SUPPLEMENTARY MANUAL TO ACCOMPANY

APPLIED ECONOMETRIC TIME SERIES (3rd edition)

Walter Enders

University of Alabama

Prepared by

Karl David Boulware

University of Alabama

Walter Enders

University of Alabama

Jared Levant

University of Alabama

Contents

Forward – Versus Backward – Looking 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

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

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

p_t = the logarithm of price level in t

p_{t+1}^e = 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 ($p_{t+1}^e - p_t$). 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:

$$p_{t+1}^{e} = p_{t+1}$$

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

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

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

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

Substituting this challenge solution into the above, we obtain

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

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

$$\begin{aligned}
& b_0 - b_0(1+\beta)/\beta = (\alpha - m)/\beta && \Rightarrow\ b_0 && = m - \alpha \\
& -\alpha_0(1+\beta)/\beta = -1/\beta && \Rightarrow\ \alpha_0 && = 1/(1+\beta) \\
& \alpha_0 - \alpha_1(1+\beta)/\beta = 0 && \Rightarrow\ \alpha_1 && = \beta/(1+\beta)^2 \\
& . \\
& . \\
& . \\
& \alpha_i - \alpha_{i+1}(1+\beta)/\beta = 0 && \Rightarrow\ \alpha_i && = \beta^i/(1+\beta)^{i+1}
\end{aligned}$$

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} - $(1+1/\beta)p_t = 0$. 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 $m - \alpha$ 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

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

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.

Stability of Higher-Order Systems

Supplement to Appendix 1.2

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

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

Denote the *n* characteristic roots by α_1, α_2, ... α_n. Given the results in Section 4, the linear combination $A_1\alpha_1^t + A_2\alpha_2^t + \ldots + A_n\alpha_n^t$ 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.

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

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

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

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.

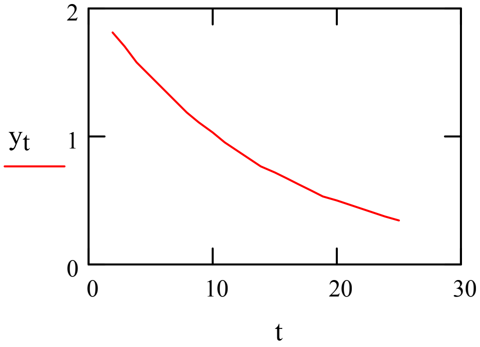
Additional Practice in Finding Homogeneous Solutions

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.

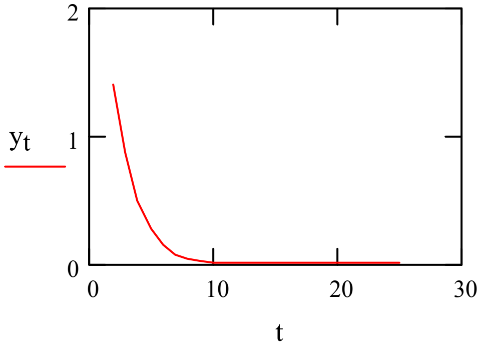


Example 2: Another AR(2) model: $y_t = a_0 + 0.9y_{t-1} - 0.2y_{t-2}$

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:

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

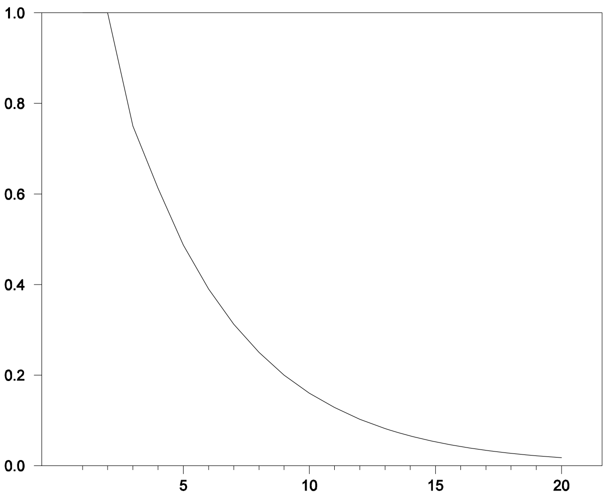
3. We can now compute the two characteristic roots:

