SUPPLEMENTARY MANUAL TO ACCOMPANY

APPLIED ECONOMETRIC TIME SERIES (4th edition)

Walter Enders

University of Alabama

CHAPTER 1

Endnotes to Chapter 1

1. Another possibility is to obtain the forward-looking solution. Since we are dealing with forecasting equations, forward looking solutions are not important for our purposes. Some of the details concerning forward-looking solutions are included in the *Supplementary Manual* available on the Wiley website or my website.

2. Alternatively, you can substitute (1.26) into (1.17). Note that when ε_t is a pure random disturbance, $y_t = a_0 + y_{t-1} + \varepsilon_t$ is called a random walk plus drift model.

3. Any linear equation in the variables z_1 through z_n is homogeneous if it has the form $a_1z_1 + a_2z_2 + \dots + a_nz_n = 0$. To obtain the homogeneous portion of (1.10), simply set the intercept term a_0 and the forcing process x_t equal to zero. Hence, the homogeneous equation for (1.10) is $y_t = a_1y_{t-1} + a_2y_{t-2} + \dots + a_ny_{t-n}$.

4. If $b > a$, the demand and supply curves do not intersect in the positive quadrant. The assumption $a > b$ guarantees that the equilibrium price is positive.

5. For example, if the forcing process is $x_t = \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \dots$, the impact multiplier is the partial derivative of y_t with respect to ε_t.

Section 1.1

Appendix 1.1: Imaginary Roots and de Moivre's Theorem

Consider a second-order difference equation $y_t = a_1 y_{t-1} + a_2 y_{t-2}$ such that the discriminant d is negative [i.e., $d = a_1^2 + 4a_2 < 0$]. From Section 6, we know that the full homogeneous solution can be written in the form

$$y_t^h = A_1\alpha_1^t + A_2\alpha_2^t \qquad \text{(A1.1)}$$

where the two imaginary characteristic roots are

$$\alpha_1 = (a_1 + i\sqrt{-d}\,)/2 \text{ and } \alpha_2 = (a_1 - i\sqrt{-d}\,)/2 \qquad \text{(A1.2)}$$

The purpose of this Section is to explain how to rewrite and interpret (A1.1) in terms of standard trigonometric functions. You might first want to refresh your memory concerning two useful trig identities. For any two angles θ_1 and θ_2,

$$\sin(\theta_1 + \theta_2) = \sin(\theta_1)\cos(\theta_2) + \cos(\theta_1)\sin(\theta_2)$$

$$\cos(\theta_1 + \theta_2) = \cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2) \qquad \text{(A1.3)}$$

If $\theta_1 = \theta_2$, we can drop subscripts and form

$$\sin(2\theta) = 2\sin(\theta)\cos(\theta)$$

$$\cos(2\theta) = \cos(\theta)\cos(\theta) - \sin(\theta)\sin(\theta) \qquad \text{(A1.4)}$$

The first task is to demonstrate how to express imaginary numbers in the complex plane. Consider Figure A1.1 in which the horizontal axis measures real numbers and the vertical axis measures imaginary numbers. The complex number $a + bi$ can be represented by the point a units from the origin along the horizontal axis and b units from the origin along the vertical axis. It is convenient to represent the distance from the origin by the length of the vector denoted by r. Consider angle θ in triangle $0ab$ and note that $\cos(\theta) = a/r$ and $\sin(\theta) = b/r$. Hence, the lengths a and b can be measured by

$$a = r\cos(\theta) \qquad \text{and} \qquad b = r\sin(\theta)$$

In terms of (A1.2), we can define $a = a_1/2$ and $b = \sqrt{-d}\,/\,2$. Thus, the characteristic roots α_1 and α_2 can be written as:

$$\alpha_1 = a + bi = r[\,\cos(\theta) + i\sin(\theta)]$$

$$\alpha_2 = a - bi = r[\,\cos(\theta) - i\sin(\theta)] \qquad \text{(A1.5)}$$

The next step is to consider the expressions $\alpha_1{}^t$ and $\alpha_2{}^t$. Begin with the expression $\alpha_1{}^2$ and recall that $i^2 = -1$:

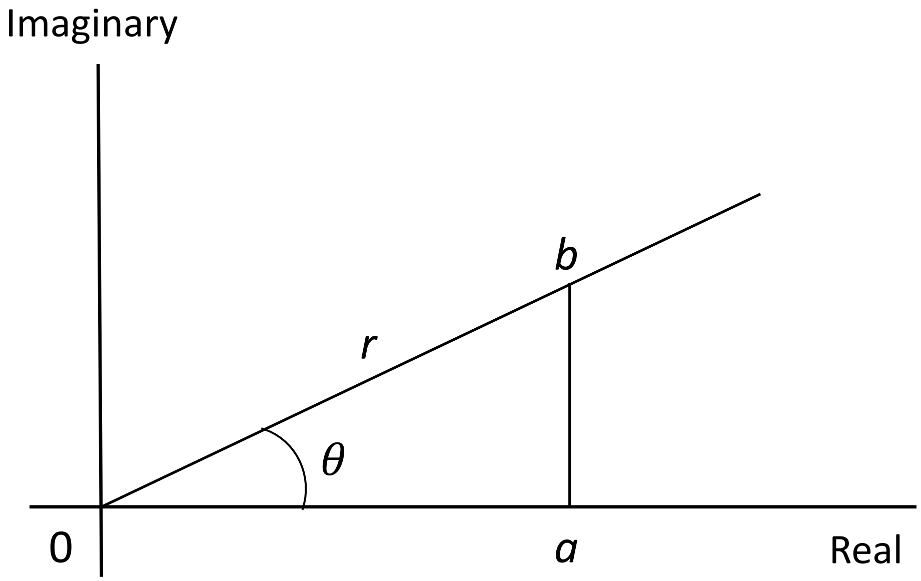


Figure A1.1 Graphical Representation of Complex Numbers

$$\alpha_1{}^2 = \{r[\cos(\theta) + i\sin(\theta)]\}\{r[\cos(\theta) + i\sin(\theta)]\}$$

$$= r^2[\cos(\theta)\cos(\theta) - \sin(\theta)\sin(\theta) + 2i\sin(\theta)\cos(\theta)]$$

From (A1.4),

$$\alpha_1{}^2 = r^2[\cos(2\theta) + i\sin(2\theta)]$$

If we continue in this fashion, it is straightforward to demonstrate that

$$\alpha_1{}^t = r^t[\cos(t\theta) + i\sin(t\theta)] \quad \text{and} \quad \alpha_2{}^t = r^t[\cos(t\theta) - i\sin(t\theta)]$$

Since y_t^h is a real number and α_1 and α_2 are complex, it follows that A_1 and A_2 must be complex. Although A_1 and A_2 are arbitrary complex numbers, they must have the form

$$A_1 = B_1[\cos(B_2) + i\sin(B_2)] \text{ and } A_2 = B_1[\cos(B_2) - i\sin(B_2)] \qquad \text{(A1.7)}$$

where B_1 and B_2 are arbitrary real numbers measured in radians.

In order to calculate $A_1(\alpha_1{}^t)$, use (A1.6) and (A1.7) to form

$$A_1\alpha_1{}^t = B_1[\cos(B_2) + i\sin(B_2)]r^t[\cos(t\theta) + i\sin(t\theta)]$$

$$= B_1r^t[\cos(B_2)\cos(t\theta) - \sin(B_2)\sin(t\theta) + i\cos(t\theta)\sin(B_2) + i\sin(t\theta)\cos(B_2)]$$

Using (A1.3), we obtain

$$A_1\alpha_1^{\,t} = B_1 r^t[\ \cos(t\theta + B_2) + i\ \sin(t\theta + B_2)\ ] \qquad \text{(A1.8)}$$

You should use the same technique to convince yourself that

$$A_2\alpha_2^{\,t} = B_1 r^t[\ \cos(t\theta + B_2) - i\ \sin(t\theta + B_2)\ ] \qquad \text{(A1.9)}$$

Since the homogeneous solution y_t^h is the sum of (A1.8) and (A1.9),

$$y_t^h = B_1 r^t[\ \cos(t\theta + B_2) + i\ \sin(t\theta + B_2)\ ] + B_1 r^t[\ \cos(t\theta + B_2) - i\ \sin(t\theta + B_2)\ ] \qquad \text{(A1.10)}$$

$$= 2B_1 r^t \cos(t\theta + B_2)$$

Since B_1 is arbitrary, the homogeneous solution can be written in terms of the arbitrary constants B_2 and B_3

$$y_t^h = B_3 r^t \cos(t\theta + B_2) \qquad \text{(A1.11)}$$

Now imagine a circle with a radius of unity superimposed on Figure A1.1. The stability condition is for the distance $r = 0b$ to be less than unity. Hence, in the literature it is said that the stability condition is for the characteristic root(s) to lie within this unit circle.

Section 1.2

Appendix 1.2: Characteristic Roots in Higher-Order Equations

The characteristic equation to an *n*th-order difference equation is

$$\alpha^n - a_1\alpha^{n-1} - a_2\alpha^{n-2} \ldots - a_n = 0 \tag{A1.12}$$

As stated in Section 6, the *n* values of α which solve this characteristic equation are called the **characteristic roots**. Denote the *n* solutions by α_1, α_2, … α_n. Given the results in Section 4, the linear combination $A_1\alpha_1^t + A_2\alpha_2^t + \ldots + A_n\alpha_n^t$ is also a solution to (A1.12)

A priori, the characteristic roots can take on any values. There is no restriction that they be real versus complex nor any restriction concerning their sign or magnitude. Consider the possibilities:

1. **All the α_i are real and distinct.** There are several important subcases. First suppose that each value of α_i is less than unity in absolute value. In this case, the homogeneous solution (A1.12) converges since the limit of each α^t_i equals zero as *t* approaches infinity. For a negative value of α_i, the expression α^t_i is positive for even values of *t* and negative for odd values of *t*. Thus, if any of the α_i are negative (but less than one in absolute value), the solution will tend to exhibit some oscillation. If any of the α_i are greater than unity in absolute value, the solution will diverge.

1. **All of the α_i are real but $m \le n$ of the roots are repeated.** Let the solution be such that $\alpha_1 = \alpha_2 = \ldots = \alpha_m$. Call the single distinct value of this root $\bar{\alpha}$ and let the other *n-m* roots be denoted by α_{m+1} through α_n. In the case of a second-order equation with a repeated root, you saw that one solution was $A_1\bar{\alpha}^t$ and the other was $A_2t\bar{\alpha}^t$. With *m* repeated roots, it is easily verified that $t\bar{\alpha}^t$, $t^2\bar{\alpha}^t$, … , $t^{m-1}\bar{\alpha}^t$ are also solutions to the homogeneous equation. With *m* repeated roots, the linear combination of all these solutions is

$$A_1\bar{\alpha}^t + A_2t\bar{\alpha}^t + A_3t^2\bar{\alpha}^t + \ldots + A_mt^{m-1}\bar{\alpha}^t + A_{m+1}\alpha_{m+1}^t + \ldots + A_n\alpha_n^t$$

2. **Some of the roots are complex.** Complex roots (which necessarily come in conjugate pairs) have the form $\alpha_i \pm i\theta$, where α_i and θ are real numbers and *i* is defined to be $\sqrt{-1}$. For any such pair, a solution to the homogeneous equation is: $A_1(\alpha_1 + i\theta)^t + A_2(\alpha_1 - i\theta)^t$ where A_1 and A_2 are arbitrary constants. Transforming to polar coordinates, the associated two solutions can be written in the form: $\beta_1 r^t\cos(\theta t + \beta_2)$ with arbitrary constants β_1 and β_2. Here stability hinges on the magnitude of r^t; if $|r| < 1$, the system converges. However, even if there is convergence, convergence is not direct because the sine and cosine functions impart oscillatory behavior to the time path of y_t. For example, if there are three roots, two of which are complex, the homogeneous solution has the form

$$\beta_1 r^t\cos(\theta t + \beta_2) + A_3(\alpha_3)^t$$

Stability of Higher-Order Systems

From equation (A1.12) of Section 1.2, the characteristic equation of an *n*th-order difference equation is

$$\alpha^n - a_1\alpha^{n-1} - a_2\alpha^{n-2} ... - a_n = 0 \qquad \text{(A1.12)}$$

Denote the *n* characteristic roots by α_1, α_2, ... α_n. Given the results in Section 4, the linear combination $A_1\alpha_1^t + A_2\alpha_2^t + ... + A_n\alpha_n^t$ is also a solution to (A1.12).

In practice, it is difficult to find the actual values of the characteristic roots. Unless the characteristic equation is easily factored, it is necessary to use numerical methods to obtain the characteristic roots. However, for most purposes it is sufficient to know the qualitative properties of the solution; usually it is sufficient to know whether all of the roots lie within the unit circle. The **Schur Theorem** gives the necessary and sufficient conditions for stability. Given the characteristic equation of (A1.12), the theorem states that if all of the *n* determinants below are positive, the real parts of all characteristic roots are less than one in absolute value.

$$\Delta_1 = \begin{vmatrix} 1 & -a_n \\ -a_n & 1 \end{vmatrix}$$

$$\Delta_2 = \begin{vmatrix} 1 & 0 & -a_n & -a_{n-1} \\ -a_1 & 1 & 0 & -a_n \\ -a_n & 0 & 1 & -a_1 \\ -a_{n-1} & -a_n & 0 & 1 \end{vmatrix} \qquad \Delta_3 = \begin{vmatrix} 1 & 0 & 0 & -a_n & -a_{n-1} & -a_{n-2} \\ -a_1 & 1 & 0 & 0 & -a_n & -a_{n-1} \\ -a_2 & -a_1 & 1 & 0 & 0 & -a_n \\ -a_n & 0 & 0 & 1 & -a_1 & -a_2 \\ -a_{n-1} & -a_n & 0 & 0 & 1 & -a_1 \\ -a_{n-2} & -a_{n-1} & -a_n & 0 & 0 & 1 \end{vmatrix} \quad ...$$

$$\Delta_n = \begin{vmatrix} 1 & 0 & 0 & .\ . & 0 & -a_n & -a_{n-1} & .\ .\ . & & -a_1 \\ -a_1 & 1 & 0 & .\ . & 0 & 0 & -a_n & .\ .\ . & & -a_2 \\ -a_2 & -a_1 & 1 & .\ . & 0 & 0 & 0 & -a_n & .\ . & -a_3 \\ . & . & . & .\ . & . & . & . & .\ . & . & . \\ -a_{n-1} & -a_{n-2} & -a_{n-3} & .\ . & 1 & 0 & 0 & 0 & .\ . & -a_n \\ -a_n & 0 & 0 & .\ . & 0 & 1 & -a_1 & -a_2 & .\ . & -a_{n-1} \\ -a_{n-1} & -a_n & 0 & .\ . & 0 & 0 & 1 & .\ .\ . & & -a_{n-2} \\ . & . & . & .\ . & . & . & . & .\ . & . & . \\ -a_2 & -a_3 & -a_4 & .\ . & 0 & 0 & 0 & .\ . & 1 & -a_1 \\ -a_1 & -a_2 & -a_3 & .\ . & -a_n & 0 & 0 & .\ .\ . & & 1 \end{vmatrix}$$

To understand the way each determinant is formed, note that each can be partitioned into four subareas. Each subarea of Δ_i is a triangular $i \times i$ matrix. The northwest subarea has the value 1 on the diagonal and all zeros above the diagonal. The subscript increases by one as we move down any column beginning from the diagonal. The southeast subarea is the transpose of the northwest subarea. Notice that the northeast subarea has a_n on the diagonal and all zeros below the diagonal. The subscript decreases by one as we move up any column beginning from the diagonal. The southwest subarea is the transpose of the northeast subarea. As defined above, the value of a_0 is unity.

Special Cases: As stated above, the **Schur Theorem** gives the necessary and sufficient conditions for all roots to lie in the unit circle. Rather than calculate all of these determinants, it is often possible to use the simple rules discussed in Section 6. Those of you familiar with matrix algebra may wish to consult Samuelson (1941) for formal proofs of these conditions.

Section 1.3: Forward Versus Backward Solutions

This Material Follows Section 9 of Chapter 1

Note that the equations are numbered consecutively following those in the text.

As suggested by equation (1.82), there is a **forward-looking** solution to any linear difference equation. This text will not make much use of the forward-looking solution since future realizations of stochastic variables are not directly observable. However, knowing how to obtain forward-looking solutions is useful for solving rational expectations models. Let's return to the simple iterative technique to consider the forward-looking solution to the first-order equation $y_t = a_0 + a_1y_{t-1} + \varepsilon_t$. Solving for y_{t-1}, we obtain

$$y_{t\text{-}1} = -(a_0 + \varepsilon_t)/a_1 + y_t/a_1 \qquad (1.83)$$

Updating one period

$$y_t = -(a_0 + \varepsilon_{t+1})/a_1 + y_{t+1}/a_1 \qquad (1.84)$$

Since $y_{t+1} = (y_{t+2} - a_0 - \varepsilon_{t+2})/a_1$, begin iterating forward:

$$\begin{aligned} y_t &= -(a_0 + \varepsilon_{t+1})/a_1 + (y_{t+2} - a_0 - \varepsilon_{t+2})/(a_1)^2 \\ &= -(a_0 + \varepsilon_{t+1})/a_1 - (a_0 + \varepsilon_{t+2})/(a_1)^2 + y_{t+2}/(a_1)^2 \\ &= -(a_0 + \varepsilon_{t+1})/a_1 - (a_0 + \varepsilon_{t+2})/(a_1)^2 + (y_{t+3} - a_0 - \varepsilon_{t+3})/(a_1)^3 \end{aligned}$$

Therefore, after n iterations,

$$y_t = -a_0\sum_{i=1}^{n} a_1^{-i} - \sum_{i=1}^{n} a_1^{-i}\varepsilon_{t+i} + y_{t+n}/a_1^n \qquad (1.85)$$

If we maintain that $|a_1| < 1$, this forward-looking solution will diverge as n gets infinitely large. However, if $|a_1| > 1$, the expression a_1^{-n} goes to zero while $-a_0(a_1^{-1} + a_1^{-2} + a_1^{-3} + \ldots)$ converges to $a_0/(1-a_1)$. Hence, we can write the forward-looking particular solution for y_t as

$$y_t = a_0/(1-a_1) - \sum_{i=1}^{n} a_1^{-i}\varepsilon_{t+i} \qquad (1.86)$$

Note that (1.86) is identical to (1.82). The key point is that the *future* values of the disturbances affect the present. Clearly, if $|a_1| > 1$ the summation is convergent so that (1.86) is a legitimate particular solution to the difference equation. Given an initial condition, a stochastic difference equation will have a forward- and a backward-looking solution. To illustrate the technique using lag operators, we can write the particular solution to $y_t = a_0 + a_1y_{t-1} + \varepsilon_t$ as $(a_0 + \varepsilon_t)/(1\text{-}a_1L)$. Now multiply the numerator and denominator by $-a_1^{-1}L^{-1}$ to form

$$y_t = a_0/(1-a_1) - a_1^{-1}L^{-1}\varepsilon_t/(1-a_1^{-1}L^{-1})$$

so that

$$y_t = a_0 /(1 - a_1) - \sum_{i=1}^{\infty} a_i^{-i} \varepsilon_{t+i} \qquad (1.87)$$

More generally, we can always obtain a forward-looking solution for any *n*th-order equation. (For practice in using the alternative methods of solving difference equations, try to obtain this forward looking solution using the method of undetermined coefficients.)

Properties of the Alternative Solutions

The backward- and forward-looking solutions are two mathematically valid solutions to any *n*-th order difference equation. In fact, since the equation itself is linear, it is simple to show that any linear combination of the forward- and backward-looking solutions is also a solution. For economic analysis, however, the distinction is important since the time paths implied by these alternative solutions are quite different. First consider the backward looking solution. If $|a_1| < 1$, the expression a_1^i converges towards zero as $i \to \infty$. Also, notice that the effect of ε_{t-i} on y_t is a_1^i; if $|a_1| < 1$, the effects of the past ε_t also diminish over time. Suppose instead that $|a_1| > 1$; in this instance, the backward-looking solution for y_t explodes.

The situation is reversed using the forward solution. Here, if $|a_1| < 1$, the expression a_1^{-i} becomes infinitely large as i approaches ∞. Instead, if $|a_1| > 1$, the forward- looking solution leads to a finite sequence for $\{y_t\}$. The reason is that a_1^{-i} converges to zero as i increases. Note that the effect of ε_{t+i} on y_t is a_1^{-i}; if $|a_1| > 1$, the effects of the future values of ε_{t+i} have a diminishing influence on the current value of y_t.

From a purely mathematical point of view, there is no "most appropriate" solution. However, economic theory may suggest that a sequence be **bounded** in the sense that the limiting value for any value in the sequence is finite. Real interest rates, real per capita income, and many other economic variables can hardly be expected to approach either plus or minus infinity. Imposing boundary restrictions entails using the backward-looking solution if $|a_1| < 1$ and using the forward-looking solution if $|a_1| > 1$. Similar remarks hold for higher-order equations.

An Example: Cagan's Money Demand Function

Cagan's model of hyperinflation provides an excellent example of illustrating the appropriateness of forward- versus backward-looking solutions. Let the demand for money take the form

$$m_t - p_t = \alpha - \beta (p^e_{t+1} - p_t) \qquad \beta > 0 \qquad (1.88)$$

where: m_t = logarithm of the nominal money supply in t

p_t = the logarithm of price level in t

p^e_{t+1} = the logarithm of the price level expected in period t+1

The key point of the model is that the demand for real money balances (m_t - p_t) is negatively related to the expected rate of inflation ($p^e_{t+1} - p_t$). Because Cagan was interested in the relationship between inflation and money demand, all other variables were subsumed into the constant α. Since our task is to work with forward-looking solutions, let the money supply function simply be the process:

$$m_t = m + \varepsilon_t$$

where m = the average value of the money supply

ε_t = a disturbance term with a mean value of zero

As opposed to the cobweb model, let individuals have forward-looking perfect foresight so the expected price for t+1 equals the price that actually prevails:

$$p^e_{t+1} = p_{t+1}$$

Under perfect foresight, agents in period t are assumed to know the price level in t+1. In the context of the example, agents are able to solve difference equations and can simply "figure out" the time path of prices. Thus, we can write the money market equilibrium condition as

$$m + \varepsilon_t - p_t = \alpha - \beta\,(\,p_{t+1} - p_t\,)$$

or

$$p_{t+1} - (1+1/\beta)p_t = -(m - \alpha)\,/\beta - \varepsilon_t/\beta \tag{1.89}$$

For practice, we use the method of undetermined coefficients to obtain the particular solution. (You should check your abilities by repeating the exercise using lag operators.) We use the forward-looking solution because the coefficient (1+1/β) is greater than unity in absolute value. Try the challenge solution

$$p^p_t = b_0 + \sum_{i=0}^{\infty} \alpha_i \; \varepsilon_{t+i}$$

Substituting this challenge solution into the above, we obtain

$$b_0 + \sum_{i=0}^{\infty} \alpha_i \; \varepsilon_{t+1+i} - \frac{1+\beta}{\beta}\left(b_0 + \sum_{i=0}^{\infty} \alpha_i \; \varepsilon_{t+i} \right) = \frac{\alpha - m - \varepsilon_t}{\beta} \tag{1.90}$$

For (1.90) to be an identity for all possible realizations of $\{\varepsilon_t\}$, it must be the case that

$$b_0 - b_0(1+\beta)/\beta = (\alpha - m)/\beta \quad \Rightarrow \quad b_0 = m - \alpha$$

$$-\alpha_0(1+\beta)/\beta = -1/\beta \quad \Rightarrow \quad \alpha_0 = 1/(1+\beta)$$

$$\alpha_0 - \alpha_1(1+\beta)/\beta = 0 \quad \Rightarrow \quad \alpha_1 = \beta/(1+\beta)^2$$

.

.

.

$$\alpha_i - \alpha_{i+1}(1+\beta)/\beta = 0 \quad \Rightarrow \quad \alpha_i = \beta^i/(1+\beta)^{i+1}$$

In compact form, the particular solution can be written as

$$p_t^p = m - \alpha + \frac{1}{\beta}\sum_{i=0}^{\infty}\left(\frac{\beta}{1+\beta}\right)^{1+i} \varepsilon_{t+i} \tag{1.91}$$

The next step is to find the homogeneous solution. Form the homogeneous equation p_{t+1} - (1+1/β)p_t = 0. For any arbitrary constant A, it is easy to verify that the solution is

$$p^h{}_t = A\ (1+1/\beta)^t$$

Therefore, the general solution is

$$p_t = m - \alpha + \frac{1}{\beta}\sum_{i=0}^{\infty}(\frac{\beta}{1+\beta})^{1+i} \varepsilon_{t+i} + A(1+1/\beta)^t \tag{1.92}$$

If you examine (1.92) closely, you will note that the impulse response function is convergent; the expression $[\beta/(1+\beta)]^{1+i}$ converges to zero as i approaches infinity. However, the homogeneous portion of the solution is divergent. For (1.92) to yield a non-explosive price sequence, we must be able to set the arbitrary constant equal to zero. To understand the economic implication of setting A = 0, suppose that the initial condition is such that the price level in period zero is p_0. Imposing this initial condition, (1.92) becomes

$$p_0 = m - \alpha + \frac{1}{\beta}\sum_{i=0}^{\infty}(\frac{\beta}{1+\beta})^{1+i} \varepsilon_i + A$$

Solving for A yields

$$A = p_0 + \alpha - m - \frac{1}{\beta}\sum_{i=0}^{\infty}(\frac{\beta}{1+\beta})^{1+i} \varepsilon_i$$

Thus, the initial condition must be such that

$$A = 0 \quad or \quad p_0 = m - \alpha + \frac{1}{\beta}\sum_{i=0}^{\infty}(\frac{\beta}{1+\beta})^{1+i} \varepsilon_i \tag{1.93}$$

Examine the three separate components of (1.92). The deterministic expression m - α is the same type of long-run "equilibrium" condition encountered on several other occasions; a stable sequence tends to converge toward the deterministic portion of its particular solution. The second component of the particular solution consists of the short-run responses induced by the various ε_t shocks. These movements are necessarily of a short-term duration because the coefficients of the impulse response function must decay. The point is that the particular solution captures the overall long-run and short-run equilibrium behavior of the system. Finally, the homogeneous solution can be viewed as a measure of disequilibrium in the initial period. Since (1.91) is the overall equilibrium solution for period t, it should be clear that the value of p_0 in (1.93) is the equilibrium value of the price for period zero. After all, (1.93) is nothing more than (1.91) with the time subscript lagged t periods. Thus, the expression $A(1+1/\beta)^t$ must be zero if the deviation from equilibrium in the initial period is zero.

Imposing the requirement that the $\{p_t\}$ sequence be bounded necessitates that the general solution be

$$p_t = m - \alpha + \frac{1}{\beta}\sum_{i=0}^{\infty}[\frac{\beta}{1+\beta}]^{1+i} \quad \varepsilon_{t+i}$$

Notice that the price in each and every period *t* is proportional to the mean value of the money supply; this point is easy to verify since all variables are expressed in logarithms and $\partial p_t/\partial m = 1$. Temporary changes in the money supply behave in an interesting fashion. The impulse response function indicates that *future* increases in the money supply, represented by the various ε_{t+i}, serve to increase the price level in the *current* period. The idea is that future money supply increases imply higher prices in the future. Forward-looking agents reduce their current money holdings, with a consequent increase in the current price level, in response to this anticipated inflation.

Practice question: Consider the Cagan demand for money function: $m_t - p_t = \alpha - \beta[p_{t+1} - p_t]$. Show that the backward-looking particular solution for p_t is divergent.

Answer: Using lag operators, rewrite the equation as $\beta p_{t+1} - (1 + \beta)p_t = \alpha - m_t$. Combining terms yields $[1 - (1 + 1/\beta)L]p_{t+1} = (\alpha - m_t)/\beta$ so that lagging by one period results in

$[1 - (1 + 1/\beta)L]p_t = (\alpha - m_{t-1})/\beta$

Since β is assumed to be positive, the expression $(1 + 1/\beta)$ is greater than unity. Hence, the backward-looking solution for p_t is divergent.

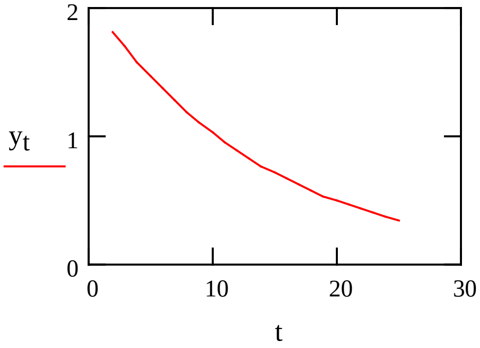
Section 1.4: Practice in Finding Homogeneous Solutions

Example 1: The AR(2) case: $y_t = a_0 + 0.5y_{t-1} + 0.4y_{t-2}$

Try $y_t = A_0r^t$ as the homogeneous solution. Hence, substitute $y_t = A_0r^t$ into $y_t - 0.5y_{t-1} - 0.4y_{t-2} = 0$ to obtain

$$A_0r^t - 0.5A_0r^{t-1} - 0.4A_0r^{t-2} = 0.$$

There are two solutions for r: $r_1 = -0.43$ and $r_2 = 0.93$. Given the initial conditions, $y_{t-2} = 0$ and $y_{t-1} = 2$, the time path of the series is shown in the figure below.

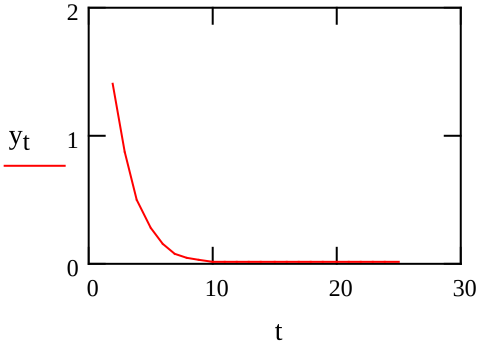


Example 2: Another AR(2) model: $y_t = a_0 + 0.9y_{t-1} - 0.2y_{t-2}$

Again, try $y_t = A_0r^t$ for the solution to the homogeneous part of the equation. Substitute $y_t = A_0r^t$ into $y_t - 0.9y_{t-1} + 0.2y_{t-2} = 0$ to obtain

$$A_0r^t - 0.9A_0r^{t-1} + 0.2A_0r^{t-2} = 0$$

There are two solutions for r: $r_1 = 0.4$ and $r_2 = 0.5$. For the initial conditions given in exercise 1, the time path of the series is:



Example 3: A third AR(2) model: $y_t = .55y_{t-1} + 0.2y_{t-2}$

Form the homogeneous equation:

$$y_t - 0.55y_{t-1} - 0.2y_{t-2} = 0$$

After forming the homogenous equation we check the discriminant (*d*) to see if the roots will be real and distinct or, alternatively, imaginary. Using our definition of the discriminant, and **Table 1**, we find that $d = (0.55)^2 + 4(0.2) = 1.1025$. Thus we conclude that because *d* is greater than zero, the roots to this particular equation will be real and distinct.

Table 1: Discriminant = $d = a_1^2 + 4a_2$

$d > 0$	$d < 0$
Roots are real and distinct	Roots are imaginary

1. We know that the trial solution will have the form $y_t \;\; = \alpha^t$ and we use this information to obtain

$$\alpha^t - .55\alpha^{t-1} - .2\alpha^{t-2} = 0$$

2. By dividing by α^{t-2} we obtain the characteristic equation:

$$\alpha^2 - .55\alpha^t - .2\alpha \quad = 0$$

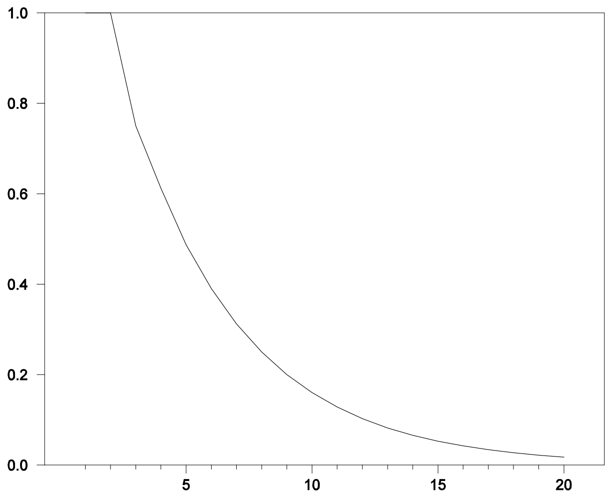
3. We can now compute the two characteristic roots:

$$\alpha_1 = 0.5\left(a_1 + d^{1/2}\right) = \ .8 \qquad\qquad \alpha_2 = 0.5\left(a_2 - d^{1/2}\right) = \ -.25$$

4. The last step is to write out the homogenous solution:

$$A_1(.8)^t + A_2(-.25)^t$$

The following graph shows the time path of this equation for the case in which the arbitrary constants equal unity and t runs from 1 to 20.



Example 5: An example with complex roots

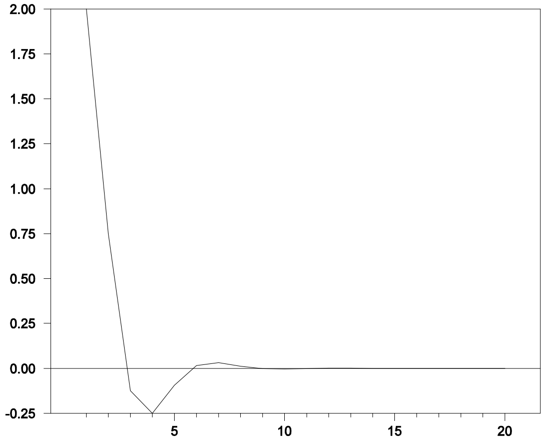
Let us analyze the homogenous solution to a second-order differential equation with complex roots and no initial conditions

$$y_t = -\frac{1}{2}y_{t-1} - \frac{1}{4}y_{t-2}$$

Calculating the discriminant (d) with $a_1 = \frac{1}{2}$ and $a_2 = \frac{1}{4}$ yields $d = -\frac{3}{4}$. This indicates that the characteristic roots to this difference equation will be complex. The homogenous solution to the difference equation will then have the form $y_t^h = \beta_1 r^t \cos(\theta t + \beta_2)$ where $r = \sqrt{(\frac{a_1}{2})^2 + (i \cdot \frac{d^{\frac{1}{2}}}{2})^2}$ and $\cos\theta = \frac{a_1}{2r}$. After solving for the values of r and θ we get $\frac{1}{2}$ and $\frac{\pi}{3}$, respectively. Therefore the homogenous solution is

$$y_t^h = \beta_1 \cdot \frac{1}{2^t}\cos(\frac{\pi}{3}t + \beta_2)$$

The following graph shows the time path of the above for the case in which the arbitrary constants equal unity and t runs from 1 to 20.



