Using [10.A.5] and [10.A.7], this can be expressed as

$$
\begin{aligned}
E(y_{it} - \mu_i)(y_{j,t-s} - \mu_j) &= E\left\{\left[\sum_{l=1}^{n} z_t(i, l)\right]\left[\sum_{m=1}^{n} z_{t-s}(j, m)\right]\right\} \\
&= \sum_{l=1}^{n}\sum_{m=1}^{n} E\{z_t(i, l)\cdot z_{t-s}(j, m)\} \\
&= \sum_{l=1}^{n}\sum_{m=1}^{n}\sum_{v=0}^{\infty} \psi_{il}^{(s+v)}\psi_{jm}^{(v)}\sigma_{lm} \\
&= \sum_{v=0}^{\infty}\sum_{l=1}^{n}\sum_{m=1}^{n} \psi_{il}^{(s+v)}\psi_{jm}^{(v)}\sigma_{lm}.
\end{aligned}
\qquad [10.A.8]
$$

But $\Sigma_{l=1}^{n}\Sigma_{m=1}^{n}\psi_{il}^{(s+v)}\psi_{jm}^{(v)}\sigma_{lm}$ is the row i, column j element of $\mathbf{\Psi}_{v+s}\mathbf{\Omega}\mathbf{\Psi}_v'$. Thus, [10.A.8] states that the row i, column j element of $\mathbf{\Gamma}_s$ is given by the row i, column j element of $\Sigma_{v=0}^{\infty}\mathbf{\Psi}_{v+s}\mathbf{\Omega}\mathbf{\Psi}_v'$, as asserted in part (a).

Proof of (b). Define $h_s(\cdot)$ to be the moment in [10.A.7]:

$$h_s(i, j, l, m) \equiv E\{z_t(i, l)\cdot z_{t-s}(j, m)\} = \sum_{v=0}^{\infty}\psi_{il}^{(s+v)}\psi_{jm}^{(v)}\sigma_{lm};$$

and notice that the sequence $\{h_s(\cdot)\}_{s=0}^{\infty}$ is absolutely summable:

$$
\begin{aligned}
\sum_{s=0}^{\infty}|h_s(i, j, l, m)| &\le \sum_{s=0}^{\infty}\sum_{v=0}^{\infty}|\psi_{il}^{(s+v)}|\cdot|\psi_{jm}^{(v)}|\cdot|\sigma_{lm}| \\
&= |\sigma_{lm}|\sum_{v=0}^{\infty}|\psi_{jm}^{(v)}|\sum_{s=0}^{\infty}|\psi_{il}^{(s+v)}| \\
&\le |\sigma_{lm}|\sum_{v=0}^{\infty}|\psi_{jm}^{(v)}|\sum_{s=0}^{\infty}|\psi_{il}^{(s)}| \\
&< \infty.
\end{aligned}
\qquad [10.A.9]
$$

Furthermore, the row i, column j element of $\mathbf{\Gamma}_s$ was seen in [10.A.8] to be given by

$$\gamma_{ij}^{(s)} = \sum_{l=1}^{n}\sum_{m=1}^{n} h_s(i, j, l, m).$$

Hence,

$$\sum_{s=0}^{\infty}|\gamma_{ij}^{(s)}| \le \sum_{s=0}^{\infty}\sum_{l=1}^{n}\sum_{m=1}^{n}|h_s(i, j, l, m)| = \sum_{l=1}^{n}\sum_{m=1}^{n}\sum_{s=0}^{\infty}|h_s(i, j, l, m)|. \qquad [10.A.10]$$

From [10.A.9], there exists an $M < \infty$ such that

$$\sum_{s=0}^{\infty}|h_s(i, j, l, m)| < M$$

for any value of i, j, l, or m. Hence, [10.A.10] implies

$$\sum_{s=0}^{\infty}|\gamma_{ij}^{(s)}| < \sum_{l=1}^{n}\sum_{m=1}^{n} M = n^2M < \infty,$$

confirming that the row i, column j element of $\{\mathbf{\Gamma}_s\}_{s=0}^{\infty}$ is absolutely summable, as claimed by part (b).

Proof of (c). Essentially the identical algebra as in the proof of Proposition 7.10 establishes that

$$
\begin{aligned}
&E|z_{t_1}(i_1,l_1)\cdot z_{t_2}(i_2,l_2)\cdot z_{t_3}(i_3,l_3)\cdot z_{t_4}(i_4,l_4)| \\
&= E\left|\left\{\sum_{v_1=0}^{\infty}\psi_{i_1,l_1}^{(v_1)}\varepsilon_{l_1,t_1-v_1}\right\}\cdot\left\{\sum_{v_2=0}^{\infty}\psi_{i_2,l_2}^{(v_2)}\varepsilon_{l_2,t_2-v_2}\right\}\right. \\
&\quad\left.\cdot\left\{\sum_{v_3=0}^{\infty}\psi_{i_3,l_3}^{(v_3)}\varepsilon_{l_3,t_3-v_3}\right\}\cdot\left\{\sum_{v_4=0}^{\infty}\psi_{i_4,l_4}^{(v_4)}\varepsilon_{l_4,t_4-v_4}\right\}\right| \qquad [10.A.11] \\
&\le \sum_{v_1=0}^{\infty}\sum_{v_2=0}^{\infty}\sum_{v_3=0}^{\infty}\cdot\sum_{v_4=0}^{\infty}|\psi_{i_1,l_1}^{(v_1)}\psi_{i_2,l_2}^{(v_2)}\psi_{i_3,l_3}^{(v_3)}\psi_{i_4,l_4}^{(v_4)}| \\
&\quad\times E|\varepsilon_{l_1,t_1-v_1}\varepsilon_{l_2,t_2-v_2}\varepsilon_{l_3,t_3-v_3}\varepsilon_{l_4,t_4-v_4}| \\
&< \infty.
\end{aligned}
$$

Now,

$$
\begin{aligned}
E|y_{i_1,t_1}y_{i_2,t_2}y_{i_3,t_3}y_{i_4,t_4}| &= E\left|\mu_{i_1}+\sum_{l_1=1}^{n}z_{t_1}(i_1,l_1)\right|\cdot\left|\mu_{i_2}+\sum_{l_2=1}^{n}z_{t_2}(i_2,l_2)\right| \\
&\quad\cdot\left|\mu_{i_3}+\sum_{l_3=1}^{n}z_{t_3}(i_3,l_3)\right|\cdot\left|\mu_{i_4}+\sum_{l_4=1}^{n}z_{t_4}(i_4,l_4)\right| \\
&\le E\left\{|\mu_{i_1}|+\sum_{l_1=1}^{n}|z_{t_1}(i_1,l_1)|\right\}\cdot\left\{|\mu_{i_2}|+\sum_{l_2=1}^{n}|z_{t_2}(i_2,l_2)|\right\} \\
&\quad\cdot\left\{|\mu_{i_3}|+\sum_{l_3=1}^{n}|z_{t_3}(i_3,l_3)|\right\}\cdot\left\{|\mu_{i_4}|+\sum_{l_4=1}^{n}|z_{t_4}(i_4,l_4)|\right\}.
\end{aligned}
$$

But this is a finite sum involving terms of the form of [10.A.11]—which were seen to be finite—along with terms involving first through third moments of z, which must also be finite.

Proof of (d). Notice that

$$
z_t(i,l)\cdot z_{t-s}(j,m) = \sum_{r=0}^{\infty}\sum_{v=0}^{\infty}\psi_{il}^{(r)}\psi_{jm}^{(v)}\varepsilon_{l,t-r}\varepsilon_{m,t-s-v}.
$$

The same argument leading to [7.2.14] can be used to establish that

$$
(1/T)\sum_{t=1}^{T}z_t(i,l)\cdot z_{t-s}(j,m)\xrightarrow{p}E\{z_t(i,l)\cdot z_{t-s}(j,m)\}. \qquad [10.A.12]
$$

To see that [10.A.12] implies ergodicity for the second moments of $\mathbf{y}$, notice from [10.A.5] that

$$
\begin{aligned}
(1/T)\sum_{t=1}^{T}y_{it}y_{j,t-s} &= (1/T)\sum_{t=1}^{T}\left[\mu_i+\sum_{l=1}^{n}z_t(i,l)\right]\left[\mu_j+\sum_{m=1}^{n}z_{t-s}(j,m)\right] \\
&= \mu_i\mu_j+\mu_i\sum_{m=1}^{n}\left[(1/T)\sum_{t=1}^{T}z_{t-s}(j,m)\right]+\mu_j\sum_{l=1}^{n}\left[(1/T)\sum_{t=1}^{T}z_t(i,l)\right] \\
&\quad+\sum_{l=1}^{n}\sum_{m=1}^{n}\left[(1/T)\sum_{t=1}^{T}z_t(i,l)z_{t-s}(j,m)\right] \\
&\xrightarrow{p}\mu_i\mu_j+\mu_i\sum_{m=1}^{n}E[z_{t-s}(j,m)]+\mu_j\sum_{l=1}^{n}E[z_t(i,l)] \\
&\quad+\sum_{l=1}^{n}\sum_{m=1}^{n}E[z_t(i,l)z_{t-s}(j,m)] \\
&= E\left\{\left[\mu_i+\sum_{l=1}^{n}z_t(i,l)\right]\left[\mu_j+\sum_{m=1}^{n}z_{t-s}(j,m)\right]\right\} \\
&= E[y_{it}y_{j,t-s}],
\end{aligned}
$$

as claimed. ■

■ **Proof of Proposition 10.3.** Writing out [10.2.11] explicitly,

$$\begin{aligned}\mathbf{H}(L)\mathbf{\Psi}(L) &= (\cdots + \mathbf{H}_{-1}L^{-1} + \mathbf{H}_0L^0 + \mathbf{H}_1L^1 + \cdots)\\ &\quad\times(\mathbf{\Psi}_0L^0 + \mathbf{\Psi}_1L^1 + \mathbf{\Psi}_2L^2 + \cdots),\end{aligned}$$

from which the coefficient on L^k is

$$\mathbf{B}_k = \mathbf{H}_k\mathbf{\Psi}_0 + \mathbf{H}_{k-1}\mathbf{\Psi}_1 + \mathbf{H}_{k-2}\mathbf{\Psi}_2 + \cdots. \qquad [10.A.13]$$

Let $b_{ij}^{(k)}$ denote the row i, column j element of $\mathbf{B}_k$, and let $h_{ij}^{(k)}$ and $\psi_{ij}^{(k)}$ denote the row i, column j elements of $\mathbf{H}_k$ and $\mathbf{\Psi}_k$, respectively. Then the row i, column j element of the matrix equation [10.A.13] states that

$$b_{ij}^{(k)} = \sum_{m=1}^{n} h_{im}^{(k)}\psi_{mj}^{(0)} + \sum_{m=1}^{n} h_{im}^{(k-1)}\psi_{mj}^{(1)} + \sum_{m=1}^{n} h_{im}^{(k-2)}\psi_{mj}^{(2)} + \cdots = \sum_{v=0}^{\infty}\sum_{m=1}^{n} h_{im}^{(k-v)}\psi_{mj}^{(v)}.$$

Thus,

$$\begin{aligned}\sum_{k=-\infty}^{\infty}|b_{ij}^{(k)}| &= \sum_{k=-\infty}^{\infty}\left|\sum_{v=0}^{\infty}\sum_{m=1}^{n} h_{im}^{(k-v)}\psi_{mj}^{(v)}\right|\\ &\le \sum_{k=-\infty}^{\infty}\sum_{v=0}^{\infty}\sum_{m=1}^{n}|h_{im}^{(k-v)}\psi_{mj}^{(v)}| \qquad [10.A.14]\\ &= \sum_{m=1}^{n}\sum_{v=0}^{\infty}|\psi_{mj}^{(v)}|\sum_{k=-\infty}^{\infty}|h_{im}^{(k-v)}|.\end{aligned}$$

But since $\{\mathbf{H}_k\}_{k=-\infty}^{\infty}$ and $\{\mathbf{\Psi}_k\}_{k=-\infty}^{\infty}$ are absolutely summable,

$$\sum_{k=-\infty}^{\infty}|h_{im}^{(k-v)}| < M_1 < \infty$$

$$\sum_{v=0}^{\infty}|\psi_{mj}^{(v)}| < M_2 < \infty.$$

Thus, [10.A.14] becomes

$$\sum_{k=-\infty}^{\infty}|b_{ij}^{(k)}| < \sum_{m=1}^{n} M_1M_2 < \infty. \quad ■$$

■ **Proof of Proposition 10.4.** Let $\mathbf{A}$ be $(m \times n)$, $\mathbf{B}$ be $(n \times r)$, and $\mathbf{C}$ be $(r \times q)$. Let the $(n \times 1)$ vector $\mathbf{b}_i$ denote the ith column of $\mathbf{B}$, and let c_{ij} denote the row i, column j element of $\mathbf{C}$. Then

$$\begin{aligned}\mathbf{ABC} &= \mathbf{A}[\mathbf{b}_1\ \ \mathbf{b}_2\ \cdots\ \mathbf{b}_r]\begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1q}\\ c_{21} & c_{22} & \cdots & c_{2q}\\ \vdots & \vdots & \cdots & \vdots\\ c_{r1} & c_{r2} & \cdots & c_{rq}\end{bmatrix}\\ &= [\{\mathbf{Ab}_1c_{11} + \mathbf{Ab}_2c_{21} + \cdots + \mathbf{Ab}_rc_{r1}\}\\ &\qquad \{\mathbf{Ab}_1c_{12} + \mathbf{Ab}_2c_{22} + \cdots + \mathbf{Ab}_rc_{r2}\}\cdots\\ &\qquad\quad \{\mathbf{Ab}_1c_{1q} + \mathbf{Ab}_2c_{2q} + \cdots + \mathbf{Ab}_rc_{rq}\}]\\ &= [\{c_{11}\mathbf{Ab}_1 + c_{21}\mathbf{Ab}_2 + \cdots + c_{r1}\mathbf{Ab}_r\}\\ &\qquad \{c_{12}\mathbf{Ab}_1 + c_{22}\mathbf{Ab}_2 + \cdots + c_{r2}\mathbf{Ab}_r\}\cdots\\ &\qquad\quad \{c_{1q}\mathbf{Ab}_1 + c_{2q}\mathbf{Ab}_2 + \cdots + c_{rq}\mathbf{Ab}_r\}].\end{aligned}$$

Applying the vec operator gives

$$\begin{aligned}\mathrm{vec}(\mathbf{ABC}) &= \begin{bmatrix} c_{11}\mathbf{Ab}_1 + c_{21}\mathbf{Ab}_2 + \cdots + c_{r1}\mathbf{Ab}_r\\ c_{12}\mathbf{Ab}_1 + c_{22}\mathbf{Ab}_2 + \cdots + c_{r2}\mathbf{Ab}_r\\ \vdots\\ c_{1q}\mathbf{Ab}_1 + c_{2q}\mathbf{Ab}_2 + \cdots + c_{rq}\mathbf{Ab}_r\end{bmatrix}\\ &= \begin{bmatrix} c_{11}\mathbf{A} & c_{21}\mathbf{A} & \cdots & c_{r1}\mathbf{A}\\ c_{12}\mathbf{A} & c_{22}\mathbf{A} & \cdots & c_{r2}\mathbf{A}\\ \vdots & \vdots & \cdots & \vdots\\ c_{1q}\mathbf{A} & c_{2q}\mathbf{A} & \cdots & c_{rq}\mathbf{A}\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \vdots\\ \mathbf{b}_r\end{bmatrix}\\ &= (\mathbf{C}' \otimes \mathbf{A})\cdot\mathrm{vec}(\mathbf{B}). \quad ■\end{aligned}$$

Chapter 10 Exercises

10.1. Consider a scalar $AR(p)$ process ($n = 1$). Deduce from equation [10.2.19] that the $(p \times 1)$ vector consisting of the variance and first $(p - 1)$ autocovariances,

$$\begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{p-1} \end{bmatrix},$$

can be calculated from the first p elements in the first column of the $(p^2 \times p^2)$ matrix $\sigma^2[\mathbf{I}_{p^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1}$ for $\mathbf{F}$ the $(p \times p)$ matrix defined in equation [1.2.3] in Chapter 1.

10.2. Let $\mathbf{y}_t = (X_t, Y_t)'$ be given by

$$X_t = \varepsilon_t + \theta\varepsilon_{t-1}$$
$$Y_t = h_1 X_{t-1} + u_t,$$

where $(\varepsilon_t, u_t)'$ is vector white noise with contemporaneous variance-covariance matrix given by

$$\begin{bmatrix} E(\varepsilon_t^2) & E(\varepsilon_t u_t) \\ E(u_t\varepsilon_t) & E(u_t^2) \end{bmatrix} = \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_u^2 \end{bmatrix}.$$

(a) Calculate the autocovariance matrices $\{\Gamma_k\}_{k=-\infty}^{\infty}$ for this process.

(b) Use equation [10.4.3] to calculate the population spectrum. Find the cospectrum between X and Y and the quadrature spectrum from X to Y.

(c) Verify that your answer to part (b) could equivalently be calculated from expression [10.4.45].

(d) Verify by integrating your answer to part (b) that [10.4.49] holds; that is, show that

$$(2\pi)^{-1}\int_{-\pi}^{\pi} \frac{s_{YX}(\omega)}{s_{XX}(\omega)} e^{i\omega k}\, d\omega = \begin{cases} h_1 & \text{for } k = 1 \\ 0 & \text{for other integer } k. \end{cases}$$

Chapter 10 References

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59:817–58.

——— and J. Christopher Monahan. 1992. "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica* 60:953–66.

Fuller, Wayne A. 1976. *Introduction to Statistical Time Series*. New York: Wiley.

Gallant, A. Ronald. 1987. *Nonlinear Statistical Models*. New York: Wiley.

Hansen, Lars P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50:1029–54.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55: 703–8.

——— and ———. 1992. "Automatic Lag Selection in Covariance Matrix Estimation." University of Wisconsin, Madison, Mimeo.

Sims, Christopher A. 1980. "Macroeconomics and Reality." *Econometrica* 48:1–48.

White, Halbert. 1984. *Asymptotic Theory for Econometricians*. Orlando, Fla.: Academic Press.

11 Vector Autoregressions

The previous chapter introduced some basic tools for describing vector time series processes. This chapter looks in greater depth at vector autoregressions, which are particularly convenient for estimation and forecasting. Their popularity for analyzing the dynamics of economic systems is due to Sims's (1980) influential work. The chapter begins with a discussion of maximum likelihood estimation and hypothesis testing. Section 11.2 examines a concept of causation in bivariate systems proposed by Granger (1969). Section 11.3 generalizes the discussion of Granger causality to multivariate systems and examines estimation of restricted vector autoregressions. Sections 11.4 and 11.5 introduce impulse-response functions and variance decompositions, which are used to summarize the dynamic relations between the variables in a vector autoregression. Section 11.6 reviews how such summaries can be used to evaluate structural hypotheses. Section 11.7 develops formulas needed to calculate standard errors for impulse-response functions.

11.1. *Maximum Likelihood Estimation and Hypothesis Testing for an Unrestricted Vector Autoregression*

The Conditional Likelihood Function for a Vector Autoregression

Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector containing the values that n variables assume at date t. The dynamics of $\mathbf{y}_t$ are presumed to be governed by a pth-order Gaussian vector autoregression,

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad [11.1.1]$$

with $\boldsymbol{\varepsilon}_t \sim$ i.i.d. $N(\mathbf{0}, \mathbf{\Omega})$.

Suppose we have observed each of these n variables for $(T + p)$ time periods. As in the scalar autoregression, the simplest approach is to condition on the first p observations (denoted $\mathbf{y}_{-p+1}, \mathbf{y}_{-p+2}, \ldots, \mathbf{y}_0$) and to base estimation on the last T observations (denoted $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$). The objective then is to form the conditional likelihood

$$f_{\mathbf{Y}_T,\mathbf{Y}_{T-1},\ldots,\mathbf{Y}_1|\mathbf{Y}_0,\mathbf{Y}_{-1},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1|\mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \qquad [11.1.2]$$

and maximize with respect to $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector that contains the elements of $\mathbf{c}, \mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_p$, and $\mathbf{\Omega}$. Vector autoregressions are invariably estimated on the basis of the conditional likelihood function [11.1.2] rather than the full-sample

unconditional likelihood. For brevity, we will hereafter refer to [11.1.2] simply as the "likelihood function" and the value of $\boldsymbol{\theta}$ that maximizes [11.1.2] as the "maximum likelihood estimate."

The likelihood function is calculated in the same way as for a scalar autoregression. Conditional on the values of $\mathbf{y}$ observed through date $t-1$, the value of $\mathbf{y}$ for date t is equal to a constant,

$$\mathbf{c} + \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\Phi}_2\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p\mathbf{y}_{t-p}, \qquad [11.1.3]$$

plus a $N(\mathbf{0}, \boldsymbol{\Omega})$ variable. Thus,

$$\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1} \sim N\Big((\mathbf{c} + \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\Phi}_2\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p\mathbf{y}_{t-p}), \boldsymbol{\Omega}\Big). \qquad [11.1.4]$$

It will be convenient to use a more compact expression for the conditional mean [11.1.3]. Let $\mathbf{x}_t$ denote a vector containing a constant term and p lags of each of the elements of $\mathbf{y}$:

$$\mathbf{x}_t \equiv \begin{bmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix}. \qquad [11.1.5]$$

Thus, $\mathbf{x}_t$ is an $[(np+1) \times 1]$ vector. Let $\boldsymbol{\Pi}'$ denote the following $[n \times (np+1)]$ matrix:

$$\boldsymbol{\Pi}' \equiv [\mathbf{c} \quad \boldsymbol{\Phi}_1 \quad \boldsymbol{\Phi}_2 \quad \cdots \quad \boldsymbol{\Phi}_p]. \qquad [11.1.6]$$

Then the conditional mean [11.1.3] is equal to $\boldsymbol{\Pi}'\mathbf{x}_t$. The jth row of $\boldsymbol{\Pi}'$ contains the parameters of the jth equation in the *VAR*. Using this notation, [11.1.4] can be written more compactly as

$$\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1} \sim N(\boldsymbol{\Pi}'\mathbf{x}_t, \boldsymbol{\Omega}). \qquad [11.1.7]$$

Thus, the conditional density of the tth observation is

$$\begin{aligned} &f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\mathbf{Y}_{t-2},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &\quad = (2\pi)^{-n/2}|\boldsymbol{\Omega}^{-1}|^{1/2} \exp[(-1/2)(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)]. \end{aligned} \qquad [11.1.8]$$

The joint density of observations 1 through t conditioned on $\mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}$ satisfies

$$\begin{aligned} &f_{\mathbf{Y}_t,\mathbf{Y}_{t-1},\ldots,\mathbf{Y}_1|\mathbf{Y}_0,\mathbf{Y}_{-1},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1 \,|\, \mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &\quad = f_{\mathbf{Y}_{t-1},\ldots,\mathbf{Y}_1|\mathbf{Y}_0,\mathbf{Y}_{-1},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_{t-1}, \ldots, \mathbf{y}_1 \,|\, \mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &\qquad \times f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\mathbf{Y}_{t-2},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}). \end{aligned}$$

Applying this formula recursively, the likelihood for the full sample $\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1$ conditioned on $\mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}$ is the product of the individual conditional densities:

$$\begin{aligned} &f_{\mathbf{Y}_T,\mathbf{Y}_{T-1},\ldots,\mathbf{Y}_1|\mathbf{Y}_0,\mathbf{Y}_{-1},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1 \,|\, \mathbf{y}_0, \mathbf{y}_{-1}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &\quad = \prod_{t=1}^{T} f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\mathbf{Y}_{t-2},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}). \end{aligned} \qquad [11.1.9]$$

The sample log likelihood is found by substituting [11.1.8] into [11.1.9] and taking

logs:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{t=1}^{T} \log f_{\mathbf{Y}_t|\mathbf{Y}_{t-1},\mathbf{Y}_{t-2},\ldots,\mathbf{Y}_{-p+1}}(\mathbf{y}_t \,|\, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_{-p+1}; \boldsymbol{\theta}) \\ &= -(Tn/2)\log(2\pi) + (T/2)\log|\boldsymbol{\Omega}^{-1}| \qquad [11.1.10] \\ &\quad - (1/2)\sum_{t=1}^{T}\left[(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)\right].\end{aligned}$$

Maximum Likelihood Estimate of $\boldsymbol{\Pi}$

Consider first the *MLE* of $\boldsymbol{\Pi}$, which contains the constant term $\mathbf{c}$ and autoregressive coefficients $\boldsymbol{\Phi}_j$. This turns out to be given by

$$\underset{[n\times(np+1)]}{\hat{\boldsymbol{\Pi}}'} = \left[\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}, \qquad [11.1.11]$$

which can be viewed as the sample analog of the population linear projection of $\mathbf{y}_t$ on a constant and $\mathbf{x}_t$ (equation [4.1.23]). The jth row of $\hat{\boldsymbol{\Pi}}'$ is

$$\underset{[1\times(np+1)]}{\hat{\boldsymbol{\pi}}_j'} = \left[\sum_{t=1}^{T}y_{jt}\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}, \qquad [11.1.12]$$

which is just the estimated coefficient vector from an *OLS* regression of y_{jt} on $\mathbf{x}_t$. Thus, maximum likelihood estimates of the coefficients for the jth equation of a *VAR* are found by an *OLS* regression of y_{jt} on a constant term and p lags of all of the variables in the system.

To verify [11.1.11], write the sum appearing in the last term in [11.1.10] as

$$\begin{aligned}&\sum_{t=1}^{T}\left[(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)\right] \\ &= \sum_{t=1}^{T}\left[(\mathbf{y}_t - \hat{\boldsymbol{\Pi}}'\mathbf{x}_t + \hat{\boldsymbol{\Pi}}'\mathbf{x}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \hat{\boldsymbol{\Pi}}'\mathbf{x}_t + \hat{\boldsymbol{\Pi}}'\mathbf{x}_t - \boldsymbol{\Pi}'\mathbf{x}_t)\right] \qquad [11.1.13] \\ &= \sum_{t=1}^{T}\left[[\hat{\boldsymbol{\varepsilon}}_t + (\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t]'\boldsymbol{\Omega}^{-1}[\hat{\boldsymbol{\varepsilon}}_t + (\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t]\right],\end{aligned}$$

where the jth element of the $(n \times 1)$ vector $\hat{\boldsymbol{\varepsilon}}_t$ is the sample residual for observation t from an *OLS* regression of y_{jt} on $\mathbf{x}_t$:

$$\hat{\boldsymbol{\varepsilon}}_t \equiv \mathbf{y}_t - \hat{\boldsymbol{\Pi}}'\mathbf{x}_t. \qquad [11.1.14]$$

Expression [11.1.13] can be expanded as

$$\begin{aligned}&\sum_{t=1}^{T}\left[(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)\right] \\ &= \sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{\varepsilon}}_t + 2\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t \qquad [11.1.15] \\ &\quad + \sum_{t=1}^{T}\mathbf{x}_t'(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t.\end{aligned}$$

Consider the middle term in [11.1.15]. Since this is a scalar, it is unchanged

by applying the "trace" operator:

$$\begin{aligned}\sum_{t=1}^{T} \hat{\mathbf{e}}_t'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t &= \text{trace}\left[\sum_{t=1}^{T} \hat{\mathbf{e}}_t'\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t\right] \\ &= \text{trace}\left[\sum_{t=1}^{T} \boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t\hat{\mathbf{e}}_t'\right] \qquad [11.1.16] \\ &= \text{trace}\left[\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})' \sum_{t=1}^{T} \mathbf{x}_t\hat{\mathbf{e}}_t'\right].\end{aligned}$$

But the sample residuals from an *OLS* regression are by construction orthogonal to the explanatory variables, meaning that $\Sigma_{t=1}^{T} \mathbf{x}_t\hat{e}_{jt} = \mathbf{0}$ for all j and so $\Sigma_{t=1}^{T} \mathbf{x}_t\hat{\mathbf{e}}_t' = \mathbf{0}$. Hence, [11.1.16] is identically zero, and [11.1.15] simplifies to

$$\begin{aligned}&\sum_{t=1}^{T}\left[(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t)\right] \\ &= \sum_{t=1}^{T} \hat{\mathbf{e}}_t'\boldsymbol{\Omega}^{-1}\hat{\mathbf{e}}_t + \sum_{t=1}^{T} \mathbf{x}_t'(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t.\end{aligned} \qquad [11.1.17]$$

Since $\boldsymbol{\Omega}$ is a positive definite matrix, $\boldsymbol{\Omega}^{-1}$ is as well.[1] Thus, defining the $(n \times 1)$ vector $\mathbf{x}_t^*$ as

$$\mathbf{x}_t^* \equiv (\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t,$$

the last term in [11.1.17] takes the form

$$\sum_{t=1}^{T} \mathbf{x}_t'(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})\boldsymbol{\Omega}^{-1}(\hat{\boldsymbol{\Pi}} - \boldsymbol{\Pi})'\mathbf{x}_t = \sum_{t=1}^{T} [\mathbf{x}_t^*]'\boldsymbol{\Omega}^{-1}\mathbf{x}_t^*.$$

This is positive for any sequence $\{\mathbf{x}_t^*\}_{t=1}^{T}$ other than $\mathbf{x}_t^* = \mathbf{0}$ for all t. Thus, the smallest value that [11.1.17] can take on is achieved when $\mathbf{x}_t^* = \mathbf{0}$, or when $\boldsymbol{\Pi} = \hat{\boldsymbol{\Pi}}$. Since [11.1.17] is minimized by setting $\boldsymbol{\Pi} = \hat{\boldsymbol{\Pi}}$, it follows that [11.1.10] is maximized by setting $\boldsymbol{\Pi} = \hat{\boldsymbol{\Pi}}$, establishing the claim that *OLS* regressions provide the maximum likelihood estimates of the coefficients of a vector autoregression.

Some Useful Results on Matrix Derivatives

The next task is to calculate the maximum likelihood estimate of $\boldsymbol{\Omega}$. Here two results from matrix calculus will prove helpful. The first result concerns the derivative of a quadratic form in a matrix. Let a_{ij} denote the row i, column j element of an $(n \times n)$ matrix $\mathbf{A}$. Suppose that the matrix $\mathbf{A}$ is nonsymmetric and unrestricted (that is, the value of a_{ij} is unrelated to the value of a_{kl} when either $i \neq k$ or $j \neq l$). Consider a quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ for $\mathbf{x}$ an $(n \times 1)$ vector. The quadratic form can be written out explicitly as

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i a_{ij} x_j, \qquad [11.1.18]$$

from which

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial a_{ij}} = x_i x_j. \qquad [11.1.19]$$

[1]This follows immediately from the fact that $\boldsymbol{\Omega}^{-1}$ can be written as $\mathbf{L}'\mathbf{L}$ for $\mathbf{L}$ a nonsingular matrix as in [8.3.1].

Collecting these n^2 different derivatives into an $(n \times n)$ matrix, equation [11.1.19] can conveniently be expressed in matrix form as

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}'. \qquad [11.1.20]$$

The second result concerns the derivative of the determinant of a matrix. Let $\mathbf{A}$ be a nonsymmetric unrestricted $(n \times n)$ matrix with positive determinant. Then

$$\frac{\partial \log|\mathbf{A}|}{\partial a_{ij}} = a^{ji}, \qquad [11.1.21]$$

where a^{ji} denotes the row j, column i element of $\mathbf{A}^{-1}$. In matrix form,

$$\frac{\partial \log|\mathbf{A}|}{\partial \mathbf{A}} = (\mathbf{A}')^{-1}. \qquad [11.1.22]$$

To derive [11.1.22], recall the formula for the determinant of $\mathbf{A}$ (equation [A.4.10] in the Mathematical Review, Appendix A, at the end of the book):

$$|\mathbf{A}| = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |\mathbf{A}_{ij}|, \qquad [11.1.23]$$

where $\mathbf{A}_{ij}$ denotes the $(n-1) \times (n-1)$ matrix formed by deleting row i and column j from $\mathbf{A}$. The derivative of [11.1.23] with respect to a_{ij} is

$$\frac{\partial |\mathbf{A}|}{\partial a_{ij}} = (-1)^{i+j} |\mathbf{A}_{ij}|, \qquad [11.1.24]$$

since the parameter a_{ij} does not appear in the matrix $\mathbf{A}_{ij}$. It follows that

$$\frac{\partial \log|\mathbf{A}|}{\partial a_{ij}} = (1/|\mathbf{A}|) \cdot (-1)^{i+j} |\mathbf{A}_{ij}|,$$

which will be recognized from equation [A.4.12] as the row j, column i element of $\mathbf{A}^{-1}$, as claimed in equation [11.1.22].

The Maximum Likelihood Estimate of $\mathbf{\Omega}$

We now apply these results to find the *MLE* of $\mathbf{\Omega}$. When evaluated at the *MLE* $\hat{\mathbf{\Pi}}$, the log likelihood [11.1.10] is

$$\begin{aligned} \mathcal{L}(\mathbf{\Omega}, \hat{\mathbf{\Pi}}) = & -(Tn/2) \log(2\pi) + (T/2) \log|\mathbf{\Omega}^{-1}| \\ & - (1/2) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t' \mathbf{\Omega}^{-1} \hat{\boldsymbol{\varepsilon}}_t. \end{aligned} \qquad [11.1.25]$$

Our objective is to find a symmetric positive definite matrix $\mathbf{\Omega}$ for which this is as large as possible. It is instructive to consider first maximizing [11.1.25] by choosing $\mathbf{\Omega}$ to be any unrestricted $(n \times n)$ matrix. For that purpose we can just differentiate [11.1.25] with respect to the elements of $\mathbf{\Omega}^{-1}$ using formulas [11.1.20] and [11.1.22]:

$$\begin{aligned} \frac{\partial \mathcal{L}(\mathbf{\Omega}, \hat{\mathbf{\Pi}})}{\partial \mathbf{\Omega}^{-1}} &= (T/2) \frac{\partial \log|\mathbf{\Omega}^{-1}|}{\partial \mathbf{\Omega}^{-1}} - (1/2) \sum_{t=1}^{T} \frac{\partial \hat{\boldsymbol{\varepsilon}}_t' \mathbf{\Omega}^{-1} \hat{\boldsymbol{\varepsilon}}_t}{\partial \mathbf{\Omega}^{-1}} \\ &= (T/2)\mathbf{\Omega}' - (1/2) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'. \end{aligned} \qquad [11.1.26]$$

The likelihood is maximized when this derivative is set to zero, or when

$$\mathbf{\Omega}' = (1/T) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'. \qquad [11.1.27]$$

The matrix $\mathbf{\Omega}$ that satisfies [11.1.27] maximizes the likelihood among the class of all unrestricted $(n \times n)$ matrices. Note, however, that the optimal unrestricted value for $\mathbf{\Omega}$ that is specified by [11.1.27] turns out to be symmetric and positive definite. The *MLE*, or the value of $\mathbf{\Omega}$ that maximizes the likelihood among the class of all symmetric positive definite matrices, is thus also given by [11.1.27]:

$$\hat{\mathbf{\Omega}} = (1/T) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'. \qquad [11.1.28]$$

The row i, column i element of $\hat{\mathbf{\Omega}}$ is given by

$$\hat{\sigma}_i^2 = (1/T) \sum_{t=1}^{T} \hat{\varepsilon}_{it}^2, \qquad [11.1.29]$$

which is just the average squared residual from a regression of the ith variable in the *VAR* on a constant term and p lags of all the variables. The row i, column j element of $\hat{\mathbf{\Omega}}$ is

$$\hat{\sigma}_{ij} = (1/T) \sum_{t=1}^{T} \hat{\varepsilon}_{it} \hat{\varepsilon}_{jt}, \qquad [11.1.30]$$

which is the average product of the *OLS* residual for variable i and the *OLS* residual for variable j.

Likelihood Ratio Tests

To perform a likelihood ratio test, we need to calculate the maximum value achieved for [11.1.25]. Thus, consider

$$\begin{aligned} \mathcal{L}(\hat{\mathbf{\Omega}}, \hat{\mathbf{\Pi}}) = & -(Tn/2) \log(2\pi) + (T/2) \log|\hat{\mathbf{\Omega}}^{-1}| \\ & - (1/2) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t' \hat{\mathbf{\Omega}}^{-1} \hat{\boldsymbol{\varepsilon}}_t \end{aligned} \qquad [11.1.31]$$

for $\hat{\mathbf{\Omega}}$ given by [11.1.28]. The last term in [11.1.31] is

$$\begin{aligned} (1/2) \sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t' \hat{\mathbf{\Omega}}^{-1} \hat{\boldsymbol{\varepsilon}}_t &= (1/2) \operatorname{trace}\left[\sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t' \hat{\mathbf{\Omega}}^{-1} \hat{\boldsymbol{\varepsilon}}_t\right] \\ &= (1/2) \operatorname{trace}\left[\sum_{t=1}^{T} \hat{\mathbf{\Omega}}^{-1} \hat{\boldsymbol{\varepsilon}}_t \hat{\boldsymbol{\varepsilon}}_t'\right] \\ &= (1/2) \operatorname{trace}[\hat{\mathbf{\Omega}}^{-1}(T\hat{\mathbf{\Omega}})] \\ &= (1/2) \operatorname{trace}(T \cdot \mathbf{I}_n) \\ &= Tn/2. \end{aligned}$$

Substituting this into [11.1.31] produces

$$\mathcal{L}(\hat{\mathbf{\Omega}}, \hat{\mathbf{\Pi}}) = -(Tn/2) \log(2\pi) + (T/2) \log|\hat{\mathbf{\Omega}}^{-1}| - (Tn/2). \qquad [11.1.32]$$

This makes likelihood ratio tests particularly simple to perform. Suppose we want to test the null hypothesis that a set of variables was generated from a Gaussian *VAR* with p_0 lags against the alternative specification of $p_1 > p_0$ lags. To estimate the system under the null hypothesis, we perform a set of n *OLS* regressions of each variable in the system on a constant term and on p_0 lags of all the variables in the system. Let $\hat{\mathbf{\Omega}}_0 = (1/T)\Sigma_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t(p_0)[\hat{\boldsymbol{\varepsilon}}_t(p_0)]'$ be the variance-covariance matrix of the residuals from these regressions. The maximum value for the log likelihood

under H_0 is then

$$\mathcal{L}_0^* = -(Tn/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{\Omega}}_0^{-1}| - (Tn/2).$$

Similarly, the system is estimated under the alternative hypothesis by *OLS* regressions that include p_1 lags of all the variables. The maximized log likelihood under the alternative is

$$\mathcal{L}_1^* = -(Tn/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{\Omega}}_1^{-1}| - (Tn/2),$$

where $\hat{\mathbf{\Omega}}_1$ is the variance-covariance matrix of the residuals from this second set of regressions. Twice the log likelihood ratio is then

$$\begin{aligned} 2(\mathcal{L}_1^* - \mathcal{L}_0^*) &= 2\{(T/2)\log|\hat{\mathbf{\Omega}}_1^{-1}| - (T/2)\log|\hat{\mathbf{\Omega}}_0^{-1}|\} \\ &= T\log(1/|\hat{\mathbf{\Omega}}_1|) - T\log(1/|\hat{\mathbf{\Omega}}_0|) \\ &= -T\log|\hat{\mathbf{\Omega}}_1| + T\log|\hat{\mathbf{\Omega}}_0| \\ &= T\{\log|\hat{\mathbf{\Omega}}_0| - \log|\hat{\mathbf{\Omega}}_1|\}. \end{aligned} \qquad [11.1.33]$$

Under the null hypothesis, this asymptotically has a χ^2 distribution with degrees of freedom equal to the number of restrictions imposed under H_0. Each equation in the specification restricted by H_0 has $(p_1 - p_0)$ fewer lags on each of n variables compared with H_1; thus, H_0 imposes $n(p_1 - p_0)$ restrictions on each equation. Since there are n such equations, H_0 imposes $n^2(p_1 - p_0)$ restrictions. Thus, the magnitude calculated in [11.1.33] is asymptotically χ^2 with $n^2(p_1 - p_0)$ degrees of freedom.

For example, suppose a bivariate *VAR* is estimated with three and four lags ($n = 2$, $p_0 = 3$, $p_1 = 4$). Say that the original sample contains 50 observations on each variable (denoted $\mathbf{y}_{-3}, \mathbf{y}_{-2}, \ldots, \mathbf{y}_{46}$) and that observations 1 through 46 were used to estimate both the three- and four-lag specifications so that $T = 46$. Let $\hat{\varepsilon}_{it}(p_0)$ be the sample residual for observation t from an *OLS* regression of y_{it} on a constant, three lags of y_{1t}, and three lags of y_{2t}. Suppose that $(1/T)\Sigma_{t=1}^T[\hat{\varepsilon}_{1t}(p_0)]^2 = 2.0$, $(1/T)\Sigma_{t=1}^T[\hat{\varepsilon}_{2t}(p_0)]^2 = 2.5$, and $(1/T)\Sigma_{t=1}^T \hat{\varepsilon}_{1t}(p_0)\hat{\varepsilon}_{2t}(p_0) = 1.0$. Then

$$\hat{\mathbf{\Omega}}_0 = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 2.5 \end{bmatrix}$$

and $\log|\hat{\mathbf{\Omega}}_0| = \log 4 = 1.386$. Suppose that when a fourth lag is added to each regression, the residual covariance matrix is reduced to

$$\hat{\mathbf{\Omega}}_1 = \begin{bmatrix} 1.8 & 0.9 \\ 0.9 & 2.2 \end{bmatrix},$$

for which $\log|\hat{\mathbf{\Omega}}_1| = 1.147$. Then

$$2(\mathcal{L}_1^* - \mathcal{L}_0^*) = 46(1.386 - 1.147) = 10.99.$$

The degrees of freedom for this test are $2^2(4 - 3) = 4$. Since $10.99 > 9.49$ (the 5% critical value for a $\chi^2(4)$ variable), the null hypothesis is rejected. The dynamics are not completely captured by a three-lag *VAR*, and a four-lag specification seems preferable.

Sims (1980, p. 17) suggested a modification to the likelihood ratio test to take into account small-sample bias. He recommended replacing [11.1.33] by

$$(T - k)\{\log|\hat{\mathbf{\Omega}}_0| - \log|\hat{\mathbf{\Omega}}_1|\}, \qquad [11.1.34]$$

where $k = 1 + np_1$ is the number of parameters estimated per equation. The

adjusted test has the same asymptotic distribution as [11.1.33] but is less likely to reject the null hypothesis in small samples. For the present example, this test statistic would be

$$(46 - 9)(1.386 - 1.147) = 8.84,$$

and the earlier conclusion would be reversed (H_0 would be accepted).

Asymptotic Distribution of $\hat{\Pi}$

The maximum likelihood estimates $\hat{\Pi}$ and $\hat{\Omega}$ will give consistent estimates of the population parameters even if the true innovations are non-Gaussian. Standard errors for $\hat{\Pi}$ can be based on the usual *OLS* formulas, as the following proposition demonstrates.

Proposition 11.1: *Let*

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t$ *is independent and identically distributed with mean* $\mathbf{0}$, *variance* $\mathbf{\Omega}$, *and* $E(\varepsilon_{it}\varepsilon_{jt}\varepsilon_{lt}\varepsilon_{mt}) < \infty$ *for all* $i, j, l,$ *and* m *and where roots of*

$$|\mathbf{I}_n - \mathbf{\Phi}_1 z - \mathbf{\Phi}_2 z^2 - \cdots - \mathbf{\Phi}_p z^p| = 0 \qquad [11.1.35]$$

lie outside the unit circle. Let $k \equiv np + 1$, *and let* $\mathbf{x}_t'$ *be the* $(1 \times k)$ *vector*

$$\mathbf{x}_t' \equiv [1 \quad \mathbf{y}_{t-1}' \quad \mathbf{y}_{t-2}' \quad \cdots \quad \mathbf{y}_{t-p}'].$$

Let $\hat{\boldsymbol{\pi}}_T = \text{vec}(\hat{\Pi}_T)$ *denote the* $(nk \times 1)$ *vector of coefficients resulting from OLS regressions of each of the elements of* $\mathbf{y}_t$ *on* $\mathbf{x}_t$ *for a sample of size* T:

$$\hat{\boldsymbol{\pi}}_T = \begin{bmatrix} \hat{\boldsymbol{\pi}}_{1.T} \\ \hat{\boldsymbol{\pi}}_{2.T} \\ \vdots \\ \hat{\boldsymbol{\pi}}_{n.T} \end{bmatrix},$$

where

$$\hat{\boldsymbol{\pi}}_{i.T} = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t y_{it}\right];$$

and let $\boldsymbol{\pi}$ *denote the* $(nk \times 1)$ *vector of corresponding population coefficients. Finally, let*

$$\hat{\mathbf{\Omega}}_T = (1/T)\sum_{t=1}^{T} \hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t',$$

where

$$\hat{\boldsymbol{\varepsilon}}_t' = [\hat{\varepsilon}_{1t} \quad \hat{\varepsilon}_{2t} \quad \cdots \quad \hat{\varepsilon}_{nt}]$$
$$\hat{\varepsilon}_{it} = y_{it} - \mathbf{x}_t'\hat{\boldsymbol{\pi}}_{i.T}.$$

Then

(*a*) $(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{Q}$ *where* $\mathbf{Q} = E(\mathbf{x}_t\mathbf{x}_t')$;

(*b*) $\hat{\boldsymbol{\pi}}_T \xrightarrow{p} \boldsymbol{\pi}$;

(c) $\hat{\mathbf{\Omega}}_T \xrightarrow{p} \mathbf{\Omega}$;

(d) $\sqrt{T}(\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) \xrightarrow{L} N(\mathbf{0}, (\mathbf{\Omega} \otimes \mathbf{Q}^{-1}))$, *where* $\otimes$ *denotes the Kronecker product.*

A proof of this proposition is provided in Appendix 11.A to this chapter.

If we are interested only in $\hat{\boldsymbol{\pi}}_{i,T}$, the coefficients of the ith regression in the *VAR*, result (d) implies that

$$\sqrt{T}(\hat{\boldsymbol{\pi}}_{i,T} - \boldsymbol{\pi}_i) \xrightarrow{L} N(\mathbf{0}, \sigma_i^2\mathbf{Q}^{-1}), \qquad [11.1.36]$$

where $\sigma_i^2 = E(\varepsilon_{it}^2)$ is the variance of the innovation of the ith equation in the *VAR*. But σ_i^2 is estimated consistently by $\hat{\sigma}_i^2 = (1/T)\Sigma_{t=1}^T \hat{\varepsilon}_{it}^2$, the average squared residual from *OLS* estimation of this equation. Similarly, $\mathbf{Q}^{-1}$ is estimated consistently by $[(1/T)\Sigma_{t=1}^T \mathbf{x}_t\mathbf{x}_t']^{-1}$. Hence, [11.1.36] invites us to treat $\hat{\boldsymbol{\pi}}_i$ approximately as

$$\hat{\boldsymbol{\pi}}_i \approx N\left(\boldsymbol{\pi}_i, \hat{\sigma}_i^2 \left[\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\right). \qquad [11.1.37]$$

But this is the standard *OLS* formula for coefficient variances with $s_i^2 = [1/(T-k)]\Sigma_{t=1}^T \hat{\varepsilon}_{it}^2$ in the standard formula replaced by the maximum likelihood estimate $\hat{\sigma}_i^2$ in [11.1.37]. Clearly, s_i^2 and $\hat{\sigma}_i^2$ are asymptotically equivalent, though following Sims's argument in [11.1.34], the larger (and thus more conservative) standard errors resulting from the *OLS* formulas might be preferred. Hence, Proposition 11.1 establishes that the standard *OLS* t and F statistics applied to the coefficients of any single equation in the *VAR* are asymptotically valid and can be evaluated in the usual way.

A more general hypothesis of the form $\mathbf{R}\boldsymbol{\pi} = \mathbf{r}$ involving coefficients across different equations of the *VAR* can be tested using a generalization of the Wald form of the *OLS* χ^2 test (expression [8.2.23]). Result (d) of Proposition 11.1 establishes that

$$\sqrt{T}(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r}) \xrightarrow{L} N\left(\mathbf{0}, \mathbf{R}(\mathbf{\Omega} \otimes \mathbf{Q}^{-1})\mathbf{R}'\right).$$

In the light of results (a) and (c), the asymptotic distribution could equivalently be described as

$$\sqrt{T}(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r}) \xrightarrow{p} N\left(\mathbf{0}, \mathbf{R}(\hat{\mathbf{\Omega}}_T \otimes \mathbf{Q}_T^{-1})\mathbf{R}'\right),$$

where $\hat{\mathbf{\Omega}}_T = (1/T)\Sigma_{t=1}^T \hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'$ and $\mathbf{Q}_T = (1/T)\Sigma_{t=1}^T \mathbf{x}_t\mathbf{x}_t'$. Hence, the following statistic has an asymptotic χ^2 distribution:

$$\begin{aligned}\chi^2(m) &= T(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r})'(\mathbf{R}(\hat{\mathbf{\Omega}}_T \otimes \mathbf{Q}_T^{-1})\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r}) \\ &= (\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r})'(\mathbf{R}[\hat{\mathbf{\Omega}}_T \otimes (T\mathbf{Q}_T)^{-1}]\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r}) \qquad [11.1.38] \\ &= (\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r})'\left\{\mathbf{R}\left[\hat{\mathbf{\Omega}}_T \otimes \left(\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\right]\mathbf{R}'\right\}^{-1}(\mathbf{R}\hat{\boldsymbol{\pi}}_T - \mathbf{r}).\end{aligned}$$

The degrees of freedom for this statistic are given by the number of rows of $\mathbf{R}$, or the number of restrictions tested.

For example, suppose we wanted to test the hypothesis that the constant term in the first equation in the *VAR* (c_1) is equal to the constant term in the second equation (c_2). Then $\mathbf{R}$ is a $(1 \times nk)$ vector with unity in the first position, -1 in

the $(k + 1)$th position, and zeros elsewhere:

$$\mathbf{R} = [1 \quad 0 \quad 0 \quad \cdots \quad 0 \quad -1 \quad 0 \quad 0 \quad \cdots \quad 0].$$

To apply result [11.1.38], it is convenient to write $\mathbf{R}$ in Kronecker product form as

$$\mathbf{R} = \mathbf{R}_n \otimes \mathbf{R}_k, \qquad [11.1.39]$$

where $\mathbf{R}_n$ selects the equations that are involved and $\mathbf{R}_k$ selects the coefficients. For this example,

$$\underset{(1\times n)}{\mathbf{R}_n} = [1 \quad -1 \quad 0 \quad 0 \quad \cdots \quad 0]$$
$$\underset{(1\times k)}{\mathbf{R}_k} = [1 \quad 0 \quad 0 \quad 0 \quad \cdots \quad 0].$$

We then calculate

$$\begin{aligned}
\mathbf{R}\left[\hat{\boldsymbol{\Omega}} \otimes \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\right]\mathbf{R}' &= (\mathbf{R}_n \otimes \mathbf{R}_k)\left[\hat{\boldsymbol{\Omega}} \otimes \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\right](\mathbf{R}_n' \otimes \mathbf{R}_k') \\
&= (\mathbf{R}_n\hat{\boldsymbol{\Omega}}\mathbf{R}_n') \otimes \left[\mathbf{R}_k\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{R}_k'\right] \\
&= (\hat{\sigma}_1^2 - 2\hat{\sigma}_{12} + \hat{\sigma}_2^2) \otimes \xi^{11},
\end{aligned}$$

where $\hat{\sigma}_{12}$ is the covariance between $\hat{\varepsilon}_{1t}$ and $\hat{\varepsilon}_{2t}$ and ξ^{11} is the (1, 1) element of $(\Sigma_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t')^{-1}$. Since ξ^{11} is a scalar, the foregoing Kronecker product is a simple multiplication. The test statistic [11.1.38] is then

$$\chi^2(1) = \frac{(\hat{c}_1 - \hat{c}_2)^2}{(\hat{\sigma}_1^2 - 2\hat{\sigma}_{12} + \hat{\sigma}_2^2)\xi^{11}}.$$

Asymptotic Distribution of $\hat{\boldsymbol{\Omega}}$

In considering the asymptotic distribution of the estimates of variances and covariances, notice that since $\boldsymbol{\Omega}$ is symmetric, some of its elements are redundant. Recall that the "vec" operator transforms an $(n \times n)$ matrix into an $(n^2 \times 1)$ vector by stacking the columns. For example,

$$\text{vec}\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{31} \\ \sigma_{12} \\ \sigma_{22} \\ \sigma_{32} \\ \sigma_{13} \\ \sigma_{23} \\ \sigma_{33} \end{bmatrix}. \qquad [11.1.40]$$

An analogous "vech" operator transforms an $(n \times n)$ matrix into an $([n(n + 1)/2] \times 1)$ vector by vertically stacking those elements on or below the principal

diagonal. For example,

$$\text{vech}\begin{bmatrix}\sigma_{11} & \sigma_{12} & \sigma_{13}\\ \sigma_{21} & \sigma_{22} & \sigma_{23}\\ \sigma_{31} & \sigma_{32} & \sigma_{33}\end{bmatrix} = \begin{bmatrix}\sigma_{11}\\ \sigma_{21}\\ \sigma_{31}\\ \sigma_{22}\\ \sigma_{32}\\ \sigma_{33}\end{bmatrix}. \qquad [11.1.41]$$

Proposition 11.2: *Let*

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t \sim$ *i.i.d.* $N(\mathbf{0}, \mathbf{\Omega})$ *and where roots of*

$$|\mathbf{I}_n - \mathbf{\Phi}_1 z - \mathbf{\Phi}_2 z^2 - \cdots - \mathbf{\Phi}_p z^p| = 0$$

lie outside the unit circle. Let $\hat{\boldsymbol{\pi}}_T$, $\hat{\mathbf{\Omega}}_T$, *and* $\mathbf{Q}$ *be as defined in Proposition 11.1. Then*

$$\begin{bmatrix}\sqrt{T}[\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}]\\ \sqrt{T}[\text{vech}(\hat{\mathbf{\Omega}}_T) - \text{vech}(\mathbf{\Omega})]\end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}(\mathbf{\Omega}\otimes\mathbf{Q}^{-1}) & \mathbf{0}\\ \mathbf{0} & \mathbf{\Sigma}_{22}\end{bmatrix}\right).$$

Let σ_{ij} *denote the row* i, *column* j *element of* $\mathbf{\Omega}$; *for example,* σ_{11} *is the variance of* ε_{1t}. *Then the element of* $\mathbf{\Sigma}_{22}$ *corresponding to the covariance between* $\hat{\sigma}_{ij}$ *and* $\hat{\sigma}_{lm}$ *is given by* $(\sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl})$ *for all* $i, j, l, m = 1, 2, \ldots, n$, *including* $i = j = l = m$.

For example, for $n = 2$, Proposition 11.2 implies that

$$\sqrt{T}\begin{bmatrix}\hat{\sigma}_{11,T} - \sigma_{11}\\ \hat{\sigma}_{12,T} - \sigma_{12}\\ \hat{\sigma}_{22,T} - \sigma_{22}\end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}2\sigma_{11}^2 & 2\sigma_{11}\sigma_{12} & 2\sigma_{12}^2\\ 2\sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 & 2\sigma_{12}\sigma_{22}\\ 2\sigma_{12}^2 & 2\sigma_{12}\sigma_{22} & 2\sigma_{22}^2\end{bmatrix}\right). \qquad [11.1.42]$$

Thus, a Wald test of the null hypothesis that there is no covariance between ε_{1t} and ε_{2t} is given by

$$\frac{\sqrt{T}\hat{\sigma}_{12}}{(\hat{\sigma}_{11}\hat{\sigma}_{22} + \hat{\sigma}_{12}^2)^{1/2}} \approx N(0, 1).$$

A Wald test of the null hypothesis that ε_{1t} and ε_{2t} have the same variance is given by

$$\frac{T(\hat{\sigma}_{11} - \hat{\sigma}_{22})^2}{2\hat{\sigma}_{11}^2 - 4\hat{\sigma}_{12}^2 + 2\hat{\sigma}_{22}^2} \approx \chi^2(1),$$

where $\hat{\sigma}_{11}^2$ denotes the square of the estimated variance of the innovation for the first equation.

The matrix $\mathbf{\Sigma}_{22}$ in Proposition 11.2 can be expressed more compactly using the *duplication matrix*. Notice that since $\mathbf{\Omega}$ is symmetric, the n^2 elements of $\text{vec}(\mathbf{\Omega})$ in [11.1.40] are simple duplications of the $n(n + 1)/2$ elements of $\text{vech}(\mathbf{\Omega})$ in [11.1.41]. There exists a unique $[n^2 \times n(n + 1)/2]$ matrix $\mathbf{D}_n$ that transforms $\text{vech}(\mathbf{\Omega})$ into $\text{vec}(\mathbf{\Omega})$, that is, a unique matrix satisfying

$$\mathbf{D}_n\,\text{vech}(\mathbf{\Omega}) = \text{vec}(\mathbf{\Omega}). \qquad [11.1.43]$$

For example, for $n = 2$, equation [11.1.43] is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix}. \qquad [11.1.44]$$

Further, define $\mathbf{D}_n^+$ to be the following $[n(n+1)/2 \times n^2]$ matrix:[2]

$$\mathbf{D}_n^+ \equiv (\mathbf{D}_n'\mathbf{D}_n)^{-1}\mathbf{D}_n'. \qquad [11.1.45]$$

Notice that $\mathbf{D}_n^+\mathbf{D}_n = \mathbf{I}_{n(n+1)/2}$. Thus, premultiplying both sides of [11.1.43] by $\mathbf{D}_n^+$ reveals $\mathbf{D}_n^+$ to be a matrix that transforms $\text{vec}(\boldsymbol{\Omega})$ into $\text{vech}(\boldsymbol{\Omega})$ for symmetric $\boldsymbol{\Omega}$:

$$\text{vech}(\boldsymbol{\Omega}) = \mathbf{D}_n^+ \text{vec}(\boldsymbol{\Omega}). \qquad [11.1.46]$$

For example, for $n = 2$, equation [11.1.46] is

$$\begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} \\ \sigma_{21} \\ \sigma_{12} \\ \sigma_{22} \end{bmatrix}. \qquad [11.1.47]$$

It turns out that the matrix $\boldsymbol{\Sigma}_{22}$ described in Proposition 11.2 can be written as[3]

$$\boldsymbol{\Sigma}_{22} = 2\mathbf{D}_n^+(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\mathbf{D}_n^+)'. \qquad [11.1.48]$$

For example, for $n = 2$, expression [11.1.48] becomes

$$\begin{aligned} 2\mathbf{D}_2^+(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\mathbf{D}_2^+)' &= 2\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &\times \begin{bmatrix} \sigma_{11}\sigma_{11} & \sigma_{11}\sigma_{12} & \sigma_{12}\sigma_{11} & \sigma_{12}\sigma_{12} \\ \sigma_{11}\sigma_{21} & \sigma_{11}\sigma_{22} & \sigma_{12}\sigma_{21} & \sigma_{12}\sigma_{22} \\ \sigma_{21}\sigma_{11} & \sigma_{21}\sigma_{12} & \sigma_{22}\sigma_{11} & \sigma_{22}\sigma_{12} \\ \sigma_{21}\sigma_{21} & \sigma_{21}\sigma_{22} & \sigma_{22}\sigma_{21} & \sigma_{22}\sigma_{22} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2\sigma_{11}^2 & 2\sigma_{11}\sigma_{12} & 2\sigma_{12}^2 \\ 2\sigma_{11}\sigma_{12} & \sigma_{11}\sigma_{22} + \sigma_{12}^2 & 2\sigma_{12}\sigma_{22} \\ 2\sigma_{12}^2 & 2\sigma_{12}\sigma_{22} & 2\sigma_{22}^2 \end{bmatrix}, \end{aligned}$$

which reproduces [11.1.42].

11.2. *Bivariate Granger Causality Tests*

One of the key questions that can be addressed with vector autoregressions is how useful some variables are for forecasting others. This section discusses a particular summary of the forecasting relation between two variables proposed by Granger (1969) and popularized by Sims (1972). A more general discussion of a related question in larger vector systems is provided in the following section.

[2]It can be shown that $(\mathbf{D}_n'\mathbf{D}_n)$ is nonsingular. For more details, see Magnus and Neudecker (1988, pp. 48–49).

[3]Magnus and Neudecker (1988, p. 318) derived this expression directly from the information matrix.

Definition of Bivariate Granger Causality

The question investigated in this section is whether a scalar y can help forecast another scalar x. If it cannot, then we say that y does not Granger-cause x. More formally, *y fails to Granger-cause x* if for all $s > 0$ the mean squared error of a forecast of x_{t+s} based on $(x_t, x_{t-1}, \ldots)$ is the same as the *MSE* of a forecast of x_{t+s} that uses both $(x_t, x_{t-1}, \ldots)$ and $(y_t, y_{t-1}, \ldots)$. If we restrict ourselves to linear functions, y fails to Granger-cause x if

$$MSE[\hat{E}(x_{t+s}|x_t, x_{t-1}, \ldots)] = MSE[\hat{E}(x_{t+s}|x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)]. \qquad [11.2.1]$$

Equivalently, we say that *x is exogenous in the time series sense with respect to y* if [11.2.1] holds. Yet a third expression meaning the same thing is that *y is not linearly informative about future x.*

Granger's reason for proposing this definition was that if an event Y is the cause of another event X, then the event Y should precede the event X. Although one might agree with this position philosophically, there can be serious obstacles to practical implementation of this idea using aggregate time series data, as will be seen in the examples considered later in this section. First, however, we explore the mechanical implications of Granger causality for the time series representation of a bivariate system.

Alternative Implications of Granger Causality

In a bivariate *VAR* describing x and y, y does not Granger-cause x if the coefficient matrices $\mathbf{\Phi}_j$ are lower triangular for all j:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} \phi_{11}^{(1)} & 0 \\ \phi_{21}^{(1)} & \phi_{22}^{(1)} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \begin{bmatrix} \phi_{11}^{(2)} & 0 \\ \phi_{21}^{(2)} & \phi_{22}^{(2)} \end{bmatrix} \begin{bmatrix} x_{t-2} \\ y_{t-2} \end{bmatrix} + \cdots + \begin{bmatrix} \phi_{11}^{(p)} & 0 \\ \phi_{21}^{(p)} & \phi_{22}^{(p)} \end{bmatrix} \begin{bmatrix} x_{t-p} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \qquad [11.2.2]$$

From the first row of this system, the optimal one-period-ahead forecast of x depends only on its own lagged values and not on lagged y:

$$\hat{E}(x_{t+1}|x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots) = c_1 + \phi_{11}^{(1)}x_t + \phi_{11}^{(2)}x_{t-1} + \cdots + \phi_{11}^{(p)}x_{t-p+1}. \qquad [11.2.3]$$

Furthermore, the value of x_{t+2} from [11.2.2] is given by

$$x_{t+2} = c_1 + \phi_{11}^{(1)}x_{t+1} + \phi_{11}^{(2)}x_t + \cdots + \phi_{11}^{(p)}x_{t-p+2} + \varepsilon_{1,t+2}.$$

Recalling [11.2.3] and the law of iterated projections, it is clear that the date t forecast of this magnitude on the basis of $(x_t, x_{t-1}, \ldots, y_t, y_{t-1}, \ldots)$ also depends only on $(x_t, x_{t-1}, \ldots, x_{t-p+1})$. By induction, the same is true of an s-period-ahead forecast. Thus, for the bivariate *VAR*, y does not Granger-cause x if $\mathbf{\Phi}_j$ is lower triangular for all j, as claimed.

Recall from equation [10.1.19] that

$$\mathbf{\Psi}_s = \mathbf{\Phi}_1\mathbf{\Psi}_{s-1} + \mathbf{\Phi}_2\mathbf{\Psi}_{s-2} + \cdots + \mathbf{\Phi}_p\mathbf{\Psi}_{s-p} \qquad \text{for } s = 1, 2, \ldots,$$

with $\mathbf{\Psi}_0$ the identity matrix and $\mathbf{\Psi}_s = \mathbf{0}$ for $s < 0$. This expression implies that if

Φ_j is lower triangular for all j, then the moving average matrices Ψ_s for the fundamental representation will be lower triangular for all s. Thus, if y fails to Granger-cause x, then the $MA(\infty)$ representation can be written

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \psi_{11}(L) & 0 \\ \psi_{21}(L) & \psi_{22}(L) \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}, \qquad [11.2.4]$$

where

$$\psi_{ij}(L) = \psi_{ij}^{(0)} + \psi_{ij}^{(1)}L^1 + \psi_{ij}^{(2)}L^2 + \psi_{ij}^{(3)}L^3 + \cdots$$

with $\psi_{11}^{(0)} = \psi_{22}^{(0)} = 1$ and $\psi_{21}^{(0)} = 0$.

Another implication of Granger causality was stressed by Sims (1972).

Proposition 11.3: *Consider a linear projection of y_t on past, present, and future x's,*

$$y_t = c + \sum_{j=0}^{\infty} b_j x_{t-j} + \sum_{j=1}^{\infty} d_j x_{t+j} + \eta_t, \qquad [11.2.5]$$

where b_j and d_j are defined as population projection coefficients, that is, the values for which

$$E(\eta_t x_\tau) = 0 \qquad \textit{for all } t \textit{ and } \tau.$$

Then y fails to Granger-cause x if and only if $d_j = 0$ for $j = 1, 2, \ldots$.

Econometric Tests for Granger Causality

Econometric tests of whether a particular observed series y Granger-causes x can be based on any of the three implications [11.2.2], [11.2.4], or [11.2.5]. The simplest and probably best approach uses the autoregressive specification [11.2.2]. To implement this test, we assume a particular autoregressive lag length p and estimate

$$\begin{aligned} x_t = c_1 &+ \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \cdots + \alpha_p x_{t-p} + \beta_1 y_{t-1} \\ &+ \beta_2 y_{t-2} + \cdots + \beta_p y_{t-p} + u_t \end{aligned} \qquad [11.2.6]$$

by *OLS*. We then conduct an F test of the null hypothesis

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_p = 0. \qquad [11.2.7]$$

Recalling Proposition 8.2, one way to implement this test is to calculate the sum of squared residuals from [11.2.6],[4]

$$RSS_1 = \sum_{t=1}^{T} \hat{u}_t^2,$$

and compare this with the sum of squared residuals of a univariate autoregression for x_t,

$$RSS_0 = \sum_{t=1}^{T} \hat{e}_t^2,$$

[4]Note that in order for t to run from 1 to T as indicated, we actually need $T + p$ observations on x and y, namely, $x_{-p+1}, x_{-p+2}, \ldots, x_T$ and $y_{-p+1}, y_{-p+2}, \ldots, y_T$.

where

$$x_t = c_0 + \gamma_1 x_{t-1} + \gamma_2 x_{t-2} + \cdots + \gamma_p x_{t-p} + e_t \qquad [11.2.8]$$

is also estimated by OLS. If

$$S_1 \equiv \frac{(RSS_0 - RSS_1)/p}{RSS_1/(T - 2p - 1)} \qquad [11.2.9]$$

is greater than the 5% critical value for an $F(p, T - 2p - 1)$ distribution, then we reject the null hypothesis that y does not Granger-cause x; that is, if S_1 is sufficiently large, we conclude that y does Granger-cause x.

The test statistic [11.2.9] would have an exact F distribution for a regression with fixed regressors and Gaussian disturbances. With lagged dependent variables as in the Granger-causality regressions, however, the test is valid only asymptotically. An asymptotically equivalent test is given by

$$S_2 \equiv \frac{T(RSS_0 - RSS_1)}{RSS_1}. \qquad [11.2.10]$$

We would reject the null hypothesis that y does not Granger-cause x if S_2 is greater than the 5% critical values for a $\chi^2(p)$ variable.

An alternative approach is to base the test on the Sims form [11.2.5] instead of the Granger form [11.2.2]. A problem with the Sims form is that the error term η_t is in general autocorrelated. Thus, a standard F test of the hypothesis that $d_j = 0$ for all j in [11.2.5] will not give the correct answer. One option is to use autocorrelation-consistent standard errors for the OLS estimates as described in Section 10.5. A second option is to use a generalized least squares transformation. A third option, suggested by Geweke, Meese, and Dent (1983), is as follows. Suppose the error term η_t in [11.2.5] has Wold representation $\eta_t = \psi_{22}(L)v_{2t}$. Multiplying both sides of [11.2.5] by $h(L) \equiv [\psi_{22}(L)]^{-1}$ produces

$$y_t = c_2 - \sum_{j=1}^{\infty} h_j y_{t-j} + \sum_{j=0}^{\infty} b_j^* x_{t-j} + \sum_{j=1}^{\infty} d_j^* x_{t+j} + v_{2t}. \qquad [11.2.11]$$

The error term in [11.2.11] is white noise and uncorrelated with any of the explanatory variables. Moreover, $d_j^* = 0$ for all j if and only if $d_j = 0$ for all j. Thus, by truncating the infinite sums in [11.2.11] at some finite value, we can test the null hypothesis that y does not Granger-cause x with an F test of $d_1^* = d_2^* = \cdots = d_p^* = 0$.

A variety of other Granger-causality tests have been proposed; see Pierce and Haugh (1977) and Geweke, Meese, and Dent (1983) for selective surveys. Bouissou, Laffont, and Vuong (1986) discussed tests using discrete-valued panel data. The Monte Carlo simulations of Geweke, Meese, and Dent suggest that the simplest and most straightforward test—namely, that based on [11.2.10]—may well be the best.

The results of any empirical test for Granger causality can be surprisingly sensitive to the choice of lag length (p) or the methods used to deal with potential nonstationarity of the series. For demonstrations of the practical relevance of such issues, see Feige and Pearce (1979), Christiano and Ljungqvist (1988), and Stock and Watson (1989).

Interpreting Granger-Causality Tests

How is "Granger causality" related to the standard meaning of "causality"? We explore this question with several examples.

Example 11.1—Granger-Causality Tests and Forward-Looking Behavior

The first example uses a modification of the model of stock prices described in Chapter 2. If an investor buys one share of a stock for the price P_t at date t, then at $t + 1$ the investor will receive D_{t+1} in dividends and be able to sell the stock for P_{t+1}. The ex post rate of return from the stock (denoted r_{t+1}) is defined by

$$(1 + r_{t+1})P_t \equiv P_{t+1} + D_{t+1}. \qquad [11.2.12]$$

A simple model of stock prices holds that the expected rate of return for the stock is a constant r at all dates:[5]

$$(1 + r)P_t = E_t[P_{t+1} + D_{t+1}]. \qquad [11.2.13]$$

Here E_t denotes an expectation conditional on all information available to stock market participants at time t. The logic behind [11.2.13] is that if investors had information at time t leading them to anticipate a higher-than-normal return to stocks, they would want to buy more stocks at date t. Such purchases would drive P_t up until [11.2.13] was satisfied. This view is sometimes called the *efficient markets hypothesis*.

As noted in the discussion of equation [2.5.15] in Chapter 2, equation [11.2.13] along with a boundedness condition implies

$$P_t = E_t \sum_{j=1}^{\infty} \left[\frac{1}{1 + r}\right]^j D_{t+j}. \qquad [11.2.14]$$

Thus, according to the theory, the stock price incorporates the market's best forecast of the present value of future dividends. If this forecast is based on more information than past dividends alone, then stock prices will Granger-cause dividends as investors try to anticipate movements in dividends.

For a simple illustration of this point, suppose that

$$D_t = d + u_t + \delta u_{t-1} + v_t, \qquad [11.2.15]$$

where u_t and v_t are independent Gaussian white noise series and d is the mean dividend. Suppose that investors at time t know the values of $\{u_t, u_{t-1}, \ldots\}$ and $\{v_t, v_{t-1}, \ldots\}$. The forecast of D_{t+j} based on this information is given by

$$E_t(D_{t+j}) = \begin{cases} d + \delta u_t & \text{for } j = 1 \\ d & \text{for } j = 2, 3, \ldots. \end{cases} \qquad [11.2.16]$$

Substituting [11.2.16] into [11.2.14], the stock price would be given by

$$P_t = d/r + \delta u_t/(1 + r). \qquad [11.2.17]$$

[5]A related model was proposed by Lucas (1978):

$$U'(C_t)P_t = E_t\{\beta U'(C_{t+1})(P_{t+1} + D_{t+1})\},$$

with $U'(C_t)$ the marginal utility of consumption at date t. If we define $\tilde{P}_t$ to be the marginal-utility-weighted stock price $\tilde{P}_t \equiv U'(C_t)P_t$ and $\tilde{D}_t$ the marginal-utility-weighted dividend, then this becomes

$$\beta^{-1}\tilde{P}_t = E_t\{\tilde{P}_{t+1} + \tilde{D}_{t+1}\},$$

which is the same basic form as [11.2.13]. With risk-neutral investors, $U'(C_t)$ is a constant and the two formulations are identical. The risk-neutral version gained early support from the empirical evidence in Fama (1965).

Thus, for this example, the stock price is white noise and could not be forecast on the basis of lagged stock prices or dividends.[6] No series should Granger-cause stock prices.

On the other hand, notice from [11.2.17] that the value of u_{t-1} can be uncovered from the lagged stock price:

$$\delta u_{t-1} = (1 + r)P_{t-1} - (1 + r)d/r.$$

Recall from Section 4.7 that u_{t-1} contains additional information about D_t beyond that contained in $\{D_{t-1}, D_{t-2}, \ldots\}$. Thus, stock prices Granger-cause dividends, though dividends fail to Granger-cause stock prices. The bivariate *VAR* takes the form

$$\begin{bmatrix} P_t \\ D_t \end{bmatrix} = \begin{bmatrix} d/r \\ -d/r \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 1 + r & 0 \end{bmatrix} \begin{bmatrix} P_{t-1} \\ D_{t-1} \end{bmatrix} + \begin{bmatrix} \delta u_t/(1 + r) \\ u_t + v_t \end{bmatrix}.$$

Hence, in this model, Granger causation runs in the opposite direction from the true causation. Dividends fail to "Granger-cause" prices, even though investors' perceptions of dividends are the sole determinant of stock prices. On the other hand, prices do "Granger-cause" dividends, even though the market's evaluation of the stock in reality has no effect on the dividend process.

In general, time series that reflect forward-looking behavior, such as stock prices and interest rates, are often found to be excellent predictors of many key economic time series. This clearly does not mean that these series *cause* GNP or inflation to move up or down. Instead, the values of these series reflect the market's best information as to where GNP or inflation might be headed. Granger-causality tests for such series may be useful for assessing the efficient markets view or investigating whether markets are concerned with or are able to forecast GNP or inflation, but should not be used to infer a direction of causation.

There nevertheless are circumstances in which Granger causality may offer useful evidence about the direction of true causation. As an illustration of this theme, consider trying to measure the effects of oil price increases on the economy.

Example 11.2—Testing for Strict Econometric Exogeneity[7]

All but one of the economic recessions in the United States since World War II have been preceded by a sharp increase in the price of crude petroleum. Does this mean that oil shocks are a cause of recessions?

One possibility is that the correlation is a fluke—it happened just by chance that oil shocks and recessions appeared at similar times, even though the actual processes that generated the two series are unrelated. We can investigate this possibility by testing the null hypothesis that oil prices do not Granger-cause GNP. This hypothesis is rejected by the data—oil prices help predict the value of GNP, and their contribution to prediction is statistically significant. This argues against viewing the correlation as simply a coincidence.

To place a causal interpretation on this correlation, one must establish that oil price increases were not reflecting some other macroeconomic influence that was the true cause of the recessions. The major oil price increases have

[6]This result is due to the particular specification of the time series properties assumed for dividends. A completely general result is that the excess return series defined by $P_{t+1} + D_{t+1} - (1 + r)P_t$ (which for this example would equal $\delta u_{t+1}/(1 + r) + u_{t+1} + v_{t+1}$) should be unforecastable. The example in the text provides a simpler illustration of the general issues.

[7]This discussion is based on Hamilton (1983, 1985).

been associated with clear historical events such as the Suez crisis of 1956–57, the Arab-Israeli war of 1973–74, the Iranian revolution of 1978–79, the start of the Iran-Iraq war in 1980, and Iraq's invasion of Kuwait in 1990. One could take the view that these events were caused by forces entirely outside the U.S. economy and were essentially unpredictable. If this view is correct, then the historical correlation between oil prices and GNP could be given a causal interpretation. The view has the refutable implication that no series should Granger-cause oil prices. Empirically, one indeed finds very few macroeconomic series that help predict the timing of these oil shocks.

The theme of these two examples is that Granger-causality tests can be a useful tool for testing hypotheses that can be framed as statements about the predictability of a particular series. On the other hand, one may be skeptical about their utility as a general diagnostic for establishing the direction of causation between two arbitrary series. For this reason, it seems best to describe these as tests of whether y helps forecast x rather than tests of whether y causes x. The tests may have implications for the latter question, but only in conjunction with other assumptions.

Up to this point we have been discussing two variables, x and y, in isolation from any others. Suppose there are other variables that interact with x or y as well. How does this affect the forecasting relationship between x and y?

Example 11.3—Role of Omitted Information

Consider the following three-variable system:

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \begin{bmatrix} 1 + \delta L & L & 0 \\ 0 & 1 & 0 \\ 0 & L & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix},$$

with

$$E(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_s') = \begin{cases} \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix} & \text{for } t = s \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Thus, y_3 can offer no improvement in a forecast of either y_1 or y_2 beyond that achieved using lagged y_1 and y_2.

Let us now examine the bivariate Granger-causality relation between y_1 and y_3. First, consider the process for y_1:

$$y_{1t} = \varepsilon_{1t} + \delta\varepsilon_{1,t-1} + \varepsilon_{2,t-1}.$$

Notice that y_1 is the sum of an $MA(1)$ process $(\varepsilon_{1t} + \delta\varepsilon_{1,t-1})$ and an uncorrelated white noise process $(\varepsilon_{2,t-1})$. We know from equation [4.7.15] that the univariate representation for y_1 is an $MA(1)$ process:

$$y_{1t} = u_t + \theta u_{t-1}.$$

From [4.7.16], the univariate forecast error u_t can be expressed as

$$\begin{aligned} u_t = {} & (\varepsilon_{1t} - \theta\varepsilon_{1,t-1} + \theta^2\varepsilon_{1,t-2} - \theta^3\varepsilon_{1,t-3} + \cdots) \\ & + \delta(\varepsilon_{1,t-1} - \theta\varepsilon_{1,t-2} + \theta^2\varepsilon_{1,t-3} - \theta^3\varepsilon_{1,t-4} + \cdots) \\ & + (\varepsilon_{2,t-1} - \theta\varepsilon_{2,t-2} + \theta^2\varepsilon_{2,t-3} - \theta^3\varepsilon_{2,t-4} + \cdots). \end{aligned}$$

The univariate forecast error u_t is, of course, uncorrelated with its own lagged values. Notice, however, that it is correlated with $y_{3,t-1}$:

$$E(u_t)(y_{3,t-1}) = E(u_t)(\varepsilon_{3,t-1} + \varepsilon_{2,t-2}) = -\theta\sigma_2^2.$$

Thus, lagged y_3 could help improve a forecast of y_1 that had been based on lagged values of y_1 alone, meaning that y_3 Granger-causes y_1 in a bivariate system. The reason is that lagged y_3 is correlated with the omitted variable y_2, which is also helpful in forecasting y_1.[8]

11.3. *Maximum Likelihood Estimation of Restricted Vector Autoregressions*

Section 11.1 discussed maximum likelihood estimation and hypothesis testing on unrestricted vector autoregressions. In these systems each equation in the *VAR* had the same explanatory variables, namely, a constant term and lags of all the variables in the system. We showed how to calculate a Wald test of linear constraints but did not discuss estimation of the system subject to the constraints. This section examines estimation of a restricted *VAR*.

Granger Causality in a Multivariate Context

As an example of a restricted system that we might be interested in estimating, consider a vector generalization of the issues explored in the previous section. Suppose that the variables of a *VAR* are categorized into two groups, as represented by the $(n_1 \times 1)$ vector $\mathbf{y}_{1t}$ and the $(n_2 \times 1)$ vector $\mathbf{y}_{2t}$. The *VAR* may then be written

$$\mathbf{y}_{1t} = \mathbf{c}_1 + \mathbf{A}_1'\mathbf{x}_{1t} + \mathbf{A}_2'\mathbf{x}_{2t} + \boldsymbol{\varepsilon}_{1t} \qquad [11.3.1]$$

$$\mathbf{y}_{2t} = \mathbf{c}_2 + \mathbf{B}_1'\mathbf{x}_{1t} + \mathbf{B}_2'\mathbf{x}_{2t} + \boldsymbol{\varepsilon}_{2t}. \qquad [11.3.2]$$

Here $\mathbf{x}_{1t}$ is an $(n_1 p \times 1)$ vector containing lags of $\mathbf{y}_{1t}$, and the $(n_2 p \times 1)$ vector $\mathbf{x}_{2t}$ contains lags of $\mathbf{y}_{2t}$:

$$\mathbf{x}_{1t} \equiv \begin{bmatrix} \mathbf{y}_{1,t-1} \\ \mathbf{y}_{1,t-2} \\ \vdots \\ \mathbf{y}_{1,t-p} \end{bmatrix} \qquad \mathbf{x}_{2t} \equiv \begin{bmatrix} \mathbf{y}_{2,t-1} \\ \mathbf{y}_{2,t-2} \\ \vdots \\ \mathbf{y}_{2,t-p} \end{bmatrix}.$$

The $(n_1 \times 1)$ and $(n_2 \times 1)$ vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ contain the constant terms of the *VAR*, while the matrices $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{B}_1$, and $\mathbf{B}_2$ contain the autoregressive coefficients.

The group of variables represented by $\mathbf{y}_1$ is said to be *block-exogenous in the time series sense* with respect to the variables in $\mathbf{y}_2$ if the elements in $\mathbf{y}_2$ are of no help in improving a forecast of any variable contained in $\mathbf{y}_1$ that is based on lagged values of all the elements of $\mathbf{y}_1$ alone. In the system of [11.3.1] and [11.3.2], $\mathbf{y}_1$ is block-exogenous when $\mathbf{A}_2 = \mathbf{0}$. To discuss estimation of the system subject to this constraint, we first note an alternative form in which the unrestricted likelihood can be calculated and maximized.

[8]The reader may note that for this example the correlation between y_{1t} and $y_{3,t-1}$ is zero. However, there are nonzero correlations between (1) y_{1t} and $y_{1,t-1}$ and (2) $y_{1,t-1}$ and $y_{3,t-1}$, and these account for the contribution of $y_{3,t-1}$ to a forecast of y_{1t} that already includes $y_{1,t-1}$.

An Alternative Expression for the Likelihood Function

Section 11.1 calculated the log likelihood function for a *VAR* using the prediction-error decomposition

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f_{\mathbf{Y}_t|\mathbf{X}_t}(\mathbf{y}_t|\mathbf{x}_t;\, \boldsymbol{\theta}), \qquad [11.3.3]$$

where $\mathbf{y}_t' = (\mathbf{y}_{1t}', \mathbf{y}_{2t}')$, $\mathbf{x}_t' = (\mathbf{y}_{t-1}', \mathbf{y}_{t-2}', \ldots, \mathbf{y}_{t-p}')$, and

$$\begin{aligned}
&\log f_{\mathbf{Y}_t|\mathbf{X}_t}(\mathbf{y}_t|\mathbf{x}_t;\, \boldsymbol{\theta}) \\
&= -\frac{n_1 + n_2}{2}\log(2\pi) - \tfrac{1}{2}\log\begin{vmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{vmatrix} \\
&\quad - \tfrac{1}{2}[(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t})' \quad (\mathbf{y}_{2t} - \mathbf{c}_2 - \mathbf{B}_1'\mathbf{x}_{1t} - \mathbf{B}_2'\mathbf{x}_{2t})'] \\
&\quad \times \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t} \\ \mathbf{y}_{2t} - \mathbf{c}_2 - \mathbf{B}_1'\mathbf{x}_{1t} - \mathbf{B}_2'\mathbf{x}_{2t} \end{bmatrix}.
\end{aligned} \qquad [11.3.4]$$

Alternatively, the joint density in [11.3.4] could be written as the product of a marginal density of $\mathbf{y}_{1t}$ with the conditional density of $\mathbf{y}_{2t}$ given $\mathbf{y}_{1t}$:

$$f_{\mathbf{Y}_t|\mathbf{X}_t}(\mathbf{y}_t|\mathbf{x}_t;\, \boldsymbol{\theta}) = f_{\mathbf{Y}_{1t}|\mathbf{X}_t}(\mathbf{y}_{1t}|\mathbf{x}_t;\, \boldsymbol{\theta}) \cdot f_{\mathbf{Y}_{2t}|\mathbf{Y}_{1t},\mathbf{X}_t}(\mathbf{y}_{2t}|\mathbf{y}_{1t}, \mathbf{x}_t;\, \boldsymbol{\theta}). \qquad [11.3.5]$$

Conditional on $\mathbf{x}_t$, the density of $\mathbf{y}_{1t}$ is

$$\begin{aligned}
f_{\mathbf{Y}_{1t}|\mathbf{X}_t}(\mathbf{y}_{1t}|\mathbf{x}_t;\, \boldsymbol{\theta}) &= (2\pi)^{-n_1/2}|\boldsymbol{\Omega}_{11}|^{-1/2} \\
&\quad \times \exp[-\tfrac{1}{2}(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t})'\boldsymbol{\Omega}_{11}^{-1} \\
&\quad \times (\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t})],
\end{aligned} \qquad [11.3.6]$$

while the conditional density of $\mathbf{y}_{2t}$ given $\mathbf{y}_{1t}$ and $\mathbf{x}_t$ is also Gaussian:

$$\begin{aligned}
f_{\mathbf{Y}_{2t}|\mathbf{Y}_{1t},\mathbf{X}_t}(\mathbf{y}_{2t}|\mathbf{y}_{1t}, \mathbf{x}_t;\, \boldsymbol{\theta}) &= (2\pi)^{-n_2/2}|\mathbf{H}|^{-1/2} \\
&\quad \times \exp[-\tfrac{1}{2}(\mathbf{y}_{2t} - \mathbf{m}_{2t})'\mathbf{H}^{-1}(\mathbf{y}_{2t} - \mathbf{m}_{2t})].
\end{aligned} \qquad [11.3.7]$$

The parameters of this conditional distribution can be calculated using the results from Section 4.6. The conditional variance is given by equation [4.6.6]:

$$\mathbf{H} = \boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12};$$

while the conditional mean ($\mathbf{m}_{2t}$) can be calculated from [4.6.5]:

$$\mathbf{m}_{2t} = E(\mathbf{y}_{2t}|\mathbf{x}_t) + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}[\mathbf{y}_{1t} - E(\mathbf{y}_{1t}|\mathbf{x}_t)]. \qquad [11.3.8]$$

Notice from [11.3.1] that

$$E(\mathbf{y}_{1t}|\mathbf{x}_t) = \mathbf{c}_1 + \mathbf{A}_1'\mathbf{x}_{1t} + \mathbf{A}_2'\mathbf{x}_{2t},$$

while from [11.3.2],

$$E(\mathbf{y}_{2t}|\mathbf{x}_t) = \mathbf{c}_2 + \mathbf{B}_1'\mathbf{x}_{1t} + \mathbf{B}_2'\mathbf{x}_{2t}.$$

Substituting these expressions into [11.3.8],

$$\begin{aligned}
\mathbf{m}_{2t} &= (\mathbf{c}_2 + \mathbf{B}_1'\mathbf{x}_{1t} + \mathbf{B}_2'\mathbf{x}_{2t}) + \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}[\mathbf{y}_{1t} - (\mathbf{c}_1 + \mathbf{A}_1'\mathbf{x}_{1t} + \mathbf{A}_2'\mathbf{x}_{2t})] \\
&= \mathbf{d} + \mathbf{D}_0'\mathbf{y}_{1t} + \mathbf{D}_1'\mathbf{x}_{1t} + \mathbf{D}_2'\mathbf{x}_{2t},
\end{aligned}$$

where

$$\mathbf{d} = \mathbf{c}_2 - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{c}_1 \qquad [11.3.9]$$

$$\mathbf{D}_0' = \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} \qquad [11.3.10]$$

$$\mathbf{D}_1' = \mathbf{B}_1' - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{A}_1' \qquad [11.3.11]$$

$$\mathbf{D}_2' = \mathbf{B}_2' - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\mathbf{A}_2'. \qquad [11.3.12]$$

The log of the joint density in [11.3.4] can thus equivalently be calculated as the sum of the logs of the marginal density [11.3.6] and the conditional density [11.3.7]:

$$\log f_{Y_t|X_t}(\mathbf{y}_t|\mathbf{x}_t;\boldsymbol{\theta}) = \ell_{1t} + \ell_{2t}, \qquad [11.3.13]$$

where

$$\begin{aligned}\ell_{1t} &= (-n_1/2)\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Omega}_{11}| \\ &\quad - \tfrac{1}{2}[(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t})'\boldsymbol{\Omega}_{11}^{-1}(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t} - \mathbf{A}_2'\mathbf{x}_{2t})] \qquad [11.3.14]\end{aligned}$$

$$\begin{aligned}\ell_{2t} &= (-n_2/2)\log(2\pi) - \tfrac{1}{2}\log|\mathbf{H}| \\ &\quad - \tfrac{1}{2}[(\mathbf{y}_{2t} - \mathbf{d} - \mathbf{D}_0'\mathbf{y}_{1t} - \mathbf{D}_1'\mathbf{x}_{1t} - \mathbf{D}_2'\mathbf{x}_{2t})'\mathbf{H}^{-1} \qquad [11.3.15] \\ &\quad \times (\mathbf{y}_{2t} - \mathbf{d} - \mathbf{D}_0'\mathbf{y}_{1t} - \mathbf{D}_1'\mathbf{x}_{1t} - \mathbf{D}_2'\mathbf{x}_{2t})].\end{aligned}$$

The sample log likelihood would then be expressed as

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ell_{1t} + \sum_{t=1}^{T} \ell_{2t}. \qquad [11.3.16]$$

Equations [11.3.4] and [11.3.13] are two different expressions for the same magnitude. As long as the parameters in the second representation are related to those of the first as in [11.3.9] through [11.3.12], either calculation would produce the identical value for the likelihood. If [11.3.3] is maximized by choice of ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{c}_2$, $\mathbf{B}_1$, $\mathbf{B}_2$, $\boldsymbol{\Omega}_{11}$, $\boldsymbol{\Omega}_{12}$, $\boldsymbol{\Omega}_{22}$), the same value for the likelihood will be achieved as by maximizing [11.3.16] by choice of ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{A}_2$, $\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\boldsymbol{\Omega}_{11}$, $\mathbf{H}$).

