$

The above relation would require that $\psi_0 = 1$ and all other coefficients of powers of L vanish, and this can be used to ascertain a definite relation between the ϕ_j and ψ_j coefficients. In fact equation 9 can be used as an alternative means to find such a relationship. It turns out these coefficients are related by the relation[7],

$$\psi_j = \phi_1\psi_{j-1} + \phi_2\psi_{j-2} + \dots + \phi_p\psi_{j-p}$$

The equivalence $(1-\phi(L))^{-1} = (1 + \psi_1 L + \psi_2 L^2 + \dots)$ makes clear the meaning behind the requirement mentioned in section 1 that $\sum_{j=0}^{\infty} |\psi_j| < \infty$ where the modulus accounts for the possibility of imaginary terms.

We can now write equation 14 in the form,

$$y_t = \delta_\phi + \psi(L)\varepsilon_t \tag{15}$$

which is not *quite* in the form suitable for comparison to a unit root process (c.f. equation 2). Lets go back to equation 13,

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p)y_t = \alpha + \varepsilon_t$$

If one of the eigenvalues has value 1 (for convenience lets say $\lambda_1 = 1$) and all others are outside the unit circle then we can leave the unit root polynomial term on the left and divide by the remaining terms. Let the superscript ^{1} on the operators derived thus far denote their equivalent form when just one term (the unit root) is ommitted from their calculation. So $\delta^1_\phi = \alpha/(1-\lambda_2 L)\dots(1-\lambda_p L)$, $u^1_t = \varepsilon_t/(1-\lambda_2 L)\dots(1-\lambda_p L) = \psi^1(L)\varepsilon_t$. We have,

$$(1-L)y_t = \delta^1_\phi + \psi^1(L)\varepsilon_t \tag{16}$$

which is the general regression (equation 13) put into the form of a unit root process. Writing $(1-L)y_t = y_t - y_{t-1}$ instead gives,

$$y_t = y_{t-1} + \delta^1_\phi + \psi^1(L)\varepsilon_t$$

iterating the equation by repeatedly making substitutions for y_{t-1} gives,

$$y_t = y_0 + t\delta^1_\phi + \psi^1(L)(\varepsilon_t + \varepsilon_{t-1} + \dots \varepsilon_1) \tag{17}$$

[7] c.f. exercise 3.3 of Hamilton

which is not stationary in expected value because $E[y_t] \propto \delta_\phi^1 t$ nor is it stationary in covariance as the expression $\varepsilon_t + \varepsilon_{t-1} + \ldots \varepsilon_1$ will obviously either grow or shrink with time.

Transformations of the general form of equation 16 are used extensively in the more advanced analysis of chapters 17-20 on cointegration.

4 Hypotheis Testing

Given n observations for column vectors y and $x_1, x_2, ...x_p$ (i.e. these are vectors of length n), an ordinary least squares (OLS) approach to estimating the relation between y and $x_1, x_2, ...x_p$ is to first hypothesize a model of the form $y = \mathbf{X}\beta + u$ where $\mathbf{X} = [x_1, x_2, ...x_p]$, $\beta = [\beta_1, \beta_2, ..., \beta_p]'$, and u is a gaussian random variable with zero mean and variance σ^2. We then find an estimate $\hat{\beta}$ by minimizing the sum of the squares of u values; $\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (y_i - x_1\beta_1 - x_2\beta_2 - \ldots - x_p\beta_p)^2$. Minimization is performed with respect to β_j coefficients. The result is that,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \tag{18}$$

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \\ E[\hat{\beta}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[y] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{X}\beta + u] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E[\beta] \\ &= E[\beta] \end{aligned}$$

Calling $\hat{u} = y - \mathbf{X}\hat{\beta}$ the residual of our estimate we then have fitted the model $y = \mathbf{X}\hat{\beta} + \hat{u}$ where all quantities in this equation are known values. And from the above relation we have $E[\hat{u}] = 0$.