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

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 4: Fibonacci's Sequence

The famed sequence of Leonardo Fibonacci can be represented by the following second-order difference equation and initial conditions

$$y_t = y_{t\text{-}1} + y_{t\text{-}2};\ y_1 = 1,\ y_2 = 1$$

We know the homogeneous solution to this equation has the form$y_t^h = A\alpha^t$. Substituting this form into the above and setting the equation equal to zero yields

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

Dividing both sides by $A\alpha^{t-2}$ reduces our equation to

$$\alpha^2 - \alpha - 1 = 0$$

Using the quadratic formula we find the roots of this characteristics equation to be

$$\alpha_1, \alpha_2 = \frac{1 \pm \sqrt{5}}{2}$$

We can now substitute these characteristic roots into the functional form to get

$$y_t = A_1(\tfrac{1+\sqrt{5}}{2})^t + A_2(\tfrac{1-\sqrt{5}}{2})^t$$

Utilizing the first order conditions $y_1 = 1$ and $y_2 = 1$ we can now solve for the constants A_1 and A_2.

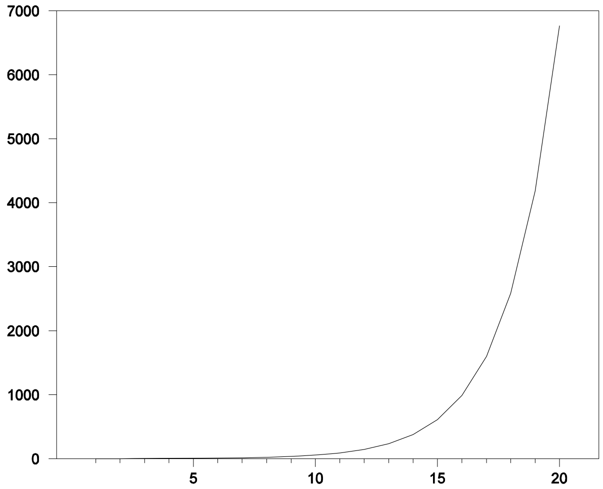
$$y_1 = 1 = A_1(\tfrac{1+\sqrt{5}}{2})^1 + A_2(\tfrac{1-\sqrt{5}}{2})^1$$

$$y_2 = 1 = A_1(\frac{1+\sqrt{5}}{2})^2 + \ A_2(\frac{1-\sqrt{5}}{2})^2$$

$$A_1, A_2 = \frac{1}{\sqrt{5}}$$

Therefore the homogeneous solution is $y_t = \frac{1}{\sqrt{5}}(\frac{1+\sqrt{5}}{2})^t + \frac{1}{\sqrt{5}}(\frac{1-\sqrt{5}}{2})^t$

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



Example 5: An example with complex roots

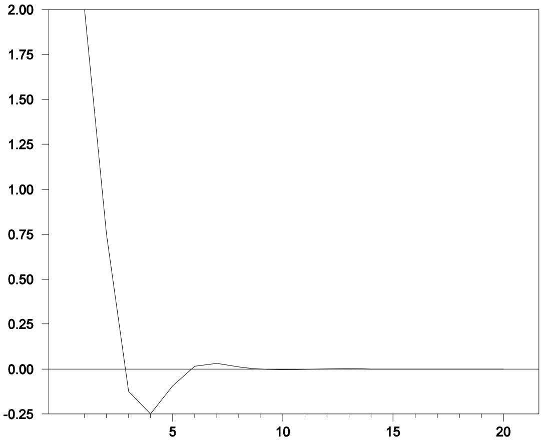
Let us analyze the homogenous solution to a second-order differential equation with complex roots and no initial conditions

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

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.