The second maximization is as easy to achieve as the first. Since the parameters ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{A}_2$) appear in [11.3.16] only through $\Sigma_{t=1}^{T}\ell_{1t}$, the *MLE*s of these parameters can be found by *OLS* regressions of the elements of $\mathbf{y}_{1t}$ on a constant and lagged values of $\mathbf{y}_1$ and $\mathbf{y}_2$, that is, by *OLS* estimation of

$$\mathbf{y}_{1t} = \mathbf{c}_1 + \mathbf{A}_1'\mathbf{x}_{1t} + \mathbf{A}_2'\mathbf{x}_{2t} + \boldsymbol{\varepsilon}_{1t}. \qquad [11.3.17]$$

The *MLE* of $\boldsymbol{\Omega}_{11}$ is the sample variance-covariance matrix of the residuals from these regressions, $\hat{\boldsymbol{\Omega}}_{11} = (1/T)\Sigma_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_{1t}\hat{\boldsymbol{\varepsilon}}_{1t}'$. Similarly, the parameters ($\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$) appear in [11.3.16] only through $\Sigma_{t=1}^{T}\ell_{2t}$, and so their *MLE*s are obtained from *OLS* regressions of the elements of $\mathbf{y}_{2t}$ on a constant, current and lagged values of $\mathbf{y}_1$, and lagged values of $\mathbf{y}_2$:

$$\mathbf{y}_{2t} = \mathbf{d} + \mathbf{D}_0'\mathbf{y}_{1t} + \mathbf{D}_1'\mathbf{x}_{1t} + \mathbf{D}_2'\mathbf{x}_{2t} + \mathbf{v}_{2t}. \qquad [11.3.18]$$

The *MLE* of $\mathbf{H}$ is the sample variance-covariance matrix of the residuals from this second set of regressions, $\hat{\mathbf{H}} = (1/T)\Sigma_{t=1}^{T}\hat{\mathbf{v}}_{2t}\hat{\mathbf{v}}_{2t}'$.

Note that the population residuals associated with the second set of regressions, $\mathbf{v}_{2t}$, are uncorrelated with the population residuals of the first regressions. This is because $\mathbf{v}_{2t} = \mathbf{y}_{2t} - E(\mathbf{y}_{2t}|\mathbf{y}_{1t}, \mathbf{x}_t)$ is uncorrelated by construction with $\mathbf{y}_{1t}$ and $\mathbf{x}_t$, whereas $\boldsymbol{\varepsilon}_{1t}$ is a linear function of $\mathbf{y}_{1t}$ and $\mathbf{x}_t$. Similarly, the *OLS* sample residuals associated with the second regressions,

$$\hat{\mathbf{v}}_{2t} \equiv \mathbf{y}_{2t} - \hat{\mathbf{d}} - \hat{\mathbf{D}}_0'\mathbf{y}_{1t} - \hat{\mathbf{D}}_1'\mathbf{x}_{1t} - \hat{\mathbf{D}}_2'\mathbf{x}_{2t},$$

are orthogonal by construction to $\mathbf{y}_{1t}$, a constant term, and $\mathbf{x}_t$. Since the *OLS* sample residuals associated with the first regressions, $\hat{\boldsymbol{\varepsilon}}_{1t}$, are linear functions of these same elements, $\hat{\mathbf{v}}_{2t}$ is orthogonal by construction to $\hat{\boldsymbol{\varepsilon}}_{1t}$.

Maximum Likelihood Estimation of a VAR *Characterized by Block Exogeneity*

Now consider maximum likelihood estimation of the system subject to the constraint that $\mathbf{A}_2 = \mathbf{0}$. Suppose we view ($\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{H}$) rather than ($\mathbf{c}_2$, $\mathbf{B}_1$,

$\mathbf{B}_2$, $\mathbf{\Omega}_{21}$, $\mathbf{\Omega}_{22}$) as the parameters of interest for the second equation and take our objective to be to choose values for ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{\Omega}_{11}$, $\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{H}$) so as to maximize the likelihood function. For this parameterization, the value of $\mathbf{A}_2$ does not affect the value of ℓ_{2t} in [11.3.15]. Thus, the full-information maximum likelihood estimates of $\mathbf{c}_1$, $\mathbf{A}_1$, and $\mathbf{\Omega}_{11}$ can be based solely on a restricted version of the regressions in [11.3.17],

$$\mathbf{y}_{1t} = \mathbf{c}_1 + \mathbf{A}_1'\mathbf{x}_{1t} + \boldsymbol{\varepsilon}_{1t}. \qquad [11.3.19]$$

Let $\hat{\mathbf{c}}_1(0)$, $\hat{\mathbf{A}}_1(0)$, $\hat{\mathbf{\Omega}}_{11}(0)$ denote the estimates from these restricted regressions. The maximum likelihood estimates of the other parameters of the system ($\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{H}$) continue to be given by unrestricted *OLS* estimation of [11.3.18], with estimates denoted ($\hat{\mathbf{d}}$, $\hat{\mathbf{D}}_0$, $\hat{\mathbf{D}}_1$, $\hat{\mathbf{D}}_2$, $\hat{\mathbf{H}}$).

The maximum value achieved for the log likelihood function can be found by applying [11.1.32] to [11.3.13]:

$$\begin{aligned}\mathcal{L}[\hat{\boldsymbol{\theta}}(0)] &= \sum_{t=1}^{T} \ell_{1t}[\hat{\mathbf{c}}_1(0), \hat{\mathbf{A}}_1(0), \hat{\mathbf{\Omega}}_{11}(0)] + \sum_{t=1}^{T} \ell_{2t}[\hat{\mathbf{d}}, \hat{\mathbf{D}}_0, \hat{\mathbf{D}}_1, \hat{\mathbf{D}}_2, \hat{\mathbf{H}}] \\ &= [-(Tn_1/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{\Omega}}_{11}^{-1}(0)| - (Tn_1/2)] \\ &\quad + [-(Tn_2/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{H}}^{-1}| - (Tn_2/2)].\end{aligned} \qquad [11.3.20]$$

By contrast, when the system is estimated with no constraints on $\mathbf{A}_2$, the value achieved for the log likelihood is

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\theta}}) &= \sum_{t=1}^{T} \ell_{1t}[\hat{\mathbf{c}}_1, \hat{\mathbf{A}}_1, \hat{\mathbf{A}}_2, \hat{\mathbf{\Omega}}_{11}] + \sum_{t=1}^{T} \ell_{2t}[\hat{\mathbf{d}}, \hat{\mathbf{D}}_0, \hat{\mathbf{D}}_1, \hat{\mathbf{D}}_2, \hat{\mathbf{H}}] \\ &= [-(Tn_1/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{\Omega}}_{11}^{-1}| - (Tn_1/2)] \\ &\quad + [-(Tn_2/2)\log(2\pi) + (T/2)\log|\hat{\mathbf{H}}^{-1}| - (Tn_2/2)],\end{aligned} \qquad [11.3.21]$$

where ($\hat{\mathbf{c}}_1$, $\hat{\mathbf{A}}_1$, $\hat{\mathbf{A}}_2$, $\hat{\mathbf{\Omega}}_{11}$) denote estimates based on *OLS* estimation of [11.3.17]. A likelihood ratio test of the null hypothesis that $\mathbf{A}_2 = \mathbf{0}$ can thus be based on

$$\begin{aligned}2\{\mathcal{L}[\hat{\boldsymbol{\theta}}] - \mathcal{L}[\hat{\boldsymbol{\theta}}(0)]\} &= T\{\log|\hat{\mathbf{\Omega}}_{11}^{-1}| - \log|\hat{\mathbf{\Omega}}_{11}^{-1}(0)|\} \\ &= T\{\log|\hat{\mathbf{\Omega}}_{11}(0)| - \log|\hat{\mathbf{\Omega}}_{11}|\}.\end{aligned} \qquad [11.3.22]$$

This will have an asymptotic χ^2 distribution with degrees of freedom equal to the number of restrictions. Since $\mathbf{A}_2$ is an $(n_1 \times n_2 p)$ matrix, the number of restrictions is $n_1 n_2 p$.

Thus, to test the null hypothesis that the n_1 variables represented by $\mathbf{y}_1$ are block-exogenous with respect to the n_2 variables represented by $\mathbf{y}_2$, perform *OLS* regressions of each of the elements of $\mathbf{y}_1$ on a constant, p lags of all of the elements of $\mathbf{y}_1$, and p lags of all of the elements of $\mathbf{y}_2$. Let $\hat{\boldsymbol{\varepsilon}}_{1t}$ denote the $(n_1 \times 1)$ vector of sample residuals for date t from these regressions and $\hat{\mathbf{\Omega}}_{11}$ their variance-covariance matrix ($\hat{\mathbf{\Omega}}_{11} = (1/T)\Sigma_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_{1t}\hat{\boldsymbol{\varepsilon}}_{1t}'$). Next perform *OLS* regressions of each of the elements of $\mathbf{y}_1$ on a constant and p lags of all the elements of $\mathbf{y}_1$. Let $\hat{\boldsymbol{\varepsilon}}_{1t}(0)$ denote the $(n_1 \times 1)$ vector of sample residuals from this second set of regressions and $\hat{\mathbf{\Omega}}_{11}(0)$ their variance-covariance matrix ($\hat{\mathbf{\Omega}}_{11}(0) = (1/T)\Sigma_{t=1}^{T}[\hat{\boldsymbol{\varepsilon}}_{1t}(0)][\hat{\boldsymbol{\varepsilon}}_{1t}(0)]'$). If

$$T\{\log|\hat{\mathbf{\Omega}}_{11}(0)| - \log|\hat{\mathbf{\Omega}}_{11}|\}$$

is greater than the 5% critical value for a $\chi^2(n_1 n_2 p)$ variable, then the null hypothesis is rejected, and the conclusion is that some of the elements of $\mathbf{y}_2$ are helpful in forecasting $\mathbf{y}_1$.

Thus, if our interest is in estimation of the parameters ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{\Omega}_{11}$, $\mathbf{d}$, $\mathbf{D}_0$, $\mathbf{D}_1$, $\mathbf{D}_2$, $\mathbf{H}$) or testing a hypothesis about block exogeneity, all that is necessary

is *OLS* regression on the affected equations. Suppose, however, that we wanted full-information maximum likelihood estimates of the parameters of the likelihood as originally parameterized ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{\Omega}_{11}$, $\mathbf{c}_2$, $\mathbf{B}_1$, $\mathbf{B}_2$, $\mathbf{\Omega}_{21}$, $\mathbf{\Omega}_{22}$). For the parameters of the first block of equations ($\mathbf{c}_1$, $\mathbf{A}_1$, $\mathbf{\Omega}_{11}$), the *MLE*s continue to be given by *OLS* estimation of [11.3.19]. The parameters of the second block can be found from the *OLS* estimates by inverting equations [11.3.9] through [11.3.12]:[9]

$$\begin{aligned}
\hat{\mathbf{\Omega}}_{21}(0) &= \hat{\mathbf{D}}_0'[\hat{\mathbf{\Omega}}_{11}(0)] \\
\hat{\mathbf{c}}_2(0) &= \hat{\mathbf{d}} + [\hat{\mathbf{\Omega}}_{21}(0)][\hat{\mathbf{\Omega}}_{11}(0)]^{-1}[\hat{\mathbf{c}}_1(0)] \\
[\hat{\mathbf{B}}_1(0)]' &= \hat{\mathbf{D}}_1' + [\hat{\mathbf{\Omega}}_{21}(0)][\hat{\mathbf{\Omega}}_{11}(0)]^{-1}[\hat{\mathbf{A}}_1(0)]' \\
[\hat{\mathbf{B}}_2(0)]' &= \hat{\mathbf{D}}_2' \\
\hat{\mathbf{\Omega}}_{22}(0) &= \hat{\mathbf{H}} + [\hat{\mathbf{\Omega}}_{21}(0)][\hat{\mathbf{\Omega}}_{11}(0)]^{-1}[\hat{\mathbf{\Omega}}_{12}(0)].
\end{aligned}$$

Thus, the maximum likelihood estimates for the original parameterization of [11.3.2] are found from these equations by combining the *OLS* estimates from [11.3.19] and [11.3.18].

Geweke's Measure of Linear Dependence

The previous subsection modeled the relation between an ($n_1 \times 1$) vector $\mathbf{y}_{1t}$ and an ($n_2 \times 1$) vector $\mathbf{y}_{2t}$ in terms of the pth-order *VAR* [11.3.1] and [11.3.2], where the innovations have a variance-covariance matrix given by

$$E\begin{bmatrix} \boldsymbol{\varepsilon}_{1t}\boldsymbol{\varepsilon}_{1t}' & \boldsymbol{\varepsilon}_{1t}\boldsymbol{\varepsilon}_{2t}' \\ \boldsymbol{\varepsilon}_{2t}\boldsymbol{\varepsilon}_{1t}' & \boldsymbol{\varepsilon}_{2t}\boldsymbol{\varepsilon}_{2t}' \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix}.$$

To test the null hypothesis that $\mathbf{y}_1$ is block exogenous with respect to $\mathbf{y}_2$, we proposed calculating the statistic in [11.3.22],

$$T\{\log|\hat{\mathbf{\Omega}}_{11}(0)| - \log|\hat{\mathbf{\Omega}}_{11}|\} \approx \chi^2(n_1 n_2 p), \qquad [11.3.23]$$

where $\hat{\mathbf{\Omega}}_{11}$ is the variance-covariance matrix of the residuals from *OLS* estimation of [11.3.1] and $\hat{\mathbf{\Omega}}_{11}(0)$ is the variance-covariance matrix of the residuals from *OLS* estimation of [11.3.1] when lagged values of $\mathbf{y}_2$ are omitted from the regression (that is, when $\mathbf{A}_2 = \mathbf{0}$ in [11.3.1]).

Clearly, to test the parallel null hypothesis that $\mathbf{y}_2$ is block-exogenous with respect to $\mathbf{y}_1$, we would calculate

$$T\{\log|\hat{\mathbf{\Omega}}_{22}(0)| - \log|\hat{\mathbf{\Omega}}_{22}|\} \approx \chi^2(n_2 n_1 p), \qquad [11.3.24]$$

where $\hat{\mathbf{\Omega}}_{22}$ is the variance-covariance matrix of the residuals from *OLS* estimation of [11.3.2] and $\hat{\mathbf{\Omega}}_{22}(0)$ is the variance-covariance matrix of the residuals from *OLS* estimation of [11.3.2] when lagged values of $\mathbf{y}_1$ are omitted from the regression (that is, when $\mathbf{B}_1 = \mathbf{0}$ in [11.3.2]).

Finally, consider maximum likelihood estimation of the *VAR* subject to the restriction that there is no relation whatsoever between $\mathbf{y}_1$ and $\mathbf{y}_2$, that is, subject

[9]To confirm that the resulting estimate $\hat{\mathbf{\Omega}}(0)$ is symmeteric and positive definite, notice that

$$\hat{\mathbf{\Omega}}_{22}(0) = \hat{\mathbf{H}} + \hat{\mathbf{D}}_0'[\hat{\mathbf{\Omega}}_{11}(0)]\hat{\mathbf{D}}_0$$

and so

$$\begin{bmatrix} \hat{\mathbf{\Omega}}_{11}(0) & \hat{\mathbf{\Omega}}_{12}(0) \\ \hat{\mathbf{\Omega}}_{21}(0) & \hat{\mathbf{\Omega}}_{22}(0) \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \hat{\mathbf{D}}_0' & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{\Omega}}_{11}(0) & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{H}} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \hat{\mathbf{D}}_0 \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix}.$$

to the restrictions that $\mathbf{A}_2 = \mathbf{0}$, $\mathbf{B}_1 = \mathbf{0}$, and $\mathbf{\Omega}_{21} = \mathbf{0}$. For this most restricted specification, the log likelihood becomes

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \Big\{ & -(n_1/2)\log(2\pi) - (1/2)\log|\mathbf{\Omega}_{11}| \\ & - (1/2)(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t})'\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_{1t} - \mathbf{c}_1 - \mathbf{A}_1'\mathbf{x}_{1t})\Big\} \\ + \sum_{t=1}^{T} \Big\{ & -(n_2/2)\log(2\pi) - (1/2)\log|\mathbf{\Omega}_{22}| \\ & - (1/2)(\mathbf{y}_{2t} - \mathbf{c}_2 - \mathbf{B}_2'\mathbf{x}_{2t})'\mathbf{\Omega}_{22}^{-1}(\mathbf{y}_{2t} - \mathbf{c}_2 - \mathbf{B}_2'\mathbf{x}_{2t})\Big\}\end{aligned}$$

and the maximized value is

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\theta}}(0)) = & \{-(Tn_1/2)\log(2\pi) - (T/2)\log|\hat{\mathbf{\Omega}}_{11}(0)| - (Tn_1/2)\} \\ & + \{-(Tn_2/2)\log(2\pi) - (T/2)\log|\hat{\mathbf{\Omega}}_{22}(0)| - (Tn_2/2)\}.\end{aligned}$$

A likelihood ratio test of the null hypothesis of no relation at all between $\mathbf{y}_1$ and $\mathbf{y}_2$ is thus given by

$$\begin{aligned}&2\{\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\hat{\boldsymbol{\theta}}(0))\} \\ &\quad = T\left\{\log|\hat{\mathbf{\Omega}}_{11}(0)| + \log|\hat{\mathbf{\Omega}}_{22}(0)| - \log\begin{vmatrix}\hat{\mathbf{\Omega}}_{11} & \hat{\mathbf{\Omega}}_{12} \\ \hat{\mathbf{\Omega}}_{21} & \hat{\mathbf{\Omega}}_{22}\end{vmatrix}\right\},\end{aligned} \tag{11.3.25}$$

where $\hat{\mathbf{\Omega}}_{12}$ is the covariance matrix between the residuals from unrestricted *OLS* estimation of [11.3.1] and [11.3.2]. This null hypothesis imposed the (n_1n_2p) restrictions that $\mathbf{A}_2 = \mathbf{0}$, the (n_2n_1p) restrictions that $\mathbf{B}_1 = \mathbf{0}$, and the (n_2n_1) restrictions that $\mathbf{\Omega}_{21} = \mathbf{0}$. Hence, the statistic in [11.3.25] has a χ^2 distribution with $(n_1n_2) \times (2p + 1)$ degrees of freedom.

Geweke (1982) proposed $(1/T)$ times the magnitude in [11.3.25] as a measure of the degree of linear dependence between $\mathbf{y}_1$ and $\mathbf{y}_2$. Note that [11.3.25] can be expressed as the sum of three terms:

$$\begin{aligned}&T\left\{\log|\hat{\mathbf{\Omega}}_{11}(0)| + \log|\hat{\mathbf{\Omega}}_{22}(0)| - \log\begin{vmatrix}\hat{\mathbf{\Omega}}_{11} & \hat{\mathbf{\Omega}}_{12} \\ \hat{\mathbf{\Omega}}_{21} & \hat{\mathbf{\Omega}}_{22}\end{vmatrix}\right\} \\ &\quad = T\{\log|\hat{\mathbf{\Omega}}_{11}(0)| - \log|\hat{\mathbf{\Omega}}_{11}|\} + T\{\log|\hat{\mathbf{\Omega}}_{22}(0)| - \log|\hat{\mathbf{\Omega}}_{22}|\} \\ &\qquad + T\left\{\log|\hat{\mathbf{\Omega}}_{11}| + \log|\hat{\mathbf{\Omega}}_{22}| - \log\begin{vmatrix}\hat{\mathbf{\Omega}}_{11} & \hat{\mathbf{\Omega}}_{12} \\ \hat{\mathbf{\Omega}}_{21} & \hat{\mathbf{\Omega}}_{22}\end{vmatrix}\right\}.\end{aligned} \tag{11.3.26}$$

The first of these three terms, $T\{\log|\hat{\mathbf{\Omega}}_{11}(0)| - \log|\hat{\mathbf{\Omega}}_{11}|\}$, is a measure of the strength of the linear feedback from $\mathbf{y}_2$ to $\mathbf{y}_1$ and is the $\chi^2(n_1n_2p)$ statistic calculated in [11.3.23]. The second term, $T\{\log|\hat{\mathbf{\Omega}}_{22}(0)| - \log|\hat{\mathbf{\Omega}}_{22}|\}$, is an analogous measure of the strength of linear feedback from $\mathbf{y}_1$ to $\mathbf{y}_2$ and is the $\chi^2(n_2n_1p)$ statistic in [11.3.24]. The third term,

$$T\left\{\log|\hat{\mathbf{\Omega}}_{11}| + \log|\hat{\mathbf{\Omega}}_{22}| - \log\begin{vmatrix}\hat{\mathbf{\Omega}}_{11} & \hat{\mathbf{\Omega}}_{12} \\ \hat{\mathbf{\Omega}}_{21} & \hat{\mathbf{\Omega}}_{22}\end{vmatrix}\right\},$$

is a measure of instantaneous feedback. This corresponds to a likelihood ratio test of the null hypothesis that $\mathbf{\Omega}_{21} = \mathbf{0}$ with $\mathbf{A}_2$ and $\mathbf{B}_1$ unrestricted and has a $\chi^2(n_1n_2)$ distribution under the null.

Thus, [11.3.26] can be used to summarize the strength of any linear relation between $\mathbf{y}_1$ and $\mathbf{y}_2$ and identify the source of that relation. Geweke showed how these measures can be further decomposed by frequency.

Maximum Likelihood Estimation Under General Coefficient Constraints

We now discuss maximum likelihood estimation of a vector autoregression in which there are constraints that cannot be expressed in a block-recursive form as in the previous example. A *VAR* subject to general exclusion restrictions can be viewed as a system of "seemingly unrelated regressions" as originally analyzed by Zellner (1962).

Let $\mathbf{x}_{1t}$ be a $(k_1 \times 1)$ vector containing a constant term and lags of the variables that appear in the first equation of the *VAR*:

$$y_{1t} = \mathbf{x}_{1t}'\boldsymbol{\beta}_1 + \varepsilon_{1t}.$$

Similarly, let $\mathbf{x}_{2t}$ denote a $(k_2 \times 1)$ vector containing the explanatory variables for the second equation and $\mathbf{x}_{nt}$ a $(k_n \times 1)$ vector containing the variables for the last equation. Hence, the *VAR* consists of the system of equations

$$\begin{aligned} y_{1t} &= \mathbf{x}_{1t}'\boldsymbol{\beta}_1 + \varepsilon_{1t} \\ y_{2t} &= \mathbf{x}_{2t}'\boldsymbol{\beta}_2 + \varepsilon_{2t} \\ &\vdots \\ y_{nt} &= \mathbf{x}_{nt}'\boldsymbol{\beta}_n + \varepsilon_{nt}. \end{aligned} \qquad [11.3.27]$$

Let $k = k_1 + k_2 + \cdots + k_n$ denote the total number of coefficients to be estimated, and collect these in a $(k \times 1)$ vector:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_n \end{bmatrix}.$$

Then the system of equations in [11.3.27] can be written in vector form as

$$\mathbf{y}_t = \mathcal{X}_t'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_t, \qquad [11.3.28]$$

where $\mathcal{X}_t'$ is the following $(n \times k)$ matrix:

$$\mathcal{X}_t' \equiv \begin{bmatrix} x_{1t}' \\ x_{2t}' \\ \vdots \\ x_{nt}' \end{bmatrix} \equiv \begin{bmatrix} \mathbf{x}_{1t}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_{2t}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{x}_{nt}' \end{bmatrix}.$$

Thus, x_{it}' is defined as a $(1 \times k)$ vector containing the k_i explanatory variables for equation i, with zeros added so as to be conformable with the $(k \times 1)$ vector $\boldsymbol{\beta}$.

The goal is to choose $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$ so as to maximize the log likelihood function

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Omega}) = &- (Tn/2)\log(2\pi) + (T/2)\log|\boldsymbol{\Omega}^{-1}| \\ &- (1/2)\sum_{t=1}^{T} (\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta}). \end{aligned} \qquad [11.3.29]$$

This calls for choosing $\boldsymbol{\beta}$ so as to minimize

$$\sum_{t=1}^{T} (\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta}). \qquad [11.3.30]$$

If Ω^{-1} is written as $\mathbf{L}'\mathbf{L}$, this becomes

$$\sum_{t=1}^{T} (\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta})'\Omega^{-1}(\mathbf{y}_t - \mathcal{X}_t'\boldsymbol{\beta}) = \sum_{t=1}^{T} (\mathbf{L}\mathbf{y}_t - \mathbf{L}\mathcal{X}_t'\boldsymbol{\beta})'(\mathbf{L}\mathbf{y}_t - \mathbf{L}\mathcal{X}_t'\boldsymbol{\beta}) \qquad [11.3.31]$$
$$= \sum_{t=1}^{T} (\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta})'(\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta}),$$

where $\tilde{\mathbf{y}}_t \equiv \mathbf{L}\mathbf{y}_t$ and

$$\tilde{\mathcal{X}}_t' \equiv \mathbf{L}\mathcal{X}_t' \equiv \begin{bmatrix} \tilde{\mathbf{x}}_{1t}' \\ \tilde{\mathbf{x}}_{2t}' \\ \vdots \\ \tilde{\mathbf{x}}_{nt}' \end{bmatrix}.$$

But [11.3.31] is simply

$$\sum_{t=1}^{T} (\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta})'(\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta})$$
$$= \sum_{t=1}^{T} \begin{bmatrix} \tilde{y}_{1t} - \tilde{\mathbf{x}}_{1t}'\boldsymbol{\beta} \\ \tilde{y}_{2t} - \tilde{\mathbf{x}}_{2t}'\boldsymbol{\beta} \\ \vdots \\ \tilde{y}_{nt} - \tilde{\mathbf{x}}_{nt}'\boldsymbol{\beta} \end{bmatrix}' \begin{bmatrix} \tilde{y}_{1t} - \tilde{\mathbf{x}}_{1t}'\boldsymbol{\beta} \\ \tilde{y}_{2t} - \tilde{\mathbf{x}}_{2t}'\boldsymbol{\beta} \\ \vdots \\ \tilde{y}_{nt} - \tilde{\mathbf{x}}_{nt}'\boldsymbol{\beta} \end{bmatrix}$$
$$= \sum_{t=1}^{T} [(\tilde{y}_{1t} - \tilde{\mathbf{x}}_{1t}'\boldsymbol{\beta})^2 + (\tilde{y}_{2t} - \tilde{\mathbf{x}}_{2t}'\boldsymbol{\beta})^2 + \cdots + (\tilde{y}_{nt} - \tilde{\mathbf{x}}_{nt}'\boldsymbol{\beta})^2],$$

which is minimized by an *OLS* regression of $\tilde{y}_{it}$ on $\tilde{\mathbf{x}}_{it}$, pooling all the equations ($i = 1, 2, \ldots, n$) into one big regression. Thus, the maximum likelihood estimate is given by

$$\hat{\boldsymbol{\beta}} = \left\{ \sum_{t=1}^{T} [(\tilde{\mathbf{x}}_{1t}\tilde{\mathbf{x}}_{1t}') + (\tilde{\mathbf{x}}_{2t}\tilde{\mathbf{x}}_{2t}') + \cdots + (\tilde{\mathbf{x}}_{nt}\tilde{\mathbf{x}}_{nt}')] \right\}^{-1}$$
$$\times \left\{ \sum_{t=1}^{T} [(\tilde{\mathbf{x}}_{1t}\tilde{y}_{1t}) + (\tilde{\mathbf{x}}_{2t}\tilde{y}_{2t}) + \cdots + (\tilde{\mathbf{x}}_{nt}\tilde{y}_{nt})] \right\}. \qquad [11.3.32]$$

Noting that the variance of the residual of this pooled regression is unity by construction,[10] the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ can be calculated from

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \cong \left\{ \sum_{t=1}^{T} [(\tilde{\mathbf{x}}_{1t}\tilde{\mathbf{x}}_{1t}') + (\tilde{\mathbf{x}}_{2t}\tilde{\mathbf{x}}_{2t}') + \cdots + (\tilde{\mathbf{x}}_{nt}\tilde{\mathbf{x}}_{nt}')] \right\}^{-1}.$$

Construction of the variables $\tilde{y}_{it}$ and $\tilde{\mathbf{x}}_{it}$ to use in this pooled *OLS* regression requires knowledge of $\mathbf{L}$ and hence Ω. The parameters in $\boldsymbol{\beta}$ and Ω can be estimated jointly by maximum likelihood through the following iterative procedure. From n *OLS* regressions of y_{it} on $\mathbf{x}_{it}$, form an initial estimate of the coefficient vector

[10]That is,

$$E(\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta})(\tilde{\mathbf{y}}_t - \tilde{\mathcal{X}}_t'\boldsymbol{\beta})' = \mathbf{L}\Omega\mathbf{L}' = \mathbf{L}(\mathbf{L}'\mathbf{L})^{-1}\mathbf{L}' = \mathbf{I}_n.$$

$\hat{\boldsymbol{\beta}}(0) = (\mathbf{b}_1' \quad \mathbf{b}_2' \quad \cdots \quad \mathbf{b}_n')'$. Use this to form an initial estimate of the variance matrix,

$$\hat{\boldsymbol{\Omega}}(0) = (1/T) \sum_{t=1}^{T} [\mathbf{y}_t - \mathcal{X}_t'\hat{\boldsymbol{\beta}}(0)][\mathbf{y}_t - \mathcal{X}_t'\hat{\boldsymbol{\beta}}(0)]'.$$

Find a matrix $\hat{\mathbf{L}}(0)$ such that $[\hat{\mathbf{L}}(0)']\hat{\mathbf{L}}(0) = [\hat{\boldsymbol{\Omega}}(0)]^{-1}$, say, by Cholesky factorization, and form $\tilde{\mathbf{y}}_t(0) = \hat{\mathbf{L}}(0)\mathbf{y}_t$ and $\tilde{\mathcal{X}}_t'(0) = \hat{\mathbf{L}}(0)\mathcal{X}_t'$. A pooled *OLS* regression of $\tilde{\mathbf{y}}_{it}(0)$ on $\tilde{\mathbf{x}}_{it}(0)$ combining $i = 1, 2, \ldots, n$ then yields the new estimate $\hat{\boldsymbol{\beta}}(1)$, from which $\hat{\boldsymbol{\Omega}}(1) = (1/T)\Sigma_{t=1}^{T}[\mathbf{y}_t - \mathcal{X}_t'\hat{\boldsymbol{\beta}}(1)][\mathbf{y}_t - \mathcal{X}_t'\hat{\boldsymbol{\beta}}(1)]'$. Iterating in this manner will produce the maximum likelihood estimates $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$, though the estimate after just one iteration has the same asymptotic distribution as the final *MLE* (see Magnus, 1978).

An alternative expression for the *MLE* in [11.3.32] is sometimes used. Notice that

$$\begin{aligned}
&[(\tilde{\mathbf{x}}_{1t}\tilde{\mathbf{x}}_{1t}') + (\tilde{\mathbf{x}}_{2t}\tilde{\mathbf{x}}_{2t}') + \cdots + (\tilde{\mathbf{x}}_{nt}\tilde{\mathbf{x}}_{nt}')] \\
&= [\tilde{\mathbf{x}}_{1t} \quad \tilde{\mathbf{x}}_{2t} \quad \cdots \quad \tilde{\mathbf{x}}_{nt}] \begin{bmatrix} \tilde{\mathbf{x}}_{1t}' \\ \tilde{\mathbf{x}}_{2t}' \\ \vdots \\ \tilde{\mathbf{x}}_{nt}' \end{bmatrix} \\
&= \tilde{\mathcal{X}}_t\tilde{\mathcal{X}}_t' \\
&= \mathcal{X}_t\mathbf{L}'\mathbf{L}\mathcal{X}_t' \\
&= \begin{bmatrix} \mathbf{x}_{1t} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2t} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{nt} \end{bmatrix} \begin{bmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{1t}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_{2t}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{x}_{nt}' \end{bmatrix} \\
&= \begin{bmatrix} \sigma^{11}\mathbf{x}_{1t}\mathbf{x}_{1t}' & \sigma^{12}\mathbf{x}_{1t}\mathbf{x}_{2t}' & \cdots & \sigma^{1n}\mathbf{x}_{1t}\mathbf{x}_{nt}' \\ \sigma^{21}\mathbf{x}_{2t}\mathbf{x}_{1t}' & \sigma^{22}\mathbf{x}_{2t}\mathbf{x}_{2t}' & \cdots & \sigma^{2n}\mathbf{x}_{2t}\mathbf{x}_{nt}' \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^{n1}\mathbf{x}_{nt}\mathbf{x}_{1t}' & \sigma^{n2}\mathbf{x}_{nt}\mathbf{x}_{2t}' & \cdots & \sigma^{nn}\mathbf{x}_{nt}\mathbf{x}_{nt}' \end{bmatrix}.
\end{aligned} \tag{11.3.33}$$

where σ^{ij} denotes the row i, column j element of $\boldsymbol{\Omega}^{-1}$. Similarly,

$$\begin{aligned}
&[(\tilde{\mathbf{x}}_{1t}\tilde{y}_{1t}) + (\tilde{\mathbf{x}}_{2t}\tilde{y}_{2t}) + \cdots + (\tilde{\mathbf{x}}_{nt}\tilde{y}_{nt})] \\
&= [\tilde{\mathbf{x}}_{1t} \quad \tilde{\mathbf{x}}_{2t} \quad \cdots \quad \tilde{\mathbf{x}}_{nt}] \begin{bmatrix} \tilde{y}_{1t} \\ \tilde{y}_{2t} \\ \vdots \\ \tilde{y}_{nt} \end{bmatrix} \\
&= \mathcal{X}_t\mathbf{L}'\mathbf{L}\mathbf{y}_t \\
&= \begin{bmatrix} \mathbf{x}_{1t} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2t} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{nt} \end{bmatrix} \begin{bmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1n} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^{n1} & \sigma^{n2} & \cdots & \sigma^{nn} \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} \\
&= \begin{bmatrix} \sigma^{11}\mathbf{x}_{1t}y_{1t} + \sigma^{12}\mathbf{x}_{1t}y_{2t} + \cdots + \sigma^{1n}\mathbf{x}_{1t}y_{nt} \\ \sigma^{21}\mathbf{x}_{2t}y_{1t} + \sigma^{22}\mathbf{x}_{2t}y_{2t} + \cdots + \sigma^{2n}\mathbf{x}_{2t}y_{nt} \\ \vdots \\ \sigma^{n1}\mathbf{x}_{nt}y_{1t} + \sigma^{n2}\mathbf{x}_{nt}y_{2t} + \cdots + \sigma^{nn}\mathbf{x}_{nt}y_{nt} \end{bmatrix}.
\end{aligned} \tag{11.3.34}$$

Substituting [11.3.33] and [11.3.34] into [11.3.32], the *MLE* satisfies

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \sigma^{11}\Sigma\mathbf{x}_{1t}\mathbf{x}'_{1t} & \sigma^{12}\Sigma\mathbf{x}_{1t}\mathbf{x}'_{2t} & \cdots & \sigma^{1n}\Sigma\mathbf{x}_{1t}\mathbf{x}'_{nt} \\ \sigma^{21}\Sigma\mathbf{x}_{2t}\mathbf{x}'_{1t} & \sigma^{22}\Sigma\mathbf{x}_{2t}\mathbf{x}'_{2t} & \cdots & \sigma^{2n}\Sigma\mathbf{x}_{2t}\mathbf{x}'_{nt} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^{n1}\Sigma\mathbf{x}_{nt}\mathbf{x}'_{1t} & \sigma^{n2}\Sigma\mathbf{x}_{nt}\mathbf{x}'_{2t} & \cdots & \sigma^{nn}\Sigma\mathbf{x}_{nt}\mathbf{x}'_{nt} \end{bmatrix}^{-1} \times \begin{bmatrix} \Sigma(\sigma^{11}\mathbf{x}_{1t}y_{1t} + \sigma^{12}\mathbf{x}_{1t}y_{2t} + \cdots + \sigma^{1n}\mathbf{x}_{1t}y_{nt}) \\ \Sigma(\sigma^{21}\mathbf{x}_{2t}y_{1t} + \sigma^{22}\mathbf{x}_{2t}y_{2t} + \cdots + \sigma^{2n}\mathbf{x}_{2t}y_{nt}) \\ \vdots \\ \Sigma(\sigma^{n1}\mathbf{x}_{nt}y_{1t} + \sigma^{n2}\mathbf{x}_{nt}y_{2t} + \cdots + \sigma^{nn}\mathbf{x}_{nt}y_{nt}) \end{bmatrix}, \qquad [11.3.35]$$

where Σ denotes summation over $t = 1, 2, \ldots, T$.

The result from Section 11.1 was that when there are no restrictions on the *VAR*, maximum likelihood estimation is achieved by *OLS* equation by equation. This result can be seen as a special case of [11.3.35] by setting $\mathbf{x}_{1t} = \mathbf{x}_{2t} = \cdots = \mathbf{x}_{nt}$, for then [11.3.35] becomes

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= [\boldsymbol{\Omega}^{-1} \otimes (\Sigma\mathbf{x}_t\mathbf{x}'_t)]^{-1}\Sigma[(\boldsymbol{\Omega}^{-1}\mathbf{y}_t) \otimes \mathbf{x}_t] \\ &= [\boldsymbol{\Omega} \otimes (\Sigma\mathbf{x}_t\mathbf{x}'_t)^{-1}]\Sigma[(\boldsymbol{\Omega}^{-1}\mathbf{y}_t) \otimes \mathbf{x}_t] \\ &= [\mathbf{I}_n \otimes (\Sigma\mathbf{x}_t\mathbf{x}'_t)^{-1}]\Sigma[\mathbf{y}_t \otimes \mathbf{x}_t] \\ &= \begin{bmatrix} (\Sigma\mathbf{x}_t\mathbf{x}'_t)^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & (\Sigma\mathbf{x}_t\mathbf{x}'_t)^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & (\Sigma\mathbf{x}_t\mathbf{x}'_t)^{-1} \end{bmatrix} \begin{bmatrix} \Sigma y_{1t}\mathbf{x}_t \\ \Sigma y_{2t}\mathbf{x}_t \\ \vdots \\ \Sigma y_{nt}\mathbf{x}_t \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix}, \end{aligned}$$

as shown directly in Section 11.1.

Maximum likelihood estimation with constraints on both the coefficients and the variance-covariance matrix was discussed by Magnus (1978).

11.4. *The Impulse-Response Function*

In equation [10.1.15] a *VAR* was written in vector *MA*(∞) form as

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots. \qquad [11.4.1]$$

Thus, the matrix $\boldsymbol{\Psi}_s$ has the interpretation

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}'_t} = \boldsymbol{\Psi}_s; \qquad [11.4.2]$$

that is, the row i, column j element of $\boldsymbol{\Psi}_s$ identifies the consequences of a one-unit increase in the jth variable's innovation at date t (ε_{jt}) for the value of the ith variable at time $t + s$ ($y_{i,t+s}$), holding all other innovations at all dates constant.

If we were told that the first element of $\boldsymbol{\varepsilon}_t$ changed by δ_1 at the same time that the second element changed by $\delta_2, \ldots,$ and the nth element by δ_n, then the

combined effect of these changes on the value of the vector $\mathbf{y}_{t+s}$ would be given by

$$\Delta\mathbf{y}_{t+s} = \frac{\partial\mathbf{y}_{t+s}}{\partial\varepsilon_{1t}}\delta_1 + \frac{\partial\mathbf{y}_{t+s}}{\partial\varepsilon_{2t}}\delta_2 + \cdots + \frac{\partial\mathbf{y}_{t+s}}{\partial\varepsilon_{nt}}\delta_n = \mathbf{\Psi}_s\boldsymbol{\delta}, \qquad [11.4.3]$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)'$.

Several analytic characterizations of $\mathbf{\Psi}_s$ were given in Section 10.1. A simple way to find these dynamic multipliers numerically is by simulation. To implement the simulation, set $\mathbf{y}_{t-1} = \mathbf{y}_{t-2} = \cdots = \mathbf{y}_{t-p} = \mathbf{0}$. Set $\varepsilon_{jt} = 1$ and all other elements of $\boldsymbol{\varepsilon}_t$ to zero, and simulate the system [11.1.1] for dates $t, t+1, t+2, \ldots$, with $\mathbf{c}$ and $\boldsymbol{\varepsilon}_{t+1}, \boldsymbol{\varepsilon}_{t+2}, \ldots$ all zero. The value of the vector $\mathbf{y}_{t+s}$ at date $t+s$ of this simulation corresponds to the jth column of the matrix $\mathbf{\Psi}_s$. By doing a separate simulation for impulses to each of the innovations ($j = 1, 2, \ldots, n$), all of the columns of $\mathbf{\Psi}_s$ can be calculated.

A plot of the row i, column j element of $\mathbf{\Psi}_s$,

$$\frac{\partial y_{i,t+s}}{\partial\varepsilon_{jt}}, \qquad [11.4.4]$$

as a function of s is called the *impulse-response function*. It describes the response of $y_{i,t+s}$ to a one-time impulse in y_{jt} with all other variables dated t or earlier held constant.

Is there a sense in which this multiplier can be viewed as measuring the causal effect of y_j on y_i? The discussion of Granger-causality tests suggests that we should be wary of such a claim. We are on surer ground with an atheoretical *VAR* if we confine ourselves to statements about forecasts. Consider, therefore, the following question. Let

$$\mathbf{x}'_{t-1} = (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_{t-p})$$

denote the information received about the system as of date $t-1$. Suppose we are then told that the date t value of the first variable in the autoregression, y_{1t}, was higher than expected, so that ε_{1t} is positive. How does this cause us to revise our forecast of $y_{i,t+s}$? In other words, what is

$$\frac{\partial\hat{E}(y_{i,t+s}\,|\,y_{1t}, \mathbf{x}_{t-1})}{\partial y_{1t}}\,? \qquad [11.4.5]$$

The answer to this question is given by [11.4.4] with $j = 1$ only in the special case when $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}'_t) = \mathbf{\Omega}$ is a diagonal matrix. In the more general case when the elements of $\boldsymbol{\varepsilon}_t$ are contemporaneously correlated with one another, the fact that ε_{1t} is positive gives us some useful new information about the values of $\varepsilon_{2t}, \ldots, \varepsilon_{nt}$. This information has further implications for the value of $y_{i,t+s}$. To summarize these implications, we need to calculate the vector

$$\frac{\partial\hat{E}(\boldsymbol{\varepsilon}_t\,|\,y_{1t}, \mathbf{x}_{t-1})}{\partial y_{1t}}$$

and then use [11.4.3] to calculate the effect of this change in all the elements of $\boldsymbol{\varepsilon}_t$ on the value of $y_{i,t+s}$.

Yet another magnitude we might propose to measure is the forecast revision resulting from new information about, say, the second variable, y_{2t}, beyond that contained in the first variable, y_{1t}. Thus, we might calculate

$$\frac{\partial\hat{E}(y_{i,t+s}\,|\,y_{2t}, y_{1t}, \mathbf{x}_{t-1})}{\partial y_{2t}}. \qquad [11.4.6]$$

Similarly, for the variable designated number 3, we might seek

$$\frac{\partial \hat{E}(y_{i,t+s} \mid y_{3t}, y_{2t}, y_{1t}, \mathbf{x}_{t-1})}{\partial y_{3t}}, \qquad [11.4.7]$$

and for variable n,

$$\frac{\partial \hat{E}(y_{i,t+s} \mid y_{nt}, y_{n-1,t}, \ldots, y_{1t}, \mathbf{x}_{t-1})}{\partial y_{nt}}. \qquad [11.4.8]$$

This last magnitude corresponds to the effect of ε_{nt} with $\varepsilon_{1t}, \ldots, \varepsilon_{n-1,t}$ constant and is given simply by the row i, column n element of $\mathbf{\Psi}_s$.

The recursive information ordering in [11.4.5] through [11.4.8] is quite commonly used. For this ordering, the indicated multipliers can be calculated from the moving average coefficients ($\mathbf{\Psi}_s$) and the variance-covariance matrix of $\boldsymbol{\varepsilon}_t$ ($\mathbf{\Omega}$) by a simple algorithm. Recall from Section 4.4 that for any real symmetric positive definite matrix $\mathbf{\Omega}$, there exists a unique lower triangular matrix $\mathbf{A}$ with 1s along the principal diagonal and a unique diagonal matrix $\mathbf{D}$ with positive entries along the principal diagonal such that

$$\mathbf{\Omega} = \mathbf{ADA}'. \qquad [11.4.9]$$

Using this matrix $\mathbf{A}$ we can construct an $(n \times 1)$ vector $\mathbf{u}_t$ from

$$\mathbf{u}_t \equiv \mathbf{A}^{-1}\boldsymbol{\varepsilon}_t. \qquad [11.4.10]$$

Notice that since $\boldsymbol{\varepsilon}_t$ is uncorrelated with its own lags or with lagged values of $\mathbf{y}$, it follows that $\mathbf{u}_t$ is also uncorrelated with its own lags or with lagged values of $\mathbf{y}$. The elements of $\mathbf{u}_t$ are furthermore uncorrelated with each other:

$$\begin{aligned} E(\mathbf{u}_t\mathbf{u}_t') &= [\mathbf{A}^{-1}]E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')[\mathbf{A}^{-1}]' \\ &= [\mathbf{A}^{-1}]\mathbf{\Omega}[\mathbf{A}']^{-1} \\ &= [\mathbf{A}^{-1}]\mathbf{ADA}'[\mathbf{A}']^{-1} \\ &= \mathbf{D}. \end{aligned} \qquad [11.4.11]$$

But $\mathbf{D}$ is a diagonal matrix, verifying that the elements of $\mathbf{u}_t$ are mutually uncorrelated. The (j, j) element of $\mathbf{D}$ gives the variance of u_{jt}.

If both sides of [11.4.10] are premultiplied by $\mathbf{A}$, the result is

$$\mathbf{Au}_t = \boldsymbol{\varepsilon}_t. \qquad [11.4.12]$$

Writing out the equations represented by [11.4.12] explicitly,

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_{1t} \\ u_{2t} \\ u_{3t} \\ \vdots \\ u_{nt} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \\ \vdots \\ \varepsilon_{nt} \end{bmatrix}. \qquad [11.4.13]$$

Thus, u_{1t} is simply ε_{1t}. The jth row of [11.4.13] states that

$$u_{jt} = \varepsilon_{jt} - a_{j1}u_{1t} - a_{j2}u_{2t} - \cdots - a_{j,j-1}u_{j-1,t}.$$

But since u_{jt} is uncorrelated with $u_{1t}, u_{2t}, \ldots, u_{j-1,t}$, it follows that u_{jt} has the interpretation as the residual from a projection of ε_{jt} on $u_{1t}, u_{2t}, \ldots, u_{j-1,t}$:

$$\hat{E}(\varepsilon_{jt} \mid u_{1t}, u_{2t}, \ldots, u_{j-1,t}) = a_{j1}u_{1t} + a_{j2}u_{2t} + \cdots + a_{j,j-1}u_{j-1,t}. \qquad [11.4.14]$$

The fact that the u_{jt} are uncorrelated further implies that the coefficient on u_{1t} in a projection of ε_{jt} on $(u_{1t}, u_{2t}, \ldots, u_{j-1,t})$ is the same as the coefficient on

u_{1t} in a projection of ε_{jt} on u_{1t} alone:

$$\hat{E}(\varepsilon_{jt}|u_{1t}) = a_{j1}u_{1t}. \qquad [11.4.15]$$

Recalling from [11.4.13] that $\varepsilon_{1t} = u_{1t}$, we see that new information about the value of ε_{1t} would cause us to revise our forecast of ε_{jt} by the amount

$$\frac{\partial \hat{E}(\varepsilon_{jt}|\varepsilon_{1t})}{\partial \varepsilon_{1t}} = \frac{\partial \hat{E}(\varepsilon_{jt}|u_{1t})}{\partial u_{1t}} = a_{j1}. \qquad [11.4.16]$$

Now ε_{1t} has the interpretation as $y_{1t} - \hat{E}(y_{1t}|\mathbf{x}_{t-1})$ and ε_{jt} has the interpretation as $y_{jt} - \hat{E}(y_{jt}|\mathbf{x}_{t-1})$. From the formula for updating a linear projection [4.5.14], the coefficient on y_{1t} in a linear projection of y_{jt} on y_{1t} and $\mathbf{x}_{t-1}$ is the same as the coefficient on ε_{1t} in a linear projection of ε_{jt} on ε_{1t}.[11] Hence,

$$\frac{\partial \hat{E}(\varepsilon_{jt}|y_{1t}, \mathbf{x}_{t-1})}{\partial y_{1t}} = a_{j1}. \qquad [11.4.17]$$

Combining these equations for $j = 1, 2, \ldots, n$ into a vector,

$$\frac{\partial \hat{E}(\boldsymbol{\varepsilon}_t|y_{1t}, \mathbf{x}_{t-1})}{\partial y_{1t}} = \mathbf{a}_1, \qquad [11.4.18]$$

where $\mathbf{a}_1$ denotes the first column of $\mathbf{A}$:

$$\mathbf{a}_1 \equiv \begin{bmatrix} 1 \\ a_{21} \\ a_{31} \\ \vdots \\ a_{n1} \end{bmatrix}.$$

Substituting [11.4.18] into [11.4.3], the consequences for $\mathbf{y}_{t+s}$ of new information about y_{1t} beyond that contained in $\mathbf{x}_{t-1}$ are given by

$$\frac{\partial \hat{E}(\mathbf{y}_{t+s}|y_{1t}, \mathbf{x}_{t-1})}{\partial y_{1t}} = \boldsymbol{\Psi}_s\mathbf{a}_1.$$

Similarly, the variable u_{2t} represents the new information in y_{2t} beyond that contained in $(y_{1t}, \mathbf{x}_{t-1})$. This information would, of course, not cause us to change our assessment of ε_{1t} (which we know with certainty from y_{1t} and $\mathbf{x}_{t-1}$), but from [11.4.14] would cause us to revise our estimate of ε_{jt} for $j = 2, 3, \ldots, n$ by

$$\frac{\partial \hat{E}(\varepsilon_{jt}|u_{2t}, u_{1t})}{\partial u_{2t}} = a_{j2}.$$

Substituting this into [11.4.3], we conclude that

$$\frac{\partial \hat{E}(\mathbf{y}_{t+s}|y_{2t}, y_{1t}, \mathbf{x}_{t-1})}{\partial y_{2t}} = \boldsymbol{\Psi}_s\mathbf{a}_2,$$

[11]That is

$$\begin{aligned}\hat{E}(y_{jt}|y_{1t}, \mathbf{x}_{t-1}) &= \hat{E}(y_{jt}|\mathbf{x}_{t-1}) \\ &\quad + \mathrm{Cov}\{[y_{jt} - \hat{E}(y_{jt}|\mathbf{x}_{t-1})], [y_{1t} - \hat{E}(y_{1t}|\mathbf{x}_{t-1})]\} \\ &\quad \times \{\mathrm{Var}[y_{1t} - \hat{E}(y_{1t}|\mathbf{x}_{t-1})]\}^{-1}[y_{1t} - \hat{E}(y_{1t}|\mathbf{x}_{t-1})] \\ &= \hat{E}(y_{jt}|\mathbf{x}_{t-1}) + \mathrm{Cov}(\varepsilon_{jt}, \varepsilon_{1t}) \cdot \{\mathrm{Var}(\varepsilon_{1t})\}^{-1} \cdot \varepsilon_{1t}.\end{aligned}$$

where

$$\mathbf{a}_2 = \begin{bmatrix} 0 \\ 1 \\ a_{32} \\ a_{42} \\ \vdots \\ a_{n2} \end{bmatrix}.$$

In general,

$$\frac{\partial \hat{E}(\mathbf{y}_{t+s} \mid y_{jt}, y_{j-1,t}, \ldots, y_{1t}, \mathbf{x}_{t-1})}{\partial y_{jt}} = \mathbf{\Psi}_s \mathbf{a}_j, \qquad [11.4.19]$$

where $\mathbf{a}_j$ denotes the jth column of the matrix $\mathbf{A}$ defined in [11.4.9].

The magnitude in [11.4.19] is a population moment, constructed from the population parameters $\mathbf{\Psi}_s$ and $\mathbf{\Omega}$ using [11.4.9]. For a given observed sample of size T, we would estimate the autoregressive coefficients $\hat{\mathbf{\Phi}}_1, \ldots, \hat{\mathbf{\Phi}}_p$ by *OLS* and construct $\hat{\mathbf{\Psi}}_s$ by simulating the estimated system. *OLS* estimation would also provide the estimate $\hat{\mathbf{\Omega}} = (1/T)\Sigma_{t=1}^{T} \hat{\mathbf{\varepsilon}}_t \hat{\mathbf{\varepsilon}}_t'$, where the ith element of $\hat{\mathbf{\varepsilon}}_t$ is the *OLS* sample residual for the ith equation in the *VAR* for date t. Matrices $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$ satisfying $\hat{\mathbf{\Omega}} = \hat{\mathbf{A}}\hat{\mathbf{D}}\hat{\mathbf{A}}'$ could then be constructed from $\hat{\mathbf{\Omega}}$ using the algorithm described in Section 4.4. Notice that the elements of the vector $\hat{\mathbf{u}}_t = \hat{\mathbf{A}}^{-1}\hat{\mathbf{\varepsilon}}_t$ are then mutually orthogonal by construction:

$$(1/T) \sum_{t=1}^{T} \hat{\mathbf{u}}_t \hat{\mathbf{u}}_t' = (1/T) \sum_{t=1}^{T} \hat{\mathbf{A}}^{-1}\hat{\mathbf{\varepsilon}}_t \hat{\mathbf{\varepsilon}}_t' (\hat{\mathbf{A}}^{-1})' = \hat{\mathbf{A}}^{-1}\hat{\mathbf{\Omega}}(\hat{\mathbf{A}}^{-1})' = \hat{\mathbf{D}}.$$

The sample estimate of [11.4.19] is then

$$\hat{\mathbf{\Psi}}_s \hat{\mathbf{a}}_j, \qquad [11.4.20]$$

where $\hat{\mathbf{a}}_j$ denotes the jth column of the matrix $\hat{\mathbf{A}}$.

A plot of [11.4.20] as a function of s is known as an *orthogonalized impulse-response function*. It is based on decomposing the original *VAR* innovations $(\varepsilon_{1t}, \ldots, \varepsilon_{nt})$ into a set of uncorrelated components $(u_{1t}, \ldots, u_{nt})$ and calculating the consequences for $\mathbf{y}_{t+s}$ of a unit impulse in u_{jt}. These multipliers describe how new information about y_{jt} causes us to revise our forecast of $\mathbf{y}_{t+s}$, though the implicit definition of "new" information is different for each variable j.

What is the rationale for treating each variable differently? Clearly, if the *VAR* is being used as a purely atheoretical summary of the dynamics of a group of variables, there can be none—we could just as easily have labeled the second variable y_{1t} and the first variable y_{2t}, in which case we would have obtained different dynamic multipliers. By choosing a particular recursive ordering of the variables, the researcher is implicitly asking a set of questions about forecasting of the form of [11.4.5] through [11.4.8]. Whether we should orthogonalize in this way and how the variables should be ordered would seem to depend on why we want to ask such questions about forecasting in the first place. We will explore this issue in more depth in Section 11.6.

Before leaving the recursive orthogonalization, we note another popular form in which it is implemented and reported. Recall that $\mathbf{D}$ is a diagonal matrix whose (j, j) element is the variance of u_{jt}. Let $\mathbf{D}^{1/2}$ denote the diagonal matrix whose (j, j) element is the standard deviation of u_{jt}. Note that [11.4.9] could be written as

$$\mathbf{\Omega} = \mathbf{A}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{A}' = \mathbf{P}\mathbf{P}', \qquad [11.4.21]$$

where

$$\mathbf{P} \equiv \mathbf{A}\mathbf{D}^{1/2}.$$

Expression [11.4.21] is the *Cholesky decomposition* of the matrix $\mathbf{\Omega}$. Note that, like $\mathbf{A}$, the $(n \times n)$ matrix $\mathbf{P}$ is lower triangular, though whereas $\mathbf{A}$ has 1s along its principal diagonal, $\mathbf{P}$ has the standard deviation of $\mathbf{u}_t$ along its principal diagonal.

In place of $\mathbf{u}_t$ defined in [11.4.10], some researchers use

$$\mathbf{v}_t \equiv \mathbf{P}^{-1}\boldsymbol{\varepsilon}_t = \mathbf{D}^{-1/2}\mathbf{A}^{-1}\boldsymbol{\varepsilon}_t = \mathbf{D}^{-1/2}\mathbf{u}_t.$$

Thus, v_{jt} is just u_{jt} divided by its standard deviation $\sqrt{d_{jj}}$. A one-unit increase in v_{jt} is the same as a one-standard-deviation increase in u_{jt}.

In place of the dynamic multiplier $\partial y_{i,t+s}/\partial u_{jt}$, these researchers then report $\partial y_{i,t+s}/\partial v_{jt}$. The relation between these multipliers is clearly

$$\frac{\partial \mathbf{y}_{t+s}}{\partial v_{jt}} = \frac{\partial \mathbf{y}_{t+s}}{\partial u_{jt}}\sqrt{d_{jj}} = \mathbf{\Psi}_s\mathbf{a}_j\sqrt{d_{jj}}.$$

But $\mathbf{a}_j\sqrt{d_{jj}}$ is just the jth column of $\mathbf{AD}^{1/2}$, which is the jth column of the Cholesky factor matrix $\mathbf{P}$. Denoting the jth column of $\mathbf{P}$ by $\mathbf{p}_j$, we have

$$\frac{\partial \mathbf{y}_{t+s}}{\partial v_{jt}} = \mathbf{\Psi}_s\mathbf{p}_j. \qquad [11.4.22]$$

Expression [11.4.22] is just [11.4.19] multiplied by the constant $\sqrt{\text{Var}(u_{jt})}$. Expression [11.4.19] gives the consequences of a one-unit increase in y_{jt}, where the units are those in which y_{jt} itself is measured. Expression [11.4.22] gives the consequences if y_{jt} were to increase by $\sqrt{\text{Var}(u_{jt})}$ units.

11.5. *Variance Decomposition*

Equations [10.1.14] and [10.1.16] identify the error in forecasting a *VAR* s periods into the future as

$$\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t} = \boldsymbol{\varepsilon}_{t+s} + \mathbf{\Psi}_1\boldsymbol{\varepsilon}_{t+s-1} + \mathbf{\Psi}_2\boldsymbol{\varepsilon}_{t+s-2} + \cdots + \mathbf{\Psi}_{s-1}\boldsymbol{\varepsilon}_{t+1}. \qquad [11.5.1]$$

The mean squared error of this s-period-ahead forecast is thus

$$\begin{aligned} MSE(\hat{\mathbf{y}}_{t+s|t}) &= E[(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})'] \\ &= \mathbf{\Omega} + \mathbf{\Psi}_1\mathbf{\Omega}\mathbf{\Psi}_1' + \mathbf{\Psi}_2\mathbf{\Omega}\mathbf{\Psi}_2' + \cdots + \mathbf{\Psi}_{s-1}\mathbf{\Omega}\mathbf{\Psi}_{s-1}', \end{aligned} \qquad [11.5.2]$$

where

$$\mathbf{\Omega} = E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'). \qquad [11.5.3]$$

Let us now consider how each of the orthogonalized disturbances $(u_{1t}, \ldots, u_{nt})$ contributes to this *MSE*. Write [11.4.12] as

$$\boldsymbol{\varepsilon}_t = \mathbf{A}\mathbf{u}_t = \mathbf{a}_1u_{1t} + \mathbf{a}_2u_{2t} + \cdots + \mathbf{a}_nu_{nt}, \qquad [11.5.4]$$

where, as before, $\mathbf{a}_j$ denotes the jth column of the matrix $\mathbf{A}$ given in [11.4.9]. Recalling that the u_{jt}'s are uncorrelated, postmultiplying equation [11.5.4] by its transpose and taking expectations produces

$$\begin{aligned} \mathbf{\Omega} &= E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') \\ &= \mathbf{a}_1\mathbf{a}_1'\cdot\text{Var}(u_{1t}) + \mathbf{a}_2\mathbf{a}_2'\cdot\text{Var}(u_{2t}) + \cdots + \mathbf{a}_n\mathbf{a}_n'\cdot\text{Var}(u_{nt}), \end{aligned} \qquad [11.5.5]$$

where $\text{Var}(u_{jt})$ is the row j, column j element of the matrix $\mathbf{D}$ in [11.4.9]. Substituting [11.5.5] into [11.5.2], the *MSE* of the s-period-ahead forecast can be written as the sum of n terms, one arising from each of the disturbances u_{jt}:

$$\begin{aligned} MSE(\hat{\mathbf{y}}_{t+s|t}) = \sum_{j=1}^{n} \{\text{Var}(u_{jt})\cdot[&\mathbf{a}_j\mathbf{a}_j' + \mathbf{\Psi}_1\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_1' \\ &+ \mathbf{\Psi}_2\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_2' + \cdots + \mathbf{\Psi}_{s-1}\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_{s-1}']\}. \end{aligned} \qquad [11.5.6]$$

With this expression, we can calculate the contribution of the jth orthogonalized innovation to the MSE of the s-period-ahead forecast:

$$\mathrm{Var}(u_{jt})\cdot[\mathbf{a}_j\mathbf{a}_j' + \mathbf{\Psi}_1\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_1' + \mathbf{\Psi}_2\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_2' + \cdots + \mathbf{\Psi}_{s-1}\mathbf{a}_j\mathbf{a}_j'\mathbf{\Psi}_{s-1}'].$$

Again, this magnitude in general depends on the ordering of the variables.

As $s \to \infty$ for a covariance-stationary VAR, $MSE(\hat{\mathbf{y}}_{t+s|t}) \to \mathbf{\Gamma}_0$, the unconditional variance of the vector $\mathbf{y}_t$. Thus, [11.5.6] permits calculation of the portion of the total variance of y_i that is due to the disturbance u_j by letting s become suitably large.

Alternatively, recalling that $\mathbf{a}_j\cdot\sqrt{\mathrm{Var}(u_{jt})}$ is equal to $\mathbf{p}_j$, the jth column of the Cholesky factor $\mathbf{P}$, result [11.5.6] can equivalently be written as

$$MSE(\hat{\mathbf{y}}_{t+s|t}) = \sum_{j=1}^{n}[\mathbf{p}_j\mathbf{p}_j' + \mathbf{\Psi}_1\mathbf{p}_j\mathbf{p}_j'\mathbf{\Psi}_1' + \mathbf{\Psi}_2\mathbf{p}_j\mathbf{p}_j'\mathbf{\Psi}_2' + \cdots + \mathbf{\Psi}_{s-1}\mathbf{p}_j\mathbf{p}_j'\mathbf{\Psi}_{s-1}'], \qquad [11.5.7]$$

11.6. *Vector Autoregressions and Structural Econometric Models*

Pitfalls in Estimating Dynamic Structural Models

The vector autoregression was introduced in Section 10.1 as a statistical description of the dynamic interrelations between n different variables contained in the vector $\mathbf{y}_t$. This description made no use of prior theoretical ideas about how these variables are expected to be related, and therefore cannot be used to test our theories or interpret the data in terms of economic principles. This section explores the relation between VARs and structural econometric models.

Suppose that we would like to estimate a money demand function that expresses the public's willingness to hold cash as a function of the level of income and interest rates. The following specification was used by some early researchers:

$$M_t - P_t = \beta_0 + \beta_1 Y_t + \beta_2 I_t + \beta_3(M_{t-1} - P_{t-1}) + v_t^D. \qquad [11.6.1]$$

Here, M_t is the log of the nominal money balances held by the public at date t, P_t is the log of the aggregate price level, Y_t is the log of real GNP, and I_t is a nominal interest rate. The parameters β_1 and β_2 represent the effect of income and interest rates on desired cash holdings. Part of the adjustment in money balances to a change in income is thought to take place immediately, with further adjustments coming in subsequent periods. The parameter β_3 characterizes this partial adjustment. The disturbance v_t^D represents factors other than income and interest rates that influence money demand.

It was once common practice to estimate such a money demand equation with Cochrane-Orcutt adjustment for first-order serial correlation. The implicit assumption behind this procedure is that

$$v_t^D = \rho v_{t-1}^D + u_t^D, \qquad [11.6.2]$$

where u_t^D is white noise. Write equation [11.6.2] as $(1 - \rho L)v_t^D = u_t^D$ and multiply both sides of [11.6.1] by $(1 - \rho L)$:

$$M_t - P_t = (1-\rho)\beta_0 + \beta_1 Y_t - \beta_1\rho Y_{t-1} + \beta_2 I_t - \beta_2\rho I_{t-1} + (\beta_3 + \rho)(M_{t-1} - P_{t-1}) - \beta_3\rho(M_{t-2} - P_{t-2}) + u_t^D. \qquad [11.6.3]$$

Equation [11.6.3] is a restricted version of

$$M_t - P_t = \alpha_0 + \alpha_1 Y_t + \alpha_2 Y_{t-1} + \alpha_3 I_t + \alpha_4 I_{t-1} + \alpha_5(M_{t-1} - P_{t-1}) + \alpha_6(M_{t-2} - P_{t-2}) + u_t^D, \quad [11.6.4]$$

where the seven parameters ($\alpha_0, \alpha_1, \ldots, \alpha_6$) are restricted in [11.6.3] to be nonlinear functions of the underlying five parameters ($\rho, \beta_0, \beta_1, \beta_2, \beta_3$). The assumption of [11.6.2] can thus be tested by comparing the fit of [11.6.3] with that from unconstrained estimation of [11.6.4].

By definition, v_t^D represents factors influencing money demand for which the researcher has no explicit theory. It therefore seems odd to place great confidence in a detailed specification of its dynamics such as [11.6.2] without testing this assumption against the data. For example, there do not seem to be clear theoretical grounds for ruling out a specification such as

$$v_t^D = \rho_1 v_{t-1}^D + \rho_2 v_{t-2}^D + u_t^D,$$

or, for that matter, a specification in which v_t^D is correlated with lagged values of Y or I.

Equation [11.6.1] further assumes that the dynamic multiplier relating money demand to income is proportional to that relating money demand to the interest rate:

$$\frac{\partial(M_{t+s} - P_{t+s})}{\partial Y_t} = \beta_1 \beta_3^s$$

$$\frac{\partial(M_{t+s} - P_{t+s})}{\partial I_t} = \beta_2 \beta_3^s.$$

Again, it seems a good idea to test this assumption before imposing it, by comparing the fit of [11.6.1] with that of a more general dynamic model. Finally, inflation may have effects on money demand that are not captured by nominal interest rates. The specification in [11.6.1] incorporates very strong assumptions about the way nominal money demand responds to the price level.

To summarize, a specification such as [11.6.1] and [11.6.2] implicitly imposes many restrictions on dynamics for which there is little or no justification on the basis of economic theory. Before relying on the inferences of [11.6.1] and [11.6.2], it seems a good idea to test that model against a more general specification such as

$$\begin{aligned} M_t = k_1 &+ \beta_{12}^{(0)} P_t + \beta_{13}^{(0)} Y_t + \beta_{14}^{(0)} I_t \\ &+ \beta_{11}^{(1)} M_{t-1} + \beta_{12}^{(1)} P_{t-1} + \beta_{13}^{(1)} Y_{t-1} + \beta_{14}^{(1)} I_{t-1} \\ &+ \beta_{11}^{(2)} M_{t-2} + \beta_{12}^{(2)} P_{t-2} + \beta_{13}^{(2)} Y_{t-2} + \beta_{14}^{(2)} I_{t-2} + \cdots \\ &+ \beta_{11}^{(p)} M_{t-p} + \beta_{12}^{(p)} P_{t-p} + \beta_{13}^{(p)} Y_{t-p} + \beta_{14}^{(p)} I_{t-p} + u_t^D. \end{aligned} \quad [11.6.5]$$

Like equation [11.6.1], the specification in [11.6.5] is regarded as a structural money demand equation; $\beta_{13}^{(0)}$ and $\beta_{14}^{(0)}$ are interpreted as the effects of current income and the interest rate on desired money holdings, and u_t^D represents factors influencing money demand other than inflation, income, and interest rates. Compared with [11.6.1], the specification in [11.6.5] generalizes the dynamic behavior for the error term v_t^D, the partial adjustment process, and the influence of the price level on desired money holdings.

Although [11.6.5] relaxes many of the dubious restrictions on the dynamics implied by [11.6.1], it is still not possible to estimate [11.6.5] by *OLS*, because of simultaneous equations bias. *OLS* estimation of [11.6.5] will summarize the correlation between money, the price level, income, and the interest rate. The public's money demand adjustments are one reason these variables will be correlated, but not the only one. For example, each period, the central bank may be adjusting the interest rate I_t to a level consistent with its policy objectives, which may depend on current and lagged values of income, the interest rate, the price level, and the money supply:

$$\begin{aligned} I_t = {} & k_4 + \beta_{41}^{(0)} M_t + \beta_{42}^{(0)} P_t + \beta_{43}^{(0)} Y_t \\ & + \beta_{41}^{(1)} M_{t-1} + \beta_{42}^{(1)} P_{t-1} + \beta_{43}^{(1)} Y_{t-1} + \beta_{44}^{(1)} I_{t-1} \\ & + \beta_{41}^{(2)} M_{t-2} + \beta_{42}^{(2)} P_{t-2} + \beta_{43}^{(2)} Y_{t-2} + \beta_{44}^{(2)} I_{t-2} + \cdots \\ & + \beta_{41}^{(p)} M_{t-p} + \beta_{42}^{(p)} P_{t-p} + \beta_{43}^{(p)} Y_{t-p} + \beta_{44}^{(p)} I_{t-p} + u_t^C. \end{aligned} \qquad [11.6.6]$$

Here, for example, $\beta_{42}^{(0)}$ captures the effect of the current price level on the interest rate that the central bank tries to achieve. The disturbance u_t^C captures changes in policy that cannot be described as a deterministic function of current and lagged money, the price level, income, and the interest rate. If the money demand disturbance u_t^D is unusually large, this will make M_t unusually large. If $\beta_{41}^{(0)} > 0$, this would cause I_t to be unusually large as well, in which case u_t^D would be positively correlated with the explanatory variable I_t in equation [11.6.5]. Thus, [11.6.5] cannot be estimated by *OLS*.

Nor is central bank policy and endogeneity of I_t the only reason to be concerned about simultaneous equations bias. Money demand disturbances and changes in central bank policy also have effects on aggregate output and the price level, so that Y_t and P_t in [11.6.5] are endogenous as well. An aggregate demand equation, for example, might be postulated that relates the level of output to the money supply, price level, and interest rate:

$$\begin{aligned} Y_t = {} & k_3 + \beta_{31}^{(0)} M_t + \beta_{32}^{(0)} P_t + \beta_{34}^{(0)} I_t \\ & + \beta_{31}^{(1)} M_{t-1} + \beta_{32}^{(1)} P_{t-1} + \beta_{33}^{(1)} Y_{t-1} + \beta_{34}^{(1)} I_{t-1} \\ & + \beta_{31}^{(2)} M_{t-2} + \beta_{32}^{(2)} P_{t-2} + \beta_{33}^{(2)} Y_{t-2} + \beta_{34}^{(2)} I_{t-2} + \cdots \\ & + \beta_{31}^{(p)} M_{t-p} + \beta_{32}^{(p)} P_{t-p} + \beta_{33}^{(p)} Y_{t-p} + \beta_{34}^{(p)} I_{t-p} + u_t^A, \end{aligned} \qquad [11.6.7]$$

with u_t^A representing other factors influencing aggregate demand. Similarly, an aggregate supply curve might relate the aggregate price level to the other variables being studied. The logical conclusion of such reasoning is that *all* of the date t explanatory variables in [11.6.5] should be treated as endogenous.

Relation Between Dynamic Structural Models and Vector Autoregressions

The system of equations [11.6.5] through [11.6.7] (along with an analogous aggregate supply equation describing P_t) can be collected and written in vector form as

$$\mathbf{B}_0 \mathbf{y}_t = \mathbf{k} + \mathbf{B}_1 \mathbf{y}_{t-1} + \mathbf{B}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{B}_p \mathbf{y}_{t-p} + \mathbf{u}_t, \qquad [11.6.8]$$

where

$$\mathbf{y}_t = (M_t, P_t, Y_t, I_t)'$$
$$\mathbf{u}_t = (u_t^D, u_t^S, u_t^A, u_t^C)'$$
$$\mathbf{B}_0 = \begin{bmatrix} 1 & -\beta_{12}^{(0)} & -\beta_{13}^{(0)} & -\beta_{14}^{(0)} \\ -\beta_{21}^{(0)} & 1 & -\beta_{23}^{(0)} & -\beta_{24}^{(0)} \\ -\beta_{31}^{(0)} & -\beta_{32}^{(0)} & 1 & -\beta_{34}^{(0)} \\ -\beta_{41}^{(0)} & -\beta_{42}^{(0)} & -\beta_{43}^{(0)} & 1 \end{bmatrix}$$
$$\mathbf{k} = (k_1, k_2, k_3, k_4)'$$

and $\mathbf{B}_s$ is a (4×4) matrix whose row i, column j element is given by $\beta_{ij}^{(s)}$ for $s = 1, 2, \ldots, p$. A large class of structural models for an $(n \times 1)$ vector $\mathbf{y}_t$ can be written in the form of [11.6.8].

Generalizing the argument in [11.6.3], it is assumed that a sufficient number of lags of p are included and the matrices $\mathbf{B}_s$ are defined so that $\mathbf{u}_t$ is vector white noise. If instead, say, $\mathbf{u}_t$ followed an rth-order *VAR*, with

$$\mathbf{u}_t = \mathbf{F}_1\mathbf{u}_{t-1} + \mathbf{F}_2\mathbf{u}_{t-2} + \cdots + \mathbf{F}_r\mathbf{u}_{t-r} + \mathbf{e}_t,$$

then we could premultiply [11.6.8] by $(\mathbf{I}_n - \mathbf{F}_1L^1 - \mathbf{F}_2L^2 - \cdots - \mathbf{F}_rL^r)$ to arrive at a system of the same basic form as [11.6.8] with p replaced by $(p + r)$ and with $\mathbf{u}_t$ replaced by the white noise disturbance $\mathbf{e}_t$.

If each side of [11.6.8] is premultiplied by $\mathbf{B}_0^{-1}$, the result is

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1\mathbf{y}_{t-1} + \mathbf{\Phi}_2\mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p\mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t, \qquad [11.6.9]$$

where

$$\mathbf{c} = \mathbf{B}_0^{-1}\mathbf{k} \qquad [11.6.10]$$
$$\mathbf{\Phi}_s = \mathbf{B}_0^{-1}\mathbf{B}_s \qquad \text{for } s = 1, 2, \ldots, p \qquad [11.6.11]$$
$$\boldsymbol{\varepsilon}_t = \mathbf{B}_0^{-1}\mathbf{u}_t. \qquad [11.6.12]$$

Assuming that [11.6.8] is parameterized sufficiently richly that $\mathbf{u}_t$ is vector white noise, then $\boldsymbol{\varepsilon}_t$ will also be vector white noise and [11.6.9] will be recognized as the vector autoregressive representation for the dynamic structural system [11.6.8]. Thus, a *VAR* can be viewed as the reduced form of a general dynamic structural model.

Interpreting Impulse-Response Functions

In Section 11.4 we calculated the impulse-response function

$$\frac{\partial \mathbf{y}_{t+s}}{\partial \varepsilon_{jt}}. \qquad [11.6.13]$$

This magnitude describes the effect of an innovation in the jth variable on future values of each of the variables in the system. According to [11.6.12], the *VAR* innovation ε_{jt} is a linear combination of the structural disturbances $\mathbf{u}_t$. For example,

it might turn out that

$$\varepsilon_{1t} = 0.3u_t^D - 0.6u_t^S + 0.1u_t^A - 0.5u_t^C.$$

In this case, if the cash held by the public is larger than would have been forecast using the *VAR* (ε_{1t} is positive), this might be because the public's demand for cash is higher than is normally associated with the current level of income and interest rate (that is, u_t^D is positive). Alternatively, ε_{1t} might be positive because the central bank has chosen to ease credit (u_t^C is negative), or a variety of other factors. In general, ε_{1t} represents a combination of all the different influences that matter for any variables in the economy. Viewed this way, it is not clear why the magnitude [11.6.13] is of particular interest.

By contrast, if we were able to calculate

$$\frac{\partial \mathbf{y}_{t+s}}{\partial u_t^C}, \qquad [11.6.14]$$

this would be of considerable interest. Expression [11.6.14] identifies the dynamic consequences for the economy if the central bank were to tighten credit more than usual and is a key magnitude for describing the effects of monetary policy on the economy.

Section 11.4 also discussed calculation of an orthogonalized impulse-response function. For $\mathbf{\Omega} = E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t')$, we found a lower triangular matrix $\mathbf{A}$ and a diagonal matrix $\mathbf{D}$ such that $\mathbf{\Omega} = \mathbf{ADA}'$. We then constructed the vector $\mathbf{A}^{-1}\boldsymbol{\varepsilon}_t$ and calculated the consequences of changes in each element of this vector for future values of $\mathbf{y}$.

Recall from [11.6.12] that the structural disturbances $\mathbf{u}_t$ are related to the *VAR* innovations $\boldsymbol{\varepsilon}_t$ by

$$\mathbf{u}_t = \mathbf{B}_0\boldsymbol{\varepsilon}_t. \qquad [11.6.15]$$

Suppose that it happened to be the case that the matrix of structural parameters $\mathbf{B}_0$ was exactly equal to the matrix $\mathbf{A}^{-1}$. Then the orthogonalized innovations would coincide with the true structural disturbances:

$$\mathbf{u}_t = \mathbf{B}_0\boldsymbol{\varepsilon}_t = \mathbf{A}^{-1}\boldsymbol{\varepsilon}_t. \qquad [11.6.16]$$

In this case, the method described in Section 11.4 could be used to find the answers to important questions such as [11.6.14].

Is there any reason to hope that $\mathbf{B}_0$ and $\mathbf{A}^{-1}$ would be the same matrix? Since $\mathbf{A}$ is lower triangular, this clearly requires $\mathbf{B}_0$ to be lower triangular. In the example [11.6.8], this would require that the current values of P, Y, and I do not influence money demand, that the current value of M but not that of Y or I enters into the aggregate supply curve, and so on. Such assumptions are rather unusual, though there may be another way to order the variables such that a recursive structure is more palatable. For example, a Keynesian might argue that prices respond to other economic variables only with a lag, so that the coefficients on current variables in the aggregate supply equation are all zero. Perhaps money and interest rates influence aggregate demand only with a lag, so that their current values are excluded from the aggregate demand equation. One might try to argue further that the interest rate affects desired money holdings only with a lag as well. Because most central banks monitor current economic conditions quite carefully, perhaps all the current values should be included in the equation for I_t. These assumptions suggest ordering the variables as $\mathbf{y}_t = (P_t, Y_t, M_t, I_t)'$, for which the structural model would

be

$$\begin{bmatrix} P_t \\ Y_t \\ M_t \\ I_t \end{bmatrix} = \begin{bmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_{21}^{(0)} & 0 & 0 & 0 \\ \beta_{31}^{(0)} & \beta_{32}^{(0)} & 0 & 0 \\ \beta_{41}^{(0)} & \beta_{42}^{(0)} & \beta_{43}^{(0)} & 0 \end{bmatrix} \begin{bmatrix} P_t \\ Y_t \\ M_t \\ I_t \end{bmatrix}$$

$$+ \begin{bmatrix} \beta_{11}^{(1)} & \beta_{12}^{(1)} & \beta_{13}^{(1)} & \beta_{14}^{(1)} \\ \beta_{21}^{(1)} & \beta_{22}^{(1)} & \beta_{23}^{(1)} & \beta_{24}^{(1)} \\ \beta_{31}^{(1)} & \beta_{32}^{(1)} & \beta_{33}^{(1)} & \beta_{34}^{(1)} \\ \beta_{41}^{(1)} & \beta_{42}^{(1)} & \beta_{43}^{(1)} & \beta_{44}^{(1)} \end{bmatrix} \begin{bmatrix} P_{t-1} \\ Y_{t-1} \\ M_{t-1} \\ I_{t-1} \end{bmatrix} + \cdots \qquad [11.6.17]$$

$$+ \begin{bmatrix} \beta_{11}^{(p)} & \beta_{12}^{(p)} & \beta_{13}^{(p)} & \beta_{14}^{(p)} \\ \beta_{21}^{(p)} & \beta_{22}^{(p)} & \beta_{23}^{(p)} & \beta_{24}^{(p)} \\ \beta_{31}^{(p)} & \beta_{32}^{(p)} & \beta_{33}^{(p)} & \beta_{34}^{(p)} \\ \beta_{41}^{(p)} & \beta_{42}^{(p)} & \beta_{43}^{(p)} & \beta_{44}^{(p)} \end{bmatrix} \begin{bmatrix} P_{t-p} \\ Y_{t-p} \\ M_{t-p} \\ I_{t-p} \end{bmatrix} + \begin{bmatrix} u_t^S \\ u_t^A \\ u_t^D \\ u_t^C \end{bmatrix}.$$

Suppose there exists such an ordering of the variables for which $\mathbf{B}_0$ is lower triangular. Write the dynamic structural model [11.6.8] as

$$\mathbf{B}_0\mathbf{y}_t = -\mathbf{\Gamma}\mathbf{x}_t + \mathbf{u}_t, \qquad [11.6.18]$$

where

$$\underset{[n\times(np+1)]}{-\mathbf{\Gamma}} \equiv [\mathbf{k} \quad \mathbf{B}_1 \quad \mathbf{B}_2 \quad \cdots \quad \mathbf{B}_p]$$

$$\underset{[(np+1)\times 1]}{\mathbf{x}_t} \equiv \begin{bmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \vdots \\ \mathbf{y}_{t-p} \end{bmatrix}.$$

Suppose, furthermore, that the disturbances in the structural equations are serially uncorrelated and uncorrelated with each other:

$$E(\mathbf{u}_t\mathbf{u}_\tau') = \begin{cases} \mathbf{D} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise,} \end{cases} \qquad [11.6.19]$$

where $\mathbf{D}$ is a diagonal matrix. The VAR is the reduced form of the dynamic structural model [11.6.18] and can be written as

$$\mathbf{y}_t = \mathbf{\Pi}'\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \qquad [11.6.20]$$

where

$$\mathbf{\Pi}' = -\mathbf{B}_0^{-1}\mathbf{\Gamma} \qquad [11.6.21]$$

$$\boldsymbol{\varepsilon}_t = \mathbf{B}_0^{-1}\mathbf{u}_t. \qquad [11.6.22]$$

Letting $\mathbf{\Omega}$ denote the variance-covariance matrix of $\boldsymbol{\varepsilon}_t$, [11.6.22] implies

$$\mathbf{\Omega} = E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \mathbf{B}_0^{-1}E(\mathbf{u}_t\mathbf{u}_t')(\mathbf{B}_0^{-1})' = \mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})'. \qquad [11.6.23]$$

Note that if the only restrictions on the dynamic structural model are that $\mathbf{B}_0$ is lower triangular with unit coefficients along the principal diagonal and that $\mathbf{D}$ is diagonal, then the structural model is just identified. To see this, note that these restrictions imply that $\mathbf{B}_0^{-1}$ must also be lower triangular with unit coefficients along the principal diagonal. Recall from Section 4.4 that given any positive definite symmetric matrix $\mathbf{\Omega}$, there exist a unique lower triangular matrix $\mathbf{A}$ with 1s along the principal diagonal and a diagonal matrix $\mathbf{D}$ with positive entries along the principal diagonal such that $\mathbf{\Omega} = \mathbf{ADA}'$. Thus, unique values $\mathbf{B}_0^{-1}$ and $\mathbf{D}$ of the required form can always be found that satisfy [11.6.23]. Moreover, any $\mathbf{B}_0$ matrix of this form is nonsingular, so that $\mathbf{\Gamma}$ in [11.6.21] can be calculated uniquely from $\mathbf{B}_0$ and $\mathbf{\Pi}$ as $\mathbf{\Gamma} = -\mathbf{B}_0\mathbf{\Pi}'$. Thus, given any allowable values for the reduced-form parameters ($\mathbf{\Pi}$ and $\mathbf{\Omega}$), there exist unique values for the structural parameters ($\mathbf{B}_0$, $\mathbf{\Gamma}$, and $\mathbf{D}$) of the specified form, establishing that the structural model is just identified.

Since the model is just identified, full-information maximum likelihood (*FIML*) estimates of ($\mathbf{B}_0$, $\mathbf{\Gamma}$, and $\mathbf{D}$) can be obtained by first maximizing the likelihood function with respect to the reduced-form parameters ($\mathbf{\Pi}$ and $\mathbf{\Omega}$) and then using the unique mapping from reduced-form parameters to find the structural parameters. The maximum likelihood estimates of $\mathbf{\Pi}$ are found from *OLS* regressions of the elements of $\mathbf{y}_t$ on $\mathbf{x}_t$, and the *MLE* of $\mathbf{\Omega}$ is obtained from the variance-covariance matrix of the residuals from these regressions. The estimates $\hat{\mathbf{B}}_0^{-1}$ and $\hat{\mathbf{D}}$ are then found from the triangular factorization of $\hat{\mathbf{\Omega}}$. This, however, is precisely the procedure described in calculating the orthogonalized innovations in Section 11.4. The estimate $\hat{\mathbf{A}}$ described there is thus the same as the *FIML* estimate of $\mathbf{B}_0^{-1}$. The vector of orthogonalized residuals $\mathbf{u}_t = \mathbf{A}^{-1}\boldsymbol{\varepsilon}_t$ would correspond to the vector of structural disturbances, and the orthogonalized impulse-response coefficients would give the dynamic consequences of the structural events represented by $\mathbf{u}_t$, provided that the structural model is lower triangular as in [11.6.17].

Nonrecursive Structural VARs

Even if the structural model cannot be written in lower triangular form, it may be possible to give a structural interpretation to a *VAR* using a similar idea to that in equation [11.6.23]. Specifically, a structural model specifies a set of restrictions on $\mathbf{B}_0$ and $\mathbf{D}$, and we can try to find values satisfying these restrictions such that $\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})' = \mathbf{\Omega}$. This point was developed by Bernanke (1986), Blanchard and Watson (1986), and Sims (1986).

For illustration, consider again the model of supply and demand discussed in equations [9.3.2] and [9.3.3]. In that specification, quantity (q_t) and price (p_t) were endogeneous variables and weather (w_t) was exogenous, and it was assumed that both disturbances were i.i.d. The structural *VAR* approach to this model would allow quite general dynamics by adding p lags of all three variables to equations [9.3.2] and [9.3.3], as well as adding a third equation to describe the dynamic behavior of weather. Weather presumably does not depend on the behavior of the market, so the third equation would for this example just be a univariate autoregression. The model would then be

$$\begin{aligned} q_t = {} & \beta p_t + \beta_{11}^{(1)} q_{t-1} + \beta_{12}^{(1)} p_{t-1} + \beta_{13}^{(1)} w_{t-1} \\ & + \beta_{11}^{(2)} q_{t-2} + \beta_{12}^{(2)} p_{t-2} + \beta_{13}^{(2)} w_{t-2} + \cdots \\ & + \beta_{11}^{(p)} q_{t-p} + \beta_{12}^{(p)} p_{t-p} + \beta_{13}^{(p)} w_{t-p} + u_t^d \end{aligned} \qquad [11.6.24]$$

$$
\begin{aligned}
q_t = {} & \gamma p_t + h w_t + \beta_{21}^{(1)} q_{t-1} + \beta_{22}^{(1)} p_{t-1} + \beta_{23}^{(1)} w_{t-1} \\
& + \beta_{21}^{(2)} q_{t-2} + \beta_{22}^{(2)} p_{t-2} + \beta_{23}^{(2)} w_{t-2} + \cdots \\
& + \beta_{21}^{(p)} q_{t-p} + \beta_{22}^{(p)} p_{t-p} + \beta_{23}^{(p)} w_{t-p} + u_t^s
\end{aligned} \qquad [11.6.25]
$$

$$
w_t = \beta_{33}^{(1)} w_{t-1} + \beta_{33}^{(2)} w_{t-2} + \cdots + \beta_{33}^{(p)} w_{t-p} + u_t^w. \qquad [11.6.26]
$$

We could then take $(u_t^d, u_t^s, u_t^w)'$ to be a white noise vector with diagonal variance-covariance matrix given by $\mathbf{D}$. This is an example of a structural model [11.6.18] in which

$$
\mathbf{B}_0 = \begin{bmatrix} 1 & -\beta & 0 \\ 1 & -\gamma & -h \\ 0 & 0 & 1 \end{bmatrix}. \qquad [11.6.27]
$$

There is no way to order the variables so as to make the matrix $\mathbf{B}_0$ lower triangular. However, equation [11.6.22] indicates that the structural disturbances $\mathbf{u}_t$ are related to the *VAR* residuals $\boldsymbol{\varepsilon}_t$ by $\boldsymbol{\varepsilon}_t = \mathbf{B}_0^{-1}\mathbf{u}_t$. Thus, if $\mathbf{B}_0$ is estimated by maximum likelihood, then the impulse-response functions could be calculated as in Section 11.4 with $\mathbf{A}$ replaced by $\mathbf{B}_0^{-1}$, and the results would give the effects of each of the structural disturbances on subsequent values of variables of the system. Specifically,

$$
\frac{\partial \boldsymbol{\varepsilon}_t}{\partial \mathbf{u}_t'} = \mathbf{B}_0^{-1},
$$

so that the effect on $\boldsymbol{\varepsilon}_t$ of the jth structural disturbance u_{jt} is given by $\mathbf{b}^j$, the jth column of $\mathbf{B}_0^{-1}$. Thus, we would calculate

$$
\frac{\partial \mathbf{y}_{t+s}}{\partial u_{jt}} = \frac{\partial \mathbf{y}_{t+s}}{\partial \boldsymbol{\varepsilon}_t'} \frac{\partial \boldsymbol{\varepsilon}_t}{\partial u_{jt}} = \boldsymbol{\Psi}_s \mathbf{b}^j
$$

for $\boldsymbol{\Psi}_s$ the $(n \times n)$ matrix of coefficients for the sth lag of the $MA(\infty)$ representation [11.4.1].