In time-series analysis it is typical we take $y = y_t$ and $x_j = y_{t-j}$ or one particular x_j may also be a column of ones corresponding to constant term included in the regression model. Were we to include a time trend we'd also include a column vector of time values. For the model,

$$y_t = \alpha + \delta t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \varepsilon_t$$

In matrix form we'd represent this as,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_p \\ y_{p-1} \\ \vdots \\ y_2 \\ y_1 \\ y_0 \end{bmatrix} = \begin{bmatrix} 1 & t & y_{t-1} & y_{t-2} & \cdots & y_{t-p} \\ 1 & t-1 & y_{t-2} & y_{t-3} & \cdots & y_{t-p-1} \\ 1 & t-2 & y_{t-3} & y_{t-4} & \cdots & y_{t-p-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & p & y_{p-1} & y_{p-2} & \cdots & y_0 \\ 1 & p-1 & y_{p-2} & y_{p-3} & \cdots & y_{-1} \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & y_1 & y_0 & \cdots & y_{-p+3} \\ 1 & 1 & y_0 & y_{-1} & \cdots & y_{-p+1} \\ 1 & 0 & y_{-1} & y_{-2} & \cdots & y_{-p} \end{bmatrix} \begin{bmatrix} \alpha \\ \delta \\ \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} \hat{u}_t \\ \hat{u}_{t-1} \\ \hat{u}_{t-2} \\ \vdots \\ \hat{u}_p \\ \hat{u}_{p-1} \\ \vdots \\ \hat{u}_2 \\ \hat{u}_1 \\ \hat{u}_0 \end{bmatrix} \tag{19}$$

Where the dashed line(s) indicate ommited values in the matrix. The reason for this is because, for example, the lag column vector y_{t-p} takes on negative time values for $t < p$. Being as we'd often prefer to use as many available and relevant data points as possible we'd call our first observed value y_0, so negative time values of y_t would have no meaning, else we are implicitly admitting that these values *are* being included

in the model. For the total number of observed values used we'd still have t observations used to calculate the regression, but the matrix X has length $t - p$. Taking the appropriate index values for lag vectors then cutting the matrix at the appropriate index, all the while accounting for pythons choice of indexing arrays from zero rather than one can be a bit tricky, but with careful thought some reliable general formulas are always to be had as is demonstrated in the coding of examples within the github repository.

t test

Substituting the true (not hypothesized) value of y into equation 18 gives,

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + u) \tag{20}$$

from which we have,

$$\hat{\beta} - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'u \tag{21}$$

if u values are gaussian i.i.d random variables with mean zero and variance σ^2 it can be shown that[8] $\hat{\beta} \sim N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$. The central limit theorem would then imply that

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2(\mathbf{X}'\mathbf{X})^{-1}_{i,i}}} \sim N[0, 1] \tag{22}$$

Using this it can be shown that[9]

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{s^2(\mathbf{X}'\mathbf{X})^{-1}_{i,i}}} \sim t(n - p - 1) \tag{23}$$

where $s^2 = [(y - \mathbf{X}\hat{\beta})/(n - p)]^2$, p being the number of parameters estimated (i.e. the length of vector β), and $t(n - p - 1)$ represents the t distribution with $n - p - 1$ degress of freedom.

To perform hypothesis testing we first ask to what degree of accuracy we are interested in. For example we may want to ask if the hypothesis we've made has a %95 probability of being true. We'd then find the critical t values which correspond to the point at which the t distribution accumulates (integrates) to a value of 0.05 and also to 0.95. These are the 'tails' of the probaility distribution. Any t value which falls within the area for which the cumulative probability is less than 0.05 or greater than 0.95 is considered to be an improbable outcome. If our hypothesis is true then we'd at least expect it not to be a highly improbable outcome, so if the t statistic falls within the tail regions (i.e. is less than $t_{0.05}$ or is greater than $t_{0.95}$) we reject whatever hypothesis (usually denoted H_0) we have made.