2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
-0.25
5
10
15
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

$$c = 0.55c + 0.2c + 2$$
$$c = \frac{2}{1-0.55-0.2} = 4$$

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

$$\lim_{t\to\infty} y_t = \bar{y}$$
$$\bar{y} = 0.55\bar{y} + 0.2\bar{y} + 2$$
$$\bar{y} = 4$$

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

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

$\mathrm{y}_{t\text{-}1} = 0.5\mathrm{y}_t - 0.5\varepsilon_t$

and updating one period:

$\mathrm{y}_t = 0.5y_{t+1} - 0.5\varepsilon_{t+1}$

Continuing to iterate forward:

$y_t = 0.5\mathrm{y}_{t+1} - 0.5\varepsilon_{t+1} \ = 0.5[0.5y_{t+2} - 0.5\varepsilon_{t+2}]$

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

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 $\mathrm{P_t}$ 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 $\mathrm{P}_t = \frac{1}{1+r}\mathrm{P}_{t+1} + \mathrm{d}_t$ more closely.

Again the backwards solution of $\mathrm{P}_{t+1} = (1+\mathrm{r})\mathrm{P}_t + (1+\mathrm{r})\mathrm{d}_t$ makes no sense. What about the forward solution? Using the method of undetermined coefficients we have:

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

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.

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 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:

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

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 $(1/T)^{0.5}$.

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, ...).

Examine the ACF and the PACF of the transformed data in order to determine the plausible models.

Estimation
Estimate the plausible models and select the best. You should entertain the possibility of several models and estimate each. The 'best' will have coefficients that are statistically signifcant and a good fit. (use the AIC or SBC to determine the fit).

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.

Heteroskedasticity-Autocorrelation-Consistent (HAC) Estimator

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 $y_t = \beta_0 + \beta_1 x_t + u_t$. The OLS estimator for this model is

$$\hat{\beta}_1 = \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

Taking the difference between the predicted estimator and the actual estimator we get

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})u_t}{\frac{1}{T}\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

Therefore when the sample is large

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{T}\sum_{t=1}^{T} \nu_t}{\sigma_x^2}$$

Where $\nu_t = (X_t - X)u_t$. Taking the variance of both sides yields

$$Var(\hat{\beta}_1) = \frac{1}{(\sigma_x^2)^2} Var(\frac{1}{T}\sum_{t=1}^{T} \nu_t)$$

In large samples. Now let us consider the simple case when T = 2

$$Var\left(\frac{1}{2}\sum_{t=1}^{2} \nu_t\right) = Var[\frac{1}{2}(\nu_1 + \nu_2)$$

$$= \tfrac{1}{4}[Var(\nu_1) + Var(\nu_2) + 2Cov(\nu_1, \nu_2)]$$

$$= \tfrac{1}{4}[2\sigma_\nu^2 + 2\rho_1\sigma_\nu^2] \quad \text{where } \rho_1\sigma_\nu^2 = 2Cov(\nu_1, \nu_2)$$

$$= \frac{1}{2}\sigma_v^2 w_2$$

Where $w_2 = (1 + \rho_1)$. It is important to note that when there is no correlation between $v_1$and v_2 $(\rho_1 = 0)$ then $w_2 = 1$ which gives the usual formula for the variance estimate in cross-section data. But in time series data $\rho_1 \neq 0$ so the usual variance formula does not apply. Therefore OLS standard errors are wrong in the presence of auto-correlated error terms.

Let us now derive an expression for the variance of estimators with general T

$$Var\left(\frac{1}{T}\sum_{t=1}^{T} v_t\right) = \frac{\sigma_v^2}{T} w_T$$

Therefore

$$Var(\hat{\beta}_1) = \frac{1}{T}\frac{\sigma_v^2}{(\sigma_x^2)^2} w_T$$

Where $w_T = 1 + 2\sum_{j=1}^{T-1}(\frac{T-j}{T})\,\rho_j$. 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.