FIML *Estimation of a Structural* VAR *with Unrestricted Dynamics*

FIML estimation is particularly simple if there are no restrictions on the coefficients $\boldsymbol{\Gamma}$ on lagged variables in [11.6.18]. For example, this would require including lagged values of p_{t-j} and q_{t-j} in the weather equation [11.6.26]. Using [11.6.23], the log likelihood function for the system [11.6.18] can be written as

$$
\begin{aligned}
\mathcal{L}(\mathbf{B}_0, \mathbf{D}, \boldsymbol{\Pi}) = {} & -(Tn/2)\log(2\pi) - (T/2)\log|\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})'| \\
& - (1/2)\sum_{t=1}^{T} [\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t]'[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}[\mathbf{y}_t - \boldsymbol{\Pi}'\mathbf{x}_t].
\end{aligned} \qquad [11.6.28]
$$

If there are no restrictions on lagged dynamics, this is maximized with respect to $\boldsymbol{\Pi}$ by *OLS* regression of $\mathbf{y}_t$ on $\mathbf{x}_t$. Substituting this estimate into [11.6.28] as in

[11.1.25] produces

$$\mathcal{L}(\mathbf{B}_0, \mathbf{D}, \hat{\mathbf{\Pi}}) = -(Tn/2)\log(2\pi) - (T/2)\log|\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})'| - (1/2)\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t'[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}\hat{\boldsymbol{\varepsilon}}_t. \qquad [11.6.29]$$

But

$$\begin{aligned}\sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t'[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}\hat{\boldsymbol{\varepsilon}}_t &= \sum_{t=1}^{T}\text{trace}\{\hat{\boldsymbol{\varepsilon}}_t'[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}\hat{\boldsymbol{\varepsilon}}_t\}\\ &= \sum_{t=1}^{T}\text{trace}\{[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'\}\\ &= \text{trace}\{[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}T\cdot\hat{\mathbf{\Omega}}\}\\ &= T \times \text{trace}\{[\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})']^{-1}\hat{\mathbf{\Omega}}\}\\ &= T \times \text{trace}\{(\mathbf{B}_0'\mathbf{D}^{-1}\mathbf{B}_0)\hat{\mathbf{\Omega}}\}.\end{aligned} \qquad [11.6.30]$$

Furthermore,

$$\log|\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})'| = \log\{|\mathbf{B}_0^{-1}|\cdot|\mathbf{D}|\cdot|\mathbf{B}_0^{-1}|\} = -\log|\mathbf{B}_0|^2 + \log|\mathbf{D}|. \qquad [11.6.31]$$

Substituting [11.6.31] and [11.6.30] into [11.6.29], *FIML* estimates of the structural parameters are found by choosing $\mathbf{B}_0$ and $\mathbf{D}$ so as to maximize

$$\mathcal{L}(\mathbf{B}_0, \mathbf{D}, \hat{\mathbf{\Pi}}) = -(Tn/2)\log(2\pi) + (T/2)\log|\mathbf{B}_0|^2 - (T/2)\log|\mathbf{D}| - (T/2)\,\text{trace}\{(\mathbf{B}_0'\mathbf{D}^{-1}\mathbf{B}_0)\hat{\mathbf{\Omega}}\}. \qquad [11.6.32]$$

Using calculations similar to those used to analyze [11.1.25], one can show that if there exist unique matrices $\mathbf{B}_0$ and $\mathbf{D}$ of the required form satisfying $\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})' = \mathbf{\Omega}$, then maximization of [11.6.32] will produce estimates $\hat{\mathbf{B}}_0$ and $\hat{\mathbf{D}}$ satisfying

$$\hat{\mathbf{B}}_0^{-1}\hat{\mathbf{D}}(\hat{\mathbf{B}}_0^{-1})' = \hat{\mathbf{\Omega}}. \qquad [11.6.33]$$

This is a nonlinear system of equations, and numerical maximization of [11.6.32] offers a convenient general approach to finding a solution to this system of equations.

Identification of Structural VARs

The existence of a unique maximum of [11.6.32] requires both an order condition and a rank condition for identification. The order condition is that $\mathbf{B}_0$ and $\mathbf{D}$ have no more unknown parameters than $\mathbf{\Omega}$. Since $\mathbf{\Omega}$ is symmetric, it can be summarized by $n(n + 1)/2$ distinct values. If $\mathbf{D}$ is diagonal, it requires n parameters, meaning that $\mathbf{B}_0$ can have no more than $n(n - 1)/2$ free parameters. For the supply-and-demand example of [11.6.24] through [11.6.26], $n = 3$, and the matrix $\mathbf{B}_0$ in [11.6.27] has $3(3 - 1)/2 = 3$ free parameters (β, γ, and h). Thus, that example satisfies the order condition for identification.

Even if the order condition is satisfied, the model may still not be identified. For example, suppose that

$$\mathbf{B}_0 = \begin{bmatrix} 1 & -\beta & 0 \\ 1 & -\gamma & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Even though this specification satisfies the order condition, it fails the rank condition, since the value of the likelihood function will be unchanged if β and γ are switched along with σ_d^2 and σ_s^2.

To characterize the rank condition, suppose that there are n_B elements of $\mathbf{B}_0$ that must be estimated; collect these in an $(n_B \times 1)$ vector $\boldsymbol{\theta}_B$. The identifying assumptions can be represented as a known $(n^2 \times n_B)$ matrix $\mathbf{S}_B$ and a known $(n^2 \times 1)$ vector $\mathbf{s}_B$ for which

$$\text{vec}(\mathbf{B}_0) = \mathbf{S}_B\boldsymbol{\theta}_B + \mathbf{s}_B. \qquad [11.6.34]$$

For example, for the dynamic model of supply and demand represented by [11.6.27],

$$\text{vec}(\mathbf{B}_0) = \begin{bmatrix} 1 \\ 1 \\ 0 \\ -\beta \\ -\gamma \\ 0 \\ 0 \\ -h \\ 1 \end{bmatrix} \qquad \boldsymbol{\theta}_B = \begin{bmatrix} \beta \\ \gamma \\ h \end{bmatrix}$$

$$\mathbf{S}_B = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{s}_B = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Similarly, collect the unknown elements of $\mathbf{D}$ in an $(n_D \times 1)$ vector $\boldsymbol{\theta}_D$, with

$$\text{vec}(\mathbf{D}) = \mathbf{S}_D\boldsymbol{\theta}_D + \mathbf{s}_D \qquad [11.6.35]$$

for $\mathbf{S}_D$ an $(n^2 \times n_D)$ matrix and $\mathbf{s}_D$ an $(n^2 \times 1)$ vector. For the supply-and-demand example,

$$\text{vec}(\mathbf{D}) = \begin{bmatrix} \sigma_d^2 \\ 0 \\ 0 \\ 0 \\ \sigma_s^2 \\ 0 \\ 0 \\ 0 \\ \sigma_w^2 \end{bmatrix} \qquad \boldsymbol{\theta}_D = \begin{bmatrix} \sigma_d^2 \\ \sigma_s^2 \\ \sigma_w^2 \end{bmatrix}$$

$$\mathbf{S}_D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \qquad \mathbf{s}_D = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Since [11.6.33] is an equation relating two symmetric matrices, there are $n^* \equiv n(n+1)/2$ separate conditions, represented by

$$\text{vech}(\mathbf{\Omega}) = \text{vech}\left([\mathbf{B}_0(\boldsymbol{\theta}_B)]^{-1}[\mathbf{D}(\boldsymbol{\theta}_D)]\{[\mathbf{B}_0(\boldsymbol{\theta}_B)]^{-1}\}'\right). \qquad [11.6.36]$$

Denote the right side of [11.6.36] by $\mathbf{f}(\boldsymbol{\theta}_B, \boldsymbol{\theta}_D)$, where $\mathbf{f}$: $(\mathbb{R}^{n^*} \times \mathbb{R}^{n^*}) \to \mathbb{R}^{n^*}$:

$$\text{vech}(\mathbf{\Omega}) = \mathbf{f}(\boldsymbol{\theta}_B, \boldsymbol{\theta}_D). \qquad [11.6.37]$$

Appendix 11.B shows that the $[n^* \times (n_B + n_D)]$ matrix of derivatives of this function is given by

$$\begin{aligned} \mathbf{J} &= \left[\frac{\partial\, \text{vech}(\mathbf{\Omega})}{\partial \boldsymbol{\theta}_B'} \quad \frac{\partial\, \text{vech}(\mathbf{\Omega})}{\partial \boldsymbol{\theta}_D'}\right] \\ &= \left[[-2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\mathbf{S}_B] \quad \mathbf{D}_n^+[(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\mathbf{S}_D\right], \end{aligned} \qquad [11.6.38]$$

where $\mathbf{D}_n^+$ is the $(n^* \times n^2)$ matrix defined in [11.1.45].

Suppose that the columns of the matrix in [11.6.38] were linearly dependent; that is, suppose there exists a nonzero $[(n_B + n_D) \times 1]$ vector $\boldsymbol{\lambda}$ such that $\mathbf{J}\boldsymbol{\lambda} = \mathbf{0}$. This would mean that if a small multiple of $\boldsymbol{\lambda}$ were added to $(\boldsymbol{\theta}_B', \boldsymbol{\theta}_D')'$, the model would imply the same probability distribution for the data. We would have no basis for distinguishing between these alternative values for $(\boldsymbol{\theta}_B', \boldsymbol{\theta}_D')$, meaning that the model would be unidentified.

Thus, the rank condition for identification of a structural *VAR* requires that the $(n_B + n_D)$ columns of the matrix $\mathbf{J}$ in [11.6.38] be linearly independent.[12] The order condition is that the number of rows of $\mathbf{J}$ ($n^* = n(n+1)/2$) be at least as great as the number of columns.

To check this condition in practice, the simplest approach is usually to make a guess as to the values of the structural parameters and check $\mathbf{J}$ numerically. Giannini (1992) derived an alternative expression for the rank condition and provided computer software for checking it numerically.

Structural VAR *with Restrictions on* $\mathbf{\Pi}$

The supply-and-demand example of [11.6.24] to [11.6.26] did not satisfy the assumptions behind the derivation of [11.6.32], because [11.6.26] imposed the restriction that lagged values of p and q did not belong in the weather equation. Where such restrictions are imposed, it is no longer that case that the *FIML* estimates of $\mathbf{\Pi}$ are obtained by *OLS*, and system parameters would have to be estimated as described in Section 11.3. As an alternative, *OLS* estimation of [11.6.24] through [11.6.26] would still give consistent estimates of $\mathbf{\Pi}$, and the variance-covariance matrix of the residuals from these regressions would provide a consistent estimate $\hat{\mathbf{\Omega}}$. One could still use this estimate in [11.6.32], and the resulting maximization problem would give reasonable estimates of $\mathbf{B}_0$ and $\mathbf{D}$.

Structural VARs *and Forward-Looking Behavior*

The supply-and-demand example assumed that lagged values of price and quantity did not appear in the equation for weather. The spirit of *VAR*s is that

[12]This condition characterizes *local* identification; it may be that even if a model satisfies both the rank and the order condition, there are two noncontiguous values of $(\boldsymbol{\theta}_B', \boldsymbol{\theta}_D')$ for which the likelihood has the same value for all realizations of the data. See Rothenberg (1971, Theorem 6, p. 585).

such assumptions ought to be tested before being imposed. What should we conclude if, contrary to our prior expectations, the price of oranges turned out to Granger-cause the weather in Florida? It certainly cannot be that the price is a cause of the weather. Instead, such a finding would suggest forward-looking behavior on the part of buyers or sellers of oranges; for example, it may be that if buyers anticipate bad weather in the future, they bid up the price of oranges today. If this should prove to be the case, the identifying assumption in [11.6.24] that demand depends on the weather only through its effect on the current price needs to be reexamined. Proper modeling of forward-looking behavior can provide an alternative way to identify *VAR*s, as explored by Flavin (1981), Hansen and Sargent (1981), and Keating (1990), among others.

Other Approaches to Identifying Structural VARs

Identification was discussed in previous subsections primarily in terms of exclusion restrictions on the matrix of structural coefficients $\mathbf{B}_0$. Blanchard and Diamond (1989, 1990) used a priori assumptions about the signs of structural parameters to identify a range of values of $\mathbf{B}_0$ consistent with the data. Shapiro and Watson (1988) and Blanchard and Quah (1989) used assumptions about long-run multipliers to achieve identification.

A Critique of Structural VARs

Structural *VAR*s have appeal for two different kinds of inquiry. The first potential user is someone who is primarily interested in estimating a structural equation such as the money demand function in [11.6.1]. If a model imposes restrictions on the dynamics of the relationship, it seems good practice to test these restrictions against a more general specification such as [11.6.5] before relying on the restricted model for inference. Furthermore, in order to estimate the dynamic consequences of, say, income on money demand, we have to take into account the fact that, historically, when income goes up, this has typically been associated with future changes in income and interest rates. What time path for these explanatory variables should be assumed in order to assess the consequences for money demand at time $t + s$ of a change in income at time t? A *VAR* offers a framework for posing this question — we use the time path that would historically be predicted for those variables following an unanticipated change in income.

A second potential user is someone who is interested in summarizing the dynamics of a vector $\mathbf{y}_t$ while imposing as few restrictions as possible. Insofar as this summary includes calculation of impulse-response functions, we need some motivation for what the statictics mean. Suppose we find that there is a temporary rise in income following an innovation in money. One is tempted to interpret this finding as suggesting that expansionary monetary policy has a positive but temporary effect on output. However, such an interpretation implicitly assumes that the orthogonalized "money innovation" is the same as the disturbance term in a description of central bank policy. Insofar as impulse-response functions are used to make statements that are structural in nature, it seems reasonable to try to use an orthogonalization that represents our understanding of these relationships as well as possible. This point has been forcefully argued by Cooley and LeRoy (1985), Leamer (1985), Bernanke (1986), and Blanchard (1989), among others.

Even so, it must be recognized that convincing identifying assumptions are hard to come by. For example, the ordering in [11.6.17] is clearly somewhat arbitrary, and the exclusion restrictions are difficult to defend. Indeed, if there were compelling identifying assumptions for such a system, the fierce debates among

macroeconomists would have been settled long ago! Simultaneous equations bias is very pervasive in the social sciences, and drawing structural inferences from observed correlations must always proceed with great care. We surely cannot always expect to find credible identifying assumptions to enable us to identify the causal relations among any arbitrary set of n variables on which we have data.

11.7. *Standard Errors for Impulse-Response Functions*

Standard Errors for Nonorthogonalized Impulse-Response Function Based on Analytical Derivatives

Section 11.4 discussed how $\boldsymbol{\Psi}_s$, the matrix of impulse-response coefficients at lag s, would be constructed from knowledge of the autoregressive coefficients. In practice, the autoregressive coefficients are not known with certainty but must be estimated by *OLS* regressions. When the estimated values of the autoregressive coefficients are used to calculate $\boldsymbol{\Psi}_s$, it is useful to report the implied standard errors for the estimates $\hat{\boldsymbol{\Psi}}_s$.[13]

Adopting the notation from Proposition 11.1, let $k = np + 1$ denote the number of coefficients in each equation of the *VAR* and let $\boldsymbol{\pi} \equiv \text{vec}(\boldsymbol{\Pi})$ denote the $(nk \times 1)$ vector of parameters for all the equations; the first k elements of $\boldsymbol{\pi}$ give the constant term and autoregressive coefficients for the first equation, the next k elements of $\boldsymbol{\pi}$ give the parameters for the second equation, and so on. Let $\boldsymbol{\psi}_s \equiv \text{vec}(\boldsymbol{\Psi}_s')$ denote the $(n^2 \times 1)$ vector of moving average coefficients associated with lag s. The first n elements of $\boldsymbol{\psi}_s$ are given by the first row of $\boldsymbol{\Psi}_s$ and identify the response of $y_{1,t+s}$ to $\boldsymbol{\varepsilon}_t$. The next n elements of $\boldsymbol{\psi}_s$ are given by the second row of $\boldsymbol{\Psi}_s$ and identify the response of $y_{2,t+s}$ to $\boldsymbol{\varepsilon}_t$, and so on. Given the values of the autoregressive coefficients in $\boldsymbol{\pi}$, the *VAR* can be simulated to calculate $\boldsymbol{\psi}_s$. Thus, $\boldsymbol{\psi}_s$ could be regarded as a nonlinear function of $\boldsymbol{\pi}$, represented by the function $\boldsymbol{\psi}_s(\boldsymbol{\pi})$, $\boldsymbol{\psi}_s$: $\mathbb{R}^{nk} \rightarrow \mathbb{R}^{n^2}$.

The impulse-response coefficients are estimated by replacing $\boldsymbol{\pi}$ with the *OLS* estimates $\hat{\boldsymbol{\pi}}_T$, generating the estimate $\hat{\boldsymbol{\psi}}_{s,T} = \boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}}_T)$. Recall that under the conditions of Proposition 11.1, $\sqrt{T}(\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) \xrightarrow{L} \mathbf{X}$, where

$$\mathbf{X} \sim N\Big(\mathbf{0}, (\boldsymbol{\Omega} \otimes \mathbf{Q}^{-1})\Big). \qquad [11.7.1]$$

Standard errors for $\hat{\boldsymbol{\psi}}_s$ can then be calculated by applying Proposition 7.4:

$$\sqrt{T}(\hat{\boldsymbol{\psi}}_{s,T} - \boldsymbol{\psi}_s) \xrightarrow{L} \mathbf{G}_s\mathbf{X},$$

where

$$\underset{(n^2 \times nk)}{\mathbf{G}_s} = \frac{\partial \boldsymbol{\psi}_s(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}'}. \qquad [11.7.2]$$

That is,

$$\sqrt{T}(\hat{\boldsymbol{\psi}}_{s,T} - \boldsymbol{\psi}_s) \xrightarrow{L} N\Big(\mathbf{0}, \mathbf{G}_s(\boldsymbol{\Omega} \otimes \mathbf{Q}^{-1})\mathbf{G}_s'\Big). \qquad [11.7.3]$$

Standard errors for an estimated impulse-response coefficient are given by the

[13]Calculations related to those developed in this section appeared in Baillie (1987), Lütkepohl (1989, 1990), and Giannini (1992). Giannini provided computer software for calculating some of these magnitudes.

square root of the associated diagonal element of $(1/T)\hat{\mathbf{G}}_{s,T}(\hat{\mathbf{\Omega}}_T \otimes \hat{\mathbf{Q}}_T^{-1})\hat{\mathbf{G}}'_{s,T}$, where

$$\hat{\mathbf{G}}_{s,T} = \left.\frac{\partial \boldsymbol{\psi}_s(\boldsymbol{\pi})}{\partial \boldsymbol{\pi}'}\right|_{\boldsymbol{\pi} = \hat{\boldsymbol{\pi}}_T}$$

$$\hat{\mathbf{Q}}_T = (1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}'_t,$$

with $\mathbf{x}_t$ and $\hat{\mathbf{\Omega}}_T$ as defined in Proposition 11.1.

To apply this result, we need an expression for the matrix $\mathbf{G}_s$ in [11.7.2]. Appendix 11.B to this chapter establishes that the sequence $\{\mathbf{G}_s\}_{s=1}^{m}$ can be calculated by iterating on

$$\begin{aligned}\mathbf{G}_s = {} & [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \mathbf{\Psi}'_{s-1} \quad \mathbf{\Psi}'_{s-2} \quad \cdots \quad \mathbf{\Psi}'_{s-p})] + (\mathbf{\Phi}_1 \otimes \mathbf{I}_n)\mathbf{G}_{s-1} \\ & + (\mathbf{\Phi}_2 \otimes \mathbf{I}_n)\mathbf{G}_{s-2} + \cdots + (\mathbf{\Phi}_p \otimes \mathbf{I}_n)\mathbf{G}_{s-p}.\end{aligned} \qquad [11.7.4]$$

Here $\mathbf{0}_{n1}$ denotes an $(n \times 1)$ vector of zeros. The iteration is initialized by setting $\mathbf{G}_0 = \mathbf{G}_{-1} = \cdots = \mathbf{G}_{-p+1} = \mathbf{0}_{n^2,nk}$. It is also understood that $\mathbf{\Psi}_0 = \mathbf{I}_n$ and $\mathbf{\Psi}_s = \mathbf{0}_{nn}$ for $s < 0$. Thus, for example,

$$\mathbf{G}_1 = [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \mathbf{I}_n \quad \mathbf{0}_{nn} \quad \cdots \quad \mathbf{0}_{nn})]$$

$$\mathbf{G}_2 = [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \mathbf{\Psi}'_1 \quad \mathbf{I}_n \quad \cdots \quad \mathbf{0}_{nn})] + (\mathbf{\Phi}_1 \otimes \mathbf{I}_n)\mathbf{G}_1.$$

A closed-form solution for [11.7.4] is given by

$$\mathbf{G}_s = \sum_{i=1}^{s} [\mathbf{\Psi}_{i-1} \otimes (\mathbf{0}_{n1} \quad \mathbf{\Psi}'_{s-i} \quad \mathbf{\Psi}'_{s-i-1} \quad \cdots \quad \mathbf{\Psi}'_{s-i-p+1})]. \qquad [11.7.5]$$

Alternative Approaches to Calculating Standard Errors for Nonorthogonalized Impulse-Response Function

The matrix of derivatives $\mathbf{G}_s$ can alternatively be calculated numerically as follows. First we use the *OLS* estimates $\hat{\boldsymbol{\pi}}$ to calculate $\boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}})$ for $s = 1, 2, \ldots, m$. We then increase the value of the ith element of $\boldsymbol{\pi}$ by some small amount Δ, holding all other elements constant, and evaluate $\boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}} + \mathbf{e}_i\Delta)$ for $s = 1, 2, \ldots, m$, where $\mathbf{e}_i$ denotes the ith column of $\mathbf{I}_{nk}$. Then the $(n^2 \times 1)$ vector

$$\frac{\boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}} + \mathbf{e}_i\Delta) - \boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}})}{\Delta}$$

gives an estimate of the ith column of $\mathbf{G}_s$. By conducting separate evaluations of the sequence $\boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}} + \mathbf{e}_i\Delta)$ for each $i = 1, 2, \ldots, nk$, all of the columns of $\mathbf{G}_s$ can be filled in.

Monte Carlo methods can also be used to infer the distribution of $\boldsymbol{\psi}_s(\hat{\boldsymbol{\pi}})$. Here we would randomly generate an $(nk \times 1)$ vector drawn from a $N(\hat{\boldsymbol{\pi}}, (1/T)(\hat{\mathbf{\Omega}} \otimes \hat{\mathbf{Q}}^{-1}))$ distribution. Denote this vector by $\boldsymbol{\pi}^{(1)}$, and calculate $\boldsymbol{\psi}_s(\boldsymbol{\pi}^{(1)})$. Draw a second vector $\boldsymbol{\pi}^{(2)}$ from the same distribution and calculate $\boldsymbol{\psi}_s(\boldsymbol{\pi}^{(2)})$. Repeat this for, say, 10,000 separate simulations. If 9500 of these simulations result in a value of the first element of $\boldsymbol{\psi}_s$ that is between $\underline{\psi}_{s1}$ and $\overline{\psi}_{s1}$, then $(\underline{\psi}_{s1}, \overline{\psi}_{s1})$ can be used as a 95% confidence interval for the first element of $\hat{\boldsymbol{\psi}}_s$.

Runkle (1987) employed a related approach based on *bootstrapping*. The idea behind bootstrapping is to obtain an estimate of the small-sample distribution of $\hat{\boldsymbol{\pi}}$ without assuming that the innovations $\boldsymbol{\varepsilon}_t$ are Gaussian. To implement this procedure, first estimate the *VAR* and save the coefficient estimates $\hat{\boldsymbol{\pi}}$ and the fitted residuals $\{\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2, \ldots, \hat{\boldsymbol{\varepsilon}}_T\}$. Then consider an artificial random variable $\mathbf{u}$ that has probability $(1/T)$ of taking on each of the particular values $\{\hat{\boldsymbol{\varepsilon}}_1, \hat{\boldsymbol{\varepsilon}}_2, \ldots, \hat{\boldsymbol{\varepsilon}}_T\}$. The

hope is that the distribution of $\mathbf{u}$ is similar to the distribution of the true population $\boldsymbol{\varepsilon}$'s. Then take a random draw from this distribution (denoted $\mathbf{u}_1^{(1)}$), and use this to construct the first innovation in an artificial sample; that is, set

$$\mathbf{y}_1^{(1)} = \hat{\mathbf{c}} + \hat{\boldsymbol{\Phi}}_1\mathbf{y}_0 + \hat{\boldsymbol{\Phi}}_2\mathbf{y}_{-1} + \cdots + \hat{\boldsymbol{\Phi}}_p\mathbf{y}_{-p+1} + \mathbf{u}_1^{(1)},$$

where $\mathbf{y}_0, \mathbf{y}_{-1}, \ldots,$ and $\mathbf{y}_{-p+1}$ denote the presample values of $\mathbf{y}$ that were actually observed in the historical data. Taking a second draw $\mathbf{u}_2^{(1)}$, generate

$$\mathbf{y}_2^{(1)} = \hat{\mathbf{c}} + \hat{\boldsymbol{\Phi}}_1\mathbf{y}_1^{(1)} + \hat{\boldsymbol{\Phi}}_2\mathbf{y}_0 + \cdots + \hat{\boldsymbol{\Phi}}_p\mathbf{y}_{-p+2} + \mathbf{u}_2^{(1)}.$$

Note that this second draw is with replacement; that is, there is a $(1/T)$ chance that $\mathbf{u}_1^{(1)}$ is exactly the same as $\mathbf{u}_2^{(1)}$. Proceeding in this fashion, a full sample $\{\mathbf{y}_1^{(1)}, \mathbf{y}_2^{(1)}, \ldots, \mathbf{y}_T^{(1)}\}$ can be generated. A *VAR* can be fitted by *OLS* to these simulated data (again taking presample values of $\mathbf{y}$ as their historical values), producing an estimate $\hat{\boldsymbol{\pi}}^{(1)}$. From this estimate, the magnitude $\psi_s(\hat{\boldsymbol{\pi}}^{(1)})$ can be calculated. Next, generate a second set of T draws from the distribution of $\mathbf{u}$, denoted $\{\mathbf{u}_1^{(2)}, \mathbf{u}_2^{(2)}, \ldots, \mathbf{u}_T^{(2)}\}$, fit $\hat{\boldsymbol{\pi}}^{(2)}$ to these data by *OLS*, and calculate $\psi_s(\hat{\boldsymbol{\pi}}^{(2)})$. A series of 10,000 such simulations could be undertaken, and a 95% confidence interval for $\psi_{s1}(\hat{\boldsymbol{\pi}})$ is then inferred from the range that includes 95% of the values for $\psi_{s1}(\hat{\boldsymbol{\pi}}^{(i)})$.

Standard Errors for Parameters of a Structural VAR

Recall from Proposition 11.2 and equation [11.1.48] that if the innovations are Gaussian,

$$\sqrt{T}[\text{vech}(\hat{\boldsymbol{\Omega}}_T) - \text{vech}(\boldsymbol{\Omega})] \xrightarrow{L} N\Big(\mathbf{0}, 2\mathbf{D}_n^+(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\mathbf{D}_n^+)'\Big).$$

The estimates of the parameters of a structural *VAR* ($\hat{\mathbf{B}}_0$ and $\hat{\mathbf{D}}$) are determined as implicit functions of $\hat{\boldsymbol{\Omega}}$ from

$$\hat{\boldsymbol{\Omega}} = \hat{\mathbf{B}}_0^{-1}\hat{\mathbf{D}}(\hat{\mathbf{B}}_0^{-1})'. \qquad [11.7.6]$$

As in equation [11.6.34], the unknown elements of $\mathbf{B}_0$ are summarized by an $(n_B \times 1)$ vector $\boldsymbol{\theta}_B$ with $\text{vec}(\mathbf{B}_0) = \mathbf{S}_B\boldsymbol{\theta}_B + \mathbf{s}_B$. Similarly, as in [11.6.35], it is assumed that $\text{vec}(\mathbf{D}) = \mathbf{S}_D\boldsymbol{\theta}_D + \mathbf{s}_D$ for $\boldsymbol{\theta}_D$ an $(n_D \times 1)$ vector. It then follows from Proposition 7.4 that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{B,T} - \boldsymbol{\theta}_B) \xrightarrow{L} N\Big(\mathbf{0}, 2\mathbf{G}_B\mathbf{D}_n^+(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\mathbf{D}_n^+)'\mathbf{G}_B'\Big) \qquad [11.7.7]$$

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{D,T} - \boldsymbol{\theta}_D) \xrightarrow{L} N\Big(\mathbf{0}, 2\mathbf{G}_D\mathbf{D}_n^+(\boldsymbol{\Omega} \otimes \boldsymbol{\Omega})(\mathbf{D}_n^+)'\mathbf{G}_D'\Big), \qquad [11.7.8]$$

where

$$\underset{(n_B \times n^*)}{\mathbf{G}_B} = \frac{\partial \boldsymbol{\theta}_B}{\partial[\text{vech}(\boldsymbol{\Omega})]'} \qquad [11.7.9]$$

$$\underset{(n_D \times n^*)}{\mathbf{G}_D} = \frac{\partial \boldsymbol{\theta}_D}{\partial[\text{vech}(\boldsymbol{\Omega})]'} \qquad [11.7.10]$$

and $n^* \equiv n(n + 1)/2$.

Equation [11.6.38] gave an expression for the $[n^* \times (n_B + n_D)]$ matrix:

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial\ \text{vech}(\boldsymbol{\Omega})}{\partial \boldsymbol{\theta}_B'} & \dfrac{\partial\ \text{vech}(\boldsymbol{\Omega})}{\partial \boldsymbol{\theta}_D'} \end{bmatrix}.$$

We noted there that if the model is to be identified, the columns of this matrix must be linearly independent. In the just-identified case, $n^* = (n_B + n_D)$ and $\mathbf{J}^{-1}$

exists, from which

$$\begin{bmatrix} \mathbf{G}_B \\ \mathbf{G}_D \end{bmatrix} = \mathbf{J}^{-1}. \qquad [11.7.11]$$

Standard Errors for Orthogonalized Impulse-Response Functions

Section 11.6 described calculation of the following $(n \times n)$ matrix:

$$\mathbf{H}_s = \mathbf{\Psi}_s \mathbf{B}_0^{-1}. \qquad [11.7.12]$$

The row i, column j element of this matrix measures the effect of the jth structural disturbance (u_{jt}) on the ith variable in the system $(y_{i,t+s})$ after a lag of s periods. Collect these magnitudes in an $(n^2 \times 1)$ vector $\mathbf{h}_s \equiv \text{vec}(\mathbf{H}_s')$. Thus, the first n elements of $\mathbf{h}_s$ give the effect of $\mathbf{u}_t$ on $y_{1,t+s}$, the next n elements give the effect of $\mathbf{u}_t$ on $y_{2,t+s}$, and so on.

Since $\hat{\mathbf{\Psi}}_s$ is a function of $\hat{\boldsymbol{\pi}}$ and since $\hat{\mathbf{B}}_0$ is a function of $\text{vech}(\hat{\mathbf{\Omega}})$, the distributions of both the autoregressive coefficients and the variances affect the asymptotic distribution of $\hat{\mathbf{h}}_s$. It follows from Proposition 11.2 that with Gaussian innovations,

$$\begin{aligned} &\sqrt{T}(\hat{\mathbf{h}}_{s,T} - \mathbf{h}_s) \\ &\xrightarrow{L} N\left(\mathbf{0}, [\mathbf{\Xi}_\pi \quad \mathbf{\Xi}_\sigma] \begin{bmatrix} \mathbf{\Omega} \otimes \mathbf{Q}^{-1} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{\Omega})(\mathbf{D}_n^+)' \end{bmatrix} \begin{bmatrix} \mathbf{\Xi}_\pi' \\ \mathbf{\Xi}_\sigma' \end{bmatrix}\right) \qquad [11.7.13] \\ &\sim N\left(\mathbf{0}, [\mathbf{\Xi}_\pi(\mathbf{\Omega} \otimes \mathbf{Q}^{-1})\mathbf{\Xi}_\pi' \dotplus 2\mathbf{\Xi}_\sigma \mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{\Omega})(\mathbf{D}_n^+)'\mathbf{\Xi}_\sigma']\right), \end{aligned}$$

where Appendix 11.B demonstrates that

$$\mathbf{\Xi}_\pi = \partial \mathbf{h}_s / \partial \boldsymbol{\pi}' = [\mathbf{I}_n \otimes (\mathbf{B}_0')^{-1}]\mathbf{G}_s \qquad [11.7.14]$$

$$\mathbf{\Xi}_\sigma = \frac{\partial \mathbf{h}_s}{\partial [\text{vech}(\mathbf{\Omega})]'} = -[\mathbf{H}_s \otimes (\mathbf{B}_0')^{-1}]\mathbf{S}_{B'}\mathbf{G}_B. \qquad [11.7.15]$$

Here $\mathbf{G}_s$ is the matrix given in [11.7.5], $\mathbf{G}_B$ is the matrix given in [11.7.11], and $\mathbf{S}_{B'}$ is an $(n^2 \times n_B)$ matrix that takes the elements of $\boldsymbol{\theta}_B$ and puts them in the corresponding position to construct $\text{vec}(\mathbf{B}_0')$:

$$\text{vec}(\mathbf{B}_0') = \mathbf{S}_{B'}\boldsymbol{\theta}_B + \mathbf{s}_{B'}.$$

For the supply-and-demand examples of [11.6.24] to [11.6.26],

$$\mathbf{S}_{B'} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Practical Experience with Standard Errors

In practice, the standard errors for dynamic inferences based on *VAR*s often turn out to be disappointingly large (see Runkle, 1987, and Lütkepohl, 1990).

Although a VAR imposes few restrictions on the dynamics, the cost of this generality is that the inferences drawn are not too precise. To gain more precision, it is necessary to impose further restrictions. One approach is to fit the multivariate dynamics using a restricted model with far fewer parameters, provided that the data allow us to accept the restrictions. A second approach is to place greater reliance on prior expectations about the system dynamics. This second approach is explored in the next chapter.

APPENDIX 11.A. *Proofs of Chapter 11 Propositions*

■ **Proof of Proposition 11.1.** The condition on the roots of [11.1.35] ensures that the $MA(\infty)$ representation is absolutely summable. Thus $\mathbf{y}_t$ is ergodic for first moments, from Propositions 10.2(b) and 10.5(a), and is also ergodic for second moments, from Proposition 10.2(d). This establishes result 11.1(a).

The proofs of results (b) and (c) are virtually identical to those for a single OLS regression with stochastic regressors (results [8.2.5] and [8.2.12]).

To verify result (d), notice that

$$\sqrt{T}(\hat{\boldsymbol{\pi}}_{i,T} - \boldsymbol{\pi}_i) = \left[(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{it}\right]$$

and so

$$\sqrt{T}(\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) = \begin{bmatrix} \mathbf{Q}_T^{-1}\,(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{1t} \\ \mathbf{Q}_T^{-1}\,(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{2t} \\ \vdots \\ \mathbf{Q}_T^{-1}\,(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{nt} \end{bmatrix}, \qquad [11.A.1]$$

where

$$\mathbf{Q}_T = \left[(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right].$$

Define $\boldsymbol{\xi}_t$ to be the following ($nk \times 1$) vector:

$$\boldsymbol{\xi}_t = \begin{bmatrix} \mathbf{x}_t\varepsilon_{1t} \\ \mathbf{x}_t\varepsilon_{2t} \\ \vdots \\ \mathbf{x}_t\varepsilon_{nt} \end{bmatrix}.$$

Notice that $\boldsymbol{\xi}_t$ is a martingale difference sequence with finite fourth moments and variance

$$\begin{aligned} E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t') &= \begin{bmatrix} E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{1t}^2) & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{1t}\varepsilon_{2t}) & \cdots & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{1t}\varepsilon_{nt}) \\ E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{2t}\varepsilon_{1t}) & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{2t}^2) & \cdots & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{2t}\varepsilon_{nt}) \\ \vdots & \vdots & \cdots & \vdots \\ E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{nt}\varepsilon_{1t}) & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{nt}\varepsilon_{2t}) & \cdots & E(\mathbf{x}_t\mathbf{x}_t')\cdot E(\varepsilon_{nt}^2) \end{bmatrix} \\ &= \begin{bmatrix} E(\varepsilon_{1t}^2) & E(\varepsilon_{1t}\varepsilon_{2t}) & \cdots & E(\varepsilon_{1t}\varepsilon_{nt}) \\ E(\varepsilon_{2t}\varepsilon_{1t}) & E(\varepsilon_{2t}^2) & \cdots & E(\varepsilon_{2t}\varepsilon_{nt}) \\ \vdots & \vdots & \cdots & \vdots \\ E(\varepsilon_{nt}\varepsilon_{1t}) & E(\varepsilon_{nt}\varepsilon_{2t}) & \cdots & E(\varepsilon_{nt}^2) \end{bmatrix} \otimes E(\mathbf{x}_t\mathbf{x}_t') \\ &= \boldsymbol{\Omega} \otimes \mathbf{Q}. \end{aligned}$$

It can further be shown that

$$(1/T)\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t' \xrightarrow{p} \boldsymbol{\Omega} \otimes \mathbf{Q} \qquad [11.A.2]$$

(see Exercise 11.1). It follows from Proposition 7.9 that

$$(1/\sqrt{T})\sum_{t=1}^{T}\xi_t \xrightarrow{L} N\Big(\mathbf{0}, (\mathbf{\Omega}\otimes\mathbf{Q})\Big). \qquad [11.A.3]$$

Now, expression [11.A.1] can be written

$$\sqrt{T}(\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) = \begin{bmatrix} \mathbf{Q}_T^{-1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_T^{-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_T^{-1} \end{bmatrix} \begin{bmatrix} (1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{1t} \\ (1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{2t} \\ \vdots \\ (1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_{nt} \end{bmatrix}$$

$$= (\mathbf{I}_n\otimes\mathbf{Q}_T^{-1})(1/\sqrt{T})\sum_{t=1}^{T}\xi_t.$$

But result (a) implies that $\mathbf{Q}_T^{-1} \xrightarrow{p} \mathbf{Q}^{-1}$. Thus,

$$\sqrt{T}(\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}) \xrightarrow{p} (\mathbf{I}_n\otimes\mathbf{Q}^{-1})(1/\sqrt{T})\sum_{t=1}^{T}\xi_t. \qquad [11.A.4]$$

But from [11.A.3], this has a distribution that is Gaussian with mean $\mathbf{0}$ and variance

$$(\mathbf{I}_n\otimes\mathbf{Q}^{-1})(\mathbf{\Omega}\otimes\mathbf{Q})(\mathbf{I}_n\otimes\mathbf{Q}^{-1}) = (\mathbf{I}_n\mathbf{\Omega}\mathbf{I}_n)\otimes(\mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}) = \mathbf{\Omega}\otimes\mathbf{Q}^{-1},$$

as claimed. ■

■ **Proof of Proposition 11.2.** Define $\hat{\mathbf{\Omega}}_T^* \equiv (1/T)\Sigma_{t=1}^{T}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'$ to be the estimate of $\mathbf{\Omega}$ based on the true residuals. We first note that $\hat{\mathbf{\Omega}}_T$ has the same asymptotic distribution as $\hat{\mathbf{\Omega}}_T^*$. To see this, observe that

$$\begin{aligned}
\hat{\mathbf{\Omega}}_T^* &= (1/T)\sum_{t=1}^{T}(\mathbf{y}_t - \mathbf{\Pi}'\mathbf{x}_t)(\mathbf{y}_t - \mathbf{\Pi}'\mathbf{x}_t)' \\
&= (1/T)\sum_{t=1}^{T}[\mathbf{y}_t - \hat{\mathbf{\Pi}}_T'\mathbf{x}_t + (\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})'\mathbf{x}_t][\mathbf{y}_t - \hat{\mathbf{\Pi}}_T'\mathbf{x}_t + (\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})'\mathbf{x}_t]' \\
&= (1/T)\sum_{t=1}^{T}(\mathbf{y}_t - \hat{\mathbf{\Pi}}_T'\mathbf{x}_t)(\mathbf{y}_t - \hat{\mathbf{\Pi}}_T'\mathbf{x}_t)' \qquad [11.A.5] \\
&\quad + (\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})'(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'(\hat{\mathbf{\Pi}}_T - \mathbf{\Pi}) \\
&= \hat{\mathbf{\Omega}}_T + (\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})'(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'(\hat{\mathbf{\Pi}}_T - \mathbf{\Pi}),
\end{aligned}$$

where cross-product terms were dropped in the third equality on the right in the light of the *OLS* orthogonality condition $(1/T)\Sigma_{t=1}^{T}(\mathbf{y}_t - \hat{\mathbf{\Pi}}_T'\mathbf{x}_t)\mathbf{x}_t' = \mathbf{0}$. Equation [11.A.5] implies that

$$\sqrt{T}(\hat{\mathbf{\Omega}}_T^* - \hat{\mathbf{\Omega}}_T) = (\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})'(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'[\sqrt{T}(\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})].$$

But Proposition 11.1 established that $(\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})' \xrightarrow{p} \mathbf{0}$, $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{Q}$, and $\sqrt{T}(\hat{\mathbf{\Pi}}_T - \mathbf{\Pi})$ converges in distribution. Thus, from Proposition 7.3, $\sqrt{T}(\hat{\mathbf{\Omega}}_T^* - \hat{\mathbf{\Omega}}_T) \xrightarrow{p} \mathbf{0}$ meaning that $\sqrt{T}(\hat{\mathbf{\Omega}}_T^* - \mathbf{\Omega}) \xrightarrow{p} \sqrt{T}(\hat{\mathbf{\Omega}}_T - \mathbf{\Omega})$.

Recalling [11.A.4],

$$\begin{bmatrix} \sqrt{T}[\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}] \\ \sqrt{T}[\text{vech}(\hat{\mathbf{\Omega}}_T) - \text{vech}(\mathbf{\Omega})] \end{bmatrix} \xrightarrow{p} \begin{bmatrix} (\mathbf{I}_n\otimes\mathbf{Q}^{-1})(1/\sqrt{T})\sum_{t=1}^{T}\xi_t \\ (1/\sqrt{T})\sum_{t=1}^{T}\boldsymbol{\lambda}_t \end{bmatrix}, \qquad [11.A.6]$$

where $\boldsymbol{\xi}_t = \boldsymbol{\varepsilon}_t \otimes \mathbf{x}_t$ and

$$\boldsymbol{\lambda}_t \equiv \text{vech}\begin{bmatrix} \varepsilon_{1t}^2 - \sigma_{11} & \varepsilon_{1t}\varepsilon_{2t} - \sigma_{12} & \cdots & \varepsilon_{1t}\varepsilon_{nt} - \sigma_{1n} \\ \varepsilon_{2t}\varepsilon_{1t} - \sigma_{21} & \varepsilon_{2t}^2 - \sigma_{22} & \cdots & \varepsilon_{2t}\varepsilon_{nt} - \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \varepsilon_{nt}\varepsilon_{1t} - \sigma_{n1} & \varepsilon_{nt}\varepsilon_{2t} - \sigma_{n2} & \cdots & \varepsilon_{nt}^2 - \sigma_{nn} \end{bmatrix}.$$

It is straightforward to show that $(\boldsymbol{\xi}_t', \boldsymbol{\lambda}_t')'$ is a martingale difference sequence that satisfies the conditions of Proposition 7.9, from which

$$\begin{bmatrix} (1/\sqrt{T})\sum_{t=1}^{T} \boldsymbol{\xi}_t \\ (1/\sqrt{T})\sum_{t=1}^{T} \boldsymbol{\lambda}_t \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right), \qquad [11.A.7]$$

where

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t') & E(\boldsymbol{\xi}_t\boldsymbol{\lambda}_t') \\ E(\boldsymbol{\lambda}_t\boldsymbol{\xi}_t') & E(\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t') \end{bmatrix}.$$

Recall from the proof of Proposition 11.1 that

$$\boldsymbol{\Sigma}_{11} = E(\boldsymbol{\xi}_t\boldsymbol{\xi}_t') = \boldsymbol{\Omega} \otimes \mathbf{Q}.$$

A typical element of $\boldsymbol{\Sigma}_{12}$ is of the form

$$E(\mathbf{x}_t\varepsilon_{it})(\varepsilon_{it}\varepsilon_{jt} - \sigma_{ij}) = E(\mathbf{x}_t)\cdot E(\varepsilon_{it}\varepsilon_{it}\varepsilon_{jt}) - \sigma_{ij}\cdot E(\mathbf{x}_t)\cdot E(\varepsilon_{it}),$$

which equals zero for all i, j, and l. Hence, [11.A.7] becomes

$$\begin{bmatrix} (1/\sqrt{T})\sum_{t=1}^{T} \boldsymbol{\xi}_t \\ (1/\sqrt{T})\sum_{t=1}^{T} \boldsymbol{\lambda}_t \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega} \otimes \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right),$$

and so, from [11.A.6],

$$\begin{bmatrix} \sqrt{T}[\hat{\boldsymbol{\pi}}_T - \boldsymbol{\pi}] \\ \sqrt{T}[\text{vech}(\hat{\boldsymbol{\Omega}}_T) - \text{vech}(\boldsymbol{\Omega})] \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Omega} \otimes \mathbf{Q}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right).$$

Hence, Proposition 11.2 will be established if we can show that $E(\boldsymbol{\lambda}_t\boldsymbol{\lambda}_t')$ is given by the matrix $\boldsymbol{\Sigma}_{22}$ described in the proposition; that is, we must show that

$$E(\varepsilon_{it}\varepsilon_{jt} - \sigma_{ij})(\varepsilon_{lt}\varepsilon_{mt} - \sigma_{lm}) = \sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl} \qquad [11.A.8]$$

for all i, j, l, and m.

To derive [11.A.8], let $\boldsymbol{\Omega} = \mathbf{PP}'$ denote the Cholesky decomposition of $\boldsymbol{\Omega}$, and define

$$\mathbf{v}_t \equiv \mathbf{P}^{-1}\boldsymbol{\varepsilon}_t. \qquad [11.A.9]$$

Then $E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{P}^{-1}\boldsymbol{\Omega}(\mathbf{P}^{-1})' = \mathbf{I}_n$. Thus, v_{it} is Gaussian with zero mean, unit variance, and fourth moment given by $E(v_{it}^4) = 3$. Moreover, v_{it} is independent of v_{jt} for $i \neq j$.

Equation [11.A.9] implies

$$\boldsymbol{\varepsilon}_t = \mathbf{P}\mathbf{v}_t. \qquad [11.A.10]$$

Let p_{ij} denote the row i, column j element of $\mathbf{P}$. Then the ith row of [11.A.10] states that

$$\varepsilon_{it} = p_{i1}v_{1t} + p_{i2}v_{2t} + \cdots + p_{in}v_{nt} \qquad [11.A.11]$$

and

$$\varepsilon_{it}\varepsilon_{jt} = (p_{i1}v_{1t} + p_{i2}v_{2t} + \cdots + p_{in}v_{nt}) \times (p_{j1}v_{1t} + p_{j2}v_{2t} + \cdots + p_{jn}v_{nt}). \qquad [11.A.12]$$

Second moments of $\boldsymbol{\varepsilon}_t$ can be found by taking expectations of [11.A.12], recalling that $E(v_{it}v_{jt}) = 1$ if $i = j$ and is zero otherwise:

$$E(\varepsilon_{it}\varepsilon_{jt}) = p_{i1}p_{j1} + p_{i2}p_{j2} + \cdots + p_{in}p_{jn}. \qquad [11.A.13]$$

Similarly, fourth moments can be found from

$$
\begin{aligned}
E(\varepsilon_{it}\varepsilon_{jt}\varepsilon_{lt}\varepsilon_{mt}) &= E[(p_{i1}v_{1t} + p_{i2}v_{2t} + \cdots + p_{in}v_{nt})(p_{j1}v_{1t} + p_{j2}v_{2t} + \cdots + p_{jn}v_{nt}) \\
&\quad \times (p_{l1}v_{1t} + p_{l2}v_{2t} + \cdots + p_{ln}v_{nt})(p_{m1}v_{1t} + p_{m2}v_{2t} + \cdots + p_{mn}v_{nt})] \\
&= [3(p_{i1}p_{j1}p_{l1}p_{m1} + p_{i2}p_{j2}p_{l2}p_{m2} + \cdots + p_{in}p_{jn}p_{ln}p_{mn})] \\
&\quad + [(p_{i1}p_{j1})(p_{l2}p_{m2} + p_{l3}p_{m3} + \cdots + p_{ln}p_{mn}) \\
&\qquad + (p_{i2}p_{j2})(p_{l1}p_{m1} + p_{l3}p_{m3} + \cdots + p_{ln}p_{mn}) + \cdots \\
&\qquad + (p_{in}p_{jn})(p_{l1}p_{m1} + p_{l2}p_{m2} + \cdots + p_{l,n-1}p_{m,n-1})] \\
&\quad + [(p_{i1}p_{l1})(p_{j2}p_{m2} + p_{j3}p_{m3} + \cdots + p_{jn}p_{mn}) \\
&\qquad + (p_{i2}p_{l2})(p_{j1}p_{m1} + p_{j3}p_{m3} + \cdots + p_{jn}p_{mn}) + \cdots \\
&\qquad + (p_{in}p_{ln})(p_{j1}p_{m1} + p_{j2}p_{m2} + \cdots + p_{j,n-1}p_{m,n-1})] \\
&\quad + [(p_{i1}p_{m1})(p_{j2}p_{l2} + p_{j3}p_{l3} + \cdots + p_{jn}p_{ln}) \\
&\qquad + (p_{i2}p_{m2})(p_{j1}p_{l1} + p_{j3}p_{l3} + \cdots + p_{jn}p_{ln}) + \cdots \\
&\qquad + (p_{in}p_{mn})(p_{j1}p_{l1} + p_{j2}p_{l2} + \cdots + p_{j,n-1}p_{l,n-1})] \\
&= [(p_{i1}p_{j1} + p_{i2}p_{j2} + \cdots + p_{in}p_{jn})(p_{l1}p_{m1} + p_{l2}p_{m2} + \cdots + p_{ln}p_{mn})] \\
&\quad + [(p_{i1}p_{l1} + p_{i2}p_{l2} + \cdots + p_{in}p_{ln})(p_{j1}p_{m1} + p_{j2}p_{m2} + \cdots + p_{jn}p_{mn})] \\
&\quad + [(p_{i1}p_{m1} + p_{i2}p_{m2} + \cdots + p_{in}p_{mn})(p_{j1}p_{l1} + p_{j2}p_{l2} + \cdots + p_{jn}p_{ln})] \\
&= \sigma_{ij}\sigma_{lm} + \sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl},
\end{aligned}
\tag{11.A.14}
$$

where the last line follows from [11.A.13]. Then

$$E[(\varepsilon_{it}\varepsilon_{jt} - \sigma_{ij})(\varepsilon_{lt}\varepsilon_{mt} - \sigma_{lm})] = E(\varepsilon_{it}\varepsilon_{jt}\varepsilon_{lt}\varepsilon_{mt}) - \sigma_{ij}\sigma_{lm} = \sigma_{il}\sigma_{jm} + \sigma_{im}\sigma_{jl},$$

as claimed in [11.A.8]. ■

■ **Proof of Proposition 11.3.** First suppose that y fails to Granger-cause x, so that the process can be written as in [11.2.4]. Define v_{2t} to be the residual from a projection of ε_{2t} on ε_{1t}, with b_0 defined to be the projection coefficient:

$$v_{2t} \equiv \varepsilon_{2t} - b_0\varepsilon_{1t}.$$

Thus, v_{2t} and ε_{1t} are uncorrelated, and, recalling that $\boldsymbol{\varepsilon}_t$ is white noise, v_{2t} must be uncorrelated with $\varepsilon_{1\tau}$ for all $t \neq \tau$ as well. From the first row of [11.2.4], this means that v_{2t} and x_τ are uncorrelated for all t and τ. With this definition of v_{2t}, the second row of [11.2.4] can be written as

$$y_t = \mu_2 + \psi_{21}(L)\varepsilon_{1t} + \psi_{22}(L)[v_{2t} + b_0\varepsilon_{1t}]. \tag{11.A.15}$$

Furthermore, from the first row of [11.2.4],

$$\varepsilon_{1t} = [\psi_{11}(L)]^{-1}(x_t - \mu_1). \tag{11.A.16}$$

Substituting [11.A.16] into [11.A.15] gives

$$y_t = c + b(L)x_t + \eta_t, \tag{11.A.17}$$

where we have defined $b(L) \equiv \{[\psi_{21}(L) + b_0\psi_{22}(L)][\psi_{11}(L)]^{-1}\}$, $c \equiv \mu_2 - b(1)\mu_1$, and $\eta_t \equiv \psi_{22}(L)v_{2t}$. But η_t, being constructed from v_{2t}, is uncorrelated with x_τ for all τ. Furthermore, only current and lagged values of x, as summarized by the operator $b(L)$, appear in equation [11.A.17]. We have thus shown that if [11.2.4] holds, then $d_j = 0$ for all j in [11.2.5].

To prove the converse, suppose that $d_j = 0$ for all j in [11.2.5]. Let

$$x_t = \mu_1 + \psi_{11}(L)\varepsilon_{1t}, \tag{11.A.18}$$

denote the univariate Wold representation for x_t; thus, $\psi_{11}^{(0)} = 1$. We will be using notation consistent with the form of [11.2.4] in anticipation of the final answer that will be derived; for now, the reader should view [11.A.18] as a new definition of $\psi_{11}(L)$ in terms of the

univariate Wold representation for x. There also exists a univariate Wold representation for the error term in [11.2.5], denoted

$$\eta_t = \psi_{22}(L)v_{2t}, \qquad [11.A.19]$$

with $\psi_{22}^{(0)} = 1$. Notice that η_t as defined in [11.2.5] is uncorrected with x_s for all t and s. It follows that v_{2t} is uncorrelated with x_τ or $\varepsilon_{1\tau}$ for all t and τ.

Substituting [11.A.18] and [11.A.19] into [11.2.5],

$$y_t = c + b(1)\mu_1 + b(L)\psi_{11}(L)\varepsilon_{1t} + \psi_{22}(L)v_{2t}. \qquad [11.A.20]$$

Define

$$\varepsilon_{2t} \equiv v_{2t} + b_0\varepsilon_{1t} \qquad [11.A.21]$$

for b_0 the coefficient on L^0 of $b(L)$ and

$$\mu_2 \equiv c + b(1)\mu_1. \qquad [11.A.22]$$

Observe that $(\varepsilon_{1t}, \varepsilon_{2t})'$ is vector white noise. Substituting [11.A.21] and [11.A.22] into [11.A.20] produces

$$y_t = \mu_2 + [b(L)\psi_{11}(L) - b_0\psi_{22}(L)]\varepsilon_{1t} + \psi_{22}(L)\varepsilon_{2t}. \qquad [11.A.23]$$

Finally, define

$$\psi_{21}(L) \equiv [b(L)\psi_{11}(L) - b_0\psi_{22}(L)],$$

noting that $\psi_{21}^{(0)} = 0$. Then, substituting this into [11.A.23] produces

$$y_t = \mu_2 + \psi_{21}(L)\varepsilon_{1t} + \psi_{22}(L)\varepsilon_{2t}.$$

This combined with [11.A.18] completes the demonstration that [11.2.5] implies [11.2.4]. ■

APPENDIX 11.B. *Calculation of Analytic Derivatives*

This appendix calculates the derivatives reported in Sections 11.6 and 11.7.

■ **Derivation of [11.6.38].** Let the scalar ξ represent some particular element of $\boldsymbol{\theta}_B$ or $\boldsymbol{\theta}_D$, and let $\partial\boldsymbol{\Omega}/\partial\xi$ denote the $(n^2 \times n^2)$ matrix that results when each element of $\boldsymbol{\Omega}$ is differentiated with respect to ξ. Thus, differentiating [11.6.33] with respect to ξ results in

$$\partial\boldsymbol{\Omega}/\partial\xi = (\partial\mathbf{B}_0^{-1}/\partial\xi)\mathbf{D}(\mathbf{B}_0^{-1})' + \mathbf{B}_0^{-1}(\partial\mathbf{D}/\partial\xi)(\mathbf{B}_0^{-1})' + (\mathbf{B}_0^{-1})\mathbf{D}[\partial(\mathbf{B}_0^{-1})'/\partial\xi]. \qquad [11.B.1]$$

Define

$$\boldsymbol{\chi} \equiv (\partial\mathbf{B}_0^{-1}/\partial\xi)\mathbf{D}(\mathbf{B}_0^{-1})' \qquad [11.B.2]$$

and notice that

$$\boldsymbol{\chi}' = (\mathbf{B}_0^{-1})\mathbf{D}[\partial(\mathbf{B}_0^{-1})'/\partial\xi],$$

since $\mathbf{D}$ is a variance-covariance matrix and must therefore be symmetric. Thus, [11.B.1] can be written

$$\partial\boldsymbol{\Omega}/\partial\xi = \boldsymbol{\chi} + \mathbf{B}_0^{-1}(\partial\mathbf{D}/\partial\xi)(\mathbf{B}_0^{-1})' + \boldsymbol{\chi}'. \qquad [11.B.3]$$

Recall from Proposition 10.4 that

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\cdot\text{vec}(\mathbf{B}). \qquad [11.B.4]$$

Thus, if the vec operator is applied to [11.B.3], the result is

$$\frac{\partial\,\text{vec}(\boldsymbol{\Omega})}{\partial\xi} = \text{vec}(\boldsymbol{\chi} + \boldsymbol{\chi}') + [(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\,\text{vec}(\partial\mathbf{D}/\partial\xi). \qquad [11.B.5]$$

Let $\mathbf{D}_n$ denote the $(n^2 \times n^*)$ duplication matrix introduced in [11.1.43]. Notice that for any $(n \times n)$ matrix χ, the elements of $\mathbf{D}_n'$ vec(χ) are of the form χ_{ii} for diagonal elements of χ and of the form $(\chi_{ij} + \chi_{ji})$ for off-diagonal elements. Hence, $\mathbf{D}_n'$ vec(χ) = $\mathbf{D}_n'$ vec(χ'). If [11.B.5] is premultiplied by $\mathbf{D}_n^+ = (\mathbf{D}_n'\mathbf{D}_n)^{-1}\mathbf{D}_n'$, the result is thus

$$\frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \xi} = 2\mathbf{D}_n^+ \,\mathrm{vec}(\chi) + \mathbf{D}_n^+[(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\,\mathrm{vec}(\partial \mathbf{D}/\partial \xi), \qquad [11.B.6]$$

since from [11.1.46] $\mathbf{D}_n^+$ vec($\mathbf{\Omega}$) = vech($\mathbf{\Omega}$).

Differentiating the identity $\mathbf{B}_0^{-1}\mathbf{B}_0 = \mathbf{I}_n$ with respect to ξ produces

$$(\partial \mathbf{B}_0^{-1}/\partial \xi)\mathbf{B}_0 + \mathbf{B}_0^{-1}(\partial \mathbf{B}_0/\partial \xi) = \mathbf{0}_{nn}$$

or

$$\partial \mathbf{B}_0^{-1}/\partial \xi = -\mathbf{B}_0^{-1}(\partial \mathbf{B}_0/\partial \xi)\mathbf{B}_0^{-1}. \qquad [11.B.7]$$

Thus, [11.B.2] can be written

$$\chi = -\mathbf{B}_0^{-1}(\partial \mathbf{B}_0/\partial \xi)\mathbf{B}_0^{-1}\mathbf{D}(\mathbf{B}_0^{-1})' = -\mathbf{B}_0^{-1}(\partial \mathbf{B}_0/\partial \xi)\mathbf{\Omega}.$$

Applying the vec operator as in [11.B.4] results in

$$\mathrm{vec}(\chi) = -(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\frac{\partial\, \mathrm{vec}(\mathbf{B}_0)}{\partial \xi}.$$

Substituting this expression into [11.B.6] gives

$$\begin{aligned}\frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \xi} &= -2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\frac{\partial\, \mathrm{vec}(\mathbf{B}_0)}{\partial \xi} + \mathbf{D}_n^+[(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\frac{\partial\, \mathrm{vec}(\mathbf{D})}{\partial \xi} \\ &= -2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\mathbf{S}_B\frac{\partial \boldsymbol{\theta}_B}{\partial \xi} + \mathbf{D}_n^+[(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\mathbf{S}_D\frac{\partial \boldsymbol{\theta}_D}{\partial \xi}.\end{aligned} \qquad [11.B.8]$$

Expression [11.B.8] is an $(n^* \times 1)$ vector that gives the effect of a change in some element of $\boldsymbol{\theta}_B$ or $\boldsymbol{\theta}_D$ on each of the n^* elements of vech($\mathbf{\Omega}$). If ξ corresponds to the first element of $\boldsymbol{\theta}_B$, then $\partial \boldsymbol{\theta}_B/\partial \xi = \mathbf{e}_1$, the first column of the $(n_B \times n_B)$ identity matrix, and $\partial \boldsymbol{\theta}_D/\partial \xi = \mathbf{0}$. If ξ corresponds to the second element of $\boldsymbol{\theta}_B$, then $\partial \boldsymbol{\theta}_B/\partial \xi = \mathbf{e}_2$. If we stack the vectors in [11.B.8] associated with $\xi = \theta_{B,1}, \xi = \theta_{B,2}, \ldots, \xi = \theta_{B,n_B}$ side by side, the result is

$$\begin{aligned}&\left[\frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \theta_{B,1}} \quad \frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \theta_{B,2}} \quad \cdots \quad \frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \theta_{B,n_B}}\right] \\ &\qquad = [-2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\mathbf{S}_B][\mathbf{e}_1 \quad \mathbf{e}_2 \quad \cdots \quad \mathbf{e}_{n_B}].\end{aligned} \qquad [11.B.9]$$

That is,

$$\frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \boldsymbol{\theta}_B'} = [-2\mathbf{D}_n^+(\mathbf{\Omega} \otimes \mathbf{B}_0^{-1})\mathbf{S}_B]. \qquad [11.B.10]$$

Similarly, letting the scalar ξ in [11.B.8] correspond to each of the elements of $\boldsymbol{\theta}_D$ in succession and stacking the resulting columns horizontally results in

$$\frac{\partial\, \mathrm{vech}(\mathbf{\Omega})}{\partial \boldsymbol{\theta}_D'} = \mathbf{D}_n^+[(\mathbf{B}_0^{-1}) \otimes (\mathbf{B}_0^{-1})]\mathbf{S}_D. \qquad [11.B.11]$$

Equation [11.6.38] then follows immediately from [11.B.10] and [11.B.11]. ■

■ **Derivation of [11.7.4].** Recall from equation [10.1.19] that

$$\mathbf{\Psi}_s = \mathbf{\Phi}_1\mathbf{\Psi}_{s-1} + \mathbf{\Phi}_2\mathbf{\Psi}_{s-2} + \cdots + \mathbf{\Phi}_p\mathbf{\Psi}_{s-p}. \qquad [11.B.12]$$

Taking transposes,

$$\mathbf{\Psi}_s' = \mathbf{\Psi}_{s-1}'\mathbf{\Phi}_1' + \mathbf{\Psi}_{s-2}'\mathbf{\Phi}_2' + \cdots + \mathbf{\Psi}_{s-p}'\mathbf{\Phi}_p'. \qquad [11.B.13]$$

Let the scalar ξ denote some particular element of $\boldsymbol{\pi}$, and differentiate [11.B.13] with respect to ξ:

$$\begin{aligned}
\frac{\partial \boldsymbol{\Psi}'_s}{\partial \xi} &= \boldsymbol{\Psi}'_{s-1}\frac{\partial \boldsymbol{\Phi}'_1}{\partial \xi} + \boldsymbol{\Psi}'_{s-2}\frac{\partial \boldsymbol{\Phi}'_2}{\partial \xi} + \cdots + \boldsymbol{\Psi}'_{s-p}\frac{\partial \boldsymbol{\Phi}'_p}{\partial \xi} \\
&\quad + \frac{\partial \boldsymbol{\Psi}'_{s-1}}{\partial \xi}\boldsymbol{\Phi}'_1 + \frac{\partial \boldsymbol{\Psi}'_{s-2}}{\partial \xi}\boldsymbol{\Phi}'_2 + \cdots + \frac{\partial \boldsymbol{\Psi}'_{s-p}}{\partial \xi}\boldsymbol{\Phi}'_p \\
&= [\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p}]\begin{bmatrix} \partial \mathbf{c}'/\partial \xi \\ \partial \boldsymbol{\Phi}'_1/\partial \xi \\ \partial \boldsymbol{\Phi}'_2/\partial \xi \\ \vdots \\ \partial \boldsymbol{\Phi}'_p/\partial \xi \end{bmatrix} \qquad [11.B.14] \\
&\quad + \frac{\partial \boldsymbol{\Psi}'_{s-1}}{\partial \xi}\boldsymbol{\Phi}'_1 + \frac{\partial \boldsymbol{\Psi}'_{s-2}}{\partial \xi}\boldsymbol{\Phi}'_2 + \cdots + \frac{\partial \boldsymbol{\Psi}'_{s-p}}{\partial \xi}\boldsymbol{\Phi}'_p \\
&= [\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p}]\frac{\partial \boldsymbol{\Pi}}{\partial \xi} \\
&\quad + \frac{\partial \boldsymbol{\Psi}'_{s-1}}{\partial \xi}\boldsymbol{\Phi}'_1 + \frac{\partial \boldsymbol{\Psi}'_{s-2}}{\partial \xi}\boldsymbol{\Phi}'_2 + \cdots + \frac{\partial \boldsymbol{\Psi}'_{s-p}}{\partial \xi}\boldsymbol{\Phi}'_p.
\end{aligned}$$

Recall result [11.B.4], and note the special case when $\mathbf{A}$ is the $(n \times n)$ identity matrix, $\mathbf{B}$ is an $(n \times r)$ matrix, and $\mathbf{C}$ is an $(r \times q)$ matrix:

$$\text{vec}(\mathbf{BC}) = (\mathbf{C}' \otimes \mathbf{I}_n)\,\text{vec}(\mathbf{B}). \qquad [11.B.15]$$

For example,

$$\text{vec}\left(\frac{\partial \boldsymbol{\Psi}'_{s-1}}{\partial \xi}\boldsymbol{\Phi}'_1\right) = (\boldsymbol{\Phi}_1 \otimes \mathbf{I}_n)\,\text{vec}\left(\frac{\partial \boldsymbol{\Psi}'_{s-1}}{\partial \xi}\right) = (\boldsymbol{\Phi}_1 \otimes \mathbf{I}_n)\left(\frac{\partial \boldsymbol{\psi}_{s-1}}{\partial \xi}\right). \qquad [11.B.16]$$

Another implication of [11.B.4] can be obtained by letting $\mathbf{A}$ be an $(m \times q)$ matrix, $\mathbf{B}$ a $(q \times n)$ matrix, and $\mathbf{C}$ the $(n \times n)$ identity matrix:

$$\text{vec}(\mathbf{AB}) = (\mathbf{I}_n \otimes \mathbf{A})\,\text{vec}(\mathbf{B}). \qquad [11.B.17]$$

For example,

$$\begin{aligned}
&\text{vec}\left([\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p}]\frac{\partial \boldsymbol{\Pi}}{\partial \xi}\right) \\
&= [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})]\left(\frac{\partial\,\text{vec}(\boldsymbol{\Pi})}{\partial \xi}\right) \qquad [11.B.18] \\
&= [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})]\left(\frac{\partial \boldsymbol{\pi}}{\partial \xi}\right).
\end{aligned}$$

Applying the vec operator to [11.B.14] and using [11.B.18] and [11.B.16] gives

$$\begin{aligned}
\frac{\partial \boldsymbol{\psi}_s}{\partial \xi} &= [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})]\left(\frac{\partial \boldsymbol{\pi}}{\partial \xi}\right) \\
&\quad + (\boldsymbol{\Phi}_1 \otimes \mathbf{I}_n)\left(\frac{\partial \boldsymbol{\psi}_{s-1}}{\partial \xi}\right) + (\boldsymbol{\Phi}_2 \otimes \mathbf{I}_n)\left(\frac{\partial \boldsymbol{\psi}_{s-2}}{\partial \xi}\right) \qquad [11.B.19] \\
&\quad + \cdots + (\boldsymbol{\Phi}_p \otimes \mathbf{I}_n)\left(\frac{\partial \boldsymbol{\psi}_{s-p}}{\partial \xi}\right).
\end{aligned}$$

Letting ξ successively represent each of the elements of $\boldsymbol{\pi}$ and stacking the resulting

equations horizontally as in [11.B.9] results in

$$\frac{\partial \boldsymbol{\psi}_s}{\partial \boldsymbol{\pi}'} = [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})]$$
$$+ (\boldsymbol{\Phi}_1 \otimes \mathbf{I}_n)\left[\frac{\partial \boldsymbol{\psi}_{s-1}}{\partial \boldsymbol{\pi}'}\right] + \cdots + (\boldsymbol{\Phi}_p \otimes \mathbf{I}_n)\left[\frac{\partial \boldsymbol{\psi}_{s-p}}{\partial \boldsymbol{\pi}'}\right],$$

as claimed in [11.7.4]. ■

■ **Derivation of [11.7.5].** Here the task is to verify that if $\mathbf{G}_s$ is given by [11.7.5], then [11.7.4] holds:

$$\mathbf{G}_s = [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})] + \sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n)\mathbf{G}_{s-k}. \qquad [11.B.20]$$

Notice that for $\mathbf{G}_s$ given by [11.7.5],

$$\sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n)\mathbf{G}_{s-k}$$
$$= \sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n) \sum_{i=1}^{s-k} [\boldsymbol{\Psi}_{i-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-k-i} \quad \boldsymbol{\Psi}'_{s-k-i-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-k-i-p+1})]$$
$$= \sum_{k=1}^{p} \sum_{i=1}^{s-k} [\boldsymbol{\Phi}_k \boldsymbol{\Psi}_{i-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-k-i} \quad \boldsymbol{\Psi}'_{s-k-i-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-k-i-p+1})].$$

For any given value for k and i, define $v \equiv k + i$. When $i = 1$, then $v = k + 1$; when $i = 2$, then $v = k + 2$; and so on:

$$\sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n)\mathbf{G}_{s-k} = \sum_{k=1}^{p} \sum_{v=k+1}^{s} [\boldsymbol{\Phi}_k \boldsymbol{\Psi}_{v-k-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})].$$

Recalling further that $\boldsymbol{\Psi}_{v-k-1} = \mathbf{0}$ for $v = 2, 3, \ldots, k$, we could equally well write

$$\sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n)\mathbf{G}_{s-k}$$
$$= \sum_{k=1}^{p} \sum_{v=2}^{s} [\boldsymbol{\Phi}_k \boldsymbol{\Psi}_{v-k-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})]$$
$$= \sum_{v=2}^{s} \sum_{k=1}^{p} [\boldsymbol{\Phi}_k \boldsymbol{\Psi}_{v-k-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})] \qquad [11.B.21]$$
$$= \sum_{v=2}^{s} \left[\left(\sum_{k=1}^{p} \boldsymbol{\Phi}_k \boldsymbol{\Psi}_{v-k-1}\right) \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})\right]$$
$$= \sum_{v=2}^{s} [\boldsymbol{\Psi}_{v-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})],$$

by virtue of [11.B.12]. If the first term on the right side of [11.B.20] is added to [11.B.21], the result is

$$[\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})] + \sum_{k=1}^{p} (\boldsymbol{\Phi}_k \otimes \mathbf{I}_n)\mathbf{G}_{s-k}$$
$$= [\mathbf{I}_n \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-1} \quad \boldsymbol{\Psi}'_{s-2} \quad \cdots \quad \boldsymbol{\Psi}'_{s-p})$$
$$+ \sum_{v=2}^{s} [\boldsymbol{\Psi}_{v-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})]$$
$$= \sum_{v=1}^{s} [\boldsymbol{\Psi}_{v-1} \otimes (\mathbf{0}_{n1} \quad \boldsymbol{\Psi}'_{s-v} \quad \boldsymbol{\Psi}'_{s-v-1} \quad \cdots \quad \boldsymbol{\Psi}'_{s-v-p+1})],$$

which is indeed the expression for $\mathbf{G}_s$ given in [11.7.5]. ■

■ **Derivation of [11.7.14] and [11.7.15].** Postmultiplying [11.7.12] by $\mathbf{B}_0$ and transposing results in

$$\mathbf{B}'_0\mathbf{H}'_s = \boldsymbol{\Psi}'_s. \qquad [11.B.22]$$

Let the scalar ξ denote some element of $\boldsymbol{\pi}$ or $\boldsymbol{\Omega}$, and differentiate [11.B.22] with respect to ξ:

$$(\partial \mathbf{B}'_0/\partial \xi)\mathbf{H}'_s + \mathbf{B}'_0(\partial \mathbf{H}'_s/\partial \xi) = \partial \boldsymbol{\Psi}'_s/\partial \xi. \qquad [11.B.23]$$

Applying the vec operator to [11.B.23] and using [11.B.15] and [11.B.17],

$$(\mathbf{H}_s \otimes \mathbf{I}_n)(\partial \operatorname{vec}(\mathbf{B}_0')/\partial\xi) + (\mathbf{I}_n \otimes \mathbf{B}_0')(\partial \operatorname{vec}(\mathbf{H}_s')/\partial\xi) = \partial \operatorname{vec}(\mathbf{\Psi}_s')/\partial\xi,$$

implying that

$$\begin{aligned}\partial\mathbf{h}_s/\partial\xi &= -(\mathbf{I}_n \otimes \mathbf{B}_0')^{-1}(\mathbf{H}_s \otimes \mathbf{I}_n)(\partial \operatorname{vec}(\mathbf{B}_0')/\partial\xi) + (\mathbf{I}_n \otimes \mathbf{B}_0')^{-1}\,\partial\boldsymbol{\psi}_s/\partial\xi \\ &= -[\mathbf{H}_s \otimes (\mathbf{B}_0')^{-1}](\partial \operatorname{vec}(\mathbf{B}_0')/\partial\xi) + [\mathbf{I}_n \otimes (\mathbf{B}_0')^{-1}]\,\partial\boldsymbol{\psi}_s/\partial\xi.\end{aligned} \qquad [11.B.24]$$

Noticing that $\mathbf{B}_0'$ does not depend on $\boldsymbol{\pi}$, if [11.B.24] is stacked horizontally for $\xi = \pi_1, \pi_2, \ldots, \pi_{nk}$, the result is

$$\partial\mathbf{h}_s/\partial\boldsymbol{\pi}' = [\mathbf{I}_n \otimes (\mathbf{B}_0')^{-1}]\,\partial\boldsymbol{\psi}_s/\partial\boldsymbol{\pi}'.$$

as claimed in [11.7.14]. Similarly, if ξ is an element of $\mathbf{\Omega}$, then ξ has no effect on $\mathbf{\Psi}_s$, and its influence on $\mathbf{B}_0'$ is given by

$$\frac{\partial \operatorname{vec}(\mathbf{B}_0')}{\partial\xi} = \mathbf{S}_B \frac{\partial\boldsymbol{\theta}_B}{\partial\xi}.$$

Stacking [11.B.24] horizontally with ξ representing each of the elements of $\operatorname{vech}(\mathbf{\Omega})$ thus produces

$$\frac{\partial\mathbf{h}_s}{\partial[\operatorname{vech}(\mathbf{\Omega})]'} = -[\mathbf{H}_s \otimes (\mathbf{B}_0')^{-1}]\mathbf{S}_B \frac{\partial\boldsymbol{\theta}_B}{\partial[\operatorname{vech}(\mathbf{\Omega})]'},$$

as claimed in [11.7.15]. ■

Chapter 11 Exercises

11.1. Verify result [11.A.2].

11.2. Consider the following three-variable *VAR*:

$$\begin{aligned} y_{1t} &= \alpha y_{1,t-1} + \beta y_{2,t-1} + \varepsilon_{1t} \\ y_{2t} &= \gamma y_{1,t-1} + \varepsilon_{2t} \\ y_{3t} &= \xi y_{1,t-1} + \zeta y_{2,t-1} + \eta y_{3,t-1} + \varepsilon_{3t}. \end{aligned}$$

(a) Is y_{1t} block-exogenous with respect to the vector $(y_{2t}, y_{3t})'$?
(b) Is the vector (y_{1t}, y_{2t}) block-exogenous with respect to y_{3t}?
(c) Is y_{3t} block-exogenous with respect to the vector $(y_{1t}, y_{2t})'$?

11.3. Consider the following bivariate *VAR*:

$$\begin{aligned} y_{1t} &= \alpha_1 y_{1,t-1} + \alpha_2 y_{1,t-2} + \cdots + \alpha_p y_{1,t-p} \\ &\quad + \beta_1 y_{2,t-1} + \beta_2 y_{2,t-2} + \cdots + \beta_p y_{2,t-p} + \varepsilon_{1t} \\ y_{2t} &= \gamma_1 y_{1,t-1} + \gamma_2 y_{1,t-2} + \cdots + \gamma_p y_{1,t-p} \\ &\quad + \delta_1 y_{2,t-1} + \delta_2 y_{2,t-2} + \cdots + \delta_p y_{2,t-p} + \varepsilon_{2t} \end{aligned}$$

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_\tau') = \begin{cases} \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Use the results of Section 11.3 to write this in the form

$$\begin{aligned} y_{1t} &= \zeta_1 y_{1,t-1} + \zeta_2 y_{1,t-2} + \cdots + \zeta_p y_{1,t-p} \\ &\quad + \eta_1 y_{2,t-1} + \eta_2 y_{2,t-2} + \cdots + \eta_p y_{2,t-p} + u_{1t} \\ y_{2t} &= \lambda_0 y_{1t} + \lambda_1 y_{1,t-1} + \lambda_2 y_{1,t-2} + \cdots + \lambda_p y_{1,t-p} \\ &\quad + \xi_1 y_{2,t-1} + \xi_2 y_{2,t-2} + \cdots + \xi_p y_{2,t-p} + u_{2t}, \end{aligned}$$

where

$$E(\mathbf{u}_t\mathbf{u}_\tau') = \begin{cases} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

What is the relation between the parameters of the first representation $(\alpha_j, \beta_j, \gamma_j, \delta_j, \Omega_{ij})$ and those of the second representation $(\zeta_j, \eta_j, \lambda_j, \xi_j, \sigma_i^2)$? What is the relation between $\boldsymbol{\varepsilon}_t$ and $\mathbf{u}_t$?

11.4. Write the result for Exercise 11.3 as

$$\begin{bmatrix} 1 - \zeta(L) & -\eta(L) \\ -\lambda_0 - \lambda(L) & 1 - \xi(L) \end{bmatrix} \begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix}$$

or

$$\mathbf{A}(L)\mathbf{y}_t = \mathbf{u}_t.$$

Premultiply this system by the adjoint of $\mathbf{A}(L)$,

$$\mathbf{A}^*(L) \equiv \begin{bmatrix} 1 - \xi(L) & \eta(L) \\ \lambda_0 + \lambda(L) & 1 - \zeta(L) \end{bmatrix},$$

to deduce that y_{1t} and y_{2t} each admit a univariate $ARMA(2p, p)$ representation. Show how the argument generalizes to establish that if the $(n \times 1)$ vector $\mathbf{y}_t$ follows a pth-order autoregression, then each individual element y_{it} follows an $ARMA[np, (n - 1)p]$ process. (See Zellner and Palm, 1974).

11.5. Consider the following bivariate VAR:

$$y_{1t} = 0.3y_{1,t-1} + 0.8y_{2,t-1} + \varepsilon_{1t}$$
$$y_{2t} = 0.9y_{1,t-1} + 0.4y_{2,t-1} + \varepsilon_{2t},$$

with $E(\varepsilon_{1t}\varepsilon_{1\tau}) = 1$ for $t = \tau$ and 0 otherwise, $E(\varepsilon_{2t}\varepsilon_{2\tau}) = 2$ for $t = \tau$ and 0 otherwise, and $E(\varepsilon_{1t}\varepsilon_{2\tau}) = 0$ for all t and τ.

(a) Is this system covariance-stationary?

(b) Calculate $\mathbf{\Psi}_s = \partial \mathbf{y}_{t+s}/\partial \boldsymbol{\varepsilon}_t'$ for $s = 0, 1$, and 2. What is the limit as $s \to \infty$?

(c) Calculate the fraction of the MSE of the two-period-ahead forecast error for variable 1,

$$E[y_{1,t+2} - \hat{E}(y_{1,t+2} | y_t, y_{t-1}, \ldots)]^2,$$

that is due to $\varepsilon_{1,t+1}$ and $\varepsilon_{1,t+2}$.

Chapter 11 References

Ashley, Richard. 1988. "On the Relative Worth of Recent Macroeconomic Forecasts." *International Journal of Forecasting* 4:363–76.

Baillie, Richard T. 1987. "Inference in Dynamic Models Containing 'Surprise' Variables." *Journal of Econometrics* 35:101–17.

Bernanke, Ben. 1986. "Alternative Explanations of the Money-Income Correlation." *Carnegie-Rochester Conference Series on Public Policy* 25:49–100.

Blanchard, Olivier. 1989. "A Traditional Interpretation of Macroeconomic Fluctuations." *American Economic Review* 79:1146–64.

——— and Peter Diamond. 1989. "The Beveridge Curve." *Brookings Papers on Economic Activity* I:1989, 1–60.

——— and ———. 1990. "The Cyclical Behavior of the Gross Flows of U.S. Workers." *Brookings Papers on Economic Activity* II:1990, 85–155.

——— and Danny Quah. 1989. "The Dynamic Effects of Aggregate Demand and Aggregate Supply Disturbances." *American Economic Review* 79:655–73.

——— and Mark Watson. 1986. "Are Business Cycles All Alike?" in Robert J. Gordon, ed., *The American Business Cycle*. Chicago: University of Chicago Press.

Bouissou, M. B., J. J. Laffont, and Q. H. Vuong. 1986. "Tests of Noncausality under Markov Assumptions for Qualitative Panel Data." *Econometrica* 54:395–414.

Christiano, Lawrence J., and Lars Ljungqvist. 1988. "Money Does Granger-Cause Output in the Bivariate Money-Output Relation." *Journal of Monetary Economics* 22:217–35.

Cooley, Thomas F., and Stephen F. LeRoy. 1985. "Atheoretical Macroeconometrics: A Critique." *Journal of Monetary Economics* 16:283–308.

Fama, Eugene F. 1965. "The Behavior of Stock Market Prices." *Journal of Business* 38:34–105.

Feige, Edgar L., and Douglas K. Pearce. 1979. "The Casual Causal Relationship between Money and Income: Some Caveats for Time Series Analysis." *Review of Economics and Statistics* 61:521–33.

Flavin, Marjorie A. 1981. "The Adjustment of Consumption to Changing Expectations about Future Income." *Journal of Political Economy* 89:974–1009.

Geweke, John. 1982. "Measurement of Linear Dependence and Feedback between Multiple Time Series." *Journal of the American Statistical Association* 77:304–13.

———, Richard Meese, and Warren Dent. 1983. "Comparing Alternative Tests of Causality in Temporal Systems: Analytic Results and Experimental Evidence." *Journal of Econometrics* 21:161–94.

Giannini, Carlo. 1992. *Topics in Structural VAR Econometrics*. New York: Springer-Verlag.

Granger, C. W. J. 1969. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica* 37:424–38.

Hamilton, James D. 1983. "Oil and the Macroeconomy since World War II." *Journal of Political Economy* 91:228–48.

———. 1985. "Historical Causes of Postwar Oil Shocks and Recessions." *Energy Journal* 6:97–116.

Hansen, Lars P., and Thomas J. Sargent. 1981. "Formulating and Estimating Dynamic Linear Rational Expectations Models," in Robert E. Lucas, Jr., and Thomas J. Sargent, eds., *Rational Expectations and Econometric Practice*, Vol. I. Minneapolis: University of Minnesota Press.

Keating, John W. 1990. "Identifying VAR Models under Rational Expectations." *Journal of Monetary Economics* 25:453–76.

Leamer, Edward. 1985. "Vector Autoregressions for Causal Inference?" *Carnegie-Rochester Conference Series on Public Policy* 22:255–303.

Lucas, Robert E., Jr. 1978. "Asset Prices in an Exchange Economy." *Econometrica* 46:1429–45.

Lütkepohl, Helmut. 1989. "A Note on the Asymptotic Distribution of Impulse Response Functions of Estimated VAR Models with Orthogonal Residuals." *Journal of Econometrics* 42:371–76.

———. 1990. "Asymptotic Distributions of Impulse Response Functions and Forecast Error Variance Decompositions of Vector Autoregressive Models." *Review of Economics and Statistics* 72:116–25.

Magnus, Jan R. 1978. "Maximum Likelihood Estimation of the GLS Model with Unknown Parameters in the Disturbance Covariance Matrix." *Journal of Econometrics* 7:281–312.

——— and Heinz Neudecker. 1988. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York: Wiley.

Pierce, David A., and Larry D. Haugh. 1977. "Causality in Temporal Systems: Characterization and a Survey." *Journal of Econometrics* 5:265–93.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39:577–91.

———. 1973. *Efficient Estimation with a Priori Information*. New Haven, Conn.: Yale University Press.

Runkle, David E. 1987. "Vector Autoregressions and Reality." *Journal of Business and Economic Statistics* 5:437–42.

Shapiro, Matthew D., and Mark W. Watson. 1988. "Sources of Business Cycle Fluctuations," in Stanley Fischer, ed., *NBER Macroeconomics Annual 1988*. Cambridge, Mass.: MIT Press.

Sims, Christopher A. 1972. "Money, Income and Causality." *American Economic Review* 62:540–52.

———. 1980. "Macroeconomics and Reality." *Econometrica* 48:1–48.

———. 1986. "Are Forecasting Models Usable for Policy Analysis?" *Quarterly Review of the Federal Reserve Bank of Minneapolis* (Winter), 2–16.

Stock, James H., and Mark W. Watson. 1989. "Interpreting the Evidence on Money-Income Causality." *Journal of Econometrics* 40:161–81.

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley.

Zellner, Arnold. 1962. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias." *Journal of the American Statistical Association* 57:348–68.

——— and Franz Palm. 1974. "Time Series Analysis and Simultaneous Equation Econometric Models." *Journal of Econometrics* 2:17–54.

12 Bayesian Analysis

The previous chapter noted that because so many parameters are estimated in a vector autoregression, the standard errors for inferences can be large. The estimates can be improved if the analyst has any information about the parameters beyond that contained in the sample. Bayesian estimation provides a convenient framework for incorporating prior information with as much weight as the analyst feels it merits.

Section 12.1 introduces the basic principles underlying Bayesian analysis and uses them to analyze a standard regression model or univariate autoregression. Vector autoregressions are discussed in Section 12.2. For the specifications in Sections 12.1 and 12.2, the Bayesian estimators can be found analytically. Numerical methods that can be used to analyze more general statistical problems from a Bayesian framework are reviewed in Section 12.3.

12.1. *Introduction to Bayesian Analysis*

Let $\boldsymbol{\theta}$ be an $(a \times 1)$ vector of parameters to be estimated from a sample of observations. For example, if $y_t \sim$ i.i.d. $N(\mu, \sigma^2)$, then $\boldsymbol{\theta} = (\mu, \sigma^2)'$ is to be estimated on the basis of $\mathbf{y} \equiv (y_1, y_2, \ldots, y_T)'$. Much of the discussion up to this point in the text has been based on the *classical* statistical perspective that there exists some true value of $\boldsymbol{\theta}$. This true value is regarded as an unknown but fixed number. An estimator $\hat{\boldsymbol{\theta}}$ is constructed from the data, and $\hat{\boldsymbol{\theta}}$ is therefore a random variable. In classical statistics, the mean and plim of the random variable $\hat{\boldsymbol{\theta}}$ are compared with the true value $\boldsymbol{\theta}$. The efficiency of the estimator is judged by the mean squared error of the random variable, $E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'$. A popular classical estimator is the value $\hat{\boldsymbol{\theta}}$ that maximizes the sample likelihood, which for this example would be

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - \mu)^2}{2\sigma^2}\right]. \qquad [12.1.1]$$

In Bayesian statistics, by contrast, $\boldsymbol{\theta}$ itself is regarded as a random variable. All inference about $\boldsymbol{\theta}$ takes the form of statements of probability, such as "there is only a 0.05 probability that θ_1 is greater than zero." The view is that the analyst will always have some uncertainty about $\boldsymbol{\theta}$, and the goal of statistical analysis is to describe this uncertainty in terms of a probability distribution. Any information the analyst had about $\boldsymbol{\theta}$ before observing the data is represented by a *prior density*

$f(\boldsymbol{\theta})$.[1] Probability statements that the analyst might have made about $\boldsymbol{\theta}$ before observing the data can be expressed as integrals of $f(\boldsymbol{\theta})$; for example, the previous statement would be expressed as $\int_0^\infty f(\theta_1)\, d\theta_1 = 0.05$ where $f(\theta_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{\theta})\, d\theta_2\, d\theta_3 \cdots d\theta_n$. The sample likelihood [12.1.1] is viewed as the density of $\mathbf{y}$ conditional on the value of the random variable $\boldsymbol{\theta}$, denoted $f(\mathbf{y}|\boldsymbol{\theta})$. The product of the prior density and the sample likelihood gives the joint density of $\mathbf{y}$ and $\boldsymbol{\theta}$:

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta}). \qquad [12.1.2]$$

Probability statements that would be made about $\boldsymbol{\theta}$ after the data $\mathbf{y}$ have been observed are based on the *posterior density* of $\boldsymbol{\theta}$, which is given by

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}, \boldsymbol{\theta})}{f(\mathbf{y})}. \qquad [12.1.3]$$

Recalling [12.1.2] and the fact that $f(\mathbf{y}) = \int_{-\infty}^{\infty} f(\mathbf{y}, \boldsymbol{\theta})\, d\boldsymbol{\theta}$, equation [12.1.3] can be written as

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta}) \cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \qquad [12.1.4]$$

which is known as *Bayes's law*. In practice, the posterior density can sometimes be found simply by rearranging the elements in [12.1.2] as

$$f(\mathbf{y}, \boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{y}) \cdot f(\mathbf{y}),$$

where $f(\mathbf{y})$ is a density that does not involve $\boldsymbol{\theta}$; the other factor, $f(\boldsymbol{\theta}|\mathbf{y})$, is then the posterior density.

Estimating the Mean of a Gaussian Distribution with Known Variance

To illustrate the Bayesian approach, let $y_t \sim$ i.i.d. $N(\mu, \sigma^2)$ as before and write the sample likelihood [12.1.1] as

$$f(\mathbf{y}|\mu; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mu \cdot \mathbf{1})'(\mathbf{y} - \mu \cdot \mathbf{1})\right\}, \qquad [12.1.5]$$

where $\mathbf{1}$ denotes a $(T \times 1)$ vector of 1s. Here μ is regarded as a random variable. To keep the example simple, we will assume that the variance σ^2 is known with certainty. Suppose that prior information about μ is represented by the prior distribution $\mu \sim N(m, \sigma^2/\nu)$:

$$f(\mu; \sigma^2) = \frac{1}{(2\pi\sigma^2/\nu)^{1/2}} \exp\left[\frac{-(\mu - m)^2}{2\sigma^2/\nu}\right]. \qquad [12.1.6]$$

Here m and ν are parameters that describe the nature and quality of prior information about μ. The parameter m can be interpreted as the estimate of μ the analyst would have made before observing $\mathbf{y}$, with σ^2/ν the *MSE* of this estimate. Expressing this *MSE* as a multiple $(1/\nu)$ of the variance of the distribution for y_t turns out to simplify some of the expressions that follow. Greater confidence in the prior information would be represented by larger values of ν.

[1]Throughout this chapter we will omit the subscript that indicates the random variable whose density is being described; for example, $f_\Theta(\boldsymbol{\theta})$ will simply be denoted $f(\boldsymbol{\theta})$. The random variable whose density is being described should always be clear from the context and the argument of $f(\cdot)$.

To make the idea of a prior distribution more concrete, suppose that before observing $\mathbf{y}$ the analyst had earlier obtained a sample of N separate observations $\{z_i, i = 1, 2, \ldots, N\}$ from the $N(\mu, \sigma^2)$ distribution. It would then be natural to take m to be the mean of this earlier sample ($m = \bar{z} = (1/N)\sum_{i=1}^{N} z_i$) and σ^2/ν to be the variance of $\bar{z}$, that is, to take $\nu = N$. The larger this earlier sample (N), the greater the confidence in the prior information.