Obviously, t tests are good for testing hypothesis which involve only one β_j variable.

F test

An alternative method to conduct hypothesis testing when we want to make hypothesis involving multiple β_j variables at once is the F test. Define the hypothesis as,

$$H_0 : \mathbf{R}\beta = \mathbf{r}$$

[8] c.f. Hamilton pg. 203
[9] c.f. *Introduction to Probability and Statistics* by Bain and Engelhardt, pg. 522

where $\mathbf{R}$ is some matrix and $\mathbf{r}$ is some column vector which make whatever hypothesis we want to make true. For example, if we wanted to test whether the true value of $\beta_2 = 0$ while $\beta_1 = \beta_3$ we would use,

$$[\mathbf{R}] = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

and,

$$[\mathbf{r}] = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Using the results of example 7.5 of Hamilton which gives the distributional effect of a constant matrix multiplied by a guassian variabe we have,

$$\mathbf{Rb} \sim N[r, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$$

Finally, using proposition 8.1 in Hamilton which states that given $z \sim N[0, \Omega]$ then $z'\Omega^{-1}z \sim \chi^2(n)$ it can be shown[10] that

$$F = (\mathbf{Rb} - \mathbf{r})' \left[s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} (\mathbf{Rb} - \mathbf{r}) \sim \chi^2(m)$$

But for reference purposes we take F/m where m is the length of r, or the number of restrictions we've imposed.

In the case that $H_0 : \mathbf{R}\beta = \mathbf{r}$ is a hypothesis which involves only one outcome ($m = 1$) for some β_j variable ten the F statistic reduces to the square of the t statistic.

Hypothesis tests for non stationary series

Though OLS estimates are used extensively in time-series analysis, in fact it is the projection theorem being used. The power of the projection theorem is that it does not presume causality in the sense that it makes no difference whether the observations x_j caused y or y caused the x_j observations. An analysis of the equivalence between OLS and the projection theorem is offered on my github rep in the $ACF\_Innovations\_READ\_ME$ file contained in the $Time - Series - Analysis$ folder. There I did not give adequate accounting of a subtle point of which one should be cognizent of; the projection method reduces to OLS only when the process it is being applied to is A) covariance stationary and B) ergotic for second moments.

For the second moments of a process y_t to be ergotic we require that,

$$\frac{1}{T-j} \sum_{t=j+1}^{T} (y_t - \mu)(y_{t-j} - \mu) \xrightarrow{\text{p}} \gamma_j \qquad for\ all\ j \tag{24}$$

where μ stands for the expected value of y_t. The right arrow with the symbol 'p' above it stands for 'approaches with probabilty $p = 1$'.

We have seen that a unit root process is not covariance stationary, so we would not expect it to be ergotic in second moments so that an OLS t test would be appropriate. As for the time trend model, were we to arbitrarily add a time trend δt to the right side of the general $AR(p)$ process discussed thus far we could then write it in a form similar to equation 15;

$$y_t = \psi(1)\alpha + \psi(L)(\delta t + \varepsilon_t) \tag{25}$$

[10] see page 205 of Hamilton

Though this model may not have a stationary mean, it *is* covariance stationary as can be seen by the fact that $E[y_t] = \psi(1)\alpha + \psi(L)\delta t = \mu$ so $\gamma_j = E(y_t - \mu)(y_{t-j} - \mu) = E(\psi(L)\varepsilon_t)(\psi(L)\varepsilon_{t-j})$ which can be shown to be the same for all values of t[11]. However, we still face the facts that A) the error terms $\psi(L)\varepsilon_t$ are correlated with the matrix $\mathbf{X}$ (e.g. y_{t-1} is dependent on ε_{t-1}, a condition known as *serial correlation*) and B) the error terms are not gaussian. The latter could be remedied with a normalization procedure which will be discussed shortly, but note that the term $\psi(L)\delta t$ has infinite terms. Even were we to truncate the series, the number of columns of $\mathbf{X}$ would be large, and the normalization procedure needed to find the appropriate asymptotic distribution would be cumbersome. In the end we'd likely not find any coresponding tables for such a model from which we could pull critical t values[12], so we'd have to derive them ourselves.