Backward Solution with Stochastic Term

Investigating difference equations with stochastic terms is very important in time-series. The stochastic terms are i.i.d and normally distributed $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$. Let us add a stochastic term, ε_t, to a Example 3 above. Consider:

$$y_t = 2 + 0.55y_{t-1} + 0.2y_{t-2} + \varepsilon_t$$

The solution to this second-order difference equation with a stochastic term takes the form

$$y_t = c + \sum_{i=0}^{\infty} c_i \cdot \varepsilon_{t-i}$$

where c and c_i are constants for all i. The question now becomes what are the values for these constants. To solve for these constants we will employ the ***method of undetermined coefficients***, which is tantamount to equating like terms (according to the stochastic term and its lags) on both sides of the equation and solving for the constant in question.

$$c + \ c_0\varepsilon_t + c_1\varepsilon_{t-1} + c_2\varepsilon_{t-2} + \cdots = 0.55[c + c_0\varepsilon_{t-1} + c_1\varepsilon_{t-2} + c_2\varepsilon_{t-3} + \cdots]$$

$$+0.2[[c + c_0\varepsilon_{t-2} + c_1\varepsilon_{t-3} + c_2\varepsilon_{t-4} \cdots] + \varepsilon_t + 2$$

Now we can start grouping according to the constants

$$c = 0.55c + 0.2c + 2$$
$$c = \frac{2}{1-0.55-0.2} = 8$$

This is the same solution if we were finding the particular solution for this difference equation

$$\lim_{t\to\infty} y_t = \bar{y}$$
$$\bar{y} = 0.55\bar{y} + 0.2\bar{y} + 2$$
$$\bar{y} = 8$$

The other constant terms can be found in the same manner as c was found

$$c_0\varepsilon_t = \ \varepsilon_t$$
$$c_0 = 1$$

$$c_1\varepsilon_{t-1} = 0.55c_0\varepsilon_{t-1}$$
$$c_1 = 0.55 \cdot 1 = 0.55$$

$$c_2\varepsilon_{t-2} = \ 0.55c_1\varepsilon_{t-2} + 0.2c_0\varepsilon_{t-2}$$
$$c_2 = 0.55 \cdot 0.55 + 0.2 \cdot 1 = \ 0.5025$$

$$c_3\varepsilon_{t-3} = 0.55c_2\varepsilon_{t-3} + 0.2c_1\varepsilon_{t-3}$$
$$c_3 = 0.55 \cdot 0.5025 + 0.2 \cdot 0.55 = \ 0.386375$$
$$\vdots$$
$$c_i\varepsilon_{t-i} = \ 0.55c_{i-1}\varepsilon_{t-i} + 0.2c_{i-2}\varepsilon_{t-i} \Rightarrow c_i = 0.55c_{i-1} + 0.2c_{i-2}$$

This last equation should look familiar. It is of the same form as our non-stochastic AR(2) model example. Therefore it should have the same form of homogenous solution as found in Example 3 above.

$$c_i = A_1(0.8)^i + A_2(-0.25)^i$$

A Forward-Looking Model with a Stochastic Term

Consider the model: $y_t = 2\, y_{t-1} + \varepsilon_t$.

It should be clear that the backward looking-solution is explosive. However, we can obtain the forward-looking solution as follows. Consider:

$y_{t-1} = 0.5y_t – 0.5\varepsilon_t$

and updating one period:

$y_t = 0.5y_{t+1} – 0.5\varepsilon_{t+1}$

Continuing to iterate forward:

$y_t = 0.5y_{t+1} – 0.5\varepsilon_{t+1} \ = 0.5[0.5y_{t+2} – 0.5\varepsilon_{t+2}]$

$= 0.25y_{t+2} – 0.25\varepsilon_{t+2} – 0.5\varepsilon_{t+1}$

You should be able to convince yourself that the continued forward iteration yield $(0.5)^i y_{t+i.}$ so that the coefficient on the “future” values of y_{t+i} converge to zero. This type of model is often used to model stock prices. Using a well known identity we have the following formula:

$P_t = \frac{E_t[P_{t+1}]}{1+r} + d_t$

where P_t is the market price of a stock in period t, d_t is the dividend, and r is the one-period interest rate. In other words the current price of a stock is equal to the expected price in the next period, discounted by the interest rate plus any current dividends. Let’s look at $P_t = \frac{1}{1+r}P_{t+1} + d_t$ more closely.

Again the backwards solution of $P_{t+1} = (1+r)P_t + (1+r)d_t$ makes no sense. What about the forward solution? Using the method of undetermined coefficients we have:

$$P_t = \sum c_i d_{t+i} = c_0 d_t + c_1 d_{t+1} + c_2 d_{t+2} + \cdots = \frac{1}{1+r}(c_0 d_{t+1} + c_1 d_{t+2} + c_2 d_{t+3} + \cdots)$$

With,

$$c_1 = \frac{c_0}{1+r}$$

$$c_2 = \frac{c_0}{(1+r)^2}$$

$$\vdots$$

$$\vdots$$

$$c_i = \frac{c_0}{(1+r)^i}$$

Therefore,

$P_t = \frac{1}{1+r} E_t[P_{t+1}] + d_t$, where $d_t = d_0 + \varepsilon_t$

Using our results from above, we can solve for the specific coefficient values using substitution. We know, $P_t = c_0 + c_1 d_t$, and therefore, $E[P_{t+1}] = E[c_0 + c_1 d_{t+1}]$

Using this equality we can now solve for the value of c_0 and c_1

$$c_0 + c_1 d_t = \frac{1}{1+r}(c_0 + c_1 d_0) + d_t$$

$$c_0 = \frac{c_0 + c_1 d_0}{1+r} = c_0 = \frac{1}{1+r}(c_0 + d_0)$$

$$c_1 d_t = d_t \rightarrow 1 = c_1$$

$c_0 = \frac{1}{r} d_0 \approx$ the value of a perpetuity

Combining terms we are left with a final solution of:

$$P_t = \frac{d_0}{r} + d_t$$

Hence, the market price of the stock is equal to the current dividend plus the present discounted value of the dividend stream.

CHAPTER 2

Endnotes to Chapter 2

1. Often, the variance is estimated as $\hat{\sigma}^2 = (T-1)^{-1}\sum(y_t - \overline{y})^2$.

2. As discussed in Appendix 2.1, the estimation of lagged MA coefficients does not entail a loss of any usable observations. Hence, the two models are estimated over the same sample period.

3. Some software programs report the Durbin-Watson test statistic as a check for first-order serial correlation. This well-known test statistic is biased toward finding no serial correlation in the presence of lagged dependent variables. Hence, it is usually not used in ARMA models.

4. Some researchers prefer to drop the first observation when adding an additional observation. As such, the model is always estimated using a fixed number of observations. A potential advantage of this "rolling window" method is that structural change occurring early sample will not affect all of the forecasts. Of course, the disadvantage is that some of the data is not used when estimating the model.

5. The details of the X-11 and X-12 procedures are not important for our purposes. Respectively, the technical details along with several versions of the seasonal adjustment procedures can be downloaded from the Bureau of the Census Web page: www.census.gov/srd/www/x12a/ and www.census.gov/srd/www/x13as/.

6. As formulated, the test can also detect a break in the variance of the error process. Estimation of an AR(p) model usually entails a loss of the number of usable observations. Hence, to estimate a model using T usable observations it will be necessary to have a total of ($T + p$) observations. Also note that the procedure outlined necessitates that the second subsample period incorporate the lagged values t_m, t_{m-1}, … t_{m-p+1}.

SECTION 2.1: Appendix 2.1

ESTIMATION OF AN MA(1) PROCESS

How do you estimate an MA or an ARMA process? When you estimate a regression using ordinary least squares (OLS), you have a dependent variable and a set of independent variables. In an AR model, the list of regressors is simply the lagged values of the $\{y_t\}$ series. Estimating an MA process is different because you do not know the values of the $\{\varepsilon_t\}$ sequence. Since you cannot directly estimate a regression equation, maximum-likelihood estimation is used. Suppose that $\{\varepsilon_t\}$ is a white-noise sequence drawn from a normal distribution. The likelihood of any realization ε_t is

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\varepsilon_t^2}{2\sigma^2}\right)$$

Since the ε_t are independent, the likelihood of the joint realizations $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T$ is

$$\prod_{t=1}^{T}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\varepsilon_t^2}{2\sigma^2}\right)$$

If you take the log of the likelihood, you obtain

$$\ln L = \frac{-T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2$$

Now suppose that we observe T values of the MA(1) series $y_t = \beta\varepsilon_{t-1} + \varepsilon_t$. The problem is to construct the $\{\varepsilon_t\}$ sequence from the observed values of $\{y_t\}$. If we knew the true value of β and knew that $\varepsilon_0 = 0$, we could construct $\varepsilon_1, \dots, \varepsilon_T$ recursively. Given that $\varepsilon_0 = 0$, it follows that

$$\varepsilon_1 = y_1$$
$$\varepsilon_2 = y_2 - \beta\varepsilon_1 = y_2 - \beta y_1$$
$$\varepsilon_3 = y_3 - \beta\varepsilon_2 = y_3 - \beta(y_2 - \beta y_1)$$
$$\varepsilon_4 = y_4 - \beta\varepsilon_3 = y_4 - \beta[y_3 - \beta(y_2 - \beta y_1)]$$

In general, $\varepsilon_t = y_t - \beta\varepsilon_{t-1}$ so that if L is the lag operator

$$\varepsilon_t = y_t/(1+\beta L) = \sum_{i=0}^{t-1}(-\beta)^i y_{t-i}$$

As long as $|\beta| < 1$, the values of ε_t will represent a convergent process. This is the justification for the assumption that the MA process be invertible. If $|\beta| > 1$, we cannot represent the $\{\varepsilon_t\}$ series in terms of the observed $\{y_t\}$ series. If we now substitute the solution for ε_t into the formula for the log likelihood, we obtain

$$\ln L = \frac{-T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(\sum_{i=0}^{t-1}(-\beta)^i y_{t-i}\right)^2$$

Now that we have expressed ln L in terms of the observable $\{y_t\}$ series, it is possible to select the values of β and σ^2 that maximize the value of ln L. Unlike OLS, if you actually take the partial derivatives of ln L with respect to β and σ^2 you will not obtain a simple set of first-order conditions. Moreover, the formula becomes much more complicated in higher-order MA(q) processes. Nevertheless, computers can use a number of available search algorithms to find the values of β and σ^2 that maximize ln L. As indicated in the text, numerical optimization routines cannot guarantee exact solutions for the estimated coefficients. Instead, various “hill-climbing” methods are used to find the parameter values that maximize ln L. If the partial derivatives of the likelihood function are close to zero (so that the likelihood function is flat), the algorithms may not be able to find a maximum.

SECTION 2.2: Appendix 2.2
Model Selection Criteria

Hypothesis testing is not particularly well suited to testing nonnested models. For example, if you wanted to chose between an AR(1) and an MA(2) you could estimate and ARMA(1, 2) and then try to restrict the MA(2) coefficients to equal zero. Alternatively, you could try to restrict the AR(1) coefficient to equal zero. Nevertheless, the method is unsatisfactory because it necessitates estimating the overparameterized ARMA(1, 2) model. Instead, model selection criteria, such as the AIC and the SBC, can be used to choose between alternative models. Such model selection criteria can be viewed as measures of *goodness-of-fit* that include a cost, or penalty, for each parameter estimated.

One reason it is not desirable to have an overparameterized model is that forecast error variance increases as a result of errors arising from parameter estimation. In other words, small models tend have better out-of-sample performance than large models. Suppose that the actual data-generating process (DGP) is the AR(1) model:

$$y_t = ay_{t-1} + \varepsilon_t$$

If a is known, the one-step-ahead forecast of y_{t+1} is $E_t y_{t+1} = ay_t$. Hence, the mean squared forecast error is $E_t(y_{t+1} - ay_t)^2 = E_t \varepsilon_{t+1}^2 = \sigma^2$. However, when a is estimated from the data, the one-step-ahead forecast of y_{t+1} is:

$$E_t y_{t+1} = \hat{a} y_t$$

where $\hat{a}$ is the estimated value of a.

Hence, the mean squared forecast, or prediction, error is:

$$\text{MSPE} = E_t(y_{t+1} - \hat{a} y_t)^2 = E_t[\ (ay_t - \hat{a} y_t) + \varepsilon_{t+1}\ ]^2$$

Since ε_{t+1} is independent of $\hat{a}$ and y_t, it follows that:

$$\begin{aligned} E_t(y_{t+1} - \hat{a} y_t)^2 &= E_t[(\ a - \hat{a}\ )y_t]^2 + \sigma^2 \\ &\approx E_t[(\ a - \hat{a}\ )]^2\ (y_t)^2 + \sigma^2 \end{aligned}$$

Since $E_t[(\ a - \hat{a}\ )]^2$ is strictly positive, parameter uncertainty contributes to forecast error variance in that the mean squared forecast error exceeds σ^2. The point is that errors in parameter estimation contribute to forecast error variance. Moreover, the more parameters estimated, the greater the parameter uncertainty. It is easy to show that the problem is particularly acute in small samples. Since $\text{var}(y_t) = \sigma^2 / (1 - a^2)$ and, in large samples, $\text{var}(\hat{a}) = E_t[(\ a - \hat{a}\ )]^2 \approx (1 - a^2)/T$, it follows that

$$E_t[(a - \hat{a})]^2 (y_t)^2 + \sigma^2 \approx [(1 - a^2)/T](1 - a^2)^{-1} \sigma^2 + \sigma^2$$
$$= [1 + (1/T)]\sigma^2$$

Thus, as T increases, the MSPE approaches σ^2.

The Finite Prediction Error (FPE) Criterion

The FPE criterion seeks to minimize the one-step ahead mean squared prediction error. Now consider the AR(p) process:

$$y_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + \varepsilon_t$$

If you use the argument in the previous section, the MSPE can be shown to be equal to

$$[1 + (p/T)]\sigma^2$$

We do not know the true variance σ^2. However, σ^2 can be replaced by its unbiased estimate $SSR/(T - p)$ to get

$$\text{FPE} = [1 + (p/T)][SSR/(T\text{-}p)]$$

Select p so as to minimize FPE. We can use logs and note that $\ln(1 + p/T)$ can be approximated by p/T. Hence, it is possible to select p to minimize

$$p/T + \ln(SSR) - \ln(T - p)$$

which is the same as minimizing

$$p + T\ln(SSR) - T\ln(T - p)$$

Since $\ln(T - p) \cong \ln T - p/T$, the problem can be written in terms of selecting p so as to minimize

$$p + T\ln(SSR) - \ln(T) + p$$

which has the same solution as minimizing

$$T\ln(SSR) + 2p$$

The AIC and the SBC

The more general AIC selects the $(1 + p + q)$ parameters of an ARMA model so as to maximize the log likelihood function including a penalty for each parameter estimated:

$$\text{AIC} = -2 \text{ ln maximized value of log likelihood} + (1 + p + q)/T$$

For a given value of *T*, selecting the values of *p* and *q* so as to minimize the AIC is equivalent to selecting *p* and *q* so as to minimize the sum:

$$T \ln (SSR) + 2(1 + p + q)$$

Notice that is $q = 0$ and there is no intercept, this is the result obtained using the FPE. Minimizing the value of the AIC implies that each estimated parameter entails a benefit and a cost. Clearly, a benefit of adding another parameter is that the value of *SSR* is reduced. The cost is that degrees of freedom are reduced and there is added parameter uncertainty. Thus, adding additional parameters will decrease ln (*SSR*) but will increase $(1 + p + q)$. The AIC allows you to add parameters until the marginal cost (i.e., the marginal cost is 2 for each parameter estimated) equals the marginal benefit. The SBC incorporates the larger penalty $(1 + p + q) \ln T$. To use the SBC, select the values of *p* and *q* so as to minimize

$$T \ln (SSR) + (1 + p + q) \ln(T)$$

For any reasonable sample size, $\ln(T) > 2$ so that the marginal cost of adding parameters using the SBC exceeds that of the AIC. Hence, the SBC will select a more parsimonious model than the AIC. As indicated in the text, the SBC has superior large sample properties. It is possible to prove that the SBC is asymptotically consistent while the AIC is biased toward selecting an overparameterized model. However, Monte Carlo studies have shown that in small samples, the AIC can work better than the SBC.

Section 2.3: Review of Expected Values and Variance

This material is key to understanding the material in Chapter 2

1. **Expected value of a discrete random variable**.

A random variable x is defined to be discrete if the range of x is countable. If x is discrete, there is a finite set of numbers x_1, x_2...x_n such that x takes on values only in that set. Let $f(x_j)$ = the probability that $x=x_j$. The mean or **expected value** of x is defined to be:

$$E(x) = \sum_{j=1}^{n} x_j f(x_j)$$

Note:

1. We can let *n* go to infinity; the notion of a discrete variable is that the set be "denumerable" or a countable infinity. For example, the set of all positive integers is discrete.

2. If $\Sigma x_j f(x_j)$ does not converge, the mean is said not to exist.

3. $E(x)$ is an "average" of the possible values of x; in the sum, each possible value of x_j is weighted by the probability that $x = x_j$; i.e.,

$$E(x) = w_1x_1 + w_2x_2 + ... + w_nx_n \quad \text{where } \Sigma w_j = 1$$

2. **Expected value of a continuous random variable**.

Now let x be a continuous random variable. Denote the probability that x is in the interval (x_0, x_1) be denoted by $f(x_0 \le x \le x_1)$. It follows that:

$$f(x_0 \le x \le x_1) = \int_{x_0}^{x_1} f(x)dx$$

The mean, or expected value, of x is:

$$E(x) = \sum_{-\infty}^{\infty} xf(x)dx$$

3. **Expected value of a function**.

Let x be a random variable and let $g(x)$ be a function. The mean or expected value of $g(x)$ is:

$$E[g(x)] = \sum_{j=1}^{n} g(x_j) f(x_j) \text{ for discrete } x$$

or

$$E[g(x)] = \int_{-\infty}^{\infty} g(x) f(x)dx \text{ for continuous x.}$$

Note: if $g(x_j) \equiv x_j$, we obtain the simple mean.

4. **Properties of the expectations operator**:

a. The expected value of a constant c is the value of the constant: i.e., $E[c] = c$.

Proof: Since we can let $c = g(x)$,

$$E(c) = \int_{-\infty}^{\infty} cf(x)dx = c\int_{-\infty}^{\infty} f(x)dx = c$$

b. The expected value of a constant times a function is the constant times the expected value of the function:

Proof: $E[cg(x)] =$ $E[cg(x)] = \int_{-\infty}^{\infty} cg(x)f(x)dx = c\int_{-\infty}^{\infty} g(x)f(x)dx = cE[g(x)]$

c. The expected value of a sum is the sum of the expectations:

$$E[c_1g_1(x) \pm c_2g_2(x)] = c_1Eg_1(x) \pm c_2Eg_2(x)$$

Proof:

$$\int_{-\infty}^{\infty}[c_1g_1(x) \pm c_2g_2(x)]f(x)dx = \int_{-\infty}^{\infty} c_1g_1(x)f(x)dx \pm \int_{-\infty}^{\infty} c_2g_2(x)f(x)dx$$

$$= c_1Eg_1(x) \pm c_2Eg_2(x)$$

5. **The Variance of a Random Variable**:

The variance of x is defined such that $\text{var}(x) = E\{[x - E(x)]^2\}$ so that:

$$\text{var}(x) = E\{x^2 - 2x\,E(x) + E(x)\,E(x)\}$$

Since $E(x)$ is a constant, $E[E(x)] = E(x)$ and $E[xE(x)] = [E(x)]^2$. Using these results and the property that expectation of a sum is the sum of the expectations:

$$\text{var}(x) = E(x^2) - 2E\{xE(x)\} + E(x)^2$$

$$= E(x^2) - [E(x)]^2$$

6. Jointly Distributed Discrete Random Variables

Let x and y be random variables such that x takes on values $x_1, x_2, \ldots, x_n$ and y takes on values $y_1, y_2, \ldots, y_m$. Also let f_{ij} denote the probability that $x = x_i$ **and** $y = y_j$. If $g(x, y)$ denotes a function of x and y, the expected value of the function is:

$$E[g(x,y)] = \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}\,g(x_i, y_j)$$

Expected value of a sum: Let the function $g(x, y)$ be $x + y$. The expected value of $x + y$ is:

$$E(x+y)=\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}(x_i+y_j)$$

$$E(x+y)=\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}x_i+\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}y_j$$

$$=\sum_{j=1}^{m}(f_{1j}x_1+f_{2j}x_2+\ldots+f_{nj}x_n)+\sum_{i=1}^{n}(f_{i1}y_1+f_{i2}y_2+\ldots+f_{im}y_m)$$

Note that ($f_{11}+f_{12}+f_{13}+\ldots+f_{1m}$) is the probability that x takes on the value x_1 denoted by f_1. More generally, ($f_{i1}+f_{i2}+f_{i3}+\ldots+f_{im}$) is the probability that x takes on the value x_i denoted by f_i or $f(x_i)$. Since ($f_{1i}+f_{2i}+f_{3i}+\ldots+f_{ni}$) is the probability that $y = y_i$ [denoted by $f(y_i)$], the two summations above can be written as:

$$E[x+y]=\Sigma x_i f(x_i)+\Sigma y_i f(y_i)$$

$$=E(x)+E(y)$$

Hence, we have generalized the result of 4c above to show that the expected value of a sum is the sum of the expectations.

7. Covariance and Correlation

The covariance between x and y, denoted by cov(x, y)––is defined to be:

$$\text{cov}(x, y)=E\{[x-E(x)]\,[y-E(y)]\}\equiv\sigma_{xy}$$

Multiply [$x - E(x)$] by [$y - E(y)$] and use the property that the expected value of a sum is the sum of the expectations:

$$\text{cov}(x, y)=E[x\,y]-E[x\,E(y)]-E[y\,E(x)]+E\,[E(x)\,E(y)]$$

$$=E(x\,y)-E(x)\,E(y)$$

The **correlation coefficient** between x and y is defined to be:

$$\rho_{\text{xy}}=\text{cov}(x, y)/[\text{var}(x)\ \text{var}(y)]^{1/2}$$

Since $\text{cov}(x, y) = E(xy) - E(x)E(y)$, we can express the expectation of the product of x and y--$E(xy)$--as:

$$E(xy)=E(x)E(y)+\text{cov}(x, y)$$

$$=E(x)E(y)+\rho_{xy}\ \sigma_x\sigma_y$$

where the standard deviation of variable *z* (denoted by σ_z) is the positive square root of *z*.

8. Conditional Expectation

Let *x* and *y* be jointly distributed random variables where f_{ij} denotes the probability that $x = x_i$ **and** $y = y_j$. Each of the f_{ij} values is a **conditional probability**; each is the probability that *x* takes on the value x_i *given* that *y* takes on the specific value y_j.

The expected value of *x* conditional on *y* taking on the value y_j is:

$$E[\ x \mid y_j\ ] = f_{1j}x_1 + f_{2j}x_2 + \ldots + f_{nj}x_n$$

9. Statistical Independence

If *x* and *y* are **statistically independent**, the probability of $x = x_i$ *and* $y = y_j$ is the probability that $x = x_i$ multiplied by the probability that $y = y_j$: using the notation in section 6, *two events are statistically independent if and only if* $f_{ij} = f(x_i)f(y_j)$. For example, if we simultaneously toss a fair coin and roll a fair die, the probability of obtaining a head *and* a three is 1/12; the probability of a head is 1/2 and the probability of obtaining a three is 1/6.

An extremely important implication follows directly from this definition. If x and y are independent events, the expected value of the product of the outcomes is the product of the expected outcomes:

$$E[\ x\,y\ ] = E(x)E(y).$$

The proof is straightforward. Form $E[\ x\,y\ ]$ as:

$$E[x\,y] = f_{11}x_1y_1 + f_{12}x_1y_2 + f_{13}x_1y_3 + ... + f_{1m}x_1y_m + f_{21}x_2y_1 + f_{22}x_2y_2 + f_{23}x_2y_3 + ... + f_{2m}x_1y_m$$
$$+ .... + f_{n1}x_ny_1 + f_{n2}x_ny_2 + f_{n3}x_ny_3 + ... + f_{nm}x_ny_m$$

or more compactly:

Since *x* and *y* are independent, $f_{ij} = f(x_i)f(y_j)$ so that:

$$E[xy] = \sum_{i=1}^{n} f(x_i)f(y_1)x_iy_1 + \sum_{i=1}^{n} f(x_i)f(y_2)x_iy_2 + ... + \sum_{i=1}^{n} f(x_i)f(y_m)x_iy_m$$

Recall $\Sigma f(x_i)x_i = E(x)$:

$$E[xy] = E(x)[f(y_1)y_1 + f(y_2)y_2 + ... + f(y_m)y_m]$$

so that $E[x\,y] = E(x)E(y)$.

Since $\text{cov}(x, y) = E(x\,y) - E(x)E(y)$, it immediately follows that the covariance and correlation coefficient of two independent events is zero.

10. An Example of Conditional Expectation

Since the concept of conditional expectation plays such an important role in time-series econometrics, it is worthwhile to consider the specific example of tossing dice. Let x denote the number of spots showing on die 1, y the number of spots on die 2, and S the sum of the spots ($S = x + y$). Each die is fair so that the probability of any face turning up is 1/6. Since the outcome on die 1 and die 2 are independent events, the probability of any specific values for x and y is the product of the probabilities. The possible outcomes and the probability associated with each outcome S are:

S	2	3	4	5	6	7	8	9	10	11	12
$f(S)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

To find the expected value of the sum S, multiply each possible outcome by the probability associated with that outcome. As you well know if you have been to Las Vegas, the expected value is 7. Suppose that you roll the dice sequentially and that the first roll turns up 3 spots. What is the expected value of the sum given that $x = 3$? We know that y can take on values 1 through 6 each with a probability of 1/6. Given $x = 3$, the possible outcomes for S are 4 through 9 each with a probability of 1/6. Hence, the conditional probability of S given three spots on die 1 is:

$$E[\, S \mid x = 3] = (1/6)4 + (1/6)5 + (1/6)6 + (1/6)7 + (1/6)8 + (1/6)9 = 6.5$$

11. Testing the significance of ρ_i

Under the null hypothesis of $\rho_i = 0$, the sample distribution of $\hat{\rho}$ is:

a. approximately **normal** (but bounded at -1.0 and +1.0) when T is **large**

b. distributed as a student**s-*t*** when T is **small**.

The standard formula for computing the appropriate ***t* value** to test significance of a correlation coefficient is:

$$t = \hat{\rho}_i \sqrt{\frac{T-2}{1-\hat{\rho}_i^2}} \quad \text{with df} = T - 2$$

In reasonably large samples, the test for the null that $\rho_i = 0$ is simplified to $\hat{\rho}_i\ T^{1/2}$. Alternatively, the standard deviation of the correlation coefficient is $(1/T)^{0.5}$.

Section 2.4: Improving Your Forecasts and the Presentation of Your Results

1. It is important for you and your reader to know the type of data you are using. There are many ways to measure certain variables. Stock prices may be opening, closing, or daily average values. Unemployment may or may not be seasonally adjusted. The point is that it is necessary to tell your reader what data you are using and where it comes from.

2. Looking at the time path of a series is the single most important step in forecasting the series. Examining the series allows you to see if it has a clear trend and to get a reasonable idea if the trend is linear or nonlinear. Similarly, a series may or may not have periods of 'excess' volatility. Graphs should be properly labeled and dates on the 'time' axis should be clear.

3. There usually are several plausible models that confirm to the data. Such models should be compared as to their in-sample fit and their forecasts.

4. It is standard to plot the forecasts in the same graph as the series being forecasted. Sometimes it is desirable to place confidence intervals around the forecasted values. If you chose a transformation of the series [e.g., log(x)] you should forecast the values of the series, not the transformed values.

5. The steps in the Box-Jenkins methodology entail:

Identification
Graph the data–see (2) above–in order to determine if any transformations are necessary (logarithms, differencing, ...). Also look for outliers, missing values, breaks, …

Nonstationary variables may have a pronounced trend or appear to meander without a constant mean and/or variance.

Examine the ACF and the PACF of the transformed data and compare them to the theoretical ACF and PACF of ARMA processes.

Estimation
Estimate the plausible models and select the best. In this second stage, the goal is to select a parsomonous model. You should entertain the possibility of several models and estimate each. Forecasters do not want to project poorly estimated coefficients into the future, The aim is to approximate the DGP but not to pind down the exact process.

The 'best' will have coefficients that are statistically signifcant and a good fit. (use the AIC or SBC to determine the fit).

Several different models will have similar properties. As as extreme example, not that an AR(1) model has an infinite order representation.

Be aware of the common factor problem.

Stationarity: The distribution theory underlying the procedure assumes that the $\{y_t\}$ series is stationary. The estimated AR coefficients should inply stationarity. The MA coreeicients should imply invertibility.

Diagnostic Checking

The residuals of a properly estimated model cannot contain any significant autocorrelations. Examine the ACF and PACF of the residuals to check for significant autocorrelations. Use the *Q*-statistics to determine if groups of autocorrelations are statistically significant.

Other diagnostic checks include splitting the sample, and overfitting (adding a lagged value that should be insignificant). Be sure to check for coefficient instability. Check to see that the variance of the residuals is constant.

Forecasting

Forecast using several plausible modes. Compare the out-of-sample forecast accuracy of the alternatives.

Section 2.5: Heteroskedasticity-Autocorrelation-Consistent (HAC) Estimators

Within the framework of the **Distributed Lag Model Assumption**, ordinary least squares yields consistent estimators and a normal sampling distribution of the estimators. Unfortunately, the variance of the sampling distribution suffers from autocorrelation and therefore OLS standard errors are wrong. The solution to this problem rests in standard errors that are robust to autocorrelation as well as heteroskedasticity. Let us return to a no lag framework. Our model takes the form y_t = βx_t + ε_t . Consider the OLS estimator for β_1:

$$\hat{\beta} = \sum_{t=1}^{T} x_t y_t \Big/ \sum_{t=1}^{T} x_t^2$$

The difference between the estimated value and the actual value of β is:

$$\hat{\beta} - \beta = \sum_{t=1}^{T} x_t \varepsilon_t \Big/ \sum_{t=1}^{T} x_t^2$$

The the sample size be large so that $\sum x_t^2 \to \sigma_x^2$ and define v_t = $\sum x_t \varepsilon_t$. As such

$$\hat{\beta} - \beta = \left(\sum_{t=1}^{T} v_t / T \right) \Big/ \sigma_x^2$$

Given that $\hat{\beta}$ is unbiased, we can take the variance of each side:

$$\operatorname{var}(\hat{\beta}) = \frac{1}{\sigma_x^4} \left[\frac{1}{T^2} \right] \operatorname{var}\left(\sum_{t=1}^{T} v_t \right)$$

Note that we can construct w_T such that

$$\operatorname{var}\left(\sum_{t=1}^{T} v_t / T \right) = w_T \sigma_v^2 / T$$

where $w_T = 1 + 2\sum_{j=1}^{T-1} \left(\frac{T-j}{T} \right) \rho_j$ and ρ_j is the j-th autocorrelation coefficient.

Hence

$$\operatorname{var}(\hat{\beta}) = \left[\frac{1}{T} \right] \frac{\sigma_v^2}{\sigma_x^4} w_T$$

The key to creating standard errors that are robust to autocorrelation as well as heteroskedasticity is finding the appropriate estimates of the weights, w_T. It is not possible to find the actual weights since these weights depend upon unknown autocorrelations. In essence, the Heteroskedasticity Autocorrelation Consistent Estimator (HAC) finds these appropriate estimates of the weights.

The most commonly used weight estimates are sometimes referred to as the ‘Newey-West” weights:

$$w_T^* = 1 + 2\sum_{j=1}^{m-1}(\frac{m-j}{m})\,\tilde{\rho}_j$$

Where $\tilde{\rho}_j$ is an estimator of ρ_j and m is called the truncation parameter which is left up to the practitioner to choose its magnitude.

CHAPTER 3

ENDNOTES TO CHAPTER 3

1. Letting $\alpha(L)$ and $\beta(L)$ be polynomials in the lag operator L, we can rewrite h_t in the for $h_t = \alpha_0 + \alpha(L)\varepsilon_t^2 + \beta(L)h_t$. The notation $\alpha(1)$ denotes the polynomial $\alpha(L)$ evaluated at $L = 1$; i.e., $\alpha(1) = \alpha_1 + \alpha_2 + \ldots + \alpha_q$. Bollerslev (1986) showed that the GARCH process is stationary with $E\varepsilon_t = 0$, $\text{var}(\varepsilon_t) = \alpha_0/(1 - \alpha(1) - \beta(1))$, and $\text{cov}(\varepsilon_t, \varepsilon_{t-s}) = 0$ for $s \neq 0$ if $\alpha(1) + \beta(1) < 1$.

2. From introductory statistics, if a correlation coefficient ρ_i is formed from T observations, then under the null hypothesis $\rho_i = 0$, in very large samples, the value $\rho_i/[(1 - \rho_i^2)/(T - 2)]^{0.5}$ has t-distribution with $T - 2$ degrees of freedom. The standard deviation is $[(1 - \rho_i^2)/(T - 2)]^{-0.5}$. For reasonably small values of ρ_i (so that ρ_i^2 is very close to zero) and with the sample sizes used in time series analysis (so that $T - 2 \cong T$), the standard deviation of ρ_i is approximately $T^{-0.5}$.

3. Unfortunately, there is no available method to test the null of white-noise errors versus the specific alternative of GARCH(p, q) errors. As indicated in Question 3, Bollerslev (1986) proved that the ACF of the squared residuals resulting from (3.9) are an ARMA(m, p) model where $m = \max(p, q)$.