Value at Risk

This material is supplementary to the GARCH modeling presented in Chapter 3

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:

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

where the conditional variance h_t follows an IGARCH process. Let

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

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:

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

and that

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

Also suppose that the values of r_t, $\varepsilon_{t\text{-}1}$ and $h_{t\text{-}1}$ 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.

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:

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

$$y_t = z_t/m$$

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.

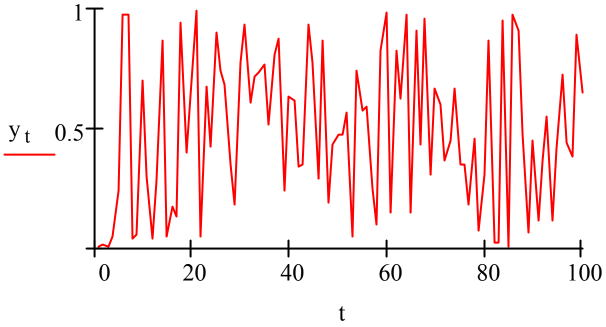
The point is that not all parameter choices for α, m, and λ are well-behaved. Note that λ is called the multiplier. Clearly, λ needs to be greater than unity so that the numbers do not converge to zero. Nevertheless, some values of $\lambda > 1$ will lead to poorly behaved sequences. Also note that m 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 $y_2 = 1.36536 \text{ x } 10^{-6}$
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.

Phillips-Perron Test Statistics

This Material Supplements the Discussion on Pages 219 – 222 of Chapter 4

Given the discussion on pages 219 – 222, 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:

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

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.,

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

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:

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

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

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

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,

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

and ω is the number of estimated autocorrelations.

Note that S^2 and $\sigma_{T\varpi}^2$ 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_ts_{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:

$$E_ts_{t+90} = f_t.$$

or:

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

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

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:

$$y_t = a_0 + a_1y_{t-1} + a_2t + \mu_t$$

The efficient market hypothesis asserts that <u>ex ante</u> 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 <u>ex ante</u> expected returns may be zero, the <u>ex post</u> 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.

Unobserved Component Models

Supplement to Section 4.1

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:

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

The first-order condition is

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

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 $\sigma_\eta^{\,2}$, $b \rightarrow 1$; as σ^2 becomes very small relative to $\sigma_\eta^{\,2}$, $b \rightarrow 0$. 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 σ^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:

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

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, μ_y, σ_{xy}, and σ^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 $t \leq 0$, 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:

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

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:

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

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:

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

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.

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 $\rho, \rho^2, \rho^3, \ldots$, 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

$$\begin{aligned} y_t &= \beta x_t + \eta_t \\ x_t &= \rho x_{t-1} + \varepsilon_t \end{aligned} \qquad (18)$$

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:

$$k_1 = 1/(1+1) = 0.5$$

$$x_{1|1} = 0.5(2.0579) = 1.029\,0$$

$$P_{1|1} = 0.5P_{1|0} = 0.5$$

We next use this information to form $x_{2|1}$ and $P_{2|1}$. From the prediction equations, we obtain:

$$x_{2|1} = 0.5x_{1|1} = 0.5(1.029\,0) = 0.5145$$

$$P_{2|1} = 0.25P_{1|1} + 1 = 0.25(0.5) + 1 = 1.1250$$

Once we observe y_2, we use the updating equations to obtain:

$$k_2 = P_{2|1}/(P_{2|1} + 1) = 1.1250/(1.1250 + 1) = 0.529\,4.$$

$$x_{2|2} = x_{2|1} + k_2(y_2 - x_{2|1}) = 0.514\,5 + 0.529\,4(0.4984 - 0.514\,5) = 0.505\,9.$$