Were a time trend not to be included it could be shown that as the sample size grows the standard t test becomes a approximately accurate test statistic[13].

Rather than transforming our regression equation into the simplified form of equation 15 the procedure most often followed in order to find the appropriate asymptotic distribution for testing is to perform some other clever transformation – one which does not involve inversion (making use of $(1 - \phi(L))^{-1} = \psi(L)$ as in equation 15) – of the regression equation such that it is in a stationary form, normalizing the regression equation according to the asymptotic orders of its values, then using this to derive the asymptotic distribution of y_t. If such a transformation is also ergotic in second moments, then OLS regression can be used, but the asymptotic distribution may be different than the standard t or F distributions, and this is a subtle point which has somewhat profound implications, as will be demonstrated shortly.

As for the transformation required to produce a stationary process, what transformation is appropriate depends on the specific model being considered. For example, including a constant or a time trend in the model would require a different transformation than were these terms not included. For an example of transforming a model of the form of equation 7 with a time trend included see page 464 of Hamilton. For an example of transforming a unit root process with a time trend included see pages 497-500 which provides a numerical example which we can use to calculate the t statistic using either the original model or the transformed regression model. Though it was omitted from the codes, as a sanity check the t statistic was also calculated using the transformed regression. The resulting t values of the two models are indeed identicle. However, though the calculation of the t value may be the same in these examples, it is important to note once again that ***the asymptotic distribution of the t value for stochastic processes is NOT [always] the same as the standard t distribution!***. Ergo, we may need to reference different tables for this derived t statistic. The appropriate tables are referred to as 'Dickey-Fuller Tests based on estimated OLS t [or F] statistic'. We now turn to a basic example of a random walk process which highlights what makes for the difference in outcomes of the asymptotic distributions.

Random walk

H_0: True process is random walk with drift ($\alpha \neq 0$):

As previouslly mentioned, after transforming the regression model into a stationary series the next step is to normalize the transformed regression so that all quantities considered converge. This is best demonstrated through a simple example. Consider the model,

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t$$

In the case that $\rho = 1$the above process would be a random walk process with drift. We can go about constructing a test for the hypothesis that $\rho = 1$ (a unit root) by constructing the regression matrix equation

[11] See pages 192-193 of Hamilton. Also note we'd need to multiply the ergotic condition by $\frac{T}{T}$, move the T term in the denominator into Hamiltons expression, then take the limit of the remaining term $\frac{T}{T-j} \to 1$ as $T \to \infty$

[12] on second thought someone probably has done it, but it would be hard to find.

[13] c.f. pages 215-216 of Hamilton

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_1 \\ y_0 \end{bmatrix} = \begin{bmatrix} 1 & y_{t-1} \\ 1 & y_{t-2} \\ 1 & y_{t-3} \\ \vdots & \vdots \\ 1 & y_0 \\ 1 & y_{-1} \end{bmatrix} \begin{bmatrix} \alpha \\ \rho \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \\ \vdots \\ \varepsilon_1 \\ \varepsilon_0 \end{bmatrix}$$

Equation 18 applied to the above model gives,

$$\begin{bmatrix} \hat{\alpha} - \alpha \\ \hat{\rho} - 1 \end{bmatrix} = \begin{bmatrix} T & \sum y_{s-1} \\ \sum y_{s-1} & \sum y_{s-1}^2 \end{bmatrix} \begin{bmatrix} \sum \varepsilon_s \\ \sum y_{s-1}\varepsilon_s \end{bmatrix} \tag{26}$$

where the summations are taken to run from $s = 1$ to $s = t$. Under the null hypothesis that $\rho = 1$ we have $y_t = \alpha + y_{t-1} + \varepsilon_t$ which can be iterated to achieve the form

$$y_t = \alpha(t-1) + \sum_{s=2}^{t} \varepsilon_s = \alpha t + S_t \tag{27}$$

The term $\sum_{s=2}^{t} \varepsilon_s$ is equivalent to the function S_t in the *Wiener_READ_ME.pdf* file within the github repo, and we will use S_t as a substitution whenever convenient.