4. The unconditional mean of y_t is altered by changing only the value δ. Changing β and δ commensurately maintains the mean value of the $\{y_t\}$ sequence.

5. If you are not particularly interested in the tails of the distribution, in large samples, it is reasonable to ignore the issue of a fat-tailed distribution. Quasi-maximum likelihood estimates use the normal distribution even though the actual distribution of the $\{\varepsilon_t\}$ sequence is fat-tailed. Under fairly weak assumptions, the parameter estimates for the model of the mean and the conditional variance are consistent and normally distributed.

6. In constructing the data set, no attempt was made to account for the fact that the market was closed on holidays and on important key dates such as September 11, 2001. For simplicity, we interpolated to obtain values for non-weekend dates when the market was closed.

SECTION 3.1

APPENDIX 3.1 MULTIVARIATE GARCH MODELS

The Log-Likelihood Function

In the multivariate case, the likelihood function presented in Section 8 needs to be modified. For the 2-variable case, suppose that ε_{1t} and ε_{2t} are zero-mean random variables that are jointly normally distributed. For the time being, we can keep the analysis simple if we assume the variances and the covariance terms are constant. As such, we can drop the time subscripts on the h_{ijt}. In such a circumstance, the log-likelihood function for the joint realization of ε_{1t} and ε_{2t} is

$$L_t = \frac{1}{2\pi\sqrt{h_{11}h_{22}(1-\rho_{12}^2)}}\exp\left[-\frac{1}{2(1-\rho_{12}^2)}\left(\frac{\varepsilon_{1t}^2}{h_{11}}+\frac{\varepsilon_{2t}^2}{h_{22}}-\frac{2\rho_{12}\varepsilon_{1t}\varepsilon_{2t}}{(h_{11}h_{22})^{0.5}}\right)\right] \qquad \text{(A3.1)}$$

where ρ_{12} is the correlation coefficient between ε_{1t} and ε_{2t}; $\rho_{12} = h_{12}/(h_{11}h_{22})^{0.5}$.

Now if we define the matrix H such that

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix}$$

the likelihood function can be written in the compact form

$$L_t = \frac{1}{2\pi|H|^{1/2}}\exp\left[-\frac{1}{2}\varepsilon_t' H^{-1}\varepsilon_t\right] \qquad \text{(A3.2)}$$

where $\varepsilon_t = (\ \varepsilon_{1t},\ \varepsilon_{2t}\ )'$, and $|\,H\,|$ is the determinant of H. To see that the two representations given by (A3.1) and (A3.2) are equivalent, note that $|\,H\,| = h_{11}h_{22} - (h_{12})^2$. Since $h_{12} = \rho_{12}(h_{11}h_{22})^{0.5}$, it follows that $|\,H\,| = (1 - (\rho_{12})^2)h_{11}h_{22}$. Moreover,

$$\varepsilon_t' H^{-1}\varepsilon_t = \frac{\varepsilon_{1t}^2 h_{22} - 2\varepsilon_{1t}\varepsilon_{2t}h_{12} + \varepsilon_{2t}^2 h_{11}}{h_{11}h_{22} - h_{12}^2}$$

Since $h_{12} = \rho_{12}(h_{11}h_{22})^{0.5}$,

$$\varepsilon_t' H^{-1}\varepsilon_t = \left[\frac{1}{(1-\rho_{12}^2)}\left(\frac{\varepsilon_{1t}^2}{h_{11}}+\frac{\varepsilon_{2t}^2}{h_{22}}-\frac{2\rho_{12}\varepsilon_{1t}\varepsilon_{2t}}{(h_{11}h_{22})^{0.5}}\right)\right]$$

Now, suppose that the realizations of $\{\varepsilon_t\}$ are independent, so that the likelihood of the joint realizations of ε_1, ε_2, … ε_T is the product in the individual likelihoods. Hence, if all have the same variance, the conditional likelihood of the joint realizations is

$$L = \prod_{t=1}^{T}\frac{1}{2\pi|H|^{1/2}}\exp\left[-\frac{1}{2}\varepsilon_t' H^{-1}\varepsilon_t\right]$$

It is far easier to work with a sum than with a product. As such, it is convenient to take the natural log of each side so as to obtain

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln|H| - \frac{1}{2}\sum_{t=1}^{T}\varepsilon_t' H^{-1}\varepsilon_t$$

The procedure used in maximum-likelihood estimation is to select the distributional parameters so as to maximize the likelihood of drawing the observed sample. Given the realizations in ε_t, it is possible to select h_{11}, h_{12}, and h_{22} so as to maximize the likelihood function.

For our purposes, we want allow the values of h_{ij} to be time-varying. If you worked through Section 8, it should be clear how to modify this equation if h_{11}, h_{22}, and h_{12} are time varying. Consider

$$L = \prod_{t=1}^{T}\frac{1}{2\pi|H_t|^{1/2}}\exp\left[-\frac{1}{2}\varepsilon_t' H_t^{-1}\varepsilon_t\right]$$

where

$$H_t = \begin{bmatrix} h_{11t} & h_{12t} \\ h_{12t} & h_{22t} \end{bmatrix}$$

Now, if we take the log of the likelihood function,

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(\ln|H_t| + \varepsilon_t' H_t^{-1}\varepsilon_t) \qquad \text{(A3.3)}$$

The convenience of working with (A3.2) and (A3.3) is that the form of the likelihood function is identical for models with k variables. In such circumstances, H is a symmetric k x k matrix, ε_t is a k x 1 column vector, and the constant term (2π) is raised to the power k.

Multivariate GARCH Specifications

Given the log-likelihood function given by (A3.3), it is necessary to specify the functional forms for each of the h_{ijt}. The most familiar specifications are given below:

1. The vech Model: The *vech* operator transforms the upper (lower) triangle of a symmetric matrix into a column vector. Consider the symmetric covariance matrix

$$H_t = \begin{bmatrix} h_{11t} & h_{12t} \\ h_{12t} & h_{22t} \end{bmatrix}$$

so that

$$vech(H_t) = [\ h_{11t}, h_{12t}, h_{22t}\ ]'$$

Now consider the vector $\varepsilon_t = [\varepsilon_{1t}, \varepsilon_{2t}]'$. The product $\varepsilon_t\varepsilon_t' = [\varepsilon_{1t}, \varepsilon_{2t}]'[\varepsilon_{1t}, \varepsilon_{2t}]$ is the 2 x 2 matrix

$$\begin{bmatrix} \varepsilon_{1t}^2 & \varepsilon_{1t}\varepsilon_{2t} \\ \varepsilon_{1t}\varepsilon_{2t} & \varepsilon_{1t}^2 \end{bmatrix}$$

Hence, $vech(\varepsilon_t\varepsilon_t')$ = $\left[\varepsilon_{1t}^2, \varepsilon_{1t}\varepsilon_{2t}, \varepsilon_{2t}^2\right]'$. If we now let C = [c_1, c_2, c_3]$'$, A = the 3 x 3 matrix with elements α_{ij}, and B = the 3 x 3 matrix with elements β_{ij}, we can write

$$vech(H_t) = C + A\ vech(\varepsilon_{t-1}\varepsilon_{t-1}') + B\ vech(H_{t-1})$$

If you are familiar with matrix operations, it should be clear that this is precisely the system represented by equations (3.42) through (3.44). The diagonal *vech* uses only the diagonal elements of A and B and sets all values of $\alpha_{ij} = \beta_{ij} = 0$ for $i \neq j$.

2. The BEK Model: In a system with k variables, the BEK specification has the form

$$H_t = C'C + A'\varepsilon_{t-1}\varepsilon_{t-1}'A + B'H_{t-1}B$$

where A and B are k x k matrices. However, C must be a symmetric k x k matrix in order to ensure that the intercepts of the off-diagonal elements h_{ijt} are identical. As suggested in the text for the 2-variable case,

$$h_{11t} = (c_{11}^2 + c_{12}^2) + (\alpha_{11}^2\varepsilon_{1t-1}^2 + 2\alpha_{11}\alpha_{21}\varepsilon_{1t-1}\varepsilon_{2t-1} + \alpha_{21}^2\varepsilon_{2t-1}^2) + (\beta_{11}^2 h_{11t-1} + 2\beta_{11}\beta_{21}h_{12t-1} + \beta_{21}^2 h_{22t-1})$$

$$h_{12t} = c_{12}(c_{11} + c_{22}) + \alpha_{12}\alpha_{11}\varepsilon_{1t-1}^2 + (\alpha_{11}\alpha_{22} + \alpha_{12}\alpha_{21})\varepsilon_{1t-1}\varepsilon_{2t-1} + \alpha_{21}\alpha_{22}\varepsilon_{2t-1}^2$$

$$+\beta_{11}\beta_{12}h_{11t-1} + (\beta_{11}\beta_{22} + \beta_{12}\beta_{21})h_{12t-1} + \beta_{21}\beta_{22}h_{22t-1}$$

$$h_{22t} = (c_{22}^2 + c_{12}^2) + (\alpha_{12}^2\varepsilon_{1t-1}^2 + 2\alpha_{12}\alpha_{22}\varepsilon_{1t-1}\varepsilon_{2t-1} + \alpha_{22}^2\varepsilon_{2t-1}^2) + (\beta_{12}^2 h_{11t-1} + 2\beta_{12}\beta_{22}h_{12t-1} + \beta_{22}^2 h_{22t-1})$$

3. Constant Conditional Correlations: The CCC formulation is clearly a special case of the more general multivariate GARCH model. In the 2 variable case, we can write H_t as

$$H_t = \begin{bmatrix} h_{11t} & \rho_{12}(h_{11t}h_{22t})^{0.5} \\ \rho_{12}(h_{11t}h_{22t})^{0.5} & h_{22t} \end{bmatrix}$$

Now, if h_{11t} and h_{22t} are both GARCH(1, 1) processes, there are seven parameters to estimate (the six values of c_i, α_{ii} and β_{ii}, and ρ_{12}).

4. Dynamic Conditional Correlations: Engle (2002) shows how to generalize the CCC model so that the correlations vary over time. Instead of estimating the all of the parameters simultaneously, the Dynamic Conditional Correlation (DCC) model uses a two-step estimation process. The first step is to use Bollerslev's CCC model to obtain the GARCH estimates of the variances and the standardized residuals. Note that the standardized residuals, $s_{it} = \hat{\varepsilon}_{it} / \hat{h}_{iit}^{0.5}$, are estimates of the v_{it}. The second step uses the standardized residuals to estimate the conditional covariances. Specifically, in the second step you create the correlations by smoothing the series of standardized residuals obtained from the first step. Engle examines several smoothing methods. The simplest is the exponential smoother $q_{ijt} = (1 - \lambda)s_{it}s_{jt} + \lambda q_{ijt□1}$ for $\lambda < 1$. Hence, each $\{q_{iit}\}$ series is an exponentially weighted moving average of the cross-products of the standardized residuals. The dynamic conditional correlations are created from the q_{ijt} as

$$\rho_{ijt} = q_{ijt} / (q_{iit} q_{jjt})^{0.5} \tag{A3.4}$$

Engle shows that a two-step procedure yields consistent estimates of the time varying correlation coefficients. However, the estimates are not as efficient as those from one-step procedures such as the BEK and diagonal *vech* models. Restricting the coefficient on $\overline{s}_{ij}$ to equal $(1 - \alpha - \beta)$ ensures that the q_{ijt} converge to the unconditional covariances.

An alternative smoothing function is to estimate $q_{ijt} = (1-\alpha-\beta)\overline{s}_{ij} + \alpha s_{it}s_{jt} + \beta q_{ijt\square 1}$ where $\overline{s}_{ij}$ is the unconditional covariance between s_{it} and s_{jt} by maximum-likelihood estimation. Plug the estimated coefficients from the first step (i.e., from the CCC model) into the likelihood function so that only α and β need to be estimated.

For those of you wanting a formal proof that the 2-step procedure is feasible, you should be able to convince yourself that it is possible to write the H_t matrix as

$$H_t = D_t R_t D_t$$

where D_t = the diagonal matrix with $(h_{iit})^{0.5}$ on the diagonals and R_t is the matrix of time-varying correlations. This follows from the definition of a correlation coefficient; R_t consists of the elements $r_{ijt} = (h_{ijt})/(h_{iit}h_{jjt})^{0.5}$. For example, in the 2-variable case it is easy to verify $H_t = D_t R_t D_t$ or $R_t = (D_t)^{-1} H_t (D_t)^{-1}$ since

$$R_t = \begin{pmatrix} h_{11t}^{0.5} & 0 \\ 0 & h_{22t}^{0.5} \end{pmatrix}^{-1} \begin{pmatrix} h_{11t} & h_{12t} \\ h_{12t} & h_{22t} \end{pmatrix} \begin{pmatrix} h_{11t}^{0.5} & 0 \\ 0 & h_{22t}^{0.5} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} 1 & h_{12t}/(h_{11t}h_{22t})^{0.5} \\ h_{12t}/(h_{11t}h_{22t})^{0.5} & 1 \end{pmatrix}$$

Now write the likelihood function (A3.3) by substituting $D_t R_t D_t$ for H_t as

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(\ln|D_t R_t D_t| + \varepsilon_t'(D_t R_t D_t)^{-1}\varepsilon_t)$$

$$= -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(2\ln|D_t| + \ln|R_t| + \varepsilon_t'(R_t)^{-1}\varepsilon_t) \tag{A3.5}$$

Notice that D_t and R_t enter the likelihood separately and that $\varepsilon_t' R_t \varepsilon_t$ represents the squared standardized residuals. The final step is to add and subtract the sum of the squared standardized residuals to (A3.5). If we represent the standardized residuals by v_t, the sum of the squared standardized residuals is $v_t'v_t$. It is also possible to show that $v_t'v_t = \varepsilon_t' D_t^{-1} D_t^{-1} \varepsilon_t$. For example, in the 2-variable case,

$$\varepsilon_t' D_t^{-1} D_t^{-1} \varepsilon_t = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}' \begin{pmatrix} h_{11t}^{0.5} & 0 \\ 0 & h_{11t}^{0.5} \end{pmatrix}^{-1} \begin{pmatrix} h_{11t}^{0.5} & 0 \\ 0 & h_{11t}^{0.5} \end{pmatrix}^{-1} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix} = \frac{\varepsilon_{1t}^2}{h_{11t}} + \frac{\varepsilon_{2t}^2}{h_{22t}}$$

Thus, we can write the likelihood function as

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{1}{2}\sum_{t=1}^{T}(2\ln|D_t| + \ln|R_t| + \varepsilon_t'(R_t)^{-1}\varepsilon_t - v_t'v_t + \varepsilon_t' D_t^{-1}D_t^{-1}\varepsilon_t)$$

The point of the exercise is to show that the 2-step procedure is appropriate. Notice that D_t and R_t enter the equation separately. As such, the parameters of the two matrices can be estimated separately. You can use the CCC model to estimate the parameters of D_t; this can be done without any knowledge of the values of R_t. Use these estimates to construct the values of | D_t | and the standardized residuals. Plug these values into the likelihood function and then select the optimal values of R_t. In essence, in the first stage, you maximize

$$-\frac{1}{2}\sum_{t=1}^{T}(2\ln|D_t| + \varepsilon_t' D_t^{-1}D_t^{-1}\varepsilon_t)$$

and in the second stage you maximize

$$-\frac{1}{2}\sum_{t=1}^{T}(\ln|R_t| + \varepsilon_t'(R_t)^{-1}\varepsilon_t - v_t'v_t)$$

Section 3.2: Value at Risk

Value at Risk (VaR) is a concept used by portfolio managers to measure the downside risk of a particular portfolio of financial instruments. For any pre-specified probability level *p*, the VaR is the value of the loss that will occur with probability *p*. Usually, the time period is a single day, but other time horizons are possible. For example, if a portfolio of stocks has a one-day 5% VaR of $10 million, there is a 5% probability that the portfolio will fall in value by more than $10 million over a one day period.

One way to calculate VaR is to use a GARCH model. Suppose that the continually compounded daily return of a portfolio (r_t) follows a conditional normal distribution such that:

$$E_{t-1}r_t \sim N(0, h_t)$$

where the conditional variance h_t follows an IGARCH process. Let

$$h_t = \alpha_0 + \alpha_1(e_{t-1})^2 + (1 - \alpha_1)(h_{t-1})^2$$

Now suppose that you want to know the value at risk of a portfolio using a 5% probability. As such, you can 1.64 standard deviations [$= 1.64(h_{t+1})^{1/2}$] to measure the risk of the portfolio. In general, the Value at Risk for one day is:

VaR = Amount of Position x $1.64(h_{t+1})^{1/2}$ and for *k* days is

and the Value at Risk for *k* days is

VaR(*k*) = Amount of Position x $1.64(k\ h_{t+1})^{1/2}$

To take a specific example, suppose that the model of the mean for the return on a particular stock (or a portfolio of stocks) is:

$$r_t = 0.001 + 0.02r_{t-1} + \varepsilon_t$$

and that

$$h_t = 0.004 + 0.1(\varepsilon_{t-1})^2 + 0.9(h_{t-1})^2$$

Also suppose that the values of r_t, ε_{t-1} and h_{t-1} are such that

$$E_t(r_{t+1}) = 0.025$$

and

$$E_t(h_{t+1}) = 0.005$$

Now, the issue is to find the amount that is 1.64 standard deviations below the expected return. The 5% quantile is calculated to be

$$0.025 – 1.64*(0.005)^{1/2} = –0.091$$

As such, –0.091 is the value that is 1.64 standard deviations below the expected return of 0.025. Thus, is you had $1 invested in this stock, you would expect a 0.025 return but there would be a 5% chance of a return less than or equal to –0.091. The VaR for a portfolio size of $10,000,000 with probability 0.05 is ($10,000,000)(0.091) = $910,000. As such, with 95% chance, the potential loss of the portfolio is $910,000 or less.

CHAPTER 4

ENDNOTES TO CHAPTER 4

1. Many treatments use the representation y_t = trend + seasonal + cyclical + noise. In this text, the term *cyclical* is avoided because it implies that cyclical economic components are deterministic.

2. For the same reason, it is also inappropriate to use one variable that is trend stationary and another that is difference stationary. In such instances, “time” can be included as a so–called *explanatory* variable, or the variable in question can be detrended.

3. Suppose that the estimated value of γ is –1.9 (so that the estimate of a_1 is –0.9) with a standard error of 0.04. Since the estimated value of γ is 2.5 standard errors from –2 [(2 – 1.9)/0.04 = 2.5], the Dickey–Fuller statistics indicate that we cannot reject the null hypothesis $a_1 = -2$ at the 95 percent significance level. Unless stated otherwise, the discussion in this text assumes that a_1 is positive.

4. When the distribution for v_t is more complicated, the distribution of the mean may not be normal with variance σ^2/T.

5. The ERS procedure is called **Generalized Least Squares** (GLS) detrending because of the way that the near-differencing is performed. Suppose $B(L)$ is the first-order autoregressive process: $\varepsilon_t + \alpha\varepsilon_{t-1}$. Forming $y_t - \alpha y_{t-1}$ yields the serially uncorrelated error structure used in GLS. In the problem at hand, the actual α is unknown. However, if the y_t series is persistent, such differencing should mean that the ACF of $B(L)\varepsilon_t - \alpha B(L)\varepsilon_{t-1}$ is close to that of a white noise process.

6. To explain, if the error process were such that $B(L) = \sum_{i=0}^{\infty}\alpha^i\varepsilon_{t-i}$, the variance of the error term would be $\sigma^2/(1 - \alpha^2)$. Treating $\sigma^2 = 1$, and dividing y_1 by its standard deviation yields $y_1(1 - \alpha^2)^{0.5}$.

Section 4.1 More on Unobserved Component Models

The purpose of this section is to expand the discussion of unobserved component models. Harvey (1989) contains a detailed treatment of the issue. The random walk plus noise model and the general trend plus irregular model are examples of processes with several unobserved components. Reconsider the general trend plus irregular model of (4.9). The variable y_t might represent real GDP, ε_t might represent a productivity shock and η_t might represent a demand-side shock. The specification in (4.9) implies that productivity shocks, but not demand shocks, have permanent effects on real GDP.

The local linear trend (LLT) model is built by combining several random walk plus noise processes. Let $\{\varepsilon_t\}$, $\{\eta_t\}$ and $\{\delta_t\}$ be three mutually uncorrelated white noise processes. The local linear trend model can be represented by

$$y_t = \mu_t + \eta_t$$

$$\mu_t = \mu_{t-1} + a_t + \varepsilon_t$$

$$a_t = a_{t-1} + \delta_t$$

The local linear trend model consists of the noise term η_t plus the stochastic trend term μ_t. What is interesting about the model is that the *change in the trend* is a random walk plus noise: that is, $\Delta\mu_t$ is equal to the random walk term a_t plus the noise term ε_t. Since this is the most detailed model thus far, it is useful to show that the other processes are special cases of the local linear trend model. For example:

1. **The random walk plus noise**: If all values of the $\{a_t\}$ sequence are equal to zero, the LLT model degenerates into a random walk ($\mu_t = \mu_{t-1} + \varepsilon_t$) plus noise ($\eta_t$). Let var($\delta$) = 0, so that $a_t = a_{t-1} = ... = a_0$. If $a_0 = 0$, $\mu_t = \mu_{t-1} + \varepsilon_t$ so that y_t is the random walk μ_t plus noise term η_t.
2. **The random walk plus drift**: Again, let var(δ) = 0, so that $a_t = a_{t-1} = ... = a_0$. Now if a_0 differs from zero, the trend is the random walk plus drift: $\mu_t = \mu_{t-1} + a_0 + \varepsilon_t$. Thus, the LLT model becomes trend plus noise model. If we further restrict the model such that var(η_t) = 0, the model becomes the pure random-walk plus drift model.

The solution for y_t can easily be found as follows. First, solve for a_t as:

$$a_t = a_0 + \sum_{i=1}^{t} \delta_i$$

Next, use this solution to write μ_t as

$$\mu_t = \mu_{t-1} + a_0 + \sum_{i=1}^{t} \delta_i + \varepsilon_t$$

so that

$$\mu_t = \mu_0 + \sum_{i=1}^{t} \varepsilon_i + t(a_0 + \delta_1) + \delta_2(t-1) + \delta_3(t-3) + ... + \delta_t$$

Since, $y_0 = \mu_0 + \eta_0$, the solution for y_t is

$$y_t = y_0 + (\eta_t - \eta_0) + \sum_{i=1}^{t} \varepsilon_i + t(a_0 + \delta_1) + (t-1)\delta_2 + (t-2)\delta_3 + ... + \delta_t$$

Here we can see the combined properties of all the other models. Each element in the $\{y_t\}$ sequence contains a deterministic trend, a stochastic trend, and an irregular term. The stochastic trend is $\Sigma\varepsilon_i$ and the irregular term is η_t. Of course, in a more general version of the model, the irregular term could be given by $A(L)\eta_t$. What is most interesting about the model is the form of the deterministic time trend. Rather than being deterministic, the *coefficient* on the time depends on the current and past realizations of the $\{\delta_t\}$ sequence. If in period t, the realized value of the sum $a_0 + \delta_1 + .. + \delta_t$ happens to be positive, the coefficient of t will be positive. Of course, this sum can be positive for some values of t and negative for others.

Signal Extraction

Signal extraction issues arise when we try to decompose a series into its individual components. Suppose we observe the realizations of a stationary sequence $\{y_t\}$ and want to find the optimal predictor of its components. Phrasing the problem this way, it is clear that the decomposition can be performed using the minimum MSE criterion discussed above. As an example of the technique, consider a sequence composed of two independent white-noise components:

$$y_t = \varepsilon_t + \eta_t$$

where $E\varepsilon_t = 0$

$E\eta_t = 0$

$E\varepsilon_t\eta_t = 0$

$E\varepsilon_t^2 = \sigma^2$

$E\eta_t^2 = \sigma_\eta^2$.

Here the correlation between the innovations is assumed to be equal to zero; it is straightforward to allow non-zero values of $E\varepsilon_t\eta_t$. The problem is to find the optimal prediction, or forecast, of ε_t (called ε_t*) conditioned of the observation of y_t. The linear forecast has the form

$$\varepsilon_t^* = a + by_t$$

In this problem, the intercept term a will be zero so that the MSE can be written as

$$\begin{aligned} \text{MSE} &= E(\varepsilon_t - \varepsilon_t^*)^2 \\ &= E(\varepsilon_t - by_t)^2 \\ &= E[\varepsilon_t - b(\varepsilon_t + \eta_t)]^2 \end{aligned}$$

Hence the optimization problem is to select b so as to minimize:

$$\text{MSE} = E[\ (1\text{-}b)\varepsilon_t - b\eta_t\ ]^2$$
$$= (1\text{-}b)^2E\varepsilon_t^2 + b^2E\eta_t^2 \qquad \text{since } E\varepsilon_t\eta_t = 0.$$

The first-order condition is

$$-2(1\text{-}b)\sigma^2 + 2b\sigma_\eta^2 = 0$$

so that

$$b = \sigma^2/(\sigma^2 + \sigma_\eta^2)$$

Here, b partitions y_t in accord with the relative variance of ε_t; i.e., $\sigma^2 /(\sigma^2 + \sigma_\eta^2)$. As σ^2 becomes very large relative to σ_η^2, $b \rightarrow 1$; as σ^2 becomes very small relative to σ_η^2, $b \rightarrow 0$. Having extracted ε_t, the predicted value of η_t is: $\eta_t^* = y_t - \varepsilon_t^*$. However, this optimal value of b depends on the assumption that the two innovations are uncorrelated. Although the computation becomes far more complex with a model like the *LLT*, the methodology is the same.

Signal Extraction and Least-Squares Projection

The problem for the econometric forecaster is to select an optimal forecast of a random variable y conditional on the observation of a second variable x. Since the theory is quite general, for the time being we ignore time subscripts. Call this conditional forecast y^* so that the forecast error is $(y\text{-}y^*)$ and the mean square forecast error (MSE) is $E(y - y^*)^2$. One criterion used to compare forecast functions is the MSE; the optimal forecast function is that with the smallest MSE.

Suppose x and y are jointly distributed random variables with known distributions. Let the mean and variance of x be μ_x and $\sigma^2{}_x$, respectively. Also, suppose the value of x is observed before having to predict y. A *linear* forecast will be such that the forecast y^* is a linear function of x. The optimal forecast will necessarily be linear if x and y are linearly related, and/or if they are bivariate normally distributed variables. In this text, only linear relationships are considered; hence, the optimal forecast of y^* has the form

$$y^* = a + b(x - \mu_x)$$

The problem is to select the values of a and b so as to minimize the MSE:

$$\underset{\{a,\,b\}}{\text{Min}}\ E(y - y^*)^2 = E[y - a - b(x\text{-}\mu_x)]^2$$
$$= E[y^2 + a^2 + b^2\,(x\text{-}\mu_x)^2 - 2ay + 2ab\,(x\text{-}\mu_x) - 2by(x\text{-}\mu_x)]$$

Since $E(x - \mu_x) = 0$, $Ey = \mu_y$, $E(x\text{-}\mu_x)^2 = \sigma^2{}_x$, and $E(xy) - \mu_x\mu_y = \text{Cov}(x\ ,\ y) = \sigma_{xy}$, it follows that

$$E(y - y^*)^2 = Ey^2 + a^2 + b^2\sigma^2{}_x - 2a\mu_y - 2b\sigma_{xy}$$

Minimizing with respect to a and b yields

$$a = \mu_y, \qquad b = \sigma_{xy}/\sigma^2{}_x$$

Thus, the optimal prediction formula is

$$y^* = \mu_y - (\sigma_{xy}/\sigma^2{}_x)\mu_x + (\sigma_{xy}/\sigma^2{}_x)x$$

The forecast is unbiased in the sense that the mean value of the forecast is equal to the mean value of y. Take the expected value of y^* to obtain:

$$Ey^* = E[\mu_y - (\sigma_{xy}/\sigma^2{}_x)\mu_x + (\sigma_{xy}/\sigma^2{}_x)x]$$

Since, $\mu_{y,}$ σ_{xy}, and $\sigma^2{}_x$ are constants, and that $\mu_x = Ex$, it follows that

$$Ey^* = \mu_y$$

You should recognize this formula from standard regression analysis; a regression equation is the minimum mean square error, linear, unbiased forecast of y^*. The argument easily generalizes forecasting y conditional on the observation of the n variables x_1 through x_n and to forecasting y_{t+s} conditional on the observation of y_t, $y_{t\text{-}1}$, ... For example, if $y_t = a_0 + a_1y_{t\text{-}1} + \varepsilon_t$ the conditional forecast of y_{t+1} is: $E_ty_{t+1} = a_0 + a_1y_t$. The forecasts of y_{t+s} can be obtained using the forecast function (or iterative forecasts) discussed in section 11 of Chapter 2.

Forecasts of a Non-stationary Series Based on Observables

Muth (1960) considers the situation in which a researcher wants to find the optimal forecast of y_t conditional on the observed values of $y_{t\text{-}1}$, $y_{t\text{-}2}$, ... Let $\{y_t\}$ be a random-walk plus noise. If all realizations of $\{\varepsilon_t\}$ are zero for $t \le 0$, the solution for y_t is:

$$y_t = \sum_{i=1}^{t} \varepsilon_i + \eta_t \qquad \text{(A4.1)}$$

where y_0 is given and $\mu_0 = 0$.

Let the forecast of y_t be a linear function of the past values of the series so that:

$$y_t{*} = \sum_{i=1}^{\infty} v_i \; y_{t-i} \tag{A4.2}$$

where the various values of v_i are selected so as to minimize the mean square forecast error.

Use (A4.1) to find each value of y_{t-i} and substitute into (A4.2) so that:

$$y_t{*} = v_1\left(\sum_{i=1}^{t-1} \varepsilon_i + \eta_{t-1} \right) + v_2\left(\sum_{i=1}^{t-2} \varepsilon_i + \eta_{t-2} \right) + v_3\left(\sum_{i=1}^{t-3} \varepsilon_i + \eta_{t-3} \right) + \ldots$$

Thus, optimization problem is to select the v_j so as to minimize the MSE:

$$E\,[y_t - y_t{*}\,]^2 = E\left[\sum_{i=1}^{t} \varepsilon_i + \eta_t - v_1\left(\sum_{i=1}^{t-1} \varepsilon_i + \eta_{t-1} \right) - v_2\left(\sum_{i=1}^{t-2} \varepsilon_i + \eta_{t-2} \right) - \ldots \right]^2$$

Since the expected value of all cross products are zero, the problem is to select the v_j so as to minimize

$$\text{MSE} = t\sigma_\varepsilon^2 + \sigma_\eta^2 + \sigma_\varepsilon^2 \sum_{i=1}^{\infty}\left[1 - \sum_{j=1}^{i} v_j \right] + \sigma_\eta^2 \sum_{j=1}^{\infty} v_j^2$$

For each value of v_k, the first-order conditions is:

$$2\,\sigma_\eta^2\, v_k - 2\,\sigma_\varepsilon^2 \sum_{j=k}^{\infty}\left(1 - \sum_{i=1}^{j} v_i \right) = 0 \quad k = 1, 2, \ldots \tag{A4.3}$$

All $\{v_k\}$ will satisfy the difference equation given by (A4.3). To characterize the nature of the solution, set $k = 1$, so that the first equation of (A4.3) is

$$2\sigma_\eta^2 v_1 - 2\sigma_\varepsilon^2 \sum_{j=1}^{\infty}\left(1 - \sum_{i=1}^{j} v_i \right) = 0$$

and for $k = 2$,

$$2\sigma_{\eta}^{2} v_2 - 2\sigma_{\varepsilon}^{2} \sum_{j=2}^{\infty} \left(1 - \sum_{i=1}^{j} v_i \right) = 0$$

so that by subtraction,

$$\sigma_{\varepsilon}^{2} \ (1 - v_1) + \sigma_{\eta}^{2} \ (v_2 - v_1) = 0 \qquad \text{(A4.4)}$$

Now take the second-difference of (A4.3) to obtain:

$$- v_{k-1} + \left(2 + \frac{\sigma_{\varepsilon}^{2}}{\sigma_{\eta}^{2}} \right) v_k - v_{k+1} = 0 \quad \text{for } k = 2, 3, \ldots$$

The solution to this homogeneous second-order difference equation has the form: $v_k = A_1\lambda_1^k + A_2\lambda_2^k$ where A_1 and A_2 are arbitrary constants and λ_1 and λ_2 are the characteristic roots. If you use the quadratic formula, you will find that the larger root (say λ_2) is greater than unity; hence, if the $\{v_k\}$ sequence is to be convergent, A_2 must equal zero. The smaller root satisfies

$$\lambda_1^2 - (2 + \sigma^2_{\varepsilon}/\sigma^2_{\eta})\lambda_1 + 1 = 0 \qquad \text{(A4.5)}$$

To find the value of A_1, substitute $v_1 = A_1\lambda_1$ and $v_2 = A_1\lambda_1^2$ into (A4.4):

$$\sigma^2_{\varepsilon}(1 - A_1\lambda_1) - \sigma^2_{\eta}A_1(\lambda_1^2 - \lambda_1) = 0$$

If you solve (A4.5) for λ_1, it is possible to verify:

$$A_1 = (1 - \lambda_1)/\lambda_1$$

Hence the v_k are determined by:

$$v_k = (1 - \lambda_1)\lambda_1^{k-1}$$

The one-step ahead forecast of y_t is

$$y_t{}^* = (1 - \lambda_1) \sum_{j=1}^{\infty} \lambda_1^{j\text{-}1} y_{t-j}$$

Since $|\lambda_1| < 1$, the summation is such that: $(1\text{-}\lambda_1)\Sigma\lambda_1{}^{j\text{-}1} = 1$. Hence, the optimal forecast of y_t can be formed as a geometrically weighted average of the past realizations of the series.

Section 4.2: Random Number Generation

Random number generation is an essential feature of the Monte Carlo methods described in Chapter 4.

Computers are not capable of generating truly random numbers--any sequence generated is actually a deterministic sequence. If you are aware of the algorithm used to generate the sequence all values of the sequence can be calculated by the outside observer. Computers generate **pseudo-random numbers--**the numbers generated are indistinguishable from those obtained from independent draws from a uniform distribution.