The posterior distribution for μ after observing the sample $\mathbf{y}$ is described by the following proposition.

Proposition 12.1: *The product of [12.1.5] and [12.1.6] can be written in the form $f(\mu|\mathbf{y}; \sigma^2) \cdot f(\mathbf{y}; \sigma^2)$, where*

$$f(\mu|\mathbf{y}; \sigma^2) = \frac{1}{[2\pi\sigma^2/(\nu + T)]^{1/2}} \exp\left[\frac{-(\mu - m^*)^2}{2\sigma^2/(\nu + T)}\right] \qquad [12.1.7]$$

$$\begin{aligned} f(\mathbf{y}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu|^{-1/2} \\ &\quad \times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - m\cdot\mathbf{1})'(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)^{-1}(\mathbf{y} - m\cdot\mathbf{1})\right\} \end{aligned} \qquad [12.1.8]$$

$$\begin{aligned} m^* &= \left(\frac{\nu}{\nu + T}\right)m + \left(\frac{T}{\nu + T}\right)\bar{y} \\ \bar{y} &= (1/T)\sum_{t=1}^{T} y_t. \end{aligned} \qquad [12.1.9]$$

In other words, the distribution of μ conditional on the data $(y_1, y_2, \ldots, y_T)$ is $N(m^*, \sigma^2/(\nu + T))$, while the marginal distribution of $\mathbf{y}$ is $N(m\cdot\mathbf{1}, \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu))$.

With a quadratic loss function, the Bayesian estimate of μ is the value $\hat{\mu}$ that minimizes $E(\mu - \hat{\mu})^2$. Although this is the same expression as the classical MSE, its interpretation is different. From the Bayesian perspective, μ is a random variable with respect to whose distribution the expectation is taken, and $\hat{\mu}$ is a candidate value for the estimate. The optimal value for $\hat{\mu}$ is the mean of the posterior distribution described in Proposition 12.1:

$$\hat{\mu} = \left(\frac{\nu}{\nu + T}\right)m + \left(\frac{T}{\nu + T}\right)\bar{y}.$$

This is a weighted average of the estimate the classical statistician would use ($\bar{y}$) and an estimate based on prior information alone (m). Larger values of ν correspond to greater confidence in prior information, and this would make the Bayesian estimate closer to m. On the other hand, as ν approaches zero, the Bayesian estimate approaches the classical estimate $\bar{y}$. The limit of [12.1.6] as $\nu \to 0$ is known as a *diffuse* or *improper prior* density. In this case, the quality of prior information is so poor that prior information is completely disregarded in forming the estimate $\hat{\mu}$.

The uncertainty associated with the posterior estimate $\hat{\mu}$ is described by the variance of the posterior distribution. To use the data to evaluate the plausibility of the claim that $\mu_0 < \mu < \mu_1$, we simply calculate the probability $\int_{\mu_0}^{\mu_1} f(\mu|\mathbf{y}; \sigma^2)\, d\mu$. For example, the Bayesian would assert that the probability that μ is within the range $\hat{\mu} \pm 2\sigma/\sqrt{\nu + T}$ is 0.95.

Estimating the Coefficients of a Regression Model with Known Variance

Next, consider the linear regression model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t,$$

where $u_t \sim$ i.i.d. $N(0, \sigma^2)$, $\mathbf{x}_t$ is a $(k \times 1)$ vector of exogenous explanatory variables, and $\boldsymbol{\beta}$ is a $(k \times 1)$ vector of coefficients. Let

$$\underset{(T\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \qquad \underset{(T\times k)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix}.$$

Treating $\boldsymbol{\beta}$ as random but σ^2 as known, we have the likelihood

$$\begin{aligned} f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}; \sigma^2) &= \prod_{t=1}^{T} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](y_t - \mathbf{x}_t'\boldsymbol{\beta})^2\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \end{aligned} \qquad [12.1.10]$$

Suppose that prior information about $\boldsymbol{\beta}$ is represented by a $N(\mathbf{m}, \sigma^2\mathbf{M})$ distribution:

$$f(\boldsymbol{\beta}; \sigma^2) = \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}|^{-1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}. \qquad [12.1.11]$$

Thus, prior to observation of the sample, the analyst's best guess as to the value of $\boldsymbol{\beta}$ is represented by the $(k \times 1)$ vector $\mathbf{m}$, and the confidence in this guess is summarized by the $(k \times k)$ matrix $\sigma^2\mathbf{M}$; less confidence is represented by larger diagonal elements of $\mathbf{M}$. Knowledge about the exogenous variables $\mathbf{X}$ is presumed to have no effect on the prior distribution, so that [12.1.11] also describes $f(\boldsymbol{\beta}|\mathbf{X}; \sigma^2)$.

Proposition 12.1 generalizes as follows.

Proposition 12.2: *The product of [12.1.10] and [12.1.11] can be written in the form* $f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}; \sigma^2) \cdot f(\mathbf{y}|\mathbf{X}; \sigma^2)$, *where*

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X}|^{1/2} \\ &\quad \times \exp\left\{[-1/(2\sigma^2)](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \mathbf{m}^*)\right\} \end{aligned} \qquad [12.1.12]$$

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}; \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2} \\ &\quad \times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\right\} \end{aligned} \qquad [12.1.13]$$

$$\mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}). \qquad [12.1.14]$$

In other words, the distribution of $\boldsymbol{\beta}$ *conditional on the observed data is* $N(\mathbf{m}^*, \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1})$ *and the marginal distribution of* $\mathbf{y}$ *given* $\mathbf{X}$ *is* $N(\mathbf{X}\mathbf{m}, \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'))$.

Poor prior information about $\boldsymbol{\beta}$ corresponds to a large variance $\mathbf{M}$, or equivalently a small value for $\mathbf{M}^{-1}$. The diffuse prior distribution for this problem is often represented by the limit as $\mathbf{M}^{-1} \to \mathbf{0}$, for which the posterior mean [12.1.14] becomes $\mathbf{m}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, the *OLS* estimator. The variance of the posterior distribution becomes $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Thus, classical regression inference is reproduced as a special case of Bayesian inference with a diffuse prior distribution. At the other extreme, if $\mathbf{X}'\mathbf{X} = \mathbf{0}$, the sample contains no information about $\boldsymbol{\beta}$ and the posterior distribution is $N(\mathbf{m}, \sigma^2\mathbf{M})$, the same as the prior distribution.

If the analyst's prior expectation is that all coefficients are zero ($\mathbf{m} = \mathbf{0}$) and this claim is made with the same confidence for each coefficient ($\mathbf{M}^{-1} = \lambda \cdot \mathbf{I}_k$ for some $\lambda > 0$), then the Bayesian estimator [12.1.14] is

$$\mathbf{m}^* = (\lambda \cdot \mathbf{I}_k + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \qquad [12.1.15]$$

which is the *ridge regression* estimator proposed by Hoerl and Kennard (1970). The effect of ridge regression is to shrink the parameter estimates toward zero.

Bayesian Estimation of a Regression Model with Unknown Variance

Propositions 12.1 and 12.2 assumed that the residual variance σ^2 was known with certainty. Usually, both σ^2 and $\boldsymbol{\beta}$ would be regarded as random variables, and Bayesian analysis requires a prior distribution for σ^2. A convenient prior distribution for this application is provided by the gamma distribution. Let $\{Z_i\}_{i=1}^N$ be a sequence of i.i.d. $N(0, \tau^2)$ variables. Then $W = \Sigma_{i=1}^N Z_i^2$ is said to have a gamma distribution with N degrees of freedom and scale parameter λ, indicated $W \sim \Gamma(N, \lambda)$, where $\lambda = 1/\tau^2$. Thus, W has the distribution of τ^2 times a $\chi^2(N)$ variable. The mean of W is given by

$$E(W) = N \cdot E(Z_i^2) = N\tau^2 = N/\lambda, \qquad [12.1.16]$$

and the variance is

$$\begin{aligned} E(W^2) - [E(W)]^2 &= N \cdot \{E(Z_i^4) - [E(Z_i^2)]^2\} \\ &= N \cdot (3\tau^4 - \tau^4) = 2N\tau^4 = 2N/\lambda^2. \end{aligned} \qquad [12.1.17]$$

The density of W takes the form

$$f(w) = \frac{(\lambda/2)^{N/2} w^{[(N/2)-1]} \exp[-\lambda w/2]}{\Gamma(N/2)}, \qquad [12.1.18]$$

where $\Gamma(\cdot)$ denotes the gamma function. If N is an even integer, then

$$\Gamma(N/2) = 1 \cdot 2 \cdot 3 \cdots [(N/2) - 1],$$

with $\Gamma(2/2) = 1$; whereas if N is an odd integer, then

$$\Gamma(N/2) = \sqrt{\pi} \cdot \tfrac{1}{2} \cdot \tfrac{3}{2} \cdot \tfrac{5}{2} \cdots [(N/2) - 1],$$

with $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Following DeGroot (1970) and Leamer (1978), it is convenient to describe the prior distribution not in terms of the variance σ^2 but rather in terms of the reciprocal of the variance, σ^{-2}, which is known as the *precision*. Thus, suppose that the prior distribution is specified as $\sigma^{-2} \sim \Gamma(N, \lambda)$, where N and λ are parameters that describe the analyst's prior information:

$$f(\sigma^{-2}|\mathbf{X}) = \frac{(\lambda/2)^{N/2} \sigma^{-2[(N/2)-1]} \exp[-\lambda\sigma^{-2}/2]}{\Gamma(N/2)}. \qquad [12.1.19]$$

Recalling [12.1.16], the ratio N/λ is the value expected for σ^{-2} on the basis of prior information. As we will see shortly in Proposition 12.3, if the prior information is based on an earlier sample of observations $\{z_1, z_2, \ldots, z_N\}$, the parameter N turns out to describe the size of this earlier sample and λ is the earlier sample's sum of squared residuals. For a given ratio of N/λ, larger values for N imply greater confidence in the prior information.

The prior distribution of $\boldsymbol{\beta}$ conditional on the value for σ^{-2} is the same as in [12.1.11]:

$$f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}|^{-1/2} \times \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}. \qquad [12.1.20]$$

Thus, $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$, the joint prior density for $\boldsymbol{\beta}$ and σ^{-2}, is given by the product of [12.1.19] and [12.1.20]. The posterior distribution $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X})$ is described by the following proposition.

Proposition 12.3: *Let the prior density $f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$ be given by the product of [12.1.19] and [12.1.20], and let the sample likelihood be*

$$f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{T/2}}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \qquad [12.1.21]$$

Then the following hold:

(*a*) *The joint posterior density of $\boldsymbol{\beta}$ and σ^{-2} is given by*

$$f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X}) = f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})\cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}), \qquad [12.1.22]$$

where the posterior distribution of $\boldsymbol{\beta}$ conditional on σ^{-2} is $N(\mathbf{m}^*, \sigma^2\mathbf{M}^*)$:

$$f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X}) = \frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}, \qquad [12.1.23]$$

with

$$\mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}) \qquad [12.1.24]$$

$$\mathbf{M}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}. \qquad [12.1.25]$$

Furthermore, the marginal posterior distribution of σ^{-2} is $\Gamma(N^*, \lambda^*)$:

$$f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = \frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)}\exp[-\lambda^*\sigma^{-2}/2], \qquad [12.1.26]$$

with

$$N^* = N + T \qquad [12.1.27]$$

$$\lambda^* = \lambda + (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \mathbf{m})'\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{m}) \qquad [12.1.28]$$

for $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ the OLS estimator.

(*b*) *The marginal posterior distribution for $\boldsymbol{\beta}$ is a k-dimensional t distribution with N^* degrees of freedom, mean $\mathbf{m}^*$, and scale matrix $(\lambda^*/N^*)\cdot\mathbf{M}^*$:*

$$f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$$
$$= \left\{ \frac{\Gamma[(k + N^*)/2]}{(\pi N^*)^{k/2}\Gamma(N^*/2)} |(\lambda^*/N^*)\mathbf{M}^*|^{-1/2} \right. \qquad [12.1.29]$$
$$\left. \times [1 + (1/N^*)(\boldsymbol{\beta} - \mathbf{m}^*)'[(\lambda^*/N^*)\mathbf{M}^*]^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2} \right\}.$$

(c) *Let* $\mathbf{R}$ *be a known* $(m \times k)$ *matrix with linearly independent rows, and define*

$$Q \equiv \frac{[\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]'[\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \cdot [\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]/m}{\lambda^*/N^*}. \qquad [12.1.30]$$

Then Q *has a marginal posterior distribution that is* $F(m, N^*)$:

$$f(q|\mathbf{y}, \mathbf{X}) = \frac{m^{m/2}(N^*)^{N^*/2}\Gamma[(N^* + m)/2]q^{[(m/2)-1]}}{\Gamma(m/2)\Gamma(N^*/2)(N^* + mq)^{[(N^*+m)/2]}}. \qquad [12.1.31]$$

Recalling [12.1.16], result (a) implies that the Bayesian estimate of the precision is

$$E(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = N^*/\lambda^*. \qquad [12.1.32]$$

Diffuse prior information is sometimes represented as $N = \lambda = 0$ and $\mathbf{M}^{-1} = \mathbf{0}$. Substituting these values into [12.1.27] and [12.1.28] implies that $N^* = T$ and $\lambda^* = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$. For these values, the posterior mean [12.1.32] would be

$$E(\sigma^{-2}|\mathbf{y}, \mathbf{X}) = T/(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}),$$

which is the maximum likelihood estimate of σ^{-2}. This is the basis for the earlier claim that the parameter N for the prior distribution might be viewed as the number of presample observations on which the prior information is based and that λ might be viewed as the sum of squared residuals for these observations.

Result (b) implies that the Bayesian estimate of the coefficient vector is

$$E(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \mathbf{m}^* = (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}^{-1}\mathbf{m} + \mathbf{X}'\mathbf{y}), \qquad [12.1.33]$$

which is identical to the estimate derived in Proposition 12.2 for the case where σ^2 is known. Again, for diffuse prior information, $\mathbf{m}^* = \mathbf{b}$, the *OLS* estimate.

Result (c) describes the Bayesian perspective on a hypothesis about the value of $\mathbf{R}\boldsymbol{\beta}$, where the matrix $\mathbf{R}$ characterizes which linear combinations of the elements of $\boldsymbol{\beta}$ are of interest. A classical statistician would test the hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ by calculating an *OLS* F statistic,

$$\frac{(\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/m}{s^2},$$

and evaluating the probability that an $F(m, T - k)$ variable could equal or exceed this magnitude. This represents the probability that the estimated value of $\mathbf{Rb}$ could be as far as it is observed to be from $\mathbf{r}$ given that the true value of $\boldsymbol{\beta}$ satisfies $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. By contrast, a Bayesian regards $\mathbf{R}\boldsymbol{\beta}$ as a random variable, the distribution for which is described in result (c). According to [12.1.30], the probability that $\mathbf{R}\boldsymbol{\beta}$ would equal $\mathbf{r}$ is related to the probability that an $F(m, N^*)$ variable would assume the value

$$\frac{(\mathbf{r} - \mathbf{Rm}^*)'[\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rm}^*)/m}{\lambda^*/N^*}.$$

The probability that an $F(m, N^*)$ variable could exceed this magnitude represents the probability that the random variable $\mathbf{R\beta}$ might be as far from the posterior mean $\mathbf{Rm}^*$ as is represented by the point $\mathbf{R\beta} = \mathbf{r}$. In the case of a diffuse prior distribution, the preceding expression simplifies to

$$\frac{(\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb})/m}{(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/T},$$

which is to be compared in this case with an $F(m, T)$ distribution. Recalling that

$$s^2 = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})/(T - k),$$

it appears that, apart from a minor difference in the denominator degrees of freedom, the classical statistician and the Bayesian with a diffuse prior distribution would essentially be calculating the identical test statistic and comparing it with the same critical value in evaluating the plausibility of the hypothesis represented by $\mathbf{R\beta} = \mathbf{r}$.

Bayesian Analysis of Regressions with Lagged Dependent Variables

In describing the sample likelihood (expression [12.1.10] or [12.1.21]), the assumption was made that the vector of explanatory variables $\mathbf{x}_t$ was strictly exogenous. If $\mathbf{x}_t$ contains lagged values of y, then as long as we are willing to treat presample values of y as deterministic, the algebra goes through exactly the same. The only changes needed are some slight adjustments in notation and in the description of the results. For example, consider a pth-order autoregression with $\mathbf{x}_t = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p})'$. In this case, the expression on the right side of [12.1.21] describes the likelihood of $(y_1, y_2, \ldots, y_T)$ conditional on $y_0, y_{-1}, \ldots, y_{-p+1}$; that is, it describes $f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{x}_1)$. The prior distributions [12.1.19] and [12.1.20] are then presumed to describe $f(\sigma^{-2}|\mathbf{x}_1)$ and $f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{x}_1)$, and the posterior distributions are all as stated in Proposition 12.3.

Note in particular that results (b) and (c) of Proposition 12.3 describe the exact small-sample posterior distributions, even when $\mathbf{x}_t$ contains lagged dependent variables. By contrast, a classical statistician would consider the usual t and F tests to be valid only asymptotically.

Calculation of the Posterior Distribution Using a GLS Regression

It is sometimes convenient to describe the prior information in terms of certain linear combinations of coefficients, such as

$$\mathbf{R\beta}|\sigma^{-2} \sim N(\mathbf{r}, \sigma^2\mathbf{V}). \qquad [12.1.34]$$

Here $\mathbf{R}$ denotes a known nonsingular $(k \times k)$ matrix whose rows represent linear combinations of $\boldsymbol{\beta}$ in terms of which it is convenient to describe the analyst's prior information. For example, if the prior expectation is that $\beta_1 = \beta_2$, then the first row of $\mathbf{R}$ could be $(1, -1, 0, \ldots, 0)$ and the first element of $\mathbf{r}$ would be zero. The (1, 1) element of $\mathbf{V}$ reflects the uncertainty of this prior information. If $\boldsymbol{\beta} \sim N(\mathbf{m}, \sigma^2\mathbf{M})$, then $\mathbf{R\beta} \sim N(\mathbf{Rm}, \sigma^2\mathbf{RMR}')$. Thus, the relation between the parameters for the prior distribution as expressed in [12.1.34] ($\mathbf{R}$, $\mathbf{r}$, and $\mathbf{V}$) and the parameters for the prior distribution as expressed in [12.1.20] ($\mathbf{m}$ and $\mathbf{M}$) is given by

$$\mathbf{r} = \mathbf{Rm} \qquad [12.1.35]$$

$$\mathbf{V} = \mathbf{RMR}'. \qquad [12.1.36]$$

Equation [12.1.36] implies

$$\mathbf{V}^{-1} = (\mathbf{R}')^{-1}\mathbf{M}^{-1}\mathbf{R}^{-1}. \qquad [12.1.37]$$

If equation [12.1.37] is premultiplied by $\mathbf{R}'$ and postmultiplied by $\mathbf{R}$, the result is

$$\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} = \mathbf{M}^{-1}. \qquad [12.1.38]$$

Using equations [12.1.35] and [12.1.38], the posterior mean [12.1.33] can be rewritten as

$$\mathbf{m}^* = (\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}'\mathbf{V}^{-1}\mathbf{r} + \mathbf{X}'\mathbf{y}). \qquad [12.1.39]$$

To obtain another perspective on [12.1.39], notice that the prior distribution [12.1.34] can be written

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad [12.1.40]$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$. This is of the same form as the observation equations of the regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \qquad [12.1.41]$$

with $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$. The mixed estimation strategy described by Theil (1971, pp. 347–49) thus regards the prior information as a set of k additional observations, with r_i treated as if it were another observation on y_t and the ith row of $\mathbf{R}$ corresponding to its vector of explanatory variables $\mathbf{x}_t'$. Specifically, equations [12.1.40] and [12.1.41] are stacked to form the system

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*, \qquad [12.1.42]$$

where

$$\underset{(T+k)\times 1}{\mathbf{y}^*} = \begin{bmatrix}\mathbf{r}\\ \mathbf{y}\end{bmatrix} \qquad \underset{(T+k)\times k}{\mathbf{X}^*} = \begin{bmatrix}\mathbf{R}\\ \mathbf{X}\end{bmatrix}$$

$$E(\mathbf{u}^*\mathbf{u}^{*\prime}) = \sigma^2\mathbf{V}^* = \sigma^2\begin{bmatrix}\mathbf{V} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}.$$

The *GLS* estimator for the stacked system is

$$\begin{aligned}\hat{\mathbf{b}} &= [\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{X}^*]^{-1}[\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{y}^*]\\ &= \left\{[\mathbf{R}' \quad \mathbf{X}']\begin{bmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}\begin{bmatrix}\mathbf{R}\\ \mathbf{X}\end{bmatrix}\right\}^{-1} \times \left\{[\mathbf{R}' \quad \mathbf{X}']\begin{bmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_T\end{bmatrix}\begin{bmatrix}\mathbf{r}\\ \mathbf{y}\end{bmatrix}\right\}\\ &= (\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}'\mathbf{V}^{-1}\mathbf{r} + \mathbf{X}'\mathbf{y}).\end{aligned}$$

Thus the posterior mean [12.1.39] can be calculated by *GLS* estimation of [12.1.42]. For known σ^2, the usual formula for the variance of the *GLS* estimator,

$$\sigma^2[\mathbf{X}^{*\prime}(\mathbf{V}^*)^{-1}\mathbf{X}^*]^{-1} = \sigma^2(\mathbf{R}'\mathbf{V}^{-1}\mathbf{R} + \mathbf{X}'\mathbf{X})^{-1},$$

gives a correct calculation of the variance of the Bayesian posterior distribution, $\sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}$.

The foregoing discussion assumed that $\mathbf{R}$ was a nonsingular $(k \times k)$ matrix. On some occasions the analyst might have valuable information about some linear combinations of coefficients but not others. Thus, suppose that the prior distribution

[12.1.34] is written as

$$\begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} \boldsymbol{\beta} \sim N\left(\begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \end{bmatrix}, \sigma^2 \begin{bmatrix} \mathbf{V}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2 \end{bmatrix} \right),$$

where $\mathbf{R}_1$ is an $(m \times k)$ matrix consisting of those linear combinations for which the prior information is good and $\mathbf{R}_2$ is a $[(k - m) \times k]$ matrix of the remaining linear combinations. Then diffuse prior information about those linear combinations described by $\mathbf{R}_2$ could be represented by the limit as $\mathbf{V}_2^{-1} \to \mathbf{0}$, for which

$$\mathbf{R}'\mathbf{V}^{-1} = [\mathbf{R}_1' \quad \mathbf{R}_2'] \begin{bmatrix} \mathbf{V}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_2^{-1} \end{bmatrix} \to [\mathbf{R}_1'\mathbf{V}_1^{-1} \quad \mathbf{0}].$$

The Bayesian estimate [12.1.39] then becomes

$$(\mathbf{R}_1'\mathbf{V}_1^{-1}\mathbf{R}_1 + \mathbf{X}'\mathbf{X})^{-1}(\mathbf{R}_1'\mathbf{V}_1^{-1}\mathbf{r}_1 + \mathbf{X}'\mathbf{y}),$$

which can be calculated from *GLS* estimation of a $[(T + m) \times 1]$ system of the form of [12.1.42] in which only the linear combinations for which there is useful prior information are added as observations.

12.2. *Bayesian Analysis of Vector Autoregressions*

Litterman's Prior Distribution for Estimation of an Equation of a VAR

This section discusses prior information that might help improve the estimates of a single equation of a *VAR*. Much of the early econometric research with dynamic relations was concerned with estimation of distributed lag relations of the form

$$y_t = c + \omega_0 x_t + \omega_1 x_{t-1} + \cdots + \omega_p x_{t-p} + u_t. \qquad [12.2.1]$$

For this specification, ω_s has the interpretation as $\partial y_t/\partial x_{t-s}$, and some have argued that this should be a smooth function of s; see Almon (1965) and Shiller (1973) for examples. Whatever the merit of this view, it is hard to justify imposing a smoothness condition on the sequences $\{\omega_s\}_{s=1}^p$ or $\{\phi_s\}_{s=1}^p$ in a model with autoregressive terms such as

$$\begin{aligned} y_t = c &+ \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} \\ &+ \omega_0 x_t + \omega_1 x_{t-1} + \cdots + \omega_p x_{t-p} + u_t, \end{aligned}$$

since here the dynamic multiplier $\partial y_t/\partial x_{t-s}$ is a complicated nonlinear function of the ϕ's and ω's.

Litterman (1986) suggested an alternative representation of prior information based on the belief that the change in the series is impossible to forecast:

$$y_t - y_{t-1} = c + \varepsilon_t, \qquad [12.2.2]$$

where ε_t is uncorrelated with lagged values of any variable. Economic theory predicts such behavior for many time series. For example, suppose that y_t is the log of the real price of some asset at time t, that is, the price adjusted for inflation. Then $y_t - y_{t-1}$ is approximately the real rate of return from buying the asset at $t - 1$ and selling it at t. In an extension of Fama's (1965) efficient markets argument described in Section 11.2, speculators would have bought more of the asset at time $t - 1$ if they had expected unusually high returns, driving y_{t-1} up in

relation to the anticipated value of y_t. The time path for $\{y_t\}$ that results from such speculation would exhibit price changes that are unforecastable. Thus, we might expect the real prices of items such as stocks, real estate, or precious metals to satisfy [12.2.2]. Hall (1978) argued that the level of spending by consumers should also satisfy [12.2.2], while Barro (1979) and Mankiw (1987) developed related arguments for the taxes levied and new money issued by the government. Changes in foreign exchange rates are argued by many to be unpredictable as well; see the evidence reviewed in Diebold and Nason (1990).

Write the ith equation in a VAR as

$$\begin{aligned} y_{it} = c_i &+ \phi_{i1}^{(1)} y_{1,t-1} + \phi_{i2}^{(1)} y_{2,t-1} + \cdots + \phi_{in}^{(1)} y_{n,t-1} \\ &+ \phi_{i1}^{(2)} y_{1,t-2} + \phi_{i2}^{(2)} y_{2,t-2} + \cdots + \phi_{in}^{(2)} y_{n,t-2} + \cdots \\ &+ \phi_{i1}^{(p)} y_{1,t-p} + \phi_{i2}^{(p)} y_{2,t-p} + \cdots + \phi_{in}^{(p)} y_{n,t-p} + \varepsilon_{it}, \end{aligned} \qquad [12.2.3]$$

where $\phi_{ij}^{(s)}$ gives the coefficient relating y_{it} to $y_{j,t-s}$. The restriction [12.2.2] requires $\phi_{ii}^{(1)} = 1$ and all other $\phi_{ij}^{(s)} = 0$. These values (0 or 1) then characterize the mean of the prior distribution for the coefficients. Litterman used a diffuse prior distribution for the constant term c_i.

Litterman took the variance-covariance matrix for the prior distribution to be diagonal, with γ denoting the standard deviation of the prior distribution for $\phi_{ii}^{(1)}$:

$$\phi_{ii}^{(1)} \sim N(1, \gamma^2).$$

Although each equation $i = 1, 2, \ldots, n$ of the VAR is estimated separately, typically the same number γ is used for each i. A smaller value for γ represents greater confidence in the prior information and will force the parameter estimates to be closer to the values predicted in [12.2.2]. A value of $\gamma = 0.20$ means that, before seeing the data, the analyst had 95% confidence that $\phi_{ii}^{(1)}$ is no smaller than 0.60 and no larger than 1.40.

The coefficients relating y_{it} to further lags are predicted to be zero, and Litterman argued that the analyst should have more confidence in this prediction the greater the lag. He therefore suggested taking $\phi_{ii}^{(2)} \sim N(0, (\gamma/2)^2)$, $\phi_{ii}^{(3)} \sim N(0, (\gamma/3)^2), \ldots$, and $\phi_{ii}^{(p)} \sim N(0, (\gamma/p)^2)$, tightening the prior distribution with a harmonic series for the standard deviation as the lag increases.

Note that the coefficients $\phi_{ii}^{(s)}$ are scale-invariant; if each value of y_{it} is multiplied by 100, the values of $\phi_{ii}^{(s)}$ will be the same. The same is not true of $\phi_{ij}^{(s)}$ for $i \neq j$; if series i is multiplied by 100 but series j is not, then $\phi_{ij}^{(s)}$ will be multiplied by 100. Thus, in calculating the weight to be given the prior information about $\phi_{ij}^{(s)}$, an adjustment for the units in which the data are measured is necessary. Litterman proposed using the following standard deviation of the prior distribution for $\phi_{ij}^{(s)}$:

$$\frac{w \cdot \gamma \cdot \hat{\tau}_i}{s \cdot \hat{\tau}_j}. \qquad [12.2.4]$$

Here $(\hat{\tau}_i/\hat{\tau}_j)$ is a correction for the scale of series i compared with series j. Litterman suggested that $\hat{\tau}_i$ could be estimated from the standard deviation of the residuals from an OLS regression of y_{it} on a constant and on p of its own lagged values. Apart from this scale correction, [12.2.4] simply multiplies γ/s (which was the standard deviation for the prior distribution for $\phi_{ii}^{(s)}$) by a parameter w. Common experience with many time series is that the own lagged values $y_{i,t-s}$ are likely to

be of more help in forecasting y_{it} than will be values of other variables $y_{j,t-s}$. Hence we should have more confidence in the prior belief that $\phi_{ij}^{(s)} = 0$ than the prior belief that $\phi_{ii}^{(s)} = 0$, suggesting a value for w that is less than 1. Doan (1990) recommended a value of $w = 0.5$ in concert with $\gamma = 0.20$.

Several cautions in employing this prior distribution should be noted. First, for some series the natural prior expectation might be that the series is white noise rather than an autoregression with unit coefficient. For example, if y_{it} is a series such as the *change* in stock prices, then the mean of $\phi_{ii}^{(1)}$ should be 0 rather than 1. Second, many economic series display seasonal behavior. In such cases, $\phi_{ij}^{(s)}$ is likely to be nonzero for $s = 12$ and 24 with monthly data, for example. Litterman's prior distribution is not well suited for seasonal data, and some researchers suggest using seasonally adjusted data or including seasonal dummy variables in the regression before employing this prior distribution. Finally, the prior distribution is not well suited for systems that exhibit cointegration, a topic discussed in detail in Chapter 19.

Full-Information Bayesian Estimation of a VAR

Litterman's approach to Bayesian estimation of a *VAR* considered a single equation in isolation. It is possible to analyze all of the equations in a *VAR* together in a Bayesian framework, though the analytical results are somewhat more complicated than for the single-equation case; see Zellner (1971, Chapter 8) and Rothenberg (1973, pp. 139–44) for discussion.

12.3. *Numerical Bayesian Methods*

In the previous examples, the class of densities used to represent the prior information was carefully chosen in order to obtain a simple analytical characterization for the posterior distribution. For many specifications of interest, however, it may be impossible to find such a class, or the density that best reflects the analyst's prior information may not be possible to represent with this class. It is therefore useful to have computer-based methods to calculate or approximate posterior moments for a quite general class of problems.

Approximating the Posterior Mean by the Posterior Mode

One option is to use the mode rather than the mean of the posterior distribution, that is, to take the Bayesian estimate $\hat{\boldsymbol{\theta}}$ to be the value that maximizes $f(\boldsymbol{\theta}|\mathbf{y})$. For symmetric unimodal distributions, the mean and the mode will be the same, as turned out to be the case for the coefficient vector $\boldsymbol{\beta}$ in Proposition 12.2. Where the mean and mode differ, with a quadratic loss function the mode is a suboptimal estimator, though typically the posterior mode will approach the posterior mean as the sample size grows (see DeGroot, 1970, p. 236).

Recall from [12.1.2] and [12.1.3] that the posterior density is given by

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})}{f(\mathbf{y})}, \qquad [12.3.1]$$

and therefore the log of the posterior density is

$$\log f(\boldsymbol{\theta}|\mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta}) - \log f(\mathbf{y}). \qquad [12.3.2]$$

Note that if the goal is to maximize [12.3.2] with respect to $\boldsymbol{\theta}$, it is not necessary

to calculate $f(\mathbf{y})$, since this does not depend on $\boldsymbol{\theta}$. The posterior mode can thus be found by maximizing

$$\log f(\boldsymbol{\theta}, \mathbf{y}) = \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta}). \qquad [12.3.3]$$

To evaluate [12.3.2], we need only to be able to calculate the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ and the density that describes the prior information, $f(\boldsymbol{\theta})$. Expression [12.3.2] can be maximized by numerical methods, and often the same particular algorithms that maximize the log likelihood will also maximize [12.3.2]. For example, the log likelihood for a Gaussian regression model such as [12.1.21] can be maximized by a *GLS* regression, just as the posterior mode [12.1.39] can be calculated with a *GLS* regression.

Tierney and Kadane's Approximation for Posterior Moments

Alternatively, Tierney and Kadane (1986) noted that the curvature of the likelihood surface can be used to estimate the distance of the posterior mode from the posterior mean. Suppose that the objective is to calculate

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \int_{-\infty}^{\infty} g(\boldsymbol{\theta})\cdot f(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta}, \qquad [12.3.4]$$

where $\boldsymbol{\theta}$ is an $(a \times 1)$ vector of parameters and g: $\mathbb{R}^a \rightarrow \mathbb{R}^1$ is a function of interest. For example, if $g(\boldsymbol{\theta}) = \theta_1$, then [12.3.4] is the posterior mean of the first parameter, while $g(\boldsymbol{\theta}) = \theta_1^2$ gives the second moment. Expression [12.3.1] can be used to write [12.3.4] as

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{f(\mathbf{y})} = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \qquad [12.3.5]$$

Define

$$h(\boldsymbol{\theta}) \equiv (1/T)\log\{g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\} \qquad [12.3.6]$$

and

$$k(\boldsymbol{\theta}) \equiv (1/T)\log\{f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\}. \qquad [12.3.7]$$

This allows [12.3.5] to be written

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} \exp[T\cdot h(\boldsymbol{\theta})]\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} \exp[T\cdot k(\boldsymbol{\theta})]\, d\boldsymbol{\theta}}. \qquad [12.3.8]$$

Let $\boldsymbol{\theta}^*$ be the value that maximizes [12.3.6], and consider a second-order Taylor series approximation to $h(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$:

$$\begin{aligned} h(\boldsymbol{\theta}) \cong h(\boldsymbol{\theta}^*) &+ \left.\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\cdot(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \\ &+ \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'\left\{\left.\frac{\partial^2 h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right\}(\boldsymbol{\theta} - \boldsymbol{\theta}^*). \end{aligned} \qquad [12.3.9]$$

Assuming that $\boldsymbol{\theta}^*$ is an interior optimum of $h(\cdot)$, the first derivative $[\partial h(\boldsymbol{\theta})/\partial \boldsymbol{\theta}']|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$ is $\mathbf{0}$. Then [12.3.9] could be expressed as

$$h(\boldsymbol{\theta}) \cong h(\boldsymbol{\theta}^*) - (1/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*), \qquad [12.3.10]$$

where

$$\boldsymbol{\Sigma}^* \equiv -\left[\frac{\partial^2 h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right]^{-1}. \qquad [12.3.11]$$

When [12.3.10] is substituted into the numerator of [12.3.8], the result is

$$\begin{aligned}
&\int_{-\infty}^{\infty} \exp[T\cdot h(\boldsymbol{\theta})]\, d\boldsymbol{\theta} \\
&\cong \int_{-\infty}^{\infty} \exp\left\{T\cdot h(\boldsymbol{\theta}^*) - (T/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T\cdot h(\boldsymbol{\theta}^*)] \int_{-\infty}^{\infty} \exp\left\{(-T/2)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T\cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2}|\boldsymbol{\Sigma}^*/T|^{1/2} \\
&\quad \times \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{a/2}|\boldsymbol{\Sigma}^*/T|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)'(\boldsymbol{\Sigma}^*/T)^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\} d\boldsymbol{\theta} \\
&= \exp[T\cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2}|\boldsymbol{\Sigma}^*/T|^{1/2}.
\end{aligned} \qquad [12.3.12]$$

The last equality follows because the expression being integrated is a $N(\boldsymbol{\theta}^*, \boldsymbol{\Sigma}^*/T)$ density and therefore integrates to unity.

Similarly, the function $k(\boldsymbol{\theta})$ can be approximated with an expansion around the posterior mode $\hat{\boldsymbol{\theta}}$,

$$k(\boldsymbol{\theta}) \cong k(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

where $\hat{\boldsymbol{\theta}}$ maximizes [12.3.7] and

$$\hat{\boldsymbol{\Sigma}} \equiv -\left[\frac{\partial^2 k(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}. \qquad [12.3.13]$$

The denominator in [12.3.8] is then approximated by

$$\int_{-\infty}^{\infty} \exp[T\cdot k(\boldsymbol{\theta})]\, d\boldsymbol{\theta} \cong \exp[T\cdot k(\hat{\boldsymbol{\theta}})]\, (2\pi)^{a/2}|\hat{\boldsymbol{\Sigma}}/T|^{1/2}. \qquad [12.3.14]$$

Tierney and Kadane's approximation is obtained by substituting [12.3.12] and [12.3.14] into [12.3.8]:

$$\begin{aligned}
E[g(\boldsymbol{\theta})|\mathbf{y}] &\cong \frac{\exp[T\cdot h(\boldsymbol{\theta}^*)]\, (2\pi)^{a/2}|\boldsymbol{\Sigma}^*/T|^{1/2}}{\exp[T\cdot k(\hat{\boldsymbol{\theta}})]\, (2\pi)^{a/2}|\hat{\boldsymbol{\Sigma}}/T|^{1/2}} \\
&= \frac{|\boldsymbol{\Sigma}^*|^{1/2}}{|\hat{\boldsymbol{\Sigma}}|^{1/2}} \exp\{T\cdot[h(\boldsymbol{\theta}^*) - k(\hat{\boldsymbol{\theta}})]\}.
\end{aligned} \qquad [12.3.15]$$

To calculate this approximation to the posterior mean of $g(\boldsymbol{\theta})$, we first find the value $\boldsymbol{\theta}^*$ that maximizes $(1/T)\cdot\{\log g(\boldsymbol{\theta}) + \log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta})\}$. Then $h(\boldsymbol{\theta}^*)$ in [12.3.15] is the maximum value attained for this function and $\boldsymbol{\Sigma}^*$ is the negative of the inverse of the matrix of second derivatives of this function. Next we find the value $\hat{\boldsymbol{\theta}}$ that maximizes $(1/T)\cdot\{\log f(\mathbf{y}|\boldsymbol{\theta}) + \log f(\boldsymbol{\theta})\}$, with $k(\hat{\boldsymbol{\theta}})$ the maximum value attained and $\hat{\boldsymbol{\Sigma}}$ the negative of the inverse of the matrix of second derivatives.

The required maximization and second derivatives could be calculated analytically or numerically. Substituting the resulting values into [12.3.15] gives the Bayesian posterior estimate of $g(\boldsymbol{\theta})$.

Monte Carlo Estimation of Posterior Moments

Posterior moments can alternatively be estimated using the Monte Carlo approach suggested by Hammersley and Handscomb (1964, Section 5.4) and Kloek and van Dijk (1978). Again, the objective is taken to be calculation of the posterior mean of $g(\boldsymbol{\theta})$. Let $I(\boldsymbol{\theta})$ be some density function defined on $\boldsymbol{\theta}$ with $I(\boldsymbol{\theta}) > 0$ for all $\boldsymbol{\theta}$. Then [12.3.5] can be written

$$E[g(\boldsymbol{\theta})|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{\int_{-\infty}^{\infty} \{g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int_{-\infty}^{\infty} \{f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}}. \qquad [12.3.16]$$

The numerator in [12.3.16] can be interpreted as the expectation of the random variable $\{g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\}$, where this expectation is taken with respect to the distribution implied by the density $I(\boldsymbol{\theta})$. If $I(\boldsymbol{\theta})$ is a known density such as multivariate Gaussian, it may be simple to generate N separate Monte Carlo draws from this distribution, denoted $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)}\}$. We can then calculate the average realized value of the random variable across these Monte Carlo draws:

$$\sum_{i=1}^{N} (1/N)\cdot\{g(\boldsymbol{\theta}^{(i)})\cdot f(\mathbf{y}|\boldsymbol{\theta}^{(i)})\cdot f(\boldsymbol{\theta}^{(i)})/I(\boldsymbol{\theta}^{(i)})\}. \qquad [12.3.17]$$

From the law of large numbers, as $N \to \infty$, this will yield a consistent estimate of

$$E_{I(\boldsymbol{\theta})}\{g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} = \int_{-\infty}^{\infty} \{g(\boldsymbol{\theta})\cdot f(\mathbf{y}|\boldsymbol{\theta})\cdot f(\boldsymbol{\theta})/I(\boldsymbol{\theta})\} I(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \qquad [12.3.18]$$

provided that the integral in [12.3.18] exists. The denominator of [12.3.16] is similarly estimated from

$$\sum_{i=1}^{N} (1/N)\cdot\{f(\mathbf{y}|\boldsymbol{\theta}^{(i)})\cdot f(\boldsymbol{\theta}^{(i)})/I(\boldsymbol{\theta}^{(i)})\}.$$

The integral in [12.3.18] need not exist if the importance density $I(\boldsymbol{\theta})$ goes to zero in the tails faster than the sample likelihood $f(\mathbf{y}|\boldsymbol{\theta})$. Even if [12.3.18] does exist, the Monte Carlo average [12.3.17] may give a poor estimate of [12.3.18] for moderate N if $I(\boldsymbol{\theta})$ is poorly chosen. Geweke (1989) provided advice on specifying $I(\boldsymbol{\theta})$. If the set of allowable values for $\boldsymbol{\theta}$ forms a compact set, then letting $I(\boldsymbol{\theta})$ be the density for the asymptotic distribution of the maximum likelihood estimator is usually a good approach.

A nice illustration of the versatility of Bayesian Monte Carlo methods for analyzing dynamic models is provided by Geweke (1988a). This approach was extended to multivariate dynamic systems in Geweke (1988b).

APPENDIX 12.A. *Proofs of Chapter 12 Propositions*

■ **Proof of Proposition 12.1.** Note that the product of [12.1.5] and [12.1.6] can be written

$$f(\mathbf{y}, \mu; \sigma^2) = \frac{1}{(2\pi)^{(T+1)/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right\}, \qquad [12.A.1]$$

where

$$\underset{(T+1)\times 1}{\boldsymbol{\alpha}} \equiv \begin{bmatrix} \mu - m \\ \mathbf{y} - \mu\cdot\mathbf{1} \end{bmatrix} \qquad [12.A.2]$$

$$\underset{(T+1)\times(T+1)}{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \sigma^2/\nu & \mathbf{0}' \\ \mathbf{0} & \sigma^2\mathbf{I}_T \end{bmatrix}. \qquad [12.A.3]$$

The goal is to rearrange $\boldsymbol{\alpha}$ so that μ appears only in the first element. Define

$$\underset{(T+1)\times(T+1)}{\mathbf{A}} \equiv \begin{bmatrix} \nu/(\nu + T) & -\mathbf{1}'/(\nu + T) \\ \mathbf{1} & \mathbf{I}_T \end{bmatrix}. \qquad [12.A.4]$$

Since $\mathbf{1}'\mathbf{1} = T$ and $\mathbf{1}'\mathbf{y} = T\bar{y}$, we have

$$\begin{aligned} \mathbf{A}\boldsymbol{\alpha} &= \begin{bmatrix} [\nu/(\nu + T)](\mu - m) - \mathbf{1}'\mathbf{y}/(\nu + T) + [T/(\nu + T)]\mu \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix} \\ &= \begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix} \\ &\equiv \boldsymbol{\alpha}^* \end{aligned} \qquad [12.A.5]$$

and

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' &= \sigma^2\begin{bmatrix} 1/(\nu + T) & -\mathbf{1}'/(\nu + T) \\ \mathbf{1}/\nu & \mathbf{I}_T \end{bmatrix}\begin{bmatrix} \nu/(\nu + T) & \mathbf{1}' \\ -\mathbf{1}/(\nu + T) & \mathbf{I}_T \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{bmatrix} \\ &\equiv \boldsymbol{\Sigma}^*. \end{aligned} \qquad [12.A.6]$$

Thus,

$$\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha} = \boldsymbol{\alpha}'\mathbf{A}'(\mathbf{A}')^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{A}^{-1}\mathbf{A}\boldsymbol{\alpha} = (\mathbf{A}\boldsymbol{\alpha})'(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')^{-1}(\mathbf{A}\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{*\prime}(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\alpha}^*. \qquad [12.A.7]$$

Moreover, observe that $\mathbf{A}$ can be expressed as

$$\mathbf{A} = \begin{bmatrix} 1 & -\mathbf{1}'/(\nu + T) \\ \mathbf{0} & \mathbf{I}_T \end{bmatrix}\begin{bmatrix} 1 & \mathbf{0}' \\ \mathbf{1} & \mathbf{I}_T \end{bmatrix}.$$

Each of these triangular matrices has 1s along the principal diagonal and so has unit determinant, implying that $|\mathbf{A}| = 1$. Hence,

$$|\boldsymbol{\Sigma}^*| = |\mathbf{A}|\cdot|\boldsymbol{\Sigma}|\cdot|\mathbf{A}'| = |\boldsymbol{\Sigma}|. \qquad [12.A.8]$$

Substituting [12.A.5] through [12.A.8] into [12.A.1] gives

$$\begin{aligned} &f(\mathbf{y}, \mu; \sigma^2) \\ &= \frac{1}{(2\pi)^{(T+1)/2}} |\boldsymbol{\Sigma}^*|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}^{*\prime}(\boldsymbol{\Sigma}^*)^{-1}\boldsymbol{\alpha}^*\right\} \\ &= \frac{1}{(2\pi)^{(T+1)/2}} \begin{vmatrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{vmatrix}^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}\begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix}'\begin{bmatrix} \sigma^2/(\nu + T) & \mathbf{0}' \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu) \end{bmatrix}^{-1}\right. \\ &\quad \left.\times \begin{bmatrix} \mu - m^* \\ \mathbf{y} - m\cdot\mathbf{1} \end{bmatrix}\right\} \end{aligned} \qquad [12.A.9]$$

$$= \frac{1}{(2\pi)^{(T+1)/2}} \left[\frac{\sigma^2}{\nu + T}\right]^{-1/2} \cdot \left|\sigma^2(\mathbf{I}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)\right|^{-1/2}$$

$$\times \exp\left\{\frac{-(\mu - m^*)^2}{2\sigma^2/(\nu + T)} - \frac{(\mathbf{y} - m\cdot\mathbf{1})'(\mathbf{1}_T + \mathbf{1}\cdot\mathbf{1}'/\nu)^{-1}(\mathbf{y} - m\cdot\mathbf{1})}{2\sigma^2}\right\},$$

from which the factorization in Proposition 12.1 follows immediately. ■

■ **Proof of Proposition 12.2.** The product of [12.1.10] and [12.1.11] can be written as

$$f(\mathbf{y}, \boldsymbol{\beta}|\mathbf{X}; \sigma^2) = \frac{1}{(2\pi)^{(T+k)/2}} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\alpha}\right\}$$

with

$$\underset{(T+k)\times 1}{\boldsymbol{\alpha}} \equiv \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \end{bmatrix}$$

$$\underset{(T+k)\times(T+k)}{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \sigma^2\mathbf{M} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I}_T \end{bmatrix}.$$

As in the proof of Proposition 12.1, define

$$\underset{(T+k)\times(T+k)}{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_k & -(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{0} & \mathbf{I}_T \end{bmatrix} \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{X} & \mathbf{I}_T \end{bmatrix}$$

$$= \begin{bmatrix} (\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{-1} & -(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ \mathbf{X} & \mathbf{I}_T \end{bmatrix}.$$

Thus, $\mathbf{A}$ has unit determinant and

$$\mathbf{A}\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}$$

with

$$\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \begin{bmatrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{bmatrix}.$$

Thus, as in equation [12.A.9],

$$f(\mathbf{y}, \boldsymbol{\beta}|\mathbf{X}; \sigma^2) = \frac{1}{(2\pi)^{(T+k)/2}} \begin{vmatrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{vmatrix}^{-1/2}$$

$$\times \exp\left\{-\frac{1}{2}\begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}' \begin{bmatrix} \sigma^2(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma^2(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}') \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\beta} - \mathbf{m}^* \\ \mathbf{y} - \mathbf{X}\mathbf{m} \end{bmatrix}\right\}. \quad ■$$

■ **Proof of Proposition 12.3(a).** We have that

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) = f(\mathbf{y}|\boldsymbol{\beta}, \sigma^{-2}, \mathbf{X})\cdot f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{X})\cdot f(\sigma^{-2}|\mathbf{X}). \qquad [12.A.10]$$

The first two terms on the right side are identical to [12.1.10] and [12.1.11]. Thus, Proposition 12.2 can be used to write [12.A.10] as

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X})$$

$$= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}} |\mathbf{M}^*|^{1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\}$$

$$\times \left\{\frac{1}{(2\pi\sigma^2)^{T/2}} |\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}'|^{-1/2}\right. \qquad [12.A.11]$$

$$\left.\times \exp\left\{[-1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\mathbf{m})'(\mathbf{I}_T + \mathbf{X}\mathbf{M}\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\mathbf{m})\right\}\right\}$$

$$\times \left\{\frac{(\lambda/2)^{N/2}\sigma^{-2[(N/2)-1]} \exp[-\lambda\sigma^{-2}/2]}{\Gamma(N/2)}\right\}.$$

Define

$$\lambda^* \equiv \lambda + (\mathbf{y} - \mathbf{Xm})'(\mathbf{I}_T + \mathbf{XMX}')^{-1}(\mathbf{y} - \mathbf{Xm}); \qquad [12.A.12]$$

we will show later that this is the same as the value λ^* described in the proposition. For $N^* \equiv N + T$, the density [12.A.11] can be written as

$$\begin{aligned} f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) &= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\ &\quad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)}|\mathbf{I}_T + \mathbf{XMX}'|^{-1/2}\exp\left[-\frac{\lambda^*\sigma^{-2}}{2}\right]\right\} \\ &= \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2}\exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\ &\quad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)}\exp\left[-\frac{\lambda^*\sigma^{-2}}{2}\right]\right\} \\ &\quad \times \left\{\frac{\Gamma(N^*/2)(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)(\lambda^*/2)^{N^*/2}}|\mathbf{I}_T + \mathbf{XMX}'|^{-1/2}\right\}. \end{aligned} \qquad [12.A.13]$$

The second term does not involve $\boldsymbol{\beta}$, and the third term does not involve $\boldsymbol{\beta}$ or σ^{-2}. Thus, [12.A.13] provides the factorization

$$f(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2}|\mathbf{X}) = \{f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})\}\cdot\{f(\sigma^{-2}|\mathbf{y}, \mathbf{X})\}\cdot\{f(\mathbf{y}|\mathbf{X})\},$$

where $f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})$ is a $N(\mathbf{m}^*, \sigma^2\mathbf{M}^*)$ density, $f(\sigma^{-2}|\mathbf{y}, \mathbf{X})$ is a $\Gamma(N^*, \lambda^*)$ density, and $f(\mathbf{y}|\mathbf{X})$ can be written as

$$\begin{aligned} f(\mathbf{y}|\mathbf{X}) &= \left\{\frac{\Gamma(N^*/2)(\lambda/2)^{N/2}}{(2\pi)^{T/2}\Gamma(N/2)(\lambda^*/2)^{N^*/2}}|\mathbf{I}_T + \mathbf{XMX}'|^{-1/2}\right\} \\ &= \left\{\frac{\Gamma[(N + T)/2]\lambda^{N/2}|\mathbf{I}_T + \mathbf{XMX}'|^{-1/2}}{\pi^{T/2}\Gamma(N/2)\{\lambda + (\mathbf{y} - \mathbf{Xm})'(\mathbf{I}_T + \mathbf{XMX}')^{-1}(\mathbf{y} - \mathbf{Xm})\}^{(N+T)/2}}\right\} \\ &= c\cdot\{1 + (1/N)(\mathbf{y} - \mathbf{Xm})'[(\lambda/N)(\mathbf{I}_T + \mathbf{XMX}')]^{-1}(\mathbf{y} - \mathbf{Xm})\}^{-(N+T)/2}, \end{aligned}$$

where

$$c = \frac{\Gamma[(N + T)/2](1/N)^{T/2}|(\lambda/N)(\mathbf{I}_T + \mathbf{XMX}')|^{-1/2}}{\pi^{T/2}\Gamma(N/2)}.$$

Thus, $f(\mathbf{y}|\mathbf{X})$ is a T-dimensional Student's t density with N degrees of freedom, mean $\mathbf{Xm}$, and scale matrix $(\lambda/N)(\mathbf{I}_T + \mathbf{XM}'\mathbf{X}')$. Hence, the distributions of $(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X})$ and $(\sigma^{-2}|\mathbf{y}, \mathbf{X})$ are as claimed in Proposition 12.3, provided that the magnitude λ^* defined in [12.A.12] is the same as the expression in [12.1.28]. To verify that this is indeed the case, notice that

$$(\mathbf{I}_T + \mathbf{XMX}')^{-1} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}', \qquad [12.A.14]$$

as can be verified by premultiplying [12.A.14] by $(\mathbf{I}_T + \mathbf{XMX}')$:

$$\begin{aligned} &(\mathbf{I}_T + \mathbf{XMX}')[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'] \\ &\quad = \mathbf{I}_T + \mathbf{XMX}' - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' - \mathbf{XM}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' \\ &\quad = \mathbf{I}_T + \mathbf{X}\left\{\mathbf{M}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1}) - \mathbf{I}_k - \mathbf{M}(\mathbf{X}'\mathbf{X})\right\}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}' \\ &\quad = \mathbf{I}_T. \end{aligned}$$

Using [12.A.14], we see that

$$\begin{aligned} &(\mathbf{y} - \mathbf{Xm})'(\mathbf{I}_T + \mathbf{XMX}')^{-1}(\mathbf{y} - \mathbf{Xm}) \\ &\quad = (\mathbf{y} - \mathbf{Xm})'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'](\mathbf{y} - \mathbf{Xm}) \\ &\quad = (\mathbf{y} - \mathbf{Xb} + \mathbf{Xb} - \mathbf{Xm})'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'](\mathbf{y} - \mathbf{Xb} + \mathbf{Xb} - \mathbf{Xm}) \\ &\quad = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) + (\mathbf{b} - \mathbf{m})'\mathbf{X}'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}']\mathbf{X}(\mathbf{b} - \mathbf{m}), \end{aligned} \qquad [12.A.15]$$

where cross-product terms have disappeared because of the *OLS* orthogonality condition $(\mathbf{y} - \mathbf{Xb})'\mathbf{X} = \mathbf{0}'$. Furthermore,

$$\begin{aligned}
&\mathbf{X}'[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}']\mathbf{X} \\
&\quad = [\mathbf{I}_k - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}]\mathbf{X}'\mathbf{X} \\
&\quad = [(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1} - (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}]\mathbf{X}'\mathbf{X} \\
&\quad = \mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}.
\end{aligned}$$

This allows [12.A.15] to be written as

$$\begin{aligned}
&(\mathbf{y} - \mathbf{Xm})'(\mathbf{I}_T + \mathbf{XMX}')^{-1}(\mathbf{y} - \mathbf{Xm}) \\
&\qquad = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) + (\mathbf{b} - \mathbf{m})'\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X} + \mathbf{M}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{m}),
\end{aligned}$$

establishing the equivalence of [12.A.12] and [12.1.28].

Proof of (b). The joint posterior density of $\boldsymbol{\beta}$ and σ^{-2} is given by

$$\begin{aligned}
&f(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&\quad = f(\boldsymbol{\beta}|\sigma^{-2}, \mathbf{y}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&\quad = \left\{\frac{1}{(2\pi\sigma^2)^{k/2}}|\mathbf{M}^*|^{-1/2} \exp\left\{\left[-\frac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m}^*)'(\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)\right\}\right\} \\
&\qquad \times \left\{\frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)} \exp[-\lambda^*\sigma^{-2}/2]\right\} \\
&\quad = \left(\frac{\sigma^{-2[((k+N^*)/2)-1]}}{\Gamma[(k + N^*)/2]} \times \left\{\frac{\lambda^*}{2}\cdot[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]\right\}^{(k+N^*)/2}\right. \\
&\qquad \left. \times \exp\left\{-\frac{\lambda^*}{2}\cdot[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]\sigma^{-2}\right\}\right) \\
&\qquad \times \left\{\frac{\Gamma[(k + N^*)/2]}{(\lambda^*)^{k/2}\pi^{k/2}\Gamma(N^*/2)}|\mathbf{M}^*|^{-1/2}[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2}\right\} \\
&\quad = \{f(\sigma^{-2}|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})\} \cdot \{f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})\},
\end{aligned}$$

where $f(\sigma^{-2}|\boldsymbol{\beta}, \mathbf{y}, \mathbf{X})$ will be recognized as a $\Gamma((k + N^*), \lambda^*[1 + (\boldsymbol{\beta} - \mathbf{m}^*)'(\lambda^*\mathbf{M}^*)^{-1} \times (\boldsymbol{\beta} - \mathbf{m}^*)])$ density, while $f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})$ can be written as

$$\begin{aligned}
f(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \left\{\frac{\Gamma[(k + N^*)/2]}{(N^*)^{k/2}\pi^{k/2}\Gamma(N^*/2)}|(\lambda^*/N^*)\mathbf{M}^*|^{-1/2}\right. \\
\left. \times [1 + (1/N^*)(\boldsymbol{\beta} - \mathbf{m}^*)'[(\lambda^*/N^*)\mathbf{M}^*]^{-1}(\boldsymbol{\beta} - \mathbf{m}^*)]^{-(k+N^*)/2}\right\},
\end{aligned}$$

which is a k-dimensional t density with N^* degrees of freedom, mean $\mathbf{m}^*$, and scale matrix $(\lambda^*/N^*)\mathbf{M}^*$.

Proof of (c). Notice that conditional on $\mathbf{y}$, $\mathbf{X}$, and σ^2, the variable

$$Z = [\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]'[\sigma^2\mathbf{R}(\mathbf{M}^{-1} + \mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \cdot [\mathbf{R}(\boldsymbol{\beta} - \mathbf{m}^*)]$$

is distributed $\chi^2(m)$, from Proposition 8.1. The variable Q in [12.1.30] is equal to $Z \cdot \sigma^2 N^*/(m\lambda^*)$, and so conditional on $\mathbf{y}$, $\mathbf{X}$, and σ^2, the variable Q is distributed $\Gamma(m, (m\lambda^*)/(\sigma^2 N^*))$:

$$f(q|\sigma^{-2}, \mathbf{y}, \mathbf{X}) = \frac{[m\lambda^*/(2\sigma^2 N^*)]^{m/2} q^{[(m/2)-1]} \exp[-m\lambda^* q/(2\sigma^2 N^*)]}{\Gamma(m/2)}. \qquad [12.A.16]$$

The joint posterior density of q and σ^{-2} is

$$
\begin{aligned}
f(q, \sigma^{-2}|\mathbf{y}, \mathbf{X}) &= f(q|\sigma^{-2}, \mathbf{y}, \mathbf{X}) \cdot f(\sigma^{-2}|\mathbf{y}, \mathbf{X}) \\
&= \left\{ \frac{[m\lambda^*/(2\sigma^2 N^*)]^{m/2} q^{[(m/2)-1]} \exp[-m\lambda^* q/(2\sigma^2 N^*)]}{\Gamma(m/2)} \right\} \\
&\quad \times \left\{ \frac{\sigma^{-2[(N^*/2)-1]}(\lambda^*/2)^{N^*/2}}{\Gamma(N^*/2)} \exp[-\lambda^* \sigma^{-2}/2] \right\} \\
&= \left\{ \frac{\{(N^* + mq) \cdot [\lambda^*/(2N^*)]\}^{[(N^*+m)/2]}}{\Gamma[(N^* + m)/2]} \right. \qquad [12.A.17] \\
&\quad \left. \times \sigma^{-2[(m+N^*)/2]-1} \exp[-(N^* + mq)(\lambda^*/N^*)\sigma^{-2}/2] \right\} \\
&\quad \times \left\{ \frac{m^{m/2}(N^*)^{N^*/2}\Gamma[(N^* + m)/2] q^{[(m/2)-1]}}{\Gamma(m/2)\Gamma(N^*/2)(N^* + mq)^{[(N^*+m)/2]}} \right\} \\
&= \{f(\sigma^{-2}|q, \mathbf{y}, \mathbf{X})\} \cdot \{f(q|\mathbf{y}, \mathbf{X})\},
\end{aligned}
$$

where $f(\sigma^{-2}|q, \mathbf{y}, \mathbf{X})$ is a $\Gamma\big((m + N^*), (N^* + mq)(\lambda^*/N^*)\big)$ density and $f(q|\mathbf{y}, \mathbf{X})$ is an $F(m, N^*)$ density. ■

Chapter 12 Exercise

12.1. Deduce Proposition 12.1 as a special case of Proposition 12.2.

Chapter 12 References

Almon, Shirley. 1965. "The Distributed Lag between Capital Appropriations and Expenditures." *Econometrica* 33:178–96.

Barro, Robert J. 1979. "On the Determination of the Public Debt." *Journal of Political Economy* 87:940–71.

DeGroot, Morris H. 1970. *Optimal Statistical Decisions*. New York: McGraw-Hill.

Diebold, Francis X., and James A. Nason. 1990. "Nonparametric Exchange Rate Prediction?" *Journal of International Economics* 28:315–32.

Doan, Thomas A. 1990. *RATS User's Manual*. VAR Econometrics, Suite 612, 1800 Sherman Ave., Evanston, IL 60201.

Fama, Eugene F. 1965. "The Behavior of Stock Market Prices." *Journal of Business* 38:34–105.

Geweke, John. 1988a. "The Secular and Cyclical Behavior of Real GDP in 19 OECD Countries, 1957–1983." *Journal of Business and Economic Statistics* 6:479–86.

———. 1988b. "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference." *Journal of Econometrics* 38:73–89.

———. 1989. "Bayesian Inference in Econometric Models Using Monte Carlo Integration." *Econometrica* 57:1317–39.

Hall, Robert E. 1978. "Stochastic Implications of the Life Cycle–Permanent Income Hypothesis: Theory and Evidence." *Journal of Political Economy* 86:971–87.

Hammersley, J. M., and D. C. Handscomb. 1964. *Monte Carlo Methods*, 1st ed. London: Methuen.

Hoerl, A. E., and R. W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12:55–82.

Kloek, T., and H. K. van Dijk. 1978. "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo." *Econometrica* 46:1–19.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.

Litterman, Robert B. 1986. "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience." *Journal of Business and Economic Statistics* 4:25–38.

Mankiw, N. Gregory. 1987. "The Optimal Collection of Seigniorage: Theory and Evidence." *Journal of Monetary Economics* 20:327–41.

Rothenberg, Thomas J. 1973. *Efficient Estimation with A Priori Information*. New Haven, Conn.: Yale University Press.

Shiller, Robert J. 1973. "A Distributed Lag Estimator Derived from Smoothness Priors." *Econometrica* 41:775–88.

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley.

Tierney, Luke, and Joseph B. Kadane. 1986. "Accurate Approximations for Posterior Moments and Marginal Densities." *Journal of the American Statistical Association* 81:82–86.

Zellner, Arnold. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

13 The Kalman Filter

This chapter introduces some very useful tools named for the contributions of R. E. Kalman (1960, 1963). The idea is to express a dynamic system in a particular form called the *state-space representation*. The Kalman filter is an algorithm for sequentially updating a linear projection for the system. Among other benefits, this algorithm provides a way to calculate exact finite-sample forecasts and the exact likelihood function for Gaussian *ARMA* processes, to factor matrix autocovariance-generating functions or spectral densities, and to estimate vector autoregressions with coefficients that change over time.

Section 13.1 describes how a dynamic system can be written in a form that can be analyzed using the Kalman filter. The filter itself is derived in Section 13.2, and its use in forecasting is described in Section 13.3. Section 13.4 explains how to estimate the population parameters by maximum likelihood. Section 13.5 analyzes the properties of the Kalman filter as the sample size grows, and explains how the Kalman filter is related in the limit to the Wold representation and factoring an autocovariance-generating function. Section 13.6 develops a smoothing algorithm, which is a way to use all the information in the sample to form the best inference about the unobserved state of the process at any historical date. Section 13.7 describes standard errors for smoothed inferences and forecasts. The use of the Kalman filter for estimating systems with time-varying parameters is investigated in Section 13.8.

13.1. *The State-Space Representation of a Dynamic System*

Maintained Assumptions

Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector of variables observed at date t. A rich class of dynamic models for $\mathbf{y}_t$ can be described in terms of a possibly unobserved $(r \times 1)$ vector $\boldsymbol{\xi}_t$ known as the *state vector*. The *state-space representation* of the dynamics of $\mathbf{y}$ is given by the following system of equations:

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1} \qquad [13.1.1]$$

$$\mathbf{y}_t = \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\boldsymbol{\xi}_t + \mathbf{w}_t, \qquad [13.1.2]$$

where $\mathbf{F}$, $\mathbf{A}'$, and $\mathbf{H}'$ are matrices of parameters of dimension $(r \times r)$, $(n \times k)$, and $(n \times r)$, respectively, and $\mathbf{x}_t$ is a $(k \times 1)$ vector of exogenous or predetermined variables. Equation [13.1.1] is known as the *state equation*, and [13.1.2] is known

as the *observation equation*. The $(r \times 1)$ vector $\mathbf{v}_t$ and the $(n \times 1)$ vector $\mathbf{w}_t$ are vector white noise:

$$E(\mathbf{v}_t\mathbf{v}_\tau') = \begin{cases} \mathbf{Q} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise} \end{cases} \qquad [13.1.3]$$

$$E(\mathbf{w}_t\mathbf{w}_\tau') = \begin{cases} \mathbf{R} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise,} \end{cases} \qquad [13.1.4]$$

where $\mathbf{Q}$ and $\mathbf{R}$ are $(r \times r)$ and $(n \times n)$ matrices, respectively. The disturbances $\mathbf{v}_t$ and $\mathbf{w}_t$ are assumed to be uncorrelated at all lags:

$$E(\mathbf{v}_t\mathbf{w}_\tau') = \mathbf{0} \qquad \text{for all } t \text{ and } \tau. \qquad [13.1.5]$$

The statement that $\mathbf{x}_t$ is predetermined or exogenous means that $\mathbf{x}_t$ provides no information about $\boldsymbol{\xi}_{t+s}$ or $\mathbf{w}_{t+s}$ for $s = 0, 1, 2, \ldots$ beyond that contained in $\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1$. Thus, for example, $\mathbf{x}_t$ could include lagged values of $\mathbf{y}$ or variables that are uncorrelated with $\boldsymbol{\xi}_\tau$ and $\mathbf{w}_\tau$ for all τ.

The system of [13.1.1] through [13.1.5] is typically used to describe a finite series of observations $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ for which assumptions about the initial value of the state vector $\boldsymbol{\xi}_1$ are needed. We assume that $\boldsymbol{\xi}_1$ is uncorrelated with any realizations of $\mathbf{v}_t$ or $\mathbf{w}_t$:

$$E(\mathbf{v}_t\boldsymbol{\xi}_1') = \mathbf{0} \qquad \text{for } t = 1, 2, \ldots, T \qquad [13.1.6]$$

$$E(\mathbf{w}_t\boldsymbol{\xi}_1') = \mathbf{0} \qquad \text{for } t = 1, 2, \ldots, T. \qquad [13.1.7]$$

The state equation [13.1.1] implies that $\boldsymbol{\xi}_t$ can be written as a linear function of $(\boldsymbol{\xi}_1, \mathbf{v}_2, \mathbf{v}_3, \ldots, \mathbf{v}_t)$:

$$\boldsymbol{\xi}_t = \mathbf{v}_t + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{F}^2\mathbf{v}_{t-2} + \cdots + \mathbf{F}^{t-2}\mathbf{v}_2 + \mathbf{F}^{t-1}\boldsymbol{\xi}_1 \qquad [13.1.8]$$
$$\text{for } t = 2, 3, \ldots, T.$$

Thus, [13.1.6] and [13.1.3] imply that $\mathbf{v}_t$ is uncorrelated with lagged values of $\boldsymbol{\xi}$:

$$E(\mathbf{v}_t\boldsymbol{\xi}_\tau') = \mathbf{0} \qquad \text{for } \tau = t-1, t-2, \ldots, 1. \qquad [13.1.9]$$

Similarly,

$$E(\mathbf{w}_t\boldsymbol{\xi}_\tau') = \mathbf{0} \qquad \text{for } \tau = 1, 2, \ldots, T \qquad [13.1.10]$$

$$\begin{aligned} E(\mathbf{w}_t\mathbf{y}_\tau') &= E[\mathbf{w}_t(\mathbf{A}'\mathbf{x}_\tau + \mathbf{H}'\boldsymbol{\xi}_\tau + \mathbf{w}_\tau)'] \\ &= \mathbf{0} \qquad \text{for } \tau = t-1, t-2, \ldots, 1 \end{aligned} \qquad [13.1.11]$$

$$E(\mathbf{v}_t\mathbf{y}_\tau') = \mathbf{0} \qquad \text{for } \tau = t-1, t-2, \ldots, 1. \qquad [13.1.12]$$

The system of [13.1.1] through [13.1.7] is quite flexible, though it is straightforward to generalize the results further to systems in which $\mathbf{v}_t$ is correlated with $\mathbf{w}_t$.[1] The various parameter matrices ($\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, or $\mathbf{R}$) could be functions of time, as will be discussed in Section 13.8. The presentation will be clearest, however, if we focus on the basic form in [13.1.1] through [13.1.7].

[1]See, for example, Anderson and Moore (1979, pp. 105–8).

Examples of State-Space Representations

Consider a univariate $AR(p)$ process,

$$y_{t+1} - \mu = \phi_1(y_t - \mu) + \phi_2(y_{t-1} - \mu) + \cdots + \phi_p(y_{t-p+1} - \mu) + \varepsilon_{t+1}, \qquad [13.1.13]$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

This could be written in state-space form as follows:

State Equation ($r = p$):

$$\begin{bmatrix} y_{t+1} - \mu \\ y_t - \mu \\ \vdots \\ y_{t-p+2} - \mu \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad [13.1.14]$$

Observation Equation ($n = 1$):

$$y_t = \mu + [1 \quad 0 \quad \cdots \quad 0] \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix}. \qquad [13.1.15]$$

That is, we would specify

$$\boldsymbol{\xi}_t = \begin{bmatrix} y_t - \mu \\ y_{t-1} - \mu \\ \vdots \\ y_{t-p+1} - \mu \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

$$\mathbf{v}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\mathbf{y}_t = y_t \qquad \mathbf{A}' = \mu \qquad \mathbf{x}_t = 1$$

$$\mathbf{H}' = [1 \quad 0 \quad \cdots \quad 0] \qquad \mathbf{w}_t = 0 \qquad \mathbf{R} = 0.$$

Note that the state equation here is simply the first-order vector difference equation introduced in equation [1.2.5]; $\mathbf{F}$ is the same matrix appearing in equation [1.2.3]. The observation equation here is a trivial identity. Thus, we have already seen that the state-space representation [13.1.14] and [13.1.15] is just another way of summarizing the $AR(p)$ process [13.1.13]. The reason for rewriting an $AR(p)$ process in such a form was to obtain a convenient summary of the system's dynamics, and this is the basic reason to be interested in the state-space representation of any system. The analysis of a vector autoregression using equation [10.1.11] employed a similar state-space representation.

As another example, consider a univariate $MA(1)$ process,

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}. \qquad [13.1.16]$$

This could be written in state-space form as follows:

State Equation ($r = 2$):

$$\begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix} \qquad [13.1.17]$$

Observation Equation ($n = 1$):

$$y_t = \mu + [1 \quad \theta] \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix}; \qquad [13.1.18]$$

that is,

$$\xi_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad \mathbf{v}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \qquad \mathbf{y}_t = y_t \qquad \mathbf{A}' = \mu \qquad \mathbf{x}_t = 1$$

$$\mathbf{H}' = [1 \quad \theta] \qquad \mathbf{w}_t = 0 \qquad \mathbf{R} = 0.$$

There are many ways to write a given system in state-space form. For example, the $MA(1)$ process [13.1.16] can also be represented in this way:

State Equation ($r = 2$):

$$\begin{bmatrix} \varepsilon_{t+1} + \theta\varepsilon_t \\ \theta\varepsilon_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_t + \theta\varepsilon_{t-1} \\ \theta\varepsilon_t \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ \theta\varepsilon_{t+1} \end{bmatrix} \qquad [13.1.19]$$

Observation Equation ($n = 1$):

$$y_t = \mu + [1 \quad 0] \begin{bmatrix} \varepsilon_t + \theta\varepsilon_{t-1} \\ \theta\varepsilon_t \end{bmatrix}. \qquad [13.1.20]$$

Note that the original $MA(1)$ representation of [13.1.16], the first state-space representation of [13.1.17] and [13.1.18], and the second state-space representation of [13.1.19] and [13.1.20] all characterize the same process. We will obtain the identical forecasts of the process or value for the likelihood function from any of the three representations and can feel free to work with whichever is most convenient.

More generally, a univariate $ARMA(p, q)$ process can be written in state-space form by defining $r \equiv \max\{p, q + 1\}$:

$$\begin{aligned} y_t - \mu = {} & \phi_1(y_{t-1} - \mu) + \phi_2(y_{t-2} - \mu) + \cdots + \phi_r(y_{t-r} - \mu) \\ & + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_{r-1}\varepsilon_{t-r+1}, \end{aligned} \qquad [13.1.21]$$

where we interpret $\phi_j = 0$ for $j > p$ and $\theta_j = 0$ for $j > q$. Consider the following state-space representation;

State Equation ($r = \max\{p, q + 1\}$):

$$\xi_{t+1} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_{r-1} & \phi_r \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \xi_t + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad [13.1.22]$$

Observation Equation ($n = 1$):

$$y_t = \mu + [1 \quad \theta_1 \quad \theta_2 \quad \cdots \quad \theta_{r-1}]\xi_t. \qquad [13.1.23]$$

To verify that [13.1.22] and [13.1.23] describe the same process as [13.1.21], let ξ_{jt} denote the jth element of ξ_t. Thus, the second row of the state equation asserts

that

$$\xi_{2,t+1} = \xi_{1t}.$$

The third row asserts that

$$\xi_{3,t+1} = \xi_{2t} = \xi_{1,t-1},$$

and in general the jth row implies that

$$\xi_{j,t+1} = L^{j-1}\xi_{1,t+1}.$$

Thus, the first row of the state equation implies that

$$\xi_{1,t+1} = (\phi_1 + \phi_2 L + \phi_3 L^2 + \cdots + \phi_r L^{r-1})\xi_{1t} + \varepsilon_{t+1}$$

or

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_r L^r)\xi_{1,t+1} = \varepsilon_{t+1}. \qquad [13.1.24]$$

The observation equation states that

$$y_t = \mu + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_{r-1} L^{r-1})\xi_{1t}. \qquad [13.1.25]$$

Multiplying [13.1.25] by $(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_r L^r)$ and using [13.1.24] gives

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_r L^r)(y_t - \mu)$$
$$= (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_{r-1} L^{r-1})\varepsilon_t,$$

which indeed reproduces [13.1.21].

The state-space form can also be very convenient for modeling sums of stochastic processes or the consequences of measurement error. For example, Fama and Gibbons (1982) wanted to study the behavior of the ex ante real interest rate (the nominal interest rate i_t minus the expected inflation rate π_t^e). This variable is unobserved, because the econometrician does not have data on the rate of inflation anticipated by the bond market. Thus, the state variable for this application was the scalar $\xi_t = i_t - \pi_t^e - \mu$, where μ denotes the average ex ante real interest rate. Fama and Gibbons assumed that the ex ante real rate follows an $AR(1)$ process:

$$\xi_{t+1} = \phi\xi_t + v_{t+1}. \qquad [13.1.26]$$

The econometrician has observations on the ex post real rate (the nominal interest rate i_t minus actual inflation π_t), which can be written as

$$i_t - \pi_t = (i_t - \pi_t^e) + (\pi_t^e - \pi_t) = \mu + \xi_t + w_t, \qquad [13.1.27]$$

where $w_t \equiv (\pi_t^e - \pi_t)$ is the error that people make in forecasting inflation. If people form these forecasts optimally, then w_t should be uncorrelated with its own lagged values or with the ex ante real interest rate. Thus, [13.1.26] and [13.1.27] are the state equation and observation equation for a state-space model with $r = n = 1$, $\mathbf{F} = \phi$, $y_t = i_t - \pi_t$, $\mathbf{A}'\mathbf{x}_t = \mu$, $\mathbf{H} = 1$, and $w_t = (\pi_t^e - \pi_t)$.

In another interesting application of the state-space framework, Stock and Watson (1991) postulated the existence of an unobserved scalar C_t that represents the state of the business cycle. A set of n different observed macroeconomic variables $(y_{1t}, y_{2t}, \ldots, y_{nt})$ are each assumed to be influenced by the business cycle and also to have an idiosyncratic component (denoted χ_{it}) that is unrelated to movements in y_{jt} for $i \neq j$. If the business cycle and each of the idiosyncratic components could be described by univariate $AR(1)$ processes, then the $[(n + 1) \times 1]$ state vector would be

$$\xi_t = \begin{bmatrix} C_t \\ \chi_{1t} \\ \chi_{2t} \\ \vdots \\ \chi_{nt} \end{bmatrix} \qquad [13.1.28]$$

with state equation

$$\begin{bmatrix} C_{t+1} \\ \chi_{1,t+1} \\ \chi_{2,t+1} \\ \vdots \\ \chi_{n,t+1} \end{bmatrix} = \begin{bmatrix} \phi_C & 0 & 0 & \cdots & 0 \\ 0 & \phi_1 & 0 & \cdots & 0 \\ 0 & 0 & \phi_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \phi_n \end{bmatrix} \begin{bmatrix} C_t \\ \chi_{1t} \\ \chi_{2t} \\ \vdots \\ \chi_{nt} \end{bmatrix} + \begin{bmatrix} v_{C,t+1} \\ v_{1,t+1} \\ v_{2,t+1} \\ \vdots \\ v_{n,t+1} \end{bmatrix} \qquad [13.1.29]$$

and observation equation

$$\begin{bmatrix} y_{1t} \\ y_{2t} \\ \vdots \\ y_{nt} \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} + \begin{bmatrix} \gamma_1 & 1 & 0 & \cdots & 0 \\ \gamma_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_n & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} C_t \\ \chi_{1t} \\ \chi_{2t} \\ \vdots \\ \chi_{nt} \end{bmatrix}. \qquad [13.1.30]$$

Thus, γ_i is a parameter that describes the sensitivity of the ith series to the business cycle. To allow for pth-order dynamics, Stock and Watson replaced C_t and χ_{it} in [13.1.28] with the $(p \times 1)$ vectors $(C_t, C_{t-1}, \ldots, C_{t-p+1})'$ and $(\chi_{it}, \chi_{i,t-1}, \ldots, \chi_{i,t-p+1})'$ so that ξ_t is an $[(n + 1)p \times 1]$ vector. The scalars ϕ_i in [13.1.29] are then replaced by $(p \times p)$ matrices $\mathbf{F}_i$ with the structure of the matrix $\mathbf{F}$ in [13.1.14], and $[n \times (p - 1)]$ blocks of zeros are added between the columns of $\mathbf{H}'$ in the observation equation [13.1.30].

13.2. *Derivation of the Kalman Filter*

Overview of the Kalman Filter

Consider the general state-space system [13.1.1] through [13.1.7], whose key equations are reproduced here for convenience:

$$\underset{(r\times 1)}{\xi_{t+1}} = \underset{(r\times r)(r\times 1)}{\mathbf{F}\cdot\xi_t} + \underset{(r\times 1)}{\mathbf{v}_{t+1}} \qquad [13.2.1]$$

$$\underset{(n\times 1)}{\mathbf{y}_t} = \underset{(n\times k)(k\times 1)}{\mathbf{A}'\cdot\mathbf{x}_t} + \underset{(n\times r)(r\times 1)}{\mathbf{H}'\cdot\xi_t} + \underset{(n\times 1)}{\mathbf{w}_t} \qquad [13.2.2]$$

$$E(\mathbf{v}_t\mathbf{v}_\tau') = \begin{cases} \underset{(r\times r)}{\mathbf{Q}} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise} \end{cases} \qquad [13.2.3]$$

$$E(\mathbf{w}_t\mathbf{w}_\tau') = \begin{cases} \underset{(n\times n)}{\mathbf{R}} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise.} \end{cases} \qquad [13.2.4]$$

The analyst is presumed to have observed $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. One of the ultimate objectives may be to estimate the values of any unknown parameters in the system on the basis of these observations. For now, however, we will assume that the particular numerical values of $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ are known with certainty; Section 13.4 will give details on how these parameters can be estimated from the data.

There are many uses of the Kalman filter. It is motivated here as an algorithm for calculating linear least squares forecasts of the state vector on the basis of data observed through date t,

$$\hat{\xi}_{t+1|t} \equiv \hat{E}(\xi_{t+1}|\mathcal{Y}_t),$$

where

$$\mathcal{Y}_t \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_1', \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_1')' \qquad [13.2.5]$$

and $\hat{E}(\xi_{t+1}|\mathcal{Y}_t)$ denotes the linear projection of ξ_{t+1} on $\mathcal{Y}_t$ and a constant. The Kalman filter calculates these forecasts recursively, generating $\hat{\xi}_{1|0}, \hat{\xi}_{2|1}, \ldots,$

$\hat{\xi}_{T|T-1}$ in succession. Associated with each of these forecasts is a mean squared error (*MSE*) matrix, represented by the following $(r \times r)$ matrix:

$$\mathbf{P}_{t+1|t} \equiv E[(\xi_{t+1} - \hat{\xi}_{t+1|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'].$$ [13.2.6]

Starting the Recursion

The recursion begins with $\hat{\xi}_{1|0}$, which denotes a forecast of ξ_1 based on no observations of $\mathbf{y}$ or $\mathbf{x}$. This is just the unconditional mean of ξ_1,

$$\hat{\xi}_{1|0} = E(\xi_1),$$

with associated *MSE*

$$\mathbf{P}_{1|0} = E\{[\xi_1 - E(\xi_1)][\xi_1 - E(\xi_1)]'\}.$$

For example, for the state-space representation of the $MA(1)$ system given in [13.1.17] and [13.1.18], the state vector was

$$\xi_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix},$$

for which

$$\hat{\xi}_{1|0} = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$ [13.2.7]

$$\mathbf{P}_{1|0} = E\left(\begin{bmatrix} \varepsilon_1 \\ \varepsilon_0 \end{bmatrix}[\varepsilon_1 \quad \varepsilon_0]\right) = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix},$$ [13.2.8]

where $\sigma^2 = E(\varepsilon_t^2)$.

More generally, if eigenvalues of $\mathbf{F}$ are all inside the unit circle, then the process for ξ_t in [13.2.1] is covariance-stationary. The unconditional mean of ξ_t can be found by taking expectations of both sides of [13.2.1], producing

$$E(\xi_{t+1}) = \mathbf{F} \cdot E(\xi_t),$$

or, since ξ_t is covariance-stationary,

$$(\mathbf{I}_r - \mathbf{F}) \cdot E(\xi_t) = \mathbf{0}.$$

Since unity is not an eigenvalue of $\mathbf{F}$, the matrix $(\mathbf{I}_r - \mathbf{F})$ is nonsingular, and this equation has the unique solution $E(\xi_t) = \mathbf{0}$. The unconditional variance of ξ can similarly be found by postmultiplying [13.2.1] by its transpose and taking expectations:

$$E(\xi_{t+1}\xi_{t+1}') = E[(\mathbf{F}\xi_t + \mathbf{v}_{t+1})(\xi_t'\mathbf{F}' + \mathbf{v}_{t+1}')] = \mathbf{F} \cdot E(\xi_t\xi_t') \cdot \mathbf{F}' + E(\mathbf{v}_{t+1}\mathbf{v}_{t+1}').$$

Cross-product terms have disappeared in light of [13.1.9]. Letting $\boldsymbol{\Sigma}$ denote the variance-covariance matrix of ξ, this equation implies

$$\boldsymbol{\Sigma} = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}' + \mathbf{Q},$$

whose solution was seen in [10.2.18] to be given by

$$\text{vec}(\boldsymbol{\Sigma}) = [\mathbf{I}_{r^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q}).$$

Thus, in general, provided that the eigenvalues of $\mathbf{F}$ are inside the unit circle, the Kalman filter iterations can be started with $\hat{\xi}_{1|0} = \mathbf{0}$ and $\mathbf{P}_{1|0}$ the $(r \times r)$ matrix whose elements expressed as a column vector are given by

$$\text{vec}(\mathbf{P}_{1|0}) = [\mathbf{I}_{r^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q}).$$

If instead some eigenvalues of $\mathbf{F}$ are on or outside the unit circle, or if the initial state ξ_1 is not regarded as an arbitrary draw from the process implied by [13.2.1], then $\hat{\xi}_{1|0}$ can be replaced with the analyst's best guess as to the initial value of ξ_1, where $\mathbf{P}_{1|0}$ is a positive definite matrix summarizing the confidence in

this guess. Larger values for the diagonal elements of $\mathbf{P}_{1|0}$ register greater uncertainty about the true value of ξ_1.

Forecasting $\mathbf{y}_t$

Given starting values $\hat{\xi}_{1|0}$ and $\mathbf{P}_{1|0}$, the next step is to calculate analogous magnitudes for the following date, $\hat{\xi}_{2|1}$ and $\mathbf{P}_{2|1}$. The calculations for $t = 2, 3, \ldots, T$ all have the same basic form, so we will describe them in general terms for step t; given $\hat{\xi}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$, the goal is to calculate $\hat{\xi}_{t+1|t}$ and $\mathbf{P}_{t+1|t}$.

First note that since we have assumed that $\mathbf{x}_t$ contains no information about ξ_t beyond that contained in $\mathcal{Y}_{t-1}$,

$$\hat{E}(\xi_t|\mathbf{x}_t, \mathcal{Y}_{t-1}) = \hat{E}(\xi_t|\mathcal{Y}_{t-1}) = \hat{\xi}_{t|t-1}.$$

Next consider forecasting the value of $\mathbf{y}_t$:

$$\hat{\mathbf{y}}_{t|t-1} \equiv \hat{E}(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}).$$

Notice from [13.2.2] that

$$\hat{E}(\mathbf{y}_t|\mathbf{x}_t, \xi_t) = \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\xi_t,$$

and so, from the law of iterated projections,

$$\hat{\mathbf{y}}_{t|t-1} = \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\cdot\hat{E}(\xi_t|\mathbf{x}_t, \mathcal{Y}_{t-1}) = \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\hat{\xi}_{t|t-1}. \qquad [13.2.9]$$

From [13.2.2], the error of this forecast is

$$\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1} = \mathbf{A}'\mathbf{x}_t + \mathbf{H}'\xi_t + \mathbf{w}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\xi}_{t|t-1} = \mathbf{H}'(\xi_t - \hat{\xi}_{t|t-1}) + \mathbf{w}_t$$

with *MSE*

$$\begin{aligned} &E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'] \\ &\qquad = E[\mathbf{H}'(\xi_t - \hat{\xi}_{t|t-1})(\xi_t - \hat{\xi}_{t|t-1})'\mathbf{H}] + E[\mathbf{w}_t\mathbf{w}_t']. \end{aligned} \qquad [13.2.10]$$

Cross-product terms have disappeared, since

$$E[\mathbf{w}_t(\xi_t - \hat{\xi}_{t|t-1})'] = \mathbf{0}. \qquad [13.2.11]$$

To justify [13.2.11], recall from [13.1.10] that $\mathbf{w}_t$ is uncorrelated with ξ_t. Furthermore, since $\hat{\xi}_{t|t-1}$ is a linear function of $\mathcal{Y}_{t-1}$, by [13.1.11] it too must be uncorrelated with $\mathbf{w}_t$.

Using [13.2.4] and [13.2.6], equation [13.2.10] can be written

$$E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'] = \mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R}. \qquad [13.2.12]$$

Updating the Inference About ξ_t

Next the inference about the current value of ξ_t is updated on the basis of the observation of $\mathbf{y}_t$ to produce

$$\hat{\xi}_{t|t} = \hat{E}(\xi_t|\mathbf{y}_t, \mathbf{x}_t, \mathcal{Y}_{t-1}) = \hat{E}(\xi_t|\mathcal{Y}_t).$$

This can be evaluated using the formula for updating a linear projection, equation [4.5.30]:[2]

$$\begin{aligned} \hat{\xi}_{t|t} = \hat{\xi}_{t|t-1} &+ \{E[(\xi_t - \hat{\xi}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})']\} \\ &\times \{E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})']\}^{-1} \times (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1}). \end{aligned} \qquad [13.2.13]$$

[2]Here ξ_t corresponds to $\mathbf{Y}_3$, $\mathbf{y}_t$ corresponds to $\mathbf{Y}_2$, and $(\mathbf{x}_t', \mathcal{Y}_{t-1}')'$ corresponds to $\mathbf{Y}_1$ in equation [4.5.30].