We want to know to what orders of t do the terms contained in the right-most vector of equation 26 converge to. Using the aforementioned substitution of S_t and equation 27 we get

$$\begin{aligned} \sum_{s=2}^{t} \varepsilon_s &= S_t \\ \sum_{s=2}^{t} y_{s-1}\varepsilon_s &= \sum_{s=1}^{t} (\alpha(t-1) + S_{t-1})\varepsilon_s \end{aligned} \tag{28}$$

Taking the variance of these expressions reveals that,

$$\begin{aligned} VAR\left[S_{t-1}\right] &= \sum_{s=1}^{t} \sigma^2 \propto t \\ VAR\left[\sum_{s=2}^{t} y_{s-1}\varepsilon_{s-1}\right] &= VAR\left[\sum_{s=2}^{t} (\alpha(t-1) + S_{t-1})\varepsilon_s\right] \\ &\propto \sum_{s=2}^{t} t^2 \to \frac{t^3}{3} \propto t^3 \end{aligned} \tag{29}$$

where the proportional ($\propto$) symbol was used because we are only concerned with whatever the highest power of t is in the resulting expression. In the second equation this happens to be $\sum_{s=1}^{t} t^2 \to t^3/3$ for which we used the general relation[14]

$$\sum_{s=1}^{t} s^v \to \frac{t^{v+1}}{v+1} \tag{30}$$

[14] this only represents the *highest* order term which $\sum_{s=1}^{t} s^v$ is proportional to.

Noting that it is the variance which is proportional to the derived powers of t, accordingly we take the square root of these terms to construct a matrix $\mathbf{M}_t$,

$$\mathbf{M}_t = \begin{bmatrix} t^{1/2} & 0 \\ 0 & t^{1/3} \end{bmatrix}$$

Multiplying equation 26 by $\mathbf{M}_t$ and using $\mathbf{M}_t\mathbf{M}_t^{-1} = \mathbf{I}$ gives,

$$\begin{aligned}
\begin{bmatrix} t^{1/2}(\hat{\alpha}-\alpha) \\ t^{3/2}(\hat{\rho}-1) \end{bmatrix} &= \mathbf{M}_t \left[\begin{bmatrix} T & \sum y_{s-1} \\ \sum y_{s-1} & \sum y_{s-1}^2 \end{bmatrix} \right]^{-1} \mathbf{M}_t \mathbf{M}_t^{-1} \begin{bmatrix} \sum \varepsilon_s \\ \sum y_{s-1}\varepsilon_s \end{bmatrix} \\
&= \left[\mathbf{M}_t^{-1} \begin{bmatrix} T & \sum y_{s-1} \\ \sum y_{s-1} & \sum y_{s-1}^2 \end{bmatrix} \mathbf{M}_t^{-1} \right]^{-1} \mathbf{M}_t^{-1} \begin{bmatrix} \sum \varepsilon_s \\ \sum y_{s-1}\varepsilon_s \end{bmatrix} \\
&= \begin{bmatrix} 1 & t^{-2}\sum y_{s-1} \\ t^{-2}\sum y_{s-1} & t^{-3}\sum y_{s-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} t^{-1/2}\sum \varepsilon_s \\ t^{-3/2}\sum y_{s-1}\varepsilon_s \end{bmatrix} \\
&= \begin{bmatrix} 1 & t^{-2}\sum y_{s-1} \\ t^{-2}\sum y_{s-1} & t^{-3}\sum y_{s-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} t^{-1/2}\sum \varepsilon_s \\ t^{-3/2}\sum y_{s-1}\varepsilon_s \end{bmatrix}
\end{aligned} \tag{31}$$