A common algorithm used in random number generation involved the mod( ) function: mod(x, z) means divide x by z and keep only the remainder. For example, mod (3, 5) = 3, mod(6, 5) = mod(11, 5) = 1, and mod(11.3, 5) = 1.3. Of course, in a computer, 1/3 will be an approximate value since it is not possible to write a decimal equivalent of 1/3 using a finite number of digits.

Consider the nonlinear difference equation:

$$z_{t+1} = \text{mod}(\lambda z_t + \alpha, m)$$

$$y_t = z_t/m$$

where: m, λ, and α are parameters.

If we use $z_1 = 1$, $\lambda = 2$, $m = 10$ and $\alpha = 5$, the next 5 values of the $\{z_t\}$ and $\{y_t\}$ sequences are:

z_2 = mod(2*1 + 5, 10) = 7 so that $y_2 = 0.7$
z_3 = mod(2* 7 + 5, 10) = 9 so that $y_3 = 0.9$
z_4 = mod(2*9 + 5, 10) = 3 so that $y_4 = 0.3$
z_5 = mod(2*3 + 5, 10) = 1 so that $y_5 = 0.1$

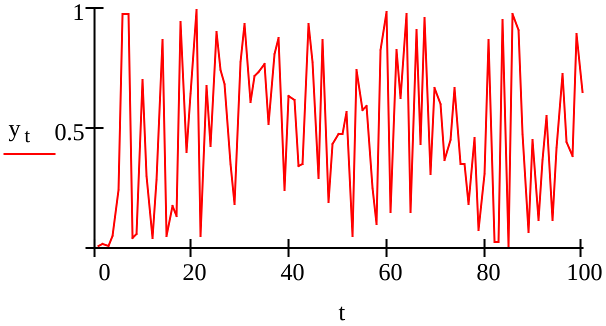
so that the series repeats itself.

The point is that not all parameter choices for α, m, and λ are well-behaved. Note that λ is called the multiplier. Clearly, λ needs to be greater than unity so that the numbers do not converge to zero. Nevertheless, some values of $\lambda > 1$ will lead to poorly behaved sequences. Also note that m is should be a very large number to ensure that the sequence does not repeat itself very quickly. A series produced by this type of random number generator will repeat itself in no more than m steps. In addition, some values of m will display serial correlation; it is important to select a value of m such that the degree of serial correlation is small. A random number generation module for Mathcad uses the values $m = 732289$, $\lambda = 9947$, and $\alpha = 67$. If we start using the seed value $z_1 = 1$, it follows that

z_2 = mod(9947*732289*1+ 67, 732289) = 10014 so that $y_2 = 1.36536 \times 10^{-6}$
z_3 = mod(9947*732289*10014+ 67, 732289) = 4421 so that $y_3 = 0.01367$
z_4 = mod(9947*732289*4421+ 67, 732289) = 32414 so that $y_4 = 0.04426$
z_5 = mod(9947*732289*32414+ 67, 732289) = 170965 so that $y_5 = 0.23343$

The numbers generated by this set of parameter values will closely approximate a set of serially uncorrelated uniformly distributed random variables over the interval [0, 1]. The time path of the first 100 values of the $\{y_t\}$ series is given by:

Figure 1: 100 Pseudo-Random Numbers



By construction, z_t must be less than m. As such each value of y_t is between zero and unity. For this particular value of m, the correlation coefficient between y_t and y_{t-1} is 0.02617. However, if m = 992 is selected, the correlation coefficient will be 0.30176.

Given the values of $\{y_t\}$, it is possible to make other transformations of the series so as to generate distributions other than a uniform distribution.

Note the important difference between correlation and independence. Each pseudo-random number is perfectly predicable if you know the initial seed value and the algorithm generating the numbers. Nevertheless, it is possible to generate sets of numbers that are serially uncorrelated. Recall that correlation is simply a measure of linear dependence. The random number generating routine described here is clearly nonlinear.

SECTION 4.3: The Bootstrap

Bootstrapping is similar to a Monte Carlo experiment with one essential difference. In a Monte Carlo study, you generate the random variables from a given distribution such as the Normal. The bootstrap takes a different approach—the random variables are drawn from their observed distribution. In essence, the bootstrap uses the **plug–in principle**—the observed distribution of the random variables is the best estimate of their actual distribution.

The idea of the bootstrap was developed in Efron (1979). The key point made by Efron is that the observed data set is a random sample of size *T* drawn from the actual probability distribution generating the data. In a sense, the empirical distribution of the data is the best estimate of the actual distribution of the data. As such, the empirical distribution function is defined to be the discrete distribution that places a probability of 1/*T* on each of the observed values. It is the empirical distribution function—and not some prespecified distribution such as the Normal—that is used to generate the random variables. The **bootstrap sample** is a random sample of size *T* drawn *with* replacement from the observed data putting a probability of 1/*T* on each of the observed values.

Example: To be more specific, suppose that we have the following 10 values of x_t:

1	2	3	4	5	6	7	8	9	10
0.8	3.5	0.5	1.7	7.0	0.6	1.3	2.0	1.8	–0.5

The sample mean is 1.87 and the standard deviation is 2.098. The following data show three different bootstrap samples. Each bootstrap sample consists of 10 randomly selected values of x_t drawn with replacement—each of the 10 values listed above is drawn with a probability of 0.1. It might seem that this resampling repeatedly selects the same sample. However, by sampling with replacement, some elements of x_t will appear more than once in the bootstrap sample. The first three bootstrap samples might look like this:

t	1	2	3	4	5	6	7	8	9	10		μ_i
											*	
x		1	–	0	1	2	1	0	0	7.0		1.
$_1^*$	3.5	.7	0.5	.5	.8	.0	.7	.6	.6		89	
x	–	0		0	1	7	1	3	1	0.8		1.
$_2^*$	0.5	.6	0.6	.8	.7	.0	.8	.5	.8		81	
x		0		1	1	7	1	1	3	0.6		2.
$_3^*$	0.5	.6	7.0	.3	.3	.0	.3	.8	.5		49	

where: x_i^* denotes bootstrap sample *i* and μ_i^* is the sample mean.

Notice that 0.6 and 1.7 appear twice in the first bootstrap sample, 0.6, 0.8 and 1.8 appear twice in the second and that 1.3 appears three times in the third bootstrap sample. Unless there is a large outlier, Efron (1979) shows that the moments of the bootstrap samples converge to the population moments as the number of bootstrap samples goes to infinity.

Bootstrapping Regression Coefficients

Suppose you have a data set with T observations and want to estimate the effects of variable x on variable y. Towards this end, you might estimate the linear regression:

$$y_t = a_0 + a_1 x_t + \varepsilon_t$$

Although the properties of the estimators are well known, you might not be confident about using standard *t-test*s if the estimated residuals do not appear to be normally distributed. You could perform a Monte Carlo study concerning the statistical properties of $\hat{a}_0$ and $\hat{a}_1$. However, instead of selecting various values of ε_t from a normal distribution, you can use the actual regression residuals. The technique is called the method of **bootstrapped residuals.**[1] To use the procedure, perform the following steps:

STEP 1: Estimate the model and calculate the residuals as: $e_t = y_t - \hat{a}_0 - \hat{a}_1 x_t$.

STEP 2: Generate a bootstrap sample of the error terms containing the elements e_1^*, e_2^*, … , e_T^*. Use the bootstrap sample to calculate a bootstrapped y series (called y^*). For each value of t running from 1 to T, calculate y_t^* as:

$$y_t^* = \hat{a}_0 + \hat{a}_1 x_t + e_t^*$$

Note that the estimated values of the coefficients are treated as fixed. Moreover, the values of x_t are treated as fixed quantities so that they remain the same across samples.

STEP 3: Use the bootstrap sample of the y_t^* series generated in Step 1 to estimate new values of a_0 and a_1 calling the resulting values a_0^* and a_1^*.

STEP 4: Repeat Steps 2 and 3 many times and calculate the sample statistics for a_0^* and a_1^*. These should be distributed in the same way as $\hat{a}_0$ and $\hat{a}_1$. For example, you can find the 95 percent confidence interval for $\hat{a}_1$ as interval between the lowest 2.5 percent and the highest 97.5 percent of the values of a_1^*.

Step 2 needs to be modified for a time series model due to the presence of lagged dependent variables. As such, the bootstrap { y_t^* } sequence is constructed in a slightly different manner. Consider the simple AR(1) model:

$$y_t = a_0 + a_1 y_{t-1} + \varepsilon_t$$

As in Step 2, we can construct a bootstrap sample of the error terms containing the elements e_1^*, e_2^*, … , e_T^*. Now, construct the bootstrap { y_t^* } sequence using this sample of error terms. In particular, given the estimates of a_0, and a_1 and an initial conditions for y_1^*, the remaining values of the { y_t^* } sequence can be constructed using:

$$y_t^* = \hat{a}_0 + \hat{a}_1 y_{t-1}^* + e_t^*$$

For an AR(p) model, the initial conditions for y_1^* through y_p^* are usually selected by random draws from the actual $\{y_t\}$ sequence.[2] To avoid problems concerning the selection of the initial condition, it is typical to construct a bootstrap sample with T + 50 elements and to discard the first 50.

Example 1: Constructing a Confidence Interval. In Sections 2 and 12, the logarithmic change in real U.S. GDP was estimated as

$$\underset{}{\Delta lrgdp_t} = \underset{(6.80)}{0.0049} + \underset{(6.44)}{0.3706}\Delta lrgdp_{t-1}$$

Suppose that you want to obtain a 90% confidence interval for the AR(1) coefficient. Given that the t-statistic is 6.44, the standard deviation is 0.0576. Thus, if the estimated value of a_1 does have a t-distribution, a confidence interval that spans ± 1.65 standard deviations on each side of 0.3706 runs from 0.276 to 0.466. However, the residuals do not appear to be normally distributed so that, the t-distribution may not be appropriate.[25] As such, we may want to construct bootstrap confidence intervals for the coefficient 0.3309. For Step 2, we need to generate a bootstrap sample of the regression residuals. This series, denoted by $\{ e_t^* \}$, consists of 262 randomly drawn values from the actual residual sequence. We also need an initial condition; we can draw a random value of the $\{\Delta lrgdp_t\}$ sequence to serve as the initial condition. As such, we can create the following bootstrap sequence for $\{ y_t^* \}$ containing a total of 262 observations.

$$y_t^* = 0.0049 + 0.3706\, y_{t-1}^* + e_t^*$$

For Step 3, we pretend that we do not know the actual data-generating process for $\{ y_t^* \}$ and estimate the series as an AR(1) process. Suppose you find that the estimated AR(1) coefficient is 0.35. This estimate is the first value of a_1^*. Repeat this process several thousand times to obtain the distribution of a_1^*. After performing the 10,000 replications of the experiment, you should find that approximately 5 percent of the estimates lie below 0.2652 and 5 percent lie above 0.4524. These values serve as a 90 percent confidence interval. In this case, it turns out that the bootstrap confidence interval is similar to that obtained using a standard t-distribution.

Example 2: Bootstrapping a Test Statistic. Equation (4.29) was used to perform a Dickey–Fuller test on the real GDP series. Recall that the estimated model was

$$\Delta lrgdp_t = \underset{(1.58)}{0.1248} + \underset{(1.31)}{0.0001t} - \underset{(-1.49)}{0.0156}lrgdp_{t-1} + \underset{(6.26)}{0.3663}\Delta lrgdp_{t-1}$$

The t-statistic for γ (i.e., the coefficient on $lrgdp_{t-1}$) of –1.49 is not significant at conventional levels. However, a concern might be that the residuals from model are not normally distributed. Moreover, only in large samples do the coefficients of the augmented terms have no bearing on the appropriate critical values. Hence, it seems reasonable to bootstrap the τ_τ-statistic for (4.29). If we bootstrap the series under the null hypothesis of no deterministic time trend and $\gamma = 0$, we need only construct the bootstrap series.

$$y_t^* = 0.0049 + 0.3706\, y_{t-1}^* + e_t^*$$

Since, y_t^* represents a first difference, it is necessary to create the level of the series as Y_t^* = Y_{t-1}^* + y_t^*. Now, for Step 3, estimate the bootstrap series in the form

$$y_t^* = a_0^* + \gamma^* Y_{t-1}^* + \lambda^* t + a_1^* y_{t-1}^*$$

Replicate the experiment many times; on each trial, record the *t*-statistic for the null hypothesis $\gamma^* = 0$. The value of the *t*-statistic marking the fifth percentile of all of the bootstrapped *t*-statistics is the 5 percent critical value. In addition, to obtain the bootstrap ϕ_3-statistic, we can obtain the sample values of *F* for the null hypothesis $\gamma^* = \lambda^* = 0$.

You can select the initial value of Y_1^* as log($RGDP_1$) plus an error term or a randomly drawn value of real GDP. However, as mentioned above, most researchers would actually construct a series with an extra 50 observations and then discard the first 50 of the realizations. When this process was repeated 10,000 times, only 500 (i.e., 5 percent) of the *t*-statistics for the null hypothesis $\gamma^* = 0$ had values below –3.43. Given the actual value of the *t*-statistic of –1.49, we cannot reject the null hypothesis $\gamma = 0$ at the 5 percent significance level. Moreover, 95 percent of the *F*-statistics for the null hypothesis $\gamma^* = \lambda^* = 0$ were less than 6.38. Recall that the sample value ϕ_3 from equation (4.29) was only 2.97. Hence, at the 5 percent significance level, we cannot reject the null hypothesis $\gamma = \lambda = 0$.

It is also possible to use the bootstrap results to test the null hypothesis $\lambda^* = 0$. Of course, if we know that the series is stationary, the *t*-statistic of 1.31 indicates that the time trend does not belong in the regression. However, we cannot use the usual *t*-distribution if real GDP is not stationary. In total, 5% of the bootstrapped *t*-statistics for the null hypothesis $\lambda^* = 0$ were below 0.615 and 5% were above 3.69. Since the actual value of the *t*-statistic is 1.31, if we use a 90% confidence interval we cannot reject the null hypothesis that the time trend does not appear in (4.29).

Example 3: Bootstrapping Correlated Residuals in a Panel. If the residuals from a panel unit root test are highly correlated across equations, you should use bootstrapped critical values. The only modification needed from Example 2 is that in a panel unit root test, you estimate a vector of equations. The essential requirement is that the residuals need to be sampled in such a way that preserves the contemporaneous correlations present in the data. Let $\{e_{it}\}$ denote the regression residual from (4.45) for regression *i* for time period *t*. If e_t denotes the vector of residuals for time period *t*, we can write

$e_1 = (e_{11}, e_{21}, e_{31}, \dots , e_{n1})$
$e_2 = (e_{12}, e_{22}, e_{32}, \dots , e_{n2})$
$e_3 = (e_{13}, e_{23}, e_{33}, \dots , e_{n3})$
…
$e_T = (e_{1T}, e_{2T}, e_{3T}, \dots , e_{nT})$

Thus, the first bootstrap sample of the residuals might be

$= (e_{13}, e_{23}, e_{33}, \dots , e_{n3})$
$= (e_{15}, e_{25}, e_{35}, \dots , e_{n5})$
…

$= (e_{12}, e_{22}, e_{32}, \ldots, e_{n2})$

The point is that you resample in such a way as to maintain the contemporaneous relationships among the regression residuals. As in Example 2, construct a bootstrap series using the resampled values of $\{e_t^*\}$. Once you obtain the average value of the *t*-statistics for the first bootstrap sample, repeat the entire process several thousand times.

Efron and Tibshirani (1993) is an extremely accessible treatment of bootstrapping. You may also download a programming manual that illustrates some bootstrapping techniques (at no charge) from my website or from Wiley Web site for this text.

Footnotes

[1] Another possibility discussed by Efron is to resample the paired (y_i, x_i) combinations.

[2] A second, although less common, bootstrapping technique used in time series models is called "Moving Blocks." For an AR(*p*) process, select a length *L* that is longer than *p*; *L* is the length of the block. To construct the bootstrap *y** series, randomly select a group of *L* adjacent data points to represent the first *L* observations of the bootstrap sample. In total, you need to select *T/L* of these samples to form a bootstrap sample with *T* observations. The idea of selecting a block is to preserve the time–dependence of the data. However, observations more than *L* apart will be nearly independent. Use this bootstrap sample to estimate the bootstrap coefficients a_0^* and a_1^*.

Bootstrapping Exercise

Monte Carlo experiments allow you to replicate a random process so as to calculate the probability of a particular type of outcome. In a properly designed experiment, you can obtain the entire probability distribution. Consider each of the following:

a. Suppose you toss a coin and a tetrahedron. For the coin, you get 1 point for a tail and 2 points for a head. The faces of the tetrahedron are labeled 1 through 4. For the tetrahedron, you get the number of points shown on the downward face. Your total score equals the number of points received for the coin and the tetrahedron. Of course, it is impossible to have a score of zero or 1. It is straightforward to calculate that the probabilities of scores 3, 4, and 5 equal 0.25 while the probabilities of scores 2 and 6 equal 0.125.

i. Describe how you can design a Monte Carlo experiment to calculate these probabilities.

ii. If your software package allows you to program such an experiment, find out how close the calculated probabilities come to the actual probabilities using 100 replications. How does your answer change if you use 1,000 replications?

b. Replicate the Monte Carlo results that were reported in Section 4 for the *t*-distribution of a unit root.

c. As discussed in Section 4.2**,** Rule 2 of Sims, Stock and Watson (1990) states:

> If the data-generating process contains any deterministic regressors (i.e., an intercept or a time trend) and the estimating equation contains these deterministic regressors, inference on all coefficients can be conducted using a *t*–test or an *F*–test.

Suppose that a series is generated by the equation $\Delta y_t = a_0 + a_2 t + \varepsilon_t$. A researcher unaware of the actual nature of the series estimates the equation in the form $\Delta y_t = a_0 + \rho y_{t-1} + a_2 t + \varepsilon_t$. Since there is a trend and an intercept in the DGP, Rule 2 indicates that it is appropriate to test the null hypothesis $\rho = 0$ using a normal distribution.

i. Perform the following Monte Carlo experiment to verify this result.

First, select values of a_0 and a_2. If the assertion is correct, you must be able to select any nonzero values for these two coefficients. Generate a series of 100 observations using the equation $y_t = a_0 + y_{t-1} + a_2 t + \varepsilon_t$. Let the initial value of the series be a_0.

Second, estimate the series in the form $\Delta y_t = a_0 + \rho y_{t-1} + a_2 t + \varepsilon_t$ and obtain the *t*-statistic for the null hypothesis $\rho = 0$.

Third, repeat the first two steps 2,000 times. Obtain the distribution for the calculated *t*-statistics.

ii. John obtained the following results using 10,000 Monte Carlo replications

First percentile:	–2.38	Ninetieth percentile:	1.29
Fifth percentile:	–1.67	Ninety-fifth percentile:	1.67
Tenth percentile:	–1.30	Ninety-ninth percentile:	2.33
Twenty-fifth:	–0.66		

Explain how John's findings are consistent with the claim that it is appropriate to test the null hypothesis $\rho = 0$ using a normal distribution.

d. In contrast to part *c*, suppose that a series is generated by the equation $\Delta y_t = \varepsilon_t$. A researcher unaware of the actual nature of the series estimates the equation in the form $\Delta y_t = a_0 + \rho y_{t-1} + a_2 t + \varepsilon_t$. Repeat steps 1 to 3 from part c above using $a_0 = a_2 = 0$. How close do your results come to the Dickey–Fuller τ_τ-statistic?

Section 4.4. Determination of the Deterministic Regressors

Sometimes the appropriate null and alternative hypotheses are unclear. As indicated in the text, you do not want to lose power by including a superfluous deterministic regressor in a unit root test. However, omitting a regressor that is actually in the data-generating process leads to a misspecification error. Fortunately, Sims, Stock and Watson (1990) provide a second rule that is helpful in selecting the appropriate set of regressors:

> **Rule 2:** If the data-generating process contains any deterministic regressors (*i.e.*, an intercept or a time trend) and the estimating equation contains these deterministic regressors, inference on all coefficients can be conducted using a *t*-test or an *F*-test. This is because a test involving a single restriction across parameters with different rates of convergence is dominated asymptotically by the parameters with the slowest rates of convergence.

While the proof is beyond the scope of this text, the point is that the nonstandard Dickey–Fuller distributions are needed only when you include deterministic regressors not in the actual data-generating process. Hence, if the DGP is known to contain the deterministic trend term a_2t, the null hypothesis $\gamma = 0$ can be tested using a *t*-distribution if you estimate the model in the form of (4.25). However, if the superfluous deterministic trend is included, there is a substantial loss of power. As such, papers such as Dolado, Jenkinson, and Sosvilla–Rivero (1990) suggest a procedure to test for a unit root when the form of the data-generating process is *completely* unknown. The following is a straightforward modification of the method:

STEP 1: As shown in Figure A2.1, start with the least restrictive of the plausible models (which will generally include a trend and drift) and use the τ_τ statistic to test the null hypothesis $\gamma = 0$. Thus, in the most general case, you estimate the model in the form of (4.25) so that $\Delta y_t = a_0 + \gamma y_{t-1} + a_2t + \Sigma\beta_i\Delta y_{t-i}$. Unit root tests have low power to reject the null hypothesis; hence, if the null hypothesis of a unit root is *rejected*, there is no need to proceed. Conclude that the $\{y_t\}$ sequence does not contain a unit root.

STEP 2: If the null hypothesis is *not rejected*, the next step is to determine whether the trend belongs in the estimating equation. Towards this end, you test the null hypothesis $a_2 = \gamma = 0$ using the ϕ_3 statistic. If you do not reject the null hypothesis, assume the absence of a trend and proceed to Step 3.

If you have reached this point, it is because the τ_τ test indicates that there is a unit root and the ϕ_3 test indicates that γ and/or a_2 differs from zero. As such, it would seem that there is a unit root and a trend. You can gain additional support for this result by assuming that there is a unit root and estimating $\Delta y_t = a_0 + a_2t + \Sigma\beta_i\Delta y_{t-i}$. Since there are no $I(1)$ regressors in this specification, the test for the null hypothesis $a_2 = 0$ can be conducted using a standard *t*-test. If you conclude $a_2 = 0$, go to Step 3 since it does not appear that there is a trend. If you find that $a_2 \neq 0$, use (4.25) to test the null hypothesis $\gamma = 0$ using a *t*-distribution. Given that the trend belongs in the regression equation, Rule 2 indicates that the test for $\gamma = 0$ can be conducted using a *t*-distribution. Go no further; if $\gamma \neq 0$, conclude that the sequence is trend stationary. If $\gamma = 0$, conclude that there is a unit root (and the $\{y_t\}$ sequence contains a quadratic trend).

STEP 3: Estimate the model without the trend [i.e., estimate a model in the form of (4.24)]. Test for the presence of a unit root using the τ_μ statistic. If the null is rejected, conclude that the model does not contain a unit root. If the null hypothesis of a unit root is not rejected, test for the significance of the intercept by testing the hypothesis $a_0 = \gamma = 0$ using the ϕ_1 statistic. If you do not reject the null hypothesis $a_0 = \gamma = 0$, assume that the intercept is zero and proceed to Step 4. Otherwise, estimate $\Delta y_t = a_0 + \Sigma\beta_i\Delta y_{t-i}$ and test whether $a_0 = 0$ using a *t*-distribution. If you find $a_0 = 0$, proceed to Step 4. Otherwise, conclude that the process contains an intercept term. In accord with Rule 2, you can use a *t*-distribution to test whether $\gamma = 0$ in the regression $\Delta y_t = a_0 + \gamma y_{t-1} + \Sigma\beta_i\Delta y_{t-i}$. If the null hypothesis of a unit root is rejected, conclude that the $\{y_t\}$ sequence stationary around a non-zero mean. Otherwise, conclude that the $\{y_t\}$ sequence contains a unit root and a drift.

STEP 4: Estimate a model without the trend or drift; i.e., estimate a model in the form of (4.23). Use τ to test for the presence of a unit root. If the null hypothesis of a unit root is rejected, conclude that the $\{y_t\}$ sequence does not contain a unit root. Otherwise, conclude that the $\{y_t\}$ sequence contains a unit root.

[Insert Figure A2.1 Here]

Remember, no procedure can be expected to work well if it is used in a completely mechanical fashion. Plotting the data is usually an important indicator of the presence of deterministic regressors. The interest rate series shown in Figure 4.2 can hardly be said to contain a deterministic trend. Moreover, theoretical considerations might suggest the appropriate regressors. The efficient market hypothesis is inconsistent with the presence of a deterministic trend in an asset’s return. Similarly, in testing for PPP you should not begin using a deterministic time trend. However, the procedure is a sensible way to test for unit roots when the form of the data-generating process is completely unknown.

GDP and Unit Roots

Although the methodology outlined in Figure A2.1 can be very useful, it does have its problems. Each step in the procedure involves a test that is conditioned on all the previous tests being correct; the significance level of each of the cascading tests is impossible to ascertain.

The procedure and its inherent dangers are nicely illustrated by trying to determine if the real GDP data shown in Figure 4.1 has a unit root. It is a good idea to replicate the results reported below using the data in RGDP.XLS. Using quarterly data over the 1947*Q*1 – 2012*Q*4 period, the correlogram of the logarithm of real GDP exhibits slow decay. At the end of Section 2, the logarithmic first difference of the series was estimated as

$$\Delta lrgdp_t = \underset{(6.80)}{0.0049} + \underset{(6.44)}{0.3706}\Delta lrgdp_{t-1} \qquad \text{(A2.1)}$$

The model is well estimated in that the residuals appear to be white noise and all coefficients are of high quality. For our purposes, the interesting point is that the $\{\Delta lrgdp_t\}$ series appears to be a stationary process. Integrating suggests that $\{lrgdp_t\}$ has a stochastic and a deterministic trend. The issue here is to determine whether it was appropriate to difference the log of real GDP. Towards this end, consider the augmented Dickey–Fuller equation with *t*-statistics in parentheses:

$$\Delta lrgdp_t = \underset{(1.58)}{0.1248} + \underset{(1.31)}{0.0001t} - \underset{(-1.49)}{0.0156 lrgdp_{t-1}} + \underset{(6.26)}{0.3663\Delta lrgdp_{t-1}} \qquad \text{(A2.2)}$$

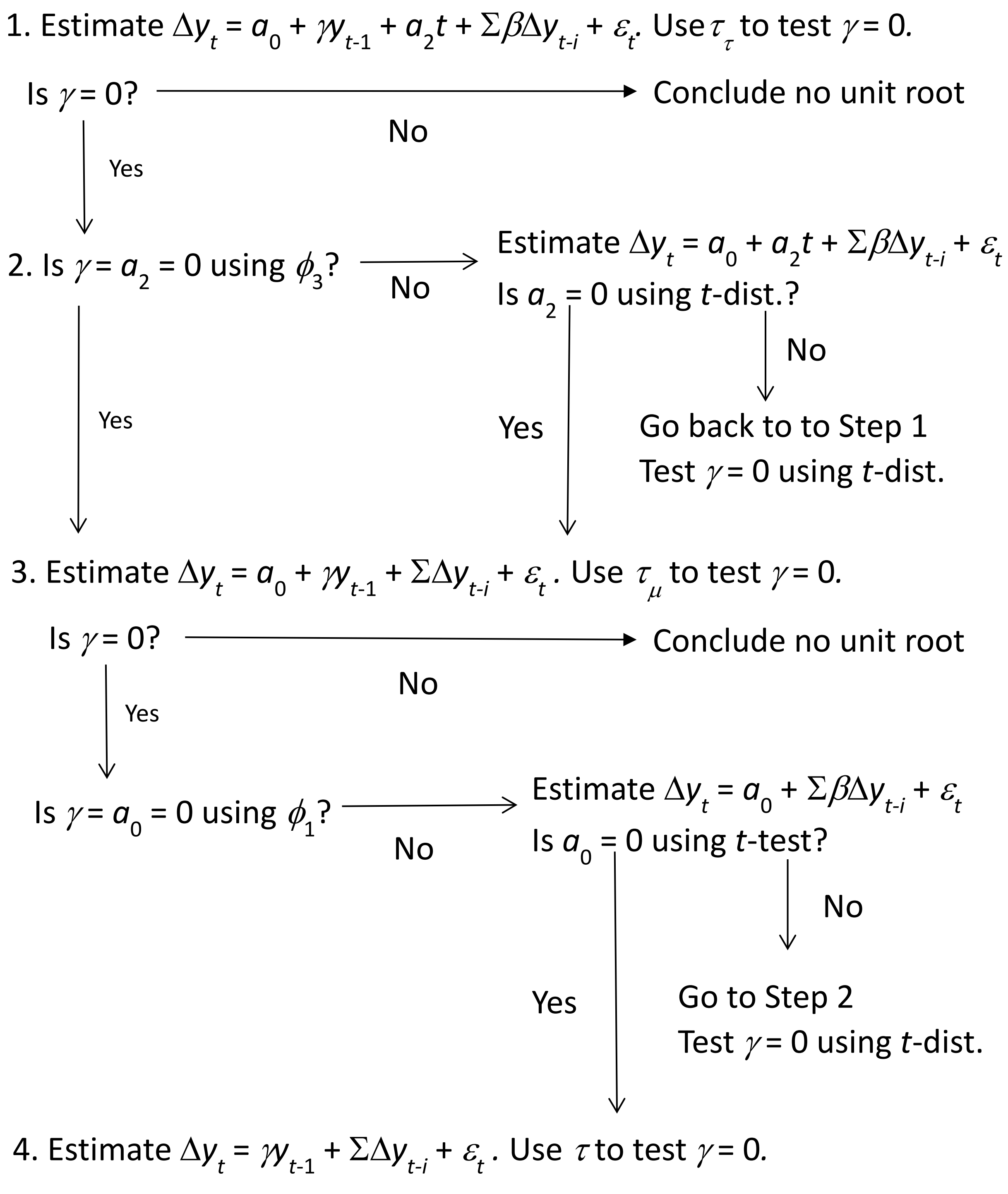
As in Step 1, we estimate the least restrictive model; as such, (A2.2) contains an intercept and a deterministic trend. The point estimates of (A2.2) suggest that the real GDP is trend stationary. However, the issue is to formally test the statistical significance of the null hypothesis $\gamma = 0$. The *t*-statistic for the null hypothesis $\gamma = 0$ is –1.49. Critical values with exactly 244 usable observations are not reported in the Dickey–Fuller table. However, with 262 observations, the critical value of τ_τ at the 10 percent and 5 percent significance levels are –3.13 and –3.43, respectively. At the 5 and 10 percent levels, we cannot reject the null of a unit root. However, the power of the test may have been reduced due to the presence of an unnecessary time trend and/or drift term. In Step 2, we use the ϕ_3 statistic to test the joint hypothesis $a_2 = \gamma = 0$. The sample value of *F* for the restriction is 2.97. Since the critical value of ϕ_3 is 6.34 at the 5 percent significance level, it is possible to conclude that the restriction $a_2 = \gamma = 0$ is not binding. Thus, we can proceed to Step 3 and estimate the model without the trend. Consider the following equation:

$$\Delta lrgdp_t = \underset{(2.64)}{0.0215} - \underset{(-2.05)}{0.0019\, lrgdp_{t-1}} + \underset{(6.12)}{0.3539\Delta lrgdp_{t-1}} \qquad \text{(A2.3)}$$

In (A2.3), the *t*-statistic for the null hypothesis $\gamma = 0$ is –2.05. Since the critical value of the τ_μ statistic is –2.88 at the 5 percent significance level, the null hypothesis of a unit root is not rejected at conventional significance levels. Again, the power of this test will have been reduced if the drift term does not belong in the model. To test for the presence of the drift, use the ϕ_1 statistic. The sample value of *F* for the restriction $a_0 = \gamma = 0$ is 25.49. Since 25.49 exceeds the critical value of 4.63, we conclude that the restriction is binding. Either $\gamma \neq 0$ (so there is not a unit root), $a_0 \neq 0$ (so there is not an intercept term), or both γ and a_0 differ from zero.

In reality, any sensible researcher would stop at this point. However, since the point of this section is to illustrate the procedure, test for the presence of the drift using (A2.1). Given the *t*-statistic of 6.80, we reject $a_0 = 0$ using a *t*-distribution. Hence, (A2.3) is our final testing regression. Since we are sure that the intercept belongs in the model, Rule 2 indicates that we can test the null hypothesis $\gamma = 0$ using a *t*-distribution. This is where the methodology runs into a bit of trouble. Given that the *t*-statistic for the coefficient of $lrgdp_{t-1}$ is –2.05, at the 5% level we can conclude that the series does not have a unit root. Nevertheless, the result is nonsense since the implication is that $lrgdp_t$ is stationary around a time invariant mean.

Figure A2.1: A Procedure to Test for Unit Roots



1. Estimate $\Delta y_t = a_0 + \gamma y_{t\text{-}1} + a_2 t + \Sigma\beta\Delta y_{t\text{-}i} + \varepsilon_t$. Use τ_τ to test $\gamma = 0$.

Is $\gamma = 0$? No → Conclude no unit root

Yes ↓

2. Is $\gamma = a_2 = 0$ using ϕ_3? No → Estimate $\Delta y_t = a_0 + a_2 t + \Sigma\beta\Delta y_{t\text{-}i} + \varepsilon_t$

Is $a_2 = 0$ using t-dist.?

No ↓ Go back to to Step 1

Test $\gamma = 0$ using t-dist.

Yes ↓

3. Estimate $\Delta y_t = a_0 + \gamma y_{t\text{-}1} + \Sigma\Delta y_{t\text{-}i} + \varepsilon_t$. Use τ_μ to test $\gamma = 0$.

Is $\gamma = 0$? No → Conclude no unit root

Yes ↓

Is $\gamma = a_0 = 0$ using ϕ_1? No → Estimate $\Delta y_t = a_0 + \Sigma\beta\Delta y_{t\text{-}i} + \varepsilon_t$

Is $a_0 = 0$ using t-test?

No ↓ Go to Step 2

Test $\gamma = 0$ using t-dist.

Yes ↓

4. Estimate $\Delta y_t = \gamma y_{t\text{-}1} + \Sigma\Delta y_{t\text{-}i} + \varepsilon_t$. Use τ to test $\gamma = 0$.