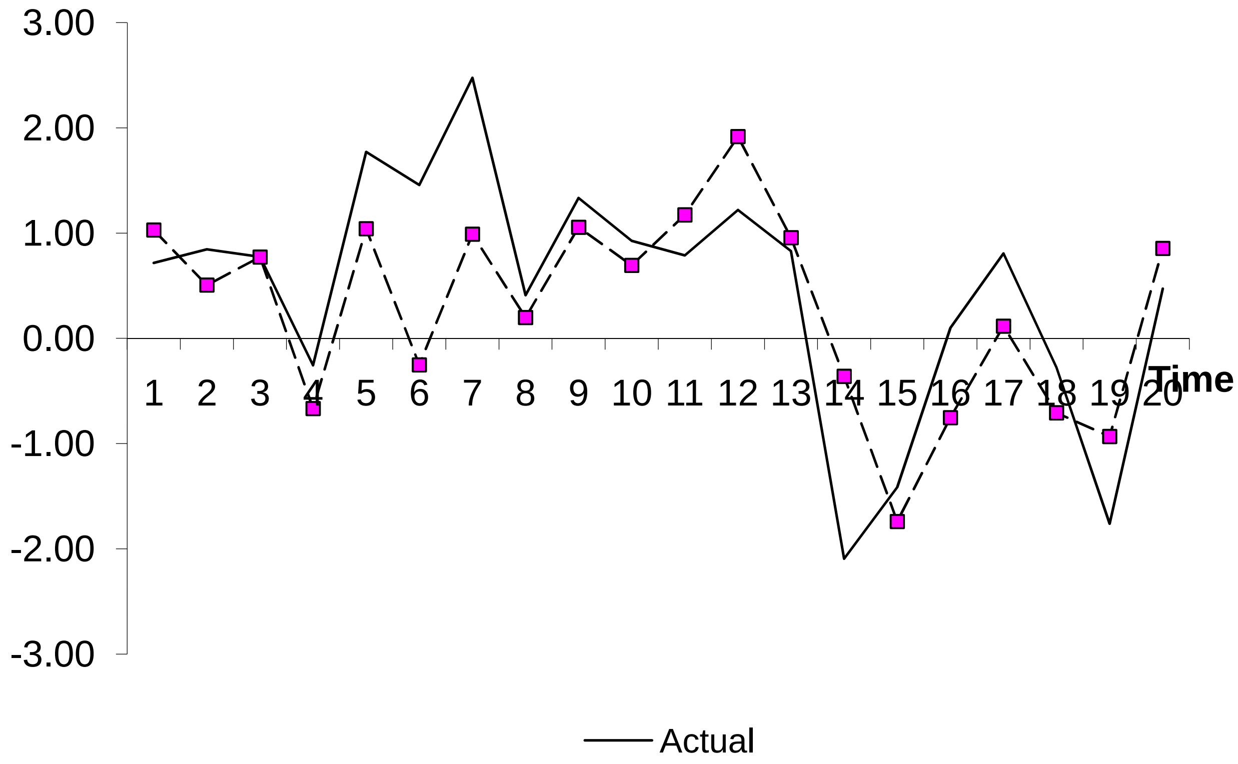
$$P_{2|2} = (1 - k_2)P_{2|1} = (1 - 0.529\,4)1.1250 = 0.529\,4.$$

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	$\varepsilon_{t\|t-1}$	$P_{t\|t-1}$	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 $\varepsilon_t = 0.5\varepsilon_{t-1} + v_t$ and values of y_t are $\varepsilon_t + \eta_t$. The values of $\varepsilon_{t|t-1}$ and $P_{t|t-1}$ 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: $P_{t|t} = (1 - k_t)P_{t|t-1} = (1 - 0.5311)(1.1328) = 0.5311$.