Lastly, we need to find the asymptotic distributions of all terms contained in the matrix $(\mathbf{X}\mathbf{X}')^{-1}$ (the first matrix on the right). The reasons for this are clear from the results on t and F tests, either of which has $(\mathbf{X}\prime\mathbf{X})^{-1}$ in the variance of the testing distribution. We follow a similar procedure as in equation(s) 29 but this time we want to be as explicit as possible, so the constant term in the denominator of equation 30 will not be disregarded.

$$\begin{aligned}
t^{-2}\sum_{s=1}^{t} y_{s-1} &= t^{-2}\sum_{s=1}^{t}(\alpha(s-1)+S_{t-1}) \to \frac{\alpha}{2} \\
t^{-3}\sum_{s=1}^{t} y_{s-1}^2 &= t^{-3}\sum_{s=1}^{t}(\alpha(t-1)+S_{t-1})^2 \\
&\propto \alpha^2 t^{-3}\sum_{s=1}^{t} t^2 \\
&\to \frac{\alpha^2}{3}
\end{aligned} \tag{32}$$

This gives for the asymptotic distribution of the matrix $(\mathbf{X}\mathbf{X}')^{-1}$,

$$\mathbf{M}_t(\mathbf{X}\mathbf{X}')^{-1}\mathbf{M}_t \xrightarrow[t\to\infty]{} \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/2 \end{bmatrix}$$

When deterministic terms dominate the asymptotic results there is no need to deviate from the previous analysis on t and F tests; the standard t and F distributions/ tables can be referenced.

Random Walk:
H_0: True process is random walk without drift ($\alpha = 0$):

For the case that the true process has no drift, the model we use to construct the regression matrix equation is the same;

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t$$

what changes is that when we go to substitute for y_t we instead use $y_t = y_{t-1} + \varepsilon_t$ as opposed to $y_t = \alpha + y_{t-1} + \varepsilon_t$ as was the case in the previous analysis. This will have a dramatic effect on the resulting t statistic because there is no deterministic term which ends up dominating the asymptotic values. Consider the iteration of y_t under the null hypothesis $\rho = 1, \alpha = 0$

$$\begin{aligned} y_t &= y_{t-1} + \varepsilon_t = y_0 + \sum_{s=1}^{t} \varepsilon_s \\ &= y_0 + S_t = S_t \end{aligned}$$

Where we took $y_0 = 0$ for convenience. In contrast to the previous analysis, the term αt does not arise. The effect on the asymptotic resusts are as follows,

$$\begin{aligned} &VAR[S_{t-1}] \propto t \\ &VAR\left[\sum_{s=1}^{t} y_{s-1}\varepsilon_{s-1}\right] = VAR\left[\sum_{s=1}^{t} (S_{t-1})\varepsilon_{s-1}\right] \propto t^2 \end{aligned} \tag{33}$$

Accordingly, we construct the normalization matrix,

$$\mathbf{M}_t = \begin{bmatrix} t^{1/2} & 0 \\ 0 & t \end{bmatrix}$$

Going through the same steps as before gives,

$$\begin{aligned} \mathbf{M}_t(\mathbf{X}\mathbf{X}')^{-1}\mathbf{M}_t &= \begin{bmatrix} 1 & t^{-3/2}\sum y_{s-1} \\ t^{-3/2}\sum y_{s-1} & t^{-2}\sum y_{s-1}^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & t^{-3/2}\sum S_{t-1} \\ t^{-3/2}\sum S_{t-1} & t^{-2}\sum S_{t-1}^2 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & \sigma\int W(r)dr \\ \sigma\int W(r)dr & \sigma^2\int [W(r)]^2 dr \end{bmatrix} \end{aligned}$$

Where the last line resulted from a non-trivial analysis of the asymptotic form of the expressions in $\mathbf{M}_t(\mathbf{X}\mathbf{X}')^{-1}\mathbf{M}_t$. Again, note that this term plays a decisive role in determining the variance of the asymptotic distribution. The distributional results can be found on page 486 of Hamilton as well as in numerous other texts. In the file *Wiener_READ_ME.pdf* contained within the github repo the reader will find a derivation of of the Wiener process, the end result of which is a random walk process.