Section 4.5 The Unobserved Components Decomposition

The Beveridge and Nelson (1981) decomposition has proven especially useful in that it provides a straightforward method to decompose any ARIMA(p, 1, q) process into a temporary and a permanent component. However, it is important to note that *the Beveridge and Nelson decomposition is not unique.* Equations (4.53) and (4.54) provide an example in which the Beveridge and Nelson decomposition forces the innovation in the trend and the stationary components to have a perfect negative correlation.

In fact, this result applies to the more general ARIMA(p, 1, q) model. Obtaining the irregular component as the difference between y_t and its trend forces the correlation coefficient between the innovations to equal –1. However, there is no reason to constrain the two innovations in the two components to be perfectly correlated. To illustrate the point, consider the random-walk plus noise plus drift (i.e., the trend plus noise model) introduced in Section 1:

$$y_t = y_0 + a_0 t + \sum_{i=1}^{t} \varepsilon_i + \eta_t$$

The deterministic portion of the trend is $y_0 + a_0 t$, the stochastic trend is $\Sigma\varepsilon_i$ and the noise component is η_t. The stochastic trend and the noise components are uncorrelated since we assumed that $E(\varepsilon_i \eta_t) = 0$. Thus, the Beveridge and Nelson methodology would incorrectly identify the trend and the irregular components because it would force the two innovations to have a correlation of –1.

Now consider an alternative identification scheme that relies on an **unobserved components** (UC) model. A UC model posits that a series is comprised of several distinct, but unobservable components (such as the ε_t and η_t shocks in the random-walk plus noise model). The goal is identify the shocks by matching the moments of the UC specification to those from an estimated ARIMA model. In Section 2, the trend plus noise model was shown to have an equivalent ARIMA(0, 1, 1) representation such that

$$E\Delta y_t = 0; \qquad \text{var}(\Delta y_t) = \sigma^2 + 2\sigma_\eta^2 \ , \quad \text{and} \quad \text{cov}(\Delta y_t, \Delta y_{t-1}) = -\sigma_\eta^2 \qquad (4.58)$$

Hence, it is possible to represent the first difference of the trend plus noise model as the MA(1) process:

$$\Delta y_t = a_0 + e_t + \beta_1 e_{t-1} \qquad (4.59)$$

where e_t is an independent white-noise disturbance. The notation e_t is designed to indicate that shocks to Δy_t come from two sources: ε_t and η_τ. The problem is to decompose the estimated values of $\{e_t\}$ into these two source components.

In this instance, it is possible to recover, or *identify*, the individual $\{\varepsilon_t\}$ and $\{\eta_t\}$ shocks from the estimation of (4.59). The appropriate use of the Box–Jenkins methodology will yield estimates of a_0, β_1 and the elements of the $\{e_t\}$ sequence. Given the coefficient estimates, it is possible to form

$$\text{var}(\Delta y_t) = \text{var}(e_t + \beta_1 e_{t-1}) = (1 + \beta_1)^2\text{var}(e_t)$$

and

$$\text{cov}(\Delta y_t, \Delta y_{t-1}) = \beta_1\text{var}(e_t)$$

However, these estimates of the variance and covariance are not arbitrary; for (4.59) to satisfy the restrictions of (4.58) it must be the case that

$$(1 + \beta_1)^2\text{var}(e_t) = \sigma^2 + 2\sigma_\eta^2$$

and

$$\beta_1\text{var}(e_t) = -\sigma_\eta^2$$

Now that we have estimated β_1 and var(e_t), it is possible to recover σ^2 and σ_η^2 from the data. The individual values of the $\{\varepsilon_t\}$ and $\{\eta_t\}$ sequences can be recovered as well. From the forecast function, $E_t y_{t+1} = y_t + a_0 - \eta_t$. Hence, it is possible to use one-step-ahead forecasts from (4.59) to find $E_t\Delta y_{t+1} = a_0 + \beta_1 e_t$, so that $E_t y_{t+1} = y_t + a_0 + \beta_1 e_t$. Since the two forecasts must be equivalent, it follows that

$$\beta_1 e_t = -\eta_t$$

Thus, the estimated values of $\beta_1 e_t$ can be used to identify the entire $\{\eta_t\}$ sequence. Given $\{e_t\}$ and $\{\eta_t\}$, the values of $\{\varepsilon_t\}$ can be obtained from $\Delta y_t = a_0 + \varepsilon_t + \Delta\eta_t$. For each value of t, form $\varepsilon_t = \Delta y_t - a_0 - \Delta\eta_t$ using the known values of Δy_t and the estimated values of a_0 and $\Delta\eta_t$.

The point is that it is possible to decompose a series such that the correlation between the trend and irregular components is zero. The example illustrates an especially important point. To use this method to decompose a series into a random walk plus drift and a stationary irregular component, it is necessary to specify the correlation coefficient between innovations in the trend and the irregular components. We have seen two ways to decompose an ARIMA(0, 1, 1) model. The Beveridge and Nelson technique assumes that are trend and cycle are perfectly correlated, while the UC decomposition adds the restriction

$$E\varepsilon_t\eta_t = 0$$

In fact, the correlation coefficient between the two components can be any number in the interval –1 to +1. Without the extra restriction concerning the correlation between the innovations, the trend and stationary components cannot be identified; in a sense, we are an equation short. The assumption that ε_t and η_t are uncorrelated places restrictions on the autoregressive and moving average coefficients of Δy_t. For example, in the pure random walk plus noise model, β_1 must be negative. To avoid estimating a constrained ARIMA model, Watson (1986) estimates the trend and the irregular terms as unobserved components

Watson (1986) decomposes the logarithm of GNP using the Beveridge and Nelson decomposition and using the UC method. Using a Beveridge and Nelson decomposition, he estimates the following ARIMA(1, 1, 0) model (with standard errors in parenthesis):

$$\underset{}{\Delta y_t = \underset{(0.001)}{0.005} + \underset{(0.077)}{0.406}\Delta y_{t-1} + \varepsilon_t} \qquad \text{var}(\varepsilon_t) = 0.0103^2$$

Using the UC method such that the innovations in the trend and irregular components are uncorrelated, Watson estimates GNP as the sum of a trend (τ_t) plus a cyclical term (c_t). The trend is simple a random walk plus drift and the cyclical component is an AR(2) process.

$$\underset{(0.001)}{\tau_t = 0.008} + \tau_{t-1} + \varepsilon_\tau \qquad \text{var}(\varepsilon_t) = 0.0057^2$$

$$c_t = \underset{(0.121)}{1.501}c_{t-1} - \underset{(0.125)}{0.577}c_{t-2} \qquad \text{var}(\eta_t) = 0.0076^2$$

The short-term forecasts of the two models are quite similar. The standard error of the one-step-ahead forecast of UC model is slightly smaller than that from the Beveridge and Nelson decomposition: $(\sigma^2 + \sigma_\eta{}^2)^{1/2} \cong 0.0095$ is slightly smaller than 0.0103. However, the long-run properties of the two models are quite different. For example, writing $\Delta y_t = (0.005 + \varepsilon_t)/(1-0.406L)$ yields the impulse response function using Beveridge and Nelson decomposition. The sum of the coefficients for this impulse response function is 1.68. Hence, a one-unit innovation will eventually increase log(GNP) by a full 1.68 units. Since all coefficients are positive, following the initial shock, y_t steadily increases to its new level. In contrast, the sum of the impulse response coefficients in the UC model is about 0.57. All coefficients beginning with lag 4 are negative. As such, a one-unit innovation in y_t has a larger effect in the short run than in the long run. Most importantly, the Beveridge and Nelson cycle has a small amplitude and is less persistent than the UC cycle.

Morley, Nelson and Zivot (2003) update Watson’s (1986) study and find similar results using data through 1998*Q*2. They also show how to use the Kalman filter to estimate the correlation between ε_t and η_t. This is a clear advantage over imposing a particular value for $E\varepsilon_t\eta_t$. It turns out that the estimated correlation is –0.9062 so that the Beveridge and Nelson cycle quite reasonable. It is going too far afield to explain the Kalman filter and state-space modeling here.

Section 4.6: Phillips-Perron Test Statistics

Most people now use the Dickey-Fuller test in conjunction with the MAIC when a large and negative MA term is suspected to be in the data generating process. However, since the Phillips-Perron (1988) test is still popular in such circumstances this modification of the Dickey-Fuller test merits some discussion.

The distribution theory supporting the Dickey-Fuller tests assumes that the errors are statistically independent and have a constant variance. In using test, care must be taken to ensure that these assumptions are not violated. Phillips and Perron (1988) developed a generalization of the Dickey-Fuller procedure which allows for fairly mild assumptions concerning the distribution of the errors.

The Phillips-Perron (1988) statistics modify the Dickey-Fuller *t*-statistics to account for heterogeneity in the error process. The Phillips-Perron (1988) test was a popular unit root test for the case of a large and negative moving average term in the data generating process. Suppose that we observe the first 1, 2, ... , *T* realizations of the $\{y_t\}$ sequence and estimate the regression equation:

$$y_t = \mu + \beta(t - T/2) + \alpha y_{t-1} + \mu_t$$

where μ, β, and α are the conventional least squares regression coefficients. The error term is denoted by μ_t to indicate that the series may be serial correlated. Phillips and Perron (1984) derive test statistics for the regression coefficients under the null hypothesis that the data is generated by:

$$y_t = y_{t-1} + \mu_t$$

Do not be deceived by the apparent simplicity of these two equations. In actually, they are far more general than the type of data generating process allowable by the Dickey-Fuller procedure. For example, suppose that the $\{\mu_t\}$ sequence is generated by the autoregressive process $\mu_t = [C(L)/B(L)]\varepsilon_t$ where $B(L)$ and $C(L)$ are polynomials in the lag operator. Given this form of the error process, we can write the first equation in the form used in the Dickey-Fuller tests; i.e.,

$$B(L)y_t = \mu B(L) + B(L)\beta(t - T/2) + \alpha B(L)y_{t-1} + C(L)\varepsilon_t.$$

Thus, the Phillips-Perron procedure can be applied to ARIMA order processes in the same way as the Dickey-Fuller tests. The difference between the two tests is that there is *no* requirement that the disturbance term be serially uncorrelated or homogeneous. Instead, the Phillips-Perron test allows the disturbances to be weakly dependent and heterogeneously distributed.

Let t_μ, t_α, and t_β be the usual *t*-test statistics for the null hypotheses $\mu = 0$, $\alpha = 1$, and $\beta = 0$, respectively. In essence, Phillips and Perron (1988) use robust standard errors so as to modify the

Dickey-Fuller statistics to allow for weakly dependent errors. The expressions are extremely complex; to actually derive them would take us far beyond the scope of this book. However, many statistical time-series software packages now calculate these statistics so they are directly available. The modified statistics are:

$$Z(t_\alpha) = (S/\sigma_{T\omega})t_\alpha - (T^3/4\sqrt{3}D^{1/2}\sigma_{T\omega})(\sigma_{T\omega}^2 - S^2)$$

$$Z(t_\mu) = (S/\sigma_{T\omega})t_\mu - (T^3/24D^{1/2}E_x\sigma_{T\omega})(\sigma_{T\omega}^2 - S^2)(T^{-3/2}\Sigma y_{t-1})$$

$$Z(t_\beta) = (S/\sigma_{T\omega})t_\beta - (T^3/2D^{1/2}E\sigma_{T\omega})[T^{-2}\Sigma(y_{t-1} - \overline{y}_{-1})^2]^{-1/2}(\sigma_{T\omega}^2 - S^2)(0.5T^{-3/2}\Sigma y_{t-1} - T^{-5/2}\Sigma t y_{t-1}])$$

where $D = \det(x'x)$, the determinant of the regressor matrix x,

$$E_X = \left[T^{-6}D_X + (1/12)\left(T^{-3/2}\Sigma y_{t-1}\right)^2\right]^{1/2} 1$$

*S*2 is the standard error of the regression,

$$\sigma_{T\omega}^2 = T^{-1}\sum_1^T \mu_l^2 + 2T^{-1}\sum_{s=l}^T \sum_{t=s+1}^T \mu_t \mu_{t-s} \; 3$$

and ω is the number of estimated autocorrelations.

Note that S^2 and $\sigma_{T\omega}^2$ are consistent estimates of $\sigma_\mu^2 = \lim \mathrm{E}\left(\mathrm{u}_\mathrm{t}^2\right)$ and $\sigma^2 = \lim \mathrm{E}\left(\mathrm{T}^{-1}\mathrm{S}_\mathrm{T}^2\right)$ where $S_T = \sum \mu_T$ and all summations run over t. For the joint hypothesis $\beta = 0$ and $\alpha = 1$, use their $Z(\varphi_3)$ statistic. Fortunately, many software packages calculate these statistics. The critical values for the Phillips-Perron statistics are precisely those given for the Dickey-Fuller tests. For example, the critical values for $Z(t_\alpha)$ and $Z(t_\beta)$ are those given in the Dickey-Fuller tables under the headings τ_μ and τ_τ, respectively. The critical values of $Z(\varphi_3)$ are given by the Dickey-Fuller φ_3 statistic.

Foreign Exchange Market Efficiency. Corbae and Ouliaris (1986) used Phillips-Perron tests to determine whether exchange rates follow a random walk and whether the return to forward exchange market speculation contains a unit root. Denote the spot dollar price of foreign exchange on day *t* as s_t. An individual at *t* can also buy or sell foreign exchange forward. A 90-day forward contract requires, that on day *t*+90, the individual take delivery (or make payment) of a specified amount of foreign exchange in return for a specified amount of dollars. Let f_t denote the 90-day forward market price of foreign exchange purchased on day *t*. On day t, suppose that an individual speculator buys forward pounds at the price: f_t = \$2.00/pound. Thus, in 90 days the individual is obligated to provide \$200,000 in return for £100,000. Of course, the agent may choose to immediately sell these pounds on the spot market. If on day *t*+90, the spot price happens to be s_{t+90} = \$2.01/pound, the individual can sell the £100,000 for \$201,000; ignoring any transactions costs, the individual earns a profit of \$1000. In general, the profit on such a transaction will be $s_{t+90} - f_t$

multiplied by the number of pounds transacted. (Note that profits will be negative if $s_{t+90} < f_t$). Of course, it is possible to speculate by selling forward pounds too. An individual selling 90-day forward pounds on day t will be able to buy them on the spot market at s_{t+90}. Here, profits will be $f_t - s_{t+90}$ multiplied by the number of pounds transacted. The efficient market hypothesis maintains that the expected profit or loss from such speculative behavior must be zero. Let $E_t s_{t+90}$ denote the expectation of the spot rate for day t+90 conditioned on the information available on day t. Since we actually know f_t on day t, the efficient market hypothesis for forward exchange market speculation can be written as:

$$E_t s_{t+90} = f_t.$$

or:

$$s_{t+90} - f_t = p_t.$$

where: p_t = per unit profit from speculation; and $E_t p_t = 0$.

Thus, the efficient market hypothesis requires that for any time period t, the 90-day forward rate (i.e., f_t) be an unbiased estimator of the spot rate 90 days from t. Suppose that a researcher collected weekly data of spot and forward exchange rates. The data set would consist for the forward rates f_t, f_{t+7}, f_{t+14}, ... and the spot rates s_t, s_{t+7}, s_{t+14}, ... Using these exchange rates, it is possible to construct the sequence: $s_{t+90} - f_t = p_t$, $s_{t+7+90} - f_{t+7} = p_{t+7}$, $s_{t+14+90} - f_{t+14} = p_{t+14}$, ... Normalize the time period to 1 week so that $y_1 = p_t$, $y_2 = p_{t+7}$, $y_3 = p_{t+14}$, ... and consider the regression equation:

$$y_t = a_0 + a_1 y_{t-1} + a_2 t + \mu_t$$

The efficient market hypothesis asserts that ex ante expected profit must equal zero; hence, using quarterly data it should be the case that $a_0 = a_1 = a_2 = 0$. However, the way that the data set was constructed means that the residuals will be correlated. As Corbae and Ouliaris (1986) point out, suppose that there is relevant exchange market "news" at date T. Agents will incorporate this news into all forward contracts signed in periods subsequent to T. However, the realized returns for all pre-existing contracts will be affected by the news. Since there are approximately 13 weeks in a 90 day period, we can expect the μ_t sequence to be an MA(12) process. Although ex ante expected returns may be zero, the ex post returns from speculation at t will be correlated with the returns from those engaging forward contracts at weeks t+1 through t+12.

Meese and Singleton (1982) assumed white noise disturbances in using a Dickey-Fuller test to study the returns from forward market speculation. One surprising result was that the return from forward speculation in the Swiss franc contained a unit root. This finding contradicts the efficient market hypothesis since it implies the existence of a permanent component in the sequence of returns. However, the assumption of white noise disturbances is inappropriate if the $\{\mu_t\}$ sequence is an MA(12) process. Instead, Corbae and Ouliaris use the more appropriate Phillips-Perron procedure to analyze foreign exchange market efficiency; some of their results are contained in the table below

First consider the test for the unit root hypothesis (i.e., $a_1 = 1$). All estimated values of a_1 exceed 0.9; the first-order autocorrelation of the returns from speculation appear to be quite high. Yet, given the small standard errors, all estimated values are over four standard deviations from unity. At the 5% significance level, the critical values for a test of $a_1 = 1$, is -3.43. Note that this critical value is the Dickey-Fuller τ_τ statistic with 250 observations. Hence, as opposed to Meese and Singleton (1982), Corbae and Ouliaris are able to reject the null of a unit root in all series examined. Thus, shocks to the return from forward exchange market speculation do not have permanent effects.

A second necessary condition for the efficient market hypothesis to hold is that the intercept term a_0 equal zero. A non-zero intercept term suggests a predictable gap between the forward rate and the spot rate in the future. If $a_0 \neq 0$, on average, there are unexploited profit opportunities. It may be that agents are risk averse or that profit maximizing speculators are not fully utilizing all available information in determining their forward exchange positions. In absolute value, all of the Z-statistics are **less than** the critical value so that Corbae and Ouliaris cannot reject the null $a_0 = 0$. In the same way, they are not able to reject the null hypothesis of no deterministic time trend (i.e., that $a_2 = 0$). The calculated $Z(t_\beta)$ statistics indicate that the estimated coefficients of the time trend are never more than 1.50 standard errors from zero.

Returns To Forward Speculation

	a_0	a_1	a_2
Switzerland	-0.117E-2 (0.106E-2) $Z(t_\mu)$= -1.28	0.941 (0.159E-1) $Z(t_\alpha)$ = -4.06	-0.111E-4 (0.834E-5) $Z(t_\beta)$ = -1.07
Canada	-0.651E-3 (0.409E-3) $Z(t_\mu)$= -1.73	0.907 (0.191E-1) $Z(t_\alpha)$ = -5.45	0.116E-5 (0.298E-5) $Z(t_\beta)$ = -1.42
United Kingdom	-0.779E-3 (0.903E-3) $Z(t_\mu)$= -.995	0.937 (0.163E-1) $Z(t_\alpha)$ = -4.69	-0.132E-4 (0.720E-5) $Z(t_\beta)$ = -1.50

Notes: Standard errors are in parenthesis and $Z(t_\mu)$ and $Z(t_\beta)$ are the Phillips-Perron adjusted t-statistics for the hypothesis that $a_0 = 0$ and $a_2 = 0$, respectively. $Z(t_\alpha)$ is the Phillips-Perron adjusted *t*-statistic for the hypothesis that $a_1 = 1$.

At this point, you might wonder whether it would be possible to perform the same sort of analysis using an Augmented Dickey-Fuller (ADF) test. After all, Said and Dickey (1984) showed that the ADF test can be used when the error process is a moving average. The desirable feature of the Phillips-Perron test is that it allows for a weaker set of assumptions concerning the error process. Also, Monte Carlo studies find that the Phillips-Perron test has greater power reject a false null hypothesis of a unit root. However, there is a cost entailed with the use of weaker assumptions. Monte Carlo studies have also shown that in the presence of *negative* moving average terms, the Phillips-Perron test tends to reject the null of a unit root whether or not the actual data generating process contains a negative unit root. It is preferable to use the ADF test when the true model contains negative moving average terms and to use the Phillips-Perron test when the true model contains positive moving average terms.

In practice, the choice of the most appropriate test can be difficult since you never know the true data generating process. A safe choice is to use both types of unit roots tests. If they reinforce each other, you can have confidence in the results. Sometimes economic theory will be helpful in that it suggests the most appropriate test. In the Corbae and Ouliaris example, excess returns should be positively correlated; hence, the Phillips-Perron test is a reasonable choice.

CHAPTER 5

ENDNOTES TO CHAPTER 5

1. In the identification process, we are primarily interested in the shape, not the height, of the cross-correlation function. It is useful to standardize the covariance by dividing through by σ_z^2; the shape of the correlogram is proportional to the standardized covariance. The text follows the procedure used by most software packages by plotting the standardized cross-covariances.

2. We were able to obtain quarterly data from 1970Q1 to 1988Q4 for Austria, Canada, Denmark, Finland, France, West Germany, Greece, Italy, the Netherlands, Norway, the United Kingdom, and the United States. The International Monetary Fund's *Balance of Payments Statistics* reports all data in Special Drawing Rights (SDR). The dependent variables were the logarithms of each nation's revenues divided by the sum of the revenues for all twelve countries.

3. Tourism is highly seasonal; we tried several alternative deseasonalization techniques. The results reported here were obtained using seasonal dummy variables. Hence, y_t represents the deseasonalized logarithmic share of tourism receipts. The published paper reports results using quarterly differencing. Using either type of deseasonalization, the final results were similar.

4. It is easily verified that this representation implies that $\rho_{12} = 0.8$. By definition, the correlation coefficient ρ_{12} is defined to be $\sigma_{12}/(\sigma_1\sigma_2)$ and the covariance is $Ee_{1t}e_{2t} = \sigma_{12}$. If we use the numbers in the example, $Ee_{1t}e_{2t} = E[\varepsilon_{zt}(\varepsilon_{yt} + 0.8\varepsilon_{zt})] = 0.8\sigma_z^2$. Since the decomposition equates var(e_{2t}) with σ_z^2, it follows that $\rho_{12} = 0.8$ if $\sigma_1^2 = \sigma_2^2$.

5. Note that γ_1 cannot equal zero if $\{y_t\}$ is $I(1)$. If $\gamma_1 = 0$, $y_t = a_{12}\Delta y_{t-1} + b_{12}\Delta z_{t-1} + \varepsilon_t$, the equation is unbalanced. The left-hand side contains the $I(1)$ variable y_t while the right-hand side contains only the three stationary variables Δy_t, Δz_t, and ε_t.

6. Since a key assumption of the model of the technique is that $E(\varepsilon_{1t}\varepsilon_{2t}) = 0$, you might wonder how it is possible to assume that aggregate demand and supply shocks are independent. After all, if the stabilization authorities follow a feedback rule, aggregate demand will change in response to aggregate supply shocks. The key to understanding this apparent contradiction is that ε_{1t} is intended to be the orthogonalized portion of the demand shock, i.e., the portion of the demand shock that does not change in response to aggregate supply. Cover, Enders and Hueng (2006) and Enders and Hurn (2007) show how this assumption can be relaxed.

7. Since two of the restrictions contain squared terms, there will be a positive value and an equal but opposite negative value for some of the coefficients. In Blanchard and Quah's example, if c11(0) is positive, positive demand shocks have a positive effect on output, and if c11(0) is negative, the positive shock has a negative effect on output. Taylor (2003) considers the problem of selecting among the alternative sets of solutions.

EXAMPLE OF A SIMPLE VAR: TERRORISM AND TOURISM IN SPAIN

In Enders and Sandler (1991), we used the VAR methodology to estimate the impact of terrorism on tourism in Spain during the period from 1970 to 1988. Most transnational terrorist incidents in Spain during this period were perpetrated by left-wing groups, which included the Anti-Fascist Resistance Group of October 1 (GRAPO), the ETA, the now defunct International Revolutionary Armed Front (FRAP), and Iraultza. Most incidents are attributed to the ETA (Basque Fatherland and Liberty) and its splinter groups, such as the Autonomous Anti-Capitalist Commandos (CAA). Rightwing terrorist groups included the Anti-Terrorist Liberation Group (GAL), the Anti-Terrorism ETA, and the Warriors of Christ the King. Catalan independence groups, such as Free Land (Terra Lliure) and Catalan Socialist Party for National Liberation, were active in the late 1980s and often targeted U.S. businesses.

The transfer function model of Section 3 may not be appropriate because of feedback between terrorism and tourism. If high levels of tourism induce terrorist activities, the basic assumption of the transfer function methodology is violated. In fact, there is some evidence that the terrorist organizations in Spain target tourist hotels in the summer season. Since increases in tourism may generate terrorist acts, the VAR methodology allows us to examine the reactions of tourists to terrorism *and* the reactions of terrorists to tourism. We can gain some additional insights into the interrelation between the two series by performing Granger causality tests both of terrorism on tourism and of tourism on terrorism. Impulse-response analysis can quantify and graphically depict the time path of the effects of a typical terrorist incident on tourism.

We assembled a time series of all publicly available transnational terrorist incidents that took place in Spain from 1970 through 1988. In total, there are 228 months of observation in the time series; each observation is the number of terrorist incidents occurring in that month. The tourism data are taken from various issues of the National Statistics Institute’s (Estadistic Institute Nacional) quarterly reports. In particular, we assemble a time series of the number of foreign tourists per month in Spain for the 1970–1988 period.

Empirical Methodology

Our basic methodology involves estimating tourism and terrorism in a vector autoregression (VAR) framework. Consider the following system of equations:

$$n_t = \alpha_{10} + A_{11}(L)n_{t-1} + A_{12}(L)i_{t-1} + e_{1t} \tag{5.47}$$

$$i_t = \alpha_{20} + A_{21}(L)n_{t-1} + A_{22}(L)i_{t-1} + e_{2t} \tag{5.48}$$

where n_t = the number of tourists visiting Spain during time period t

i_t = the number of transnational terrorist incidents in Spain during t

α_{i0} = the vectors containing a constant, eleven seasonal (monthly) dummy variables, and a time trend

A_{ij} = the polynomials in the lag operator L

e_{it} = independent and identically distributed disturbance terms such that $E(e_{1t}e_{2t})$ is not necessarily zero

Although Sims (1980) recommended against the use of a deterministic time trend, we decided not to heed this advice. We experimented with several alternative ways to model the series; the model including the time trend had the best diagnostic statistics. Other variants included differencing (5.47) and (5.48) and simply eliminating the trend and letting the random walk plus drift terms capture any nonstationary behavior. We were also concerned that the number of incidents had a large number of zeroes (and could not be negative), so that the normality assumption was violated.

The polynomials $A_{12}(L)$ and $A_{21}(L)$ in (5.47) and (5.48) are of particular interest. If all of the coefficients of $A_{21}(L)$ are zero, then knowledge of the tourism series does not reduce the forecast error variance of terrorist incidents. Formally, tourism would not Granger cause terrorism. Unless there is a contemporaneous response of terrorism to tourism, the terrorism series evolves independently of tourism. In the same way, if all of the coefficients of $A_{12}(L)$ are zero, then terrorism does not Granger cause tourism. The absence of a statistically significant contemporaneous correlation of the error terms would then imply that terrorism cannot affect tourism. If, instead, any of the coefficients in these polynomials differ from zero, there are interactions between the two series. In the case of negative coefficients of $A_{12}(L)$, terrorism would have a negative effect on the number of foreign tourist visits to Spain.

Each equation was estimated using lag lengths of 24, 12, 6, and 3 months (i.e., for four estimations, we set p = 24, 12, 6, and 3). Because each equation has identical right-hand side variables, ordinary least squares (OLS) is an efficient estimation technique. Using χ^2 tests, we determined that a lag length of 12 months was most appropriate (reducing lag length from 24 to 12 months had a χ^2 value that was significant at the 0.56 level, whereas reducing the lag length to 6 months had a χ^2 value that was significant at the 0.049 level). The AIC indicated that 12 lags were appropriate, whereas the SBC suggested we could use only 6 lags. Since we were using monthly data, we decided to use the 12 lags.

To ascertain the importance of the interactions between the two series, we obtained the variance decompositions. The moving-average representations of Equations (5.47) and (5.48) express n_t and i_t as dependent on the current and past values of both $\{e_{1t}\}$ and $\{e_{2t}\}$ sequences:

$$n_t = c_0 + \sum_{j=1}^{\infty}(c_{1j}e_{1t-j} + c_{2j}e_{2t-j}) + e_{1t} \tag{5.49}$$

$$i_t = d_0 + \sum_{j=1}^{\infty}(d_{1j}e_{1t-j} + d_{2j}e_{2t-j}) + e_{2t} \tag{5.50}$$

where c_0 and d_0 are vectors containing constants, the 11 seasonal dummies, and a trend; and c_{1j}, c_{2j}, d_{1j}, and d_{2j} are parameters.

Because we cannot estimate (5.49) and (5.50) directly, we used the residuals of (5.47) and (5.48) and then decomposed the variances of n_t and i_t into the percentages attributable to each type of innovation. We used the orthogonalized innovations obtained from a Choleski decomposition; the order of the variables in the factorization had no qualitative effects on our results (the contemporaneous correlation between e_{1t} and e_{2t} was –0.0176).

Empirical Results

The variance decompositions for a 24-month forecasting horizon, with significance levels in parentheses, are reported in Table A5.1. As expected, each time series explains the preponderance of its own past values; n_t explains over 91 percent of its forecast error variance, while i_t explains nearly 98 percent of its forecast error variance. It is interesting that terrorist incidents explain 8.7 percent of the forecast error variance of Spain's tourism, while tourism explains only 2.2 percent of the forecast error variance of terrorist incidents. More important, Granger causality tests indicate that the effects of terrorism on tourism are significant at the 0.006 level, whereas the effects of tourism on terrorism are not significant at conventional levels. Thus, causality is unidirectional: Terrorism affects tourism but not the reverse. We also note that the terrorism series appears to be autonomous in the sense that neither series significantly explains the forecast error variance of i_t. This result is consistent with the notion that terrorists randomize their incidents so that any one incident is not predictable on a month-to-month basis.

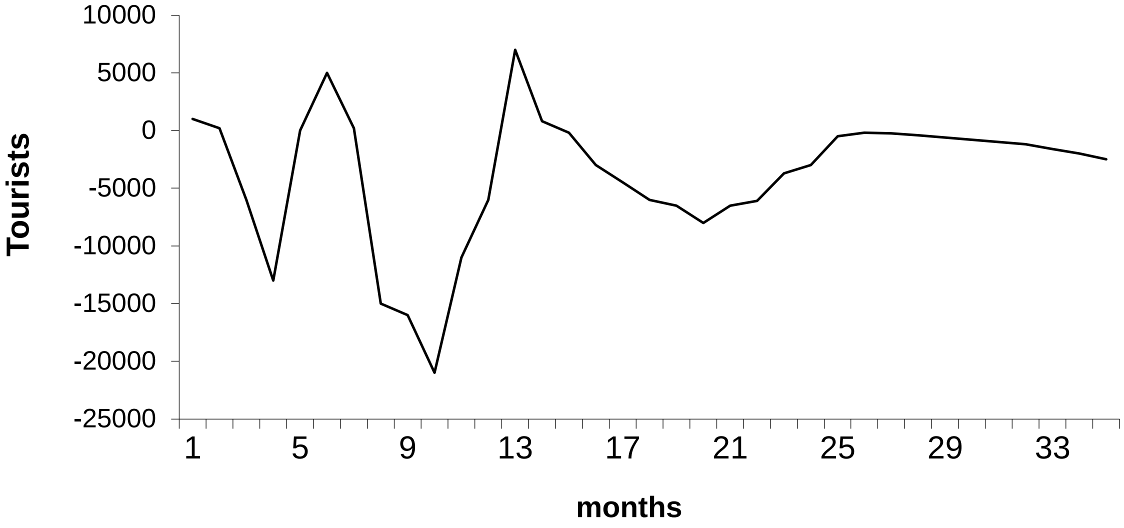
Table A5.1: Variance Decomposition Percentage of 24-Month Error Variance

Percent of forecast error variance in	Typical shock in	
	n_t	i_t
n_t	91.3 (3 x E-15)	8.7 (0.006)
i_t	2.2 (17.2)	97.8 (93.9)

Note: The numbers in parentheses indicate the significance level for the joint hypothesis that all lagged coefficients of the variable in question can be set equal to zero.

Forecasts from an unrestricted VAR are known to suffer from overparameterization. Given the results of the variance decompositions and the Granger causality tests, we reestimated (5.47) and (5.48) restricting all of the coefficients of $A_{21}(L)$ to equal zero. Because the right-hand variables were no longer identical, we reestimated the equations with seemingly unrelated regressions (SUR). From the resulting coefficients from the SUR estimates, the effects of a typical terrorist incident on Spain's tourism can be depicted. In terms of the restricted version of (5.50), we set all e_{1t-j} and e_{2t-j} equal to zero for $j > 0$. We then simulated the time paths resulting from the effects of a one-unit shock to e_{2t}. The time path is shown in Figure 5.8, where the vertical axis measures the monthly impact on the number of foreign tourists and the horizontal axis shows the months following the shock. To smooth out the series, we present the time path of a three-month moving average of the simulated tourism response function.

Figure A5.1: Tourism response to a terrorist incident



After a “typical” terrorist incident, tourism to Spain begins to decline in the third month. After the sixth month, tourism begins to revert to its original level. There does appear to be a rebound in months eight and nine. There follows another drop in tourism in month nine, reaching the maximum decline about one year after the original incident. Obviously, some of this pattern is due to the seasonality in the series. However, tourism slowly recovers and generally remains below its preincident level for a substantial period of time. Aggregating all 36 monthly impacts, we estimate that the combined effects of a typical transnational terrorist incident in Spain are to decrease the total number of foreign visits by 140,847 people. By comparison, a total of 5,392,000 tourists visited Spain in 1988 alone.

SECTION 5.2: A SYMMETRY RESTRICTION

A common assumption in the open-economy macroeconomics literature is that global shocks have little influence on current account balances, relative output levels, and real exchange rates. The notion underlying this assumption is that global shocks affect all nations equally; in a sense, global shocks are like the tides that "cause all boats to rise and fall together." However, this might not be true if nations have different technologies, preferences, and/or factor supplies. In Souki and Enders (2008) we use a four-variable structural VAR to obtain a global shock and three country-specific shocks. The nature of the VAR is that we allow the global shock to have asymmetric effects on the U.S., Japanese and German economies.