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:

$$y_t = \beta x_t + \eta_t \tag{3}$$

$$x_t = \rho x_{t\text{-}1} + v_t \tag{4}$$

[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:

$$X_t = \begin{bmatrix} x_t \\ x_{t-1} \\ x_{t-2} \end{bmatrix}, A_0 = \begin{bmatrix} a_0 \\ 0 \\ 0 \end{bmatrix}, A_1 = \begin{bmatrix} a_1 & a_2 & a_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ 0 \end{bmatrix} \tag{5}$$

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 $y_t = [\ 1 \quad 0\ ] X_t$. 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:

$$X_t = \begin{bmatrix} x_t \\ x_{t-1} \\ \vdots \\ x_{t-p} \end{bmatrix}, A_0 = \begin{bmatrix} a_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, A_1 = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

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:

$$X_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix}, A_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix}$$

Hence, it is possible to write:

$$\begin{aligned} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_{t-1} \\ \varepsilon_{t-2} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ 0 \end{bmatrix} \\ X_t &= A_1 X_{t-1} + v_t \end{aligned}$$

The observation equation is $y_t = [\ 1 \quad \beta_1\ ] X_t$. 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:

$$X_t = \begin{bmatrix} x_t \\ \varepsilon_t \end{bmatrix}, A_1 = \begin{bmatrix} 0 & \beta_1 \\ 0 & 0 \end{bmatrix}, v_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_t \end{bmatrix}$$

As such, it is possible to write the measurement equation as $y_t = \begin{bmatrix} 1 & 0 \end{bmatrix} X_t$ and state equation as $X_t = A_1 X_{t-1} + v_t$, or:

$$\begin{bmatrix} x_t \\ \varepsilon_t \end{bmatrix} = \begin{bmatrix} 0 & \beta_1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ \varepsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_t \\ \varepsilon_t \end{bmatrix}$$

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: $X_t = \begin{bmatrix} \varepsilon_t & \varepsilon_{t-1} & \varepsilon_{t-1} \end{bmatrix}^T$ so that the state equation becomes:

$$\begin{aligned}\begin{bmatrix}\varepsilon_t\\ \varepsilon_{t-1}\\ \varepsilon_{t-2}\end{bmatrix} &= \begin{bmatrix}-\beta_1 & -\beta_2 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\end{bmatrix}\begin{bmatrix}\varepsilon_{t-1}\\ \varepsilon_{t-2}\\ \varepsilon_{t-3}\end{bmatrix}\\ X_t &= A_1 X_{t-1}\end{aligned}$$

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

$$\max_{\mu} \Lambda(\mu|y_1, y_2, ..., y_T) \qquad (6)$$

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^2_ε 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

$$\begin{aligned}
\varepsilon_1 &= y_1 \\
\varepsilon_2 &= y_2 - a_1 y_1 \\
\varepsilon_3 &= y_3 - a_1 y_2 \\
&... \\
\varepsilon_T &= y_T - a_1 y_{T-1}
\end{aligned}$$

Given that we impose $y_0 = 0$, it follows that

$$\Lambda = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - a_1 y_{t-1})^2$$

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

$$\begin{aligned}
\widehat{a}_1 &= \sum_{t=1}^{T} y_t y_{t-1} / \sum_{t=1}^{T} y_t^2 \\
\widehat{\sigma}^2 &= (\frac{1}{T})\sum_{t=1}^{T}(y_t - \widehat{a}_1 y_{t-1})^2 && (9)\ (10)
\end{aligned}$$

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

$$\begin{aligned}
\varepsilon_1 &= y_1 \\
\varepsilon_2 &= y_2 - \beta\varepsilon_1 = y_2 - \beta y_1 \\
\varepsilon_3 &= y_3 - \beta\varepsilon_2 = y_3 - \beta(y_2 - \beta y_1) = y_3 - \beta y_2 + \beta^2 y_1 \\
&\dots \\
\varepsilon_t &= \sum_{i=1}^{t-1}(-\beta)^i y_{t-i} \qquad (11)
\end{aligned}$$

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.