So we have seen a dramatic difference in outcome when stochastic rather than deterministic terms asymptotically dominate the variance. Though the calculation of the t and F statistics from our regression model remain the same, the critical values we reference for comparison will change.

5 Differencing Data for Unit Root Models

Here we will briefly discuss a transformation which is used prevalently in some of the more advanced analysis presented by Hamilton in chapters 17-19 and within the examples covered in the github repo.

Consider once again the generalized $AR(p)$ process,

$$(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p) y_t = \alpha + \varepsilon_t \tag{34}$$

defining the variables,

$$\begin{aligned} \zeta_j &= \phi_{j+1} + \phi_{j+2} + ... + \phi_p \\ \rho &= \phi_1 + \phi_2 + ... + \phi_p \end{aligned} \tag{35}$$

the reader may verify that the above variables allow for the transformation,

$$(1-\rho L)y_t - (\xi_1 L + \xi_2 L^2 + \ldots + \xi_{p-1}L^{p-1})(1-L)y_t = \alpha + \varepsilon_t \tag{36}$$

or equivalently,

$$y_t = \xi_1 \Delta y_{t-1} + \xi_2 \Delta y_{t-2} + \ldots + \xi_{p-1}\Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \varepsilon_t \tag{37}$$

Notice that the above equation is in a form suitable for testing for a unit root; $\rho = \phi_1 + \phi_2 + ... + \phi_p = 1$. In matrix form we have,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_p \\ y_{p-1} \\ \vdots \\ y_2 \\ y_1 \\ y_0 \end{bmatrix} = \begin{bmatrix} \Delta y_{t-1} & \Delta y_{t-2} & \ldots & \Delta y_{t-p+1} & y_{t-1} & 1 \\ \Delta y_{t-2} & \Delta y_{t-3} & \ldots & \Delta y_{t-p} & y_{t-2} & 1 \\ \Delta y_{t-3} & \Delta y_{t-4} & \ldots & \Delta y_{t-p-1} & y_{t-3} & 1 \\ \ddots & \vdots & \vdots & \vdots & \vdots & 1 \\ \Delta y_{p-1} & \Delta y_{p-2} & \ldots & \Delta y_1 & y_{t-p} & 1 \\ \Delta y_{p-2} & \Delta y_{p-3} & \ldots & \Delta y_{-1} & y_{p-2} & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & 1 \\ \Delta y_1 & \Delta y_0 & \ldots & \Delta y_{-p+3} & y_1 & 1 \\ \Delta y_0 & \Delta y_{-1} & \ldots & \Delta y_{-p+2} & y_0 & 1 \\ \Delta y_{-1} & \Delta y_{-2} & \ldots & \Delta y_{-p+1} & y_{-1} & 1 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{p-1} \\ \rho \\ \alpha \end{bmatrix} + \begin{bmatrix} \hat{u}_t \\ \hat{u}_{t-1} \\ \hat{u}_{t-2} \\ \vdots \\ \hat{u}_p \\ \hat{u}_{p-1} \\ \vdots \\ \hat{u}_2 \\ \hat{u}_1 \\ \hat{u}_0 \end{bmatrix} \tag{38}$$

Again, properly indexing the differenced lagged vectors and truncating the matrix at the appropriate index can be a bit tricky. For an example of the appropriate method for indexing lagged difference values to construct the above matrix see the file *Cointegration_test_CPI_US_vs_Italy.py* within the github repo.