But

$$
\begin{aligned}
&E\{(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})'\} \\
&\quad = E\{[\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1}][\mathbf{H}'(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1}) + \mathbf{w}_t]'\} \\
&\quad = E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})'\mathbf{H}] \\
&\quad = \mathbf{P}_{t|t-1}\mathbf{H}
\end{aligned}
\tag{13.2.14}
$$

by virtue of [13.2.11] and [13.2.6]. Substituting [13.2.14], [13.2.12], and [13.2.9] into [13.2.13] gives

$$
\hat{\boldsymbol{\xi}}_{t|t} = \hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}). \tag{13.2.15}
$$

The *MSE* associated with this updated projection, which is denoted $\mathbf{P}_{t|t}$, can be found from [4.5.31]:

$$
\begin{aligned}
\mathbf{P}_{t|t} &\equiv E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})'] \\
&= E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})'] \\
&\quad - \{E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})']\} \\
&\quad \times \{E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})']\}^{-1} \\
&\quad \times \{E[(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})']\} \\
&= \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}.
\end{aligned}
\tag{13.2.16}
$$

Producing a Forecast of $\boldsymbol{\xi}_{t+1}$

Next, the state equation [13.2.1] is used to forecast $\boldsymbol{\xi}_{t+1}$:

$$
\begin{aligned}
\hat{\boldsymbol{\xi}}_{t+1|t} &= \hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t) \\
&= \mathbf{F}\cdot\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_t) + \hat{E}(\mathbf{v}_{t+1}|\mathcal{Y}_t) \\
&= \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t} + \mathbf{0}.
\end{aligned}
\tag{13.2.17}
$$

Substituting [13.2.15] into [13.2.17],

$$
\begin{aligned}
\hat{\boldsymbol{\xi}}_{t+1|t} &= \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t-1} \\
&\quad + \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}).
\end{aligned}
\tag{13.2.18}
$$

The coefficient matrix in [13.2.18] is known as the *gain matrix* and is denoted $\mathbf{K}_t$:

$$
\mathbf{K}_t \equiv \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}, \tag{13.2.19}
$$

allowing [13.2.18] to be written

$$
\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}). \tag{13.2.20}
$$

The *MSE* of this forecast can be found from [13.2.17] and the state equation [13.2.1]:

$$
\begin{aligned}
\mathbf{P}_{t+1|t} &= E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})'] \\
&= E[(\mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1} - \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t})(\mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1} - \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t})'] \\
&= \mathbf{F}\cdot E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})']\cdot\mathbf{F}' + E[\mathbf{v}_{t+1}\mathbf{v}_{t+1}'] \\
&= \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q},
\end{aligned}
\tag{13.2.21}
$$

with cross-product terms again clearly zero. Substituting [13.2.16] into [13.2.21] produces

$$
\mathbf{P}_{t+1|t} = \mathbf{F}[\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}]\mathbf{F}' + \mathbf{Q}. \tag{13.2.22}
$$

Summary and Remarks

To summarize, the Kalman filter is started with the unconditional mean and variance of $\boldsymbol{\xi}_1$:

$$\hat{\boldsymbol{\xi}}_{1|0} = E(\boldsymbol{\xi}_1)$$
$$\mathbf{P}_{1|0} = E\{[\boldsymbol{\xi}_1 - E(\boldsymbol{\xi}_1)][\boldsymbol{\xi}_1 - E(\boldsymbol{\xi}_1)]'\}.$$

Typically, these are given by $\hat{\boldsymbol{\xi}}_{1|0} = \mathbf{0}$ and $\text{vec}(\mathbf{P}_{1|0}) = [\mathbf{I}_{r^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q})$. We then iterate on

$$\begin{aligned}\hat{\boldsymbol{\xi}}_{t+1|t} &= \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t-1} \\ &\quad + \mathbf{F}\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1})\end{aligned} \qquad [13.2.23]$$

and [13.2.22] for $t = 1, 2, \ldots, T$. The value $\hat{\boldsymbol{\xi}}_{t+1|t}$ denotes the best forecast of $\boldsymbol{\xi}_{t+1}$ based on a constant and a linear function of $(\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1)$. The matrix $\mathbf{P}_{t+1|t}$ gives the *MSE* of this forecast. The forecast of $\mathbf{y}_{t+1}$ is given by

$$\hat{\mathbf{y}}_{t+1|t} \equiv \hat{E}(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}, \mathcal{Y}_t) = \mathbf{A}'\mathbf{x}_{t+1} + \mathbf{H}'\hat{\boldsymbol{\xi}}_{t+1|t} \qquad [13.2.24]$$

with associated *MSE*

$$E[(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})'] = \mathbf{H}'\mathbf{P}_{t+1|t}\mathbf{H} + \mathbf{R}. \qquad [13.2.25]$$

It is worth noting that the recursion in [13.2.22] could be calculated without ever evaluating [13.2.23]. The values for $\mathbf{P}_{t|t-1}$ in [13.2.22] and $\mathbf{K}_t$ in [13.2.19] are not functions of the data, but instead are determined entirely by the population parameters of the process.

An alternative way of writing the recursion for $\mathbf{P}_{t+1|t}$ is sometimes useful. Subtracting the Kalman updating equation [13.2.20] from the state equation [13.2.1] produces

$$\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1}) - \mathbf{K}_t(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}) + \mathbf{v}_{t+1}. \qquad [13.2.26]$$

Further substituting the observation equation [13.2.2] into [13.2.26] results in

$$\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t} = (\mathbf{F} - \mathbf{K}_t\mathbf{H}')(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1}) - \mathbf{K}_t\mathbf{w}_t + \mathbf{v}_{t+1}. \qquad [13.2.27]$$

Postmultiplying [13.2.27] by its transpose and taking expectations,

$$\begin{aligned}&E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})'] \\ &\quad = (\mathbf{F} - \mathbf{K}_t\mathbf{H}')E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})'](\mathbf{F}' - \mathbf{H}\mathbf{K}_t') + \mathbf{K}_t\mathbf{R}\mathbf{K}_t' + \mathbf{Q};\end{aligned}$$

or, recalling the definition of $\mathbf{P}_{t+1|t}$ in equation [13.2.6],

$$\mathbf{P}_{t+1|t} = (\mathbf{F} - \mathbf{K}_t\mathbf{H}')\mathbf{P}_{t|t-1}(\mathbf{F}' - \mathbf{H}\mathbf{K}_t') + \mathbf{K}_t\mathbf{R}\mathbf{K}_t' + \mathbf{Q}. \qquad [13.2.28]$$

Equation [13.2.28] along with the definition of $\mathbf{K}_t$ in [13.2.19] will produce the same sequence generated by equation [13.2.22].

13.3. *Forecasts Based on the State-Space Representation*

The Kalman filter computations in [13.2.22] through [13.2.25] are normally calculated by computer, using the known numerical values of $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ along with the actual data. To help make the ideas more concrete, however, we now explore analytically the outcome of these calculations for a simple example.

Example—Using the Kalman Filter to Find Exact Finite-Sample Forecasts for an MA(*1*) *Process*

Consider again a state-space representation for the $MA(1)$ process:

State Equation ($r = 2$):

$$\begin{bmatrix} \varepsilon_{t+1} \\ \varepsilon_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix} \qquad [13.3.1]$$

Observation Equation ($n = 1$):

$$y_t = \mu + [1 \quad \theta] \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} \qquad [13.3.2]$$

$$\boldsymbol{\xi}_t = \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix} \qquad [13.3.3]$$

$$\mathbf{F} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \qquad [13.3.4]$$

$$\mathbf{v}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix} \qquad [13.3.5]$$

$$\mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix} \qquad [13.3.6]$$

$$\mathbf{y}_t = y_t \qquad [13.3.7]$$

$$\mathbf{A}' = \mu \qquad [13.3.8]$$

$$\mathbf{x}_t = 1 \qquad [13.3.9]$$

$$\mathbf{H}' = [1 \quad \theta] \qquad [13.3.10]$$

$$\mathbf{w}_t = 0 \qquad [13.3.11]$$

$$\mathbf{R} = 0. \qquad [13.3.12]$$

The starting values for the filter were described in [13.2.7] and [13.2.8]:

$$\hat{\boldsymbol{\xi}}_{1|0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\mathbf{P}_{1|0} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}.$$

Thus, from [13.2.24], the period 1 forecast is

$$\hat{y}_{1|0} = \mu + \mathbf{H}'\hat{\boldsymbol{\xi}}_{1|0} = \mu,$$

with MSE given by [13.2.25]:

$$E(y_1 - \hat{y}_{1|0})^2 = \mathbf{H}'\mathbf{P}_{1|0}\mathbf{H} + \mathbf{R} = [1 \quad \theta] \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 \\ \theta \end{bmatrix} + 0 = \sigma^2(1 + \theta^2).$$

These, of course, are just the unconditional mean and variance of y.

To see the structure of the recursion for $t = 2, 3, \ldots, T$, consider the basic form of the updating equation [13.2.23]. Notice that since the first row of $\mathbf{F}$ consists entirely of zeros, the first element of the vector $\hat{\boldsymbol{\xi}}_{t+1|t}$ will always equal zero, for all t. We see why if we recall the meaning of the state vector in [13.3.3]:

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \begin{bmatrix} \hat{\varepsilon}_{t+1|t} \\ \hat{\varepsilon}_{t|t} \end{bmatrix}. \qquad [13.3.13]$$

Naturally, the forecast of the future white noise, $\hat{\varepsilon}_{t+1|t}$, is always zero. The forecast of y_{t+1} is given by [13.2.24]:

$$\hat{y}_{t+1|t} = \mu + [1 \quad \theta]\begin{bmatrix} \hat{\varepsilon}_{t+1|t} \\ \hat{\varepsilon}_{t|t} \end{bmatrix} = \mu + \theta\hat{\varepsilon}_{t|t}. \qquad [13.3.14]$$

The Kalman filter updating equation for the *MSE*, equation [13.2.21], for this example becomes

$$\mathbf{P}_{t+1|t} = \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\mathbf{P}_{t|t}\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}. \qquad [13.3.15]$$

Thus, $\mathbf{P}_{t+1|t}$ is a diagonal matrix of the form

$$\mathbf{P}_{t+1|t} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & p_{t+1} \end{bmatrix}, \qquad [13.3.16]$$

where the (2, 2) element of $\mathbf{P}_{t+1|t}$ (which we have denoted by p_{t+1}) is the same as the (1, 1) element of $\mathbf{P}_{t|t}$. Recalling [13.2.6] and [13.3.13], this term has the interpretation as the *MSE* of $\hat{\varepsilon}_{t|t}$:

$$p_{t+1} = E(\varepsilon_t - \hat{\varepsilon}_{t|t})^2. \qquad [13.3.17]$$

The (1, 1) element of $\mathbf{P}_{t+1|t}$ has the interpretation as the *MSE* of $\hat{\varepsilon}_{t+1|t}$. We have seen that this forecast is always zero, and its *MSE* in [13.3.16] is σ^2 for all t. The fact that $\mathbf{P}_{t+1|t}$ is a diagonal matrix means that the forecast error $(\varepsilon_{t+1} - \hat{\varepsilon}_{t+1|t})$ is uncorrelated with $(\varepsilon_t - \hat{\varepsilon}_{t|t})$.

The *MSE* of the forecast of y_{t+1} is given by [13.2.25]:

$$\begin{aligned} E(y_{t+1} - \hat{y}_{t+1|t})^2 &= \mathbf{H}'\mathbf{P}_{t+1|t}\mathbf{H} + \mathbf{R} \\ &= [1 \quad \theta]\begin{bmatrix} \sigma^2 & 0 \\ 0 & p_{t+1} \end{bmatrix}\begin{bmatrix} 1 \\ \theta \end{bmatrix} + 0 \qquad [13.3.18] \\ &= \sigma^2 + \theta^2 p_{t+1}. \end{aligned}$$

Again, the intuition can be seen from the nature of the forecast in [13.3.14]:

$$\begin{aligned} E(y_{t+1} - \hat{y}_{t+1|t})^2 &= E[(\mu + \varepsilon_{t+1} + \theta\varepsilon_t) - (\mu + \theta\hat{\varepsilon}_{t|t})]^2 \\ &= E(\varepsilon_{t+1}^2) + \theta^2 E(\varepsilon_t - \hat{\varepsilon}_{t|t})^2, \end{aligned}$$

which, from [13.3.17], reproduces [13.3.18].

From [13.2.23], the series for $\hat{\varepsilon}_{t|t}$ is generated recursively from

$$\begin{aligned} \begin{bmatrix} 0 \\ \hat{\varepsilon}_{t|t} \end{bmatrix} &= \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ \hat{\varepsilon}_{t-1|t-1} \end{bmatrix} \\ &+ \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & p_t \end{bmatrix}\begin{bmatrix} 1 \\ \theta \end{bmatrix}\{1/[\sigma^2 + \theta^2 p_t]\}\cdot\{y_t - \mu - \theta\hat{\varepsilon}_{t-1|t-1}\} \end{aligned}$$

or

$$\hat{\varepsilon}_{t|t} = \{\sigma^2/[\sigma^2 + \theta^2 p_t]\}\cdot\{y_t - \mu - \theta\hat{\varepsilon}_{t-1|t-1}\} \qquad [13.3.19]$$

starting from the initial value $\hat{\varepsilon}_{0|0} = 0$. Note that the value for $\hat{\varepsilon}_{t|t}$ differs from the approximation suggested in equations [4.2.36] and [4.3.2],

$$\hat{\varepsilon}_t = y_t - \mu - \theta\hat{\varepsilon}_{t-1} \qquad \hat{\varepsilon}_0 = 0,$$

in that [13.3.19] shrinks the inference $\hat{\varepsilon}_t$ toward zero to take account of the nonzero variance p_t of $\hat{\varepsilon}_{t-1|t-1}$ around the true value ε_{t-1}.

The gain matrix $\mathbf{K}_t$ in equation [13.2.19] is given by

$$\mathbf{K}_t = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & p_t \end{bmatrix}\begin{bmatrix} 1 \\ \theta \end{bmatrix}\left(\frac{1}{\sigma^2 + \theta^2 p_t}\right) = \begin{bmatrix} 0 \\ \sigma^2/[\sigma^2 + \theta^2 p_t] \end{bmatrix}. \qquad [13.3.20]$$

Finally, notice from [13.2.16] that

$$\mathbf{P}_{t|t} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & p_t \end{bmatrix} - \left(\frac{1}{\sigma^2 + \theta^2 p_t}\right)\begin{bmatrix} \sigma^2 & 0 \\ 0 & p_t \end{bmatrix}\begin{bmatrix} 1 \\ \theta \end{bmatrix}[1 \quad \theta]\begin{bmatrix} \sigma^2 & 0 \\ 0 & p_t \end{bmatrix}.$$

The (1, 1) element of $\mathbf{P}_{t|t}$ (which we saw equals p_{t+1}) is thus given by

$$p_{t+1} = \sigma^2 - \{1/[\sigma^2 + \theta^2 p_t]\}\cdot\sigma^4 = \frac{\sigma^2\theta^2 p_t}{\sigma^2 + \theta^2 p_t}. \qquad [13.3.21]$$

The recursion in [13.3.21] is started with $p_1 = \sigma^2$ and thus has the solution

$$p_{t+1} = \frac{\sigma^2\theta^{2t}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2t}}. \qquad [13.3.22]$$

It is interesting to note what happens to the filter as t becomes large. First consider the case when $|\theta| \leq 1$. Then, from [13.3.22],

$$\lim_{t\to\infty} p_{t+1} = 0,$$

and so, from [13.3.17],

$$\hat{\varepsilon}_{t|t} \xrightarrow{p} \varepsilon_t.$$

Thus, given a sufficient number of observations on y, the Kalman filter inference $\hat{\varepsilon}_{t|t}$ converges to the true value ε_t, and the forecast [13.3.14] converges to that of the Wold representation for the process. The Kalman gain in [13.3.20] converges to $(0, 1)'$.

Alternatively, consider the case when $|\theta| > 1$. From [13.3.22], we have

$$p_{t+1} = \frac{\sigma^2\theta^{2t}(1 - \theta^2)}{1 - \theta^{2(t+1)}} = \frac{\sigma^2(1 - \theta^2)}{\theta^{-2t} - \theta^2}$$

and

$$\lim_{t\to\infty} p_{t+1} = \frac{\sigma^2(1 - \theta^2)}{-\theta^2} > 0.$$

No matter how many observations are obtained, it will not be possible to know with certainty the value of the nonfundamental innovation ε_t associated with date t on the basis of $(y_t, y_{t-1}, \ldots, y_1)$. The gain is given by

$$\frac{\sigma^2}{\sigma^2 + \theta^2 p_t} \to \frac{\sigma^2}{\sigma^2 - \sigma^2(1 - \theta^2)} = \frac{1}{\theta^2},$$

and the recursion [13.3.19] approaches

$$\hat{\varepsilon}_{t|t} = (1/\theta^2)\cdot(y_t - \mu - \theta\hat{\varepsilon}_{t-1|t-1})$$

or

$$\theta\hat{\varepsilon}_{t|t} = (1/\theta)\cdot(y_t - \mu - \theta\hat{\varepsilon}_{t-1|t-1}).$$

Recalling [13.3.14], we thus have

$$\hat{y}_{t+1|t} - \mu = (1/\theta)\cdot[(y_t - \mu) - (\hat{y}_{t|t-1} - \mu)]$$

or

$$\hat{y}_{t+1|t} - \mu = (1/\theta)\cdot(y_t - \mu) - (1/\theta)^2\cdot(y_{t-1} - \mu) + (1/\theta)^3\cdot(y_{t-2} - \mu) - \cdots,$$

which again is the $AR(\infty)$ forecast associated with the invertible $MA(1)$ representation. Indeed, the forecasts of the Kalman filter with θ replaced by θ^{-1} and σ^2 replaced by $\theta^2\sigma^2$ will be identical for any t; see Exercise 13.5.

Calculating s*-Period-Ahead Forecasts with the Kalman Filter*

The forecast of $\mathbf{y}_t$ calculated in [13.2.24] is an exact finite-sample forecast of $\mathbf{y}_t$ on the basis of $\mathbf{x}_t$ and $\mathcal{Y}_{t-1} \equiv (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_1, \mathbf{x}'_{t-1}, \mathbf{x}'_{t-2}, \ldots, \mathbf{x}'_1)'$. If $\mathbf{x}_t$

is deterministic, it is also easy to use the Kalman filter to calculate exact finite-sample s-period-ahead forecasts.

The state equation [13.2.1] can be solved by recursive substitution to yield

$$\boldsymbol{\xi}_{t+s} = \mathbf{F}^s\boldsymbol{\xi}_t + \mathbf{F}^{s-1}\mathbf{v}_{t+1} + \mathbf{F}^{s-2}\mathbf{v}_{t+2} + \cdots + \mathbf{F}^1\mathbf{v}_{t+s-1} + \mathbf{v}_{t+s} \quad \text{for } s = 1, 2, \ldots. \qquad [13.3.23]$$

The projection of $\boldsymbol{\xi}_{t+s}$ on $\boldsymbol{\xi}_t$ and $\mathcal{Y}_t$ is given by

$$\hat{E}(\boldsymbol{\xi}_{t+s}|\boldsymbol{\xi}_t, \mathcal{Y}_t) = \mathbf{F}^s\boldsymbol{\xi}_t. \qquad [13.3.24]$$

From the law of iterated projections,

$$\hat{\boldsymbol{\xi}}_{t+s|t} \equiv \hat{E}(\boldsymbol{\xi}_{t+s}|\mathcal{Y}_t) = \mathbf{F}^s\hat{\boldsymbol{\xi}}_{t|t}. \qquad [13.3.25]$$

Thus, from [13.3.23] the s-period-ahead forecast error for the state vector is

$$\boldsymbol{\xi}_{t+s} - \hat{\boldsymbol{\xi}}_{t+s|t} = \mathbf{F}^s(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t}) + \mathbf{F}^{s-1}\mathbf{v}_{t+1} + \mathbf{F}^{s-2}\mathbf{v}_{t+2} + \cdots + \mathbf{F}^1\mathbf{v}_{t+s-1} + \mathbf{v}_{t+s} \qquad [13.3.26]$$

with *MSE*

$$\mathbf{P}_{t+s|t} = \mathbf{F}^s\mathbf{P}_{t|t}(\mathbf{F}')^s + \mathbf{F}^{s-1}\mathbf{Q}(\mathbf{F}')^{s-1} + \mathbf{F}^{s-2}\mathbf{Q}(\mathbf{F}')^{s-2} + \cdots + \mathbf{F}\mathbf{Q}\mathbf{F}' + \mathbf{Q}. \qquad [13.3.27]$$

To forecast the observed vector $\mathbf{y}_{t+s}$, recall from the observation equation that

$$\mathbf{y}_{t+s} = \mathbf{A}'\mathbf{x}_{t+s} + \mathbf{H}'\boldsymbol{\xi}_{t+s} + \mathbf{w}_{t+s}. \qquad [13.3.28]$$

There are advantages if the state vector is defined in such a way that $\mathbf{x}_t$ is deterministic, so that the dynamics of any exogenous variables can be represented through $\boldsymbol{\xi}_t$. If $\mathbf{x}_t$ is deterministic, the s-period-ahead forecast of $\mathbf{y}$ is

$$\hat{\mathbf{y}}_{t+s|t} \equiv \hat{E}(\mathbf{y}_{t+s}|\mathcal{Y}_t) = \mathbf{A}'\mathbf{x}_{t+s} + \mathbf{H}'\hat{\boldsymbol{\xi}}_{t+s|t}. \qquad [13.3.29]$$

The forecast error is

$$\begin{aligned}\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t} &= (\mathbf{A}'\mathbf{x}_{t+s} + \mathbf{H}'\boldsymbol{\xi}_{t+s} + \mathbf{w}_{t+s}) - (\mathbf{A}'\mathbf{x}_{t+s} + \mathbf{H}'\hat{\boldsymbol{\xi}}_{t+s|t}) \\ &= \mathbf{H}'(\boldsymbol{\xi}_{t+s} - \hat{\boldsymbol{\xi}}_{t+s|t}) + \mathbf{w}_{t+s}\end{aligned}$$

with *MSE*

$$E[(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})(\mathbf{y}_{t+s} - \hat{\mathbf{y}}_{t+s|t})'] = \mathbf{H}'\mathbf{P}_{t+s|t}\mathbf{H} + \mathbf{R}. \qquad [13.3.30]$$

13.4. *Maximum Likelihood Estimation of Parameters*

Using the Kalman Filter to Evaluate the Likelihood Function

The Kalman filter was motivated in Section 13.2 in terms of linear projections. The forecasts $\hat{\boldsymbol{\xi}}_{t|t-1}$ and $\hat{\mathbf{y}}_{t|t-1}$ are thus optimal within the set of forecasts that are linear in $(\mathbf{x}_t, \mathcal{Y}_{t-1})$, where $\mathcal{Y}_{t-1} \equiv (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_1, \mathbf{x}'_{t-1}, \mathbf{x}'_{t-2}, \ldots, \mathbf{x}'_1)'$. If the initial state $\boldsymbol{\xi}_1$ and the innovations $\{\mathbf{w}_t, \mathbf{v}_t\}_{t=1}^T$ are multivariate Gaussian, then we can make the stronger claim that the forecasts $\hat{\boldsymbol{\xi}}_{t|t-1}$ and $\hat{\mathbf{y}}_{t|t-1}$ calculated by the Kalman filter are optimal among any functions of $(\mathbf{x}_t, \mathcal{Y}_{t-1})$. Moreover, if $\boldsymbol{\xi}_1$ and $\{\mathbf{w}_t, \mathbf{v}_t\}_{t=1}^T$ are Gaussian, then the distribution of $\mathbf{y}_t$ conditional on $(\mathbf{x}_t, \mathcal{Y}_{t-1})$ is Gaussian with mean given by [13.2.24] and variance given by [13.2.25]:

$$\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1} \sim N((\mathbf{A}'\mathbf{x}_t + \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}), (\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R}));$$

that is,

$$\begin{aligned}&f_{\mathbf{Y}_t|\mathbf{X}_t,\mathcal{Y}_{t-1}}(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}) \\ &= (2\pi)^{-n/2}|\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R}|^{-1/2} \\ &\quad \times \exp\{-\tfrac{1}{2}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1})'(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1} \\ &\quad \times (\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1})\} \qquad \text{for } t = 1, 2, \ldots, T.\end{aligned} \qquad [13.4.1]$$

From [13.4.1], it is a simple matter to construct the sample log likelihood,

$$\sum_{t=1}^{T} \log f_{\mathbf{Y}_t|\mathbf{X}_t,\mathcal{Y}_{t-1}}(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}). \qquad [13.4.2]$$

Expression [13.4.2] can then be maximized numerically with respect to the unknown parameters in the matrices $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$; see Burmeister and Wall (1982) for an illustrative application.

As stressed by Harvey and Phillips (1979), this representation of the likelihood is particularly convenient for estimating regressions involving moving average terms. Moreover, [13.4.2] gives the exact log likelihood function, regardless of whether the moving average representation is invertible.

As an illustrative example, suppose we wanted to estimate a bivariate regression model whose equations were

$$y_{1t} = \mathbf{a}_1'\mathbf{x}_t + u_{1t}$$
$$y_{2t} = \mathbf{a}_2'\mathbf{x}_t + u_{2t},$$

where $\mathbf{x}_t$ is a $(k \times 1)$ vector of exogenous explanatory variables and $\mathbf{a}_1$ and $\mathbf{a}_2$ are $(k \times 1)$ vectors of coefficients; if the two regressions have different explanatory variables, the variables from both regressions are included in $\mathbf{x}_t$ with zeros appropriately imposed on $\mathbf{a}_1$ and $\mathbf{a}_2$. Suppose that the disturbance vector follows a bivariate $MA(1)$ process:

$$\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix} + \begin{bmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{bmatrix} \begin{bmatrix} \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix},$$

with $(\varepsilon_{1t}, \varepsilon_{2t})' \sim$ i.i.d. $N(\mathbf{0}, \mathbf{\Omega})$. This model can be written in state-space form by defining

$$\boldsymbol{\xi}_t = \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{1,t-1} \\ \varepsilon_{2,t-1} \end{bmatrix} \qquad \mathbf{F} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{v}_{t+1} = \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{Q} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{A}' = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{bmatrix}$$

$$\mathbf{H}' = \begin{bmatrix} 1 & 0 & \theta_{11} & \theta_{12} \\ 0 & 1 & \theta_{21} & \theta_{22} \end{bmatrix} \qquad \mathbf{R} = \mathbf{0},$$

where $\sigma_{ij} = E(\varepsilon_{it}\varepsilon_{jt})$. The Kalman filter iteration is started from

$$\hat{\boldsymbol{\xi}}_{1|0} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{P}_{1|0} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_{22} & 0 & 0 \\ 0 & 0 & \sigma_{11} & \sigma_{12} \\ 0 & 0 & \sigma_{21} & \sigma_{22} \end{bmatrix}.$$

Maximization of [13.4.2] is started by making an initial guess as to the numerical values of the unknown parameters. One obvious way to do this is to regress y_{1t} on the elements of $\mathbf{x}_t$ that appear in the first equation to get an initial guess for $\mathbf{a}_1$. A similar OLS regression for y_2 yields a guess for $\mathbf{a}_2$. Setting $\theta_{11} = \theta_{12} = \theta_{21} = \theta_{22} = 0$ initially, a first guess for $\mathbf{\Omega}$ could be the estimated variance-covariance matrix of the residuals from these two OLS regressions. For these initial numerical values for the population parameters, we could construct $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ from the expressions just given and iterate on [13.2.22] through [13.2.25] for $t = 1, 2, \ldots, T - 1$. The sequences $\{\hat{\boldsymbol{\xi}}_{t|t-1}\}_{t=1}^{T}$ and $\{\mathbf{P}_{t|t-1}\}_{t=1}^{T}$ resulting from these iter-

ations could then be used in [13.4.1] and [13.4.2] to calculate the value for the log likelihood function that results from these initial parameter values. The numerical optimization methods described in Section 5.7 can then be employed to make better guesses as to the value of the unknown parameters until [13.4.2] is maximized. As noted in Section 5.9, the numerical search will be better behaved if $\mathbf{\Omega}$ is parameterized in terms of its Cholesky factorization.

As a second example, consider a scalar Gaussian $ARMA(1, 1)$ process,

$$y_t - \mu = \phi(y_{t-1} - \mu) + \varepsilon_t + \theta\varepsilon_{t-1},$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. This can be written in state-space form as in [13.1.22] and [13.1.23] with $r = 2$ and

$$\mathbf{F} = \begin{bmatrix} \phi & 0 \\ 1 & 0 \end{bmatrix} \qquad \mathbf{v}_{t+1} = \begin{bmatrix} \varepsilon_{t+1} \\ 0 \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{A}' = \mu \qquad \mathbf{x}_t = 1 \qquad \mathbf{H}' = [1 \quad \theta] \qquad \mathbf{R} = 0$$

$$\hat{\boldsymbol{\xi}}_{1|0} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \mathbf{P}_{1|0} = \begin{bmatrix} \sigma^2/(1 - \phi^2) & \phi\sigma^2/(1 - \phi^2) \\ \phi\sigma^2/(1 - \phi^2) & \sigma^2/(1 - \phi^2) \end{bmatrix}.$$

This value for $\mathbf{P}_{1|0}$ was obtained by recognizing that the state equation [13.1.22] describes the behavior of $\boldsymbol{\xi}_t = (z_t, z_{t-1}, \ldots, z_{t-r+1})'$, where $z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \cdots + \phi_r z_{t-r} + \varepsilon_t$ follows an $AR(r)$ process. For this example, $r = 2$, so that $\mathbf{P}_{1|0}$ is the variance-covariance matrix of two consecutive draws from an $AR(2)$ process with parameters $\phi_1 = \phi$ and $\phi_2 = 0$. The expressions just given for $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ are then used in the Kalman filter iterations. Thus, expression [13.4.2] allows easy computation of the exact likelihood function for an $ARMA(p, q)$ process. This computation is valid regardless of whether the moving average parameters satisfy the invertibility condition. Similarly, expression [13.3.29] gives the exact finite-sample s-period-ahead forecast for the process and [13.3.30] its MSE, again regardless of whether the invertible representation is used.

Typically, numerical search procedures for maximizing [13.4.2] require the derivatives of the log likelihood. These can be calculated numerically or analytically. To characterize the analytical derivatives of [13.4.2], collect the unknown parameters to be estimated in a vector $\boldsymbol{\theta}$, and write $\mathbf{F}(\boldsymbol{\theta})$, $\mathbf{Q}(\boldsymbol{\theta})$, $\mathbf{A}(\boldsymbol{\theta})$, $\mathbf{H}(\boldsymbol{\theta})$, and $\mathbf{R}(\boldsymbol{\theta})$. Implicitly, then, $\hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta})$ and $\mathbf{P}_{t|t-1}(\boldsymbol{\theta})$ will be functions of $\boldsymbol{\theta}$ as well, and the derivative of the log of [13.4.1] with respect to the ith element of $\boldsymbol{\theta}$ will involve $\partial\hat{\boldsymbol{\xi}}_{t|t-1}(\boldsymbol{\theta})/\partial\theta_i$ and $\partial\mathbf{P}_{t|t-1}(\boldsymbol{\theta})/\partial\theta_i$. These derivatives can also be generated recursively by differentiating the Kalman filter recursion, [13.2.22] and [13.2.23], with respect to θ_i; see Caines (1988, pp. 585–86) for illustration.

For many state-space models, the EM algorithm of Dempster, Laird, and Rubin (1977) offers a particularly convenient means for maximizing [13.4.2], as developed by Shumway and Stoffer (1982) and Watson and Engle (1983).

Identification

Although the state-space representation gives a very convenient way to calculate the exact likelihood function, a word of caution should be given. In the absence of restrictions on $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$, the parameters of the state-space representation are unidentified—more than one set of values for the parameters can give rise to the identical value of the likelihood function, and the data give us no guide for choosing among these. A trivial example is the following system:

State Equation ($r = 2$):

$$\boldsymbol{\xi}_{t+1} = \begin{bmatrix} \varepsilon_{1,t+1} \\ \varepsilon_{2,t+1} \end{bmatrix} \qquad [13.4.3]$$

Observation Equation ($n = 1$):

$$y_t = \varepsilon_{1t} + \varepsilon_{2t}. \qquad [13.4.4]$$

Here, $\mathbf{F} = \mathbf{0}$, $\mathbf{Q} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$, $\mathbf{A}' = 0$, $\mathbf{H}' = [1 \quad 1]$, and $\mathbf{R} = 0$. This model asserts that y_t is white noise, with mean zero and variance given by $(\sigma_1^2 + \sigma_2^2)$. The reader is invited to confirm in Exercise 13.4 that the log of the likelihood function from [13.4.1] and [13.4.2] simplifies to

$$\begin{aligned} &\log f_{\mathbf{Y}_T,\mathbf{Y}_{T-1},\ldots,\mathbf{Y}_1}(\mathbf{y}_T, \mathbf{y}_{T-1}, \ldots, \mathbf{y}_1) \\ &= -(T/2)\log(2\pi) - (T/2)\log(\sigma_1^2 + \sigma_2^2) - \sum_{t=1}^{T} y_t^2/[2(\sigma_1^2 + \sigma_2^2)]. \end{aligned} \qquad [13.4.5]$$

Clearly, any values for σ_1^2 and σ_2^2 that sum to a given constant will produce the identical value for the likelihood function.

The $MA(1)$ process explored in Section 13.3 provides a second example of an unidentified state-space representation. As the reader may verify in Exercise 13.5, the identical value for the log likelihood function [13.4.2] would result if θ is replaced by θ^{-1} and σ^2 by $\theta^2\sigma^2$.

These two examples illustrate two basic forms in which absence of identification can occur. Following Rothenberg (1971), a model is said to be *globally identified* at a particular parameter value $\boldsymbol{\theta}_0$ if for any value of $\boldsymbol{\theta}$ there exists a possible realization $\mathcal{Y}_T$ for which the value of the likelihood at $\boldsymbol{\theta}$ is different from the value of the likelihood at $\boldsymbol{\theta}_0$. A model is said to be *locally identified* at $\boldsymbol{\theta}_0$ if there exists a $\delta > 0$ such that for any value of $\boldsymbol{\theta}$ satisfying $(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) < \delta$, there exists a possible realization of $\mathcal{Y}_T$ for which the value of the likelihood at $\boldsymbol{\theta}$ is different from the value of the likelihood at $\boldsymbol{\theta}_0$. Thus, global identification implies local identification. The first example, [13.4.3] and [13.4.4], is neither globally nor locally identified, while the $MA(1)$ example is locally identified but globally unidentified.

Local identification is much easier to test for than global identification. Rothenberg (1971) showed that a model is locally identified at $\boldsymbol{\theta}_0$ if and only if the information matrix is nonsingular in a neighborhood around $\boldsymbol{\theta}_0$. Thus, a common symptom of trying to estimate an unidentified model is difficulty with inverting the matrix of second derivatives of the log likelihood function. One approach to checking for local identification is to translate the state-space representation back into a vector *ARMA* model and check for satisfaction of the conditions in Hannan (1971); see Hamilton (1985) for an example of this approach. A second approach is to work directly with the state-space representation, as is done in Gevers and Wertz (1984) and Wall (1987). For an illustration of the second approach, see Burmeister, Wall, and Hamilton (1986).

Asymptotic Properties of Maximum Likelihood Estimates

If certain regularity conditions are satisfied, then Caines (1988, Chapter 7) showed that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_T$ based on a sample of size T is consistent and asymptotically normal. These conditions include the following: (1) the model must be identified; (2) eigenvalues of $\mathbf{F}$ are all inside the unit circle; (3)

apart from a constant term, the variables $\mathbf{x}_t$ behave asymptotically like a full-rank linearly indeterministic covariance-stationary process; and (4) the true value of $\boldsymbol{\theta}$ does not fall on a boundary of the allowable parameter space. Pagan (1980, Theorem 4) and Ghosh (1989) examined special cases of state-space models for which

$$\sqrt{T}\mathcal{I}_{2D,T}^{1/2}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{I}_a), \qquad [13.4.6]$$

where a is the number of elements of $\boldsymbol{\theta}$ and $\mathcal{I}_{2D,T}$ is the $(a \times a)$ information matrix for a sample of size T as calculated from second derivatives of the log likelihood function:

$$\mathcal{I}_{2D,T} = -\frac{1}{T}E\left(\sum_{t=1}^{T} \frac{\partial^2 \log f(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right). \qquad [13.4.7]$$

A common practice is to assume that the limit of $\mathcal{I}_{2D,T}$ as $T \to \infty$ is the same as the plim of

$$\hat{\mathcal{I}}_{2D,T} = -\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 \log f(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}, \qquad [13.4.8]$$

which can be calculated analytically or numerically by differentiating [13.4.2]. Reported standard errors for $\hat{\boldsymbol{\theta}}_T$ are then square roots of diagonal elements of $(1/T)(\hat{\mathcal{I}}_{2D,T})^{-1}$.

Quasi-Maximum Likelihood Estimation

Even if the disturbances $\mathbf{v}_t$ and $\mathbf{w}_t$ are non-Gaussian, the Kalman filter can still be used to calculate the linear projection of $\mathbf{y}_{t+s}$ on past observables. Moreover, we can form the function [13.4.2] and maximize it with respect to $\boldsymbol{\theta}$ even for non-Gaussian systems. This procedure will still yield consistent and asymptotically Normal estimates of the elements of $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$, with the variance-covariance matrix constructed as described in equation [5.8.7]. Watson (1989, Theorem 2) presented conditions under which the quasi-maximum likelihood estimates satisfy

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, [\mathcal{I}_{2D}\mathcal{I}_{OP}^{-1}\mathcal{I}_{2D}]^{-1}), \qquad [13.4.9]$$

where $\mathcal{I}_{2D}$ is the plim of [13.4.8] when evaluated at the true value $\boldsymbol{\theta}_0$ and $\mathcal{I}_{OP}$ is the outer-product estimate of the information matrix,

$$\mathcal{I}_{OP} = \text{plim } (1/T)\sum_{t=1}^{T} [\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)]',$$

where

$$\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t) \equiv \frac{\partial \log f(\mathbf{y}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

13.5. *The Steady-State Kalman Filter*

Convergence Properties of the Kalman Filter

Section 13.3 applied the Kalman filter to an $MA(1)$ process and found that when $|\theta| \leq 1$,

$$\lim_{t\to\infty} \mathbf{P}_{t+1|t} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\lim_{t\to\infty} \mathbf{K}_t = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

whereas when $|\theta| > 1$,

$$\lim_{t\to\infty} \mathbf{P}_{t+1|t} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2(\theta^2 - 1)/\theta^2 \end{bmatrix}$$

$$\lim_{t\to\infty} \mathbf{K}_t = \begin{bmatrix} 0 \\ 1/\theta^2 \end{bmatrix}.$$

It turns out to be a property of a broad class of state-space models that the sequences $\{\mathbf{P}_{t+1|t}\}_{t=1}^{T}$ and $\{\mathbf{K}_t\}_{t=1}^{T}$ converge to fixed matrices, as the following proposition shows.

Proposition 13.1: Let $\mathbf{F}$ *be an* $(r \times r)$ *matrix whose eigenvalues are all inside the unit circle, let* $\mathbf{H}'$ *denote an arbitrary* $(n \times r)$ *matrix, and let* $\mathbf{Q}$ *and* $\mathbf{R}$ *be positive semidefinite symmetric* $(r \times r)$ *and* $(n \times n)$ *matrices, respectively. Let* $\{\mathbf{P}_{t+1|t}\}_{t=1}^{T}$ *be the sequence of MSE matrices calculated by the Kalman filter,*

$$\mathbf{P}_{t+1|t} = \mathbf{F}[\mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1}]\mathbf{F}' + \mathbf{Q}, \qquad [13.5.1]$$

where iteration on [13.5.1] *is initialized by letting* $\mathbf{P}_{1|0}$ *be the positive semidefinite* $(r \times r)$ *matrix satisfying*

$$\text{vec}(\mathbf{P}_{1|0}) = [\mathbf{I}_{r^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1} \cdot \text{vec}(\mathbf{Q}). \qquad [13.5.2]$$

Then $\{\mathbf{P}_{t+1|t}\}_{t=1}^{T}$ *is a monotonically nonincreasing sequence and converges as* $T \to \infty$ *to a steady-state matrix* $\mathbf{P}$ *satisfying*

$$\mathbf{P} = \mathbf{F}[\mathbf{P} - \mathbf{P}\mathbf{H}(\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}]\mathbf{F}' + \mathbf{Q}. \qquad [13.5.3]$$

Moreover, the steady-state value for the Kalman gain matrix, defined by

$$\mathbf{K} \equiv \mathbf{F}\mathbf{P}\mathbf{H}(\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R})^{-1}, \qquad [13.5.4]$$

has the property that the eigenvalues of $(\mathbf{F} - \mathbf{K}\mathbf{H}')$ *all lie on or inside the unit circle.*

The claim in Proposition 13.1 that $\mathbf{P}_{t+1|t} \leq \mathbf{P}_{t|t-1}$ means that for any real $(r \times 1)$ vector $\mathbf{h}$, the scalar inequality $\mathbf{h}'\mathbf{P}_{t+1|t}\mathbf{h} \leq \mathbf{h}'\mathbf{P}_{t|t-1}\mathbf{h}$ holds.

Proposition 13.1 assumes that the Kalman filter is started with $\mathbf{P}_{1|0}$ equal to the unconditional variance-covariance matrix of the state vector $\boldsymbol{\xi}_t$. Although the sequence $\{\mathbf{P}_{t+1|t}\}$ converges to a matrix $\mathbf{P}$, the solution to [13.5.3] need not be unique; a different starting value for $\mathbf{P}_{1|0}$ might produce a sequence that converges to a different matrix $\mathbf{P}$ satisfying [13.5.3]. Under the slightly stronger assumption that either $\mathbf{Q}$ or $\mathbf{R}$ is strictly positive definite, then iteration on [13.5.1] will converge to a unique solution to [13.5.3], where the starting value for the iteration $\mathbf{P}_{1|0}$ can be any positive semidefinite symmetric matrix.

Proposition 13.2: Let $\mathbf{F}$ *be an* $(r \times r)$ *matrix whose eigenvalues are all inside the unit circle, let* $\mathbf{H}'$ *denote an arbitrary* $(n \times r)$ *matrix, and let* $\mathbf{Q}$ *and* $\mathbf{R}$ *be positive semidefinite symmetric* $(r \times r)$ *and* $(n \times n)$ *matrices, respectively, with either* $\mathbf{Q}$ *or* $\mathbf{R}$ *strictly positive definite. Then the sequence of Kalman MSE matrices* $\{\mathbf{P}_{t+1|t}\}_{t=1}^{T}$ *determined by [13.5.1] converges to a unique positive semidefinite steady-state matrix* $\mathbf{P}$ *satisfying [13.5.3], where the value of* $\mathbf{P}$ *is the same for any positive semidefinite symmetric starting value for* $\mathbf{P}_{1|0}$. *Moreover, the steady-state value for the Kalman gain matrix* $\mathbf{K}$ *in [13.5.4] has the property that the eigenvalues of* $(\mathbf{F} - \mathbf{K}\mathbf{H}')$ *are all strictly inside the unit circle.*

We next discuss the relevance of the results in Propositions 13.1 and 13.2 concerning the eigenvalues of $(\mathbf{F} - \mathbf{K}\mathbf{H}')$.

Using the Kalman Filter to Find the Wold Representation and Factor an Autocovariance-Generating Function

Consider a system in which the explanatory variables ($\mathbf{x}_t$) consist solely of a constant term. Without loss of generality, we simplify the notation by assuming that $\mathbf{A}'\mathbf{x}_t \equiv \mathbf{0}$. For such systems, the Kalman filter forecast of the state vector can be written as in [13.2.20]:

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}). \qquad [13.5.5]$$

The linear projection of $\mathbf{y}_{t+1}$ on the observed finite sample of its own lagged values is then calculated from

$$\hat{\mathbf{y}}_{t+1|t} = \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1) = \mathbf{H}'\hat{\boldsymbol{\xi}}_{t+1|t}, \qquad [13.5.6]$$

with *MSE* given by [13.2.25]:

$$E[(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})(\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1|t})'] = \mathbf{H}'\mathbf{P}_{t+1|t}\mathbf{H} + \mathbf{R}. \qquad [13.5.7]$$

Consider the result from applying the Kalman filter to a covariance-stationary process that started up at a time arbitrarily distant in the past. From Proposition 13.1, the difference equation [13.5.5] will converge to

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{K}(\mathbf{y}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}), \qquad [13.5.8]$$

with $\mathbf{K}$ given by [13.5.4]. The forecast [13.5.6] will approach the forecast of $\mathbf{y}_{t+1}$ based on the infinite history of its own lagged values:

$$\hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots) = \mathbf{H}'\hat{\boldsymbol{\xi}}_{t+1|t}. \qquad [13.5.9]$$

The *MSE* of this forecast is given by the limiting value of [13.5.7],

$$E\{[\mathbf{y}_{t+1} - \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)][\mathbf{y}_{t+1} - \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)]'\} = \mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}, \qquad [13.5.10]$$

where $\mathbf{P}$ is given by [13.5.3].

Equation [13.5.8] can be written

$$\hat{\boldsymbol{\xi}}_{t+1|t} = (\mathbf{F} - \mathbf{K}\mathbf{H}')L\hat{\boldsymbol{\xi}}_{t+1|t} + \mathbf{K}\mathbf{y}_t \qquad [13.5.11]$$

for L the lag operator. Provided that the eigenvalues of $(\mathbf{F} - \mathbf{K}\mathbf{H}')$ are all inside the unit circle, [13.5.11] can be expressed as

$$\begin{aligned}\hat{\boldsymbol{\xi}}_{t+1|t} &= [\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')L]^{-1}\mathbf{K}\mathbf{y}_t \\ &= [\mathbf{I}_r + (\mathbf{F} - \mathbf{K}\mathbf{H}')L + (\mathbf{F} - \mathbf{K}\mathbf{H}')^2L^2 + (\mathbf{F} - \mathbf{K}\mathbf{H}')^3L^3 + \cdots]\mathbf{K}\mathbf{y}_t.\end{aligned} \qquad [13.5.12]$$

Substituting [13.5.12] into [13.5.9] gives a steady-state rule for forecasting $\mathbf{y}_{t+1}$ as a linear function of its lagged values:

$$\hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots) = \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')L]^{-1}\mathbf{K}\mathbf{y}_t. \qquad [13.5.13]$$

Expression [13.5.13] implies a $VAR(\infty)$ representation for $\mathbf{y}_t$ of the form

$$\mathbf{y}_{t+1} = \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')L]^{-1}\mathbf{K}\mathbf{y}_t + \boldsymbol{\varepsilon}_{t+1}, \qquad [13.5.14]$$

where

$$\boldsymbol{\varepsilon}_{t+1} \equiv \mathbf{y}_{t+1} - \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots). \qquad [13.5.15]$$

Thus, $\boldsymbol{\varepsilon}_{t+1}$ is the fundamental innovation for $\mathbf{y}_{t+1}$. Since $\boldsymbol{\varepsilon}_{t+1}$ is uncorrelated with $\mathbf{y}_{t-j}$ for any $j \geq 0$, it is also uncorrelated with $\boldsymbol{\varepsilon}_{t-j} = \mathbf{y}_{t-j} - \hat{E}(\mathbf{y}_{t-j}|\mathbf{y}_{t-j-1}, \mathbf{y}_{t-j-2}, \ldots)$ for any $j \geq 0$. The variance-covariance matrix of $\boldsymbol{\varepsilon}_{t+1}$ can be calculated using [13.5.15] and [13.5.10]):

$$\begin{aligned}E(\boldsymbol{\varepsilon}_{t+1}\boldsymbol{\varepsilon}_{t+1}') &= E\{[\mathbf{y}_{t+1} - \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)] \\ &\quad \times [\mathbf{y}_{t+1} - \hat{E}(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots)]'\} \\ &= \mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}.\end{aligned} \qquad [13.5.16]$$

Note that [13.5.14] can be written as

$$\{\mathbf{I}_n - \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{KH}')L]^{-1}\mathbf{K}L\}\mathbf{y}_{t+1} = \boldsymbol{\varepsilon}_{t+1}. \qquad [13.5.17]$$

The following result helps to rewrite the $VAR(\infty)$ representation [13.5.17] in the Wold $MA(\infty)$ form.

Proposition 13.3: Let $\mathbf{F}$, $\mathbf{H}'$, *and* $\mathbf{K}$ *be matrices of dimension* $(r \times r)$, $(n \times r)$, *and* $(r \times n)$, *respectively, such that eigenvalues of* $\mathbf{F}$ *and of* $(\mathbf{F} - \mathbf{KH}')$ *are all inside the unit circle, and let* z *be a scalar on the complex unit circle. Then*

$$\{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{I}_n - \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{KH}')z]^{-1}\mathbf{K}z\} = \mathbf{I}_n.$$

Applying Proposition 13.3, if both sides of [13.5.17] are premultiplied by $(\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}L)^{-1}\mathbf{K}L)$, the result is the Wold representation for $\mathbf{y}$:

$$\mathbf{y}_{t+1} = \{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}L)^{-1}\mathbf{K}L\}\boldsymbol{\varepsilon}_{t+1}. \qquad [13.5.18]$$

To summarize, the Wold representation can be found by iterating on [13.5.1] until convergence. The steady-state value for $\mathbf{P}$ is then used to construct $\mathbf{K}$ in [13.5.4]. If the eigenvalues of $(\mathbf{F} - \mathbf{KH}')$ are all inside the unit circle, then the Wold representation is given by [13.5.18].

The task of finding the Wold representation is sometimes alternatively posed as the question of factoring the autocovariance-generating function of $\mathbf{y}$. Applying result [10.3.7] to [13.5.16] and [13.5.18], we would anticipate that the autocovariance-generating function of $\mathbf{y}$ can be written in the form

$$\begin{aligned}\mathbf{G}_{\mathbf{Y}}(z) = {} & \{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{H}'\mathbf{PH} + \mathbf{R}\} \\ & \times \{\mathbf{I}_n + \mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1}\}.\end{aligned} \qquad [13.5.19]$$

Compare [13.5.19] with the autocovariance-generating function that we would have written down directly from the structure of the state-space model. From [10.3.5], the autocovariance-generating function of $\boldsymbol{\xi}$ is given by

$$\mathbf{G}_{\boldsymbol{\xi}}(z) = [\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{Q}[\mathbf{I}_r - \mathbf{F}'z^{-1}]^{-1},$$

while from [10.3.6] the autocovariance-generating function of $\mathbf{y}_t = \mathbf{H}'\boldsymbol{\xi}_t + \mathbf{w}_t$ is

$$\mathbf{G}_{\mathbf{Y}}(z) = \mathbf{H}'[\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{Q}[\mathbf{I}_r - \mathbf{F}'z^{-1}]^{-1}\mathbf{H} + \mathbf{R}. \qquad [13.5.20]$$

Comparing [13.5.19] with [13.5.20] suggests that the limiting values for the Kalman gain and MSE matrices $\mathbf{K}$ and $\mathbf{P}$ can be used to factor an autocovariance-generating function. The following proposition gives a formal statement of this result.

Proposition 13.4: Let $\mathbf{F}$ *denote an* $(r \times r)$ *matrix whose eigenvalues are all inside the unit circle; let* $\mathbf{Q}$ *and* $\mathbf{R}$ *denote symmetric positive semidefinite matrices of dimension* $(r \times r)$ *and* $(n \times n)$, *respectively; and let* $\mathbf{H}'$ *denote an arbitrary* $(n \times r)$ *matrix. Let* $\mathbf{P}$ *be a positive semidefinite matrix satisfying [13.5.3] and let* $\mathbf{K}$ *be given by [13.5.4]. Suppose that eigenvalues of* $(\mathbf{F} - \mathbf{KH}')$ *are all inside the unit circle. Then*

$$\begin{aligned}&\mathbf{H}'[\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{Q}[\mathbf{I}_r - \mathbf{F}'z^{-1}]^{-1}\mathbf{H} + \mathbf{R} \\ &= \{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{H}'\mathbf{PH} + \mathbf{R}\}\{\mathbf{I}_n + \mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1}\}.\end{aligned} \qquad [13.5.21]$$

A direct demonstration of this claim is provided in Appendix 13.A at the end of this chapter.

As an example of using these results, consider observations on a univariate $AR(1)$ process subject to white noise measurement error, such as the state-space system of [13.1.26] and [13.1.27] with $\mu = 0$. For this system, $\mathbf{F} = \phi$, $\mathbf{Q} = \sigma_V^2$, $\mathbf{A} = \mathbf{0}$, $\mathbf{H} = 1$, and $\mathbf{R} = \sigma_W^2$. The conditions of Proposition 13.2 are satisfied as long as $|\phi| < 1$, establishing that $|F - KH| = |\phi - K| < 1$. From equation

[13.5.14], the $AR(\infty)$ representation for this process can be found from

$$y_{t+1} = [1 - (\phi - K)L]^{-1}Ky_t + \varepsilon_{t+1},$$

which can be written

$$[1 - (\phi - K)L]y_{t+1} = Ky_t + [1 - (\phi - K)L]\varepsilon_{t+1}$$

or

$$y_{t+1} = \phi y_t + \varepsilon_{t+1} - (\phi - K)\varepsilon_t. \qquad [13.5.22]$$

This is an $ARMA(1, 1)$ process with AR parameter given by ϕ and MA parameter given by $-(\phi - K)$. The variance of the innovation for this process can be calculated from [13.5.16]:

$$E(\varepsilon_{t+1}^2) = \sigma_W^2 + P. \qquad [13.5.23]$$

The value of P can be found by iterating on [13.5.1]:

$$\begin{aligned} P_{t+1|t} &= \phi^2[P_{t|t-1} - P_{t|t-1}^2/(\sigma_W^2 + P_{t|t-1})] + \sigma_V^2 \\ &= \phi^2 P_{t|t-1}\sigma_W^2/(\sigma_W^2 + P_{t|t-1}) + \sigma_V^2, \end{aligned} \qquad [13.5.24]$$

starting from $P_{1|0} = \sigma_V^2/(1 - \phi^2)$, until convergence. The steady-state Kalman gain is given by [13.5.4]:

$$K = \phi P/(\sigma_W^2 + P). \qquad [13.5.25]$$

As a second example, consider adding an $MA(q_1)$ process to an $MA(q_2)$ process with which the first process is uncorrelated at all leads and lags. This could be represented in state-space form as follows:

State Equation $(r = q_1 + q_2 + 2)$:

$$\begin{bmatrix} u_{t+1} \\ u_t \\ \vdots \\ u_{t-q_1+1} \\ v_{t+1} \\ v_t \\ \vdots \\ v_{t-q_2+1} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{I}_{q_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{q_2} & \mathbf{0} \end{bmatrix}}_{(q_1+q_2+2)\times(q_1+q_2+2)} \begin{bmatrix} u_t \\ u_{t-1} \\ \vdots \\ u_{t-q_1} \\ v_t \\ v_{t-1} \\ \vdots \\ v_{t-q_2} \end{bmatrix} + \begin{bmatrix} u_{t+1} \\ 0 \\ \vdots \\ 0 \\ v_{t+1} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad [13.5.26]$$

Observation Equation $(n = 1)$:

$$y_t = [1 \quad \delta_1 \quad \delta_2 \quad \cdots \quad \delta_{q_1} \quad 1 \quad \kappa_1 \quad \kappa_2 \quad \cdots \quad \kappa_{q_2}] \begin{bmatrix} u_t \\ u_{t-1} \\ \vdots \\ u_{t-q_1} \\ v_t \\ v_{t-1} \\ \vdots \\ v_{t-q_2} \end{bmatrix}. \qquad [13.5.27]$$

Note that all eigenvalues of $\mathbf{F}$ are equal to zero. Write equation [13.5.18] in the form

$$\begin{aligned} \mathbf{y}_{t+1} &= \{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}L)^{-1}\mathbf{K}L\}\varepsilon_{t+1} \\ &= \{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r + \mathbf{F}L + \mathbf{F}^2L^2 + \mathbf{F}^3L^3 + \cdots)\mathbf{K}L\}\varepsilon_{t+1}. \end{aligned} \qquad [13.5.28]$$

Let $q \equiv \max\{q_1, q_2\}$, and notice from the structure of $\mathbf{F}$ that $\mathbf{F}^{q+j} = \mathbf{0}$ for $j = 1, 2, \ldots$. Furthermore, from [13.5.4], $\mathbf{F}^q\mathbf{K} = \mathbf{F}^{q+1}\mathbf{P}\mathbf{H}(\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R})^{-1} = \mathbf{0}$. Thus

[13.5.28] takes the form

$$\begin{aligned} y_{t+1} &= \{1 + \mathbf{H}'(\mathbf{I}_r + \mathbf{F}L + \mathbf{F}^2L^2 + \mathbf{F}^3L^3 \\ &\quad + \cdots + \mathbf{F}^{q-1}L^{q-1})\mathbf{K}L\}\varepsilon_{t+1} \qquad [13.5.29] \\ &= \{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\}\varepsilon_{t+1}, \end{aligned}$$

where

$$\theta_j = \mathbf{H}'\mathbf{F}^{j-1}\mathbf{K} \qquad \text{for } j = 1, 2, \ldots, q.$$

This provides a constructive demonstration of the claim that an $MA(q_1)$ process plus an $MA(q_2)$ process with which it is uncorrelated can be described as an $MA(\max\{q_1, q_2\})$ process.

The Kalman filter thus provides a general algorithm for finding the Wold representation or factoring an autocovariance-generating function—we simply iterate on [13.5.1] until convergence, and then use the steady-state gain from [13.5.4] either in [13.5.14] (for the $AR(\infty)$ form) or in [13.5.18] (for the $MA(\infty)$ form).

Although the convergent values provide the Wold representation, for any finite t the Kalman filter forecasts have the advantage of calculating the exact optimal forecast of $\mathbf{y}_{t+1}$ based on a linear function of $\{\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1\}$.

13.6. *Smoothing*

The Kalman filter was motivated in Section 13.2 as an algorithm for calculating a forecast of the state vector $\boldsymbol{\xi}_t$ as a linear function of previous observations,

$$\hat{\boldsymbol{\xi}}_{t|t-1} \equiv \hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_{t-1}), \qquad [13.6.1]$$

where $\mathcal{Y}_{t-1} \equiv (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_1, \mathbf{x}'_{t-1}, \mathbf{x}'_{t-2}, \ldots, \mathbf{x}_1)'$. The matrix $\mathbf{P}_{t|t-1}$ represented the MSE of this forecast:

$$\mathbf{P}_{t|t-1} \equiv E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t-1})']. \qquad [13.6.2]$$

For many uses of the Kalman filter these are the natural magnitudes of interest. In some settings, however, the state vector $\boldsymbol{\xi}_t$ is given a structural interpretation, in which case the value of this unobserved variable might be of interest for its own sake. For example, in the model of the business cycle by Stock and Watson, it would be helpful to know the state of the business cycle at any historical date t. A goal might then be to form an inference about the value of $\boldsymbol{\xi}_t$ based on the full set of data collected, including observations on $\mathbf{y}_t, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_T, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_T$. Such an inference is called the *smoothed* estimate of $\boldsymbol{\xi}_t$, denoted

$$\hat{\boldsymbol{\xi}}_{t|T} \equiv \hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_T). \qquad [13.6.3]$$

For example, data on GNP from 1954 through 1990 might be used to estimate the value that $\boldsymbol{\xi}$ took on in 1960. The *MSE* of this smoothed estimate is denoted

$$\mathbf{P}_{t|T} \equiv E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T})']. \qquad [13.6.4]$$

In general, $\mathbf{P}_{t|\tau}$ denotes the *MSE* of an estimate of $\boldsymbol{\xi}_t$ that is based on observations of $\mathbf{y}$ and $\mathbf{x}$ through date τ.

For the reader's convenience, we reproduce here the key equations for the Kalman filter:

$$\hat{\boldsymbol{\xi}}_{t|t} = \hat{\boldsymbol{\xi}}_{t|t-1} + \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\boldsymbol{\xi}}_{t|t-1}) \qquad [13.6.5]$$

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}\hat{\boldsymbol{\xi}}_{t|t} \qquad [13.6.6]$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{H}'\mathbf{P}_{t|t-1}\mathbf{H} + \mathbf{R})^{-1}\mathbf{H}'\mathbf{P}_{t|t-1} \qquad [13.6.7]$$

$$\mathbf{P}_{t+1|t} = \mathbf{F}\mathbf{P}_{t|t}\mathbf{F}' + \mathbf{Q}. \qquad [13.6.8]$$

Consider the estimate of ξ_t based on observations through date t, $\hat{\xi}_{t|t}$. Suppose we were subsequently told the true value of ξ_{t+1}. From the formula for updating a linear projection, equation [4.5.30], the new estimate of ξ_t could be expressed as[3]

$$\begin{aligned}\hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_t) = \hat{\xi}_{t|t} &+ \{E[(\xi_t - \hat{\xi}_{t|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})']\} \\ &\times \{E[(\xi_{t+1} - \hat{\xi}_{t+1|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})']\}^{-1} \\ &\times (\xi_{t+1} - \hat{\xi}_{t+1|t}).\end{aligned} \tag{13.6.9}$$

The first term in the product on the right side of [13.6.9] can be written

$$E[(\xi_t - \hat{\xi}_{t|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'] = E[(\xi_t - \hat{\xi}_{t|t})(\mathbf{F}\xi_t + \mathbf{v}_{t+1} - \mathbf{F}\hat{\xi}_{t|t})'],$$

by virtue of [13.2.1] and [13.6.6]. Furthermore, $\mathbf{v}_{t+1}$ is uncorrelated with ξ_t and $\hat{\xi}_{t|t}$. Thus,

$$E[(\xi_t - \hat{\xi}_{t|t})(\xi_{t+1} - \hat{\xi}_{t+1|t})'] = E[(\xi_t - \hat{\xi}_{t|t})(\xi_t - \hat{\xi}_{t|t})'\mathbf{F}'] = \mathbf{P}_{t|t}\mathbf{F}'. \tag{13.6.10}$$

Substituting [13.6.10] and the definition of $\mathbf{P}_{t+1|t}$ into [13.6.9] produces

$$\hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_t) = \hat{\xi}_{t|t} + \mathbf{P}_{t|t}\mathbf{F}'\mathbf{P}_{t+1|t}^{-1}(\xi_{t+1} - \hat{\xi}_{t+1|t}).$$

Defining

$$\mathbf{J}_t \equiv \mathbf{P}_{t|t}\mathbf{F}'\mathbf{P}_{t+1|t}^{-1}, \tag{13.6.11}$$

we have

$$\hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_t) = \hat{\xi}_{t|t} + \mathbf{J}_t(\xi_{t+1} - \hat{\xi}_{t+1|t}). \tag{13.6.12}$$

Now, the linear projection in [13.6.12] turns out to be the same as

$$\hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_T); \tag{13.6.13}$$

that is, knowledge of $\mathbf{y}_{t+j}$ or $\mathbf{x}_{t+j}$ for $j > 0$ would be of no added value if we already knew the value of ξ_{t+1}. To see this, note that $\mathbf{y}_{t+j}$ can be written as

$$\mathbf{y}_{t+j} = \mathbf{A}'\mathbf{x}_{t+j} + \mathbf{H}'(\mathbf{F}^{j-1}\xi_{t+1} + \mathbf{F}^{j-2}\mathbf{v}_{t+2} + \mathbf{F}^{j-3}\mathbf{v}_{t+3} + \cdots + \mathbf{v}_{t+j}) + \mathbf{w}_{t+j}.$$

But the error

$$\xi_t - \hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_t) \tag{13.6.14}$$

is uncorrelated with ξ_{t+1}, by the definition of a linear projection, and uncorrelated with $\mathbf{x}_{t+j}$, $\mathbf{w}_{t+j}$, $\mathbf{v}_{t+j}$, $\mathbf{v}_{t+j-1}$, . . . , $\mathbf{v}_{t+2}$ under the maintained assumptions. Thus, the error [13.6.14] is uncorrelated with $\mathbf{y}_{t+j}$ or $\mathbf{x}_{t+j}$ for $j > 0$, meaning that [13.6.13] and [13.6.12] are the same, as claimed:

$$\hat{E}(\xi_t|\xi_{t+1}, \mathcal{Y}_T) = \hat{\xi}_{t|t} + \mathbf{J}_t(\xi_{t+1} - \hat{\xi}_{t+1|t}). \tag{13.6.15}$$

It follows from the law of iterated projections that the smoothed estimate, $\hat{E}(\xi_t|\mathcal{Y}_T)$, can be obtained by projecting [13.6.15] on $\mathcal{Y}_T$. In calculating this projection, we need to think carefully about the nature of the magnitudes in [13.6.15]. The first term, $\hat{\xi}_{t|t}$, indicates a particular *exact* linear function of $\mathcal{Y}_t$; the coefficients of this function are constructed from population moments, and these coefficients should be viewed as deterministic constants from the point of view of performing a subsequent projection. The projection of $\hat{\xi}_{t|t}$ on $\mathcal{Y}_T$ is thus still $\hat{\xi}_{t|t}$, this same

[3]Here, $\mathbf{Y}_3 = \xi_t$, $\mathbf{Y}_2 = \xi_{t+1}$, and $\mathbf{Y}_1 = \mathcal{Y}_t$.

linear function of $\mathcal{Y}_t$—we can't improve on a perfect fit![4] The term $\mathbf{J}_t$ in [13.6.11] is also a function of population moments, and so is again treated as deterministic for purposes of any linear projection. The term $\hat{\boldsymbol{\xi}}_{t+1|t}$ is another exact linear function of $\mathcal{Y}_t$. Thus, projecting [13.6.15] on $\mathcal{Y}_T$ turns out to be trivial:

$$\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_T) = \hat{\boldsymbol{\xi}}_{t|t} + \mathbf{J}_t[\hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_T) - \hat{\boldsymbol{\xi}}_{t+1|t}],$$

or

$$\hat{\boldsymbol{\xi}}_{t|T} = \hat{\boldsymbol{\xi}}_{t|t} + \mathbf{J}_t(\hat{\boldsymbol{\xi}}_{t+1|T} - \hat{\boldsymbol{\xi}}_{t+1|t}). \qquad [13.6.16]$$

Thus, the sequence of smoothed estimates $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^{T}$ is calculated as follows. First, the Kalman filter, [13.6.5] to [13.6.8], is calculated and the sequences $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$, $\{\hat{\boldsymbol{\xi}}_{t+1|t}\}_{t=0}^{T-1}$, $\{\mathbf{P}_{t|t}\}_{t=1}^{T}$, and $\{\mathbf{P}_{t+1|t}\}_{t=0}^{T-1}$ are stored. The smoothed estimate for the final date in the sample, $\hat{\boldsymbol{\xi}}_{T|T}$, is just the last entry in $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$. Next, [13.6.11] is used to generate $\{\mathbf{J}_t\}_{t=1}^{T-1}$. From this, [13.6.16] is used for $t = T - 1$ to calculate

$$\hat{\boldsymbol{\xi}}_{T-1|T} = \hat{\boldsymbol{\xi}}_{T-1|T-1} + \mathbf{J}_{T-1}(\hat{\boldsymbol{\xi}}_{T|T} - \hat{\boldsymbol{\xi}}_{T|T-1}).$$

Now that $\hat{\boldsymbol{\xi}}_{T-1|T}$ has been calculated, [13.6.16] can be used for $t = T - 2$ to evaluate

$$\hat{\boldsymbol{\xi}}_{T-2|T} = \hat{\boldsymbol{\xi}}_{T-2|T-2} + \mathbf{J}_{T-2}(\hat{\boldsymbol{\xi}}_{T-1|T} - \hat{\boldsymbol{\xi}}_{T-1|T-2}).$$

Proceeding backward through the sample in this fashion permits calculation of the full set of smoothed estimates, $\{\hat{\boldsymbol{\xi}}_{t|T}\}_{t=1}^{T}$.

Next, consider the mean squared error associated with the smoothed estimate. Subtracting both sides of [13.6.16] from $\boldsymbol{\xi}_t$ produces

$$\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T} = \boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t} - \mathbf{J}_t\hat{\boldsymbol{\xi}}_{t+1|T} + \mathbf{J}_t\hat{\boldsymbol{\xi}}_{t+1|t}$$

or

$$\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T} + \mathbf{J}_t\hat{\boldsymbol{\xi}}_{t+1|T} = \boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t} + \mathbf{J}_t\hat{\boldsymbol{\xi}}_{t+1|t}.$$

Multiplying this equation by its transpose and taking expectations,

$$\begin{aligned} &E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T})'] + \mathbf{J}_t E[(\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T})]\mathbf{J}'_t \\ &\quad = E[(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})(\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t})'] + \mathbf{J}_t E[(\hat{\boldsymbol{\xi}}_{t+1|t}\hat{\boldsymbol{\xi}}'_{t+1|t})]\mathbf{J}'_t. \end{aligned} \qquad [13.6.17]$$

The cross-product terms have disappeared from the left side because $\hat{\boldsymbol{\xi}}_{t+1|T}$ is a linear function of $\mathcal{Y}_T$ and so is uncorrelated with the projection error $\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|T}$. Similarly, on the right side, $\hat{\boldsymbol{\xi}}_{t+1|t}$ is uncorrelated with $\boldsymbol{\xi}_t - \hat{\boldsymbol{\xi}}_{t|t}$.

Equation [13.6.17] states that

$$\mathbf{P}_{t|T} = \mathbf{P}_{t|t} + \mathbf{J}_t\{-E[(\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T})] + E[(\hat{\boldsymbol{\xi}}_{t+1|t}\hat{\boldsymbol{\xi}}'_{t+1|t})]\}\mathbf{J}'_t. \qquad [13.6.18]$$

[4]The law of iterated projections states that

$$\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_t) = \hat{E}[\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_T)|\mathcal{Y}_t].$$

The law of iterated projections thus allows us to go from a larger information set to a smaller. Of course, the same operation does not work in reverse:

$$\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_T) \neq \hat{E}[\hat{E}(\boldsymbol{\xi}_t|\mathcal{Y}_t)|\mathcal{Y}_T].$$

We cannot go from a smaller information set to a larger.

An example may clarify this point. Let y_t be an i.i.d. zero-mean sequence with

$$\xi_t = \mu + y_{t+1}.$$

Then

$$\hat{E}(\xi_t|y_t) = \mu$$

and

$$\hat{E}[\hat{E}(\xi_t|y_t)|y_t, y_{t+1}] = \hat{E}[\mu|y_t, y_{t+1}] = \mu.$$

The bracketed term in [13.6.18] can be expressed as

$$
\begin{aligned}
&-E[(\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T})] + E[(\hat{\boldsymbol{\xi}}_{t+1|t}\hat{\boldsymbol{\xi}}'_{t+1|t})] \\
&= \{E[(\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}'_{t+1})] - E[(\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T})]\} - \{E[(\boldsymbol{\xi}_{t+1}\boldsymbol{\xi}'_{t+1})] - E[(\hat{\boldsymbol{\xi}}_{t+1|t}\hat{\boldsymbol{\xi}}'_{t+1|t})]\} \\
&= \{E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|T})(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|T})']\} - \{E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|t})']\} \\
&= \mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t}.
\end{aligned}
\tag{13.6.19}
$$

The second-to-last equality used the fact that

$$
\begin{aligned}
E[\boldsymbol{\xi}_{t+1}\hat{\boldsymbol{\xi}}'_{t+1|T}] &= E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|T} + \hat{\boldsymbol{\xi}}_{t+1|T})\hat{\boldsymbol{\xi}}'_{t+1|T}] \\
&= E[(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|T})\hat{\boldsymbol{\xi}}'_{t+1|T}] + E[\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T}] \\
&= E[\hat{\boldsymbol{\xi}}_{t+1|T}\hat{\boldsymbol{\xi}}'_{t+1|T}],
\end{aligned}
$$

since the projection error $(\boldsymbol{\xi}_{t+1} - \hat{\boldsymbol{\xi}}_{t+1|T})$ is uncorrelated with $\hat{\boldsymbol{\xi}}_{t+1|T}$. Similarly, $E(\boldsymbol{\xi}_{t+1}\hat{\boldsymbol{\xi}}'_{t+1|t}) = E(\hat{\boldsymbol{\xi}}_{t+1|t}\hat{\boldsymbol{\xi}}'_{t+1|t})$. Substituting [13.6.19] into [13.6.18] establishes that the smoothed estimate $\hat{\boldsymbol{\xi}}_{t|T}$ has *MSE* given by

$$
\mathbf{P}_{t|T} = \mathbf{P}_{t|t} + \mathbf{J}_t(\mathbf{P}_{t+1|T} - \mathbf{P}_{t+1|t})\mathbf{J}'_t. \tag{13.6.20}
$$

Again, this sequence is generated by moving through the sample backward starting with $t = T - 1$.

13.7. *Statistical Inference with the Kalman Filter*

The calculation of the mean squared error

$$
\mathbf{P}_{\tau|t} = E[(\boldsymbol{\xi}_\tau - \hat{\boldsymbol{\xi}}_{\tau|t})(\boldsymbol{\xi}_\tau - \hat{\boldsymbol{\xi}}_{\tau|t})']
$$

described earlier assumed that the parameters of the matrices $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ were known with certainty. Section 13.4 showed how these parameters could be estimated from the data by maximum likelihood. There would then be some sampling uncertainty about the true values of these parameters, and the calculation of $\mathbf{P}_{\tau|t}$ would need to be modified to obtain the true mean squared errors of the smoothed estimates and forecasts.[5]

Suppose the unknown parameters are collected in a vector $\boldsymbol{\theta}$. For any given value of $\boldsymbol{\theta}$, the matrices $\mathbf{F}(\boldsymbol{\theta})$, $\mathbf{Q}(\boldsymbol{\theta})$, $\mathbf{A}(\boldsymbol{\theta})$, $\mathbf{H}(\boldsymbol{\theta})$, and $\mathbf{R}(\boldsymbol{\theta})$ could be used to construct $\hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta})$ and $\mathbf{P}_{\tau|T}(\boldsymbol{\theta})$ in the formulas presented earlier; for $\tau \leq T$, these are the smoothed estimate and *MSE* given in [13.6.16] and [13.6.20], respectively; while for $\tau > T$, these are the forecast and its *MSE* in [13.3.25] and [13.3.27]. Let $\mathcal{Y}_T \equiv (\mathbf{y}'_T, \mathbf{y}'_{T-1}, \ldots, \mathbf{y}'_1, \mathbf{x}'_T, \mathbf{x}'_{T-1}, \ldots, \mathbf{x}'_1)'$ denote the observed data, and let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. The earlier derivations assumed that the true value of $\boldsymbol{\theta}$ was used to construct $\hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta}_0)$ and $\mathbf{P}_{\tau|T}(\boldsymbol{\theta}_0)$.

Recall that the formulas for updating a linear projection and its *MSE*, [4.5.30] and [4.5.31], yield the conditional mean and conditional *MSE* when applied to Gaussian vectors; see equation [4.6.7]. Thus, if $\{\mathbf{v}_t\}$, $\{\mathbf{w}_t\}$, and $\boldsymbol{\xi}_1$ are truly Gaussian, then the linear projection $\hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta}_0)$ has the interpretation as the expectation of $\boldsymbol{\xi}_\tau$ conditional on the data,

$$
\hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta}_0) = E(\boldsymbol{\xi}_\tau|\mathcal{Y}_T); \tag{13.7.1}
$$

while $\mathbf{P}_{\tau|T}(\boldsymbol{\theta}_0)$ can be described as the conditional *MSE*:

$$
\mathbf{P}_{\tau|T}(\boldsymbol{\theta}_0) = E\{[\boldsymbol{\xi}_\tau - \hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta}_0)][\boldsymbol{\xi}_\tau - \hat{\boldsymbol{\xi}}_{\tau|T}(\boldsymbol{\theta}_0)]'|\mathcal{Y}_T\}. \tag{13.7.2}
$$

Let $\hat{\boldsymbol{\theta}}$ denote an estimate of $\boldsymbol{\theta}$ based on $\mathcal{Y}_T$, and let $\hat{\boldsymbol{\xi}}_{\tau|T}(\hat{\boldsymbol{\theta}})$ denote the estimate that results from using $\hat{\boldsymbol{\theta}}$ to construct the smoothed inference or forecast in [13.6.16]

[5]This discussion is based on Hamilton (1986).

or [13.3.25]. The conditional mean squared error of this estimate is

$$\begin{aligned}
&E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\theta})][\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\} \\
&\quad = E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0) + \hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})] \\
&\qquad \times [\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0) + \hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\} \qquad [13.7.3] \\
&\quad = E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0)][\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0)]'|\mathcal{Y}_T\} \\
&\qquad + E\{[\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})][\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\}.
\end{aligned}$$

Cross-product terms have disappeared from [13.7.3], since

$$\begin{aligned}
&E\{[\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})][\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0)]'|\mathcal{Y}_T\} \\
&\quad = [\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})] \times E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\theta_0)]'|\mathcal{Y}_T\} \\
&\quad = [\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})] \times 0'.
\end{aligned}$$

The first equality follows because $\hat{\xi}_{\tau|T}(\theta_0)$ and $\hat{\xi}_{\tau|T}(\hat{\theta})$ are known nonstochastic functions of $\mathcal{Y}_T$, and the second equality is implied by [13.7.1]. Substituting [13.7.2] into [13.7.3] results in

$$\begin{aligned}
&E\{[\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\theta})][\xi_\tau - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\} \\
&\quad = \mathbf{P}_{\tau|T}(\theta_0) + E\{[\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})][\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\}. \qquad [13.7.4]
\end{aligned}$$

Equation [13.7.4] decomposes the mean squared error into two components. The first component, $\mathbf{P}_{\tau|T}(\theta_0)$, might be described as the "filter uncertainty." This is the term calculated from the smoothing iteration [13.6.20] or forecast *MSE* [13.3.27] and represents uncertainty about ξ_τ that would be present even if the true value θ_0 were known with certainty. The second term in [13.7.4],

$$E\{[\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})][\hat{\xi}_{\tau|T}(\theta_0) - \hat{\xi}_{\tau|T}(\hat{\theta})]'\},$$

might be called "parameter uncertainty." It reflects the fact that in a typical sample, $\hat{\theta}$ will differ from the true value θ_0.

A simple way to estimate the size of each source of uncertainty is by Monte Carlo integration. Suppose we adopt the Bayesian perspective that θ itself is a random variable. From this perspective, [13.7.4] describes the *MSE* conditional on $\theta = \theta_0$. Suppose that the posterior distribution of θ conditional on the data $\mathcal{Y}_T$ is known; the asymptotic distribution for the *MLE* in [13.4.6] suggests that $\theta|\mathcal{Y}_T$ might be regarded as approximately distributed $N(\hat{\theta}, (1/T)\cdot\mathcal{I}^{-1})$, where $\hat{\theta}$ denotes the *MLE*. We might then generate a large number of values of θ, say, $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(2000)}$, drawn from a $N(\hat{\theta}, (1/T)\cdot\mathcal{I}^{-1})$ distribution. For each draw (j), we could calculate the smoothed estimate or forecast $\hat{\xi}_{\tau|T}(\theta^{(j)})$. The deviations of these estimates across Monte Carlo draws from the estimate $\hat{\xi}_{\tau|T}(\hat{\theta})$ can be used to describe how sensitive the estimate $\hat{\xi}_{\tau|T}(\hat{\theta})$ is to parameter uncertainty about θ:

$$\frac{1}{2000}\sum_{j=1}^{2000}[\hat{\xi}_{\tau|T}(\theta^{(j)}) - \hat{\xi}_{\tau|T}(\hat{\theta})][\hat{\xi}_{\tau|T}(\theta^{(j)}) - \hat{\xi}_{\tau|T}(\hat{\theta})]'. \qquad [13.7.5]$$

This affords an estimate of

$$E\{[\hat{\xi}_{\tau|T}(\theta) - \hat{\xi}_{\tau|T}(\hat{\theta})][\hat{\xi}_{\tau|T}(\theta) - \hat{\xi}_{\tau|T}(\hat{\theta})]'|\mathcal{Y}_T\},$$

where this expectation is understood to be with respect to the distribution of θ conditional on $\mathcal{Y}_T$.

For each Monte Carlo realization $\theta^{(j)}$, we can also calculate $\mathbf{P}_{\tau|T}(\theta^{(j)})$ from [13.6.20] or [13.3.27]. Its average value across Monte Carlo draws,

$$\frac{1}{2000}\sum_{j=1}^{2000}\mathbf{P}_{\tau|T}(\theta^{(j)}), \qquad [13.7.6]$$

provides an estimate of the filter uncertainty in [13.7.4],

$$E[\mathbf{P}_{\tau|T}(\boldsymbol{\theta})|\mathcal{Y}_T].$$

Again, this expectation is with respect to the distribution of $\boldsymbol{\theta}|\mathcal{Y}_T$.

The sum of [13.7.5] and [13.7.6] is then proposed as an *MSE* for the estimate $\hat{\boldsymbol{\xi}}_{\tau|T}(\hat{\boldsymbol{\theta}})$ around the true value $\boldsymbol{\xi}_\tau$.

13.8. *Time-Varying Parameters*

State-Space Model with Stochastically Varying Coefficients

Up to this point we have been assuming that the matrices **F**, **Q**, **A**, **H**, and **R** were all constant. The Kalman filter can also be adapted for more general state-space models in which the values of these matrices depend on the exogenous or lagged dependent variables included in the vector $\mathbf{x}_t$. Consider

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}(\mathbf{x}_t)\boldsymbol{\xi}_t + \mathbf{v}_{t+1} \qquad [13.8.1]$$

$$\mathbf{y}_t = \mathbf{a}(\mathbf{x}_t) + [\mathbf{H}(\mathbf{x}_t)]'\boldsymbol{\xi}_t + \mathbf{w}_t. \qquad [13.8.2]$$

Here $\mathbf{F}(\mathbf{x}_t)$ denotes an $(r \times r)$ matrix whose elements are functions of $\mathbf{x}_t$; $\mathbf{a}(\mathbf{x}_t)$ similarly describes an $(n \times 1)$ vector-valued function, and $\mathbf{H}(\mathbf{x}_t)$ an $(r \times n)$ matrix-valued function. It is assumed that conditional on $\mathbf{x}_t$ and on data observed through date $t - 1$, denoted

$$\mathcal{Y}_{t-1} \equiv (\mathbf{y}'_{t-1}, \mathbf{y}'_{t-2}, \ldots, \mathbf{y}'_1, \mathbf{x}'_{t-1}, \mathbf{x}'_{t-2}, \ldots, \mathbf{x}'_1)',$$

the vector $(\mathbf{v}'_{t+1}, \mathbf{w}'_t)'$ has the Gaussian distribution

$$\left[\begin{matrix}\mathbf{v}_{t+1}\\ \mathbf{w}_t\end{matrix}\,\middle|\,\mathbf{x}_t, \mathcal{Y}_{t-1}\right] \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{Q}(\mathbf{x}_t) & \mathbf{0}\\ \mathbf{0} & \mathbf{R}(\mathbf{x}_t)\end{bmatrix}\right). \qquad [13.8.3]$$

Note that although [13.8.1] to [13.8.3] generalize the earlier framework by allowing for stochastically varying parameters, it is more restrictive in that a Gaussian distribution is assumed in [13.8.3]; the role of the Gaussian requirement will be explained shortly.

Suppose it is taken as given that $\boldsymbol{\xi}_t|\mathcal{Y}_{t-1} \sim N(\hat{\boldsymbol{\xi}}_{t|t-1}, \mathbf{P}_{t|t-1})$. Assuming as before that $\mathbf{x}_t$ contains only strictly exogenous variables or lagged values of $\mathbf{y}$, this also describes the distribution of $\boldsymbol{\xi}_t|\mathbf{x}_t, \mathcal{Y}_{t-1}$. It follows from the assumptions in [13.8.1] to [13.8.3] that

$$\left[\begin{matrix}\boldsymbol{\xi}_t\\ \mathbf{y}_t\end{matrix}\,\middle|\,\mathbf{x}_t, \mathcal{Y}_{t-1}\right] \sim N\left(\begin{bmatrix}\hat{\boldsymbol{\xi}}_{t|t-1}\\ \mathbf{a}(\mathbf{x}_t) + [\mathbf{H}(\mathbf{x}_t)]'\hat{\boldsymbol{\xi}}_{t|t-1}\end{bmatrix}, \begin{bmatrix}\mathbf{P}_{t|t-1} & \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t)\\ \mathbf{H}'(\mathbf{x}_t)\mathbf{P}_{t|t-1} & [\mathbf{H}(\mathbf{x}_t)]'\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) + \mathbf{R}(\mathbf{x}_t)\end{bmatrix}\right). \qquad [13.8.4]$$

Conditional on $\mathbf{x}_t$, the terms $\mathbf{a}(\mathbf{x}_t)$, $\mathbf{H}(\mathbf{x}_t)$, and $\mathbf{R}(\mathbf{x}_t)$ can all be treated as deterministic. Thus, the formula for the conditional distribution of Gaussian vectors [4.6.7] can be used to deduce that[6]

$$\boldsymbol{\xi}_t|\mathbf{y}_t, \mathbf{x}_t, \mathcal{Y}_{t-1} \equiv \boldsymbol{\xi}_t|\mathcal{Y}_t \sim N(\hat{\boldsymbol{\xi}}_{t|t}, \mathbf{P}_{t|t}), \qquad [13.8.5]$$

[6]Here $\mathbf{Y}_1 = \mathbf{y}_t$, $\mathbf{Y}_2 = \boldsymbol{\xi}_t$, $\boldsymbol{\mu}_1 = \mathbf{a}(\mathbf{x}_t) + [\mathbf{H}(\mathbf{x}_t)]'\hat{\boldsymbol{\xi}}_{t|t-1}$, $\boldsymbol{\mu}_2 = \hat{\boldsymbol{\xi}}_{t|t-1}$, $\boldsymbol{\Omega}_{11} = \{[\mathbf{H}(\mathbf{x}_t)]'\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) + \mathbf{R}(\mathbf{x}_t)\}$, $\boldsymbol{\Omega}_{22} = \mathbf{P}_{t|t-1}$, and $\boldsymbol{\Omega}_{21} = \mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t)$.

where

$$\hat{\boldsymbol{\xi}}_{t|t} = \hat{\boldsymbol{\xi}}_{t|t-1} + \Big\{\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t)[[\mathbf{H}(\mathbf{x}_t)]'\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) + \mathbf{R}(\mathbf{x}_t)]^{-1} \times [\mathbf{y}_t - \mathbf{a}(\mathbf{x}_t) - [\mathbf{H}(\mathbf{x}_t)]'\hat{\boldsymbol{\xi}}_{t|t-1}]\Big\} \qquad [13.8.6]$$

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \Big\{\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) \times [[\mathbf{H}(\mathbf{x}_t)]'\mathbf{P}_{t|t-1}\mathbf{H}(\mathbf{x}_t) + \mathbf{R}(\mathbf{x}_t)]^{-1}[\mathbf{H}(\mathbf{x}_t)]'\mathbf{P}_{t|t-1}\Big\}. \qquad [13.8.7]$$

It then follows from [13.8.1] and [13.8.3] that $\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t \sim N(\hat{\boldsymbol{\xi}}_{t+1|t}, \mathbf{P}_{t+1|t})$, where

$$\hat{\boldsymbol{\xi}}_{t+1|t} = \mathbf{F}(\mathbf{x}_t)\hat{\boldsymbol{\xi}}_{t|t} \qquad [13.8.8]$$

$$\mathbf{P}_{t+1|t} = \mathbf{F}(\mathbf{x}_t)\mathbf{P}_{t|t}[\mathbf{F}(\mathbf{x}_t)]' + \mathbf{Q}(\mathbf{x}_t). \qquad [13.8.9]$$

Equations [13.8.6] through [13.8.9] are just the Kalman filter equations [13.2.15], [13.2.16], [13.2.17], and [13.2.21] with the parameter matrices $\mathbf{F}$, $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{H}$, and $\mathbf{R}$ replaced by their time-varying analogs. Thus, as long as we are willing to treat the initial state $\boldsymbol{\xi}_1$ as $N(\hat{\boldsymbol{\xi}}_{1|0}, \mathbf{P}_{1|0})$, the Kalman filter iterations go through the same as before. The obvious generalization of [13.4.1] can continue to be used to evaluate the likelihood function.

Note, however, that unlike the constant-parameter case, the inference [13.8.6] is a nonlinear function of $\mathbf{x}_t$. This means that although [13.8.6] gives the optimal inference if the disturbances and initial state are Gaussian, it cannot be interpreted as the linear projection of $\boldsymbol{\xi}_t$ on $\mathcal{Y}_t$ with non-Gaussian disturbances.

Linear Regression Models with Time-Varying Coefficients

One important application of the state-space model with stochastically varying parameters is a regression in which the coefficient vector changes over time. Consider

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_t + w_t, \qquad [13.8.10]$$

where $\mathbf{x}_t$ is a $(k \times 1)$ vector that can include lagged values of y or variables that are independent of the regression disturbance w_τ for all τ. The parameters of the coefficient vector are presumed to evolve over time according to

$$(\boldsymbol{\beta}_{t+1} - \bar{\boldsymbol{\beta}}) = \mathbf{F}(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}) + \mathbf{v}_{t+1}. \qquad [13.8.11]$$

If the eigenvalues of the $(k \times k)$ matrix $\mathbf{F}$ are all inside the unit circle, then $\bar{\boldsymbol{\beta}}$ has the interpretation as the average or steady-state value for the coefficient vector. If it is further assumed that

$$\begin{bmatrix} \mathbf{v}_{t+1} \\ w_t \end{bmatrix} \Bigg| \mathbf{x}_t, \mathcal{Y}_{t-1} \Bigg] \sim N\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0}' & \sigma^2 \end{bmatrix}\right), \qquad [13.8.12]$$

then [13.8.10] to [13.8.12] will be recognized as a state-space model of the form of [13.8.1] to [13.8.3] with state vector $\boldsymbol{\xi}_t = \boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}}$. The regression in [13.8.10] can be written as

$$y_t = \mathbf{x}_t'\bar{\boldsymbol{\beta}} + \mathbf{x}_t'\boldsymbol{\xi}_t + w_t, \qquad [13.8.13]$$

which is an observation equation of the form of [13.8.2] with $\mathbf{a}(\mathbf{x}_t) = \mathbf{x}_t'\bar{\boldsymbol{\beta}}$, $\mathbf{H}(\mathbf{x}_t) = \mathbf{x}_t$, and $\mathbf{R}(\mathbf{x}_t) = \sigma^2$. These values are then used in the Kalman filter iterations [13.8.6] to [13.8.9]. A one-period-ahead forecast for [13.8.10] can be calculated from [13.8.4] as

$$E(y_t|\mathbf{x}_t, \mathcal{Y}_{t-1}) = \mathbf{x}_t'\bar{\boldsymbol{\beta}} + \mathbf{x}_t'\hat{\boldsymbol{\xi}}_{t|t-1},$$

where $\{\hat{\boldsymbol{\xi}}_{t|t-1}\}_{t=1}^{T}$ is calculated from [13.8.6] and [13.8.8]. The *MSE* of this forecast can also be inferred from [13.8.4]:

$$E[(y_t - \mathbf{x}_t'\bar{\boldsymbol{\beta}} - \mathbf{x}_t'\hat{\boldsymbol{\xi}}_{t|t-1})^2|\mathbf{x}_t, \mathcal{Y}_{t-1}] = \mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2,$$

where $\{\mathbf{P}_{t|t-1}\}_{t=1}^{T}$ is calculated from [13.8.7] and [13.8.9]. The sample log likelihood is therefore

$$\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_t, \mathcal{Y}_{t-1}) = -(T/2)\log(2\pi) - (1/2)\sum_{t=1}^{T}\log(\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2)$$
$$- (1/2)\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\bar{\boldsymbol{\beta}} - \mathbf{x}_t'\hat{\boldsymbol{\xi}}_{t|t-1})^2/(\mathbf{x}_t'\mathbf{P}_{t|t-1}\mathbf{x}_t + \sigma^2).$$

The specification in [13.8.11] can easily be generalized to allow for a pth-order *VAR* for the coefficient vector $\boldsymbol{\beta}_t$ by defining $\boldsymbol{\xi}_t' \equiv [(\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})', (\boldsymbol{\beta}_{t-1} - \bar{\boldsymbol{\beta}})', \ldots, (\boldsymbol{\beta}_{t-p+1} - \bar{\boldsymbol{\beta}})']$ and replacing [13.8.11] with

$$\boldsymbol{\xi}_{t+1} = \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \cdots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p \\ \mathbf{I}_k & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_k & \mathbf{0} \end{bmatrix} \boldsymbol{\xi}_t + \begin{bmatrix} \mathbf{v}_{t+1} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

Estimation of a VAR *with Time-Varying Coefficients*

Section 12.2 described Litterman's approach to Bayesian estimation of an equation of a vector autoregression with constant but unknown coefficients. A related approach to estimating a *VAR* with time-varying coefficients was developed by Doan, Litterman, and Sims (1984). Although efficiency might be improved by estimating all the equations of the *VAR* jointly, their proposal was to infer the parameters for each equation in isolation from the others.