As a first step, we performed unit-root tests on the log levels and on the logarithmic first-differences of the variables. All six variables contain a unit root but are stationary in first-differences. We then estimated a three-country model represented by the appropriately differenced four-variable VAR:

$$\begin{bmatrix} \Delta rgus_t \\ \Delta rjus_t \\ \Delta ygus_t \\ \Delta yjus_t \end{bmatrix} = \begin{bmatrix} A_{11}(L) & A_{12}(L) & A_{13}(L) & A_{14}(L) \\ A_{21}(L) & A_{22}(L) & A_{23}(L) & A_{24}(L) \\ A_{31}(L) & A_{32}(L) & A_{33}(L) & A_{34}(L) \\ A_{41}(L) & A_{42}(L) & A_{43}(L) & A_{44}(L) \end{bmatrix} \begin{bmatrix} \Delta rgus_{t-1} \\ \Delta rjus_{t-1} \\ \Delta ygus_{t-1} \\ \Delta yjus_{t-1} \end{bmatrix} + \begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ e_{4t} \end{bmatrix}$$

where $rgus_t$ is the log of the real exchange rate between Germany and the U.S., $rjus_t$ is the log of the real exchange rate between Japan and the U.S., $ygus_t$ is the log of German/U.S. output, $yjus_t$ is the log of Japanese/U.S. output, Δ is the difference operator, the $A_{ij}(L)$ are polynomials in the lag operator L, and the e_{it} are the regression residuals. Note that the responses of the German/Japanese real exchange rate and relative income levels can be obtained from $\Delta rgus_t - \Delta rjus_t$ and $\Delta ygus_t - \Delta yjus_t$, respectively. The estimated VAR includes a constant and four lags of the first-difference of each variable (the lag length selection is based on a likelihood ratio test). The estimation period runs from March 1973 to June 2004.

We classify the shocks by their consequences, not by their source. After all, almost any shock emanates from some particular country. In our classification system, shocks--such as 9/11 or the financial crisis beginning in the U.S. housing market--with immediate worldwide consequences are global, not country-specific, shocks. The shock is global because of its immediate worldwide consequences, not because of its source. The sharp rise and then fall in the price of oil is a global shock. In order to ensure that country-specific shocks do not have any immediate worldwide consequences, it is necessary to assume that country-specific shocks are orthogonal to each other and to the global shock. The discussion implies that it makes sense to decompose the regression residuals using the six restrictions $\alpha_{12} = \alpha_{21} = \alpha_{32} = \alpha_{41} = 0$, $\alpha_{13} = \alpha_{23}$, and $\alpha_{33} = \alpha_{34}$. Hence:

$$\begin{bmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ e_{4t} \end{bmatrix} = \begin{bmatrix} \alpha_{11} & 0 & \alpha_{13} & \alpha_{14} \\ 0 & \alpha_{22} & \alpha_{13} & \alpha_{24} \\ \alpha_{31} & 0 & \alpha_{33} & \alpha_{34} \\ 0 & \alpha_{42} & \alpha_{33} & \alpha_{44} \end{bmatrix} \begin{bmatrix} \varepsilon_{gt} \\ \varepsilon_{jt} \\ \varepsilon_{ut} \\ \varepsilon_{wt} \end{bmatrix}$$

where ε_{wt} is the global (or worldwide) shock in period t, and ε_{it} is the country-specific shock for i in period t. The nature of the four ε_{it} shocks is that they are all *i.i.d.* zero-mean random variables that are mutually uncorrelated in the sense that $E_{t\text{-}1}\varepsilon_{it}\varepsilon_{kt} = 0$ for $i \neq k$. Moreover, we normalize units so that the variance of each structural shock is unity. As such, in this four-variable VAR, we have imposed 6 additional restrictions to obtain an exactly identified system.

To explain, note that an ε_{jt} shock has no contemporaneous effect on $\Delta rgus_t$ if $\alpha_{12} = 0$ and has no contemporaneous effect on $\Delta ygus_t$ if $\alpha_{32} = 0$. In the same way, an ε_{gt} shock has no contemporaneous effect on $\Delta rjus_t$ if $\alpha_{21} = 0$ and has no contemporaneous effect on $\Delta yjus_t$ if $\alpha_{41} = 0$. To explain the last two restrictions, notice that the log of the real exchange rate between Japan and Germany is $rgus_t - rjus_t$ and the log of the German/Japanese output is $ygus_t - yjus_t$. Hence, if $\alpha_{13} = \alpha_{23}$, the U.S. shock will have no contemporaneous effect on the German/Japanese real exchange rate, and if $\alpha_{33} = \alpha_{34}$, the U.S. shock will have no contemporaneous effect on the German/Japanese output.

Since we do not restrict α_{14}, α_{24}, α_{34}, or α_{44} to zero, our identification scheme allows global shocks to change relative output levels and real exchange rates. Nevertheless, we do not *force* global shocks to have asymmetric effects. If the standard assumption is correct (so that global shocks have only symmetric effects), we should find that all values of α_{i4} are equal to zero. Moreover, the lag structure should be such that global shocks explain none of the forecast error variance of real exchange rates and relative outputs. Hence, any findings that our identified global shocks affect relative output levels and/or real exchange rates are necessarily due to non-proportional effects of global shocks.

The variance decompositions are shown in Table A5.1. The key points are:

- We find little evidence that third-country effects are important. The maximal impact is that the U.S.-shock explains 12% of the forecast error variance of the Japanese/German industrial production ratio.

- Global shocks have little effect on relative output levels. As such, the conventional wisdom is correct in that global shocks do tend to affect industrial production levels proportionately.

- Global shocks explain almost all of the movements in the DM/Dollar real exchange rate and sizable portions of the movements in the other two real rates. As such, our identified global shocks alter relative prices but not relative outputs.

A natural interpretation is that preferences differ across nations. Even if global productivity shocks cause output levels to move together, differences in preferences can induce relative price changes. After all, residents of different nations will use their altered income levels to buy different baskets of goods and services. For our purposes, the main point is that a combination of coefficient restrictions and symmetry restrictions can be used to identify structural shocks.

Table A5.1: Variance Decompositions Using Structural Shocks

	Percent of Forecast Error Variances due to Germany shock					
Horizon	$\Delta rgus$	$\Delta rjus$	Δrgj	$\Delta ygus$	$\Delta yjus$	Δygj
1-quarter	1.977	0.000	1.002	63.495	0.000	99.486
4-quarter	4.427	0.630	7.459	61.496	0.589	90.711
8-quarter	5.287	4.851	9.742	63.795	1.563	87.010

	Percent of Forecast Error Variances due to Japanese shock					
Horizon	$\Delta rgus$	$\Delta rjus$	Δrgj	$\Delta ygus$	$\Delta yjus$	Δygj
1-quarter	0.000	72.366	81.365	0.000	1.483	0.420
4-quarter	2.648	68.731	75.597	2.125	4.539	0.473
8-quarter	3.875	64.193	73.175	2.196	4.603	0.833

	Percent of Forecast Error Variances due to U.S.shock					
Horizon	$\Delta rgus$	$\Delta rjus$	Δrgj	$\Delta ygus$	$\Delta yjus$	Δygj
1-quarter	1.948	1.645	0.000	35.600	95.024	0.000
4-quarter	2.713	2.588	2.029	34.709	92.028	7.711
8-quarter	2.782	3.294	2.741	32.388	90.914	11.161

	Percent of Forecast Error Variances due to Global shock					
Horizon	$\Delta rgus$	$\Delta rjus$	Δrgj	$\Delta ygus$	$\Delta yjus$	Δygj
1-quarter	96.076	25.990	17.633	0.905	3.492	0.094
4-quarter	90.212	28.052	14.915	1.669	2.844	1.106
8-quarter	88.056	27.661	14.342	1.621	2.920	0.996

CHAPTER 6

ENDNOTES TO CHAPTER 6

1. To include an intercept term, simply set all realizations of one $\{x_{it}\}$ sequence equal to unity. In the text, the long-run relationship with an intercept will be denoted by $\beta_0 + \beta_1 x_{1t} + \ldots + \beta_n x_{nt} = 0$. Also note that the definition rules out the trivial case in which all elements of β equal zero. Obviously if all the $\beta_i = 0$, $\beta x_t' = 0$.

2. Equation (6.18) can be written as $\lambda^2 = a_1\lambda + a_2$ where $a_1 = (a_{11} + a_{22})$ and $a_2 = (a_{12}a_{21} - a_{11}a_{22})$. Now refer all the way back to Figure 1.5 in Chapter 1. For $\lambda_1 = 1$, the coefficients of (6.18) must lie along line segment BC. Hence, $a_1 + a_2 = 1$, or $a_{11} + a_{22} + a_{12}a_{21} - a_{11}a_{22} = 1$. Solving for a_{11} yields (6.19). For $|\lambda_2| < 1$, the coefficients must lie inside region A0BC. Given (6.19), the condition $a_2 - a_1 = 1$ is equivalent to that in (6.21).

3. Another interesting way to obtain this result is to refer back to (6.14). If $a_{12} = 0$, $y_t = a_{11}y_{t-1} + \varepsilon_{yt}$. Imposing the condition that $\{y_t\}$ is a unit root process is equivalent to setting $a_{11} = 1$ so that $\Delta y_t = \varepsilon_{yt}$.

4. As discussed in Section 6.1 of the *Supplementary Manual*, the Johansen procedure consists of the matrix of vectors of the squared canonical correlations between the residuals of x_t and Δx_{t-1} regressed on lagged values of Δx_t. The cointegrating vectors are the rows of the normalized eigenvectors.

5. Johansen shows that this two-step procedure has the following properties: (i) if the rank of π is r and there are no $I(2)$ components, the procedure picks out the true value of r with a high probability, (ii) a value of r that is too low is selected with a limiting probability of zero, and (iii) if there are $I(2)$ components, the procedure will accept no $I(2)$ components with a small probability. Jorgensen, Kongsted, and Rahbek (1996) show how to simultaneously select the values of r and s. As such, the characteristic roots are those of $\alpha_\perp \Gamma \beta_\perp$

6. As summarized in Ericsson and MacKinnon (2002), there are other variants of the ECM test. For example, if $\alpha_1 = 0$, it follows that $\beta_1 = \alpha_1$ and $\beta_2 = \alpha_1\beta$ should both equal zero. As such, it is also possible to test whether the restriction $\beta_1 = \beta_2 = 0$ is binding on (6.67). However, this type of test becomes more difficult when z_t is actually a vector of weakly exogenous variables.

SECTION 6.1:

APPENDIX 6.1: Characteristic Roots, Stability, and Rank

Characteristic Roots Defined

Let A be an $(n \cdot n)$ square matrix with elements a_{ij} and let x be an $(n \cdot 1)$ vector. The scalar λ is called a characteristic root of A if

$$Ax = \lambda x \tag{A6.1}$$

Let I be an $(n \cdot n)$ identity matrix so that we can rewrite (A6.1) as

$$(A - \lambda I)x = 0 \tag{A6.2}$$

Since x is a vector containing values not identically equal to zero, (A6.2) requires that the rows of $(A - \lambda I)$ be linearly dependent. Equivalently, (A6.2) requires that the determinant $| A - \lambda I | = 0$. Thus, we can find the characteristic root(s) of (A6.1) by finding the values of λ that satisfy

$$| A - \lambda I | = 0 \tag{A6.3}$$

Example 1

Let A be the matrix:

$$A = \begin{bmatrix} 0.5 & -0.2 \\ -0.2 & 0.5 \end{bmatrix}$$

so that

$$| A - \lambda I | = \begin{vmatrix} 0.5 - \lambda & -0.2 \\ -0.2 & 0.5 - \lambda \end{vmatrix}$$

Solving for the value of λ such that $| A - \lambda I | = 0$ yields the quadratic equation:

$$\lambda^2 - \lambda + 0.21 = 0$$

The two values of λ which solve the equation are $\lambda = 0.7$ and $\lambda = 0.3$. Hence, 0.7 and 0.3 are the two characteristic roots.

Example 2

Now change A such that each element in column 2 is twice the corresponding value in column 1. Specifically,

$$A = \begin{bmatrix} 0.5 & 1 \\ -0.2 & -0.4 \end{bmatrix}$$

Now,

$$|A-\lambda I|=\begin{bmatrix} 0.5-\lambda & 1 \\ -0.2 & -0.4-\lambda \end{bmatrix}$$

Again, there are two values of λ which solve $|A-\lambda I| = 0$. Solving the quadratic equation $\lambda^2 - 0.1\lambda = 0$ yields the two characteristic roots $\lambda_1 = 0$ and $\lambda_2 = 0.1$.

Characteristic Equations

Equation (A6.3) is called the characteristic equation of the square matrix A. Notice that the characteristic equation will be an nth-order polynomial in λ. The reason is that the determinant $|A - \lambda I| = 0$ contains the nth degree term λ^n resulting from the expression:

$$(a_{11}-\lambda)(a_{22}-\lambda)(a_{33}-\lambda)\ \ldots\ (a_{nn}-\lambda)$$

As such, the characteristic equation will be an nth-order polynomial of the form:

$$\lambda^n + b_1\lambda^{n-1} + b_2\lambda^{n-2} + b_3\lambda^{n-3} + \ldots + b_{n-1}\lambda + b_n = 0 \qquad \text{(A6.4)}$$

From (A6.4) it immediately follows that an $(n \cdot n)$ square matrix will necessarily have n characteristic roots. As we saw in Chapter 1, some of the roots may be repeating and some may be complex. In practice, it is not necessary to actually calculate the values of the roots solving (A6.4). The necessary and sufficient conditions for all characteristic roots to lie within the unit circle are given in Chapter 1 and in the *Supplementary Manual*.

Notice that the term b_n is of particular relevance because $b_n = (-1)^n|A|$. After all, b_n is the only expression resulting from $|A - \lambda I|$ that is not multiplied by λ. In terms of (A6.4), the expressions λ^n and b_n will have the same sign if n is even and opposite signs if n is odd. In Example 1, the characteristic equation is $\lambda^2 - \lambda + 0.21 = 0$ so that $b_2 = 0.21$. Since $|A| = 0.21$, it follows that $b_2 = (-1)^2(0.21)$. Similarly, in Example 2, the characteristic equation is $\lambda^2 - 0.1\lambda = 0$, so that $b_2 = 0$. Since it is also the case that $|A| = 0$, it also follows that $b_2 = (-1)^2|A|$. In Example 3 below, we consider the case in which $n = 3$.

Example 3

Let A be such that

$$|A-\lambda I|=\begin{bmatrix} 0.5-\lambda & 0.2 & 0.2 \\ 0.2 & 0.5-\lambda & 0.2 \\ 0.2 & 0.2 & 0.5-\lambda \end{bmatrix}$$

The characteristic equation is

$$\lambda^3 - 1.5\lambda^2 + 0.63\lambda - 0.081 = 0$$

and the characteristic roots are

$$\lambda_1 = 0.9,\ \lambda_2 = 0.3, \text{ and } \lambda_3 = 0.3$$

The determinant of A is 0.081 so that $b_3 = -0.081 = (-1)^3|A|$.

Determinants and Characteristic Roots

The determinant of an ($n \cdot n$) matrix is equal to the product of its characteristic roots, that is

$$|A| = \prod_{i=1}^{n} \lambda_i \tag{A6.5}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n$ are the n characteristic roots of the ($n \cdot n$) matrix A.

The proof of this important proposition is straightforward since the values $\lambda_1, \lambda_2, \ldots, \lambda_n$ solve (A6.4). Yet, from the algebra of polynomials, the product of the factors of (A6.4) is equal to $(-1)^n b_n$:

$$\prod_{i=1}^{n} \lambda_i = (-1)^n b_n$$

From the discussion above, we also know that $(-1)^n b_n = |A|$. Hence (A6.5) must hold in that the product $(\lambda_1)(\lambda_2) \ldots (\lambda_n) = (-1)^n b_n = |A|$.

Examples 1 to 3 Continued

In Examples 1 and 2, the characteristic equation is quadratic of the form $\lambda^2 + b_1\lambda + b_2 = 0$. To find the roots of this quadratic equation, we seek the factors λ_1 and λ_2 such that

$$(\lambda - \lambda_1)(\lambda - \lambda_2) = 0$$

or

$$\lambda^2 - (\lambda\lambda_1 + \lambda\lambda_2) + \lambda_1\lambda_2 = 0$$

or

$$\lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2 = 0$$

Clearly, the value of $\lambda_1\lambda_2$ must equal b_2. To check the formulas in Example 1, recall that the characteristic equation is $\lambda^2 - \lambda + 0.21 = 0$. In this problem, the value of b_2 is 0.21, the product of the characteristic roots is $\lambda_1\lambda_2 = (0.7)(0.3) = 0.21$, and the determinant of A is $(0.5)^2 - (0.2)^2 = 0.21$. In Example 2, the characteristic equation is $\lambda^2 - 0.1\lambda = 0$ so that $b_2 = 0$. The product of the characteristic roots is $\lambda_1\lambda_2 = (0.0)(0.1) = 0.0$, and the determinant of A is $(0.5)(0.4) - (0.2) = 0$.

In Example 3, the characteristic equation is cubic: $\lambda^3 - 1.5\lambda^2 + 0.63\lambda - 0.081 = 0$. The value of b_3 is -0.081, product of the characteristic roots is $(0.9)(0.3)(0.3) = 0.081$, and the determinant of A is 0.081.

Characteristic Roots and Rank

The rank of a square ($n \cdot n$) matrix A is the number of linearly independent rows (columns) in the matrix. The notation rank(A) = r means that the rank of A is equal to r. The matrix A is said to be of full rank if rank(A) = n.

From the discussion above, it follows that *the rank of A is equal to the number of its nonzero characteristic roots.* Certainly, if all rows of A are linearly independent, the determinant of A is not equal to zero. From (A6.5) it follows that none of the characteristic roots can equal zero if $\mid A \mid \neq 0$. At the other extreme, if rank(A) = 0, each element of A must equal zero. When rank(A) = 0, the characteristic equation degenerates into $\lambda^n = 0$ with the solutions $\lambda_1 = \lambda_2 = \ldots = \lambda_n = 0$. Consider the intermediate cases wherein $0 < \text{rank}(A) = r < n$. Since interchanging the various rows of a matrix does not alter the absolute value of its determinant, we can always rewrite $\mid A - \lambda I \mid = 0$ such that the first r rows comprise the r linearly independent rows of A. The determinant of these first r rows will contain r characteristic roots. The other $(n–r)$ roots will be zeroes.

In Example 2, rank(A) = 1 since each element in row 1 equals –2.5 times the corresponding element in row 2. For this case, $\mid A \mid = 0$ and exactly one characteristic root is equal to zero. In the other two examples, A is of full rank and all characteristic roots differ from zero.

Example 4

Now consider a (3· 3) matrix A such that rank(A) = 1. Let

$$|A - \lambda I| = \begin{bmatrix} 0.5-\lambda & 0.2 & 0.2 \\ 1 & 0.4-\lambda & 0.4 \\ -0.25 & -0.1 & -0.1-\lambda \end{bmatrix}$$

The rank of A is unity since row 2 is twice row 1 and row 3 is –0.5 times row 1. The determinant of A equals zero and the characteristic equation is given by

$$\lambda^3 - 0.8\lambda^2 = 0$$

The three characteristic roots are $\lambda_1 = 0.8$, $\lambda_2 = 0$, and $\lambda_3 = 0$.

Stability of a First-order VAR

Let x_t be the $(n \cdot 1)$ vector $(x_{1t}, x_{2t}, \ldots, x_{nt})'$ and consider the first-order VAR

$$x_t = A_0 + A_1 x_{t-1} + \varepsilon_t \tag{A6.6}$$

where $A_0 = (n \cdot 1)$ vector with elements a_{i0}

$A_1 = (n \cdot n)$ square matrix with elements a_{ij}

$\varepsilon_t = (n \cdot 1)$ vector of white-noise disturbances $(\varepsilon_{1t}, \varepsilon_{2t}, \ldots, \varepsilon_{nt})'$

To check the stability of the system, we need only examine the homogeneous equation:

$$x_t = A_1 x_{t-1} \tag{A6.7}$$

We can use the method of undetermined coefficients and for each x_{it} posit a solution of the form:

$$x_{it} = c_i \lambda^t \tag{A6.8}$$

where c_i is an arbitrary constant.

If (A6.8) is to be a solution, it must satisfy all each of the n equations represented by (A6.7). Substituting $x_{it} = c_i\lambda^t$ and $x_{it-1} = c_i\lambda^{t-1}$ for each of the x_{it} in (A6.7), we get

$$c_1\lambda^t = a_{11}c_1\lambda^{t-1} + a_{12}c_2\lambda^{t-1} + \ldots + a_{1n}c_n\lambda^{t-1}$$

$$c_2\lambda^t = a_{21}c_1\lambda^{t-1} + a_{22}c_2\lambda^{t-1} + \ldots + a_{2n}c_n\lambda^{t-1}$$

$$c_3\lambda^t = a_{31}c_1\lambda^{t-1} + a_{32}c_2\lambda^{t-1} + \ldots + a_{3n}c_n\lambda^{t-1}$$

$$\ldots$$

$$c_n\lambda^t = a_{n1}c_1\lambda^{t-1} + a_{n2}c_2\lambda^{t-1} + \ldots + a_{nn}c_n\lambda^{t-1}$$

Now, divide each equation by λ^{t-1} and collect terms to form

$$\begin{array}{lllll}
c_1(a_{11}-\lambda) & + c_2a_{12} & + c_3a_{13} & \ldots + c_na_{1n} & = 0 \\
c_1a_{21} & + c_2(a_{22}-\lambda) & + c_3a_{23} & \ldots + c_na_{2n} & = 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
c_1a_{n1} & + c_2a_{n2} & + c_3a_{n3} & \ldots + c_n(a_{nn}-\lambda) & = 0
\end{array}$$

so that the following system of equations must be satisfied:

$$\begin{bmatrix} (a_{11}-\lambda) & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & (a_{22}-\lambda) & a_{23} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & (a_{nn}-\lambda) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \cdots \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

For a nontrivial solution to the system of equations, the following determinant must equal zero:

$$\begin{vmatrix} (a_{11}-\lambda) & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & (a_{22}-\lambda) & a_{23} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & (a_{nn}-\lambda) \end{vmatrix} = 0$$

The determinant will be an nth-order polynomial that is satisfied by n values of λ. Denote these n characteristic roots by λ_1, λ_2, … λ_n. Since each is a solution to the homogeneous equation, we know that the following linear combination of the homogeneous solutions is also a homogeneous solution:

$$x_{it} = d_1\lambda_1^{\,t} + d_2\lambda_2^{\,t} + \ldots + d_n\lambda_n^{\,t}$$

Note that each $\{x_{it}\}$ sequence will have the same roots. The necessary and sufficient condition for stability is that all characteristic roots lie within the unit circle.

Cointegration and Rank

The relationship between the rank of a matrix and its characteristic roots is critical in the Johansen procedure. Using the notation from Section 7, let

$$x_t = A_1 x_{t-1} + \varepsilon_t$$

so that

$$\Delta x_t = (A_1 - I)x_{t-1} + \varepsilon_t$$
$$= \pi x_{t-1} + \varepsilon_t$$

If the rank of π is unity, all rows of π can be written as a scalar multiple of the first. Thus, each of the $\{\Delta x_{it}\}$ sequences can be written as

$$\Delta x_{it} = s_i\,(\pi_{11}x_{1t-1} + \pi_{12}x_{2t-1} + \ldots + \pi_{1n}x_{nt-1}) + \varepsilon_{it}$$

where $s_1 = 1$ and $s_i = \pi_{ij}/\pi_{1j}$.

Hence, the linear combination: $\pi_{11}x_{1t-1} + \pi_{12}x_{2t-1} + \ldots + \pi_{1n}x_{nt-1} = (\Delta x_{it} - \varepsilon_{it})/s_i$ is stationary since Δx_{it} and ε_{it} are both stationary.

The rank of π equals the number of cointegrating vectors. If rank(π) = r, there are r linearly independent combinations of the $\{x_{it}\}$ sequences that are stationary. If rank(π) = n, all variables are stationary.

The rank of π is equal to the number of its characteristic roots that differ from zero. Order the roots such that $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. The Johansen methodology allows you to determine the number of roots that are statistically different from zero. The relationship between A_1 and π is such that if all characteristic roots of A_1 are in the unit circle, π is of full rank.

Calculating the Characteristic Roots in Johansen's Method

Although commercially available software packages can obtain the characteristic roots of π, you might be interested in programming the method yourself (or at least understanding the method). First select the most appropriate lag length p in for the VAR

$$x_t = A_1 x_{t-1} + \ldots + A_p x_{t-p} + \varepsilon_t$$

STEP 1: Estimate the VAR in first differences, that is, estimate

$$\Delta x_t = B_1\Delta x_{t-1} + \ldots + B_{p-1}\Delta x_{t-p+1} + e_{1t}$$

STEP 2: Regress x_{t-1} on the lagged changes; that is, estimate a VAR of the form

$$x_{t-1} = C_1\Delta x_{t-1} + \ldots + C_{p-1}\Delta x_{t-p+1} + e_{2t}$$

STEP 3: Compute the squares of the canonical correlations between e_{1t} and e_{2t}. In an n-equation VAR, the n canonical correlations are the n values of λ_i. The λ_i are obtained as the solutions to

$$\left|\lambda_i S_{22} - S_{12}S_{11}^{-1}S_{12}^{'}\right| = 0$$

where $S_{ii} = T^{-1}\sum_{t=1}^{T} e_{it}(e_{it})'$, $S_{12} = T^{-1}\sum_{t=1}^{T} e_{2t}(e_{1t})'$

and e_{1t} and e_{2t} are the column vectors of residuals obtained in Steps 1 and 2.

STEP 4: The maximum likelihood estimates of the cointegrating vectors are the *n* columns that are nontrivial solutions for

$$\lambda_i S_{22}\pi_i = S_{12}S_{11}^{-1}S_{12}^{'}\pi_i$$

SECTION 6.2

APPENDIX 6.2: INFERENCE ON A COINTEGRATING VECTOR

The Johansen procedure allows you to test restrictions on one or more cointegrating vectors. However, it is very tempting to use the *t*-statistics on a coefficient of a cointegrating vector estimated by OLS in the Engle–Granger methodology. Nevertheless, you must avoid this temptation since the coefficients *do not* have asymptotic *t*-distributions except in one special circumstance. The problem is that the coefficients are super-consistent but the standard errors are not. Nevertheless, it is typical for a published study to report the coefficients of the cointegrating vector and the associated *t*-statistics or standard errors. For example, the cointegrating relationship between y_t, z_t, and w_t in Section 5 was reported as (with *t*-statistics in parentheses)

$$\begin{array}{c} y_t = -0.0484 - 0.9273z_t + 0.97687w_t + e_{yt} \\ (-0.575)\ \ (-38.095)\ \ \ (53.462) \end{array}$$

However, $\{e_{yt}\}$ may be serially correlated and z_t and w_t may not be exogenous variables. As in a traditional regression with stationary variables, you need to correct for serial correlation and the problem of endogenous regressors. To illustrate the *fully modified least squares* procedure developed by Phillips and Hansen (1990), consider the simple two-variable example

$$y_t = \beta_0 + \beta_1 z_t + e_{1t}$$

$$\Delta z_t = e_{2t}$$

The first equation is the cointegrating relationship and the second indicates that $\{z_t\}$ is the stochastic trend. The notation e_{1t} and e_{2t} is designed to illustrate the point that the residuals from both equations are stationary. However, they may be serially correlated and may be correlated with each other. As such, the second equation is actually quite general since Δz_t can be correlated with its own lags and with values of y_t.

Clearly, the relationship between the two errors is crucial. We begin with the simple case wherein:

$$\begin{bmatrix} e_{1t} \\ e_{2t} \end{bmatrix} = N\ i.i.d.\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

Case 1: In this circumstance, the errors are serially uncorrelated and the cross-correlations are zero. Hence, the OLS regression of y_t on z_t and a constant is such that the explanatory variable (i.e., z_t) is independent of the error term e_{1t}. As indicated in the text, the OLS estimates of β_0 and β_1 can be tested using the normal distribution. Hence, *t*-tests and *F*-tests are appropriate. If the disturbances are not normally distributed, the asymptotic results are such that *t*-tests and *F*-tests are appropriate.

Case 2: In general, e_{1t} and e_{2t} will be correlated with each other so that $Ee_{1t}e_{2t} \neq 0$. In order to conduct inference on the parameters of the cointegrating vector, it is necessary to correct for the endogeneity of z_t. You do this by including leads and lags of $\{\Delta z_t\}$ in the cointegrating relationship. Hence, you estimate the equation

$$y_t = \beta_0 + \beta_1 z_t + \ldots + \gamma_{-1}\Delta z_{t+1} + \gamma_0 \Delta z_t + \gamma_1 \Delta z_{t-1} + \ldots + e_{1t}$$

In essence, you are controlling for innovations in z_t since the equation is equivalent to

$$y_t = \beta_0 + \beta_1 z_\mathrm{t} + \ldots + \gamma_{-1} e_{2t+1} + \gamma_0 e_{2t} + \gamma_1 e_{2t-1} + \ldots + e_{1t}$$

Let var(e_{1t}) be denoted by σ_e^2. If $\{e_{1t}\}$ is serially uncorrelated, you can form a t-statistic to determine whether the estimated value of β_1 (i.e, $\hat{\beta}_1$) equals the hypothesized value β_1 using the t–statistic

$$t = (\hat{\beta}_1 - \beta_1) / \sigma_e$$

Case 3: In the most general case, $Ee_{1t}e_{2t} \neq 0$ and the residuals from the cointegrating vector (i.e., the estimated values of e_{1t}) are likely to be serially correlated. Hence, you also need to modify the t–statistic so that you use the appropriate estimate of the variance of e_{1t}. If the $\{e_{1t}\}$ series is serially correlated, you adjust the t-statistic using the following procedure:

STEP 1: Estimate the equation for y_t and obtain the estimated $\{e_{1t}\}$ series. Denote the t-statistic for the null hypothesis $\hat{\beta}_1 = \beta_1$ as t_0.

STEP 2: Estimate the $\{e_{1t}\}$ series as an AR(p) process to correct for autocorrelation. In particular, use the residuals from Step 1 to estimate the equation

$$e_{1t} = \alpha_1 e_{1t-1} + \ldots + \alpha_p e_{1t-p} + \varepsilon_t$$

Let σ^2 denote the estimated variance of ε_t so that σ is the standard deviation. Construct the value λ as

$$\lambda = \sigma/(1 - \alpha_1 - \ldots - \alpha_p).$$

STEP 3: Multiply t_0 by σ_e/λ. The resulting value is the appropriate t-statistic for the null hypothesis $= \beta_1$. Compare the corrected t-statistic to that in a t-table. As you can see, the corrected t-statistic uses a more appropriate estimator for var(e_{1t}).

Little is altered if we allow z_t to be a vector of variables. However, a word of caution is in order. There are many possible sources of error in the three-step methodology outlined above. You could use too few or too many lags in Step 1. A similar problem arises in Step 2 because p is unknown. The Johansen procedure circumvents many of these problems in that all variables are treated as jointly endogenous and the VAR residuals are not serially correlated. Hence, you can conduct inference on the cointegrating vector(s) directly.

The procedure is not always as difficult as it sounds. As a practical matter, many researchers correct for serial correlation by adding lagged changes of Δy_t to the estimated equation. If the augmented equation eliminated the serial correlation, Steps 2 and 3 are unnecessary. The estimated equation has the form

$$y_t = \beta_0 + \beta_1 z_\mathrm{t} + A_1(L)\Delta y_{t-1} + \ldots + \gamma_{-1}\Delta z_{t+1} + \gamma_0 \Delta z_t + \gamma_1 \Delta z_{t-1} + \ldots + \varepsilon_t$$

where $A_1(L)$ is a polynomial in the lag operator L and $\{\varepsilon_t\}$ is serially uncorrelated.

CHAPTER 7

ENDNOTES TO CHAPTER 7

1. Recall that the third-order Taylor series expansion of $y = f(x)$ around the point $x = x_0$ is $y = f(x_0) + f'(x_0)(x - x_0) + (1/2)f''(x_0)(x - x_0)^2 + (1/6)f'''(x_0)(x - x_0)^3$ where the symbol $'$ denotes differentiation.

2. Hansen (1999) and Enders, Falk and Siklos (2007) consider the issue of inference on the coefficients (and the threshold τ) in a TAR model. Although the confidence intervals obtained from a conventional *t*-distribution are only approximations to the actual distributions, oftentimes the distributions obtained by bootstrapping do not perform better. Note that the results here differ from those reported in the paper.

3. Petrucelli and Woolford (1984) showed that a weaker set of sufficient conditions for the stationarity of $\{y_t\}$ is $\rho_1 < 0$, $\rho_2 < 0$, and $(1 + \rho_1)(1 + \rho_2) < 1$.

4. The Wald test, LM test and Likelihood ratio tests are all asymptotically equivalent. Andrews (1993) develops the critical values using an LM test. In essence, the regressors from the null model plus the breaking terms are regressed on the residuals from the no break model. Instead of using a single supremum value, Andrews and Ploeberger (1994) develop an optimal test that uses an exponentially weighted average of the individual *F*-statistics.

Section 7.1 Introduction to the Kalman Filter

The Signal Extraction Problem

Many researchers now use unobserved components models and the Kalman filter to estimate nonlinear processes. Before reading this section, you might want to reread some of the supplementary material to Chapter 4.