One benefit of differencing data is that it may eliminate trends in the data we are not interested in. The reader can verify for example, that all the the differenced regressors Δy_{t-j} are $I[0]$ processes while y_{t-1} is an $I[1]$ process, and this may be useful if we are only concerned with examining the relation which the present value of y_t has with its immediate past value. This property also proves useful in asymptotic analysis. Lastly, the form of 38 may avoid problems with to do with serial correlation and spurious regression which is a phenomena in which unrelated vectors appear to be related by the resulting t statistic when in fact they are not, and this obviously is not a desirable thing when our interest is in testing for cointegration.

6 Spurious Regression

Recall that OLS tests are only valid when applied to covariance stationary processes. Also recall in the consumer price index example we defined the variable,

$$z_t = p_t - s_t - p_t^*$$

If z_t is not covariance stationary then any t or F tests we perform for cointegration will not be reliable. Recall that the hypothesized relation was actually that $p_t = s_t + p_t^*$ (the natural log of the purchase power parity hypothesis $P_t = S_t P_t$). Were we to perform a regression of the form $p_t = \alpha + \beta_1 s_t + \beta_2 p_t^*$ then our hypothesis would be that $\alpha = 0$ and $\beta_1 = \beta_2 = 1$. The residual error $\hat{u}_t$ in our regression is,

$$\hat{u}_t = p_t - \alpha - \beta_1 s_t - \beta_2 p_t^*$$

which if our hypothesis that $\alpha = 0$ and $\beta_1 = \beta_2 = 1$ were true would imply that $z_t = \hat{u}_t$.

Before applying any t test to see if our hypothesis is true, we at least first ought to ask whether the process we are applying it to is covariance stationary. This can be accomplished by performing a regression on $\hat{u}_t$ then checking whether the $\hat{u}_t$ has a unit root, which as we've seen implies $\hat{u}_t$ is nonstationary. The regression is of the form

$$\hat{u}_t = \rho \hat{u}_{t-1} + \varepsilon_t$$

if $\rho \approx 1$ then we conclude that our t and F test results on our original hypothesis are not reliable. But the fact that the OLS estimates of $\hat{u}_t$ are non-stationary also implies that no cointegration relation exists between between the vectors p_t, s_t, p_t^* in the first place. A proof of this statement would involve something called the triangular representation of cointegrating vectors for the system p_t, s_t, p_t^*. For details see Hamilton, pg 590. The file *residual_test_leads_AND_lags.py* in the github repo performs residual regression testing on personal disposable income expenditures vs. personal disposable income in the U.S. for the years 1947-89 to test whether a cointegration relation exists between them.

7 A Note on the Github Codes

I have a tendency to code first and understand what I'm coding second. Though I made an effort to organize the code so that it is legible, and by making numerous references to page numbers and specific equations in Hamiltons text throughout the code I think the interested reader can deduce what is happening without too much trouble so long as they have Hamiltons text available.

Before testing for cointegration, one should test whether the vectors being tested for cointegration are themselves $I[1]$ processes in the first place. Such preliminary tests are not performed in any of the examples presented in the github repo, but the codes can be easily tweaked to this ends.

Finally, note that cointegration tests may produce varying results depending on which vector is chosen to be the regression variable, e.g. in the above example we chose $p_t = \alpha + \beta_1 s_t + \beta_2 p_t^*$ which may produce different results than were we to choose $p_t^* = \alpha + \beta_1 s_t^* + \beta_2 p_t$ (where s_t^* would be the Italy to U.S. exchange rate). This problem can be avoided in the full information maximum likilehood approach to cointegration testing (chapter 20 of Hamilton) which was not discussed here.

Cointegration analysis in its full form is non-trivial. Here I have provided the interested reader with an overview of the basics, an understanding of which should make some of the more advanced theoretical treatments easier to comprehend.

References

[1] James D. Hamilton *Time Series Analysis*. Princeton University Press, New Jersey 1994