Suppose for illustration that equation [13.8.10] describes the first equation from a *VAR*, so that the dependent variable (y_t) is y_{1t} and the ($k \times 1$) vector of explanatory variables is $\mathbf{x}_t = (1, \mathbf{y}_{t-1}', \mathbf{y}_{t-2}', \ldots, \mathbf{y}_{t-p}')'$, where $\mathbf{y}_t = (y_{1t}, y_{2t}, \ldots, y_{nt})'$ and $k = np + 1$. The coefficient vector is

$$\boldsymbol{\beta}_t = (c_{1,t}, \phi_{11,t}^{(1)}, \phi_{12,t}^{(1)}, \ldots, \phi_{1n,t}^{(1)}, \phi_{11,t}^{(2)}, \phi_{12,t}^{(2)}, \ldots, \phi_{1n,t}^{(2)}, \ldots,$$
$$\phi_{11,t}^{(p)}, \phi_{12,t}^{(p)}, \ldots, \phi_{1n,t}^{(p)})',$$

where $\phi_{1j,t}^{(s)}$ is the coefficient relating y_{1t} to $y_{j,t-s}$. This coefficient is allowed to be different for each date t in the sample.

Doan, Litterman, and Sims specified a Bayesian prior distribution for the initial value of the coefficient vector at date 1:

$$\boldsymbol{\beta}_1 \sim N(\bar{\boldsymbol{\beta}}, \mathbf{P}_{1|0}). \qquad [13.8.14]$$

The prior distribution is independent across coefficients, so that $\mathbf{P}_{1|0}$ is a diagonal matrix. The mean of the prior distribution, $\bar{\boldsymbol{\beta}}$, is that used by Litterman (1986) for a constant-coefficient *VAR*. This prior distribution holds that changes in y_{1t} are probably difficult to forecast, so that the coefficient on $y_{1,t-1}$ is likely to be near unity and all other coefficients are expected to be near zero:

$$\bar{\boldsymbol{\beta}} = (0, 1, 0, 0, \ldots, 0)'. \qquad [13.8.15]$$

As in Section 12.2, let γ characterize the analyst's confidence in the prediction that $\phi_{11,1}^{(1)}$ is near unity:

$$\phi_{11,1}^{(1)} \sim N(1, \gamma^2).$$

Smaller values of γ imply more confidence in the prior conviction that $\phi_{11,1}^{(1)}$ is near unity.

The coefficient $\phi_{11,1}^{(s)}$ relates the value of variable 1 at date 1 to its own value

s periods earlier. Doan, Litterman, and Sims had more confidence in the prior conviction that $\phi_{11,1}^{(s)}$ is zero the greater the lag, or the larger the value of s. They represented this with a harmonic series for the variance,

$$\phi_{11,1}^{(s)} \sim N(0, \gamma^2/s) \qquad \text{for } s = 2, 3, \ldots, p.$$

The prior distribution for the coefficient relating variable 1 to lags of other variables was taken to be

$$\phi_{1j,1}^{(s)} \sim N\left(0, \frac{w^2 \cdot \gamma^2 \cdot \hat{\tau}_1^2}{s \cdot \hat{\tau}_j^2}\right) \qquad \begin{matrix} j = 2, 3, \ldots, n \\ s = 1, 2, \ldots, p \end{matrix}. \qquad [13.8.16]$$

As in expression [12.2.4], this includes a correction $(\hat{\tau}_1^2/\hat{\tau}_j^2)$ for the scale of y_{1t} relative to y_{jt}, where $\hat{\tau}_j^2$ is the estimated variance of the residuals for a univariate fixed-coefficient $AR(p)$ process fitted to series j. The variance in [13.8.16] also includes a factor $w^2 < 1$ representing the prior expectation that lagged values of y_j for $j \neq 1$ are less likely to be of help in forecasting y_1 than would be the lagged values of y_1 itself; hence, a tighter prior is used to set coefficients on y_j to zero.

Finally, let g describe the variance of the prior distribution for the constant term:

$$c_{1,1} \sim N(0, g \cdot \hat{\tau}_1^2).$$

To summarize, the matrix $\mathbf{P}_{1|0}$ is specified to be

$$\mathbf{P}_{1|0} = \begin{bmatrix} g \cdot \hat{\tau}_1^2 & \mathbf{0}' \\ \mathbf{0} & (\mathbf{B} \otimes \mathbf{C}) \end{bmatrix}, \qquad [13.8.17]$$

where

$$\underset{(p \times p)}{\mathbf{B}} = \begin{bmatrix} \gamma^2 & 0 & 0 & \cdots & 0 \\ 0 & \gamma^2/2 & 0 & \cdots & 0 \\ 0 & 0 & \gamma^2/3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \gamma^2/p \end{bmatrix}$$

$$\underset{(n \times n)}{\mathbf{C}} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & w^2\hat{\tau}_1^2/\hat{\tau}_2^2 & 0 & \cdots & 0 \\ 0 & 0 & w^2\hat{\tau}_1^2/\hat{\tau}_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & w^2\hat{\tau}_1^2/\hat{\tau}_n^2 \end{bmatrix}.$$

For typical economic time series, Doan, Litterman, and Sims recommended using $\gamma^2 = 0.07$, $w^2 = 1/74$, and $g = 630$. This last value ensures that very little weight is given to the prior expectation that the constant term is zero.

Each of the coefficients in the VAR is then presumed to evolve over time according to a first-order autoregression:

$$\boldsymbol{\beta}_{t+1} = \pi_8 \cdot \boldsymbol{\beta}_t + (1 - \pi_8) \cdot \bar{\boldsymbol{\beta}} + \mathbf{v}_{t+1}. \qquad [13.8.18]$$

Thus, the same scalar π_8 is used to describe a univariate $AR(1)$ process for each element of $\boldsymbol{\beta}_t$; Doan, Litterman, and Sims recommended a value of $\pi_8 = 0.999$. The disturbance $\mathbf{v}_t$ is assumed to have a diagonal variance-covariance matrix:

$$E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{Q}. \qquad [13.8.19]$$

For all coefficients except the constant term, the variance of the ith element of $\mathbf{v}_t$ was assumed to be proportional to the corresponding element of $\mathbf{P}_{1|0}$. Thus, for $i = 2, 3, \ldots, k$, the row i, column i element of $\mathbf{Q}$ is taken to be π_7 times the row i, column i element of $\mathbf{P}_{1|0}$. The (1, 1) element of $\mathbf{Q}$ is taken to be π_7 times the (2, 2) element of $\mathbf{P}_{1|0}$. This adjustment is used because the (1, 1) element of $\mathbf{P}_{1|0}$ represents an effectively infinite variance corresponding to prior ignorance about

the value for the constant term. Doan, Litterman, and Sims recommended $\pi_7 = 10^{-7}$ as a suitable value for the constant of proportionality.

Equation [13.8.18] can be viewed as a state equation of the form

$$\boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1}, \qquad [13.8.20]$$

where the state vector is given by $\boldsymbol{\xi}_t = (\boldsymbol{\beta}_t - \bar{\boldsymbol{\beta}})$ and $\mathbf{F} = \pi_8 \cdot \mathbf{I}_k$. The observation equation is

$$y_{1t} = \mathbf{x}_t'\bar{\boldsymbol{\beta}} + \mathbf{x}_t'\boldsymbol{\xi}_t + w_{1t}. \qquad [13.8.21]$$

The one parameter yet to be specified is the variance of w_{1t}, the residual in the *VAR*. Doan, Litterman, and Sims suggested taking this to be 0.9 times $\hat{\tau}_1^2$.

Thus, the sequence of estimated state vectors $\{\hat{\boldsymbol{\xi}}_{t|t}\}_{t=1}^{T}$ is found by iterating on [13.8.6] through [13.8.9] starting from $\hat{\boldsymbol{\xi}}_{1|0} = \mathbf{0}$ and $\mathbf{P}_{1|0}$ given by [13.8.17], with $\mathbf{F}(\mathbf{x}_t) = \pi_8 \cdot \mathbf{I}_k$, $\mathbf{Q}(\mathbf{x}_t) = \pi_7 \cdot \mathbf{P}_{1|0}$, $a(\mathbf{x}_t) = \mathbf{x}_t'\bar{\boldsymbol{\beta}}$ with $\bar{\boldsymbol{\beta}}$ given by [13.8.15], $\mathbf{H}(\mathbf{x}_t) = \mathbf{x}_t$, and $\mathbf{R}(\mathbf{x}_t) = 0.9 \cdot \hat{\tau}_1^2$. The estimated coefficient vector is then $\hat{\boldsymbol{\beta}}_{t|t} = \bar{\boldsymbol{\beta}} + \hat{\boldsymbol{\xi}}_{t|t}$. Optimal one-period-ahead forecasts are given by $\hat{y}_{1,t+1|t} = \mathbf{x}_t'\hat{\boldsymbol{\beta}}_{t|t}$.

Optimal s-period-ahead forecasts are difficult to calculate. However, Doan, Litterman, and Sims suggested a simple approximation. The approximation takes the optimal one-period-ahead forecasts for each of the n variables in the *VAR*, $\hat{\mathbf{y}}_{t+1|t}$, and then treats these forecasts as if they were actual observations on $\mathbf{y}_{t+1}$. Then $E(\mathbf{y}_{t+2}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1)$ is approximated by $E(\mathbf{y}_{t+2}|\mathbf{y}_{t+1}, \mathbf{y}_t, \ldots, \mathbf{y}_1)$ evaluated at $\mathbf{y}_{t+1} = E(\mathbf{y}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_1)$. The law of iterated expectations does not apply here, since $E(\mathbf{y}_{t+2}|\mathbf{y}_{t+1}, \mathbf{y}_t, \ldots, \mathbf{y}_1)$ is a nonlinear function of $\mathbf{y}_{t+1}$. However, Doan, Litterman, and Sims argued that this simple approach gives a good approximation to the optimal forecast.

APPENDIX 13.A. *Proofs of Chapter 13 Propositions*

■ **Proof of Proposition 13.1.**[7] Recall that $\mathbf{P}_{t+1|t}$ has the interpretation as the *MSE* of the linear projection of $\boldsymbol{\xi}_{t+1}$ on $\mathcal{Y}_t \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_1', \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_1')'$,

$$\mathbf{P}_{t+1|t} = MSE[\hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t)]. \qquad [13.A.1]$$

Suppose for some reason we instead tried to forecast $\boldsymbol{\xi}_{t+1}$ using only observations $2, 3, \ldots, t$, discarding the observation for date $t = 1$. Thus, define $\mathcal{Y}_t^* \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_2', \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_2')'$, and let

$$\mathbf{P}_{t+1|t}^* \equiv MSE[\hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t^*)]. \qquad [13.A.2]$$

Then clearly, [13.A.2] cannot be smaller than [13.A.1], since the linear projection $\hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t)$ made optimal use of $\mathcal{Y}_t^*$ along with the added information in $(\mathbf{y}_1', \mathbf{x}_1')'$. Specifically, if $\mathbf{h}$ is any $(r \times 1)$ vector, the linear projection of $z_{t+1} \equiv \mathbf{h}'\boldsymbol{\xi}_{t+1}$ on $\mathcal{Y}_t$ has *MSE* given by

$$\begin{aligned} E[z_{t+1} - \hat{E}(z_{t+1}|\mathcal{Y}_t)]^2 &= E[\mathbf{h}'\boldsymbol{\xi}_{t+1} - \mathbf{h}' \cdot \hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t)]^2 \\ &= \mathbf{h}' \cdot E\{[\boldsymbol{\xi}_{t+1} - \hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t)][\boldsymbol{\xi}_{t+1} - \hat{E}(\boldsymbol{\xi}_{t+1}|\mathcal{Y}_t)]'\} \cdot \mathbf{h} \\ &= \mathbf{h}'\mathbf{P}_{t+1|t}\mathbf{h}. \end{aligned}$$

Similarly, the linear projection of z_{t+1} on $\mathcal{Y}_t^*$ has *MSE* $\mathbf{h}'\mathbf{P}_{t+1|t}^*\mathbf{h}$, with

$$\mathbf{h}'\mathbf{P}_{t+1|t}\mathbf{h} \le \mathbf{h}'\mathbf{P}_{t+1|t}^*\mathbf{h}. \qquad [13.A.3]$$

But for a system of the form of [13.2.1] and [13.2.2] with eigenvalues of $\mathbf{F}$ inside the unit circle and time-invariant coefficients, it will be the case that

$$\begin{aligned} &MSE[\hat{E}(\boldsymbol{\xi}_{t+1}|\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{y}_2, \mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_2)] \\ &\qquad = MSE[\hat{E}(\boldsymbol{\xi}_t|\mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_1)], \end{aligned}$$

that is,

$$\mathbf{P}_{t+1|t}^* = \mathbf{P}_{t|t-1}.$$

Hence, [13.A.3] implies that

$$\mathbf{h}'\mathbf{P}_{t+1|t}\mathbf{h} \le \mathbf{h}'\mathbf{P}_{t|t-1}\mathbf{h}$$

[7]The arguments in the proofs of Propositions 13.1 and 13.2 are adapted from Anderson and Moore (1979, pp. 76–82).

for any $(r \times 1)$ vector $\mathbf{h}$. The sequence of scalars $\{\mathbf{h}'\mathbf{P}_{t+1|t}\mathbf{h}\}_{t=1}^{T}$ is thus monotonically nonincreasing and is bounded below by zero. It therefore converges to some fixed nonnegative value. Since this is true for any $(r \times 1)$ vector $\mathbf{h}$ and since the matrix $\mathbf{P}_{t+1|t}$ is symmetric, it follows that the sequence $\{\mathbf{P}_{t+1|t}\}_{t=1}^{T}$ converges to some fixed positive semidefinite matrix $\mathbf{P}$.

To verify the claims about the eigenvalues of the matrix $(\mathbf{F} - \mathbf{KH}')$, note that if $\mathbf{P}$ is a fixed point of [13.5.3], then it must also be a fixed point of the equivalent difference equation [13.2.28]:

$$\mathbf{P} = (\mathbf{F} - \mathbf{KH}')\mathbf{P}(\mathbf{F} - \mathbf{KH}')' + \mathbf{KRK}' + \mathbf{Q}. \qquad [13.A.4]$$

Let $\mathbf{x}$ denote an eigenvector of $(\mathbf{F} - \mathbf{KH}')'$ and λ its eigenvalue:

$$(\mathbf{F} - \mathbf{KH}')'\mathbf{x} = \lambda\mathbf{x}. \qquad [13.A.5]$$

Although $\mathbf{F}$, $\mathbf{K}$, and $\mathbf{H}$ are all real, the eigenvalue λ and eigenvector $\mathbf{x}$ could be complex. If $\mathbf{x}''$ denotes the conjugate transpose of $\mathbf{x}$, then

$$\begin{aligned}\mathbf{x}''(\mathbf{F} - \mathbf{KH}')\mathbf{P}(\mathbf{F} - \mathbf{KH}')'\mathbf{x} &= [(\mathbf{F} - \mathbf{KH}')'\mathbf{x}]''\mathbf{P}[(\mathbf{F} - \mathbf{KH}')'\mathbf{x}] \\ &= [\lambda\mathbf{x}]''\mathbf{P}[\lambda\mathbf{x}] \\ &= |\lambda|^2\mathbf{x}''\mathbf{P}\mathbf{x}.\end{aligned}$$

Thus, if [13.A.4] is premultiplied by $\mathbf{x}''$ and postmultiplied by $\mathbf{x}$, the result is

$$\mathbf{x}''\mathbf{P}\mathbf{x} = |\lambda|^2\mathbf{x}''\mathbf{P}\mathbf{x} + \mathbf{x}''(\mathbf{KRK}' + \mathbf{Q})\mathbf{x},$$

or

$$(1 - |\lambda|^2)\mathbf{x}''\mathbf{P}\mathbf{x} = \mathbf{x}''(\mathbf{KRK}' + \mathbf{Q})\mathbf{x}. \qquad [13.A.6]$$

Now, $(\mathbf{KRK}' + \mathbf{Q})$ is positive semidefinite, so the right side of [13.A.6] is nonnegative. Likewise, $\mathbf{P}$ is positive semidefinite, so $\mathbf{x}''\mathbf{P}\mathbf{x}$ is nonnegative. Expression [13.A.6] then requires $|\lambda| \leq 1$, meaning that any eigenvalue of $(\mathbf{F} - \mathbf{KH}')$ must be on or inside the unit circle, as claimed. ■

■ **Proof of Proposition 13.2.** First we establish the final claim of the proposition, concerning the eigenvalues of $(\mathbf{F} - \mathbf{KH}')$. Let $\mathbf{P}$ denote any positive semidefinite matrix that satisfies [13.A.4], and let $\mathbf{K}$ be given by [13.5.4]. Notice that if $\mathbf{Q}$ is positive definite, then the right side of [13.A.6] is strictly positive for any nonzero $\mathbf{x}$, meaning from the left side of [13.A.6] that any eigenvalue λ of $(\mathbf{F} - \mathbf{KH}')$ is strictly inside the unit circle. Alternatively, if $\mathbf{R}$ is positive definite, then the only way that the right side of [13.A.6] could fail to be strictly positive would be if $\mathbf{K}'\mathbf{x} = \mathbf{0}$. But from [13.A.5], this would imply that $\mathbf{F}'\mathbf{x} = \lambda\mathbf{x}$, that is, that $\mathbf{x}$ is an eigenvector and λ is an eigenvalue of $\mathbf{F}'$. This, in turn, means that λ is an eigenvalue of $\mathbf{F}$, in which case $|\lambda| < 1$, by the assumption of stability of $\mathbf{F}$. Thus, there cannot be an eigenvector $\mathbf{x}$ of $(\mathbf{F} - \mathbf{KH}')'$ associated with an eigenvalue whose modulus is greater than or equal to unity if $\mathbf{R}$ is positive definite.

Turning next to the rest of Proposition 13.2, let $\{\mathbf{P}_{t+1|t}\}$ denote the sequence that results from iterating on [13.5.1] starting from an arbitrary positive semidefinite initial value $\mathbf{P}_{1|0}$. We will show that there exist two other sequences of matrices, to be denoted $\{\underline{\mathbf{P}}_{t+1|t}\}$ and $\{\tilde{\mathbf{P}}_{t+1|t}\}$, such that

$$\underline{\mathbf{P}}_{t+1|t} \leq \mathbf{P}_{t+1|t} \leq \tilde{\mathbf{P}}_{t+1|t} \qquad \text{for all } t,$$

where

$$\lim_{t\to\infty} \underline{\mathbf{P}}_{t+1|t} = \lim \tilde{\mathbf{P}}_{t+1|t} = \mathbf{P}$$

and where $\mathbf{P}$ does not depend on $\mathbf{P}_{1|0}$. The conclusion will then be that $\{\mathbf{P}_{t+1|t}\}$ converges to $\mathbf{P}$ regardless of the value of $\mathbf{P}_{1|0}$.

To construct the matrix $\underline{\mathbf{P}}_{t+1|t}$ that is to be offered as a lower bound on $\mathbf{P}_{t+1|t}$, consider the sequence $\{\underline{\mathbf{P}}_{t+1|t}\}$ that results from iterating on [13.5.1] starting from the initial value $\underline{\mathbf{P}}_{1|0} = \mathbf{0}$. This would correspond to treating the intial state ξ_1 as if known with certainty:

$$\underline{\mathbf{P}}_{t+1|t} = MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t, \xi_1)]. \qquad [13.A.7]$$

Note that $\mathbf{y}_t$ and $\mathbf{x}_t$ are correlated with ξ_{t+1} for $t = 1, 2, \ldots$ only through the value of ξ_1, which means that we could equally well write

$$\underline{\mathbf{P}}_{t+1|t} = MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_1)], \qquad [13.A.8]$$

where $\mathcal{Y}_t^* = (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_2', \mathbf{x}_t', \mathbf{x}_{t-1}', \ldots, \mathbf{x}_2')'$. Added knowledge about ξ_2 could not hurt the forecast:

$$MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_2, \xi_1)] \leq MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_1)], \qquad [13.A.9]$$

and indeed, ξ_1 is correlated with ξ_{t+1} for $t = 2, 3, \ldots$ only through the value of ξ_2:

$$MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_2, \xi_1)] = MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_2)]. \qquad [13.A.10]$$

Because coefficients are time-invariant,

$$MSE[\hat{E}(\xi_{t+1}|\mathcal{Y}_t^*, \xi_2)] = MSE[\hat{E}(\xi_t|\mathcal{Y}_{t-1}, \xi_1)] = \underline{\mathbf{P}}_{t|t-1}. \qquad [13.A.11]$$

Thus, [13.A.10] and [13.A.11] establish that the left side of [13.A.9] is the same as $\underline{\mathbf{P}}_{t|t-1}$, while from [13.A.8] the right side of [13.A.9] is the same as $\underline{\mathbf{P}}_{t+1|t}$. Thus, [13.A.9] states that

$$\underline{\mathbf{P}}_{t|t-1} \leq \underline{\mathbf{P}}_{t+1|t},$$

so that $\{\underline{\mathbf{P}}_{t+1|t}\}$ is a monotonically nondecreasing sequence; the farther in the past is the perfect information about ξ_1, the less value it is for forecasting ξ_{t+1}.

Furthermore, a forecast based on perfect information about ξ_1, for which $\underline{\mathbf{P}}_{t+1|t}$ gives the *MSE*, must be better than one based on imperfect information about ξ_1, for which $\mathbf{P}_{t+1|t}$ gives the *MSE*:

$$\underline{\mathbf{P}}_{t+1|t} \leq \mathbf{P}_{t+1|t} \qquad \text{for all } t.$$

Thus, $\underline{\mathbf{P}}_{t+1|t}$ puts a lower bound on $\mathbf{P}_{t+1|t}$, as claimed. Moreover, since the sequence $\{\underline{\mathbf{P}}_{t+1|t}\}$ is monotonically nondecreasing and bounded from above, it converges to a fixed value $\mathbf{P}$ satisfying [13.5.3] and [13.A.4].

To construct an upper bound on $\mathbf{P}_{t+1|t}$, consider a sequence $\{\tilde{\mathbf{P}}_{t+1|t}\}$ that begins with $\tilde{\mathbf{P}}_{1|0} = \mathbf{P}_{1|0}$, the same starting value that was used to construct $\{\mathbf{P}_{t+1|t}\}$. Recall that $\mathbf{P}_{t+1|t}$ gave the *MSE* of the sequence $\hat{\xi}_{t+1|t}$ described in equation [13.2.20]:

$$\hat{\xi}_{t+1|t} = \mathbf{F}\hat{\xi}_{t|t-1} + \mathbf{K}_t(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\hat{\xi}_{t|t-1}).$$

Imagine instead using a sequence of suboptimal inferences $\{\tilde{\xi}_{t+1|t}\}$ defined by the recursion

$$\tilde{\xi}_{t+1|t} = \mathbf{F}\tilde{\xi}_{t|t-1} + \mathbf{K}(\mathbf{y}_t - \mathbf{A}'\mathbf{x}_t - \mathbf{H}'\tilde{\xi}_{t|t-1}), \qquad [13.A.12]$$

where $\mathbf{K}$ is the value calculated from [13.5.4] in which the steady-state value for $\mathbf{P}$ is taken to be the limit of the sequence $\{\underline{\mathbf{P}}_{t+1|t}\}$. Note that the magnitude $\tilde{\xi}_{t+1|t}$ so defined is a linear function of $\mathcal{Y}_t$ and so must have a greater *MSE* than the optimal inference $\hat{\xi}_{t+1|t}$:

$$\tilde{\mathbf{P}}_{t+1|t} \equiv E[(\xi_{t+1} - \tilde{\xi}_{t+1|t})(\xi_{t+1} - \tilde{\xi}_{t+1|t})'] \geq \mathbf{P}_{t+1|t}.$$

Thus, we have established that

$$\underline{\mathbf{P}}_{t+1|t} \leq \mathbf{P}_{t+1|t} \leq \tilde{\mathbf{P}}_{t+1|t}$$

and that $\underline{\mathbf{P}}_{t+1|t} \to \mathbf{P}$. The proof will be complete if we can further show that $\tilde{\mathbf{P}}_{t+1|t} \to \mathbf{P}$. Parallel calculations to those leading to [13.2.28] reveal that

$$\tilde{\mathbf{P}}_{t+1|t} = (\mathbf{F} - \mathbf{KH}')\tilde{\mathbf{P}}_{t|t-1}(\mathbf{F} - \mathbf{KH}')' + \mathbf{KRK}' + \mathbf{Q}. \qquad [13.A.13]$$

Apply the vec operator to both sides of [13.A.13] and recall Proposition 10.4:

$$\text{vec}(\tilde{\mathbf{P}}_{t+1|t}) = \mathcal{B}\,\text{vec}(\tilde{\mathbf{P}}_{t|t-1}) + c = [\mathbf{I}_{r^2} + \mathcal{B} + \mathcal{B}^2 + \cdots + \mathcal{B}^{t-1}]c + \mathcal{B}^t\,\text{vec}(\tilde{\mathbf{P}}_{1|0}),$$

where

$$\mathcal{B} \equiv (\mathbf{F} - \mathbf{KH}') \otimes (\mathbf{F} - \mathbf{KH}')$$
$$c \equiv \text{vec}(\mathbf{KRK}' + \mathbf{Q}).$$

Recall further that since either $\mathbf{R}$ or $\mathbf{Q}$ is positive definite, the value of $\mathbf{K}$ has the property that all eigenvalues of $(\mathbf{F} - \mathbf{KH}')$ are strictly less than unity in modulus. Thus, all eigenvalues of $\mathcal{B}$ are also strictly less than unity in modulus, implying that

$$\lim_{t\to\infty} \text{vec}(\tilde{\mathbf{P}}_{t+1|t}) = (\mathbf{I}_{r^2} - \mathcal{B})^{-1}c,$$

the same value regardless of the starting value for $\tilde{\mathbf{P}}_{1|0}$. In particular, if the iteration on [13.A.13] is started with $\tilde{\mathbf{P}}_{1|0} = \mathbf{P}$, this being a fixed point of the iteration, the result would be $\tilde{\mathbf{P}}_{t+1|t} = \mathbf{P}$ for all t. Thus,

$$\lim_{t\to\infty} \tilde{\mathbf{P}}_{t+1|t} = \mathbf{P},$$

regardless of the value of $\tilde{\mathbf{P}}_{1|0} = \mathbf{P}_{1|0}$ from which the iteration for $\tilde{\mathbf{P}}_{t+1|t}$ is started. ■

■ **Proof of Proposition 13.3.** Observe that

$$\begin{aligned}
&\{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{I}_n - \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1}\mathbf{K}z\} \\
&\quad = \mathbf{I}_n - \mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1}\mathbf{K}z + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z \\
&\qquad - \{\mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{H}'[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1}\mathbf{K}z\} \qquad [13.A.14] \\
&\quad = \mathbf{I}_n + \mathbf{H}'\Big\{-[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1} + [\mathbf{I}_r - \mathbf{F}z]^{-1} \\
&\qquad - [\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{K}\mathbf{H}'z[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1}\Big\}\mathbf{K}z.
\end{aligned}$$

The term in curly braces in the last line of [13.A.14] is indeed zero, as may be verified by taking the identity

$$-[\mathbf{I}_r - \mathbf{F}z] + [\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z] - \mathbf{K}\mathbf{H}'z = \mathbf{0}$$

and premultiplying by $[\mathbf{I}_r - \mathbf{F}z]^{-1}$ and postmultiplying by $[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1}$:

$$\begin{aligned}
&-[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1} + [\mathbf{I}_r - \mathbf{F}z]^{-1} \\
&\quad - [\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{K}\mathbf{H}'z[\mathbf{I}_r - (\mathbf{F} - \mathbf{K}\mathbf{H}')z]^{-1} = \mathbf{0}. \quad ■ \qquad [13.A.15]
\end{aligned}$$

■ **Proof of Proposition 13.4.** Notice that

$$\begin{aligned}
&\{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\{\mathbf{I}_n + \mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1}\} \\
&\quad = \{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\} + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}z \qquad [13.A.16] \\
&\qquad + \{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1} \\
&\qquad + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}.
\end{aligned}$$

Now [13.5.4] requires that

$$\mathbf{K}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\} = \mathbf{F}\mathbf{P}\mathbf{H} \qquad [13.A.17]$$

$$\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\mathbf{K}' = \mathbf{H}'\mathbf{P}\mathbf{F}' \qquad [13.A.18]$$

$$\begin{aligned}
\mathbf{K}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\mathbf{K}' &= \mathbf{F}\mathbf{P}\mathbf{H}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}^{-1}\mathbf{H}'\mathbf{P}\mathbf{F}' \\
&= \mathbf{F}\mathbf{P}\mathbf{F}' - \mathbf{P} + \mathbf{Q}, \qquad [13.A.19]
\end{aligned}$$

with the last equality following from [13.5.3]. Substituting [13.A.17] through [13.A.19] into [13.A.16] results in

$$\begin{aligned}
&\{\mathbf{I}_n + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{K}z\}\{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\}\{\mathbf{I}_n + \mathbf{K}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1}\} \\
&\quad = \{\mathbf{H}'\mathbf{P}\mathbf{H} + \mathbf{R}\} + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{F}\mathbf{P}\mathbf{H}z + \mathbf{H}'\mathbf{P}\mathbf{F}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H}z^{-1} \\
&\qquad + \mathbf{H}'(\mathbf{I}_r - \mathbf{F}z)^{-1}\{\mathbf{F}\mathbf{P}\mathbf{F}' - \mathbf{P} + \mathbf{Q}\}(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\mathbf{H} \qquad [13.A.20] \\
&\quad = \mathbf{R} + \mathbf{H}'\Big\{\mathbf{P} + (\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{F}\mathbf{P}z + \mathbf{P}\mathbf{F}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}z^{-1} \\
&\qquad + (\mathbf{I}_r - \mathbf{F}z)^{-1}\{\mathbf{F}\mathbf{P}\mathbf{F}' - \mathbf{P} + \mathbf{Q}\}(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}\Big\}\mathbf{H}.
\end{aligned}$$

The result in Proposition 13.4 follows provided that

$$\begin{aligned}
&\mathbf{P} + (\mathbf{I}_r - \mathbf{F}z)^{-1}\mathbf{F}\mathbf{P}z + \mathbf{P}\mathbf{F}'(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}z^{-1} \\
&\qquad + (\mathbf{I}_r - \mathbf{F}z)^{-1}\{\mathbf{F}\mathbf{P}\mathbf{F}' - \mathbf{P}\}(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1} = \mathbf{0}. \qquad [13.A.21]
\end{aligned}$$

To verify that [13.A.21] is true, start from the identity

$$\begin{aligned}
&(\mathbf{I}_r - \mathbf{F}z)\mathbf{P}(\mathbf{I}_r - \mathbf{F}'z^{-1}) + \mathbf{F}\mathbf{P}z(\mathbf{I}_r - \mathbf{F}'z^{-1}) \\
&\qquad + (\mathbf{I}_r - \mathbf{F}z)\mathbf{P}\mathbf{F}'z^{-1} + \mathbf{F}\mathbf{P}\mathbf{F}' - \mathbf{P} = \mathbf{0}. \qquad [13.A.22]
\end{aligned}$$

Premultiplying [13.A.22] by $(\mathbf{I}_r - \mathbf{F}z)^{-1}$ and postmultiplying by $(\mathbf{I}_r - \mathbf{F}'z^{-1})^{-1}$ confirms [13.A.21]. Substituting [13.A.21] into [13.A.20] produces the claim in Proposition 13.4. ■

Chapter 13 Exercises

13.1. Suppose we have a noisy indicator y on an underlying unobserved random variable ξ:

$$y = \xi + \varepsilon.$$

Suppose moreover that the measurement error (ε) is $N(0, \tau^2)$, while the true value ξ is $N(\mu, \sigma^2)$, with ε uncorrelated with ξ. Show that the optimal estimate of ξ is given by

$$E(\xi \mid y) = \mu + \frac{\sigma^2}{\tau^2 + \sigma^2}(y - \mu)$$

with associated MSE

$$E[\xi - E(\xi \mid y)]^2 = \frac{\sigma^2\tau^2}{\tau^2 + \sigma^2}.$$

Discuss the intuition for these results as $\tau^2 \to \infty$ and $\tau^2 \to 0$.

13.2. Deduce the state-space representation for an $AR(p)$ model in [13.1.14] and [13.1.15] and the state-space representation for an $MA(1)$ model given in [13.1.17] and [13.1.18] as special cases of that for the $ARMA(r, r-1)$ model of [13.1.22] and [13.1.23].

13.3. Is the following a valid state-space representation of an $MA(1)$ process?

State Equation:

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

Observation Equation:

$$y_t - \mu = [1 \quad \theta] \begin{bmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{bmatrix}$$

13.4. Derive equation [13.4.5] as a special case of [13.4.1] and [13.4.2] for the model specified in [13.4.3] and [13.4.4] by analysis of the Kalman filter recursion for this case.

13.5. Consider a particular $MA(1)$ representation of the form of [13.3.1] through [13.3.12] parameterized by (θ, σ^2) with $|\theta| < 1$. The noninvertible representation for the same process is parameterized by $(\tilde{\theta}, \tilde{\sigma}^2)$ with $\tilde{\theta} = 1/\theta$ and $\tilde{\sigma}^2 = \theta^2\sigma^2$. The forecast generated by the Kalman filter using the noninvertible representation satisfies

$$\hat{y}_{t+1|t} = \mathbf{A}'\mathbf{x}_{t+1} + \tilde{\mathbf{H}}'\hat{\boldsymbol{\xi}}_{t+1|t} = \mu + \tilde{\theta}\tilde{\varepsilon}_{t|t},$$

where $\tilde{\varepsilon}_{t|t} = \{\tilde{\sigma}^2/[\tilde{\sigma}^2 + \tilde{\theta}^2\tilde{p}_t]\} \cdot \{y_t - \mu - \tilde{\theta}\tilde{\varepsilon}_{t-1|t-1}\}$. The MSE of this forecast is

$$E(y_{t+1} - \hat{y}_{t+1|t})^2 = \tilde{\mathbf{H}}'\tilde{\mathbf{P}}_{t+1|t}\tilde{\mathbf{H}} + \tilde{\mathbf{R}} = \tilde{\sigma}^2 + \tilde{\theta}^2\tilde{p}_{t+1},$$

where $\tilde{p}_{t+1} = (\tilde{\sigma}^2\tilde{\theta}^{2t})/(1 + \tilde{\theta}^2 + \tilde{\theta}^4 + \cdots + \tilde{\theta}^{2t})$. Show that this forecast and MSE are identical to those for the process as parameterized using the invertible representation (θ, σ^2). Deduce that the likelihood function given by [13.4.1] and [13.4.2] takes on the same value at (θ, σ^2) as it does at $(\tilde{\theta}, \tilde{\sigma}^2)$.

13.6. Show that ε_t in equation [13.5.22] is fundamental for y_t. What principle of the Kalman filter ensures that this will be the case? Show that the first autocovariance of the implied $MA(1)$ error process is given by

$$-(\phi - K)E(\varepsilon_t^2) = -\phi\sigma_w^2$$

while the variance is

$$[1 + (\phi - K)^2]E(\varepsilon_t^2) = (1 + \phi^2)\sigma_w^2 + \sigma_v^2.$$

Derive these expressions independently, using the approach to sums of $ARMA$ processes in Section 4.7.

13.7. Consider again the invertible $MA(1)$ of equations [13.3.1] to [13.3.12]. We found that the steady-state value of $\mathbf{P}_{t|t-1}$ is given by

$$\mathbf{P} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}.$$

From this, deduce that the steady-state value of $\mathbf{P}_{t|t+s} = \mathbf{0}$ for $s = 0, 1, 2, \ldots$. Give the intuition for this result.

Chapter 13 References

Anderson, Brian D. O., and John B. Moore. 1979. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall.

Burmeister, Edwin, and Kent D. Wall. 1982. "Kalman Filtering Estimation of Unobserved Rational Expectations with an Application to the German Hyperinflation." *Journal of Econometrics* 20:255–84.

———, ———, and James D. Hamilton. 1986. "Estimation of Unobserved Expected Monthly Inflation Using Kalman Filtering." *Journal of Business and Economic Statistics* 4:147–60.

Caines, Peter E. 1988. *Linear Stochastic Systems*. New York: Wiley.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* Series B, 39:1–38.

Doan, Thomas, Robert B. Litterman, and Christopher A. Sims. 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions." *Econometric Reviews* 3:1–100.

Fama, Eugene F., and Michael R. Gibbons. 1982. "Inflation, Real Returns, and Capital Investment." *Journal of Monetary Economics* 9:297–323.

Gevers, M., and V. Wertz. 1984. "Uniquely Identifiable State-Space and ARMA Parameterizations for Multivariable Linear Systems." *Automatica* 20:333–47.

Ghosh, Damayanti. 1989. "Maximum Likelihood Estimation of the Dynamic Shock-Error Model." *Journal of Econometrics* 41:121–43.

Hamilton, James D. 1985. "Uncovering Financial Market Expectations of Inflation." *Journal of Political Economy* 93:1224–41.

———. 1986. "A Standard Error for the Estimated State Vector of a State-Space Model." *Journal of Econometrics* 33:387–97.

Hannan, E. J. 1971. "The Identification Problem for Multiple Equation Systems with Moving Average Errors." *Econometrica* 39:751–65.

Harvey, Andrew, and G. D. A. Phillips. 1979. "Maximum Likelihood Estimation of Regression Models with Autoregressive–Moving Average Disturbances." *Biometrika* 66:49–58.

Kalman, R. E. 1960. "A New Approach to Linear Filtering and Prediction Problems." *Journal of Basic Engineering, Transactions of the ASME* Series D, 82:35–45.

———. 1963. "New Methods in Wiener Filtering Theory," in John L. Bogdanoff and Frank Kozin, eds., *Proceedings of the First Symposium of Engineering Applications of Random Function Theory and Probability*, 270–388. New York: Wiley.

Litterman, Robert B. 1986. "Forecasting with Bayesian Vector Autoregressions—Five Years of Experience." *Journal of Business and Economic Statistics* 4:25–38.

Meinhold, Richard J., and Nozer D. Singpurwalla. 1983. "Understanding the Kalman Filter." *American Statistician* 37:123–27.

Nicholls, D. F., and A. R. Pagan. 1985. "Varying Coefficient Regression," in E. J. Hannan, P. R. Krishnaiah, and M. M. Rao, eds., *Handbook of Statistics*, Vol. 5. Amsterdam: North-Holland.

Pagan, Adrian. 1980. "Some Identification and Estimation Results for Regression Models with Stochastically Varying Coefficients." *Journal of Econometrics* 13:341–63.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39:577–91.

Shumway, R. H., and D. S. Stoffer. 1982. "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm." *Journal of Time Series Analysis* 3:253–64.

Sims, Christopher A. 1982. "Policy Analysis with Econometric Models." *Brookings Papers on Economic Activity* 1:107–52.

Stock, James H., and Mark W. Watson. 1991. "A Probability Model of the Coincident Economic Indicators," in Kajal Lahiri and Geoffrey H. Moore, eds., *Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge, England: Cambridge University Press.

Tanaka, Katsuto. 1983. "Non-Normality of the Lagrange Multiplier Statistic for Testing the Constancy of Regression Coefficients." *Econometrica* 51:1577–82.

Wall, Kent D. 1987. "Identification Theory for Varying Coefficient Regression Models." *Journal of Time Series Analysis* 8:359–71.

Watson, Mark W. 1989. "Recursive Solution Methods for Dynamic Linear Rational Expectations Models." *Journal of Econometrics* 41:65–89.

——— and Robert F. Engle. 1983. "Alternative Algorithms for the Estimation of Dynamic Factor, MIMIC, and Varying Coefficient Regression Models." *Journal of Econometrics* 23:385–400.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.

14 Generalized Method of Moments

Suppose we have a set of observations on a variable y_t whose probability law depends on an unknown vector of parameters $\boldsymbol{\theta}$. One general approach to estimating $\boldsymbol{\theta}$ is based on the principle of maximum likelihood—we choose as the estimate $\hat{\boldsymbol{\theta}}$ the value for which the data would be most likely to have been observed. A drawback of this approach is that it requires us to specify the form of the likelihood function.

This chapter explores an alternative principle for parameter estimation known as *generalized method of moments* (*GMM*). Although versions of this approach have been used for a long time, the general statement of *GMM* on which this chapter is based was only recently developed by Hansen (1982). The key advantage of *GMM* is that it requires specification only of certain moment conditions rather than the full density. This can also be a drawback, in that *GMM* often does not make efficient use of all the information in the sample.

Section 14.1 introduces the ideas behind *GMM* estimation and derives some of the key results. Section 14.2 shows how various other estimators can be viewed as special cases of *GMM*, including ordinary least squares, instrumental variable estimation, two-stage least squares, estimators for systems of nonlinear simultaneous equations, and estimators for dynamic rational expectations models. Extensions and further discussion are provided in Section 14.3. In many cases, even maximum likelihood estimation can be viewed as a special case of *GMM*. Section 14.4 explores this analogy and uses it to derive some general asymptotic properties of maximum likelihood and quasi-maximum likelihood estimation.

14.1. *Estimation by the Generalized Method of Moments*

Classical Method of Moments

It will be helpful to introduce the ideas behind *GMM* with a concrete example. Consider a random variable Y_t drawn from a standard t distribution with ν degrees of freedom, so that its density is

$$f_{Y_t}(y_t;\, \nu) = \frac{\Gamma[(\nu + 1)/2]}{(\pi\nu)^{1/2}\Gamma(\nu/2)}\,[1 + (y_t^2/\nu)]^{-(\nu+1)/2}, \qquad [14.1.1]$$

where $\Gamma(\cdot)$ is the gamma function. Suppose we have an i.i.d. sample of size T (y_1, y_2, . . . , y_T) and want to estimate the degrees of freedom parameter ν. One approach is to estimate ν by maximum likelihood. This approach calculates the

sample log likelihood

$$\mathcal{L}(\nu) = \sum_{t=1}^{T} \log f_{Y_t}(y_t;\, \nu)$$

and chooses as the estimate $\hat{\nu}$ the value for which $\mathcal{L}(\nu)$ is largest.

An alternative principle on which estimation of ν might be based reasons as follows. Provided that $\nu > 2$, a standard t variable has population mean zero and variance given by

$$\mu_2 \equiv E(Y_t^2) = \nu/(\nu - 2). \qquad [14.1.2]$$

As the degrees of freedom parameter (ν) goes to infinity, the variance [14.1.2] approaches unity and the density [14.1.1] approaches that of a standard $N(0, 1)$ variable. Let $\hat{\mu}_{2,T}$ denote the average squared value of y observed in the actual sample:

$$\hat{\mu}_{2,T} \equiv (1/T) \sum_{t=1}^{T} y_t^2. \qquad [14.1.3]$$

For large T, the sample moment ($\hat{\mu}_{2,T}$) should be close to the population moment (μ_2):

$$\hat{\mu}_{2,T} \xrightarrow{p} \mu_2.$$

Recalling [14.1.2], this suggests that a consistent estimate of ν can be obtained by finding a solution to

$$\nu/(\nu - 2) = \hat{\mu}_{2,T} \qquad [14.1.4]$$

or

$$\hat{\nu}_T = \frac{2 \cdot \hat{\mu}_{2,T}}{\hat{\mu}_{2,T} - 1}. \qquad [14.1.5]$$

This estimate exists provided that $\hat{\mu}_{2,T} > 1$, that is, provided that the sample seems to exhibit more variability than the $N(0, 1)$ distribution. If we instead observed $\hat{\mu}_{2,T} \leq 1$, the estimate of the degrees of freedom would be infinity—a $N(0, 1)$ distribution fits the sample second moment better than any member of the t family.

The estimator derived from [14.1.4] is known as a *classical method of moments* estimator. A general description of this approach is as follows. Given an unknown $(a \times 1)$ vector of parameters $\boldsymbol{\theta}$ that characterizes the density of an observed variable y_t, suppose that a distinct population moments of the random variable can be calculated as functions of $\boldsymbol{\theta}$, such as

$$E(Y_t^i) = \mu_i(\boldsymbol{\theta}) \qquad \text{for } i = i_1, i_2, \ldots, i_a. \qquad [14.1.6]$$

The classical method of moments estimate of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}_T$ for which these population moments are equated to the observed sample moments; that is, $\hat{\boldsymbol{\theta}}_T$ is the value for which

$$\mu_i(\hat{\boldsymbol{\theta}}_T) = (1/T) \sum_{t=1}^{T} y_t^i \qquad \text{for } i = i_1, i_2, \ldots, i_a.$$

An early example of this approach was provided by Pearson (1894).

Generalized Method of Moments

In the example of the t distribution just discussed, a single sample moment ($\hat{\mu}_{2,T}$) was used to estimate a single population parameter (ν). We might also have made use of other moments. For example, if $\nu > 4$, the population fourth moment of a standard t variable is

$$\mu_4 \equiv E(Y_t^4) = \frac{3\nu^2}{(\nu - 2)(\nu - 4)},$$

and we might expect this to be close to the sample fourth moment,

$$\hat{\mu}_{4,T} \equiv (1/T) \sum_{t=1}^{T} y_t^4.$$

We cannot choose the single parameter ν so as to match both the sample second moment and the sample fourth moment. However, we might try to choose ν so as to be as close as possible to both, by minimizing a criterion function such as

$$Q(\nu; y_T, y_{T-1}, \ldots, y_1) \equiv \mathbf{g}'\mathbf{W}\mathbf{g}, \qquad [14.1.7]$$

where

$$\mathbf{g} \equiv \begin{bmatrix} \left\{\hat{\mu}_{2,T} - \dfrac{\nu}{\nu - 2}\right\} \\ \left\{\hat{\mu}_{4,T} - \dfrac{3\nu^2}{(\nu - 2)(\nu - 4)}\right\} \end{bmatrix}. \qquad [14.1.8]$$

Here $\mathbf{W}$ is a (2×2) positive definite symmetric weighting matrix reflecting the importance given to matching each of the moments. The larger is the $(1, 1)$ element of $\mathbf{W}$, the greater is the importance of being as close as possible to satisfying [14.1.4].

An estimate based on minimization of an expression such as [14.1.7] was called a "minimum chi-square" estimator by Cramér (1946, p. 425), Ferguson (1958), and Rothenberg (1973) and a "minimum distance estimator" by Malinvaud (1970). Hansen (1982) provided the most general characterization of this approach and derived the asymptotic properties for serially dependent processes. Most of the results reported in this section were developed by Hansen (1982), who described this as estimation by the "generalized method of moments."

Hansen's formulation of the estimation problem is as follows. Let $\mathbf{w}_t$ be an $(h \times 1)$ vector of variables that are observed at date t, let $\boldsymbol{\theta}$ denote an unknown $(a \times 1)$ vector of coefficients, and let $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ be an $(r \times 1)$ vector-valued function, $\mathbf{h}$: $(\mathbb{R}^a \times \mathbb{R}^h) \rightarrow \mathbb{R}^r$. Since $\mathbf{w}_t$ is a random variable, so is $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and suppose this true value is characterized by the property that

$$E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\} = \mathbf{0}. \qquad [14.1.9]$$

The r rows of the vector equation [14.1.9] are sometimes described as *orthogonality conditions*. Let $\mathcal{Y}_T \equiv (\mathbf{w}_T', \mathbf{w}_{T-1}', \ldots, \mathbf{w}_1')'$ be a $(Th \times 1)$ vector containing all the observations in a sample of size T, and let the $(r \times 1)$ vector-valued function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ denote the sample average of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$:

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \equiv (1/T) \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t). \qquad [14.1.10]$$

Notice that $\mathbf{g}$: $\mathbb{R}^a \rightarrow \mathbb{R}^r$. The idea behind *GMM* is to choose $\boldsymbol{\theta}$ so as to make the sample moment $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ as close as possible to the population moment of zero; that is, the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is the value of $\boldsymbol{\theta}$ that minimizes the scalar

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\mathbf{W}_T[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)], \qquad [14.1.11]$$

where $\{\mathbf{W}_T\}_{T=1}^{\infty}$ is a sequence of $(r \times r)$ positive definite weighting matrices which may be a function of the data $\mathcal{Y}_T$. Often, this minimization is achieved numerically using the methods described in Section 5.7.

The classical method of moments estimator of ν given in [14.1.5] is a special case of this formulation with $\mathbf{w}_t = y_t$, $\boldsymbol{\theta} = \nu$, $\mathbf{W}_T = 1$, and

$$h(\boldsymbol{\theta}, \mathbf{w}_t) = y_t^2 - \nu/(\nu - 2)$$

$$g(\boldsymbol{\theta}; \mathcal{Y}_T) = (1/T) \sum_{t=1}^{T} y_t^2 - \nu/(\nu - 2).$$

Here, $r = a = 1$ and the objective function [14.1.11] becomes

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = \left\{(1/T) \sum_{t=1}^{T} y_t^2 - \nu/(\nu - 2)\right\}^2.$$

The smallest value that can be achieved for $Q(\cdot)$ is zero, which obtains when ν is the magnitude given in [14.1.5].

The estimate of ν obtained by minimizing [14.1.7] is also a *GMM* estimator with $r = 2$ and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} \left\{y_t^2 - \dfrac{\nu}{\nu - 2}\right\} \\ \left\{y_t^4 - \dfrac{3\nu^2}{(\nu - 2)(\nu - 4)}\right\} \end{bmatrix}.$$

Here, $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ and $\mathbf{W}_T$ would be as described in [14.1.7] and [14.1.8].

A variety of other estimators can also be viewed as examples of *GMM*, including ordinary least squares, instrumental variable estimation, two-stage least squares, nonlinear simultaneous equations estimators, estimators for dynamic rational expectations models, and in many cases even maximum likelihood. These applications will be discussed in Sections 14.2 through 14.4.

If the number of parameters to be estimated (a) is the same as the number of orthogonality conditions (r), then typically the objective function [14.1.11] will be minimized by setting

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = \mathbf{0}. \qquad [14.1.12]$$

If $a = r$, then the *GMM* estimator is the value $\hat{\boldsymbol{\theta}}_T$ that satisfies these r equations. If instead there are more orthogonality conditions than parameters to estimate ($r > a$), then [14.1.12] will not hold exactly. How close the ith element of $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ is to zero depends on how much weight the ith orthogonality condition is given by the weighting matrix $\mathbf{W}_T$.

For any value of $\boldsymbol{\theta}$, the magnitude of the $(r \times 1)$ vector $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ is the sample mean of T realizations of the $(r \times 1)$ random vector $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$. If $\mathbf{w}_t$ is strictly stationary and $\mathbf{h}(\cdot)$ is continuous, then it is reasonable to expect the law of large numbers to hold:

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \xrightarrow{p} E\{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\}.$$

The expression $E\{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\}$ denotes a population magnitude that depends on the value of $\boldsymbol{\theta}$ and on the probability law of $\mathbf{w}_t$. Suppose that this function is continuous in $\boldsymbol{\theta}$ and that $\boldsymbol{\theta}_0$ is the only value of $\boldsymbol{\theta}$ that satisfies [14.1.9]. Then, under fairly general stationarity, continuity, and moment conditions, the value of $\hat{\boldsymbol{\theta}}_T$ that minimizes [14.1.11] offers a consistent estimate of $\boldsymbol{\theta}_0$; see Hansen (1982), Gallant and White (1988), and Andrews and Fair (1988) for details.

Optimal Weighting Matrix

Suppose that when evaluated at the true value $\boldsymbol{\theta}_0$, the process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ is strictly stationary with mean zero and νth autocovariance matrix given by

$$\boldsymbol{\Gamma}_\nu = E\{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-\nu})]'\}. \qquad [14.1.13]$$

Assuming that these autocovariances are absolutely summable, define

$$\mathbf{S} \equiv \sum_{\nu=-\infty}^{\infty} \boldsymbol{\Gamma}_\nu. \qquad [14.1.14]$$

Recall from the discussion in Section 10.5 that $\mathbf{S}$ is the asymptotic variance of the sample mean of $\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)$:

$$\mathbf{S} = \lim_{T\to\infty} T\cdot E\{[\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)][\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)]'\}.$$

The optimal value for the weighting matrix $\mathbf{W}_T$ in [14.1.11] turns out to be given by $\mathbf{S}^{-1}$, the inverse of the asymptotic variance matrix. That is, the minimum asymptotic variance for the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is obtained when $\hat{\boldsymbol{\theta}}_T$ is chosen to minimize

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\mathbf{S}^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]. \qquad [14.1.15]$$

To see the intuition behind this claim, consider a simple linear model in which we have r different observations $(y_1, y_2, \ldots, y_r)$ with a different population mean for each observation $(\mu_1, \mu_2, \ldots, \mu_r)$. For example, y_1 might denote the sample mean in a sample of T_1 observations on some variable, y_2 the sample mean from a second sample, and so on. In the absence of restrictions, the estimates would simply be $\hat{\mu}_i = y_i$ for $i = 1, 2, \ldots, r$. In the presence of linear restrictions across the μ's, the best estimates that are linear functions of the y's would be obtained by generalized least squares. Recall that the *GLS* estimate of $\boldsymbol{\mu}$ is the value that minimizes

$$(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \qquad [14.1.16]$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_r)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_r)'$, and $\boldsymbol{\Omega}$ is the variance-covariance matrix of $\mathbf{y} - \boldsymbol{\mu}$:

$$\boldsymbol{\Omega} = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'].$$

The optimal weighting matrix to use with the quadratic form in [14.1.16] is given by $\boldsymbol{\Omega}^{-1}$. Just as $\boldsymbol{\Omega}$ in [14.1.16] is the variance of $(\mathbf{y} - \boldsymbol{\mu})$, so $\mathbf{S}$ in [14.1.15] is the asymptotic variance of $\sqrt{T}\cdot\mathbf{g}(\cdot)$.

If the vector process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ were serially uncorrelated, then the matrix $\mathbf{S}$ could be consistently estimated by

$$\mathbf{S}_T^* = (1/T)\sum_{t=1}^{T}[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)]'. \qquad [14.1.17]$$

Calculating this magnitude requires knowledge of $\boldsymbol{\theta}_0$, though it often also turns out that

$$\hat{\mathbf{S}}_T \equiv (1/T)\sum_{t=1}^{T}[\mathbf{h}(\hat{\boldsymbol{\theta}}_T, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}_T, \mathbf{w}_t)]' \xrightarrow{p} \mathbf{S} \qquad [14.1.18]$$

for $\hat{\boldsymbol{\theta}}_T$ any consistent estimate of $\boldsymbol{\theta}_0$, assuming that $\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)$ is serially uncorrelated.

Note that this description of the optimal weighting matrix is somewhat circular—before we can estimate $\boldsymbol{\theta}$, we need an estimate of the matrix $\mathbf{S}$, and before we can estimate the matrix $\mathbf{S}$, we need an estimate of $\boldsymbol{\theta}$. The practical procedure used in *GMM* is as follows. An initial estimate $\hat{\boldsymbol{\theta}}_T^{(0)}$ is obtained by minimizing [14.1.11] with an arbitrary weighting matrix such as $\mathbf{W}_T = \mathbf{I}_r$. This estimate of $\boldsymbol{\theta}$ is then used in [14.1.18] to produce an initial estimate $\hat{\mathbf{S}}_T^{(0)}$. Expression [14.1.11] is then minimized with $\mathbf{W}_T = [\hat{\mathbf{S}}_T^{(0)}]^{-1}$ to arrive at a new *GMM* estimate $\hat{\boldsymbol{\theta}}_T^{(1)}$. This process can be iterated until $\hat{\boldsymbol{\theta}}_T^{(j)} \cong \hat{\boldsymbol{\theta}}_T^{(j+1)}$, though the estimate based on a single iteration $\hat{\boldsymbol{\theta}}_T^{(1)}$ has the same asymptotic distribution as that based on an arbitrarily large number of iterations. Iterating nevertheless offers the practical advantage that the resulting estimates are invariant with respect to the scale of the data and to the initial weighting matrix for $\mathbf{W}_T$.

On the other hand, if the vector process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ is serially correlated,

the Newey-West (1987) estimate of $\mathbf{S}$ could be used:

$$\hat{\mathbf{S}}_T = \hat{\Gamma}_{0,T} + \sum_{\nu=1}^{q} \{1 - [\nu/(q+1)]\}(\hat{\Gamma}_{\nu,T} + \hat{\Gamma}'_{\nu,T}), \qquad [14.1.19]$$

where

$$\hat{\Gamma}_{\nu,T} = (1/T) \sum_{t=\nu+1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_{t-\nu})]', \qquad [14.1.20]$$

with $\hat{\boldsymbol{\theta}}$ again an initial consistent estimate of $\boldsymbol{\theta}_0$. Alternatively, the estimators proposed by Gallant (1987), Andrews (1991), or Andrews and Monahan (1992) that were discussed in Section 10.5 could also be applied in this context.

Asymptotic Distribution of the GMM *Estimates*

Let $\hat{\boldsymbol{\theta}}_T$ be the value that minimizes

$$[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)], \qquad [14.1.21]$$

with $\hat{\mathbf{S}}_T$ regarded as fixed with respect to $\boldsymbol{\theta}$ and $\hat{\mathbf{S}}_T \xrightarrow{p} \mathbf{S}$. Assuming an interior optimum, this minimization is achieved by setting the derivative of [14.1.21] with respect to $\boldsymbol{\theta}$ to zero. Thus, the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$ is typically a solution to the following system of nonlinear equations:

$$\underbrace{\left\{\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}'}_{(a\times r)} \times \underbrace{\hat{\mathbf{S}}_T^{-1}}_{(r\times r)} \times \underbrace{[\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]}_{(r\times 1)} = \underbrace{\mathbf{0}}_{(a\times 1)}. \qquad [14.1.22]$$

Here $[\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)/\partial \boldsymbol{\theta}']|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}$ denotes the $(r \times a)$ matrix of derivatives of the function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$, where these derivatives are evaluated at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$.

Since $\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)$ is the sample mean of a process whose population mean is zero, $\mathbf{g}(\cdot)$ should satisfy the central limit theorem given conditions such as strict stationarity of $\mathbf{w}_t$, continuity of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$, and restrictions on higher moments. Thus, in many instances it should be the case that

$$\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S}).$$

Not much more than this is needed to conclude that the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is asymptotically Gaussian and to calculate its asymptotic variance. The following proposition, adapted from Hansen (1982), is proved in Appendix 14.A at the end of this chapter.

Proposition 14.1: *Let $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ be differentiable in $\boldsymbol{\theta}$ for all $\mathcal{Y}_T$, and let $\hat{\boldsymbol{\theta}}_T$ be the GMM estimator satisfying* [14.1.22] *with $r \geq a$. Let $\{\hat{\mathbf{S}}_T\}_{T=1}^{\infty}$ be a sequence of positive definite $(r \times r)$ matrices such that $\hat{\mathbf{S}}_T \xrightarrow{p} \mathbf{S}$, with $\mathbf{S}$ positive definite. Suppose, further, that the following hold:*

(a) $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$;

(b) $\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S})$; *and*

(c) for any sequence $\{\boldsymbol{\theta}_T^*\}_{T=1}^{\infty}$ satisfying $\boldsymbol{\theta}_T^* \xrightarrow{p} \boldsymbol{\theta}_0$, it is the case that

$$\operatorname{plim}\left\{\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T^*}\right\} = \operatorname{plim}\left\{\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\} \equiv \underset{(r\times a)}{\mathbf{D}'}, \qquad [14.1.23]$$

with the columns of $\mathbf{D}'$ linearly independent.

Then

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \qquad [14.1.24]$$

where

$$\mathbf{V} = \{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1}.$$

Proposition 14.1 implies that we can treat $\hat{\boldsymbol{\theta}}_T$ approximately as

$$\hat{\boldsymbol{\theta}}_T \approx N(\boldsymbol{\theta}_0, \hat{\mathbf{V}}_T/T), \qquad [14.1.25]$$

where

$$\hat{\mathbf{V}}_T = \{\hat{\mathbf{D}}_T\hat{\mathbf{S}}_T^{-1}\hat{\mathbf{D}}_T'\}^{-1}.$$

The estimate $\hat{\mathbf{S}}_T$ can be constructed as in [14.1.18] or [14.1.19], while

$$\underset{(r\times a)}{\hat{\mathbf{D}}_T'} = \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}.$$

Testing the Overidentifying Restrictions

When the number of orthogonality conditions exceeds the number of parameters to be estimated ($r > a$), the model is overidentified, in that more orthogonality conditions were used than are needed to estimate $\boldsymbol{\theta}$. In this case, Hansen (1982) suggested a test of whether all of the sample moments represented by $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are as close to zero as would be expected if the corresponding population moments $E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}$ were truly zero.

From Proposition 8.1 and condition (b) in Proposition 14.1, notice that if the population orthogonality conditions in [14.1.9] were all true, then

$$[\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)]'\mathbf{S}^{-1}[\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)] \xrightarrow{L} \chi^2(r). \qquad [14.1.26]$$

In [14.1.26], the sample moment function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ is evaluated at the true value of $\boldsymbol{\theta}_0$. One's first guess might be that condition [14.1.26] also holds when [14.1.26] is evaluated at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$. However, this is not the case. The reason is that [14.1.22] implies that a different linear combinations of the $(r \times 1)$ vector $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are identically zero, these being the a linear combinations obtained when $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ is premultiplied by the $(a \times r)$ matrix

$$\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1}.$$

For example, when $a = r$, *all* linear combinations of $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are identically zero, and if $\boldsymbol{\theta}_0$ were replaced by $\hat{\boldsymbol{\theta}}_T$, the magnitude in [14.1.26] would simply equal zero in all samples.

Since the vector $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ contains $(r - a)$ nondegenerate random variables, it turns out that a correct test of the overidentifying restrictions for the case when $r > a$ can be based on the fact that

$$[\sqrt{T}\cdot\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\sqrt{T}\cdot\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)] \xrightarrow{L} \chi^2(r - a). \qquad [14.1.27]$$

Moreover, this test statistic is trivial to calculate, for it is simply the sample size T times the value attained for the objective function [14.1.21] at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$.

Unfortunately, Hansen's χ^2 test based on [14.1.27] can easily fail to detect a misspecified model (Newey, 1985). It is therefore often advisable to supplement this test with others described in Section 14.3.

14.2. *Examples*

This section shows how properties of a variety of different estimators can be obtained as special cases of Hansen's results for generalized method of moments

estimation. To facilitate this discussion, we first summarize the results of the preceding section.

Summary of GMM

The statistical model is assumed to imply a set of r orthogonality conditions of the form

$$\underset{(r\times 1)}{E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}} = \underset{(r\times 1)}{\mathbf{0}}, \qquad [14.2.1]$$

where $\mathbf{w}_t$ is a strictly stationary vector of variables observed at date t, $\boldsymbol{\theta}_0$ is the true value of an unknown $(a \times 1)$ vector of parameters, and $\mathbf{h}(\cdot)$ is a differentiable r-dimensional vector-valued function with $r \geq a$. The *GMM* estimate $\hat{\boldsymbol{\theta}}_T$ is the value of $\boldsymbol{\theta}$ that minimizes

$$\underset{(1\times r)}{[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'}\underset{(r\times r)}{\hat{\mathbf{S}}_T^{-1}}\underset{(r\times 1)}{[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]}, \qquad [14.2.2]$$

where

$$\underset{(r\times 1)}{\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)} \equiv (1/T) \sum_{t=1}^{T} \underset{(r\times 1)}{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)} \qquad [14.2.3]$$

and $\hat{\mathbf{S}}_T$ is an estimate of[1]

$$\underset{(r\times r)}{\mathbf{S}} = \lim_{T\to\infty} (1/T) \sum_{t=1}^{T} \sum_{\nu=-\infty}^{\infty} E\{\underset{(r\times 1)}{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)]}\underset{(1\times r)}{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-\nu})]'}\}. \qquad [14.2.4]$$

The *GMM* estimate can be treated as if

$$\underset{(a\times 1)}{\hat{\boldsymbol{\theta}}_T} \approx \underset{(a\times 1)}{N(\boldsymbol{\theta}_0,} \underset{(a\times a)}{\hat{\mathbf{V}}_T/T)}, \qquad [14.2.5]$$

where

$$\underset{(a\times a)}{\hat{\mathbf{V}}_T} = \{\underset{(a\times r)}{\hat{\mathbf{D}}_T} \cdot \underset{(r\times r)}{\hat{\mathbf{S}}_T^{-1}} \cdot \underset{(r\times a)}{\hat{\mathbf{D}}_T'}\}^{-1} \qquad [14.2.6]$$

and

$$\underset{(r\times a)}{\hat{\mathbf{D}}_T'} = \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}. \qquad [14.2.7]$$

We now explore how these results would be applied in various special cases.

Ordinary Least Squares

Consider the standard linear regression model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t, \qquad [14.2.8]$$

for $\mathbf{x}_t$ a $(k \times 1)$ vector of explanatory variables. The critical assumption needed to justify *OLS* regression is that the regression residual u_t is uncorrelated with the explanatory variables:

$$E(\mathbf{x}_t u_t) = \mathbf{0}. \qquad [14.2.9]$$

[1]Under strict stationarity, the magnitude

$$E\{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-\nu})]'\} = \boldsymbol{\Gamma}_\nu$$

does not depend on t. The expression in the text is more general than necessary under the stated assumptions. This expression is appropriate for a characterization of *GMM* that does not assume strict stationarity. The expression in the text is also helpful in suggesting estimates of S that can be used in various special cases described later in this section.

In other words, the true value $\boldsymbol{\beta}_0$ is assumed to satisfy the condition

$$E[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_0)] = \mathbf{0}. \qquad [14.2.10]$$

Expression [14.2.10] describes k orthogonality conditions of the form of [14.2.1], in which $\mathbf{w}_t = (y_t, \mathbf{x}_t')'$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}). \qquad [14.2.11]$$

The number of orthogonality conditions is the same as the number of unknown parameters in $\boldsymbol{\beta}$, so that $r = a = k$. Hence, the standard regression model could be viewed as a just-identified *GMM* specification. Since it is just identified, the *GMM* estimate of $\boldsymbol{\beta}$ is the value that sets the sample average value for [14.2.11] equal to zero:

$$\mathbf{0} = \mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = (1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T). \qquad [14.2.12]$$

Rearranging [14.2.12] results in

$$\sum_{t=1}^{T} \mathbf{x}_t y_t = \left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}\hat{\boldsymbol{\beta}}_T$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \mathbf{x}_t y_t\right\}, \qquad [14.2.13]$$

which is the usual *OLS* estimator. Hence, *OLS* is a special case of *GMM*.

Note that in deriving the *GMM* estimator in [14.2.13] we assumed that the residual u_t was uncorrelated with the explanatory variables, but we did not make any other assumptions about heteroskedasticity or serial correlation of the residuals. In the presence of heteroskedasticity or serial correlation, *OLS* is not as efficient as *GLS*. Because *GMM* uses the *OLS* estimate even in the presence of heteroskedasticity or serial correlation, *GMM* in general is not efficient. However, recall from Section 8.2 that one can still use *OLS* in the presence of heteroskedasticity or serial correlation. As long as condition [14.2.9] is satisfied, *OLS* yields a consistent estimate of $\boldsymbol{\beta}$, though the formulas for standard errors have to be adjusted to take account of the heteroskedasticity or autocorrelation.

The *GMM* expression for the variance of $\hat{\boldsymbol{\beta}}_T$ is given by [14.2.6]. Differentiating [14.2.11], we see that

$$\begin{aligned}\hat{\mathbf{D}}_T' &= \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} \\ &= (1/T)\sum_{t=1}^{T} \left.\frac{\partial \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_T} \qquad [14.2.14] \\ &= -(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'.\end{aligned}$$

Substituting [14.2.11] into [14.2.4] results in

$$\mathbf{S} = \lim_{T\to\infty} (1/T)\sum_{t=1}^{T}\sum_{v=-\infty}^{\infty} E\{u_t u_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'\}. \qquad [14.2.15]$$

Suppose that u_t is regarded as conditionally homoskedastic and serially uncorrelated:

$$E\{u_t u_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'\} = \begin{cases}\sigma^2 E(\mathbf{x}_t\mathbf{x}_t') & \text{for } v = 0 \\ \mathbf{0} & \text{for } v \neq 0.\end{cases}$$

In this case the matrix in [14.2.15] should be consistently estimated by

$$\hat{\mathbf{S}}_T = \hat{\sigma}_T^2 (1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t', \qquad [14.2.16]$$

where

$$\hat{\sigma}_T^2 = (1/T) \sum_{t=1}^{T} \hat{u}_t^2$$

for $\hat{u}_t \equiv y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ the *OLS* residual. Substituting [14.2.14] and [14.2.16] into [14.2.6] produces a variance-covariance matrix for the *OLS* estimate $\hat{\boldsymbol{\beta}}_T$ of

$$\begin{aligned}(1/T)\hat{\mathbf{V}}_T &= (1/T)\left\{(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \left[\hat{\sigma}_T^2 (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1} \\ &= \hat{\sigma}_T^2 \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}.\end{aligned}$$

Apart from the estimate of σ^2, this is the usual expression for the variance of the *OLS* estimator under these conditions.

On the other hand, suppose that u_t is conditionally heteroskedastic and serially correlated. In this case, the estimate of $\mathbf{S}$ proposed in [14.1.19] would be

$$\hat{\mathbf{S}}_T = \hat{\boldsymbol{\Gamma}}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q+1)]\}(\hat{\boldsymbol{\Gamma}}_{v,T} + \hat{\boldsymbol{\Gamma}}_{v,T}'),$$

where

$$\hat{\boldsymbol{\Gamma}}_{v,T} = (1/T) \sum_{t=v+1}^{T} \hat{u}_t\hat{u}_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'.$$

Under these assumptions, the *GMM* approximation for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_T$ would be

$$\begin{aligned}E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] &\cong (1/T)\left\{(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \hat{\mathbf{S}}_T^{-1} (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1} \\ &= T\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \hat{\mathbf{S}}_T \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1},\end{aligned}$$

which is the expression derived earlier in equation [10.5.21]. White's (1980) heteroskedasticity-consistent standard errors in [8.2.35] are obtained as a special case when $q = 0$.

Instrumental Variable Estimation

Consider again a linear model

$$y_t = \mathbf{z}_t'\boldsymbol{\beta} + u_t, \qquad [14.2.17]$$

where $\mathbf{z}_t$ is a $(k \times 1)$ vector of explanatory variables. Suppose now that some of the explanatory variables are endogenous, so that $E(\mathbf{z}_t u_t) \neq \mathbf{0}$. Let $\mathbf{x}_t$ be an $(r \times 1)$ vector of predetermined explanatory variables that are correlated with $\mathbf{z}_t$ but uncorrelated with u_t:

$$E(\mathbf{x}_t u_t) = \mathbf{0}.$$

The r orthogonality conditions are now

$$E[\mathbf{x}_t(y_t - \mathbf{z}_t'\boldsymbol{\beta}_0)] = \mathbf{0}. \qquad [14.2.18]$$

This again will be recognized as a special case of the *GMM* framework in which $\mathbf{w}_t = (y_t, \mathbf{z}_t', \mathbf{x}_t')'$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $a = k$, and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \mathbf{x}_t(y_t - \mathbf{z}_t'\boldsymbol{\beta}). \qquad [14.2.19]$$

Suppose that the number of parameters to be estimated equals the number of orthogonality conditions ($a = k = r$). Then the model is just identified, and the *GMM* estimator satisfies

$$\mathbf{0} = \mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = (1/T) \sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T) \qquad [14.2.20]$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right\}^{-1} \left\{\sum_{t=1}^{T} \mathbf{x}_t y_t\right\},$$

which is the usual instrumental variable estimator for this model. To calculate the standard errors implied by Hansen's (1982) general results, we differentiate [14.2.19] to find

$$\begin{aligned} \hat{\mathbf{D}}_T' &= \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} \\ &= (1/T) \sum_{t=1}^{T} \left.\frac{\partial \mathbf{x}_t(y_t - \mathbf{z}_t'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_T} \qquad [14.2.21] \\ &= -(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'. \end{aligned}$$

The requirement in Proposition 14.1 that the plim of this matrix have linearly independent columns is the same condition that was needed to establish consistency of the *IV* estimator in Chapter 9, namely, the condition that the rows of $E(\mathbf{z}_t\mathbf{x}_t')$ be linearly independent. The *GMM* variance for $\hat{\boldsymbol{\beta}}_T$ is seen from [14.2.6] to be

$$(1/T)\hat{\mathbf{V}}_T = (1/T)\left\{\left[(1/T) \sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]\hat{\mathbf{S}}_T^{-1}\left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1}, \qquad [14.2.22]$$

where $\hat{\mathbf{S}}_T$ is an estimate of

$$\mathbf{S} = \lim_{T\to\infty} (1/T) \sum_{t=1}^{T} \sum_{\nu=-\infty}^{\infty} E\{u_t u_{t-\nu}\mathbf{x}_t\mathbf{x}_{t-\nu}'\}. \qquad [14.2.23]$$

If the regression residuals $\{u_t\}$ are serially uncorrelated and homoskedastic with variance σ^2, the natural estimate of $\mathbf{S}$ is

$$\hat{\mathbf{S}}_T = \hat{\sigma}_T^2 \cdot (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \qquad [14.2.24]$$

for $\hat{\sigma}_T^2 = (1/T)\sum_{t=1}^{T}(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)^2$. Substituting this estimate into [14.2.22] yields

$$\begin{aligned} E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] &\cong \hat{\sigma}_T^2\left\{\left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1} \\ &= \hat{\sigma}_T^2\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]^{-1}, \end{aligned}$$

the same result derived earlier in [9.2.30]. On the other hand, a heteroskedasticity- and autocorrelation-consistent variance-covariance matrix for *IV* estimation is given by

$$E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] \cong T\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]^{-1} \hat{\mathbf{S}}_T\left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]^{-1}, \qquad [14.2.25]$$

where

$$\hat{\mathbf{S}}_T = \hat{\mathbf{\Gamma}}_{0,T} + \sum_{\nu=1}^{q} \{1 - [\nu/(q+1)]\}(\hat{\mathbf{\Gamma}}_{\nu,T} + \hat{\mathbf{\Gamma}}'_{\nu,T}), \qquad [14.2.26]$$

$$\hat{\mathbf{\Gamma}}_{\nu,T} = (1/T) \sum_{t=\nu+1}^{T} \hat{u}_t \hat{u}_{t-\nu} \mathbf{x}_t \mathbf{x}'_{t-\nu}$$

$$\hat{u}_t = y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T.$$

Two-Stage Least Squares

Consider again the linear model of [14.2.17] and [14.2.18], but suppose now that the number of valid instruments r exceeds the number of explanatory variables k. For this overidentified model, *GMM* will no longer set all the sample orthogonality conditions to zero as in [14.2.20], but instead will be the solution to [14.1.22],

$$\begin{aligned} \mathbf{0} &= \left\{ \frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'} \bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)] \\ &= \left\{ -(1/T) \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \hat{\mathbf{S}}_T^{-1} \left\{ (1/T) \sum_{t=1}^{T} \mathbf{x}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) \right\}, \end{aligned} \qquad [14.2.27]$$

with the last line following from [14.2.21] and [14.2.20]. Again, if u_t is serially uncorrelated and homoskedastic with variance σ^2, a natural estimate of $\mathbf{S}$ is given by [14.2.24]. Using this estimate, [14.2.27] becomes

$$(1/\hat{\sigma}_T^2) \times \left\{ \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \left\{ \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right\}^{-1} \left\{ \sum_{t=1}^{T} \mathbf{x}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) \right\} = \mathbf{0}. \qquad [14.2.28]$$

As in expression [9.2.5], define

$$\hat{\boldsymbol{\delta}}' \equiv \left\{ \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \left\{ \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right\}^{-1}.$$

Thus, $\hat{\boldsymbol{\delta}}'$ is a $(k \times r)$ matrix whose ith row represents the coefficients from an *OLS* regression of z_{it} on $\mathbf{x}_t$. Let

$$\hat{\mathbf{z}}_t \equiv \hat{\boldsymbol{\delta}}' \mathbf{x}_t$$

be the $(k \times 1)$ vector of fitted values from these regressions of $\mathbf{z}_t$ on $\mathbf{x}_t$. Then [14.2.28] implies that

$$\sum_{t=1}^{T} \hat{\mathbf{z}}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) = \mathbf{0}$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{ \sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}'_t \right\}^{-1} \left\{ \sum_{t=1}^{T} \hat{\mathbf{z}}_t y_t \right\}.$$

Thus, the *GMM* estimator for this case is simply the two-stage least squares estimator as written in [9.2.8]. The variance given in [14.2.6] would be

$$\begin{aligned} (1/T)\hat{\mathbf{V}}_T &= (1/T) \left\{ \left[(1/T) \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right] \hat{\mathbf{S}}_T^{-1} \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}'_t \right] \right\}^{-1} \\ &= \hat{\sigma}_T^2 \left\{ \left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right] \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}'_t \right] \right\}^{-1}, \end{aligned}$$

as earlier derived in expression [9.2.25]. A test of the overidentifying assumptions embodied in the model in [14.2.17] and [14.2.18] is given by

$$T[\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]$$

$$= T\left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\}'\left\{\hat{\sigma}_T^2 \cdot (1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1}$$

$$\times \left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\}$$

$$= \hat{\sigma}_T^{-2}\left\{\sum_{t=1}^{T} \hat{u}_t\mathbf{x}_t'\right\}\left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \mathbf{x}_t\hat{u}_t\right\}.$$

This magnitude will have an asymptotic χ^2 distribution with $(r - k)$ degrees of freedom if the model is correctly specified.

Alternatively, to allow for heteroskedasticity and autocorrelation for the residuals u_t, the estimate $\hat{\mathbf{S}}_T$ in [14.2.24] would be replaced by [14.2.26]. Recall the first-order condition [14.2.27]:

$$\left\{(1/T)\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right\}\hat{\mathbf{S}}_T^{-1}\left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\} = \mathbf{0}. \qquad [14.2.29]$$

If we now define

$$\tilde{\mathbf{z}}_t \equiv \tilde{\boldsymbol{\delta}}'\mathbf{x}_t$$

$$\tilde{\boldsymbol{\delta}}' \equiv \left\{(1/T)\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right\}\hat{\mathbf{S}}_T^{-1},$$

then [14.2.29] implies that the *GMM* estimator for this case is given by

$$\hat{\boldsymbol{\beta}}_T = \left\{\sum_{t=1}^{T} \tilde{\mathbf{z}}_t\mathbf{z}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \tilde{\mathbf{z}}_t y_t\right\}.$$

This characterization of $\hat{\boldsymbol{\beta}}_T$ is circular—in order to calculate $\hat{\boldsymbol{\beta}}_T$, we need to know $\tilde{\mathbf{z}}_t$ and thus $\hat{\mathbf{S}}_T$, whereas to construct $\hat{\mathbf{S}}_T$ from [14.2.26] we first need to know $\hat{\boldsymbol{\beta}}_T$. The solution is first to estimate $\boldsymbol{\beta}$ using a suboptimal weighting matrix such as $\hat{\mathbf{S}}_T = (1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$, and then to use this estimate of $\mathbf{S}$ to reestimate $\boldsymbol{\beta}$. The asymptotic variance of the *GMM* estimator is given by

$$E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] \cong T\left\{\left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]\hat{\mathbf{S}}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1}.$$

Nonlinear Systems of Simultaneous Equations

Hansen's (1982) *GMM* also provides a convenient framework for estimating the nonlinear systems of simultaneous equations analyzed by Amemiya (1974), Jorgenson and Laffont (1974), and Gallant (1977). Suppose that the goal is to estimate a system of n nonlinear equations of the form

$$\mathbf{y}_t = \mathbf{f}(\boldsymbol{\theta}, \mathbf{z}_t) + \mathbf{u}_t$$

for $\mathbf{z}_t$ a $(k \times 1)$ vector of explanatory variables and $\boldsymbol{\theta}$ an $(a \times 1)$ vector of unknown parameters. Let $\mathbf{x}_{it}$ denote a vector of instruments that are uncorrelated with the ith element of $\mathbf{u}_t$. The r orthogonality conditions for this model are

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} [y_{1t} - f_1(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{1t} \\ [y_{2t} - f_2(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{2t} \\ \vdots \\ [y_{nt} - f_n(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{nt} \end{bmatrix},$$

where $f_i(\boldsymbol{\theta}, \mathbf{z}_t)$ denotes the ith element of $\mathbf{f}(\boldsymbol{\theta}, \mathbf{z}_t)$ and $\mathbf{w}_t \equiv (\mathbf{y}_t', \mathbf{z}_t', \mathbf{x}_t')'$. The *GMM* estimate of $\boldsymbol{\theta}$ is the value that minimizes

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = \left[(1/T)\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\right]' \hat{\mathbf{S}}_T^{-1} \left[(1/T)\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\right], \qquad [14.2.30]$$

where an estimate of $\mathbf{S}$ that could be used with heteroskedasticity and serial correlation of u_t is given by

$$\hat{\mathbf{S}}_T = \hat{\boldsymbol{\Gamma}}_{0,T} + \sum_{\nu=1}^{q} \{1-[\nu/(q+1)]\}(\hat{\boldsymbol{\Gamma}}_{\nu,T} + \hat{\boldsymbol{\Gamma}}_{\nu,T}')$$

$$\hat{\boldsymbol{\Gamma}}_{\nu,T} = (1/T)\sum_{t=\nu+1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_{t-\nu})]'.$$

Minimization of [14.2.30] can be achieved numerically. Again, in order to evaluate [14.2.30], we first need an initial estimate of $\mathbf{S}$. One approach is to first minimize [14.2.30] with $\mathbf{S}_T = \mathbf{I}_r$, use the resulting estimate $\hat{\boldsymbol{\theta}}$ to construct a better estimate of $\mathbf{S}_T$, and recalculate $\hat{\boldsymbol{\theta}}$; the procedure can be iterated further, if desired. Identification requires an order condition ($r \geq a$) and the rank condition that the columns of the plim of $\hat{\mathbf{D}}_T'$ be linearly independent, where

$$\hat{\mathbf{D}}_T' = (1/T)\sum_{t=1}^{T} \left.\frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}.$$

Standard errors for $\hat{\boldsymbol{\theta}}_T$ are then readily calculated from [14.2.5] and [14.2.6].

Estimation of Dynamic Rational Expectation Models

People's behavior is often influenced by their expectations about the future. Unfortunately, we typically do not have direct observations on these expectations. However, it is still possible to estimate and test behavioral models if people's expectations are formed rationally in the sense that the errors they make in forecasting are uncorrelated with information they had available at the time of the forecast. As long as the econometrician observes a subset of the information people have actually used, the rational expectations hypothesis suggests orthogonality conditions that can be used in the *GMM* framework.

For illustration, we consider the study of portfolio decisions by Hansen and Singleton (1982). Let c_t denote the overall level of spending on consumption goods by a particular stockholder during period t. The satisfaction or utility that the stockholder receives from this spending is represented by a function $u(c_t)$, where it is assumed that

$$\frac{\partial u(c_t)}{\partial c_t} > 0 \qquad \frac{\partial^2 u(c_t)}{\partial c_t^2} < 0.$$

The stockholder is presumed to want to maximize

$$\sum_{\tau=0}^{\infty} \beta^\tau E\{u(c_{t+\tau})|\mathbf{x}_t^*\}, \qquad [14.2.31]$$

where $\mathbf{x}_t^*$ is a vector representing all the information available to the stockholder at date t and β is a parameter satisfying $0 < \beta < 1$. Smaller values of β mean that the stockholder places a smaller weight on future events. At date t, the stockholder contemplates purchasing any of m different assets, where a dollar invested in asset i at date t will yield a gross return of $(1 + r_{i,t+1})$ at date $t + 1$; in general this rate of return is not known for certain at date t. Assuming that the stockholder takes a position in each of these m assets, the stockholder's optimal portfolio will satisfy

$$u'(c_t) = \beta E\{(1 + r_{i,t+1})u'(c_{t+1})|\mathbf{x}_t^*\} \qquad \text{for } i = 1, 2, \ldots, m, \qquad [14.2.32]$$

where $u'(c_t) \equiv \partial u/\partial c_t$. To see the intuition behind this claim, suppose that condition [14.2.32] failed to hold. Say, for example, that the left side were smaller than the right. Suppose the stockholder were to save one more dollar at date t and invest the dollar in asset i, using the returns to boost period $t + 1$ consumption. Following this strategy would cause consumption at date t to fall by one dollar (reducing [14.2.31] by an amount given by the left side of [14.2.32]), while consumption at date $t + 1$ would rise by $(1 + r_{i,t+1})$ dollars (increasing [14.2.31] by an amount given by the right side of [14.2.32]). If the left side of [14.2.32] were less than the right side of [14.2.32], then the stockholder's objective [14.2.31] would be improved under this change. Only when [14.2.32] is satisfied is the stockholder as well off as possible.[2]

Suppose that the utility function is parameterized as

$$u(c_t) = \begin{cases} \dfrac{c_t^{1-\gamma}}{1-\gamma} & \text{for } \gamma > 0 \text{ and } \gamma \neq 1 \\ \log c_t & \text{for } \gamma = 1. \end{cases}$$

The parameter γ is known as the *coefficient of relative risk aversion*, which for this class of utility functions is a constant. For this function, [14.2.32] becomes

$$c_t^{-\gamma} = \beta E\{(1 + r_{i,t+1})c_{t+1}^{-\gamma} | \mathbf{x}_t^*\}. \qquad [14.2.33]$$

Dividing both sides of [14.2.33] by $c_t^{-\gamma}$ results in

$$1 = \beta E\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma} | \mathbf{x}_t^*\}, \qquad [14.2.34]$$

where c_t could be moved inside the conditional expectation operator, since it represents a decision based solely on the information contained in $\mathbf{x}_t^*$. Expression [14.2.34] requires that the random variable described by

$$1 - \beta\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma}\} \qquad [14.2.35]$$

be uncorrelated with any variable contained in the information set $\mathbf{x}_t^*$ for any asset i that the stockholder holds. It should therefore be the case that

$$E\{[1 - \beta\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t\} = \mathbf{0}, \qquad [14.2.36]$$

where $\mathbf{x}_t$ is any subset of the stockholder's information set $\mathbf{x}_t^*$ that the econometrician is also able to observe.

Let $\boldsymbol{\theta} \equiv (\beta, \gamma)'$ denote the unknown parameters that are to be estimated, and let $\mathbf{w}_t \equiv (r_{1,t+1}, r_{2,t+1}, \ldots, r_{m,t+1}, c_{t+1}/c_t, \mathbf{x}_t')'$ denote the vector of variables that are observed by the econometrician for date t. Stacking the equations in [14.2.36] for $i = 1, 2, \ldots, m$ produces a set of r orthogonality conditions that can be used to estimate $\boldsymbol{\theta}$:

$$\underset{(r\times 1)}{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)} = \begin{bmatrix} [1 - \beta\{(1 + r_{1,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \\ [1 - \beta\{(1 + r_{2,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \\ \vdots \\ [1 - \beta\{(1 + r_{m,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \end{bmatrix}. \qquad [14.2.37]$$

The sample average value of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ is

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \equiv (1/T) \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t),$$

and the *GMM* objective function is

$$Q(\boldsymbol{\theta}) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]. \qquad [14.2.38]$$

This expression can then be minimized numerically with respect to $\boldsymbol{\theta}$.

According to the theory, the magnitude in [14.2.35] should be uncorrelated with any information the stockholder has available at time t, which would include

[2]For further details, see Sargent (1987).

lagged values of [14.2.35]. Hence, the vector in [14.2.37] should be uncorrelated with its own lagged values, suggesting that S can be consistently estimated by

$$\hat{\mathbf{S}}_T = (1/T) \sum_{t=1}^{T} \{[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)]'\},$$

where $\hat{\boldsymbol{\theta}}$ is an initial consistent estimate. This initial estimate $\hat{\boldsymbol{\theta}}$ could be obtained by minimizing [14.2.38] with $\hat{\mathbf{S}}_T = \mathbf{I}_r$.

Hansen and Singleton (1982) estimated such a model using real consumption expenditures for the aggregate United States divided by the U.S. population as their measure of c_t. For r_{1t}, they used the inflation-adjusted return that an investor would earn if one dollar was invested in every stock listed on the New York Stock Exchange, while r_{2t} was a value-weighted inflation-adjusted return corresponding to the return an investor would earn if the investor owned the entire stock of each company listed on the exchange. Hansen and Singleton's instruments consisted of a constant term, lagged consumption growth rates, and lagged rates of return:

$$\mathbf{x}_t = (1, c_t/c_{t-1}, c_{t-1}/c_{t-2}, \ldots, c_{t-\ell+1}/c_{t-\ell}, r_{1t}, r_{1,t-1}, \ldots, r_{1,t-\ell+1}, r_{2,t}, r_{2,t-1}, \ldots, r_{2,t-\ell+1})'.$$

When ℓ lags are used, there are $3\ell + 1$ elements in $\mathbf{x}_t$, and thus $r = 2(3\ell + 1)$ separate orthogonality conditions are represented by [14.2.37]. Since $a = 2$ parameters are estimated, the χ^2 statistic in [14.1.27] has 6ℓ degrees of freedom.

14.3. *Extensions*

GMM *with Nonstationary Data*

The maintained assumption throughout this chapter has been that the $(h \times 1)$ vector of observed variables $\mathbf{w}_t$ is strictly stationary. Even if the raw data appear to be trending over time, sometimes the model can be transformed or reparameterized so that stationarity of the transformed system is a reasonable assumption. For example, the consumption series $\{c_t\}$ used in Hansen and Singleton's study (1982) is increasing over time. However, it was possible to write the equation to be estimated [14.2.36] in such a form that only the consumption growth rate (c_{t+1}/c_t) appeared, for which the stationarity assumption is much more plausible. Alternatively, suppose that some of the elements of the observed vector $\mathbf{w}_t$ are presumed to grow deterministically over time according to

$$\mathbf{w}_t = \boldsymbol{\alpha} + \boldsymbol{\delta} \cdot t + \mathbf{w}_t^*, \qquad [14.3.1]$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are $(h \times 1)$ vectors of constants and $\mathbf{w}_t^*$ is strictly stationary with mean zero. Suppose that the orthogonality conditions can be expressed in terms of $\mathbf{w}_t^*$ as

$$E\{\mathbf{f}(\boldsymbol{\theta}_0, \mathbf{w}_t^*)\} = \mathbf{0}.$$

Then Ogaki (1993) recommended jointly estimating $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ using

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} \mathbf{w}_t - \boldsymbol{\alpha} - \boldsymbol{\delta}t \\ \mathbf{f}(\boldsymbol{\theta}, \mathbf{w}_t - \boldsymbol{\alpha} - \boldsymbol{\delta}t) \end{bmatrix}$$

to construct the moment condition in [14.2.3].