Suppose that we observe a variable, y, and want to decompose it into two orthogonal components. Let:

$$y = x + \eta \tag{1}$$

where: x and η are the unobserved stochastic components. Although we do not observe the individual components, we know their distribution is such that $Ex = E\eta = 0, var(x) = \sigma_x^2$, $var(\eta) = \sigma_\eta^2$, and $Ex\eta = 0$. Hence, it follows that:

$$Ey = 0$$

$$var(y) = \sigma_x^2 + \sigma_\eta^2$$

Our aim is to from a prediction of x, called $\widehat{x}$, having observed the variable y. Consider the prediction equation:

$$\widehat{x} = \alpha_0 + \alpha_1 y \tag{2}$$

Notice that the prediction equation is linear in that the prediction of x is a linear function of the observed variable y. Of course, the predicted value of η, called $\widehat{\eta}$, is equal to $y - \widehat{x} = -\alpha_0 + (1 - \alpha_1)y$. The selection of the coefficients α_0 and α_1 is not arbitrary in that we want to minimize the expected value of the squared prediction error. Hence, a formal statement of the problem is:

$$\underset{\alpha_0,\alpha_1}{Min} E(x - \widehat{x})^2 = E(x - \alpha_0 - \alpha_1 y)^2 \tag{3}$$

Minimizing the expected prediction error with respect to α_0 and α_1 yields to two first order conditions:

$$\begin{aligned} -2E(x - \alpha_0 - \alpha_1 y) &= 0 \\ -2E[(x - \alpha_0 - \alpha_1 y)y] &= 0 \end{aligned} \tag{4}$$

The rest is simply arithmetic. From the first equation,

$$E(x) - \alpha_0 - \alpha_1 E(y) = 0 \tag{5}$$

Since $E(x) = 0$ and $E(y) = 0,$ it follows that $\alpha_0 = 0$. Now rewrite the second equation using the facts that $\alpha_0 = 0$ and $y = x + \eta$ so that:

$$E[\{x - \alpha_1(x + \eta)\}(x + \eta)] = 0$$

Since the cross-product term $E(x\eta) = 0,$ we can write

$$E[(1 - \alpha_1)x^2 - \alpha_1 \eta^2] = 0$$

or recognizing that $Ex^2 = \sigma_x^2$ and $E\eta^2 = \sigma_\eta^2$, we have

$$(1-\alpha_1)\sigma_x^2 - \alpha_1\sigma_\eta^2 = 0.$$

If you solve for α_1, you should find

$$\alpha_1 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2} \tag{6}$$

Thus, the optimal forecast rule is such that α_1 is the percentage of the total variance of y that is due to x. If, for example, $\sigma_x^2 = 0$, all of the variance of y is due to η. In this case, $\alpha_1 = 0$ so that the forecast of $x = 0$. On the other hand, if $\sigma_\eta^2 = 0$, all of the variation in y is due to x. As such, $\alpha_1 = 1$, and the optimal forecast of x is simply the current value of y If x and η are equally variable (so that $\sigma_x^2 = \sigma_\eta^2$), the optimal forecast rule simply splits the observed value of y in half; one half is equal to the predicted value of x and the other is equal to the predicted value of η.

Exercises

a. Derive the optimal values of α_0 and α_1 assuming that the expected values of x and η differ from zero. Specifically, let $Ex = \overline{x}$ and $E\eta = \overline{\eta}$.

b. Derive the optimal values of α_0 and α_1 under the assumption that x_t does not have a $1:1$ effect on y_t. Specifically, let the model for y_t be given by

$$y_t = \beta x_t + \eta_t$$

c. Explain the difference between the regression model $y_t = \beta x_t + \eta_t$ and the unobserved components model.

Signal Extraction for an Autoregressive Process

The problem of decomposing a series into two constituent components is more difficult when one of the processes is autoregressive. The reason is that the conditional mean of the autoregressive component will be changing over time. The optimal predictor of such a component will take these changes into account when forecasting. Consider the process:

$$\begin{aligned} y_t &= \beta x_t + \eta_t \\ x_t &= \rho x_{t-1} + \varepsilon_t \end{aligned} \tag{7}$$

Time subscripts have been introduced since the conditional mean of x_t, and hence the conditional mean of y_t, is changing over time. Although we do not observe the x_t, η_t, or ε_t directly, we know their distribution is such that $E\varepsilon = E\eta = 0$, $var(\varepsilon) = \sigma_\varepsilon^2$, $var(\eta) = \sigma_\eta^2$, and $E\varepsilon\eta = 0$. Note that the model of the previous section is the special case of an $AR(1)$ process such that $\beta = 1$, and $\rho = 0$.

The goal is to minimize the squared prediction error of x_t conditional on the observation of y_t. If you were not very careful, you might guess that it is optimal to select a forecasting rule of the form

$$\widehat{x}_t = \alpha_0 + \alpha_1 y_t \tag{8}$$

However, this would not be quite correct since the optimal value of α_1 changes over time. Remember that $\{y_t\}$ is an autoregressive process plus a noise term due to the presence of η_t. If you observed that the $\{y_t\}$ series exhibited no serial correlation, you might properly surmise that all of the shocks were due to the noise term η_t. If you observed that the $\{y_t\}$ series had autocorrelations equal to $\rho, \rho^2, \rho^3, ...$, you might infer that all of the shocks were due to x_t. The point is that from an initial observation of the series, you would want to adjust the values of α_0 and α_1 as additional values of y_t became available.

As will be explained in more detail below, the optimal forecasting rule has the form:

$$\widehat{x}_t = E_{t-1}x_t + k_t(y_t - E_{t-1}y_t)$$

where k_t is a 'weight' that changes as new information becomes available. Suppose that at the end of period $t-1$ we forecast the values of x_t and y_t. Hence, we forecast these two values before observing the realized value of y_t. Our conditional forecast of x_t is $E_{t-1}x_t$ and our conditional forecast of y_t is $E_{t-1}y_t$. These forecasts are conditional in the sense that they are made without knowledge of the realized value of y_t. The nature of the formula is such that $\widehat{x}_t$ will equal $E_{t-1}x_t$ if $y_t - E_{t-1}y_t$. Hence, if our conditional forecast of y_t turns out to be correct (so that $y_t - E_{t-1}y_t = 0$), we will not alter our forecast of of x_t. However, if $y_t - E_{t-1}y_t \neq 0,$ we will modify our conditional forecast of by k_t percent of the discrepancy. The issue is to find the optimal value of k_t.

Now we will change our notation to be consistent with that found in the literature. Let the symbol $x_{t|t}$ denote the forecast of variable x_t once y_t is realized and $x_{t|t-1}$ denote the forecast of variable x_t before y_t is realized. Hence:

$x_{t|t}$ denotes $\widehat{x}_t$

$x_{t|t-1}$ denotes $E_{t-1}x_t$

$y_{t|t-1}$ denotes $E_{t-1}y_t$

Just to ensure that you understand the notation, we can rewrite the equation for $\widehat{x}_t$ as:

$$x_{t|t} = x_{t|t-1} + k_t(y_t - y_{t|t-1}) \tag{9}$$

Now we are in a position to select the optimal value of k_t so as to minimize the mean square prediction error ($MSPE$). Suppose we enter period t having observed the values y_1 through y_{t-1} and have made the forecast for $x_{t|t-1}$ and $y_{t|t-1}$. The optimization problem for period t is:

$$\underset{k_t}{Min}E_t(x_t - x_{t|t})^2 = E_t[x_t - (x_{t|t-1} + k_t(y_t - y_{t|t-1}))]^2 \tag{10}$$

Since $y_t = \beta x_t + \eta_t$, and $\eta_{t|t-1} = 0$, it follows that $y_{t|t-1} = \beta x_{t|t-1}$. We can rewrite the optimization problem as:

$$\underset{k_t}{Min}E_t[x_t - (x_{t|t-1} + k_t(\beta x_t + \eta_t - \beta x_{t|t-1}))]^2$$

Combining terms:

$$\underset{k_t}{Min}E_t[(1 - \beta k_t)(x_t - x_{t|t-1}) + k_t\eta_t]^2$$

Since x_t and η_t are uncorrelated, we can square the term in square brackets to obtain

$$\underset{k_t}{Min}(1 - \beta k_t)^2 E_t(x_t - x_{t|t-1})^2 + k_t^2\sigma_\eta^2 \tag{11}$$

Optimizing with respect to k_t yields the first-order condition:

$$-2\beta(1-\beta k_t)E_t(x_t - x_{t|t-1})^2 + 2k_t\sigma_\eta^2 = 0$$

Let $P_{t|t-1}$ denote the expression $E_t(x_t - x_{t|t-1})^2$ so that the first-order condition becomes:

$$-2\beta(1-\beta k_t)P_{t|t-1} + 2k_t\sigma_\eta^2 = 0.$$

Solving for k_t yields:

$$k_t = \frac{\beta P_{t|t-1}}{\beta^2 P_{t|t-1} + \sigma_\eta^2} \tag{12}$$

The result is only partially helpful. If we knew the value of $P_{t|t-1} = E_t(x_t - x_{t|t-1})^2$, we would be able to calculate the optimal value of k_t. Of course, there are instances in which $P_{t|t-1}$ is known. For example, in the example above, where there is no serial correlation, it should be clear that $E_t(x_t - x_{t|t-1})^2 = \sigma_x^2$. Since x had a mean of zero and was not serially correlated, $x_{t|t-1} = 0$ and $E_t x_t^2 = \sigma_x^2$. The problem is a bit more complicated here since $x_{t|t-1}$ evolves over time.

Regrouping Equations

We know that $x_t = \rho x_{t-1} + \varepsilon_t$ so that our forecasts of x_t will be linked over time. Specifically, since $E_{t-1}\varepsilon_t = 0$, it must be the case that:

$$E_{t-1}x_t = \rho E_{t-1}x_{t-1}$$

or using the notation $x_{t|t-1} = E_{t-1}x_t$

$$x_{t|t-1} = \rho x_{t-1|t-1} \tag{13}$$

Similarly, we can take the conditional variance of each side of $x_t = \rho x_{t-1} + \varepsilon_t$ to obtain:

$$var(x_t^2) = \rho^2 var(x_{t-1}^2) + \sigma_\varepsilon^2$$

or, if we use the notation $P_{t|t-1} = E_t(x_t - x_{t|t-1})^2$ and $P_{t|t} = E_t(x_t - x_{t|t})^2$

$$P_{t|t-1} = \rho^2 P_{t-1|t-1} + \sigma_\varepsilon^2 \tag{14}$$

Equations (13) and (14) are called the **prediction** equations. The other equations we need, called the **updating** equations, are given by

$$k_t = \beta P_{t|t-1}/(\beta^2 P_{t|t-1} + \sigma_\eta^2) \tag{15}$$

$$\begin{aligned} x_{t|t} &= x_{t|t-1} + k_t(y_t - y_{t|t-1}) \\ &= x_{t|t-1} + k_t(y_t - \beta x_{t|t-1}) \end{aligned} \tag{16}$$

and

$$P_{t|t} = (1-\beta k_t)P_{t|t-1} \tag{17}$$

This last equation follows from substituting the formula for k_t into the formula for $E_t(x_t - x_{t|t})^2 = P_{t|t}$. It should be clear from equation (11) that $P_{t|t}$ can be written as

$$P_{t|t} = (1 - \beta k_t)^2 E_t(x_t - x_{t|t-1})^2 + k_t^2 \sigma_\eta^2$$

or

$$P_{t|t} = (1 - \beta k_t)^2 P_{t|t-1} + k_t^2 \sigma_\eta^2$$

Now consider the formula for $k_t = \beta P_{t|t-1}/(\beta^2 P_{t|t-1} + \sigma_\eta^2)$. Since $1 - \beta k_t = \sigma_\eta^2/(\beta^2 P_{t|t-1} + \sigma_\eta^2)$, it follows that

$$P_{t|t} = [\sigma_\eta^2/(\beta^2 P_{t|t-1} + \sigma_\eta^2)]^2 P_{t|t-1} + [\beta P_{t|t-1}/(\beta^2 P_{t|t-1} + \sigma_\eta^2)]^2 \sigma_\eta^2$$

Collecting terms, it is easy to show that:

$$\begin{aligned} P_{t|t} &= \left[\frac{\sigma_\eta^2}{\beta^2 P_{t|t-1} + \sigma_\eta^2}\right] P_{t|t-1} \\ &= (1 - \beta k_t) P_{t|t-1} \end{aligned}$$

Summary

The basic Kalman filtering problem has the form

$$\begin{aligned} y_t &= \beta x_t + \eta_t \\ x_t &= \rho x_{t-1} + \varepsilon_t \end{aligned} \quad (18)$$

Although y_t is observed, the values of x_t, η_t,and ε_t cannot be directly observed by the researcher. However, it is known that the influence of x_t on y_t is β and that η_t and ε_t are orthogonal to each other. The issue is to form the optimal predictors of x_t and y_t. If the x_t series was observed, we could view (18) as a simple autoregression and use it to forecast x_t. Once we forecasted x_t, we could use this value to forecast y_t.Given that x_t, η_t,and ε_t are unobserved, we need to use a different method. The Kalman filter allows us to decompose the y_t series into two constituent components. Given that we use the weight k_t from (15) equations (13) and (16) are the optimal predictors of x_t conditional on the information set at $t-1$ (i.e., $x_{t|t-1}$) and conditional on the information set at t (i.e., $x_{t|t}$), respectively. Equations (14) and (17) yield the mean square prediction errors. The properties of the forecasts are:

The Conditional Expectation of x_t

Since x_t is unobserved, there are two different ways to think about predicting its realized value. First, the value x_t can be predicted, or ‘estimated’ using the information set available in period $t-1$. We denoted this value as $x_{t|t-1}$. Alternatively, the value of x_t can be predicted using the information set available in t. We denoted this value as $x_{t|t}$. Of course, $x_{t|t}$ should be a better predictor of the actual value of x_t than x_{t-1} since it uses more information. It was shown that the optimal predictor of $x_{t|t}$ is

$$x_{t|t} = x_{t|t-1} + k_t(y_t - \beta x_{t|t-1})$$

where k_t is determined in (15). Of course, any predictor will not be entirely accurate. In (17), we calculated that the $MSPE$ of $x_{t|t}$ is $P_{t|t} = (1 - \beta k_t)P_{t|t-1}$. If you take the conditional expectation of the second equation in (18) with respect to the information set in t you should find

$$x_{t|t-1} = \rho x_{t-1|t-1}$$

As shown in (14), the $MSPE$ of this estimate is $P_{t|t-1} = \rho^2 P_{t-1|t-1} + \sigma_\varepsilon^2$

The Conditional Expectations of y_t

Although y_t can be observed in t, it can be forecasted in $t-1$. Simple take the conditional expectation of the first equation in (18) with respect to the information set in period t to obtain

$$y_{t|t-1} = \beta x_{t|t-1}$$

MSPE of y_t

The forecast of y_t from the perspective of period $t-1$ will contain error. The mean square prediction error of y_t can be calculated from

$$E[y_t - y_{t|t-1}]^2$$

Since $y_t = \beta x_t + \eta_t$ and $y_{t|t-1} = \beta x_{t|t-1}$,it follows that

$$\begin{aligned} E[y_t - y_{t|t-1}]^2 &= E[\beta x_t + \eta_t - \beta x_{t-1}]^2 \\ &= E[\beta(x_t - x_{t|t-1}) + \eta_t]^2 \end{aligned}$$

If you square the term in brackets and recognize that η_t is independent of x_t and $x_{t|t-1}$.you should find

$$\begin{aligned} E[y_t - y_{t|t-1}]^2 &= \beta^2 E[x_t - x_{t|t-1}) + \sigma_\eta^2 \\ &= \beta^2 P_{t|t-1} + \sigma_\eta^2 \end{aligned}$$

Thus, the $MSPE$ of y_t has two sources, $\beta^2 P_{t|t-1}$ and σ_η^2. Note that η_t is the pure noise term that is unforecastable from period $t-1$; the variance of this term is σ_η^2. The other source of forecast error variance is due to the fact that x_t itself needs to be predicted. The variance of this prediction error is $P_{t|t-1}$ and the influence of x_t on y_t is β. Hence, the influence of the prediction error of x_t on the prediction error variance of y_t is $\beta^2 P_{t|t-1}$.

Example of Kalman Filtering

The Kalman filter consists of two prediction equations and three updating equations. Although we have derived the filter for a simple $AR(1)$ process, more complicated functions all work in the same fashion. This section illustrates the use of the filter to predict the successive values of y_t generated from the same $AR(1)$ discussed in the previous section. It is important to understand that Kalman filtering is a dynamic process. You begin with a specific information set and make predictions about the current state of the system. As such, in period 1, we observe y_1 and make a prediction about the value of x_1. If you understand the notation, it should be clear that this prediction is $x_{1|1}$. We then use the observed value of y_1 to make a prediction about the value of x_2; again, if you understand the notation, this value of $x_{2|1}$ since it is the forecast of x_2 given the observation of y_1. Of course, once we enter period 2, we will be able to observe y_2 and so update our forecast of x_2–the updated forecast is $x_{2|2}$. We continue to repeat this process until the end of the data set.

To take the simplest case possible, first consider the case in which $\rho = 0$. From the first example, we already know that the optimal forecasting rule is to partition the observed values of according to the relative variance $\sigma_x^2/(\sigma_x^2 + \sigma_\eta^2)$. We can now use the prediction and updating equations of the Kalman filter to achieve this same result. Since $\rho = 0$, the two prediction equations are:

$$x_{t|t-1} = 0$$

$$P_{t|t-1} = \sigma_\varepsilon^2$$

The updating equation equations are:

$$k_t = P_{t|t-1}/(P_{t|t-1} + \sigma_\eta^2) = \sigma_\varepsilon^2/(\sigma_\varepsilon^2 + \sigma_\eta^2)$$

$$x_{t|t} = x_{t|t-1} + k_t(y_t - x_{t|t-1}) = k_t y_t$$

$$P_{t|t} = (1 - k_t)P_{t|t-1} = \sigma_\varepsilon^2\sigma_\eta^2/(\sigma_\varepsilon^2 + \sigma_\eta^2)$$

If $\sigma_\varepsilon^2 = \sigma_\eta^2$, it follows that:

$$k_t = 0.5$$

$$x_{t|t} = 0.5y_t$$

$$P_{t|t} = (1 - k_t)P_{t|t-1} = 0.5\sigma_\varepsilon^2$$

In period $t - 1$, your forecast is $x_{t|t-1} = 0$ and the variance of this forecast error is σ_ε^2. Once y_t is observed, you can update your forecasts such that $x_{t|t} = 0.5y_t$. The variance of this forecast error, $P_{t|t} = 0.5\sigma_\varepsilon^2$. The fact that $P_{t|t} < P_{t|t-1}$ follows fro the simple fact that the forecast error made after y_t is observed is smaller than that without the knowledge of y_t.

The situation is only slightly more complicated when $\rho > 0$. If we take the case in which $\rho = 0.5$, and further assume that $\sigma_\varepsilon^2 = \sigma_\eta^2 = 1$, the prediction and updating equations become:

Prediction:

$$\begin{aligned} x_{t|t-1} &= 0.5x_{t-1|t-1} \\ P_{t|t-1} &= 0.25P_{t-1|t-1} + 1 \end{aligned}$$

Updating:

$$\begin{aligned} k_t &= P_{t|t-1}/(P_{t|t-1} + 1) \\ x_{t|t} &= x_{t|t-1} + k_t(y_t - x_{t|t-1}) \\ P_{t|t} &= (1 - k_t)P_{t|t-1} \end{aligned}$$

Suppose that the first five occurrences of the y_t series are given by:

t	1	2	3	4	5
y_t	2.0579	0.4984	1.2311	−1.5974	2.2544

Although we do not know the initial conditions of the system, suppose that we are at the very beginning of period 1 and have not, as yet, observed y_1. If the system was just beginning–so that $x_0 = 0$–it might be reasonable to set $x_{0|0} = 0$ and to assign an initial value of $P_{0|0} = 0$. As such, we have the initial conditions necessary to use the Kalman filter. Now, we can consider the iterations for the Kalman filter. Given these initial conditions, we use the prediction equations to obtain $x_{1|0} = 0$ and $P_{1|0} = 1.$ In essence, we forecast a value of zero for the first realization x_1 and let variance of the forecast error be unity. Once we observe $y_1 = 2.0579,$ we use the updating equations to form:

$$k_1 = 1/(1+1) = 0.5$$

$$x_{1|1} = 0.5(2.0579) = 1.0290$$

$$P_{1|1} = 0.5P_{1|0} = 0.5$$

We next use this information to form $x_{2|1}$ and $P_{2|1}.$ From the prediction equations, we obtain:

$$x_{2|1} = 0.5x_{1|1} = 0.5(1.0290) = 0.5145$$

$$P_{2|1} = 0.25P_{1|1} + 1 = 0.25(0.5) + 1 = 1.1250$$

Once we observe y_2, we use the updating equations to obtain:

$$k_2 = P_{2|1}/(P_{2|1} + 1) = 1.1250/(1.1250 + 1) = 0.5294.$$

$$x_{2|2} = x_{2|1} + k_2(y_2 - x_{2|1}) = 0.5145 + 0.5294(0.4984 - 0.5145) = 0.5059.$$

$$P_{2|2} = (1 - k_2)P_{2|1} = (1 - 0.5294)1.1250 = 0.5294.$$

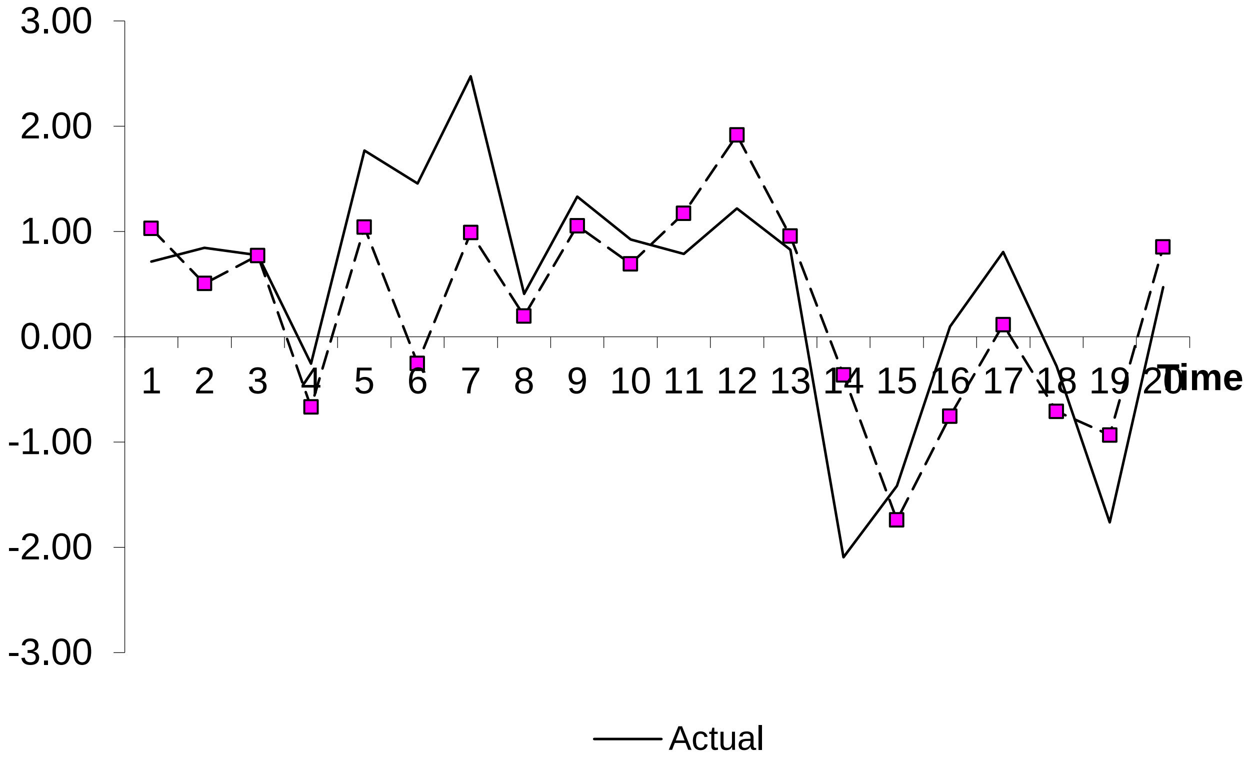
Continuing in this fashion, we can obtain the complete set of forecasts for the series. The subsequent calculations are reported in Table 1. For each time period, t, the simulated values of $\eta_t, \varepsilon_t,$ and x_t are shown in the second through fourth columns, respectively, The fifth column shows $y_t = x_t + \eta_t$.Columns 6 and 7 show the values of $x_{t|t-1}$ and $P_{t|t-1}$ calculated using the prediction equations. If you read down the entries in the sixth column, you will see that $x_{1|0} = 0, x_{2|1} = 0.514$ and $x_{3|2} = 0.775$ (Note that the entries in the table are rounded to three decimal places). Columns 8 through 10 show the values of $k_t, x_{t|t},$ and $P_{t|t}$ calculated using the updating equations. As shown in Figure 1, the Kalman filter forecasts $x_{t|t}$ are reasonable. The solid line in the figure shows the values of x_t and the dashed line shows the predicted values.

Table 1: Decomposition of the AR(1) Process

t	η_t	v_t	ε_t	y_t	$\varepsilon_{t\|t\text{-}1}$	$P_{t\|t\text{-}1}$	k_t	$\varepsilon_{t\|t}$	$P_{t\|t}$
			0					0	0
1	1.341	0.716	0.716	2.058	0.000	1.000	0.500	1.029	0.500
2	-0.347	0.487	0.845	0.498	0.514	1.125	0.529	0.506	0.529
3	0.457	0.352	0.775	1.231	0.253	1.132	0.531	0.772	0.531
4	-1.341	-0.643	-0.256	-1.597	0.386	1.133	0.531	-0.667	0.531
5	0.483	1.899	1.771	2.254	-0.334	1.133	0.531	1.041	0.531
6	-2.392	0.572	1.458	-0.934	0.520	1.133	0.531	-0.252	0.531
7	-0.502	1.747	2.475	1.974	-0.126	1.133	0.531	0.989	0.531
8	-0.473	-0.829	0.409	-0.064	0.495	1.133	0.531	0.198	0.531
9	0.565	1.129	1.334	1.899	0.099	1.133	0.531	1.055	0.531
10	-0.087	0.260	0.926	0.840	0.528	1.133	0.531	0.693	0.531
11	1.115	0.324	0.787	1.902	0.347	1.133	0.531	1.173	0.531
12	1.871	0.825	1.219	3.091	0.586	1.133	0.531	1.916	0.531
13	0.126	0.219	0.829	0.955	0.958	1.133	0.531	0.956	0.531
14	0.992	-2.509	-2.094	-1.102	0.478	1.133	0.531	-0.361	0.531
15	-1.701	-0.368	-1.416	-3.117	-0.181	1.133	0.531	-1.740	0.531
16	-0.749	0.805	0.097	-0.651	-0.870	1.133	0.531	-0.754	0.531
17	-0.254	0.757	0.806	0.551	-0.377	1.133	0.531	0.116	0.531
18	-1.106	-0.680	-0.277	-1.384	0.058	1.133	0.531	-0.708	0.531
19	0.319	-1.623	-1.762	-1.444	-0.354	1.133	0.531	-0.933	0.531
20	1.549	1.352	0.471	2.020	-0.466	1.133	0.531	0.854	0.531

Notes: η_t and v_t are uncorrelated *i.i.d.* normally distributed random variables such that σ_η^2 and σ_v^2 both equal unity. The values of ε_t were constructed as $\varepsilon_t = 0.5\varepsilon_{t\text{-}1} + v_t$ and values of y_t are $\varepsilon_t + \eta_t$. The values of $\varepsilon_{t|t\text{-}1}$ and $P_{t|t\text{-}1}$ are constructed using the prediction equations and the values of

k_t, $\varepsilon_{t|t}$, and $P_{t|t}$ are constructed using the updating equations. The twenty values of ε_t and $\varepsilon_{t|t}$ are shown in Figure 1.



Exercise

The file Table1.xls is an Excel worksheet that contains the data shown in the first five columns of Table 1. However, the entries for $x_{t|t-1}, P_{t|t-1}, k_t, x_{t|t},$ and $P_{t|t}$ shown in columns 6 through 10 of Table 1 are missing. Open the worksheet and construct the formulas for the prediction and updating equations in the appropriate columns. For example, the cell $I2$ contains the value $x_{0|0} = 0$. The formula "$= 0.5 * I2$" $x_{t|t-1}$ is entered in cell $F3$, the value of $x_{1|1}$ will equal 0. Copy this formula to the other cells in column F in order to obtain the predicted values of $x_{t|t-1}$. If you construct the formulas for the other cells properly, you should be able to completely reproduce Table 1.

Convergence

In order to use the Kalman filter, it is necessary to posit initial values for $x_{0|0}$ and $P_{0|0}$. In the example, we used $x_{0|0} = 0$ and $P_{0|0} = 0$ since it was assumed that we knew that $x_0 = 0$. With such knowledge, the period zero forecast of x_0 is obviously zero and, since there is no uncertainty about this forecast, $P_{0|0} = 0$. In general, the choice of the initial values to use in the filter may not be so obvious. What would happen if a different set of initial values for $x_{0|0}$ and $P_{0|0}$ had been chosen? If you are not sure of the answer, you can get a good hint by examining columns 6 and 10 of Table 1. Notice that the successive values of $P_{t|t-1}$ and $P_{t|t}$ both quickly converge to particular values; $P_{t|t-1}$ converges to 1.133 and $P_{t|t}$ converges to 0.531. With this hint, it should not surprise you to know that $P_{t|t-1}$ and $P_{t|t}$ would converge to the same numbers regardless of the initial values used for the filter. To show this more formally, notice that we can collapse the system in order to obtain a difference equation in $P_{t|t-1}$. Write $P_{t|t-1}$as:

$$\begin{aligned} P_{t|t-1} &= 0.25 * (1 - k_{t-1})P_{t-1|t-2} + \sigma_\eta^2 \\ &= 0.25 * (\sigma_x^2/(\sigma_x^2 + P_{t-1|t-2})) * P_{t-1|t-2} + \sigma_\eta^2 \end{aligned}$$

For notational simplicity, let w_t denote $P_{t|t-1}$, so that w_{t-1} denotes $P_{t-1|t-2}$; as such, the difference equation becomes:

$$w_t = 0.25 * (\sigma_x^2/(\sigma_x^2 + w_{t-1})) * w_{t-1} + \sigma_\eta^2$$

Note that the slope of this nonlinear difference equation, $0.25 * (\sigma_x^2/(\sigma_x^2 + w_{t-1}))$, is less than one so that the system is convergent. The steady state solution is obtained by setting $w_t = w_{t-1} = ... = \overline{w}$. Hence:

$$\overline{w} = 0.25 * (\sigma_x^2/(\sigma_x^2 + \overline{w})) * \overline{w} + \sigma_\eta^2$$

If $\sigma_x^2 = \sigma_\eta^2 = 1$, we can write:

$$\overline{w} = 0.25 * (1/(1 + \overline{w})) * \overline{w} + 1$$

The two solutions for $\overline{w}$ are: $\overline{w} = -0.882\,8$,and $1.132\,8$. Since only the positive solution is feasible for the variance, we find: $P_{t|t-1} = 1.1328$. From the first of the updating equations, we know that the solution for $k_t = P_{t|t-1}/(P_{t|t-1} + 1) = 1.1328/(1 + 1.1328) = 0.5311$. Since $P_{t|t} = (1 - k_t)P_{t|t-1}$, it follows that the convergent solution for $P_{t|t}$is such that: $P_{t|t} = (1 - k_t)P_{t|t-1} = (1 - 0.5311)(1.1328) = 0.5311$.

Section 7.2 The Kalman Filter and State Space Models

In a simple dynamic model, the variable of interest, x_t, can often be described by the $AR(1)$ process:

$$x_t = a_0 + a_1 x_{t-1} + \varepsilon_t$$

In the econometrics literature, the symbol x_t usually denotes the magnitude of the variable of interest at time period t. Here, we will call x_t the *state variable* and the equation of motion describing x_t is the *state equation*. The reason for this terminology is that Kalman filtering problems were first used in engineering applications wherein the physical position of an object in motion is usually called the *state* of the object.[1] As such, we can think of (1) as an equation of motion describing the current state of the system as a function of the state in the previous period. To be more general, we can allow the state variable to be a vector so that the state equation becomes:

$$X_t = A_0 + A_1 X_{t-1} + \varepsilon_t \tag{1}$$

where: X_t is an $n\ x\ 1$ vector of state variables, A_0 is an $n\ x\ 1$ vector of constant terms, A_1 is an $n\ x\ n$ matrix of coefficients, and ε_t is an $n\ x\ 1$ vector of random error terms. Obviously, the univariate $AR(1)$ model is a special case of (1) such that $n = 1$. Although the individual elements of ε_t–called ε_{it}–are assumed to be normally distributed and serially uncorrelated, it is generally that case that $E\varepsilon_{it}\varepsilon_{jt} \neq 0$.

The key feature of state space models is that the elements of X_t are not observed directly. As in the last chapter, suppose we observe the variable, y_t, and need to infer the value of the state X_t. To be more general, we can let the relationship between y_t and X_t be given by:

$$Y_t = \beta X_t + \eta_t \tag{2}$$

where: Y_t is an $m\ x\ 1$ vector of observed variables, β is an $m\ x\ n$ matrix of coefficients, and η_t is an $m\ x\ 1$ vector of error terms. Equation (2) is called the *observation equation,* or *measurement equation,* and η_t is called the observation error. The individual elements of the observation error–called η_{it}–are assumed to be normally distributed and serially uncorrelated. We allow for the possibility that the η_{it} are contemporaneously correlated (so that $E\eta_i\eta_j \neq 0)$ although we assume that all $E\varepsilon_{it}\eta_{jt} = 0.$

Together, equations (1) and (2) form a state space model. An essential feature of any state space model is such that the state equation for X_t must be a first-order stochastic difference equation. More general forms allow the coefficient vectors A_0, A_1, and β to be time-varying and allow the presence of exogenous variables. However, at this point, we work with the simple form of (1) and (2).

If you understand the terminology, you should be able to properly identify the two equations used in the last section. Reconsider the equations:

$$y_t = \beta x_t + \eta_t \tag{3}$$

$$x_t = \rho x_{t-1} + v_t \tag{4}$$

[1] In some texts, the state equation is called the *transition* equation.

Clearly, (3) is the observation equation in that it expresses the observed variable, y_t, as the sum of the state variable, x_t, and a noise term. The second equation is the state equation in that the state variable, x, is expressed as an $AR(1)$ process.

State Space Representations of Unobserved Components

The state space model and the Kalman filter go hand-in-hand. To use the Kalman filter, it is necessary to be able to write the model in state space form. Since any state equation must be a first-order stochastic difference equation, you might incorrectly jump to the conclusion that the Kalman filter is of limited use. However, it is often possible to rewrite a very complicated dynamic process as a vector $AR(1)$ process. Once this $AR(1)$ system has been obtained, it is possible to use the Kalman filter.