Testing for Structural Stability

Suppose we want to test the hypothesis that the $(a \times 1)$ parameter vector $\boldsymbol{\theta}$ that characterizes the first T_0 observations in the sample is different from the value

that characterizes the last $T - T_0$ observations, where T_0 is a known change point. One approach is to obtain an estimate $\hat{\boldsymbol{\theta}}_{1.T_0}$ based solely on the first T_0 observations, minimizing

$$Q(\boldsymbol{\theta}_1; \mathbf{w}_{T_0}, \mathbf{w}_{T_0-1}, \ldots, \mathbf{w}_1)$$
$$= \left[(1/T_0)\sum_{t=1}^{T_0} \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)\right]' \hat{\mathbf{S}}_{1.T_0}^{-1}\left[(1/T_0)\sum_{t=1}^{T_0} \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)\right], \qquad [14.3.2]$$

where, for example, if $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}$ is serially uncorrelated,

$$\hat{\mathbf{S}}_{1.T_0} = (1/T_0)\sum_{t=1}^{T_0} [\mathbf{h}(\hat{\boldsymbol{\theta}}_{1.T_0}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}_{1.T_0}, \mathbf{w}_t)]'.$$

Proposition 14.1 implies that

$$\sqrt{T_0}(\hat{\boldsymbol{\theta}}_{1.T_0} - \boldsymbol{\theta}_1) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_1) \qquad [14.3.3]$$

as $T_0 \to \infty$, where $\mathbf{V}_1$ can be estimated from

$$\hat{\mathbf{V}}_{1.T_0} = \{\hat{\mathbf{D}}_{1.T_0}\hat{\mathbf{S}}_{1.T_0}^{-1}\hat{\mathbf{D}}'_{1.T_0}\}^{-1}$$

for

$$\hat{\mathbf{D}}'_{1.T_0} \equiv (1/T_0)\sum_{t=1}^{T_0} \frac{\partial \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)}{\partial \boldsymbol{\theta}'_1}\bigg|_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_{1.T_0}}.$$

Similarly, a separate estimate $\hat{\boldsymbol{\theta}}_{2.T-T_0}$ can be based on the last $T - T_0$ observations, with analogous measures $\hat{\mathbf{S}}_{2.T-T_0}$, $\hat{\mathbf{V}}_{2.T-T_0}$, $\hat{\mathbf{D}}_{2.T-T_0}$, and

$$\sqrt{T - T_0}(\hat{\boldsymbol{\theta}}_{2.T-T_0} - \boldsymbol{\theta}_2) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_2) \qquad [14.3.4]$$

as $T - T_0 \to \infty$. Let $\pi \equiv T_0/T$ denote the fraction of observations contained in the first subsample. Then [14.3.3] and [14.3.4] state that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{1.T_0} - \boldsymbol{\theta}_1) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_1/\pi)$$
$$\sqrt{T}(\hat{\boldsymbol{\theta}}_{2.T-T_0} - \boldsymbol{\theta}_2) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_2/(1 - \pi))$$

as $T \to \infty$. Andrews and Fair (1988) suggested using a Wald test of the null hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, exploiting the fact that under the stationarity conditions needed to justify Proposition 14.1, $\hat{\boldsymbol{\theta}}_1$ is asymptotically independent of $\hat{\boldsymbol{\theta}}_2$:

$$\lambda_T = T(\hat{\boldsymbol{\theta}}_{1.T_0} - \hat{\boldsymbol{\theta}}_{2.T-T_0})'$$
$$\times \{\pi^{-1}\cdot\hat{\mathbf{V}}_{1.T_0} + (1 - \pi)^{-1}\cdot\hat{\mathbf{V}}_{2.T-T_0}\}^{-1}(\hat{\boldsymbol{\theta}}_{1.T_0} - \hat{\boldsymbol{\theta}}_{2.T-T_0}).$$

Then $\lambda_T \xrightarrow{L} \chi^2(a)$ under the null hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

One can further test for structural change at a variety of different possible dates, repeating the foregoing test for all T_0 between, say, $0.15T$ and $0.85T$ and choosing the largest value for the resulting test statistic λ_T. Andrews (1993) described the asymptotic distribution of such a test.

Another simple test associates separate moment conditions with the observations before and after T_0 and uses the χ^2 test suggested in [14.1.27] to test the validity of the separate sets of conditions. Specifically, let

$$d_{1t} = \begin{cases} 1 & \text{for } t \le T_0 \\ 0 & \text{for } t > T_0. \end{cases}$$

If $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ is an $(r \times 1)$ vector whose population mean is zero at $\boldsymbol{\theta}_0$, define

$$\underset{(2r\times 1)}{\mathbf{h}^*(\boldsymbol{\theta}, \mathbf{w}_t, d_{1t})} \equiv \begin{bmatrix} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\cdot d_{1t} \\ \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\cdot(1 - d_{1t}) \end{bmatrix}.$$

The a elements of $\boldsymbol{\theta}$ can then be estimated by using the $2r$ orthogonality conditions given by $E\{\mathbf{h}^*(\boldsymbol{\theta}_0, \mathbf{w}_t, d_{1t})\} = \mathbf{0}$ for $t = 1, 2, \ldots, T$, by simply replacing $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$

in [14.2.3] with $\mathbf{h}^*(\boldsymbol{\theta}, \mathbf{w}_t, d_{1t})$ and minimizing [14.2.2] in the usual way. Hansen's χ^2 test statistic described in [14.1.27] based on the $\mathbf{h}^*(\cdot)$ moment conditions could then be compared with a $\chi^2(2r - a)$ critical value to provide a test of the hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

A number of other tests for structural change have been proposed by Andrews and Fair (1988) and Ghysels and Hall (1990a, b).

GMM *and Econometric Identification*

For the portfolio decision model [14.2.34], it was argued that any variable would be valid to include in the instrument vector $\mathbf{x}_t$, as long as that variable was known to investors at date t and their expectations were formed rationally. Essentially, [14.2.34] represents an asset demand curve. In the light of the discussion of simultaneous equations bias in Section 9.1, one might be troubled by the claim that it is possible to estimate a demand curve without needing to think about the way that variables may affect the demand and supply of assets in different ways.

As stressed by Garber and King (1984), the portfolio choice model avoids simultaneous equations bias because it postulates that equation [14.2.32] holds *exactly*, with no error term. The model as written claims that if the econometrician had the same information $\mathbf{x}_t^*$ used by investors, then investors' behavior could be predicted with an R^2 of unity. If there were no error term in the demand for oranges equation [9.1.1], or if the error in the demand for oranges equation were negligible compared with the error term in the supply equation, then we would not have had to worry about simultaneous equations bias in that example, either.

It is hard to take seriously the suggestion that the observed data are exactly described by [14.2.32] with no error. There are substantial difficulties in measuring aggregate consumption, population, and rates of return on assets. Even if these aggregates could in some sense be measured perfectly, it is questionable whether they are the appropriate values to be using to test a theory about individual investor preferences. And even if we had available a perfect measure of the consumption of an individual investor, the notion that the investor's utility could be represented by a function of this precise parametric form with γ constant across time is surely hard to defend.

Once we acknowledge that an error term reasonably ought to be included in [14.2.32], then it is no longer satisfactory to say that any variable dated t or earlier is a valid instrument. The difficulties with estimation are compounded by the nonlinearity of the equations of interest. If one wants to take seriously the possibility of an error term in [14.2.32] and its correlation with other variables, the best approach currently available appears to be to linearize the dynamic rational expectations model. Any variable uncorrelated with both the forecast error people make and the specification error in the model could then be used as a valid instrument for traditional instrumental variable estimation; see Sill (1992) for an illustration of this approach.

Optimal Choice of Instruments

If one does subscribe to the view that any variable dated t or earlier is a valid instrument for estimation of [14.2.32], this suggests a virtually infinite set of possible variables that could be used. One's first thought might be that, the more orthogonality conditions used, the better the resulting estimates might be. However, Monte Carlo simulations by Tauchen (1986) and Kocherlakota (1990) strongly suggest that one should be quite parsimonious in the selection of $\mathbf{x}_t$. Nelson and

Startz (1990) in particular stress that in the linear simultaneous equations model $y_t = \mathbf{z}_t'\boldsymbol{\beta} + u_t$, a good instrument not only must be uncorrelated with u_t, but must also be strongly correlated with $\mathbf{z}_t$. See Bates and White (1988), Hall (1993), and Gallant and Tauchen (1992) for further discussion on instrument selection.

14.4. GMM *and Maximum Likelihood Estimation*

In many cases the maximum likelihood estimate of $\boldsymbol{\theta}$ can also be viewed as a *GMM* estimate. This section explores this analogy and shows how asymptotic properties of maximum likelihood estimation and quasi-maximum likelihood can be obtained from the previous general results about *GMM* estimation.

The Score and Its Population Properties

Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector of variables observed at date t, and let $\mathcal{Y}_t \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_1')'$ denote the full set of data observed through date t. Suppose that the conditional density of the tth observation is given by

$$f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta}). \qquad [14.4.1]$$

Since [14.4.1] is a density, it must integrate to unity:

$$\int_{\mathcal{A}} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = 1, \qquad [14.4.2]$$

where $\mathcal{A}$ denotes the set of possible values that $\mathbf{y}_t$ could take on and $\int d\mathbf{y}_t$ denotes multiple integration:

$$\int h(\mathbf{y}_t)\, d\mathbf{y}_t \equiv \int\int \cdots \int h(y_{1t}, y_{2t}, \ldots, y_{nt})\, dy_{1t}\, dy_{2t} \cdots dy_{nt}.$$

Since [14.4.2] holds for all admissible values of $\boldsymbol{\theta}$, we can differentiate both sides with respect to $\boldsymbol{\theta}$ to conclude that

$$\int_{\mathcal{A}} \frac{\partial f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.3]$$

The conditions under which the order of differentiation and integration can be reversed as assumed in arriving at [14.4.3] and the equations to follow are known as "regularity conditions" and are detailed in Cramér (1946). Assuming that these hold, we can multiply and divide the integrand in [14.4.3] by the conditional density of $\mathbf{y}_t$:

$$\int_{\mathcal{A}} \frac{\partial f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0},$$

or

$$\int_{\mathcal{A}} \frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.4]$$

Let $\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)$ denote the derivative of the log of the conditional density of the tth observation:

$$\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) = \frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \qquad [14.4.5]$$

If there are a elements in $\boldsymbol{\theta}$, then [14.4.5] describes an $(a \times 1)$ vector for each date t that is known as the *score* of the tth observation. Since the score is a function of $\mathcal{Y}_t$, it is a random variable. Moreover, substitution of [14.4.5] into [14.4.4] reveals

that

$$\int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.6]$$

Equation [14.4.6] indicates that if the data were really generated by the density [14.4.1], then the expected value of the score conditional on information observed through date $t - 1$ should be zero:

$$E\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) | \mathcal{Y}_{t-1}\} = \mathbf{0}. \qquad [14.4.7]$$

In other words, the score vectors $\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)\}_{t=1}^{\infty}$ should form a martingale difference sequence. This observation prompted White (1987) to suggest a general specification test for models estimated by maximum likelihood based on whether the sample scores appear to be serially correlated. Expression [14.4.7] further implies that the score has unconditional expectation of zero, provided that the unconditional first moment exists:

$$E\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)\} = \mathbf{0}. \qquad [14.4.8]$$

Maximum Likelihood and GMM

Expression [14.4.8] can be viewed as a set of a orthogonality conditions that could be used to estimate the a unknown elements of $\boldsymbol{\theta}$. The *GMM* principle suggests using as an estimate of $\boldsymbol{\theta}$ the solution to

$$\mathbf{0} = (1/T) \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t). \qquad [14.4.9]$$

But this is also the characterization of the maximum likelihood estimate, which is based on maximization of

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta}),$$

the first-order conditions for which are

$$\sum_{t=1}^{T} \frac{\partial \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \qquad [14.4.10]$$

assuming an interior maximum. Recalling [14.4.5], observe that [14.4.10] and [14.4.9] are identical conditions—the *MLE* is the same as the *GMM* estimator based on the orthogonality conditions in [14.4.8].

The *GMM* formula [14.2.6] suggests that the variance-covariance matrix of the *MLE* can be approximated by

$$E[(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)'] \cong (1/T)\{\hat{\mathbf{D}}_T \hat{\mathbf{S}}_T^{-1} \hat{\mathbf{D}}_T'\}^{-1}, \qquad [14.4.11]$$

where

$$\begin{aligned} \underset{(a \times a)}{\hat{\mathbf{D}}_T'} &= \left. \frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \\ &= (1/T) \sum_{t=1}^{T} \left. \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \qquad [14.4.12] \\ &= (1/T) \sum_{t=1}^{T} \left. \frac{\partial^2 \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T}. \end{aligned}$$

Moreover, the observation in [14.4.7] that the scores are serially uncorrelated suggests estimating S by

$$\hat{\mathbf{S}}_T = (1/T) \sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'. \qquad [14.4.13]$$

The Information Matrix Equality

Expression [14.4.12] will be recognized as -1 times the second derivative estimate of the information matrix. Similarly, expression [14.4.13] is the outer-product estimate of the information matrix. That these two expressions are indeed estimating the same matrix if the model is correctly specified can be seen from calculations similar to those that produced [14.4.6]. Differentiating both sides of [14.4.6] with respect to $\boldsymbol{\theta}'$ reveals that

$$\begin{aligned} \mathbf{0} &= \int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t + \int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) \frac{\partial f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\, d\mathbf{y}_t \\ &= \int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t \\ &\quad + \int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) \frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t \end{aligned}$$

or

$$\int_{\mathcal{A}} [\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)][\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)]' f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = -\int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t.$$

This equation implies that if the model is correctly specified, the expected value of the outer product of the vector of first derivatives of the log likelihood is equal to the negative of the expected value of the matrix of second derivatives:

$$\begin{aligned} &E\left\{ \left[\frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \left[\frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \right] \middle| \mathcal{Y}_{t-1} \right\} \\ &\quad = -E\left\{ \frac{\partial^2 \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'} \middle| \mathcal{Y}_{t-1} \right\} \qquad [14.4.14] \\ &\quad \equiv \mathcal{I}_t. \end{aligned}$$

Expression [14.4.14] is known as the *information matrix equality*. Assuming that $(1/T)\sum_{t=1}^{T} \mathcal{I}_t \xrightarrow{p} \mathcal{I}$, a positive definite matrix, we can reasonably expect that for many models, the estimate $\hat{\mathbf{S}}_T$ in [14.4.13] converges in probability to the information matrix $\mathcal{I}$ and the estimate $\hat{\mathbf{D}}'_T$ in [14.4.12] converges in probability to $-\mathcal{I}$. Thus, result [14.4.11] suggests that if the data are stationary and the estimates do not fall on the boundaries of the allowable parameter space, it will often be the case that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathcal{I}^{-1}), \qquad [14.4.15]$$

where the information matrix $\mathcal{I}$ can be estimated consistently from either $-\hat{\mathbf{D}}'_T$ in [14.4.12] or $\hat{\mathbf{S}}_T$ in [14.4.13].

In small samples, the estimates $-\hat{\mathbf{D}}'_T$ and $\hat{\mathbf{S}}_T$ will differ, though if they differ too greatly this suggests that the model may be misspecified. White (1982) developed an alternative specification test based on comparing these two magnitudes.

The Wald Test for Maximum Likelihood Estimates

Result [14.4.15] suggests a general approach to testing hypotheses about the value of a parameter vector $\boldsymbol{\theta}$ that has been estimated by maximum likelihood.

Consider a null hypothesis involving m restrictions on $\boldsymbol{\theta}$ represented as $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$ where $\mathbf{g}: \mathbb{R}^a \to \mathbb{R}^m$ is a known differentiable function. The Wald test of this hypothesis is given by

$$\underset{(1\times m)}{T[\mathbf{g}(\hat{\boldsymbol{\theta}}_T)]'}\left\{\underset{(m\times a)}{\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right]}\underset{(a\times a)}{\hat{\mathcal{J}}_T^{-1}}\underset{(a\times m)}{\left[\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right]'}\right\}^{-1}\underset{(m\times 1)}{[\mathbf{g}(\hat{\boldsymbol{\theta}}_T)]}, \qquad [14.4.16]$$

which converges in distribution to a $\chi^2(m)$ variable under the null hypothesis. Again, the estimate of the information matrix $\hat{\mathcal{J}}_T$ could be based on either $-\hat{\mathbf{D}}_T'$ in [14.4.12] or $\hat{\mathbf{S}}_T$ in [14.4.13].

The Lagrange Multiplier Test

We have seen that if the model is correctly specified, the scores $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\}_{t=1}^{\infty}$ often form a martingale difference sequence. Expression [14.4.14] indicates that the conditional variance-covariance matrix of the tth score is given by $\mathcal{J}_t$. Hence, typically,

$$T\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\right]'\hat{\mathcal{J}}_T^{-1}\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\right] \xrightarrow{L} \chi^2(a). \qquad [14.4.17]$$

Expression [14.4.17] does not hold when $\boldsymbol{\theta}_0$ is replaced by $\hat{\boldsymbol{\theta}}_T$, since, from [14.4.9], this would cause [14.4.17] to be identically zero.

Suppose, however, that the likelihood function is maximized subject to m constraints on $\boldsymbol{\theta}$, and let $\tilde{\boldsymbol{\theta}}_T$ denote the restricted estimate of $\boldsymbol{\theta}$. Then, as in the *GMM* test for overidentifying restrictions [14.1.27], we would expect that

$$T\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\tilde{\boldsymbol{\theta}}_T, \mathcal{Y}_t)\right]'\hat{\mathcal{J}}_T^{-1}\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\tilde{\boldsymbol{\theta}}_T, \mathcal{Y}_t)\right] \xrightarrow{L} \chi^2(m). \qquad [14.4.18]$$

The magnitude in [14.4.18] was called the *efficient score* statistic by Rao (1948) and the *Lagrange multiplier test* by Aitchison and Silvey (1958). It provides an extremely useful class of diagnostic tests, enabling one to estimate a restricted model and test it against a more general specification without having to estimate the more general model. Breusch and Pagan (1980), Engle (1984), and Godfrey (1988) illustrated applications of the usefulness of the Lagrange multiplier principle.

Quasi-Maximum Likelihood Estimation

Even if the data were not generated by the density $f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})$, the orthogonality conditions [14.4.8] might still provide a useful description of the parameter vector of interest. For example, suppose that we incorrectly specified that a scalar series y_t came from a Gaussian $AR(1)$ process:

$$\log f(y_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta}) = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(\sigma^2) - (y_t - \phi y_{t-1})^2/(2\sigma^2),$$

with $\boldsymbol{\theta} \equiv (\phi, \sigma^2)'$. The score vector is then

$$\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) = \begin{bmatrix} (y_t - \phi y_{t-1})y_{t-1}/\sigma^2 \\ -1/(2\sigma^2) + (y_t - \phi y_{t-1})^2/(2\sigma^4) \end{bmatrix},$$

which has expectation zero whenever

$$E[(y_t - \phi y_{t-1})y_{t-1}] = 0 \qquad [14.4.19]$$

$$E[(y_t - \phi y_{t-1})^2] = \sigma^2. \qquad [14.4.20]$$

The value of the parameter ϕ that satisfies [14.4.19] corresponds to the coefficient of a linear projection of y_t on y_{t-1} regardless of the time series process followed by y_t, while σ^2 in [14.4.20] is a general characterization of the mean squared error of this linear projection. Hence, the moment conditions in [14.4.8] hold for a broad class of possible processes, and the estimates obtained by maximizing a Gaussian likelihood function (that is, the values satisfying [14.4.9]) should give reasonable estimates of the linear projection coefficient and its mean squared error for a fairly general class of possible data-generating mechanisms.

However, if the data were not generated by a Gaussian $AR(1)$ process, then the information matrix equality no longer need hold. As long as the score vector is serially uncorrelated, the variance-covariance matrix of the resulting estimates could be obtained from [14.4.11]. Proceeding in this fashion—maximizing the likelihood function in the usual way, but using [14.4.11] rather than [14.4.15] to calculate standard errors—was first proposed by White (1982), who described this approach as *quasi-maximum likelihood estimation*.[3]

APPENDIX 14.A. *Proof of Chapter 14 Proposition*

■ **Proof of Proposition 14.1.** Let $g_i(\boldsymbol{\theta}; \mathcal{Y}_T)$ denote the ith element of $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$, so that g_i: $\mathbb{R}^a \to \mathbb{R}^1$. By the mean-value theorem,

$$g_i(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = g_i(\boldsymbol{\theta}_0; \mathcal{Y}_T) + [\mathbf{d}_i(\boldsymbol{\theta}^*_{i,T}; \mathcal{Y}_T)]'(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0), \qquad [14.A.1]$$

where

$$\underset{(a\times 1)}{\mathbf{d}_i(\boldsymbol{\theta}^*_{i,T}; \mathcal{Y}_T)} = \left.\frac{\partial g_i(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{i,T}}$$

for some $\boldsymbol{\theta}^*_{i,T}$ between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_T$; notice that $\mathbf{d}_i$: $\mathbb{R}^a \to \mathbb{R}^a$. Define

$$\underset{(r\times a)}{\mathbf{D}'_T} \equiv \begin{bmatrix} [\mathbf{d}_1(\boldsymbol{\theta}^*_{1,T}; \mathcal{Y}_T)]' \\ [\mathbf{d}_2(\boldsymbol{\theta}^*_{2,T}; \mathcal{Y}_T)]' \\ \vdots \\ [\mathbf{d}_r(\boldsymbol{\theta}^*_{r,T}; \mathcal{Y}_T)]' \end{bmatrix}. \qquad [14.A.2]$$

Stacking the equations in [14.A.1] in an $(r \times 1)$ vector produces

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = \mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) + \mathbf{D}'_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0). \qquad [14.A.3]$$

If both sides of [14.A.3] are premultiplied by the $(a \times r)$ matrix

$$\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1},$$

the result is

$$\begin{aligned} &\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)] \\ &= \left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)] \qquad [14.A.4] \\ &\quad + \left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times \mathbf{D}'_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0). \end{aligned}$$

[3]For further discussion, see Gourieroux, Monfort, and Trognon (1984), Gallant and White (1988), and Wooldridge (1991a, b).

But equation [14.1.22] implies that the left side of [14.A.4] is zero, so that

$$(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = -\left[\left\{\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times \mathbf{D}_T'\right]^{-1} \times \left\{\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)]. \qquad [14.A.5]$$

Now, $\boldsymbol{\theta}_{i,T}^*$ in [14.A.1] is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_T$, so that $\boldsymbol{\theta}_{i,T}^* \xrightarrow{p} \boldsymbol{\theta}_0$ for each i. Thus, condition (c) ensures that each row of $\mathbf{D}_T'$ converges in probability to the corresponding row of $\mathbf{D}'$. Then [14.A.5] implies that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{p} -\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} \times \{\mathbf{D}\mathbf{S}^{-1}\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)\}. \qquad [14.A.6]$$

Define

$$\mathbf{C} \equiv -\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} \times \mathbf{D}\mathbf{S}^{-1},$$

so that [14.A.6] becomes

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{C}\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T).$$

Recall from condition (b) of the proposition that

$$\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S}).$$

It follows as in Example 7.5 of Chapter 7 that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \qquad [14.A.7]$$

where

$$\mathbf{V} = \mathbf{C}\mathbf{S}\mathbf{C}' = \{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1}\mathbf{D}\mathbf{S}^{-1} \times \mathbf{S} \times \mathbf{S}^{-1}\mathbf{D}'\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} = \{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1},$$

as claimed. ■

Chapter 14 Exercise

14.1. Consider the Gaussian linear regression model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t,$$

with $u_t \sim$ i.i.d. $N(0, \sigma^2)$ and u_t independent of $\mathbf{x}_\tau$ for all t and τ. Define $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$. The log of the likelihood of $(y_1, y_2, \ldots, y_T)$ conditional on $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = -(T/2)\log(2\pi) - (T/2)\log(\sigma^2) - \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2/(2\sigma^2).$$

(a) Show that the estimate $\hat{\mathbf{D}}_T'$ in [14.4.12] is given by

$$\hat{\mathbf{D}}_T' = \begin{bmatrix} -\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'/\hat{\sigma}_T^2 & \mathbf{0} \\ \mathbf{0}' & \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{1}{2\hat{\sigma}_T^4} - \frac{\hat{u}_t^2}{\hat{\sigma}_T^6}\right\} \end{bmatrix},$$

where $\hat{u}_t \equiv (y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T)$ and $\hat{\boldsymbol{\beta}}_T$ and $\hat{\sigma}_T^2$ denote the maximum likelihood estimates.

(b) Show that the estimate $\hat{\mathbf{S}}_T$ in [14.4.13] is given by

$$\hat{\mathbf{S}}_T = \begin{bmatrix} \frac{1}{T}\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'/\hat{\sigma}_T^4 & \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{\hat{u}_t^3\mathbf{x}_t}{2\hat{\sigma}_T^6}\right\} \\ \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{\hat{u}_t^3\mathbf{x}_t'}{2\hat{\sigma}_T^6}\right\} & \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{\hat{u}_t^2}{2\hat{\sigma}_T^4} - \frac{1}{2\hat{\sigma}_T^2}\right\}^2 \end{bmatrix}.$$

(c) Show that $\text{plim}(\hat{\mathbf{S}}_T) = -\text{plim}(\hat{\mathbf{D}}_T) = \mathcal{J}$, where

$$\mathcal{J} = \begin{bmatrix} \mathbf{Q}/\sigma^2 & \mathbf{0} \\ \mathbf{0}' & 1/(2\sigma^4) \end{bmatrix}$$

for $\mathbf{Q} = \text{plim}(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$.

(d) Consider a set of m linear restrictions on $\boldsymbol{\beta}$ of the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ for $\mathbf{R}$ a known $(m \times k)$ matrix and $\mathbf{r}$ a known $(m \times 1)$ vector. Show that for $\hat{\mathcal{J}}_T = -\hat{\mathbf{D}}_T$, the Wald test statistic given in [14.4.16] is identical to the Wald form of the *OLS* χ^2 test in [8.2.23] with the *OLS* estimate of the variance s_T^2 in [8.2.23] replaced by the *MLE* $\hat{\sigma}_T^2$.

(e) Show that when the lower left and upper right blocks of $\hat{\mathbf{S}}_T$ are set to their plim of zero, then the quasi-maximum likelihood Wald test of $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is identical to the heteroskedasticity-consistent form of the *OLS* χ^2 test given in [8.2.36].

Chapter 14 References

Aitchison, J., and S. D. Silvey. 1958. "Maximum Likelihood Estimation of Parameters Subject to Restraints." *Annals of Mathematical Statistics* 29:813–28.

Amemiya, Takeshi. 1974. "The Nonlinear Two-Stage Least-Squares Estimator." *Journal of Econometrics* 2:105–10.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59:817–58.

———. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica* 61:821–56.

——— and Ray C. Fair. 1988. "Inference in Nonlinear Econometric Models with Structural Change." *Review of Economic Studies* 55:615–40.

——— and J. Christopher Monahan. 1992. "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica* 60:953–66.

Bates, Charles, and Halbert White. 1988. "Efficient Instrumental Variables Estimation of Systems of Implicit Heterogeneous Nonlinear Dynamic Equations with Nonspherical Errors," in William A. Barnett, Ernst R. Berndt, and Halbert White, eds., *Dynamic Econometric Modeling*. Cambridge, England: Cambridge University Press.

Breusch, T. S., and A. R. Pagan. 1980. "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics." *Review of Economic Studies* 47:239–53.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.

Engle, Robert F. 1984. "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2. Amsterdam: North-Holland.

Ferguson, T. S. 1958. "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities." *Annals of Mathematical Statistics* 29:1046–62.

Gallant, A. Ronald. 1977. "Three-Stage Least-Squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations." *Journal of Econometrics* 5:71–88.

———. 1987. *Nonlinear Statistical Models*. New York: Wiley.

——— and George Tauchen. 1992. "Which Moments to Match?" Duke University. Mimeo.

——— and Halbert White. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Blackwell.

Garber, Peter M., and Robert G. King. 1984. "Deep Structural Excavation? A Critique of Euler Equation Methods." University of Rochester. Mimeo.

Ghysels, Eric, and Alastair Hall. 1990a. "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator." *International Economic Review* 31:355–64.

——— and ———. 1990b. "Are Consumption-Based Intertemporal Capital Asset Pricing Models Structural?" *Journal of Econometrics* 45:121–39.

Godfrey, L. G. 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge, England: Cambridge University Press.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52:681–700.

Hall, Alastair. 1993. "Some Aspects of Generalized Method of Moments Estimation," in C. R. Rao, G. S. Maddala, and H. D. Vinod, eds., *Handbook of Statistics*, Vol. 11, *Econometrics*. Amsterdam: North-Holland.

Hansen, Lars P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50:1029–54.

——— and Kenneth J. Singleton. 1982. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50:1269–86. Errata: *Econometrica* 52:267–68.

Jorgenson, D. W., and J. Laffont. 1974. "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances." *Annals of Economic and Social Measurement* 3:615–40.

Kocherlakota, Narayana R. 1990. "On Tests of Representative Consumer Asset Pricing Models." *Journal of Monetary Economics* 26:285–304.

Malinvaud, E. 1970. *Statistical Methods of Econometrics*. Amsterdam: North-Holland.

Nelson, Charles R., and Richard Startz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58:967–76.

Newey, Whitney K. 1985. "Generalized Method of Moments Specification Testing." *Journal of Econometrics* 29:229–56.

——— and Kenneth D. West. 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55:703–8.

Ogaki, Masao. 1993. "Generalized Method of Moments: Econometric Applications," in G. S. Maddala, C. R. Rao, and H. D. Vinod, eds., *Handbook of Statistics*, Vol. 11, *Econometrics*. Amsterdam: North-Holland.

Pearson, Karl. 1894. "Contribution to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London*, Series A 185:71–110.

Rao, C. R. 1948. "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Application to Problems of Estimation." *Proceedings of the Cambridge Philosophical Society* 44:50–57.

Rothenberg, Thomas J. 1973. *Efficient Estimation with A Priori Information*. New Haven, Conn.: Yale University Press.

Sargent, Thomas J. 1987. *Dynamic Macroeconomic Theory*. Cambridge, Mass.: Harvard University Press.

Sill, Keith. 1992. *Money in the Cash-in-Advance Model: An Empirical Implementation*. Unpublished Ph.D. dissertation, University of Virginia.

Tauchen, George. 1986. "Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data." *Journal of Business and Economic Statistics* 4:397–416.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817–38.

———. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.

———. 1987. "Specification Testing in Dynamic Models," in Truman F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, Vol. II. Cambridge, England: Cambridge University Press.

Wooldridge, Jeffrey M. 1991a. "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances." *Journal of Econometrics* 47:5–46.

———. 1991b. "Specification Testing and Quasi-Maximum Likelihood Estimation." *Journal of Econometrics* 48:29–55.

15 Models of Nonstationary Time Series

Up to this point our analysis has typically been confined to stationary processes. This chapter introduces several approaches to modeling nonstationary time series and analyzes the dynamic properties of different models of nonstationarity. Consequences of nonstationarity for statistical inference are investigated in subsequent chapters.

15.1. *Introduction*

Chapters 3 and 4 discussed univariate time series models that can be written in the form

$$y_t = \mu + \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots = \mu + \psi(L)\varepsilon_t, \qquad [15.1.1]$$

where $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$, roots of $\psi(z) = 0$ are outside the unit circle, and $\{\varepsilon_t\}$ is a white noise sequence with mean zero and variance σ^2. Two features of such processes merit repeating here. First, the unconditional expectation of the variable is a constant, independent of the date of the observation:

$$E(y_t) = \mu.$$

Second, as one tries to forecast the series farther into the future, the forecast $\hat{y}_{t+s|t} \equiv \hat{E}(y_{t+s}|y_t, y_{t-1}, \ldots)$ converges to the unconditional mean:

$$\lim_{s\to\infty} \hat{y}_{t+s|t} = \mu.$$

These can be quite unappealing assumptions for many of the economic and financial time series encountered in practice. For example, Figure 15.1 plots the level of nominal gross national product for the United States since World War II. There is no doubt that this series has trended upward over time, and this upward trend should be incorporated in any forecasts of this series.

There are two popular approaches to describing such trends. The first is to include a *deterministic time trend*:

$$y_t = \alpha + \delta t + \psi(L)\varepsilon_t. \qquad [15.1.2]$$

Thus, the mean μ of the stationary[1] process [15.1.1] is replaced by a linear function of the date t. Such a process is sometimes described as *trend-stationary*, because if one subtracts the trend $\alpha + \delta t$ from [15.1.2], the result is a stationary process.

The second specification is a *unit root* process,

$$(1 - L)y_t = \delta + \psi(L)\varepsilon_t, \qquad [15.1.3]$$

[1]Recall that "stationary" is taken to mean "covariance-stationary."

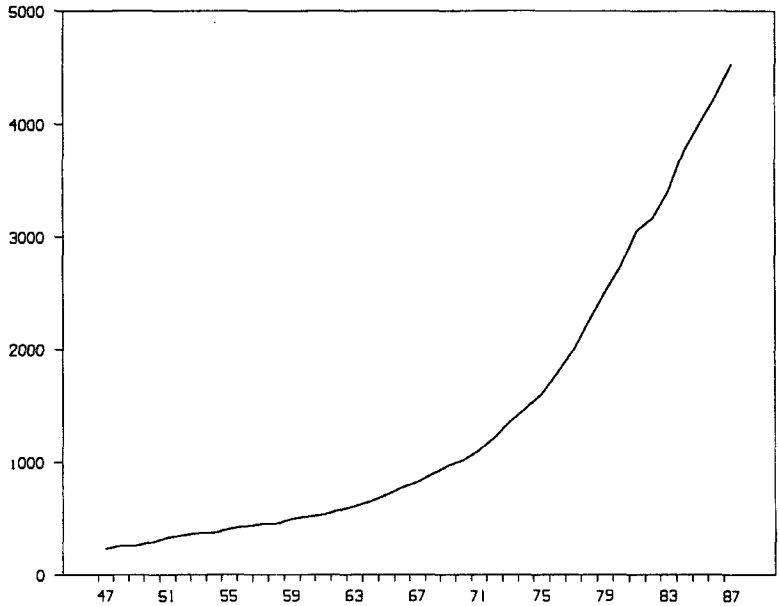


FIGURE 15.1 U.S. nominal GNP, 1947–87.

where $\psi(1) \neq 0$. For a unit root process, a stationary representation of the form of [15.1.1] describes changes in the series. For reasons that will become clear shortly, the mean of $(1 - L)y_t$ is denoted δ rather than μ.

The *first-difference* operator $(1 - L)$ will come up sufficiently often that a special symbol (the Greek letter Δ) is reserved for it:

$$\Delta y_t \equiv y_t - y_{t-1}.$$

The prototypical example of a unit root process is obtained by setting $\psi(L)$ equal to 1 in [15.1.3]:

$$y_t = y_{t-1} + \delta + \varepsilon_t. \qquad [15.1.4]$$

This process is known as a *random walk with drift* δ.

In the definition of the unit root process in [15.1.3], it was assumed that $\psi(1)$ is nonzero, where $\psi(1)$ denotes the polynomial

$$\psi(z) = 1 + \psi_1 z^1 + \psi_2 z^2 + \cdots$$

evaluated at $z = 1$. To see why such a restriction must be part of the definition of a unit root process, suppose that the original series y_t is in fact stationary with a representation of the form

$$y_t = \mu + \chi(L)\varepsilon_t.$$

If such a stationary series is differenced, the result is

$$(1 - L)y_t = (1 - L)\chi(L)\varepsilon_t \equiv \psi(L)\varepsilon_t,$$

where $\psi(L) \equiv (1 - L)\chi(L)$. This representation is in the form of [15.1.3]—if the original series y_t is stationary, then so is Δy_t. However, the moving average operator $\psi(L)$ that characterizes Δy_t has the property that $\psi(1) = (1 - 1)\cdot\chi(1) = 0$. When we stipulated that $\psi(1) \neq 0$ in [15.1.3], we were thus ruling out the possibility that the original series y_t is stationary.

It is sometimes convenient to work with a slightly different representation of

the unit root process [15.1.3]. Consider the following specification:

$$y_t = \alpha + \delta t + u_t, \qquad [15.1.5]$$

where u_t follows a zero-mean $ARMA$ process:

$$\begin{aligned}(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)u_t \\ = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t\end{aligned} \qquad [15.1.6]$$

and where the moving average operator $(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)$ is invertible. Suppose that the autoregressive operator in [15.1.6] is factored as in equation [2.4.3]:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L).$$

If all of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ are inside the unit circle, then [15.1.6] can be expressed as

$$u_t = \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{(1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L)} \varepsilon_t \equiv \psi(L)\varepsilon_t,$$

with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and roots of $\psi(z) = 0$ outside the unit circle. Thus, when $|\lambda_i| < 1$ for all i, the process [15.1.5] would just be a special case of the trend-stationary process of [15.1.2].

Suppose instead that $\lambda_1 = 1$ and $|\lambda_i| < 1$ for $i = 2, 3, \ldots, p$. Then [15.1.6] would state that

$$\begin{aligned}(1 - L)(1 - \lambda_2 L)(1 - \lambda_3 L) \cdots (1 - \lambda_p L)u_t \\ = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,\end{aligned} \qquad [15.1.7]$$

implying that

$$(1 - L)u_t = \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{(1 - \lambda_2 L)(1 - \lambda_3 L) \cdots (1 - \lambda_p L)} \varepsilon_t \equiv \psi^*(L)\varepsilon_t,$$

with $\Sigma_{j=0}^{\infty}|\psi_j^*| < \infty$ and roots of $\psi^*(z) = 0$ outside the unit circle. Thus, if [15.1.5] is first-differenced, the result is

$$(1 - L)y_t = (1 - L)\alpha + [\delta t - \delta(t - 1)] + (1 - L)u_t = 0 + \delta + \psi^*(L)\varepsilon_t,$$

which is of the form of the unit root process [15.1.3].

The representation in [15.1.5] explains the use of the term "unit root process." One of the roots or eigenvalues (λ_1) of the autoregressive polynomial in [15.1.6] is unity, and all other eigenvalues are inside the unit circle.

Another expression that is sometimes used is that the process [15.1.3] is *integrated* of order 1. This is indicated as $y_t \sim I(1)$. The term "integrated" comes from calculus; if $dy/dt = x$, then y is the integral of x. In discrete time series, if $\Delta y_t = x_t$, then y might also be viewed as the integral, or sum over t, of x.

If a process written in the form of [15.1.5] and [15.1.6] has two eigenvalues λ_1 and λ_2 that are both equal to unity with the others all inside the unit circle, then second differences of the data have to be taken before arriving at a stationary time series:

$$(1 - L)^2 y_t = \kappa + \psi(L)\varepsilon_t.$$

Such a process is said to be integrated of order 2, denoted $y_t \sim I(2)$.

A general process written in the form of [15.1.5] and [15.1.6] is called an *autoregressive integrated moving average process*, denoted $ARIMA(p, d, q)$. The first parameter (p) refers to the number of autoregressive lags (not counting the unit roots), the second parameter (d) refers to the order of integration, and the third parameter (q) gives the number of moving average lags. Taking dth differences of an $ARIMA(p, d, q)$ produces a stationary $ARMA(p, q)$ process.

15.2. *Why Linear Time Trends and Unit Roots?*

One might wonder why, for the trend-stationary specification [15.1.2], the trend is specified to be a linear function of time (δt) rather than a quadratic function ($\delta t + \gamma t^2$) or exponential ($e^{\delta t}$). Indeed, the GNP series in Figure 15.1, like many economic and financial time series, seems better characterized by an exponential trend than a linear trend. An exponential trend exhibits constant proportional growth; that is, if

$$y_t = e^{\delta t}, \qquad [15.2.1]$$

then $dy/dt = \delta \cdot y_t$. Proportional growth in the population would arise if the number of children born were a constant fraction of the current population. Proportional growth in prices (or constant inflation) would arise if the government were trying to collect a constant level of real revenues from printing money. Such stories are often an appealing starting point for thinking about the sources of time trends, and exponential growth is often confirmed by the visual appearance of the series as in Figure 15.1. For this reason, many economists simply assume that growth is of the exponential form.

Notice that if we take the natural log of the exponential trend [15.2.1], the result is a linear trend,

$$\log(y_t) = \delta t.$$

Thus, it is common to take logs of the data before attempting to describe them with the model in [15.1.2].

Similar arguments suggest taking natural logs before applying [15.1.3]. For small changes, the first difference of the log of a variable is approximately the same as the percentage change in the variable:

$$\begin{aligned}(1 - L)\log(y_t) &= \log(y_t/y_{t-1}) \\ &= \log\{1 + [(y_t - y_{t-1})/y_{t-1}]\} \\ &\cong (y_t - y_{t-1})/y_{t-1},\end{aligned}$$

where we have used the fact that for x close to zero, $\log(1 + x) \cong x$.[2] Thus, if the logs of a variable are specified to follow a unit root process, the assumption is that the rate of growth of the series is a stationary stochastic process. The same arguments used to justify taking logs before applying [15.1.2] also suggest taking logs before applying [15.1.3].

Often the units are slightly more convenient if $\log(y_t)$ is multiplied by 100. Then changes are measured directly in units of percentage change. For example, if $(1 - L)[100 \times \log(y_t)] = 1.0$, then y_t is 1% higher than y_{t-1}.

15.3. *Comparison of Trend-Stationary and Unit Root Processes*

This section compares a trend-stationary process [15.1.2] with a unit root process [15.1.3] in terms of forecasts of the series, variance of the forecast error, dynamic multipliers, and transformations needed to achieve stationarity.

[2]See result [A.3.36] in the Mathematical Review (Appendix A) at the end of the book.

Comparison of Forecasts

To forecast a trend-stationary process [15.1.2], the known deterministic component ($\alpha + \delta t$) is simply added to the forecast of the stationary stochastic component:

$$\hat{y}_{t+s|t} = \alpha + \delta(t + s) + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots. \qquad [15.3.1]$$

Here $\hat{y}_{t+s|t}$ denotes the linear projection of y_{t+s} on a constant and $y_t, y_{t-1}, \ldots$. Note that for nonstationary processes, we will follow the convention that the "constant" term in a linear projection, in this case $\alpha + \delta(t + s)$, can be different for each date $t + s$. As the forecast horizon (s) grows large, absolute summability of $\{\psi_j\}$ implies that this forecast converges in mean square to the time trend:

$$E[\hat{y}_{t+s|t} - \alpha - \delta(t + s)]^2 \to 0 \qquad \text{as} \quad s \to \infty.$$

To forecast the unit root process [15.1.3], recall that the change Δy_t is a stationary process that can be forecast using the standard formula:

$$\begin{aligned}\Delta\hat{y}_{t+s|t} &\equiv \hat{E}[(y_{t+s} - y_{t+s-1})|y_t, y_{t-1}, \ldots] \\ &= \delta + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots.\end{aligned} \qquad [15.3.2]$$

The level of the variable at date $t + s$ is simply the sum of the changes between t and $t + s$:

$$\begin{aligned}y_{t+s} &= (y_{t+s} - y_{t+s-1}) + (y_{t+s-1} - y_{t+s-2}) + \cdots \\ &\quad + (y_{t+1} - y_t) + y_t \\ &= \Delta y_{t+s} + \Delta y_{t+s-1} + \cdots + \Delta y_{t+1} + y_t.\end{aligned} \qquad [15.3.3]$$

Taking the linear projection of [15.3.3] on a constant and $y_t, y_{t-1}, \ldots$ and substituting from [15.3.2] gives

$$\begin{aligned}\hat{y}_{t+s|t} &= \Delta\hat{y}_{t+s|t} + \Delta\hat{y}_{t+s-1|t} + \cdots + \Delta\hat{y}_{t+1|t} + y_t \\ &= \{\delta + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots\} \\ &\quad + \{\delta + \psi_{s-1}\varepsilon_t + \psi_s\varepsilon_{t-1} + \psi_{s+1}\varepsilon_{t-2} + \cdots\} \\ &\quad + \cdots + \{\delta + \psi_1\varepsilon_t + \psi_2\varepsilon_{t-1} + \psi_3\varepsilon_{t-2} + \cdots\} + y_t\end{aligned}$$

or

$$\begin{aligned}\hat{y}_{t+s|t} &= s\delta + y_t + (\psi_s + \psi_{s-1} + \cdots + \psi_1)\varepsilon_t \\ &\quad + (\psi_{s+1} + \psi_s + \cdots + \psi_2)\varepsilon_{t-1} + \cdots.\end{aligned} \qquad [15.3.4]$$

Further insight into the forecast of a unit root process is obtained by analyzing some special cases. Consider first the random walk with drift [15.1.4], in which $\psi_1 = \psi_2 = \cdots = 0$. Then [15.3.4] becomes

$$\hat{y}_{t+s|t} = s\delta + y_t.$$

A random walk with drift δ is expected to grow at the constant rate of δ per period from whatever its current value y_t happens to be.

Consider next an $ARIMA(0, 1, 1)$ specification ($\psi_1 = \theta$, $\psi_2 = \psi_3 = \cdots = 0$). Then

$$\hat{y}_{t+s|t} = s\delta + y_t + \theta\varepsilon_t. \qquad [15.3.5]$$

Here, the current level of the series y_t along with the current innovation ε_t again defines a base from which the variable is expected to grow at the constant rate δ.

Notice that ε_t is the one-period-ahead forecast error:

$$\varepsilon_t = y_t - \hat{y}_{t|t-1}.$$

It follows from [15.3.5] that for $\delta = 0$ and $s = 1$,

$$\hat{y}_{t+1|t} = y_t + \theta(y_t - \hat{y}_{t|t-1}) \qquad [15.3.6]$$

or

$$\hat{y}_{t+1|t} = (1 + \theta)y_t - \theta\hat{y}_{t|t-1}. \qquad [15.3.7]$$

Equation [15.3.7] takes the form of a simple first-order difference equation, relating $\hat{y}_{t+1|t}$ to its own lagged value and to the input variable $(1 + \theta)y_t$. Provided that $|\theta| < 1$, expression [15.3.7] can be written using result [2.2.9] as

$$\begin{aligned}\hat{y}_{t+1|t} &= [(1 + \theta)y_t] + (-\theta)[(1 + \theta)y_{t-1}] \\ &\quad + (-\theta)^2[(1 + \theta)y_{t-2}] + (-\theta)^3[(1 + \theta)y_{t-3}] + \cdots \qquad [15.3.8] \\ &= (1 + \theta)\sum_{j=0}^{\infty}(-\theta)^j y_{t-j}.\end{aligned}$$

Expression [15.3.7] is sometimes described as *adaptive expectations*, and its implication [15.3.8] is referred to as *exponential smoothing*; typical applications assume that $-1 < \theta < 0$. Letting y_t denote income, Friedman (1957) used exponential smoothing to construct one of his measures of *permanent income*. Muth (1960) noted that adaptive expectations or exponential smoothing corresponds to a rational forecast of future income only if y_t follows an $ARIMA(0, 1, 1)$ process and the smoothing weight $(-\theta)$ is chosen to equal the negative of the moving average coefficient of the differenced data (θ).

For an $ARIMA(0, 1, q)$ process, the value of y_t and the q most recent values of ε_t influence the forecasts $\hat{y}_{t+1|t}, \hat{y}_{t+2|t}, \ldots, \hat{y}_{t+q|t}$, but thereafter the series is expected to grow at the rate δ. For an $ARIMA(p, 1, q)$, the forecast growth rate asymptotically approaches δ.

Thus, the parameter δ in the unit root process [15.1.3] plays a similar role to that of δ in the deterministic time trend [15.1.2]. With either specification, the forecast $\hat{y}_{t+s|t}$ in [15.3.1] or [15.3.4] converges to a linear function of the forecast horizon s with slope δ; see Figure 15.2. The key difference is in the intercept of the line. For a trend-stationary process, the forecast converges to a line whose intercept is the same regardless of the value of y_t. By contrast, the intercept of the limiting forecast for a unit root process is continually changing with each new observation on y.

Comparison of Forecast Errors

The trend-stationary and unit root specifications are also very different in their implications for the variance of the forecast error. For the trend-stationary process [15.1.2], the s-period-ahead forecast error is

$$\begin{aligned}y_{t+s} - \hat{y}_{t+s|t} &= \{\alpha + \delta(t + s) + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots \\ &\quad + \psi_{s-1}\varepsilon_{t+1} + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \cdots\} \\ &\quad - \{\alpha + \delta(t + s) + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots\} \\ &= \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots + \psi_{s-1}\varepsilon_{t+1}.\end{aligned}$$

The mean squared error (MSE) of this forecast is

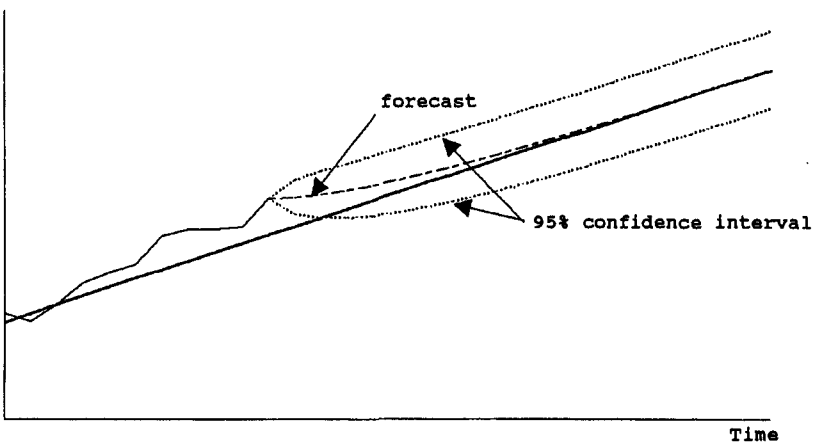
$$E[y_{t+s} - \hat{y}_{t+s|t}]^2 = \{1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{s-1}^2\}\sigma^2.$$

The MSE increases with the forecasting horizon s, though as s becomes large, the added uncertainty from forecasting farther into the future becomes negligible:

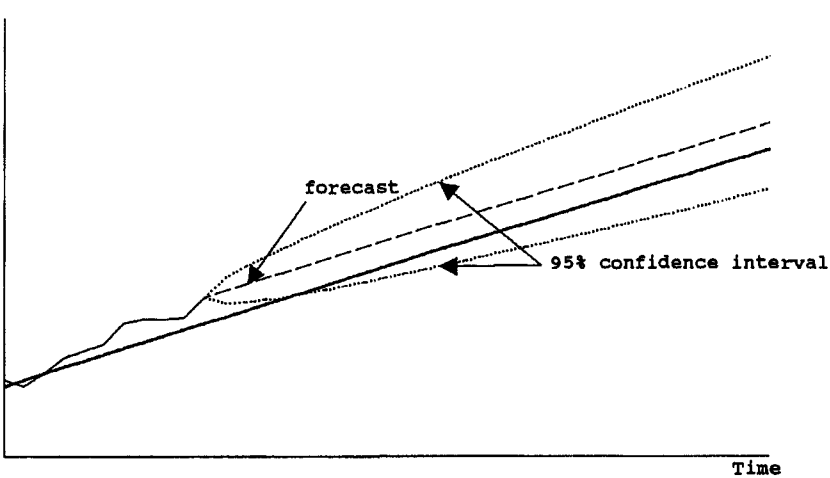
$$\lim_{s\to\infty} E[y_{t+s} - \hat{y}_{t+s|t}]^2 = \{1 + \psi_1^2 + \psi_2^2 + \cdots\}\sigma^2.$$

Note that the limiting MSE is just the unconditional variance of the stationary component $\psi(L)\varepsilon_t$.

By contrast, for the unit root process [15.1.3], the s-period-ahead forecast error is



(a) Trend-stationary process



(b) Unit root process

FIGURE 15.2 Forecasts and 95% confidence intervals.

$$
\begin{aligned}
y_{t+s} - \hat{y}_{t+s|t} &= \{\Delta y_{t+s} + \Delta y_{t+s-1} + \cdots + \Delta y_{t+1} + y_t\} \\
&\quad - \{\Delta\hat{y}_{t+s|t} + \Delta\hat{y}_{t+s-1|t} + \cdots + \Delta\hat{y}_{t+1|t} + y_t\} \\
&= \{\varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \cdots + \psi_{s-1}\varepsilon_{t+1}\} \\
&\quad + \{\varepsilon_{t+s-1} + \psi_1\varepsilon_{t+s-2} + \cdots + \psi_{s-2}\varepsilon_{t+1}\} + \cdots + \{\varepsilon_{t+1}\} \\
&= \varepsilon_{t+s} + \{1 + \psi_1\}\varepsilon_{t+s-1} + \{1 + \psi_1 + \psi_2\}\varepsilon_{t+s-2} + \cdots \\
&\quad + \{1 + \psi_1 + \psi_2 + \cdots + \psi_{s-1}\}\varepsilon_{t+1},
\end{aligned}
$$

with *MSE*

$$
\begin{aligned}
E[y_{t+s} - \hat{y}_{t+s|t}]^2 = \{&1 + (1 + \psi_1)^2 + (1 + \psi_1 + \psi_2)^2 + \cdots \\
&+ (1 + \psi_1 + \psi_2 + \cdots + \psi_{s-1})^2\}\sigma^2.
\end{aligned}
$$

The *MSE* again increases with the length of the forecasting horizon s, though in contrast to the trend-stationary case, the *MSE* does not converge to any fixed value as s goes to infinity. Instead, it asymptotically approaches a linear function of s with slope $(1 + \psi_1 + \psi_2 + \cdots)^2\sigma^2$. For example, for an *ARIMA*(0, 1, 1) process,

$$
E[y_{t+s} - \hat{y}_{t+s|t}]^2 = \{1 + (s - 1)(1 + \theta)^2\}\sigma^2. \qquad [15.3.9]
$$

To summarize, for a trend-stationary process the *MSE* reaches a finite bound as the forecast horizon becomes large, whereas for a unit root process the *MSE* eventually grows linearly with the forecast horizon. This result is again illustrated in Figure 15.2.

Note that since the *MSE* grows linearly with the forecast horizon s, the standard deviation of the forecast error grows with the square root of s. On the other hand, if $\delta > 0$, then the forecast itself grows linearly in s. Thus, a 95% confidence interval for y_{t+s} expands more slowly than the level of the series, meaning that data from a unit root process with positive drift are certain to exhibit an upward trend if observed for a sufficiently long period. In this sense the trend introduced by a nonzero drift δ asymptotically dominates the increasing variability arising over time due to the unit root component. This result is very important for understanding the asymptotic statistical results to be presented in Chapters 17 and 18.

Figure 15.3 plots realizations of a Gaussian random walk without drift and with drift. The random walk without drift, shown in panel (a), shows no tendency to return to its starting value or any unconditional mean. The random walk with drift, shown in panel (b), shows no tendency to return to a fixed deterministic trend line, though the series is asymptotically dominated by the positive drift term.

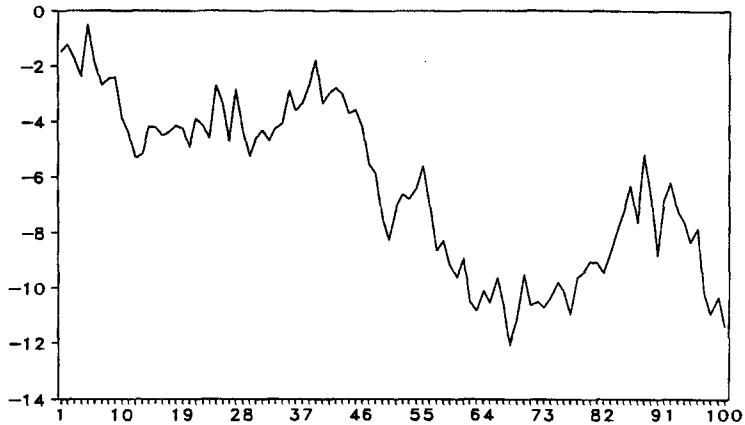
Comparison of Dynamic Multipliers

Another difference between trend-stationary and unit root processes is the persistence of innovations. Consider the consequences for y_{t+s} if ε_t were to increase by one unit with ε's for all other dates unaffected. For the trend-stationary process [15.1.2], this dynamic multiplier is given by

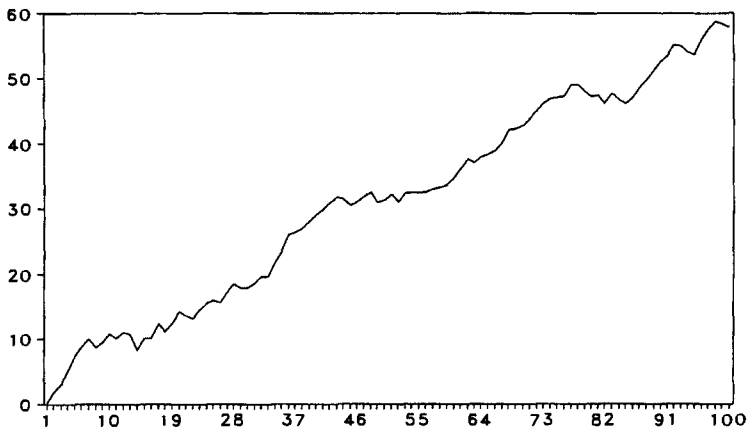
$$
\frac{\partial y_{t+s}}{\partial \varepsilon_t} = \psi_s.
$$

For a trend-stationary process, then, the effect of any stochastic disturbance eventually wears off:

$$
\lim_{s\to\infty} \frac{\partial y_{t+s}}{\partial \varepsilon_t} = 0.
$$



(a) Random walk without drift



(b) Random walk with drift

FIGURE 15.3 Sample realizations from Gaussian unit root processes.

By contrast, for a unit root process, the effect of ε_t on y_{t+s} is seen from [15.3.4] to be[3]

$$\frac{\partial y_{t+s}}{\partial \varepsilon_t} = \frac{\partial y_t}{\partial \varepsilon_t} + \psi_s + \psi_{s-1} + \cdots + \psi_1 = 1 + \psi_1 + \psi_2 + \cdots + \psi_s.$$

An innovation ε_t has a permanent effect on the level of y that is captured by

$$\lim_{s\to\infty} \frac{\partial y_{t+s}}{\partial \varepsilon_t} = 1 + \psi_1 + \psi_2 + \cdots = \psi(1). \qquad [15.3.10]$$

[3]This, of course, contrasts with the multiplier that describes the effect of ε_t on the *change* between y_{t+s} and y_{t+s-1}, which is given by

$$\frac{\partial \Delta y_{t+s}}{\partial \varepsilon_t} = \psi_s.$$

As an illustration of calculating such a multiplier, the following *ARIMA*(4, 1, 0) model was estimated for y_t equal to 100 times the log of quarterly U.S. real GNP (t = 1952:II to 1984:IV):

$$\Delta y_t = 0.555 + 0.312\,\Delta y_{t-1} + 0.122\,\Delta y_{t-2} - 0.116\,\Delta y_{t-3} - 0.081\,\Delta y_{t-4} + \varepsilon_t.$$

For this specification, the permanent effect of a one-unit change in ε_t on the level of real GNP is estimated to be

$$\psi(1) = 1/\phi(1) = 1/(1 - 0.312 - 0.122 + 0.116 + 0.081) = 1.31.$$

Transformations to Achieve Stationarity

A final difference between trend-stationary and unit root processes that deserves comment is the transformation of the data needed to generate a stationary time series. If the process is really trend stationary as in [15.1.2], the appropriate treatment is to subtract δt from y_t to produce a stationary representation of the form of [15.1.1]. By contrast, if the data were really generated by the unit root process [15.1.3], subtracting δt from y_t would succeed in removing the time-dependence of the mean but not the variance. For example, if the data were generated by [15.1.4], the random walk with drift, then

$$y_t - \delta t = y_0 + (\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t) \equiv y_0 + u_t.$$

The variance of the residual u_t is $t\sigma^2$; it grows with the date of the observation. Thus, subtracting a time trend from a unit root process is not sufficient to produce a stationary time series.

The correct treatment for a unit root process is to difference the series, and for this reason a process described by [15.1.3] is sometimes called a *difference-stationary* process. Note, however, that if one were to try to difference a trend-stationary process [15.1.2], the result would be

$$\Delta y_t = \delta + (1 - L)\psi(L)\varepsilon_t.$$

This is a stationary time series, but a unit root has been introduced into the moving average representation. Thus, the result would be a noninvertible process subject to the potential difficulties discussed in Chapters 3 through 5.

15.4. *The Meaning of Tests for Unit Roots*

Knowing whether nonstationarity in the data is due to a deterministic time trend or a unit root would seem to be a very important question. For example, macroeconomists are very interested in knowing whether economic recessions have permanent consequences for the level of future GNP, or instead represent temporary downturns with the lost output eventually made up during the recovery. Nelson and Plosser (1982) argued that many economic series are better characterized by unit roots than by deterministic time trends. A number of economists have tried to measure the size of the permanent consequences by estimating $\psi(1)$ for various time series representations of GNP growth.[4]

Although it might be very interesting to know whether a time series has a unit root, several recent papers have argued that the question is inherently un-

[4]See, for example, Watson (1986), Clark (1987), Campbell and Mankiw (1987a, b), Cochrane (1988), Gagnon (1988), Stock and Watson (1988), Durlauf (1989), and Hamilton (1989).

answerable on the basis of a finite sample of observations.[5] The argument takes the form of two observations.

The first observation is that for any unit root process there exists a stationary process that will be impossible to distinguish from the unit root representation for any given sample size T. Such a stationary process is found easily enough by setting one of the eigenvalues close to but not quite equal to unity. For example, suppose the sample consists of $T = 10{,}000$ observations that were really generated by a driftless random walk:

$$y_t = y_{t-1} + \varepsilon_t \qquad \textit{true model} \text{ (unit root).} \qquad [15.4.1]$$

Consider trying to distinguish this from the following stationary process:

$$y_t = \phi y_{t-1} + \varepsilon_t \qquad |\phi| < 1 \qquad \textit{false model} \text{ (stationary).} \qquad [15.4.2]$$

The s-period-ahead forecast of [15.4.1] is

$$\hat{y}_{t+s|t} = y_t \qquad [15.4.3]$$

with *MSE*

$$E(y_{t+s} - \hat{y}_{t+s|t})^2 = s\sigma^2. \qquad [15.4.4]$$

The corresponding forecast of [15.4.2] is

$$\hat{y}_{t+s|t} = \phi^s y_t \qquad [15.4.5]$$

with *MSE*

$$E(y_{t+s} - \hat{y}_{t+s|t})^2 = (1 + \phi^2 + \phi^4 + \cdots + \phi^{2(s-1)})\cdot\sigma^2. \qquad [15.4.6]$$

Clearly there exists a value of ϕ sufficiently close to unity such that the observable implications of the stationary representation ([15.4.5] and [15.4.6]) are arbitrarily close to those of the unit root process ([15.4.3] and [15.4.4]) in a sample of size 10,000.

More formally, the conditional likelihood function for a Gaussian process characterized by [15.1.7] is continuous in the parameter λ_1. Hence, given any fixed sample size T, any small numbers η and ε, and any unit root specification with $\lambda_1 = 1$, there exists a stationary specification with $\lambda_1 < 1$ with the property that the probability is less than ε that one would observe a sample of size T for which the value of the likelihood implied by the unit root representation differs by more than η from the value of the likelihood implied by the stationary representation.

The converse proposition is also true—for any stationary process and a given sample size T, there exists a unit root process that will be impossible to distinguish from the unit root representation. Again, consider a simple example. Suppose the true process is white noise:

$$y_t = \varepsilon_t \qquad \textit{true model} \text{ (stationary).} \qquad [15.4.7]$$

Consider trying to distinguish this from

$$(1 - L)y_t = (1 + \theta L)\varepsilon_t \qquad |\theta| < 1 \qquad \textit{false model} \text{ (unit root)} \qquad [15.4.8]$$
$$y_0 = \varepsilon_0 = 0.$$

The s-period-ahead forecast of [15.4.7] is

$$\hat{y}_{t+s|t} = 0$$

with *MSE*

$$E(y_{t+s} - \hat{y}_{t+s|t})^2 = \sigma^2.$$

[5]See Blough (1992a, b), Cochrane (1991), Christiano and Eichenbaum (1990), Stock (1990), and Sims (1989). The sharpest statement of this view, and the perspective on which the remarks in the text are based, is that of Blough.

The forecast of [15.4.8] is obtained from [15.3.5]:

$$\begin{aligned}\hat{y}_{t+s|t} &= y_t + \theta\varepsilon_t \\ &= \{\Delta y_t + \Delta y_{t-1} + \cdots + \Delta y_2 + y_1\} + \theta\varepsilon_t \\ &= \{(\varepsilon_t + \theta\varepsilon_{t-1}) + (\varepsilon_{t-1} + \theta\varepsilon_{t-2}) + \cdots + (\varepsilon_2 + \theta\varepsilon_1) + (\varepsilon_1)\} + \theta\varepsilon_t \\ &= (1 + \theta)\{\varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1\}.\end{aligned}$$

From [15.3.9], the *MSE* of the s-period-ahead forecast is

$$E(y_{t+s} - \hat{y}_{t+s|t})^2 = \{1 + (s - 1)(1 + \theta)^2\}\sigma^2.$$

Again, clearly, given any fixed sample size T, there exists a value of θ sufficiently close to -1 that the unit root process [15.4.8] will have virtually the identical observable implications to those of the stationary process [15.4.7].

Unit root and stationary processes differ in their implications at infinite time horizons, but for any given finite number of observations on the time series, there is a representative from either class of models that could account for all the observed features of the data. We therefore need to be careful with our choice of wording—testing whether a particular time series "contains a unit root," or testing whether innovations "have a permanent effect on the level of the series," however interesting, is simply impossible to do.

Another way to express this is as follows. For a unit root process given by [15.1.3], the autocovariance-generating function of $(1 - L)y_t$ is

$$g_{\Delta Y}(z) = \psi(z)\sigma^2\psi(z^{-1}).$$

The autocovariance-generating function evaluated at $z = 1$ is then

$$g_{\Delta Y}(1) = [\psi(1)]^2\sigma^2. \qquad [15.4.9]$$

Recalling that the population spectrum of Δy at frequency ω is defined by

$$s_{\Delta Y}(\omega) = \frac{1}{2\pi}\, g_{\Delta Y}(e^{-i\omega}),$$

expression [15.4.9] can alternatively be described as 2π times the spectrum at frequency zero:

$$s_{\Delta Y}(0) = \frac{1}{2\pi}\,[\psi(1)]^2\sigma^2.$$

By contrast, if the true process is the trend-stationary specification [15.1.2], the autocovariance-generating function of Δy can be calculated from [3.6.15] as

$$g_{\Delta Y}(z) = (1 - z)\psi(z)\sigma^2\psi(z^{-1})(1 - z^{-1}),$$

which evaluated at $z = 1$ is zero. Thus, if the true process is trend-stationary, the population spectrum of Δy at frequency zero is zero, whereas if the true process is characterized by a unit root, the population spectrum of Δy at frequency zero is positive.

The question of whether y_t follows a unit root process can thus equivalently be expressed as a question of whether the population spectrum of Δy at frequency zero is zero. However, there is no information in a sample of size T about cycles with period greater than T, just as there is no information in a sample of size T about the dynamic multiplier for a horizon $s > T$.

These observations notwithstanding, there are several closely related and very interesting questions that are answerable. Given enough data, we certainly can ask

whether innovations have a significant effect on the level of the series over a specified finite horizon. For a fixed time horizon (say, $s = 3$ years), there exists a sample size (say, the half century of observations since World War II) such that we can meaningfully inquire whether $\partial y_{t+s}/\partial \varepsilon_t$ is close to zero. We cannot tell whether the data were really generated by [15.4.1] or a close relative of the form of [15.4.2], but we can measure whether innovations have much persistence over a fixed interval (as in [15.4.1] or [15.4.2]) or very little persistence over that interval (as in [15.4.7] or [15.4.8]).

We can also arrive at a testable hypothesis if we are willing to restrict further the class of processes considered. Suppose the dynamics of a given sample $\{y_1, \ldots, y_T\}$ are to be modeled using an autoregression of fixed, known order p. For example, suppose we are committed to using an $AR(1)$ process:

$$y_t = \phi y_{t-1} + \varepsilon_t. \qquad [15.4.10]$$

Within this class of models, the restriction

$$H_0\colon \phi = 1$$

is certainly testable. While it is true that there exist local alternatives (such as $\phi = 0.99999$) against which a test would have essentially no power, this is true of most hypothesis tests. There are also alternatives (such as $\phi = 0.3$) that would lead to certain rejection of H_0 given enough observations. The hypothesis "$\{y_t\}$ is an $AR(1)$ process with a unit root" is potentially refutable; the hypothesis "$\{y_t\}$ is a general unit root process of the form [15.1.3]" is not.

There may be good reasons to restrict ourselves to consider only low-order autoregressive representations. Parsimonious models often perform best, and autoregressions are much easier to estimate and forecast than moving average processes, particularly moving average processes with a root near unity.

If we are indeed committed to describing the data with a low-order autoregression, knowing whether the further restriction of a unit root should be imposed can clearly be important for two reasons. The first involves a familiar trade-off between efficiency and consistency. If a restriction (in this case, a unit root) is true, more efficient estimates result from imposing it. Estimates of the other coefficients and dynamic multipliers will be more accurate, and forecasts will be better. If the restriction is false, the estimates are unreliable no matter how large the sample. Researchers differ in their advice on how to deal with this trade-off. One practical guide is to estimate the model both with and without the unit root imposed. If the key inferences are similar, so much the better. If the inferences differ, some attempt at explaining the conflicting findings (as in Christiano and Ljungqvist, 1988, or Stock and Watson, 1989) may be desirable.

In addition to the familiar trade-off between efficiency and consistency, the decision whether or not to impose unit roots on an autoregression also raises issues involving the asymptotic distribution theory one uses to test hypotheses about the process. This issue is explored in detail in later chapters.

15.5. *Other Approaches to Trended Time Series*

Although most of the analysis of nonstationarity in this book will be devoted to unit roots and time trends, this section briefly discusses two alternative approaches to modeling nonstationarity: fractionally integrated processes and processes with occasional, discrete shifts in the time trend.

Fractional Integration

Recall that an integrated process of order d can be represented in the form

$$(1 - L)^d y_t = \psi(L)\varepsilon_t, \qquad [15.5.1]$$

with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$. The normal assumption is that $d = 1$, or that the first difference of the series is stationary. Occasionally one finds a series for which $d = 2$ might be a better choice.

Granger and Joyeux (1980) and Hosking (1981) suggested that noninteger values of d in [15.5.1] might also be useful. To understand the meaning of [15.5.1] for noninteger d, consider the $MA(\infty)$ representation implied by [15.5.1]. It will be shown shortly that the inverse of the operator $(1 - L)^d$ exists provided that $d < \frac{1}{2}$. Multiplying both sides of [15.5.1] by $(1 - L)^{-d}$ results in

$$y_t = (1 - L)^{-d}\psi(L)\varepsilon_t. \qquad [15.5.2]$$

For z a scalar, define the function

$$f(z) \equiv (1 - z)^{-d}.$$

This function has derivatives given by

$$\begin{aligned}
\frac{\partial f}{\partial z} &= d\cdot(1 - z)^{-d-1} \\
\frac{\partial^2 f}{\partial z^2} &= (d + 1)\cdot d\cdot(1 - z)^{-d-2} \\
\frac{\partial^3 f}{\partial z^3} &= (d + 2)\cdot(d + 1)\cdot d\cdot(1 - z)^{-d-3} \\
&\vdots \\
\frac{\partial^j f}{\partial z^j} &= (d + j - 1)\cdot(d + j - 2)\cdots(d + 1)\cdot d\cdot(1 - z)^{-d-j}.
\end{aligned}$$

A power series expansion for $f(z)$ around $z = 0$ is thus given by

$$\begin{aligned}
(1 - z)^{-d} &= f(0) + \left.\frac{\partial f}{\partial z}\right|_{z=0}\cdot z + \frac{1}{2!}\left.\frac{\partial^2 f}{\partial z^2}\right|_{z=0}\cdot z^2 + \frac{1}{3!}\left.\frac{\partial^3 f}{\partial z^3}\right|_{z=0}\cdot z^3 + \cdots \\
&= 1 + dz + (1/2!)(d + 1)dz^2 + (1/3!)(d + 2)(d + 1)dz^3 + \cdots.
\end{aligned}$$

This suggests that the operator $(1 - L)^{-d}$ might be represented by the filter

$$\begin{aligned}
(1 - L)^{-d} &= 1 + dL + (1/2!)(d + 1)dL^2 \\
&\quad + (1/3!)(d + 2)(d + 1)dL^3 + \cdots \qquad [15.5.3] \\
&= \sum_{j=0}^{\infty} h_j L^j,
\end{aligned}$$

where $h_0 \equiv 1$ and

$$h_j \equiv (1/j!)(d + j - 1)(d + j - 2)(d + j - 3)\cdots(d + 1)(d). \qquad [15.5.4]$$

Appendix 15.A to this chapter establishes that if $d < 1$, h_j can be approximated for large j by

$$h_j \cong (j + 1)^{d-1}. \qquad [15.5.5]$$

Thus, the time series model

$$y_t = (1 - L)^{-d}\varepsilon_t = h_0\varepsilon_t + h_1\varepsilon_{t-1} + h_2\varepsilon_{t-2} + \cdots \qquad [15.5.6]$$

describes an $MA(\infty)$ representation in which the impulse-response coefficient h_j behaves for large j like $(j + 1)^{d-1}$. For comparison, recall that the impulse-response coefficient associated with the $AR(1)$ process $y_t = (1 - \phi L)^{-1}\varepsilon_t$ is given by ϕ^j. The impulse-response coefficients for a stationary $ARMA$ process decay geometrically, in contrast to the slower decay implied by [15.5.5]. Because of this slower rate of decay, Granger and Joyeux proposed the fractionally integrated process as an approach to modeling long memories in a time series.

In a finite sample, this long memory could be approximated arbitrarily well with a suitably large-order $ARMA$ representation. The goal of the fractional-difference specification is to capture parsimoniously long-run multipliers that decay very slowly.

The sequence of limiting moving average coefficients $\{h_j\}_{j=0}^{\infty}$ given in [15.5.4] can be shown to be square-summable provided that $d < \frac{1}{2}$.[6]

$$\sum_{j=0}^{\infty} h_j^2 < \infty \qquad \text{for } d < \tfrac{1}{2}.$$

Thus, [15.5.6] defines a covariance-stationary process provided that $d < \frac{1}{2}$. If $d > \frac{1}{2}$, the proposal is to difference the process before describing it by [15.5.2]. For example, if $d = 0.7$, the process of [15.5.1] implies

$$(1 - L)^{-0.3}(1 - L)y_t = \psi(L)\varepsilon_t;$$

that is, Δy_t is fractionally integrated with parameter $d = -0.3 < \frac{1}{2}$.

Conditions under which fractional integration could arise from aggregation of other processes were described by Granger (1980). Geweke and Porter-Hudak (1983) and Sowell (1992) proposed techniques for estimating d. Diebold and Rudebusch (1989) analyzed GNP data and the persistence of business cycle fluctuations using this approach, while Lo (1991) provided an interesting investigation of the persistence of movements in stock prices.

Occasional Breaks in Trend

According to the unit root specification [15.1.3], events are occurring all the time that permanently affect the course of y. Perron (1989) and Rappoport and Reichlin (1989) have argued that economic events that have large permanent effects

[6]Reasoning as in Appendix 3.A to Chapter 3.

$$\begin{aligned}\sum_{j=0}^{N-1} (j + 1)^{2(d-1)} &= \sum_{j=1}^{N} j^{2(d-1)} \\ &< 1 + \int_1^N x^{2(d-1)}\,dx \\ &= 1 + [1/(2d - 1)]x^{2d-1}\big|_{x=1}^{N} \\ &= 1 + [1/(2d - 1)]\cdot[N^{2d-1} - 1],\end{aligned}$$

which converges to $1 - [1/(2d - 1)]$ as $N \to \infty$, provided that $d < \frac{1}{2}$.

are relatively rare. The idea can be illustrated with the following model, in which y_t is stationary around a trend with a single break:

$$y_t = \begin{cases} \alpha_1 + \delta t + \varepsilon_t & \text{for } t < T_0 \\ \alpha_2 + \delta t + \varepsilon_t & \text{for } t \geq T_0. \end{cases} \qquad [15.5.7]$$

The finding is that such series would appear to exhibit unit root nonstationarity on the basis of the tests to be discussed in Chapter 17.

Another way of thinking about the process in [15.5.7] is as follows:

$$\Delta y_t = \xi_t + \delta + \varepsilon_t - \varepsilon_{t-1}, \qquad [15.5.8]$$

where $\xi_t = (\alpha_2 - \alpha_1)$ when $t = T_0$ and is zero otherwise. Suppose ξ_t is viewed as a random variable with some probability distribution—say,

$$\xi_t = \begin{cases} \alpha_2 - \alpha_1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

Evidently, p must be quite small to represent the idea that this is a relatively rare event. Equation [15.5.8] could then be rewritten as

$$\Delta y_t = \mu + \eta_t, \qquad [15.5.9]$$

where

$$\mu = p(\alpha_2 - \alpha_1) + \delta$$
$$\eta_t = \xi_t - p(\alpha_2 - \alpha_1) + \varepsilon_t - \varepsilon_{t-1}.$$

But η_t is the sum of a zero-mean white noise process $[\xi_t - p(\alpha_2 - \alpha_1)]$ and an independent $MA(1)$ process $[\varepsilon_t - \varepsilon_{t-1}]$. Therefore, an $MA(1)$ representation for η_t exists. From this perspective, [15.5.9] could be viewed as an $ARIMA(0, 1, 1)$ process,

$$\Delta y_t = \mu + v_t + \theta v_{t-1},$$

with a non-Gaussian distribution for the innovations v_t:

$$v_t = y_t - \hat{E}(y_t | y_{t-1}, y_{t-2}, \ldots).$$

The optimal linear forecasting rule,

$$\hat{E}(y_{t+s} | y_t, y_{t-1}, \ldots) = \mu s + y_t + \theta v_t,$$

puts a nonvanishing weight on each date's innovation. This weight does not disappear as $s \to \infty$, because each period essentially provides a new observation on the variable ξ_t and the realization of ξ_t has permanent consequences for the level of the series. From this perspective, a time series satisfying [15.5.7] could be described as a unit root process with non-Gaussian innovations.

Lam (1990) estimated a model closely related to [15.5.7] where shifts in the slope of the trend line were assumed to follow a Markov chain and where U.S. real GNP was allowed to follow a stationary third-order autoregression around this trend. Results of his maximum likelihood estimation are reported in Figure 15.4. These findings are very interesting for the question of the long-run consequences of economic recessions. According to this specification, events that permanently changed the level of GNP coincided with the recessions of 1957, 1973, and 1980.

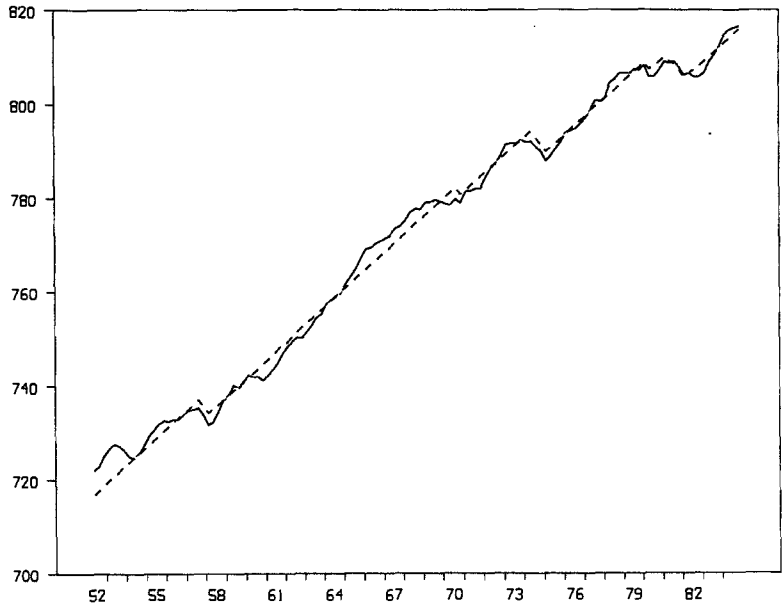


FIGURE 15.4 Discrete trend shifts estimated for U.S. real GNP, 1952–84 (Lam, 1990).

APPENDIX 15.A. *Derivation of Selected Equations for Chapter 15*

■ **Derivation of Equation [15.5.5].** Write [15.5.4] as

$$\begin{aligned} h_j &\equiv (1/j!)(d+j-1)(d+j-2)(d+j-3)\cdots(d+1)(d) \\ &= \left[\frac{d+j-1}{j}\right]\left[\frac{d+j-2}{j-1}\right]\left[\frac{d+j-3}{j-2}\right]\cdots\left[\frac{d+1}{2}\right]\left[\frac{d}{1}\right] \\ &= \left[\frac{j+d-1}{j}\right]\left[\frac{j-1+d-1}{j-1}\right]\left[\frac{j-2+d-1}{j-2}\right]\times\cdots \\ &\quad\times\left[\frac{j-(j-2)+d-1}{j-(j-2)}\right]\left[\frac{j-(j-1)+d-1}{j-(j-1)}\right] \qquad [15.A.1] \\ &= \left[1+\frac{d-1}{j}\right]\left[1+\frac{d-1}{j-1}\right]\left[1+\frac{d-1}{j-2}\right]\times\cdots \\ &\quad\times\left[1+\frac{d-1}{j-(j-2)}\right]\left[1+\frac{d-1}{j-(j-1)}\right]. \end{aligned}$$

For large j, we have the approximation

$$\left[1+\frac{d-1}{j}\right] \cong \left[1+\frac{1}{j}\right]^{d-1}. \qquad [15.A.2]$$

To justify this formally, consider the function $g(x) \equiv (1+x)^{d-1}$. Taylor's theorem states that

$$\begin{aligned} (1+x)^{d-1} &= g(0) + \left.\frac{\partial g}{\partial x}\right|_{x=0}\cdot x + \frac{1}{2}\left.\frac{\partial^2 g}{\partial x^2}\right|_{x=\delta}\cdot x^2 \qquad [15.A.3] \\ &= 1 + (d-1)x + \frac{1}{2}(d-1)(d-2)(1+\delta)^{d-3}x^2 \end{aligned}$$

for some δ between zero and x. For $x > -1$ and $d < 1$, equation [15.A.3] implies that

$$(1 + x)^{d-1} \geq 1 + (d - 1)x.$$

Letting $x = 1/j$ gives

$$1 + \frac{d-1}{j} \leq \left[1 + \frac{1}{j}\right]^{d-1} = \left[\frac{j+1}{j}\right]^{d-1} \qquad [15.A.4]$$

for all $j > 0$ and $d < 1$, with the approximation [15.A.2] improving as $j \to \infty$. Substituting [15.A.4] into [15.A.1] implies that

$$h_j \geq \left[\frac{j+1}{j}\right]^{d-1}\left[\frac{j}{j-1}\right]^{d-1}\left[\frac{j-1}{j-2}\right]^{d-1} \cdots \left[\frac{3}{2}\right]^{d-1}\left[\frac{2}{1}\right]^{d-1} = (j+1)^{d-1}. \qquad [15.A.5]$$

■

Chapter 15 References

Blough, Stephen R. 1992a. "The Relationship between Power and Level for Generic Unit Root Tests in Finite Samples." *Journal of Applied Econometrics* 7:295–308.

———. 1992b. "Near Observational Equivalence of Unit Root and Stationary Processes: Theory and Implications." Johns Hopkins University. Mimeo.

Box, G. E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, rev. ed. San Francisco: Holden-Day.

Campbell, John Y., and N. Gregory Mankiw. 1987a. "Permanent and Transitory Components in Macroeconomic Fluctuations." *American Economic Review Papers and Proceedings* 77:111–17.

——— and ———. 1987b. "Are Output Fluctuations Transitory?" *Quarterly Journal of Economics* 102:857–80.

Christiano, Lawrence J., and Martin Eichenbaum. 1990. "Unit Roots in Real GNP: Do We Know and Do We Care?" in Allan H. Meltzer, ed., *Unit Roots, Investment Measures, and Other Essays*, 7–61. Carnegie-Rochester Conference Series on Public Policy, Vol. 32. Amsterdam: North-Holland.

——— and Lars Ljungqvist. 1988. "Money Does Granger-Cause Output in the Bivariate Money-Output Relation." *Journal of Monetary Economics* 22:217–35.

Clark, Peter K. 1987. "The Cyclical Component of U.S. Economic Activity." *Quarterly Journal of Economics* 102:797–814.

Cochrane, John H. 1988. "How Big Is the Random Walk in GNP?" *Journal of Political Economy* 96:893–920.

———. 1991. "A Critique of the Application of Unit Root Tests." *Journal of Economic Dynamics and Control* 15:275–84.

Diebold, Francis X., and Glenn D. Rudebusch. 1989. "Long Memory and Persistence in Aggregate Output." *Journal of Monetary Economics* 24:189–209.

Durlauf, Steven N. 1989. "Output Persistence, Economic Structure, and Choice of Stabilization Policy." *Brookings Papers on Economic Activity* 2:1989, 69–116.

Friedman, Milton. 1957. *A Theory of the Consumption Function*. Princeton, N.J.: Princeton University Press.

Gagnon, Joseph E. 1988. "Short-Run Models and Long-Run Forecasts: A Note on the Permanence of Output Fluctuations." *Quarterly Journal of Economics* 103:415–24.

Geweke, John, and Susan Porter-Hudak. 1983. "The Estimation and Application of Long Memory Time Series Models." *Journal of Time Series Analysis* 4:221–38.

Granger, C. W. J. 1980. "Long Memory Relationships and the Aggregation of Dynamic Models." *Journal of Econometrics* 14:227–38.

——— and Roselyne Joyeux. 1980. "An Introduction to Long-Memory Time Series Models and Fractional Differencing." *Journal of Time Series Analysis* 1:15–29.

Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica* 57:357–84.

Hosking, J. R. M. 1981. "Fractional Differencing." *Biometrika* 68:165–76.

Lam, Pok-sang. 1990. "The Hamilton Model with a General Autoregressive Component: Estimation and Comparison with Other Models of Economic Time Series." *Journal of Monetary Economics* 26:409–32.

Lo, Andrew W. 1991. "Long-Term Memory in Stock Market Prices." *Econometrica* 59:1279–1313.

Muth, John F. 1960. "Optimal Properties of Exponentially Weighted Forecasts." *Journal of the American Statistical Association* 55:299–306.

Nelson, Charles R., and Charles I. Plosser. 1982. "Trends and Random Walks in Macroeconomic Time Series: Some Evidence and Implications." *Journal of Monetary Economics* 10:139–62.

Perron, Pierre. 1989. "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis." *Econometrica* 57:1361–1401.

Rappoport, Peter, and Lucrezia Reichlin. 1989. "Segmented Trends and Nonstationary Time Series." *Economic Journal* supplement 99:168–77.

Sims, Christopher A. 1989. "Modeling Trends." Yale University. Mimeo.

Sowell, Fallaw. 1992. "Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models." *Journal of Econometrics* 53:165–88.

Stock, James H. 1990. "'Unit Roots in Real GNP: Do We Know and Do We Care?' A Comment," in Allan H. Meltzer, ed., *Unit Roots, Investment Measures, and Other Essays*. 63–82. Carnegie-Rochester Conference Series on Public Policy, Vol. 32. Amsterdam: North-Holland.

——— and Mark W. Watson. 1988. "Variable Trends in Economic Time Series." *Journal of Economic Perspectives* vol. 2, no. 3, 147–74.

——— and ———. 1989. "Interpreting the Evidence on Money-Income Causality." *Journal of Econometrics* 40:161–81.

Watson, Mark W. 1986. "Univariate Detrending Methods with Stochastic Trends." *Journal of Monetary Economics* 18:49–75.

16 Processes with Deterministic Time Trends

The coefficients of regression models involving unit roots or deterministic time trends are typically estimated by ordinary least squares. However, the asymptotic distributions of the coefficient estimates cannot be calculated in the same way as are those for regression models involving stationary variables. Among other difficulties, the estimates of different parameters will in general have different asymptotic rates of convergence. This chapter introduces the idea of different rates of convergence and develops a general approach to obtaining asymptotic distributions suggested by Sims, Stock, and Watson (1990).[1] This chapter deals exclusively with processes involving deterministic time trends but no unit roots. One of the results for such processes will be that the usual *OLS* t and F statistics, calculated in the usual way, have the same asymptotic distributions as they do for stationary regressions. Although the limiting distributions are standard, the techniques used to verify these limiting distributions are different from those used in Chapter 8. These techniques will also be used to develop the asymptotic distributions for processes including unit roots in Chapters 17 and 18.

This chapter begins with the simplest example of i.i.d. innovations around a deterministic time trend. Section 16.1 derives the asymptotic distributions of the coefficient estimates for this model and illustrates a rescaling of variables that is necessary to accommodate different asymptotic rates of convergence. Section 16.2 shows that despite the different asymptotic rates of convergence, the standard *OLS* t and F statistics have the usual limiting distributions for this model. Section 16.3 develops analogous results for a covariance-stationary autoregression around a deterministic time trend. That section also introduces the Sims, Stock, and Watson technique of transforming the regression model into a canonical form for which the asymptotic distribution is simpler to describe.

16.1. *Asymptotic Distribution of* OLS *Estimates of the Simple Time Trend Model*

This section considers *OLS* estimation of the parameters of a simple time trend,

$$y_t = \alpha + \delta t + \varepsilon_t, \qquad [16.1.1]$$

for ε_t a white noise process. If $\varepsilon_t \sim N(0, \sigma^2)$, then the model [16.1.1] satisfies the classical regression assumptions[2] and the standard *OLS* t or F statistics in equations

[1] A simpler version of this theme appeared in the analysis of a univariate process with unit roots by Fuller (1976).

[2] See Assumption 8.1 in Chapter 8.

[8.1.26] and [8.1.32] would have exact small-sample t or F distributions. On the other hand, if ε_t is non-Gaussian, then a slightly different technique for finding the asymptotic distributions of the OLS estimates of α and δ would have to be used from that employed for stationary regressions in Chapter 8. This chapter introduces this technique, which will prove useful not only for studying time trends but also for analyzing estimators for a variety of nonstationary processes in Chapters 17 and 18.[3]

Recall the approach used to find asymptotic distributions for regressions with stationary explanatory variables in Chapter 8. Write [16.1.1] in the form of the standard regression model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \qquad [16.1.2]$$

where

$$\underset{(1\times 2)}{\mathbf{x}_t'} \equiv [1 \quad t] \qquad [16.1.3]$$

$$\underset{(2\times 1)}{\boldsymbol{\beta}} \equiv \begin{bmatrix} \alpha \\ \delta \end{bmatrix}. \qquad [16.1.4]$$

Let $\mathbf{b}_T$ denote the OLS estimate of $\boldsymbol{\beta}$ based on a sample of size T:

$$\mathbf{b}_T \equiv \begin{bmatrix} \hat{\alpha}_T \\ \hat{\delta}_T \end{bmatrix} = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t y_t\right]. \qquad [16.1.5]$$

Recall from equation [8.2.3] that the deviation of the OLS estimate from the true value can be expressed as

$$(\mathbf{b}_T - \boldsymbol{\beta}) = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\varepsilon_t\right]. \qquad [16.1.6]$$

To find the limiting distribution for a regression with stationary explanatory variables, the approach in Chapter 8 was to multiply [16.1.6] by $\sqrt{T}$, resulting in

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) = \left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T} \mathbf{x}_t\varepsilon_t\right]. \qquad [16.1.7]$$

The usual assumption was that $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ converged in probability to a nonsingular matrix $\mathbf{Q}$ while $(1/\sqrt{T})\Sigma_{t=1}^{T}\mathbf{x}_t\varepsilon_t$ converged in distribution to a $N(\mathbf{0}, \sigma^2\mathbf{Q})$ random variable, implying that $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$.