Some unobserved components models have very natural state space representations. Clearly, the system of equations represented by (3) and (4) are already in state space form: (3) is the observation equation and (4) is the state equation. To use another example, suppose that y_t is composed of a trend plus a noise term. The variable y_t is observed but neither the trend nor the noise term are directly observable. Specifically, let

$$\begin{aligned} y_t &= \tau_t + \eta_t \\ \tau_t &= a_0 + \tau_{t-1} + v_t \end{aligned}$$

Here, y_t consists of a trend component, τ_t, plus the pure random noise component, η_t. Notice that the trend is a random walk plus drift. Again, the state space representation is trivial since the observation equation is nothing more than $y_t = \tau_t + \eta_t$. The state equation expresses the evolution of τ_t as an $AR(1)$ process. Hence the state equation is $\tau_t = a_0 + \tau_{t-1} + v_t$.Now take a more interesting case in which the intercept of the trend is time varying so that we can write the model as

$$\begin{aligned} y_t &= \tau_t + \eta_t \\ \tau_t &= a_t + \tau_{t-1} + v_{1t} \\ a_t &= a_{t-1} + v_{2t} \end{aligned}$$

This model, called the *local linear trend* (*LLT*) model, is such that drift term of the trend is a random walk process. The observation equation is unchanged so that is can be written as $y_t = \tau_t + \eta_t$. Note that the random walk plus noise model above is a special case of the *LLT* model such that $var(v_{2t}) = 0$ implying that $a_t = a_{t-1}$.One way to work with the model is to allow the state variables to be the trend, τ_t,and the intercept, a_t. However, it is more convenient to allow the state variables to be τ_t and a_{t+1}. As such, the equation describing the evolution of the state variables can be written as

$$\begin{bmatrix} \tau_t \\ a_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ a_t \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t+1} \end{bmatrix}$$

A vexing problem in economic analysis is to decompose an observed time series variable, such as real GDP, into its trend and cyclical component. The nature of the problem is that the trend and cyclical components are not observed. Nevertheless, it is of interest to know whether GDP is above or below trend. Consider a simple formulation of the problem such that

$$\begin{aligned} y_t &= \tau_t + c_t \\ \tau_t &= \tau_0 + \tau_{t-1} + v_t \\ c_t &= a_1 c_{t-1} + a_2 c_{t-2} + \eta_t \end{aligned}$$

where y_t is the level of real GDP in period t, τ_t is the trend component, and c_t is the cyclical component.

Notice that the formulation is such that the trend is a random walk plus a drift term. As such, on average, the trend increases by τ_0 each period. Notice that a v_t shock represents a change in the intercept of the trend. There are good economic reasons to suppose that the cyclical component, c_t, follows an $AR(2)$ process. After all, the cyclical component is the deviation of GDP from its trend (i.e., $c_t = y_t - \tau_t$) that can be thought of as a recession if c_t is negative or as an expansion if c_t is positive. Since recessions and expansions are persistent, it makes sense to model the cyclical component as an AR process. There are several state space representations for this model. The transition equation needs to adapted since we need to write that $AR(2)$ process for c_t as an $AR(1)$. The technique is to actually write c_{t-1} as one of the unobserved components in the system.

$$\begin{bmatrix} \tau_t \\ c_t \\ c_{t-1} \end{bmatrix} = \begin{bmatrix} \tau_0 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & a_1 & a_2 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau_{t-1} \\ c_{t-1} \\ c_{t-2} \end{bmatrix} + \begin{bmatrix} v_t \\ \eta_t \\ 0 \end{bmatrix}$$

Clearly, this is in the form of (1) such that

$$X_t = [\tau_t, c_t, c_{t-1}]^T, A_0 = [\tau_0, 0, 0]^T, v_t = [v_t, \eta_t, 0]^T$$

and $A_1 =$ the 3 x 3 coefficient matrix.

The observation equation relates the observed variable, y_t, to the unobserved components. Hence, in matrix form, we can write the observation equation as

$$y_t = [1 \quad 1 \quad 0] \begin{bmatrix} \tau_t \\ c_t \\ c_{t-1} \end{bmatrix}$$

Another important example involves two *cointegrated* variables. According to Engle and Granger (1987), two $I(1)$ variables are cointegrated if there exists a linear combination of the variables that is $I(0)$.Another way to think about cointegrated variables is that they share a single stochastic trend. Suppose that it is possible to observe the variables x_{1t} and x_{2t} but that the trend component and the noise components are unobservable. To be specific, consider the process

$$\begin{aligned} x_{1t} &= \mu_t + \varepsilon_{1t} \\ x_{2t} &= \mu_t + \varepsilon_{2t} \\ \mu_t &= \mu_{t-1} + \varepsilon_{3t} \end{aligned}$$

Here, x_{1t} is composed of the stochastic trend component μ_t plus a pure noise term ε_{1t}. Notice that x_{2t} shares the same trend as x_{i1} although the noise components, ε_{1t} and ε_{2t}, differ. The stochastic trend, μ_t is assumed to be a pure random walk process. Clearly each of the variables is a nonstationary $I(1)$ process. However, they are cointegrated since they share the same trend–as such, it is possible to form a linear combination of the two variables that is stationary. Obviously, the difference between x_{1t} and x_{2t} is stationary since $x_{1t} - x_{2t} = \varepsilon_{1t} - \varepsilon_{2t}$.To write the system in state space form, note that the state variable is μ_t.The state equation is nothing more than

$$\mu_t = \mu_{t-1} + \varepsilon_{3t}$$

The measurement equation relates the observables to the unobservables. The measurement equation can be written as

$$\begin{bmatrix} x_{1t} \\ x_{2t} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_t \\ \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}$$

The State Space Representation of an $AR(p)$ Process

For the examples in the last section, it seemed quite natural to express the model in state space form. However, in other circumstances, appropriately transforming the model can be tricky. The best way to learn is through practice. Towards this end, the remainder of this section consists of a number of examples. There is no particular reason apply a Kalman filter to an $AR(p)$ equation since the variable of interest can be observed directly. However, transforming an $AR(p)$ process into state space form is a good illustration of the technique.

Example 1: The AR(2) model:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \varepsilon_t$$

Since x_{t-1} is identical to itself, it is always possible to write:

$$\begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ x_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}$$

As such, it is possible to define the matrices X_t, A, and ε_t such that:

$$X_t = \begin{bmatrix} x_t \\ x_{t-1} \end{bmatrix}, A_1 = \begin{bmatrix} a_1 & a_2 \\ 1 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}$$

The important point is that we have transformed the $AR(2)$ process into the state equation:

$$X_t = A_1 X_{t-1} + v_t$$

The measurement equation is trivial in that it expresses $y_t = x_t$. Since x_t is actually observed in an $AR(1)$ model, the observation error is necessarily equal to zero. Hence, we can write the measurement equation:

$$y_t = [\ 1 \quad 0\ ] X_t$$

Example 2: The AR(3) model with an intercept

$$x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + \varepsilon_t$$

Define the matrices X_t, A_0, A_1,and ε_t such that:

$$X_t = \begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{bmatrix}, A_0 = \begin{bmatrix} a_0 \\ 0 \\ 0 \end{bmatrix}, A_1 = \begin{bmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix} \qquad (5)$$

You should be able to verify that $X_t = A_0 + A_1 X_{t-1} + v_t$. If you read back the individual equations of this system, it should be clear that the first equation is $x_t = a_0 + a_1 x_{t-1} + a_2 x_{t-2} + a_3 x_{t-3} + \varepsilon_t$, the second is $x_{t-1} = x_{t-1}$, and the third is $x_{t-2} = x_{t-2}$. The measurement equation is $y_t = [\begin{array}{cc} 1 & 0 \end{array}] X_t$. At this point, you should be able to pick up the general pattern for any $AR(p)$ process. The details are given in the next example.

The general $AR(p)$ equation can be written in the form $X_t = A_0 + A_1 X_{t-1} + v_t$ where:

$$X_t = \left[\begin{array}{c} x_t \\ x_{t-1} \\ \vdots \\ x_{t-p} \end{array}\right], A_0 = \left[\begin{array}{c} a_0 \\ 0 \\ \vdots \\ 0 \end{array}\right], A_1 = \left[\begin{array}{ccccc} a_1 & a_2 & a_3 & \cdots & a_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{array}\right], v_t = \left[\begin{array}{c} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{array}\right]$$

Example 3: The State Space Representation of an $MA(q)$ Process

First consider the MA(1) model $x_t = \varepsilon_t + \beta_1 \varepsilon_{t-1}$.

Define X_t such that:

$$X_t = \left[\begin{array}{c} \varepsilon_t \\ \varepsilon_{t-1} \end{array}\right], A_1 = \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array}\right], v_t = \left[\begin{array}{c} \varepsilon_t \\ 0 \end{array}\right]$$

Hence, it is possible to write:

$$\begin{array}{rcl} \left[\begin{array}{c} \varepsilon_t \\ \varepsilon_{t-1} \end{array}\right] & = & \left[\begin{array}{cc} 0 & 0 \\ 1 & 0 \end{array}\right] \left[\begin{array}{c} \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{array}\right] + \left[\begin{array}{c} \varepsilon_t \\ 0 \end{array}\right] \\ X_t & = & A_1 X_{t-1} + v_t \end{array}$$

The observation equation is $y_t = [\begin{array}{cc} 1 & \beta_1 \end{array}] X_t$. Note that there are several state space representations of MA processes. Another way to write the $MA(1)$ model in state space form is to define X_t, A_1 and v_t as follows:

$$X_t = \left[\begin{array}{c} x_t \\ \varepsilon_t \end{array}\right], A_1 = \left[\begin{array}{cc} 0 & \beta_1 \\ 0 & 0 \end{array}\right], v_t = \left[\begin{array}{c} \varepsilon_t \\ \varepsilon_t \end{array}\right]$$

As such, it is possible to write the measurement equation as $y_t = \left[\begin{array}{cc} 1 & 0 \end{array}\right] X_t$ and state equation as $X_t = A_1 X_{t-1} + v_t$, or:

$$\left[\begin{array}{c} x_t \\ \varepsilon_t \end{array}\right] = \left[\begin{array}{cc} 0 & \beta_1 \\ 0 & 0 \end{array}\right] \left[\begin{array}{c} x_{t-1} \\ \varepsilon_{t-1} \end{array}\right] + \left[\begin{array}{c} \varepsilon_t \\ \varepsilon_t \end{array}\right]$$

Now consider the MA(2) model $x_t = \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2}$.

The ‘trick’ here is to recognize that the moving average component $\varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2}$ can be represented in the same way as an $AR(2)$ process. Let: $X_t = \left[\begin{array}{ccc} \varepsilon_t & \varepsilon_{t-1} & \varepsilon_{t-1} \end{array}\right]^T$ so that the state equation becomes:

$$\begin{array}{rcl} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix} & = & \begin{bmatrix} -\beta_1 & -\beta_2 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{t-1} \\ \varepsilon_{t-2} \\ \varepsilon_{t-3} \end{bmatrix} \\ X_t & = & A_1 X_{t-1} \end{array}$$

Now, it is trivial to put the model in state space form. If $\beta = [\ 1 \quad \beta_1 \quad \beta_2\ ]$,the observation equation is:

$$y_t = [\ 1 \quad \beta_1 \quad \beta_2\ ]X_t$$

Estimation of State Space Models

In almost every economic application, the full set of the model's parameters are unknown and need to be estimated. Even if the coefficients of $state$and measurement equations are known the variances of v_t and η_t are usually unknown. It turns out the it is possible to estimate the parameters of the model using maximum likelihood methods. Once the parameters are known, it is possible to write the model in state space form and apply the Kalman filter. As such, it is worthwhile to review the use of maximum likelihood methods for a model with unobserved components. To begin, suppose we have a series of T independently drawn observations for $y_1, y_2, ..., y_T$. The likelihood of each observation will depend on all of the parameters of the data generating process (such as the mean and variance). Obviously, if the parameters of the data generating process change, the likelihood of any particular realization will change as well. To keep the notation simple, let $p(y_t|\mu)$ denote the likelihood of y_t conditional on the value of the parameter vector μ. Since we are assuming that the observations are independent, the likelihood of the sample of observations $y_1, y_2, ..., y_T$ is the product of the likelihoods. If you understand the notation, it should be clear that this joint likelihood, Λ, is

$$\Lambda =_{t=1}^{T} p(y_t|\mu)$$

Another way to think about the issue is to recognize that Λ is an indirect function of u;a different value of μ would have lead to a different realization of $\{y_t\}$ and a different value of Λ. We can let this dependence be denoted by $\Lambda(\mu)$. Once you recognize that different values of μ make some draws for the $\{y_t\}$ sequence more likely than others, it is is natural to want to know the particular value of μ that is the most probable one to have generated the observed realization of the $\{y_t\}$ sequence. In other words, we want to know, conditional on $y_1, ..., y_T$, what is the most likely value of μ that maximizes Λ? Formally, we want to seek the value of μ that solves the following problem

$$\max_{\mu} \Lambda(\mu|y_1, y_2, ..., y_T) \tag{6}$$

The details of maximum likelhood estimation should be familiar to anyone who has taken an introductort econometric class. However, the issue becomes more difficult with processes that are not independent. To take the simplest case, suppose that you want to estimate the values of a_1 and σ_ε^2 in the $AR(1)$ model $y_t = a_1 y_{t-1} + \varepsilon_t$. Although you could estimate a regression equation directly, the goal is to illustrate some of the issues involved with maximum likelihood estimation. If you are willing to assume that the individual values of the $\{\varepsilon_t\}$ series are independently drawn from a normal distribution, it is straightforward to obtain the estimates. Recall that the log of the likelihood of each value of ε_t is

$$-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln\sigma^2-\left(\frac{\varepsilon_t^2}{2\sigma^2}\right) \tag{7}$$

Since the individual values of the ε_t are independent of each other, the log likelihood of the joint realization of the entire series $\varepsilon_1, \varepsilon_2, \varepsilon_3 ... \varepsilon_T$ is the sum of the individual log likelihoods. As such, the log of the joint likelihood (Λ) is

$$\Lambda=-\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\varepsilon_t^2 \tag{8}$$

The next step is to express Λ in terms of the observed values of the $\{y_t\}$ series. The problem is that we began to observe the series in period 1 (i.e., the first observation is $y_1)$ and this value is conditional on the value in period 0. One way to tackle the issue is to impose the initial condition $y_0 = 0$ so that

$$\begin{aligned} \varepsilon_1 &= y_1 \\ \varepsilon_2 &= y_2 - a_1 y_1 \\ \varepsilon_3 &= y_3 - a_1 y_2 \\ &\dots \\ \varepsilon_T &= y_T - a_1 y_{T-1} \end{aligned}$$

Given that we impose $y_0 = 0$, it follows that

$$\Lambda=-\frac{T}{2}\ln(2\pi)-\frac{T}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t-a_1y_{t-1}\right)^2$$

Notice that Λ can be viewed as a function of the values of the $\{y_t\}$ sequence. We seek to determine the parameter set that makes the observed sequence the most likely. Now, to obtain the first-order conditions for a maximum, find the values a_1 and σ^2 that satisfy $\partial\Lambda/\partial a_1 = 0$ and $\partial\Lambda/\partial\sigma^2 = 0$. The resultant values, $\widehat{a}_1$ and $\widehat{\sigma}^2$are the maximum likelihood estimates of σ^2 and a_1. The well-known solution to the first-order conditions is

$$\begin{aligned} \widehat{a}_1 &= \sum_{t=1}^{T} y_t y_{t-1} / \sum_{t=1}^{T} y_t^2 \\ \widehat{\sigma}^2 &= (\frac{1}{T})\sum_{t=1}^{T}(y_t-\widehat{a}_1 y_{t-1})^2 \end{aligned} \qquad (9)\ (10)$$

Similar remarks hold for the maximum likelihood estimates for β and σ_ε^2 in the $MA(1)$ model $y_t = \varepsilon_t + \beta\varepsilon_{t-1}$.If the errors are normally distributed, the log likelihood of ε_t is indentical to that in (7). However, it is not possible to estimate a linear regression equation to find the best fitting value of β because the individual values of the $\{\varepsilon_t\}$ series. As in the $AR(1)$ example, it is necessary to express Λ in terms of the observable $y_1, y_2 ..., y_T$ sequence. Again, to make the transition from the $\{\varepsilon_t\}$ sequence to the $\{y_t\}$ sequence it is necessary to impose an initial condition. Specifically, if we assume that $\varepsilon_0 = 0$, we can write the $\{\varepsilon_t\}$ sequence in terms of the $\{y_t\}$ sequence as

$$\begin{aligned}
\varepsilon_1 &= y_1 \\
\varepsilon_2 &= y_2 - \beta\varepsilon_1 = y_2 - \beta y_1 \\
\varepsilon_3 &= y_3 - \beta\varepsilon_2 = y_3 - \beta(y_2 - \beta y_1) = y_3 - \beta y_2 + \beta^2 y_1 \\
&\ldots \\
\varepsilon_t &= \sum_{i=1}^{t-1}(-\beta)^i y_{t-i} \qquad (11)
\end{aligned}$$

Note that (11) is a convergent sequence as long as the $MA(1)$ process is invertible (i.e., as long as $|\beta| < 1$). If (11) is substituted into (8), we obtain the desired expression for Λ

$$\Lambda = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(\sum_{i=1}^{t-1}(-\beta)^i y_{t-i}\right)^2$$

There are several important points to note about this example. Unlike a regression equation, if you were to actually obtain the first-order conditions for a maximum, you would not get analytic solutions for $\widehat{\sigma}^2$ and $\widehat{\beta}$. Instead of being able to directly solve the first-order equations (9) and (10), you would need to use numerical methods to find the solution. It is also important to note that it is necessary to initialize the system. In any dynamic model, it is necessary to have a set of initial conditions pertaining to the behavior of the variables in the model prior to the first observation. Finally, it is necessary to express the unobserved variables in terms of the obsevables. In models more sophisticated than an $AR(1)$ or an $MA(1)$, all of these issues can become quite difficult.

The maximium likelihood estimation of a state space model a bit more difficult in that there are more parameters to estimate. To best understand the the method, suppose that we want to forecast y_t based on all information up to and including y_{t-1}. In the last chapter, we showed

$$y_{t|t-1} = \beta x_{t|t-1}$$

As such, the one-step ahead forecast error is

$$y_t - y_{t|t-1} = y_t - \beta x_{t|t-1}.$$

In the last section, it was also shown that the variance of this error is

$$E[y_t - y_{t|t-1}]^2 = \beta^2 P_{t|t-1} + \sigma_\eta^2$$

If we are willing to maintain the assumption that the forecast error for normally distributed, the conditional distribution of $y_t - y_{t|t-1}$ is such that

$$y_t - y_{t|t-1} \sim N(y_t - \beta x_{t|t-1}, \beta^2 P_{t|t-1} + \sigma_\eta^2)$$

so that the log likelihood of the forecast error is

$$-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\beta^2 P_{t|t-1} + \sigma_\eta^2) - \frac{1}{2}\left(\frac{y_t - \beta x_{t|t-1}}{\beta^2 P_{t|t-1} + \sigma_\eta^2}\right)^2$$

Given that $x_{t|t-1} = \rho x_{t-1|t-1}$, we can write this likelihood function as

$$-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\beta^2 P_{t|t-1} + \sigma_\eta^2) - \frac{1}{2}\left(\frac{y_t - \beta\rho x_{t-1|t-1}}{\beta^2 P_{t|t-1} + \sigma_\eta^2}\right)^2$$

If we have a sequence of T such forecast errors, under the assumption that are all independent, we can write the joint log likelihood as

$$\Lambda = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}{}_{t=1}^{T} \ln(\beta^2 P_{t|t-1} + \sigma_\eta^2) - \frac{1}{2}{}_{t=1}^{T} \left(\frac{y_t - \beta\rho x_{t-1|t-1}}{\beta^2 P_{t|t-1} + \sigma_\eta^2}\right)^2$$

In the case where x_t is observable, the forecast error variance of x_t (i.e., $P_{t|t-1})$ is zero and $x_{t-1|t-1}$ would be nothing more than the actual value of x_{t-1}.As such, it would be straightforward to maximize the likelihood function to obtain estimates of $\beta,\ \rho$ and σ_η^2. Clearly, this possibility is ruled out in the unobserved components framework so that another estimation strategy needs to be employed. Before proceeding, you should take a moment to try and devise an algorithm that uses the Kalman filter to enable the maximum likelihood estimation. If you understand the logic of the method, you should have reasoned as follows:

1. Write the model in state space form and impose a set of initial conditons for $x_{0|0}$ and $P_{0|0}$
2. Select an initial set of values of $\beta,\ \rho$ and σ_η^2.For this set of initial values and the initial conditions, use the Kalman filter to obtain the subsequent values of $P_{t|t-1}$ and $x_{t-1|t-1}$. Use these values to evaluate the likelihood function Λ.
3. Select a new set of values for $\beta,\ \rho$ and σ_η^2 and use the Kalman filter to create the resultant set of values for $P_{t|t-1}$ and $x_{t-1|t-1}$. Evaluate the likelihood function Λ.
4. Continue to select values for $\beta,\ \rho$ and σ_η^2 until the likelihood function in maximized.

There are a number of numerical techniques that are able to efficiently select new values for $\beta,\ \rho$ and σ_η^2 so that the maximized value of the log likelihood function can be reached quickly. For our purposes, the details of the search strategies used in the various algorithms are not important. What is important is to note that there is no simple way to obtain a closed form solution for the parameters of the model.

Example: The Regression Model with Time Varying Parameters

An important example is the case of a regression equation with time-varying parameters. The usual regression set-up is in the form such that the dependent variable, $y_{t,}$ is linearly related to an independent variable, $x_{t,}$ such that: $y_t = a + bx_t + \varepsilon_t$. In the standard regression context, the coefficients $a_{;}$and b are assumed to be constant. Instead, suppose that theses coefficients are allowed to evolve over time. In particular, suppose that each of the coefficients is an autoregressive process such that

$$\begin{aligned} y_t &= a_t + b_t x_t + \varepsilon_t \\ a_t &= \alpha_0 + \alpha_1 a_{t-1} + v_{1t} \\ b_t &= \beta_0 + \beta_1 b_{t-1} + v_{2t} \end{aligned}$$

The state equation is straightforward to write once it is recognized that we can observe y_t and x_t but the time-varying coefficients the unobservables. The state equation describes the dynamic evolution of the unobserved state variables a_t and b_t. Let the vector of state variables be $[a_t, b_t]^T$. Hence, the state equation is

$$\begin{bmatrix} a_t \\ b_t \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \alpha_1 & 0 \\ 0 & \beta_1 \end{bmatrix} \begin{bmatrix} a_{t-1} \\ b_{t-1} \end{bmatrix} + \begin{bmatrix} v_{1t} \\ v_{2t} \end{bmatrix}$$

The measurement equations related the observables to the unobservables. Let

$$y_t = [1 \qquad x_t] \begin{bmatrix} a_t \\ b_t \end{bmatrix} + \varepsilon_t$$

Now that the model is in state space for, it can be estimated using the Kalman filter.

STATISTICAL TABLES

Table A Empirical Cumulative Distribution of τ

	Significance level			
	0.01	0.025	0.05	0.10
The τ statistic: No Constant or Time Trend ($a_0 = a_2 = 0$)				
Sample Size T				
25	–2.65	–2.26	–1.95	–1.60
50	–2.62	–2.25	–1.95	–1.61
100	–2.60	–2.24	–1.95	–1.61
250	–2.58	–2.24	–1.95	–1.62
300	–2.58	–2.23	–1.95	–1.62
∞	–2.58	–2.23	–1.95	–1.62
The τ_μ statistic: Constant but No Time Trend ($a_2 = 0$)				
25	–3.75	–3.33	–2.99	–2.62
50	–3.59	–3.22	–2.93	–2.60
100	–3.50	–3.17	–2.90	–2.59
250	–3.45	–3.14	–2.88	–2.58
500	–3.44	–3.13	–2.87	–2.57
∞	–3.42	–3.12	–2.86	–2.57
The τ_τ statistic: Constant + Time Trend				
25	–4.38	–3.95	–3.60	–3.24
50	–4.15	–3.80	–3.50	–3.18
100	–4.05	–3.73	–3.45	–3.15
250	–3.99	–3.69	–3.43	–3.13
500	–3.97	–3.67	–3.42	–3.13
∞	–3.96	–3.67	–3.41	–3.12

The table is reproduced from Fuller (1996).

Table B Empirical Distribution of Φ

Significance level	0.10	0.05	0.025	0.01
Sample size T				
		Φ_1		
25	4.12	5.18	6.30	7.88
50	3.94	4.86	5.80	7.06
100	3.86	4.71	5.57	6.70
250	3.81	4.63	5.45	6.52
500	3.79	4.61	5.41	6.47
∞	3.78	4.59	5.38	6.43
		Φ_2		
25	4.67	5.68	6.75	8.21
50	4.31	5.13	5.94	7.02
100	4.16	4.88	5.59	6.50
250	4.07	4.75	5.40	6.22
500	4.05	4.71	5.35	6.15
∞	4.03	4.68	5.31	6.09
		Φ_3		
25	5.91	7.24	8.65	10.61
50	5.61	6.73	7.81	9.31
100	5.47	6.49	7.44	8.73
250	5.39	6.34	7.25	8.43
500	5.36	6.30	7.20	8.34
∞	5.34	6.25	7.16	8.27

TABLE C: Critical Values for the Engle–Granger Cointegration Test

T	1%	5%	10%	1%	5%	10%
		Two Variables			***Three Variables***	
50	–4.123	–3.461	–3.130	–4.592	–3.915	–3.578
100	–4.008	–3.398	–3.087	–4.441	–3.828	–3.514
200	–3.954	–3.368	–3.067	–4.368	–3.785	–3.483
500	–3.921	–3.350	–3.054	–4.326	–3.760	–3.464
		Four Variables			***Five Variables***	
50	–5.017	–4.324	–3.979	–5.416	–4.700	–4.348
100	–4.827	–4.210	–3.895	–5.184	–4.557	–4.240
200	–4.737	–4.154	–3.853	–5.070	–4.487	–4.186
500	–4.684	–4.122	–3.828	–5.003	–4.446	–4.154

The critical values are for cointegrating relations (with a constant in the cointegrating vector) estimated using the Engle–Granger methodology.

Source: Critical values are interpolated using the response surface in MacKinnon (1991)

Table D: Residual Based Cointegration Test with I(1) and I(2) Variables

		Intercept Only				Linear Trend			
m_1	T	$m_2 = 1$		$m_2 = 2$		$m_2 = 1$		$m_2 = 2$	
		prob–value		*prob*–value		*prob*–value		*prob*–value	
		0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
0	50	–4.18	–3.51	–4.70	–4.02	–4.66	–4.01	–5.14	–4.45
	100	–4.09	–3.42	–4.51	–3.86	–4.55	–3.90	–4.93	–4.31
	250	–4.02	–3.38	– 4.35	–3.80	–4.41	–3.83	–4.81	–4.20
1	50	–4.65	–3.93	–5.15	–4.40	–5.11	–4.42	–5.62	–4.89
	100	–4.51	–3.89	–4.85	–4.26	–4.85	–4.26	–5.23	–4.62
	250	–4.39	–3.80	–4.71	–4.18	–4.73	–4.19	–5.11	–4.50
2	50	–4.93	–4.30	–5.54	–4.77	–5.47	–4.74	–5.98	–5.17
	100	–4.81	–4.25	–5.29	–4.59	–5.21	–4.58	–5.59	–4.93
	250	–4.77	–4.16	–5.06	–4.49	–5.07	–4.51	–5.35	–4.80
3	50	–5.38	–4.71	–5.76	–5.08	–5.89	–5.13	–6.23	–5.48
	100	–5.20	–4.56	–5.58	–4.92	–5.52	–4.91	–5.97	–5.25
	250	–5.05	–4.48	–5.44	–4.83	–5.38	–4.78	–5.69	–5.07
4	50	–5.81	–5.09	–6.24	–5.48	–6.35	–5.47	–6.64	–5.82
	100	–5.58	–4.93	–5.88	–5.20	–5.86	–5.20	–6.09	–5.50
	250	–5.39	–4.28	–5.64	–5.07	–5.66	–5.08	–5.95	–5.34

Note: m_1 is the number of $I(1)$ variables and m_2 is the number of $I(2)$ variables on the right–hand side of the multicointegrating relationship.

Source: The critical values for the intercept only case are from Haldrup (1994) and critical values for the linear trend are from Engsted, Gonzalo and Haldrup (1997).

TABLE E: Empirical Distributions of the λ_{max} and λ_{trace} Statistics

Significance level

	10%	5%	2.5%	1%	10%	5%	2.5%	1%
λ_{max} and λ_{trace} statistics without any deterministic regressors								
n–r			λ_{max}				λ_{trace}	
1	2.86	3.84	4.93	6.51	2.86	3.84	4.93	6.51
2	9.52	11.44	13.27	15.69	10.47	12.53	14.43	16.31
3	15.59	17.89	20.02	22.99	21.63	24.31	26.64	29.75
4	21.56	23.80	26.14	28.82	36.58	39.89	42.30	45.58
5	27.62	30.04	32.51	35.17	54.44	59.46	62.91	66.52
λ_{max} and λ_{trace} statistics with drift								
n–r			λ_{max}				λ_{trace}	
1	2.69	3.76	4.95	6.65	2.69	3.76	4.95	6.65
2	12.07	14.07	16.05	18.63	13.33	15.41	17.52	20.04
3	18.60	20.97	23.09	25.52	26.79	29.68	32.56	35.65
4	24.73	27.07	28.98	32.24	43.95	47.21	50.35	54.46
5	30.90	33.46	35.71	38.77	64.84	68.52	71.80	76.07
λ_{max} and λ_{trace} statistics with a constant in the cointegrating vector								
			λ_{max}				λ_{trace}	
1	7.52	9.24	10.80	12.97	7.52	9.24	10.80	12.95
2	13.75	15.67	17.63	20.20	17.85	19.96	22.05	24.60
3	19.77	22.00	24.07	26.81	32.00	34.91	37.61	41.07
4	25.56	28.14	30.32	33.24	49.65	53.12	56.06	60.16
5	31.66	34.40	36.90	39.79	71.86	76.07	80.06	84.45

Source: Osterwald–Lenum (1992).

Table F: Critical Values for $\beta_1 = 0$ in the Error–Correction Model

k		$T^a = 50$	$T^a = 100$	$T^a = 200$	$T^a = 500$
No Intercept or Trend (*d* = 0)					
2	1%	–3.309	–3.259	–3.235	–3.220
	5%	–2.625	–2.609	–2.602	–2.597
	10%	–2.273	–2.268	–2.266	–2.265
3	1%	–3.746	–3.683	–3.652	–3.633
	5%	–3.047	–3.026	–3.016	–3.009
	10%	–2.685	–2.680	–2.677	–2.675
4	1%	–4.088	–4.015	–3.979	–3.957
	5%	–3.370	–3.348	–3.337	–3.331
	10%	–3.000	–2.997	–2.995	–2.994
Intercept but no Trend (*d* = 1)					
2	1%	–3.954	–3.874	–3.834	–3.811
	5%	–3.279	–3.247	–3.231	–3.221
	10%	–2.939	–2.924	–2.916	–2.911
3	1%	–4.268	–4.181	–4.138	–4.112
	5%	–3.571	–3.538	–3.522	–3.512
	10%	–3.216	–3.205	–3.199	–3.195
4	1%	–4.537	–4.446	–4.401	–4.374
	5%	–3.819	–3.789	–3.774	–3.765
	10%	–3.453	–3.447	–3.444	–3.442
Intercept and Trend (*d* = 2)					
2	1%	–4.451	–4.350	–4.299	–4.269
	5%	–3.778	–3.733	–3.710	–3.696
	10%	–3.440	–3.416	–3.405	–3.398
3	1%	–4.712	–4.605	–4.552	–4.519
	5%	–4.014	–3.971	–3.949	–3.935
	10%	–3.662	–3.643	–3.634	–3.629
4	1%	–4.940	–4.831	–4.776	–4.743
	5%	–4.221	–4.182	–4.162	–4.150
	10%	–3.857	–3.846	–3.840	–3.837

Note: T^a is the adjusted sample size equal to $T - (2k - 1) - d$ where T is the usable sample size, d is the number of deterministic regressors, and k is the number of $I(1)$ variables in the model. The critical values are calculated using equation (26) in Ericsson and MacKinnon (2002).

Table G: Critical Values for Threshold Unit Roots

Panel (a): Consistent Estimate of the Threshold Using the TAR Model

T	No Lagged Changes				One Lagged Change				Four Lagged Changes			
	90%	95%	97.5	99%	90%	95%	97.5	99%	90%	95%	97.5	99%
50	5.15	6.19	7.25	8.64	5.55	6.62	7.66	9.10	5.49	6.55	7.59	9.00
100	5.08	6.06	6.93	8.19	5.39	6.34	7.30	8.54	5.38	6.32	7.29	8.56
250	5.11	6.03	6.88	8.04	5.26	6.12	6.99	8.14	5.36	6.29	7.15	8.35

Panel (b): Consistent Estimate of the Threshold Using the M–TAR Model

	No Lagged Changes				One Lagged Change				Four Lagged Changes			
T	90%	95%	97.5	99%	90%	95%	97.5	99%	90%	95%	97.5	99%
50	5.02	6.05	7.09	8.59	4.98	6.07	7.15	8.56	4.93	5.96	7.01	8.48
100	4.81	5.77	6.73	7.99	4.77	5.71	6.56	7.90	4.74	5.70	6.67	7.97
250	4.70	5.64	6.51	7.64	4.64	5.54	6.40	7.56	4.64	5.54	6.39	7.61

Panel (c): Known Threshold Value in the M–TAR Model

	No Lagged Changes				One Lagged Change				Four Lagged Changes			
T	90%	95%	97.5	99%	90%	95%	97.5	99%	90%	95%	97.5	99%
50	4.21	5.19	6.15	7.55	4.12	5.11	6.05	7.25	3.82	4.73	5.65	6.84
100	4.11	5.04	5.96	7.10	4.08	4.97	5.87	7.06	3.81	4.72	5.63	6.83
250	4.08	4.97	5.83	6.91	4.05	4.93	5.78	6.83	3.69	4.71	5.63	6.78