To see why this same argument cannot be used for a deterministic time trend, note that for $\mathbf{x}_t$ and $\boldsymbol{\beta}$ given in equations [16.1.3] and [16.1.4], expression [16.1.6] would be

$$\begin{bmatrix} \hat{\alpha}_T - \alpha \\ \hat{\delta}_T - \delta \end{bmatrix} = \begin{bmatrix} \Sigma 1 & \Sigma t \\ \Sigma t & \Sigma t^2 \end{bmatrix}^{-1}\begin{bmatrix} \Sigma\varepsilon_t \\ \Sigma t\varepsilon_t \end{bmatrix}, \qquad [16.1.8]$$

[3]The general approach in these chapters follows Sims, Stock, and Watson (1990).

where Σ denotes summation for $t = 1$ through T. It is straightforward to show by induction that[4]

$$\sum_{t=1}^{T} t = T(T + 1)/2 \qquad [16.1.9]$$

$$\sum_{t=1}^{T} t^2 = T(T + 1)(2T + 1)/6. \qquad [16.1.10]$$

Thus, the leading term in $\Sigma_{t=1}^{T} t$ is $T^2/2$; that is,

$$(1/T^2) \sum_{t=1}^{T} t = (1/T^2)[(T^2/2) + (T/2)] = 1/2 + 1/(2T) \to 1/2. \qquad [16.1.11]$$

Similarly, the leading term in $\Sigma_{t=1}^{T} t^2$ is $T^3/3$:

$$\begin{aligned}(1/T^3) \sum_{t=1}^{T} t^2 &= (1/T^3)[(2T^3/6) + (3T^2/6) + T/6] \\ &= 1/3 + 1/(2T) + 1/(6T^2) \\ &\to 1/3.\end{aligned} \qquad [16.1.12]$$

For future reference, we note here the general pattern—the leading term in $\Sigma_{t=1}^{T} t^\nu$ is $T^{\nu+1}/(\nu + 1)$:

$$(1/T^{\nu+1}) \sum_{t=1}^{T} t^\nu \to 1/(\nu + 1). \qquad [16.1.13]$$

To verify [16.1.13], note that

$$(1/T^{\nu+1}) \sum_{t=1}^{T} t^\nu = (1/T) \sum_{t=1}^{T} (t/T)^\nu. \qquad [16.1.14]$$

The right side of [16.1.14] can be viewed as an approximation to the area under the curve

$$f(r) = r^\nu$$

for r between zero and unity. To see this, notice that $(1/T)\cdot(t/T)^\nu$ represents the area of a rectangle with width $(1/T)$ and height r^ν evaluated at $r = t/T$ (see Figure 16.1). Thus, [16.1.14] is the sum of the area of these rectangles evaluated

[4]Clearly, [16.1.9] and [16.1.10] hold for $T = 1$. Given that [16.1.9] holds for T,

$$\sum_{t=1}^{T+1} t = \sum_{t=1}^{T} t + (T + 1) = T(T + 1)/2 + (T + 1) = (T + 1)[(T/2) + 1] = (T + 1)(T + 2)/2,$$

establishing that [16.1.9] holds for $T + 1$. Similarly, given that [16.1.10] holds for T,

$$\begin{aligned}\sum_{t=1}^{T+1} t^2 &= T(T + 1)(2T + 1)/6 + (T + 1)^2 \\ &= (T + 1)\{[T(2T + 1)/6] + (T + 1)\} \\ &= (T + 1)(2T^2 + 7T + 6)/6 \\ &= (T + 1)(T + 2)[2(T + 1) + 1]/6,\end{aligned}$$

establishing that [16.1.10] holds for $T + 1$.

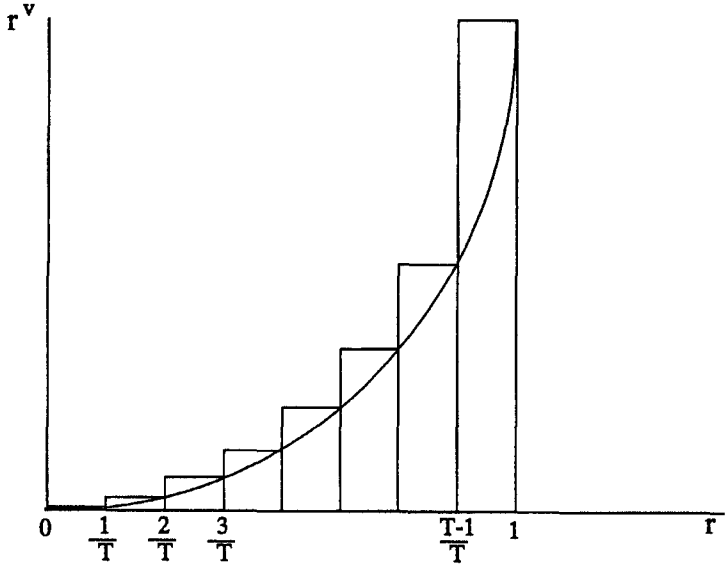


FIGURE 16.1 Demonstration that $(1/T)\,\Sigma_{t=1}^{T}\,(t/T)^v \to \int_0^1 r^v\,dr = 1/(v+1)$.

at $r = 1/T, 2/T, \ldots, 1$. As $T \to \infty$, this sum converges to the area under the curve $f(r)$:

$$(1/T)\sum_{t=1}^{T}(t/T)^v \to \int_0^1 r^v\,dr = r^{v+1}/(v+1)\big|_{r=0}^{1} = 1/(v+1). \qquad [16.1.15]$$

For $\mathbf{x}_t$ given in [16.1.3], results [16.1.9] and [16.1.10] imply that

$$\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = \begin{bmatrix}\Sigma 1 & \Sigma t\\ \Sigma t & \Sigma t^2\end{bmatrix} = \begin{bmatrix}T & T(T+1)/2\\ T(T+1)/2 & T(T+1)(2T+1)/6\end{bmatrix}. \qquad [16.1.16]$$

In contrast to the usual result for stationary regressions, for the matrix in [16.1.16], $(1/T)\,\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ diverges. To obtain a convergent matrix, [16.1.16] would have to be divided by T^3 rather than T:

$$T^{-3}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \to \begin{bmatrix}0 & 0\\ 0 & \frac{1}{3}\end{bmatrix}.$$

Unfortunately, this limiting matrix cannot be inverted, as $(1/T)\,\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ can be in the usual case. Hence, a different approach from that in the stationary case will be needed to calculate the asymptotic distribution of $\mathbf{b}_T$.

It turns out that the *OLS* estimates $\hat{\alpha}_T$ and $\hat{\delta}_T$ have different asymptotic rates of convergence. To arrive at nondegenerate limiting distributions, $\hat{\alpha}_T$ is multiplied by $\sqrt{T}$, whereas $\hat{\delta}_T$ must be multiplied by $T^{3/2}$! We can think of this adjustment as premultiplying [16.1.6] or [16.1.8] by the matrix

$$\mathbf{Y}_T \equiv \begin{bmatrix}\sqrt{T} & 0\\ 0 & T^{3/2}\end{bmatrix}, \qquad [16.1.17]$$

resulting in

$$\begin{bmatrix} \sqrt{T}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\delta}_T - \delta) \end{bmatrix} = \mathbf{Y}_T \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t \right]$$

$$= \mathbf{Y}_T \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \mathbf{Y}_T \mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t \right] \qquad [16.1.18]$$

$$= \left\{ \mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{Y}_T^{-1} \right\}^{-1} \left\{ \mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t \right] \right\}.$$

Consider the first term in the last expression of [16.1.18]. Substituting from [16.1.17] and [16.1.16],

$$\left\{ \mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{Y}_T^{-1} \right\} = \left\{ \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} \Sigma 1 & \Sigma t \\ \Sigma t & \Sigma t^2 \end{bmatrix} \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \right\}$$

$$= \begin{bmatrix} T^{-1}\Sigma 1 & T^{-2}\Sigma t \\ T^{-2}\Sigma t & T^{-3}\Sigma t^2 \end{bmatrix}.$$

Thus, it follows from [16.1.11] and [16.1.12] that

$$\left\{ \mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \mathbf{Y}_T^{-1} \right\} \to \mathbf{Q}, \qquad [16.1.19]$$

where

$$\mathbf{Q} \equiv \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}. \qquad [16.1.20]$$

Turning next to the second term in [16.1.18],

$$\mathbf{Y}_T^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t \right] = \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} \Sigma \varepsilon_t \\ \Sigma t \varepsilon_t \end{bmatrix} = \begin{bmatrix} (1/\sqrt{T})\Sigma \varepsilon_t \\ (1/\sqrt{T})\Sigma (t/T)\varepsilon_t \end{bmatrix}. \qquad [16.1.21]$$

Under standard assumptions about ε_t, this vector will be asymptotically Gaussian. For example, suppose that ε_t is i.i.d. with mean zero, variance σ^2, and finite fourth moment. Then the first element of the vector in [16.1.21] satisfies

$$(1/\sqrt{T}) \sum_{t=1}^{T} \varepsilon_t \xrightarrow{L} N(0, \sigma^2),$$

by the central limit theorem.

For the second element of the vector in [16.1.21], observe that $\{(t/T)\varepsilon_t\}$ is a martingale difference sequence that satisfies the conditions of Proposition 7.8. Specifically, its variance is

$$\sigma_t^2 = E[(t/T)\varepsilon_t]^2 = \sigma^2 \cdot (t^2/T^2),$$

where

$$(1/T) \sum_{t=1}^{T} \sigma_t^2 = \sigma^2 (1/T^3) \sum_{t=1}^{T} t^2 \to \sigma^2/3.$$

Furthermore, $(1/T)\sum_{t=1}^{T}[(t/T)\varepsilon_t]^2 \xrightarrow{p} \sigma^2/3$. To verify this last claim, notice that

$$\begin{aligned} &E\left((1/T)\sum_{t=1}^{T}[(t/T)\varepsilon_t]^2 - (1/T)\sum_{t=1}^{T}\sigma_t^2\right)^2 \\ &= E\left((1/T)\sum_{t=1}^{T}[(t/T)\varepsilon_t]^2 - (1/T)\sum_{t=1}^{T}(t/T)^2\sigma^2\right)^2 \\ &= E\left((1/T)\sum_{t=1}^{T}(t/T)^2(\varepsilon_t^2 - \sigma^2)\right)^2 \\ &= (1/T)^2\sum_{t=1}^{T}(t/T)^4E(\varepsilon_t^2 - \sigma^2)^2. \end{aligned} \qquad [16.1.22]$$

But from [16.1.13], T times the magnitude in [16.1.22] converges to

$$(1/T)\sum_{t=1}^{T}(t/T)^4E(\varepsilon_t^2 - \sigma^2)^2 \to (1/5)\cdot E(\varepsilon_t^2 - \sigma^2)^2,$$

meaning that [16.1.22] itself converges to zero:

$$(1/T)\sum_{t=1}^{T}[(t/T)\varepsilon_t]^2 - (1/T)\sum_{t=1}^{T}\sigma_t^2 \xrightarrow{m.s.} 0.$$

But this implies that

$$(1/T)\sum_{t=1}^{T}[(t/T)\varepsilon_t]^2 \xrightarrow{p} \sigma^2/3,$$

as claimed. Hence, from Proposition 7.8, $(1/\sqrt{T})\sum_{t=1}^{T}(t/T)\varepsilon_t$ satisfies the central limit theorem:

$$(1/\sqrt{T})\sum_{t=1}^{T}(t/T)\varepsilon_t \xrightarrow{L} N(0, \sigma^2/3).$$

Finally, consider the joint distribution of the two elements in the (2×1) vector described by [16.1.21]. Any linear combination of these elements takes the form

$$(1/\sqrt{T})\sum_{t=1}^{T}[\lambda_1 + \lambda_2(t/T)]\varepsilon_t.$$

Then $[\lambda_1 + \lambda_2(t/T)]\varepsilon_t$ is also a martingale difference sequence with positive variance[5] given by $\sigma^2[\lambda_1^2 + 2\lambda_1\lambda_2(t/T) + \lambda_2^2(t/T)^2]$ satisfying

$$\begin{aligned} (1/T)\sum_{t=1}^{T}\sigma^2[\lambda_1^2 + 2\lambda_1\lambda_2(t/T) + \lambda_2^2(t/T)^2] &\to \sigma^2[\lambda_1^2 + 2\lambda_1\lambda_2(\tfrac{1}{2}) + \lambda_2^2(\tfrac{1}{3})] \\ &= \sigma^2\boldsymbol{\lambda}'\mathbf{Q}\boldsymbol{\lambda} \end{aligned}$$

for $\boldsymbol{\lambda} \equiv (\lambda_1, \lambda_2)'$ and $\mathbf{Q}$ the matrix in [16.1.20]. Furthermore,

$$(1/T)\sum_{t=1}^{T}[\lambda_1 + \lambda_2(t/T)]^2\varepsilon_t^2 \xrightarrow{p} \sigma^2\boldsymbol{\lambda}'\mathbf{Q}\boldsymbol{\lambda}; \qquad [16.1.23]$$

see Exercise 16.1. Thus any linear combination of the two elements in the vector in [16.1.21] is asymptotically Gaussian, implying a limiting bivariate Gaussian dis-

[5]More accurately, a given nonzero λ_1 and λ_2 will produce a zero variance for $[\lambda_1 + \lambda_2(t/T)]\varepsilon_t$ for at most a single value of t, which does not affect the validity of the asymptotic claim.

tribution:

$$\begin{bmatrix} (1/\sqrt{T})\Sigma\varepsilon_t \\ (1/\sqrt{T})\Sigma(t/T)\varepsilon_t \end{bmatrix} \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}). \qquad [16.1.24]$$

From [16.1.19] and [16.1.24], the asymptotic distribution of [16.1.18] can be calculated as in Example 7.5 of Chapter 7:

$$\begin{bmatrix} \sqrt{T}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\delta}_T - \delta) \end{bmatrix} \xrightarrow{L} N(\mathbf{0}, [\mathbf{Q}^{-1}\cdot\sigma^2\mathbf{Q}\cdot\mathbf{Q}^{-1}]) = N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}). \qquad [16.1.25]$$

These results can be summarized as follows.

Proposition 16.1: Let y_t be generated according to the simple deterministic time trend [16.1.1] where ε_t is i.i.d. with $E(\varepsilon_t^2) = \sigma^2$ and $E(\varepsilon_t^4) < \infty$. Then

$$\begin{bmatrix} \sqrt{T}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\delta}_T - \delta) \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}^{-1}\right). \qquad [16.1.26]$$

Note that the resulting estimate of the coefficient on the time trend ($\hat{\delta}_T$) is *superconsistent*—not only does $\hat{\delta}_T \xrightarrow{p} \delta$, but even when multiplied by T, we still have

$$T(\hat{\delta}_T - \delta) \xrightarrow{p} 0; \qquad [16.1.27]$$

see Exercise 16.2.

Different rates of convergence are sometimes described in terms of *order in probability*. A sequence of random variables $\{X_T\}_{T=1}^{\infty}$ is said to be $O_p(T^{-1/2})$ if for every $\varepsilon > 0$, there exists an $M > 0$ such that

$$P\{|X_T| > M/\sqrt{T}\} < \varepsilon \qquad [16.1.28]$$

for all T; in other words, the random variable $\sqrt{T}\cdot X_T$ is almost certain to fall within $\pm M$ for any T. Most of the estimators encountered for stationary time series are $O_p(T^{-1/2})$. For example, suppose that X_T represents the mean of a sample of size T,

$$X_T = (1/T)\sum_{t=1}^{T} y_t,$$

where $\{y_t\}$ is i.i.d. with mean zero and variance σ^2. Then the variance of X_T is σ^2/T. But Chebyshev's inequality implies that

$$P\{|X_T| > M/\sqrt{T}\} \le \frac{\sigma^2/T}{M^2/T} = (\sigma/M)^2$$

for any M. By choosing M so that $(\sigma/M)^2 < \varepsilon$, condition [16.1.28] is guaranteed. Since the standard deviation of the estimator is $\sigma/\sqrt{T}$, by choosing M to be a suitable multiple of σ, the band $X_T \pm M/\sqrt{T}$ can include as much of the density as desired.

As another example, the estimator $\hat{\alpha}_T$ in [16.1.26] would also be said to be $O_p(T^{-1/2})$. Since $\sqrt{T}$ times $(\hat{\alpha}_T - \alpha)$ is asymptotically Gaussian, there exists a band $\pm M/\sqrt{T}$ around $\hat{\alpha}_T$ that contains as much of the probability distribution as desired.

In general, a sequence of random variables $\{X_T\}_{T=1}^{\infty}$ is said to be $O_p(T^{-k})$ if for every $\varepsilon > 0$ there exists an $M > 0$ such that

$$P\{|X_T| > M/(T^k)\} < \varepsilon. \qquad [16.1.29]$$

Thus, for example, the estimator $\hat{\delta}_T$ in [16.1.26] is $O_p(T^{-3/2})$, since there exists a band $\pm M$ around $T^{3/2}(\hat{\delta}_T - \delta)$ that contains as much of the probability distribution as desired.

16.2. *Hypothesis Testing for the Simple Time Trend Model*

If the innovations ε_t for the simple time trend [16.1.1] are Gaussian, then the *OLS* estimates $\hat{\alpha}_T$ and $\hat{\delta}_T$ are Gaussian and the usual *OLS* t and F tests have exact small-sample t and F distributions for all sample sizes T. Thus, despite the fact that $\hat{\alpha}_T$ and $\hat{\delta}_T$ have different asymptotic rates of convergence, the standard errors $\hat{\sigma}_{\hat{\alpha}_T}$ and $\hat{\sigma}_{\hat{\delta}_T}$ evidently have offsetting asymptotic behavior so that the statistics such as $(\hat{\delta}_T - \delta_0)/\hat{\sigma}_{\hat{\delta}_T}$ are asymptotically $N(0, 1)$ when the innovations are Gaussian. We might thus conjecture that the usual t and F tests are asymptotically valid for non-Gaussian innovations as well. This conjecture is indeed correct, as we now verify.

First consider the *OLS* t test of the null hypothesis $\alpha = \alpha_0$, which can be written as

$$t_T = \frac{\hat{\alpha}_T - \alpha_0}{\left\{s_T^2[1 \quad 0](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}1\\0\end{bmatrix}\right\}^{1/2}}. \qquad [16.2.1]$$

Here s_T^2 denotes the *OLS* estimate of σ^2:

$$s_T^2 = [1/(T-2)]\sum_{t=1}^{T}(y_t - \hat{\alpha}_T - \hat{\delta}_T t)^2; \qquad [16.2.2]$$

and $(\mathbf{X}_T'\mathbf{X}_T) = \Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ denotes the matrix in equation [16.1.16]. The numerator and denominator of [16.2.1] can further be multiplied by $\sqrt{T}$, resulting in

$$t_T = \frac{\sqrt{T}(\hat{\alpha}_T - \alpha_0)}{\left\{s_T^2[\sqrt{T} \quad 0](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}\sqrt{T}\\0\end{bmatrix}\right\}^{1/2}}. \qquad [16.2.3]$$

Note further from [16.1.17] that

$$[\sqrt{T} \quad 0] = [1 \quad 0]\mathbf{Y}_T. \qquad [16.2.4]$$

Substituting [16.2.4] into [16.2.3],

$$t_T = \frac{\sqrt{T}(\hat{\alpha}_T - \alpha_0)}{\left\{s_T^2[1 \quad 0]\mathbf{Y}_T(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T\begin{bmatrix}1\\0\end{bmatrix}\right\}^{1/2}}. \qquad [16.2.5]$$

But recall from [16.1.19] that

$$\mathbf{Y}_T(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T = [\mathbf{Y}_T^{-1}(\mathbf{X}_T'\mathbf{X}_T)\mathbf{Y}_T^{-1}]^{-1} \to \mathbf{Q}^{-1}. \qquad [16.2.6]$$

It is straightforward to show that $s_T^2 \xrightarrow{p} \sigma^2$. Recall further that $\sqrt{T}(\hat{\alpha}_T - \alpha_0) \xrightarrow{L} N(0, \sigma^2 q^{11})$ for q^{11} the (1, 1) element of $\mathbf{Q}^{-1}$. Hence, from [16.2.5],

$$t_T \xrightarrow{p} \frac{\sqrt{T}(\hat{\alpha}_T - \alpha_0)}{\left\{\sigma^2[1 \quad 0]\mathbf{Q}^{-1}\begin{bmatrix}1\\0\end{bmatrix}\right\}^{1/2}} = \frac{\sqrt{T}(\hat{\alpha}_T - \alpha_0)}{\sigma\sqrt{q^{11}}}. \qquad [16.2.7]$$

But this is an asymptotically Gaussian variable divided by the square root of its variance, and so asymptotically it has a $N(0, 1)$ distribution. Thus, the usual *OLS* t test of $\alpha = \alpha_0$ will give an asymptotically valid inference.

Similarly, consider the usual *OLS* t test of $\delta = \delta_0$:

$$t_T = \frac{\hat{\delta}_T - \delta_0}{\left\{s_T^2[0 \quad 1](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}0\\1\end{bmatrix}\right\}^{1/2}}.$$

Multiplying numerator and denominator by $T^{3/2}$,

$$t_T = \frac{T^{3/2}(\hat{\delta}_T - \delta_0)}{\left\{s_T^2[0 \quad T^{3/2}](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}0\\T^{3/2}\end{bmatrix}\right\}^{1/2}}$$

$$= \frac{T^{3/2}(\hat{\delta}_T - \delta_0)}{\left\{s_T^2[0 \quad 1]\mathbf{Y}_T(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T\begin{bmatrix}0\\1\end{bmatrix}\right\}^{1/2}}$$

$$\xrightarrow{p} \frac{T^{3/2}(\hat{\delta}_T - \delta_0)}{\sigma\sqrt{q^{22}}},$$

which again is asymptotically a $N(0, 1)$ variable. Thus, although $\hat{\alpha}_T$ and $\hat{\delta}_T$ converge at different rates, the corresponding standard errors $\hat{\sigma}_{\hat{\alpha}_T}$ and $\hat{\sigma}_{\hat{\delta}_T}$ also incorporate different orders of T, with the result that the usual *OLS* t tests are asymptotically valid.

It is interesting also to consider a test of a single hypothesis involving both α and δ,

$$H_0\colon r_1\alpha + r_2\delta = r,$$

where r_1, r_2, and r are parameters that describe the hypothesis. A t test of H_0 can be obtained from the square root of the *OLS* F test (expression [8.1.32]):[6]

$$t_T = \frac{(r_1\hat{\alpha}_T + r_2\hat{\delta}_T - r)}{\left\{s_T^2[r_1 \quad r_2](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}r_1\\r_2\end{bmatrix}\right\}^{1/2}}.$$

In this case we multiply numerator and denominator by $\sqrt{T}$, the slower rate of convergence among the two estimators $\hat{\alpha}_T$ and $\hat{\delta}_T$:

$$t_T = \frac{\sqrt{T}(r_1\hat{\alpha}_T + r_2\hat{\delta}_T - r)}{\left\{s_T^2\sqrt{T}[r_1 \quad r_2](\mathbf{X}_T'\mathbf{X}_T)^{-1}\begin{bmatrix}r_1\\r_2\end{bmatrix}\sqrt{T}\right\}^{1/2}}$$

$$= \frac{\sqrt{T}(r_1\hat{\alpha}_T + r_2\hat{\delta}_T - r)}{\left\{s_T^2\sqrt{T}[r_1 \quad r_2]\mathbf{Y}_T^{-1}\mathbf{Y}_T(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T\mathbf{Y}_T^{-1}\begin{bmatrix}r_1\\r_2\end{bmatrix}\sqrt{T}\right\}^{1/2}}$$

$$= \frac{\sqrt{T}(r_1\hat{\alpha}_T + r_2\hat{\delta}_T - r)}{\{s_T^2\mathbf{r}_T'[\mathbf{Y}_T(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T]\mathbf{r}_T\}^{1/2}},$$

where

$$\mathbf{r}_T \equiv \mathbf{Y}_T^{-1}\begin{bmatrix}r_1\\r_2\end{bmatrix}\sqrt{T} = \begin{bmatrix}r_1\\r_2/T\end{bmatrix} \to \begin{bmatrix}r_1\\0\end{bmatrix}. \qquad [16.2.8]$$

Similarly, recall from [16.1.27] that $\hat{\delta}_T$ is superconsistent, implying that

$$\sqrt{T}(r_1\hat{\alpha}_T + r_2\hat{\delta}_T - r) \xrightarrow{p} \sqrt{T}(r_1\hat{\alpha}_T + r_2\delta - r), \qquad [16.2.9]$$

[6]With a single linear restriction as here, $m = 1$ and expression [8.1.32] describes an $F(1, T - k)$ variable when the innovations are Gaussian. But an $F(1, T - k)$ variable is the square of a $t(T - k)$ variable. The test is described here in terms of a t test rather than an F test in order to facilitate comparison with the earlier results in this section.

where δ is the true population value for the time trend parameter. Again applying [16.2.6], it follows that

$$t_T \xrightarrow{p} \frac{\sqrt{T}(r_1\hat{\alpha}_T + r_2\delta - r)}{\left\{\sigma^2[r_1 \quad 0]\mathbf{Q}^{-1}\begin{bmatrix} r_1 \\ 0 \end{bmatrix}\right\}^{1/2}} = \frac{\sqrt{T}(r_1\hat{\alpha}_T + r_2\delta - r)}{\{r_1^2\sigma^2 q^{11}\}^{1/2}}. \qquad [16.2.10]$$

But notice that

$$\begin{aligned}\sqrt{T}(r_1\hat{\alpha}_T + r_2\delta - r) &= \sqrt{T}[r_1(\hat{\alpha}_T - \alpha) + r_1\alpha + r_2\delta - r] \\ &= \sqrt{T}[r_1(\hat{\alpha}_T - \alpha)]\end{aligned}$$

under the null hypothesis. Hence, under the null,

$$t_T \xrightarrow{p} \frac{\sqrt{T}[r_1(\hat{\alpha}_T - \alpha)]}{\{r_1^2\sigma^2 q^{11}\}^{1/2}} = \frac{\sqrt{T}(\hat{\alpha}_T - \alpha)}{\{\sigma^2 q^{11}\}^{1/2}},$$

which asymptotically has a $N(0, 1)$ distribution. Thus, again, the usual *OLS* t test of H_0 is valid asymptotically.

This last example illustrates the following general principle: A test involving a single restriction across parameters with different rates of convergence is dominated asymptotically by the parameters with the slowest rates of convergence. This means that a test involving both α and δ that employs the estimated value of δ would have the same asymptotic properties under the null as a test that employs the true value of δ.

Finally, consider a joint test of separate hypotheses about α and δ,

$$H_0: \begin{bmatrix} \alpha \\ \delta \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \delta_0 \end{bmatrix},$$

or, in vector form,

$$\boldsymbol{\beta} = \boldsymbol{\beta}_0.$$

The Wald form of the *OLS* χ^2 test of H_0 is found from [8.2.23] by taking $\mathbf{R} = \mathbf{I}_2$:

$$\begin{aligned}\chi_T^2 &= (\mathbf{b}_T - \boldsymbol{\beta}_0)'[s_T^2(\mathbf{X}_T'\mathbf{X}_T)^{-1}]^{-1}(\mathbf{b}_T - \boldsymbol{\beta}_0) \\ &= (\mathbf{b}_T - \boldsymbol{\beta}_0)'\mathbf{Y}_T[\mathbf{Y}_T s_T^2(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{Y}_T]^{-1}\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}_0) \\ &\xrightarrow{p} [\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}_0)]'[\sigma^2\mathbf{Q}^{-1}]^{-1}[\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}_0)].\end{aligned}$$

Recalling [16.1.25], this is a quadratic form in a two-dimensional Gaussian vector of the sort considered in Proposition 8.1, from which

$$\chi_T^2 \xrightarrow{L} \chi^2(2).$$

Thus, again, the usual *OLS* test is asymptotically valid.

16.3. *Asymptotic Inference for an Autoregressive Process Around a Deterministic Time Trend*

The same principles can be used to study a general autoregressive process around a deterministic time trend:

$$y_t = \alpha + \delta t + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t. \qquad [16.3.1]$$

It is assumed throughout this section that ε_t is i.i.d. with mean zero, variance σ^2, and finite fourth moment, and that roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

lie outside the unit circle. Consider a sample of $T + p$ observations on y, $\{y_{-p+1}, y_{-p+2}, \ldots, y_T\}$, and let $\hat{\alpha}_T$, $\hat{\delta}_T$, $\hat{\phi}_{1,T}, \ldots, \hat{\phi}_{p,T}$ denote coefficient estimates based on ordinary least squares estimation of [16.3.1] for $t = 1, 2, \ldots, T$.

A Useful Transformation of the Regressors

By adding and subtracting $\phi_j[\alpha + \delta(t - j)]$ for $j = 1, 2, \ldots, p$ on the right side, the regression model [16.3.1] can equivalently be written as

$$\begin{aligned} y_t = {} & \alpha(1 + \phi_1 + \phi_2 + \cdots + \phi_p) + \delta(1 + \phi_1 + \phi_2 + \cdots + \phi_p)t \\ & - \delta(\phi_1 + 2\phi_2 + \cdots + p\phi_p) + \phi_1[y_{t-1} - \alpha - \delta(t - 1)] \\ & + \phi_2[y_{t-2} - \alpha - \delta(t - 2)] + \cdots \\ & + \phi_p[y_{t-p} - \alpha - \delta(t - p)] + \varepsilon_t \end{aligned} \qquad [16.3.2]$$

or

$$y_t = \alpha^* + \delta^* t + \phi_1^* y_{t-1}^* + \phi_2^* y_{t-2}^* + \cdots + \phi_p^* y_{t-p}^* + \varepsilon_t, \qquad [16.3.3]$$

where

$$\begin{aligned} \alpha^* &\equiv [\alpha(1 + \phi_1 + \phi_2 + \cdots + \phi_p) - \delta(\phi_1 + 2\phi_2 + \cdots + p\phi_p)] \\ \delta^* &\equiv \delta(1 + \phi_1 + \phi_2 + \cdots + \phi_p) \\ \phi_j^* &\equiv \phi_j \qquad \text{for } j = 1, 2, \ldots, p \end{aligned}$$

and

$$y_{t-j}^* \equiv y_{t-j} - \alpha - \delta(t - j) \qquad \text{for } j = 1, 2, \ldots, p. \qquad [16.3.4]$$

The idea of transforming the regression into a form such as [16.3.3] is due to Sims, Stock, and Watson (1990).[7] The objective is to rewrite the regressors of [16.3.1] in terms of zero-mean covariance-stationary random variables (the terms y_{t-j}^* for $j = 1, 2, \ldots, p$), a constant term, and a time trend. Transforming the regressors in this way isolates components of the *OLS* coefficient vector with different rates of convergence and provides a general technique for finding the asymptotic distribution of regressions involving nonstationary variables. A general result is that, if such a transformed equation were estimated by *OLS*, the coefficients on zero-mean covariance-stationary random variables (in this case, $\hat{\phi}_{1,T}^*$, $\hat{\phi}_{2,T}^*$, $\ldots, \hat{\phi}_{p,T}^*$) would converge at rate $\sqrt{T}$ to a Gaussian distribution. The coefficients $\hat{\alpha}_T^*$ and $\hat{\delta}_T^*$ from *OLS* estimation of [16.3.3] turn out to behave asymptotically exactly like $\hat{\alpha}_T$ and $\hat{\delta}_T$ for the simple time trend model analyzed in Section 16.1 and are asymptotically independent of the $\hat{\phi}^*$'s.

It is helpful to describe this transformation in more general notation that will also apply to more complicated models in the chapters that follow. The original regression model [16.3.1] can be written

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \qquad [16.3.5]$$

[7]A simpler version of this theme appeared in the analysis of a univariate process with unit roots by Fuller (1976).

where

$$\underset{(p+2)\times 1}{\mathbf{x}_t} \equiv \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \\ 1 \\ t \end{bmatrix} \qquad \underset{(p+2)\times 1}{\boldsymbol{\beta}} \equiv \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \\ \alpha \\ \delta \end{bmatrix} \tag{16.3.6}$$

The algebraic transformation in arriving at [16.3.3] could then be described as rewriting [16.3.5] in the form

$$y_t = \mathbf{x}_t'\mathbf{G}'[\mathbf{G}']^{-1}\boldsymbol{\beta} + \varepsilon_t = [\mathbf{x}_t^*]'\boldsymbol{\beta}^* + \varepsilon_t, \tag{16.3.7}$$

where

$$\underset{(p+2)\times(p+2)}{\mathbf{G}'} \equiv \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ -\alpha+\delta & -\alpha+2\delta & \cdots & -\alpha+p\delta & 1 & 0 \\ -\delta & -\delta & \cdots & -\delta & 0 & 1 \end{bmatrix} \tag{16.3.8}$$

$$\underset{(p+2)\times(p+2)}{[\mathbf{G}']^{-1}} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ \alpha-\delta & \alpha-2\delta & \cdots & \alpha-p\delta & 1 & 0 \\ \delta & \delta & \cdots & \delta & 0 & 1 \end{bmatrix}$$

$$\mathbf{x}_t^* \equiv \mathbf{G}\mathbf{x}_t = \begin{bmatrix} y_{t-1}^* \\ y_{t-2}^* \\ \vdots \\ y_{t-p}^* \\ 1 \\ t \end{bmatrix} \tag{16.3.9}$$

$$\boldsymbol{\beta}^* \equiv [\mathbf{G}']^{-1}\boldsymbol{\beta} = \begin{bmatrix} \phi_1^* \\ \phi_2^* \\ \vdots \\ \phi_p^* \\ \alpha^* \\ \delta^* \end{bmatrix} \tag{16.3.10}$$

The system of [16.3.7] is just an algebraically equivalent representation of the regression model [16.3.5]. Notice that the estimate of $\boldsymbol{\beta}^*$ based on an *OLS* regression of y_t on $\mathbf{x}_t^*$ is given by

$$\begin{aligned} \mathbf{b}^* &= \left[\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t^* y_t\right] \\ &= \left[\mathbf{G}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)\mathbf{G}'\right]^{-1}\mathbf{G}\left(\sum_{t=1}^{T} \mathbf{x}_t y_t\right) \\ &= [\mathbf{G}']^{-1}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{G}^{-1}\mathbf{G}\left(\sum_{t=1}^{T} \mathbf{x}_t y_t\right) \\ &= [\mathbf{G}']^{-1}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T} \mathbf{x}_t y_t\right) \\ &= [\mathbf{G}']^{-1}\mathbf{b}, \end{aligned} \tag{16.3.11}$$

where $\mathbf{b}$ denotes the estimated coefficient vector from an *OLS* regression of y_t on $\mathbf{x}_t$. Thus, the coefficient estimate for the transformed regression ($\mathbf{b}^*$) is a simple linear transformation of the coefficient estimate for the original system ($\mathbf{b}$). The fitted value for date t associated with the transformed regression is

$$[\mathbf{x}_t^*]'\mathbf{b}^* = [\mathbf{G}\mathbf{x}_t]'[\mathbf{G}']^{-1}\mathbf{b} = \mathbf{x}_t'\mathbf{b}.$$

Thus, the fitted values for the transformed regression are numerically identical to the fitted values from the original regression.

Of course, given data only on $\{y_t\}$, we could not actually estimate the transformed regression by *OLS*, because construction of $\mathbf{x}_t^*$ from $\mathbf{x}_t$ requires knowledge of the true values of the parameters α and δ. It is nevertheless helpful to summarize the properties of hypothetical *OLS* estimation of [16.3.7], because [16.3.7] is easier to analyze than [16.3.5]. Moreover, once we find the asymptotic distribution of $\mathbf{b}^*$, the asymptotic distribution of $\mathbf{b}$ can be inferred by inverting [16.3.11]:

$$\mathbf{b} = \mathbf{G}'\mathbf{b}^*. \qquad [16.3.12]$$

The Asymptotic Distribution of OLS *Estimates for the Transformed Regression*

Appendix 16.A to this chapter demonstrates that

$$\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*) \xrightarrow{L} N(\mathbf{0}, \sigma^2[\mathbf{Q}^*]^{-1}), \qquad [16.3.13]$$

where

$$\underset{(p+2)\times(p+2)}{\mathbf{Y}_T} = \begin{bmatrix} \sqrt{T} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & \sqrt{T} & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{T} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & \sqrt{T} & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & T^{3/2} \end{bmatrix} \qquad [16.3.14]$$

$$\underset{(p+2)\times(p+2)}{\mathbf{Q}^*} = \begin{bmatrix} \gamma_0^* & \gamma_1^* & \gamma_2^* & \cdots & \gamma_{p-1}^* & 0 & 0 \\ \gamma_1^* & \gamma_0^* & \gamma_1^* & \cdots & \gamma_{p-2}^* & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \gamma_{p-1}^* & \gamma_{p-2}^* & \gamma_{p-3}^* & \cdots & \gamma_0^* & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & \cdots & 0 & \frac{1}{2} & \frac{1}{3} \end{bmatrix} \qquad [16.3.15]$$

for $\gamma_j^* \equiv E(y_t^* y_{t-j}^*)$. In other words, the *OLS* estimate $\mathbf{b}^*$ is asymptotically Gaussian, with the coefficient on the time trend ($\hat{\delta}^*$) converging at rate $T^{3/2}$ and all other coefficients converging at rate $\sqrt{T}$. The earlier result [16.1.26] is a special case of [16.3.13] with $p = 0$.

The Asymptotic Distribution of OLS *Estimates for the Original Regression*

What does this result imply about the asymptotic distribution of $\mathbf{b}$, the estimated coefficient vector for the *OLS* regression that is actually estimated? Writing

out [16.3.12] explicitly using [16.3.8], we have

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \vdots \\ \hat{\phi}_p \\ \hat{\alpha} \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 \\ -\alpha + \delta & -\alpha + 2\delta & \cdots & -\alpha + p\delta & 1 & 0 \\ -\delta & -\delta & \cdots & -\delta & 0 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_1^* \\ \hat{\phi}_2^* \\ \vdots \\ \hat{\phi}_p^* \\ \hat{\alpha}^* \\ \hat{\delta}^* \end{bmatrix}. \qquad [16.3.16]$$

The *OLS* estimates $\hat{\phi}_j$ of the untransformed regression are identical to the corresponding coefficients of the transformed regression $\hat{\phi}_j^*$, so the asymptotic distribution of $\hat{\phi}_j$ is given immediately by [16.3.13]. The estimate $\hat{\alpha}_T$ is a linear combination of variables that converge to a Gaussian distribution at rate $\sqrt{T}$, and so $\hat{\alpha}_T$ behaves the same way. Specifically, $\hat{\alpha}_T = \mathbf{g}_\alpha' \mathbf{b}_T^*$, where

$$\mathbf{g}_\alpha' \equiv [-\alpha + \delta \quad -\alpha + 2\delta \quad \cdots \quad -\alpha + p\delta \quad 1 \quad 0],$$

and so, from [16.3.13],

$$\sqrt{T}(\hat{\alpha}_T - \alpha) \xrightarrow{L} N(0, \sigma^2 \mathbf{g}_\alpha'[\mathbf{Q}^*]^{-1}\mathbf{g}_\alpha). \qquad [16.3.17]$$

Finally, the estimate $\hat{\delta}_T$ is a linear combination of variables converging at different rates:

$$\hat{\delta}_T = \mathbf{g}_\delta' \mathbf{b}_T^* + \hat{\delta}_T^*,$$

where

$$\mathbf{g}_\delta' \equiv [-\delta \quad -\delta \quad \cdots \quad -\delta \quad 0 \quad 0].$$

Its asymptotic distribution is governed by the variables with the slowest rate of convergence:

$$\begin{aligned} \sqrt{T}(\hat{\delta}_T - \delta) &= \sqrt{T}(\hat{\delta}_T^* + \mathbf{g}_\delta' \mathbf{b}_T^* - \delta^* - \mathbf{g}_\delta' \boldsymbol{\beta}^*) \\ &\xrightarrow{p} \sqrt{T}(\delta^* + \mathbf{g}_\delta' \mathbf{b}_T^* - \delta^* - \mathbf{g}_\delta' \boldsymbol{\beta}^*) \\ &= \mathbf{g}_\delta' \sqrt{T}(\mathbf{b}_T^* - \boldsymbol{\beta}^*) \\ &\xrightarrow{L} N(0, \sigma^2 \mathbf{g}_\delta'[\mathbf{Q}^*]^{-1}\mathbf{g}_\delta). \end{aligned}$$

Thus, each of the elements of $\mathbf{b}_T$ individually is asymptotically Gaussian and $O_p(T^{-1/2})$. The asymptotic distribution of the full vector $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})$ is multivariate Gaussian, though with a singular variance-covariance matrix. Specifically, the particular linear combination of the elements of $\mathbf{b}_T$ that recovers $\hat{\delta}_T^*$, the time trend coefficient of the hypothetical regression,

$$\hat{\delta}_T^* = -\mathbf{g}_\delta' \mathbf{b}_T^* + \hat{\delta}_T = \delta\hat{\phi}_{1,T} + \delta\hat{\phi}_{2,T} + \cdots + \delta\hat{\phi}_{p,T} + \hat{\delta}_T,$$

converges to a point mass around δ^* even when scaled by $\sqrt{T}$:

$$\sqrt{T}(\hat{\delta}_T^* - \delta^*) \xrightarrow{p} 0.$$

However, [16.3.13] establishes that

$$T^{3/2}(\hat{\delta}_T^* - \delta^*) \xrightarrow{L} N(0, \sigma^2 (q^*)^{p+2,p+2})$$

for $(q^*)^{p+2,p+2}$ the bottom right element of $[\mathbf{Q}^*]^{-1}$.

Hypothesis Tests

The preceding analysis described the asymptotic distribution of $\mathbf{b}$ in terms of the properties of the transformed regression estimates $\mathbf{b}^*$. This might seem to imply

that knowledge of the transformation matrix $\mathbf{G}$ in [16.3.8] is necessary in order to conduct hypothesis tests. Fortunately, this is not the case. The results of Section 16.2 turn out to apply equally well to the general model [16.3.1]—the usual t and F tests about $\boldsymbol{\beta}$ calculated in the usual way on the untransformed system are all asymptotically valid.

Consider the following null hypothesis about the parameters of the untransformed system:

$$H_0\text{: } \mathbf{R}\boldsymbol{\beta} = \mathbf{r}. \qquad [16.3.18]$$

Here $\mathbf{R}$ is a known $[m \times (p + 2)]$ matrix, $\mathbf{r}$ is a known $(m \times 1)$ vector, and m is the number of restrictions. The Wald form of the OLS χ^2 test of H_0 (expression [8.2.23]) is

$$\chi_T^2 = (\mathbf{R}\mathbf{b}_T - \mathbf{r})'\left[s_T^2\mathbf{R}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{R}'\right]^{-1}(\mathbf{R}\mathbf{b}_T - \mathbf{r}). \qquad [16.3.19]$$

Here $\mathbf{b}_T$ is the OLS estimate of $\boldsymbol{\beta}$ based on observation of $\{y_{-p+1}, y_{-p+2}, \ldots, y_0, y_1, \ldots, y_T\}$ and $s_T^2 = [1/(T - p - 2)]\,\Sigma_{t=1}^{T}(y_t - \mathbf{x}_t'\mathbf{b}_T)^2$.

Under the null hypothesis [16.3.18], expression [16.3.19] can be rewritten

$$\begin{aligned}\chi_T^2 &= [\mathbf{R}(\mathbf{b}_T - \boldsymbol{\beta})]'\left[s_T^2\mathbf{R}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{R}'\right]^{-1}[\mathbf{R}(\mathbf{b}_T - \boldsymbol{\beta})]\\ &= [\mathbf{R}\mathbf{G}'(\mathbf{G}')^{-1}(\mathbf{b}_T - \boldsymbol{\beta})]'\\ &\quad\times\left[s_T^2\mathbf{R}\mathbf{G}'(\mathbf{G}')^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}(\mathbf{G})^{-1}\mathbf{G}\mathbf{R}'\right]^{-1}[\mathbf{R}\mathbf{G}'(\mathbf{G}')^{-1}(\mathbf{b}_T - \boldsymbol{\beta})].\end{aligned} \qquad [16.3.20]$$

Notice that

$$(\mathbf{G}')^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}(\mathbf{G})^{-1} = \left[\mathbf{G}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)\mathbf{G}'\right]^{-1} = \left(\sum_{t=1}^{T}\mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1}$$

for $\mathbf{x}_t^*$ given by [16.3.9]. Similarly, from [16.3.10] and [16.3.11],

$$(\mathbf{b}_T^* - \boldsymbol{\beta}^*) = (\mathbf{G}')^{-1}(\mathbf{b}_T - \boldsymbol{\beta}).$$

Defining

$$\mathbf{R}^* \equiv \mathbf{R}\mathbf{G}',$$

expression [16.3.20] can be written

$$\begin{aligned}\chi_T^2 = [\mathbf{R}^*(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]'&\left[s_T^2\mathbf{R}^*\left(\sum_{t=1}^{T}\mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1}[\mathbf{R}^*]'\right]^{-1}\\ &\times[\mathbf{R}^*(\mathbf{b}_T^* - \boldsymbol{\beta}^*)].\end{aligned} \qquad [16.3.21]$$

Expression [16.3.21] will be recognized as the χ^2 test that would be calculated if we had estimated the transformed system and wanted to test the hypothesis that $\mathbf{R}^*\boldsymbol{\beta}^* = \mathbf{r}$ (recall that the fitted values for the transformed and untransformed regressions are identical, so that s_T^2 will be the same value for either representation). Observe that the transformed regression does not actually have to be estimated in order to calculate this statistic, since [16.3.21] is numerically identical to the χ^2 statistic [16.3.20] that is calculated from the untransformed system in the usual way. Nevertheless, expression [16.3.21] gives us another way of thinking about the distribution of the statistic as actually calculated in [16.3.20].

Expression [16.3.21] can be further rewritten as

$$\begin{aligned}\chi_T^2 = {} & [\mathbf{R}^*\mathbf{Y}_T^{-1}\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]' \\ & \times \left[s_T^2\mathbf{R}^*\mathbf{Y}_T^{-1}\mathbf{Y}_T\left(\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1} \mathbf{Y}_T\mathbf{Y}_T^{-1}[\mathbf{R}^*]'\right]^{-1} \\ & \times [\mathbf{R}^*\mathbf{Y}_T^{-1}\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]\end{aligned} \qquad [16.3.22]$$

for $\mathbf{Y}_T$ the matrix in [16.3.14]. Recall the insight from Section 16.2 that hypothesis tests involving coefficients with different rates of convergence will be dominated by the variables with the slowest rate of convergence. This means that some of the elements of $\mathbf{R}^*$ may be irrelevant asymptotically, so that [16.3.22] has the same asymptotic distribution as a simpler expression. To describe this expression, consider two possibilities.

Case 1. Each of the m *Hypotheses Represented by* $\mathbf{R}^*\boldsymbol{\beta}^* = \mathbf{r}$ *Involves a Parameter that Converges at Rate* $\sqrt{T}$

Of course, we could trivially rewrite any system of restrictions so as to involve $O_p(T^{-1/2})$ parameters in every equation. For example, the null hypothesis

$$H_0\colon \phi_2^* = 0, \qquad \delta^* = 0 \qquad [16.3.23]$$

could be rewritten as

$$H_0\colon \phi_2^* = 0, \qquad \delta^* = \phi_2^*, \qquad [16.3.24]$$

which seems to include ϕ_2^* in each restriction. For purposes of *implementing* a test of H_0, it does not matter which representation of H_0 is used, since either will produce the identical value for the test statistic.[8] For purposes of *analyzing the properties* of the test, we distinguish a hypothesis such as [16.3.23] from a hypothesis involving only ϕ_2^* and ϕ_3^*. For this distinction to be meaningful, we will assume that H_0 would be written in the form of [16.3.23] rather than [16.3.24].

[8]More generally, let $\mathbf{H}$ be any nonsingular $(m \times m)$ matrix. Then the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ can equivalently be written as $\dot{\mathbf{R}}\boldsymbol{\beta} = \dot{\mathbf{r}}$, where $\dot{\mathbf{R}} \equiv \mathbf{HR}$ and $\dot{\mathbf{r}} \equiv \mathbf{Hr}$. The χ^2 statistic constructed from the second parameterization is

$$\begin{aligned}\chi^2 &= (\dot{\mathbf{R}}\mathbf{b} - \dot{\mathbf{r}})'\left[s_T^2\dot{\mathbf{R}}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\dot{\mathbf{R}}'\right]^{-1}(\dot{\mathbf{R}}\mathbf{b} - \dot{\mathbf{r}}) \\ &= (\mathbf{Rb} - \mathbf{r})'\mathbf{H}'[\mathbf{H}']^{-1}\left[s_T^2\mathbf{R}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{R}'\right]^{-1}\mathbf{H}^{-1}\mathbf{H}(\mathbf{Rb} - \mathbf{r}) \\ &= (\mathbf{Rb} - \mathbf{r})'\left[s_T^2\mathbf{R}\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\mathbf{R}'\right]^{-1}(\mathbf{Rb} - \mathbf{r}),\end{aligned}$$

which is identical to the χ^2 statistic constructed from the first parameterization. The representation [16.3.24] is an example of such a transformation of [16.3.23], with

$$\mathbf{H} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}.$$

In general terms, this means that $\mathbf{R}^*$ is "upper triangular."[9] "Case 1" describes the situation in which the first $p + 1$ elements of the last row of $\mathbf{R}^*$ are not all zero.

For case 1, even though some of the hypotheses may involve $\hat{\delta}_T^*$, a test of the null hypothesis will be asymptotically equivalent to a test that treated δ^* as if known with certainty. This is a consequence of $\hat{\delta}_T^*$ being superconsistent. To develop this result rigorously, notice that

$$\mathbf{R}^*\mathbf{Y}_T^{-1} = \begin{bmatrix} r_{11}^*/\sqrt{T} & r_{12}^*/\sqrt{T} & \cdots & r_{1,p+1}^*/\sqrt{T} & r_{1,p+2}^*/T^{3/2} \\ r_{21}^*/\sqrt{T} & r_{22}^*/\sqrt{T} & \cdots & r_{2,p+1}^*/\sqrt{T} & r_{2,p+2}^*/T^{3/2} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m1}^*/\sqrt{T} & r_{m2}^*/\sqrt{T} & \cdots & r_{m,p+1}^*/\sqrt{T} & r_{m,p+2}^*/T^{3/2} \end{bmatrix},$$

and define

$$\underset{(m\times m)}{\tilde{\mathbf{Y}}_T} \equiv \sqrt{T}\mathbf{I}_m$$

$$\tilde{\mathbf{R}}_T^* \equiv \begin{bmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1,p+1}^* & r_{1,p+2}^*/T \\ r_{21}^* & r_{22}^* & \cdots & r_{2,p+1}^* & r_{2,p+2}^*/T \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m1}^* & r_{m2}^* & \cdots & r_{m,p+1}^* & r_{m,p+2}^*/T \end{bmatrix}.$$

These matrices were chosen so that

$$\mathbf{R}^*\mathbf{Y}_T^{-1} = \tilde{\mathbf{Y}}_T^{-1}\tilde{\mathbf{R}}_T^*. \qquad [16.3.25]$$

The matrix $\tilde{\mathbf{R}}_T^*$ has the further property that

$$\tilde{\mathbf{R}}_T^* \to \tilde{\mathbf{R}}^*, \qquad [16.3.26]$$

where $\tilde{\mathbf{R}}^*$ involves only those restrictions that affect the asymptotic distribution:

$$\tilde{\mathbf{R}}^* = \begin{bmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1,p+1}^* & 0 \\ r_{21}^* & r_{22}^* & \cdots & r_{2,p+1}^* & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m1}^* & r_{m2}^* & \cdots & r_{m,p+1}^* & 0 \end{bmatrix}.$$

[9] "Upper triangular" means that if the set of restrictions in H_0 involves parameters $\beta_{i_1}^*, \beta_{i_2}^*, \ldots, \beta_{i_n}^*$ with $i_1 < i_2 < \cdots < i_n$, then elements of $\mathbf{R}^*$ in rows 2 through m and columns 1 through i_1 are all zero. This is simply a normalization—any hypothesis $\mathbf{R}^*\boldsymbol{\beta}^* = \mathbf{r}$ can be written in such a form by selecting a restriction involving $\beta_{i_1}^*$ to be the first row of $\mathbf{R}^*$ and then multiplying the first row of this system of equations by a suitable constant and subtracting it from each of the following rows. If the system of restrictions represented by rows 2 through m of the resulting matrix involves parameters $\beta_{l_1}^*, \beta_{l_2}^*, \ldots, \beta_{l_l}^*$ with $j_1 < j_2 < \cdots < j_l$, then it is assumed that the elements in rows 3 through m and columns 1 through j_1 are all zero. An example of an upper triangular system is

$$\mathbf{R}^* = \begin{bmatrix} 0 & r_{1,l_1}^* & r_{1,l_2}^* & 0 & \cdots & 0 & r_{1,l_n}^* \\ 0 & 0 & 0 & r_{2,j_1}^* & \cdots & r_{2,l_l}^* & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & r_{m,k_{z-1}}^* & r_{m,k_z}^* \end{bmatrix}.$$

Substituting [16.3.25] into [16.3.22],

$$\chi_T^2 = [\tilde{\mathbf{Y}}_T^{-1}\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]'$$

$$\times \left[s_T^2\tilde{\mathbf{Y}}_T^{-1}\tilde{\mathbf{R}}_T^*\mathbf{Y}_T\left(\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1} \mathbf{Y}_T[\tilde{\mathbf{Y}}_T^{-1}\tilde{\mathbf{R}}_T^*]' \right]^{-1} [\tilde{\mathbf{Y}}_T^{-1}\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]$$

$$= [\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]'\tilde{\mathbf{Y}}_T^{-1}$$

$$\times \tilde{\mathbf{Y}}_T\left[s_T^2\tilde{\mathbf{R}}_T^*\mathbf{Y}_T\left(\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1} \mathbf{Y}_T[\tilde{\mathbf{R}}_T^*]' \right]^{-1} \tilde{\mathbf{Y}}_T\tilde{\mathbf{Y}}_T^{-1}[\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]$$

$$= [\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]'$$

$$\times \left[s_T^2\tilde{\mathbf{R}}_T^*\mathbf{Y}_T\left(\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right)^{-1} \mathbf{Y}_T[\tilde{\mathbf{R}}_T^*]' \right]^{-1} [\tilde{\mathbf{R}}_T^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]$$

$$\xrightarrow{p} [\tilde{\mathbf{R}}^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)]'[\sigma^2\tilde{\mathbf{R}}^*[\mathbf{Q}^*]^{-1}[\tilde{\mathbf{R}}^*]']^{-1}[\tilde{\mathbf{R}}^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*)] \qquad [16.3.27]$$

by virtue of [16.3.26] and [16.A.4].

Now [16.3.13] implies that

$$\tilde{\mathbf{R}}^*\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*) \xrightarrow{L} N(\mathbf{0}, \tilde{\mathbf{R}}^*\sigma^2[\mathbf{Q}^*]^{-1}[\tilde{\mathbf{R}}^*]'),$$

and so [16.3.27] is a quadratic form in an asymptotically Gaussian variable of the kind covered in Proposition 8.1. It is therefore asymptotically $\chi^2(m)$. Since [16.3.27] is numerically identical to [16.3.19], the Wald form of the *OLS* χ^2 test, calculated in the usual way from the untransformed regression [16.3.1], has the usual $\chi^2(m)$ distribution.

Case 2. One of the Hypotheses Involves Only the Time Trend Parameter δ^*

Again assuming for purposes of discussion that $\mathbf{R}^*$ is upper triangular, for case 2 the hypothesis about δ^* will be the sole entry in the mth row of $\mathbf{R}^*$:

$$\mathbf{R}^* = \begin{bmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1,p+1}^* & r_{1,p+2}^* \\ r_{21}^* & r_{22}^* & \cdots & r_{2,p+1}^* & r_{2,p+2}^* \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m-1,1}^* & r_{m-1,2}^* & \cdots & r_{m-1,p+1}^* & r_{m-1,p+2}^* \\ 0 & 0 & \cdots & 0 & r_{m,p+2}^* \end{bmatrix}.$$

For this case, define

$$\underset{(m\times m)}{\tilde{\mathbf{Y}}_T} \equiv \begin{bmatrix} \sqrt{T} & 0 & \cdots & 0 & 0 \\ 0 & \sqrt{T} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{T} & 0 \\ 0 & 0 & \cdots & 0 & T^{3/2} \end{bmatrix}$$

and

$$\tilde{\mathbf{R}}_T^* \equiv \begin{bmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1,p+1}^* & r_{1,p+2}^*/T \\ r_{21}^* & r_{22}^* & \cdots & r_{2,p+1}^* & r_{2,p+2}^*/T \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m-1,1}^* & r_{m-1,2}^* & \cdots & r_{m-1,p+1}^* & r_{m-1,p+2}^*/T \\ 0 & 0 & \cdots & 0 & r_{m,p+2}^* \end{bmatrix}.$$

Notice that these matrices again satisfy [16.3.25] and [16.3.26] with

$$\tilde{\mathbf{R}}^* = \begin{bmatrix} r_{11}^* & r_{12}^* & \cdots & r_{1,p+1}^* & 0 \\ r_{21}^* & r_{22}^* & \cdots & r_{2,p+1}^* & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ r_{m-1,1}^* & r_{m-1,2}^* & \cdots & r_{m-1,p+1}^* & 0 \\ 0 & 0 & \cdots & 0 & r_{m,p+2}^* \end{bmatrix}.$$

The analysis of [16.3.27] thus goes through for this case as well with no change.

Summary

Any standard *OLS* χ^2 test of the null hypothesis $\mathbf{R\beta} = \mathbf{r}$ for the regression model [16.3.1] can be calculated and interpreted in the usual way. The test is asymptotically valid for any hypothesis about any subset of the parameters in $\boldsymbol{\beta}$. The elements of $\mathbf{R}$ do not have to be ordered or expressed in any particular form for this to be true.

APPENDIX 16.A. *Derivation of Selected Equations for Chapter 16*

■ **Derivation of [16.3.13].** As in [16.1.6],

$$\mathbf{b}_T^* - \boldsymbol{\beta}^* = \left[\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t^*\varepsilon_t\right], \qquad [16.A.1]$$

since the population residuals ε_t are identical for the transformed and untransformed representations. As in [16.1.18], premultiply by $\mathbf{Y}_T$ to write

$$\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*) = \left\{\mathbf{Y}_T^{-1}\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\mathbf{Y}_T^{-1}\right\}^{-1}\left\{\mathbf{Y}_T^{-1}\sum_{t=1}^{T}\mathbf{x}_t^*\varepsilon_t\right\}. \qquad [16.A.2]$$

From [16.3.9],

$$\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]' = \begin{bmatrix} \Sigma(y_{t-1}^*)^2 & \Sigma y_{t-1}^* y_{t-2}^* & \cdots & \Sigma y_{t-1}^* y_{t-p}^* & \Sigma y_{t-1}^* & \Sigma t y_{t-1}^* \\ \Sigma y_{t-2}^* y_{t-1}^* & \Sigma(y_{t-2}^*)^2 & \cdots & \Sigma y_{t-2}^* y_{t-p}^* & \Sigma y_{t-2}^* & \Sigma t y_{t-2}^* \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \Sigma y_{t-p}^* y_{t-1}^* & \Sigma y_{t-p}^* y_{t-2}^* & \cdots & \Sigma(y_{t-p}^*)^2 & \Sigma y_{t-p}^* & \Sigma t y_{t-p}^* \\ \Sigma y_{t-1}^* & \Sigma y_{t-2}^* & \cdots & \Sigma y_{t-p}^* & \Sigma 1 & \Sigma t \\ \Sigma t y_{t-1}^* & \Sigma t y_{t-2}^* & \cdots & \Sigma t y_{t-p}^* & \Sigma t & \Sigma t^2 \end{bmatrix}$$

and

$$\mathbf{Y}_T^{-1}\sum_{t=1}^{T} \mathbf{x}_t^*[\mathbf{x}_t^*]'\mathbf{Y}_T^{-1}$$

$$= \begin{bmatrix} T^{-1}\Sigma(y_{t-1}^*)^2 & T^{-1}\Sigma y_{t-1}^* y_{t-2}^* & \cdots & T^{-1}\Sigma y_{t-1}^* y_{t-p}^* & T^{-1}\Sigma y_{t-1}^* & T^{-2}\Sigma t y_{t-1}^* \\ T^{-1}\Sigma y_{t-2}^* y_{t-1}^* & T^{-1}\Sigma(y_{t-2}^*)^2 & \cdots & T^{-1}\Sigma y_{t-2}^* y_{t-p}^* & T^{-1}\Sigma y_{t-2}^* & T^{-2}\Sigma t y_{t-2}^* \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ T^{-1}\Sigma y_{t-p}^* y_{t-1}^* & T^{-1}\Sigma y_{t-p}^* y_{t-2}^* & \cdots & T^{-1}\Sigma(y_{t-p}^*)^2 & T^{-1}\Sigma y_{t-p}^* & T^{-2}\Sigma t y_{t-p}^* \\ T^{-1}\Sigma y_{t-1}^* & T^{-1}\Sigma y_{t-2}^* & \cdots & T^{-1}\Sigma y_{t-p}^* & T^{-1}\cdot T & T^{-2}\cdot\Sigma t \\ T^{-2}\Sigma t y_{t-1}^* & T^{-2}\Sigma t y_{t-2}^* & \cdots & T^{-2}\Sigma t y_{t-p}^* & T^{-2}\cdot\Sigma t & T^{-3}\cdot\Sigma t^2 \end{bmatrix}. \qquad [16.A.3]$$

For the first p rows and columns, the row i, column j element of this matrix is

$$T^{-1} \sum_{t=1}^{T} y^*_{t-i} y^*_{t-j}.$$

But y^*_t follows a zero-mean stationary $AR(p)$ process satisfying the conditions of Exercise 7.7. Thus, these terms converge in probability to $\gamma^*_{|i-j|}$. The first p elements of row $p + 1$ (or the first p elements of column $p + 1$) are of the form

$$T^{-1} \sum_{t=1}^{T} y^*_{t-i},$$

which converge in probability to zero. The first p elements of row $p + 2$ (or the first p elements of column $p + 2$) are of the form

$$T^{-1} \sum_{t=1}^{T} (t/T) y^*_{t-i},$$

which can be shown to converge in probability to zero with a ready adaptation of the techniques in Chapter 7 (see Exercise 16.3). Finally, the (2×2) matrix in the bottom right corner of [16.A.3] converges to

$$\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}.$$

Thus

$$\mathbf{Y}_T^{-1} \sum_{t=1}^{T} \mathbf{x}^*_t [\mathbf{x}^*_t]' \mathbf{Y}_T^{-1} \xrightarrow{p} \mathbf{Q}^* \qquad [16.A.4]$$

for $\mathbf{Q}^*$ the matrix in [16.3.15].

Turning next to the second term in [16.A.2],

$$\mathbf{Y}_T^{-1} \sum_{t=1}^{T} \mathbf{x}^*_t \varepsilon_t = \begin{bmatrix} T^{-1/2}\Sigma y^*_{t-1}\varepsilon_t \\ T^{-1/2}\Sigma y^*_{t-2}\varepsilon_t \\ \vdots \\ T^{-1/2}\Sigma y^*_{t-p}\varepsilon_t \\ T^{-1/2}\Sigma \varepsilon_t \\ T^{-1/2}\Sigma (t/T)\varepsilon_t \end{bmatrix} = T^{-1/2} \sum_{t=1}^{T} \boldsymbol{\xi}_t, \qquad [16.A.5]$$

where

$$\boldsymbol{\xi}_t \equiv \begin{bmatrix} y^*_{t-1}\varepsilon_t \\ y^*_{t-2}\varepsilon_t \\ \vdots \\ y^*_{t-p}\varepsilon_t \\ \varepsilon_t \\ (t/T)\varepsilon_t \end{bmatrix}.$$

But $\boldsymbol{\xi}_t$ is a martingale difference sequence with variance

$$E(\boldsymbol{\xi}_t \boldsymbol{\xi}_t') = \sigma^2 \mathbf{Q}^*_t,$$

where

$$\mathbf{Q}^*_t = \begin{bmatrix} \gamma^*_0 & \gamma^*_1 & \gamma^*_2 & \cdots & \gamma^*_{p-1} & 0 & 0 \\ \gamma^*_1 & \gamma^*_0 & \gamma^*_1 & \cdots & \gamma^*_{p-2} & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \gamma^*_{p-1} & \gamma^*_{p-2} & \gamma^*_{p-3} & \cdots & \gamma^*_0 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & t/T \\ 0 & 0 & 0 & \cdots & 0 & t/T & t^2/T^2 \end{bmatrix}$$

and

$$(1/T)\sum_{t=1}^{T} \mathbf{Q}_t^* \to \mathbf{Q}^*.$$

Applying the arguments used in Exercise 8.3 and in [16.1.24], it can be shown that

$$\mathbf{Y}_T^{-1}\sum_{t=1}^{T} \mathbf{x}_t^*\varepsilon_t \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}^*). \qquad [16.A.6]$$

It follows from [16.A.4], [16.A.6], and [16.A.2] that

$$\mathbf{Y}_T(\mathbf{b}_T^* - \boldsymbol{\beta}^*) \xrightarrow{L} N(\mathbf{0}, [\mathbf{Q}^*]^{-1}\sigma^2\mathbf{Q}^*[\mathbf{Q}^*]^{-1}) = N(\mathbf{0}, \sigma^2[\mathbf{Q}^*]^{-1}),$$

as claimed in [16.3.13]. ■

Chapter 16 Exercises

16.1. Verify result [16.1.23].

16.2. Verify expression [16.1.27].

16.3. Let y_t be covariance-stationary with mean zero and absolutely summable autocovariances:

$$\sum_{j=-\infty}^{\infty} |\gamma_j| < \infty$$

for $\gamma_j = E(y_t y_{t-j})$. Adapting the argument in expression [7.2.6], show that

$$T^{-1}\sum_{t=1}^{T} (t/T)y_t \xrightarrow{m.s.} 0.$$

Chapter 16 References

Fuller, Wayne A. 1976. *Introduction to Statistical Time Series*. New York: Wiley.

Sims, Christopher A., James H. Stock, and Mark W. Watson. 1990. "Inference in Linear Time Series Models with Some Unit Roots." *Econometrica* 58:113–44.

17 Univariate Processes with Unit Roots

This chapter discusses statistical inference for univariate processes containing a unit root. Section 17.1 gives a brief explanation of why the asymptotic distributions and rates of convergence for the estimated coefficients of unit root processes differ from those for stationary processes. The asymptotic distributions for unit root processes can be described in terms of functionals on Brownian motion. The basic idea behind Brownian motion is introduced in Section 17.2. The technical tools used to establish that the asymptotic distributions of certain statistics involving unit root processes can be represented in terms of such functionals are developed in Section 17.3, though it is not necessary to master these tools in order to read Sections 17.4 through 17.9. Section 17.4 derives the asymptotic distribution of the estimated coefficient for a first-order autoregression when the true process is a random walk. This distribution turns out to depend on whether a constant or time trend is included in the estimated regression and whether the true random walk is characterized by nonzero drift.

Section 17.5 extends the results of Section 17.3 to cover unit root processes whose differences exhibit general serial correlation. These results can be used to develop two different classes of tests for unit roots. One approach, due to Phillips and Perron (1988), adjusts the statistics calculated from a simple first-order autoregression to account for serial correlation of the differenced data. The second approach, due to Dickey and Fuller (1979), adds lags to the autoregression. These approaches are reviewed in Sections 17.6 and 17.7, respectively. Section 17.7 further derives the properties of all of the estimated coefficients for a pth-order autoregression when one of the roots is unity.

Readers interested solely in how these results are applied in practice may want to begin with the summaries in Table 17.2 or Table 17.3 and with the empirical applications described in Examples 17.6 through 17.9.

17.1. *Introduction*

Consider OLS estimation of a Gaussian $AR(1)$ process,

$$y_t = py_{t-1} + u_t, \qquad [17.1.1]$$

where $u_t \sim$ i.i.d. $N(0, \sigma^2)$, and $y_0 = 0$. The OLS estimate of p is given by

$$\hat{p}_T = \frac{\sum_{t=1}^{T} y_{t-1}y_t}{\sum_{t=1}^{T} y_{t-1}^2}. \qquad [17.1.2]$$

We saw in Chapter 8 that if the true value of ρ is less than 1 in absolute value, then

$$\sqrt{T}(\hat{\rho}_T - \rho) \xrightarrow{L} N(0, (1 - \rho^2)). \qquad [17.1.3]$$

If [17.1.3] were also valid for the case when $\rho = 1$, it would seem to claim that $\sqrt{T}(\hat{\rho}_T - \rho)$ has zero variance, or that the distribution collapses to a point mass at zero:

$$\sqrt{T}(\hat{\rho}_T - 1) \xrightarrow{p} 0. \qquad [17.1.4]$$

As we shall see shortly, [17.1.4] is indeed a valid statement for unit root processes, but it obviously is not very helpful for hypothesis tests. To obtain a nondegenerate asymptotic distribution for $\hat{\rho}_T$ in the unit root case, it turns out that we have to multiply $\hat{\rho}_T$ by T rather than by $\sqrt{T}$. Thus, the unit root coefficient converges at a faster rate (T) than a coefficient for a stationary regression (which converges at $\sqrt{T}$), but at a slower rate than the coefficient on a time trend in the regressions analyzed in the previous chapter (which converged at $T^{3/2}$).

To get a better sense of why scaling by T is necessary when the true value of ρ is unity, recall that the difference between the estimate $\hat{\rho}_T$ and the true value can be expressed as in equation [8.2.3]:[1]

$$(\hat{\rho}_T - 1) = \frac{\sum_{t=1}^{T} y_{t-1}u_t}{\sum_{t=1}^{T} y_{t-1}^2}, \qquad [17.1.5]$$

so that

$$T(\hat{\rho}_T - 1) = \frac{(1/T)\sum_{t=1}^{T} y_{t-1}u_t}{(1/T^2)\sum_{t=1}^{T} y_{t-1}^2}. \qquad [17.1.6]$$

Consider first the numerator in [17.1.6]. When the true value of ρ is unity, equation [17.1.1] describes a random walk with

$$y_t = u_t + u_{t-1} + \cdots + u_1, \qquad [17.1.7]$$

since $y_0 = 0$. It follows from [17.1.7] that

$$y_t \sim N(0, \sigma^2 t). \qquad [17.1.8]$$

Note further that for a random walk,

$$y_t^2 = (y_{t-1} + u_t)^2 = y_{t-1}^2 + 2y_{t-1}u_t + u_t^2,$$

implying that

$$y_{t-1}u_t = (1/2)\{y_t^2 - y_{t-1}^2 - u_t^2\}. \qquad [17.1.9]$$

If [17.1.9] is summed over $t = 1, 2, \ldots, T$, the result is

$$\sum_{t=1}^{T} y_{t-1}u_t = (1/2)\{y_T^2 - y_0^2\} - (1/2)\sum_{t=1}^{T} u_t^2. \qquad [17.1.10]$$

Recalling that $y_0 = 0$, equation [17.1.10] establishes that

$$(1/T)\sum_{t=1}^{T} y_{t-1}u_t = (1/2)\cdot(1/T)y_T^2 - (1/2)\cdot(1/T)\sum_{t=1}^{T} u_t^2, \qquad [17.1.11]$$

[1]This discussion is based on Fuller (1976, p. 369).

and if each side of [17.1.11] is divided by σ^2, the result is

$$\left(\frac{1}{\sigma^2 T}\right)\sum_{t=1}^{T} y_{t-1}u_t = \left(\frac{1}{2}\right)\left(\frac{y_T}{\sigma\sqrt{T}}\right)^2 - \left(\frac{1}{2\sigma^2}\right)\left(\frac{1}{T}\right)\cdot\sum_{t=1}^{T} u_t^2. \qquad [17.1.12]$$

But [17.1.8] implies that $y_T/(\sigma\sqrt{T})$ is a $N(0, 1)$ variable, so that its square is $\chi^2(1)$:

$$[y_T/(\sigma\sqrt{T})]^2 \sim \chi^2(1). \qquad [17.1.13]$$

Also, $\Sigma_{t=1}^{T} u_t^2$ is the sum of T i.i.d. random variables, each with mean σ^2, and so, by the law of large numbers,

$$(1/T)\cdot\sum_{t=1}^{T} u_t^2 \xrightarrow{p} \sigma^2. \qquad [17.1.14]$$

Using [17.1.13] and [17.1.14], it follows from [17.1.12] that

$$[1/(\sigma^2 T)]\sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{L} (1/2)\cdot(X - 1), \qquad [17.1.15]$$

where $X \sim \chi^2(1)$.

Turning next to the denominator of [17.1.6], consider

$$\sum_{t=1}^{T} y_{t-1}^2. \qquad [17.1.16]$$

Recall from [17.1.8] that $y_{t-1} \sim N(0, \sigma^2(t-1))$, so $E(y_{t-1}^2) = \sigma^2(t-1)$. Consider the mean of [17.1.16],

$$E\left[\sum_{t=1}^{T} y_{t-1}^2\right] = \sigma^2\sum_{t=1}^{T}(t-1) = \sigma^2(T-1)T/2.$$

In order to construct a random variable that could have a convergent distribution, the quantity in [17.1.16] will have to be divided by T^2, as was done in the denominator of [17.1.6].

To summarize, if the true process is a random walk, then the deviation of the *OLS* estimate from the true value $(\hat{\rho}_T - 1)$ must be multiplied by T rather than $\sqrt{T}$ to obtain a variable with a useful asymptotic distribution. Moreover, this asymptotic distribution is not the usual Gaussian distribution but instead is a ratio involving a $\chi^2(1)$ variable in the numerator and a separate, nonstandard distribution in the denominator.

The asymptotic distribution of $T(\hat{\rho}_T - 1)$ will be fully characterized in Section 17.4. In preparation for this, the idea of Brownian motion is introduced in Section 17.2, followed by a discussion of the functional central limit theorem in Section 17.3.

17.2. *Brownian Motion*

Consider a random walk,

$$y_t = y_{t-1} + \varepsilon_t, \qquad [17.2.1]$$

in which the innovations are standard Normal variables:

$$\varepsilon_t \sim \text{i.i.d } N(0, 1).$$

If the process is started with $y_0 = 0$, then it follows as in [17.1.7] and [17.1.8] that

$$y_t = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t$$
$$y_t \sim N(0, t).$$

Moreover, the change in the value of y between dates t and s,

$$y_s - y_t = \varepsilon_{t+1} + \varepsilon_{t+2} + \cdots + \varepsilon_s,$$

is itself $N(0, (s - t))$ and is independent of the change between dates r and q for any dates $t < s < r < q$.

Consider the change between y_{t-1} and y_t. This innovation ε_t was taken to be $N(0, 1)$. Suppose we view ε_t as the sum of two independent Gaussian variables:

$$\varepsilon_t = e_{1t} + e_{2t},$$

with $e_{it} \sim$ i.i.d. $N(0, \frac{1}{2})$. We might then associate e_{1t} with the change between y_{t-1} and the value of y at some interim point (say, $y_{t-(1/2)}$),

$$y_{t-(1/2)} - y_{t-1} = e_{1t}, \qquad [17.2.2]$$

and e_{2t} with the change between $y_{t-(1/2)}$ and y_t:

$$y_t - y_{t-(1/2)} = e_{2t}. \qquad [17.2.3]$$

Sampled at integer dates $t = 1, 2, \ldots$, the process of [17.2.2] and [17.2.3] will have exactly the same properties as [17.2.1], since

$$y_t - y_{t-1} = e_{1t} + e_{2t} \sim \text{i.i.d. } N(0, 1).$$

In addition, the process of [17.2.2] and [17.2.3] is defined also at the noninteger dates $\{t + \frac{1}{2}\}_{t=0}^{\infty}$ and retains the property for both integer and noninteger dates that $y_s - y_t \sim N(0, s - t)$ with $y_s - y_t$ independent of the change over any other nonoverlapping interval.

By the same reasoning, we could imagine partitioning the change between $t - 1$ and t into N separate subperiods:

$$y_t - y_{t-1} = e_{1t} + e_{2t} + \cdots + e_{Nt},$$

with $e_{it} \sim$ i.i.d. $N(0, 1/N)$. The result would be a process with all the same properties as [17.2.1], defined at a finer and finer grid of dates as we increase N. The limit as $N \to \infty$ is a *continuous-time* process known as *standard Brownian motion*. The value of this process at date t is denoted $W(t)$.[2] A continuous-time process is a random variable that takes on a value for any nonnegative real number t, as distinct from a discrete-time process, which is only defined at integer values of t. To emphasize the distinction, we will put the date in parentheses when describing the value of a continuous-time variable at date t (as in $W(t)$) and use subscripts for a discrete-time variable (as in y_t). A discrete-time process was represented as a countable sequence of random variables, denoted $\{y_t\}_{t=1}^{\infty}$. A realization of a continuous-time process can be viewed as a stochastic function, denoted $W(\cdot)$, where $W: t \in [0, \infty) \to \mathbb{R}^1$.

A particular realization of Brownian motion turns out to be a continuous function of t. To see why it would be continuous, recall that the change between t and $t + \Delta$ is distributed $N(0, \Delta)$. Such a change is essentially certain to be arbitrarily small as the interval Δ goes to zero.

Definition: *Standard Brownian motion $W(\cdot)$ is a continuous-time stochastic process, associating each date $t \in [0, 1]$ with the scalar $W(t)$ such that:*

(*a*) $W(0) = 0$;

[2]Brownian motion is sometimes also referred to as a *Wiener process*.

(b) *For any dates* $0 \leq t_1 < t_2 < \cdots < t_k \leq 1$, *the changes* $[W(t_2) - W(t_1)]$, $[W(t_3) - W(t_2)], \ldots, [W(t_k) - W(t_{k-1})]$ *are independent multivariate Gaussian with* $[W(s) - W(t)] \sim N(0, s - t)$;

(c) *For any given realization,* $W(t)$ *is continuous in* t *with probability 1.*

There are advantages to restricting the analysis to dates t within a closed interval. All of the results in this text relate to the behavior of Brownian motion for dates within the unit interval ($t \in [0, 1]$), and in anticipation of this we have simply defined $W(\cdot)$ to be a function mapping $t \in [0, 1]$ into $\mathbb{R}^1$.

Other continuous-time processes can be generated from standard Brownian motion. For example, the process

$$Z(t) = \sigma \cdot W(t)$$

has independent increments and is distributed $N(0, \sigma^2 t)$ across realizations. Such a process is described as *Brownian motion with variance* σ^2. Thus, standard Brownian motion could also be described as Brownian motion with unit variance.

As another example,

$$Z(t) = [W(t)]^2 \qquad [17.2.4]$$

would be distributed as t times a $\chi^2(1)$ variable across realizations.

Although $W(t)$ is continuous in t, it cannot be differentiated using standard calculus; the direction of change at t is likely to be completely different from that at $t + \Delta$, no matter how small we make Δ.[3]

17.3. *The Functional Central Limit Theorem*

One of the uses of Brownian motion is to permit more general statements of the central limit theorem than those in Chapter 7. Recall the simplest version of the central limit theorem: if $u_t \sim$ i.i.d. with mean zero and variance σ^2, then the sample mean $\bar{u}_T \equiv (1/T)\Sigma_{t=1}^{T} u_t$ satisfies

$$\sqrt{T}\bar{u}_T \xrightarrow{L} N(0, \sigma^2).$$

Consider now an estimator based on the following principle: When given a sample of size T, we calculate the mean of the first half of the sample and throw out the rest of the observations:

$$\bar{u}_{[T/2]^*} = (1/[T/2]^*) \sum_{t=1}^{[T/2]^*} u_t.$$

Here $[T/2]^*$ denotes the largest integer that is less than or equal to $T/2$; that is, $[T/2]^* = T/2$ for T even and $[T/2]^* = (T - 1)/2$ for T odd. This strange estimator would also satisfy the central limit theorem:

$$\sqrt{[T/2]^*}\,\bar{u}_{[T/2]^*} \xrightarrow[T\to\infty]{L} N(0, \sigma^2). \qquad [17.3.1]$$

Moreover, this estimator would be independent of an estimator that uses only the second half of the sample.

More generally, we can construct a variable $X_T(r)$ from the sample mean of

[3]For an introduction to differentiation and integration of Brownian motion, see Malliaris and Brock (1982, Chapter 2).

the first rth fraction of observations, $r \in [0, 1]$, defined by

$$X_T(r) \equiv (1/T) \sum_{t=1}^{[Tr]^*} u_t. \qquad [17.3.2]$$

For any given realization, $X_T(r)$ is a step function in r, with

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \le r < 1/T \\ u_1/T & \text{for } 1/T \le r < 2/T \\ (u_1 + u_2)/T & \text{for } 2/T \le r < 3/T \\ \vdots & \\ (u_1 + u_2 + \cdots + u_T)/T & \text{for } r = 1. \end{cases} \qquad [17.3.3]$$

Then

$$\sqrt{T} \cdot X_T(r) = (1/\sqrt{T}) \sum_{t=1}^{[Tr]^*} u_t = (\sqrt{[Tr]^*}/\sqrt{T})(1/\sqrt{[Tr]^*}) \sum_{t=1}^{[Tr]^*} u_t. \qquad [17.3.4]$$

But

$$(1/\sqrt{[Tr]^*}) \sum_{t=1}^{[Tr]^*} u_t \xrightarrow{L} N(0, \sigma^2),$$

by the central limit theorem as in [17.3.1], while $(\sqrt{[Tr]^*}/\sqrt{T}) \to \sqrt{r}$. Hence, the asymptotic distribution of $\sqrt{T} \cdot X_T(r)$ in [17.3.4] is that of $\sqrt{r}$ times a $N(0, \sigma^2)$ random variable, or

$$\sqrt{T} \cdot X_T(r) \xrightarrow{L} N(0, r\sigma^2)$$

and

$$\sqrt{T} \cdot [X_T(r)/\sigma] \xrightarrow{L} N(0, r). \qquad [17.3.5]$$

If we were similarly to consider the behavior of a sample mean based on observations $[Tr_1]^*$ through $[Tr_2]^*$ for $r_2 > r_1$, we would conclude that this too is asymptotically Normal,

$$\sqrt{T} \cdot [X_T(r_2) - X_T(r_1)]/\sigma \xrightarrow{L} N(0, r_2 - r_1),$$

and is independent of the estimator in [17.3.5], provided that $r < r_1$. It thus should not be surprising that the sequence of stochastic functions $\{\sqrt{T} \cdot X_T(\cdot)/\sigma\}_{T=1}^{\infty}$ has an asymptotic probability law that is described by standard Brownian motion $W(\cdot)$:

$$\sqrt{T} \cdot X_T(\cdot)/\sigma \xrightarrow{L} W(\cdot). \qquad [17.3.6]$$

Note the difference between the claims in [17.3.5] and [17.3.6]. The expression $X_T(\cdot)$ denotes a random function while $X_T(r)$ denotes the value that function assumes at date r; thus, $X_T(\cdot)$ is a function, while $X_T(r)$ is a random variable.

Result [17.3.6] is known as the *functional central limit theorem*. The derivation here assumed that u_t was i.i.d. A more general statement will be provided in Section 17.5.

Evaluated at $r = 1$, the function $X_T(r)$ in [17.3.2] is just the sample mean:

$$X_T(1) = (1/T) \sum_{t=1}^{T} u_t.$$

Thus, when the functions in [17.3.6] are evaluated at $r = 1$, the conventional central limit theorem [7.1.6] obtains as a special case of [17.3.6]:

$$\sqrt{T} X_T(1)/\sigma = [1/(\sigma\sqrt{T})] \sum_{t=1}^{T} u_t \xrightarrow{L} W(1) \sim N(0, 1). \qquad [17.3.7]$$

We earlier defined convergence in law for random variables, and now we need to extend the definition to cover random functions. Let $S(\cdot)$ represent a continuous-time stochastic process with $S(r)$ representing its value at some date r for $r \in [0, 1]$. Suppose, further, that for any given realization, $S(\cdot)$ is a continuous function of r with probability 1. For $\{S_T(\cdot)\}_{T=1}^{\infty}$ a sequence of such continuous functions, we say that $S_T(\cdot) \xrightarrow{L} S(\cdot)$ if all of the following hold:[4]

(a) For any finite collection of k particular dates,

$$0 \leq r_1 < r_2 < \cdots < r_k \leq 1,$$

the sequence of k-dimensional random vectors $\{\mathbf{y}_T\}_{T=1}^{\infty}$ converges in distribution to the vector $\mathbf{y}$, where

$$\mathbf{y}_T \equiv \begin{bmatrix} S_T(r_1) \\ S_T(r_2) \\ \vdots \\ S_T(r_k) \end{bmatrix} \qquad \mathbf{y} \equiv \begin{bmatrix} S(r_1) \\ S(r_2) \\ \vdots \\ S(r_k) \end{bmatrix};$$

(b) For each $\varepsilon > 0$, the probability that $S_T(r_1)$ differs from $S_T(r_2)$ for any dates r_1 and r_2 within δ of each other goes to zero uniformly in T as $\delta \to 0$;

(c) $P\{|S_T(0)| > \lambda\} \to 0$ uniformly in T as $\lambda \to \infty$.

This definition applies to sequences of continuous functions, though the function in [17.3.2] is a discontinuous step function. Fortunately, the discontinuities occur at a countable set of points. Formally, $S_T(\cdot)$ can be replaced with a similar continuous function, interpolating between the steps (as in Hall and Heyde, 1980). Alternatively, the definition of convergence of random functions can be generalized to allow for discontinuities of the type in [17.3.2] (as in Chapter 3 of Billingsley, 1968).

It will also be helpful to extend the earlier definition of convergence in probability to sequences of random functions. Let $\{S_T(\cdot)\}_{T=1}^{\infty}$ and $\{V_T(\cdot)\}_{T=1}^{\infty}$ denote sequences of random continuous functions with S_T: $r \in [0, 1] \to \mathbb{R}^1$ and V_T: $r \in [0, 1] \to \mathbb{R}^1$. Let the scalar Y_T represent the largest amount by which $S_T(r)$ differs from $V_T(r)$ for any r:

$$Y_T \equiv \sup_{r \in [0,1]} |S_T(r) - V_T(r)|.$$

Thus, $\{Y_T\}_{T=1}^{\infty}$ is a sequence of random variables, and we could talk about its probability limit using the standard definition given in [7.1.2]. If the sequence of scalars $\{Y_T\}_{T=1}^{\infty}$ converges in probability to zero, then we say that the sequence of functions $S_T(\cdot)$ converges in probability to $V_T(\cdot)$. That is, the expression

$$S_T(\cdot) \xrightarrow{p} V_T(\cdot)$$

is interpreted to mean that

$$\sup_{r \in [0,1]} |S_T(r) - V_T(r)| \xrightarrow{p} 0.$$

With this definition, result (a) of Proposition 7.3 can be generalized to apply

[4]The sequence of probability measures induced by $\{S_T(\cdot)\}_{T=1}^{\infty}$ weakly converges (in the sense of Billingsley, 1968) to the probability measure induced by $S(\cdot)$ if and only if conditions (a) through (c) hold; see Theorem A.2, p. 275, in Hall and Heyde (1980).

to sequences of functions. Specifically, if $\{S_T(\cdot)\}_{T=1}^{\infty}$ and $\{V_T(\cdot)\}_{T=1}^{\infty}$ are sequences of continuous functions with $V_T(\cdot) \xrightarrow{p} S_T(\cdot)$ and $S_T(\cdot) \xrightarrow{L} S(\cdot)$ for $S(\cdot)$ a continuous function, then $V_T(\cdot) \xrightarrow{L} S(\cdot)$; see, for example, Stinchcombe and White (1993).

Example 17.1

Let $\{x_T\}_{T=1}^{\infty}$ be a sequence of random scalars with $x_T \xrightarrow{p} 0$, and let $\{S_T(\cdot)\}_{T=1}^{\infty}$ be a sequence of random continuous functions, S_T: $r \in [0, 1] \to \mathbb{R}^1$ with $S_T(\cdot) \xrightarrow{L} S(\cdot)$. Then the sequence of functions $\{V_T(\cdot)\}_{T=1}^{\infty}$ defined by $V_T(r) \equiv S_T(r) + x_T$ has the property that $V_T(\cdot) \xrightarrow{L} S(\cdot)$. To see this, note that $V_T(r) - S_T(r) = x_T$ for all r, so that

$$\sup_{r \in [0,1]} |S_T(r) - V_T(r)| = |x_T|,$$

which converges in probability to zero. Hence, $V_T(\cdot) \xrightarrow{p} S_T(\cdot)$, and therefore $V_T(\cdot) \xrightarrow{L} S(\cdot)$.

Example 17.2

Let η_t be a strictly stationary time series with finite fourth moment, and let $S_T(r) = (1/\sqrt{T}) \cdot \eta_{[Tr]}$. Then $S_T(\cdot) \xrightarrow{p} 0$. To see this, note that

$$\begin{aligned}
&P\left\{\sup_{r \in [0,1]} |S_T(r)| > \delta\right\} \\
&= P\{[|(1/\sqrt{T}) \cdot \eta_1| > \delta] \quad \text{or} \quad [|(1/\sqrt{T}) \cdot \eta_2| > \delta] \quad \text{or} \quad \cdots \\
&\quad \text{or} \quad [|(1/\sqrt{T}) \cdot \eta_T| > \delta]\} \\
&\leq T \cdot P\{|(1/\sqrt{T}) \cdot \eta_t| > \delta]\} \\
&\leq T \cdot \frac{E\{(1/\sqrt{T}) \cdot \eta_t\}^4}{\delta^4} \\
&= \frac{E(\eta_t^4)}{T\delta^4},
\end{aligned}$$

where the next-to-last line follows from Chebyshev's inequality. Since $E(\eta_t^4)$ is finite, this probability goes to zero as $T \to \infty$, establishing that $S_T(\cdot) \xrightarrow{p} 0$, as claimed.

Continuous Mapping Theorem

In Chapter 7 we saw that if $\{x_T\}_{T=1}^{\infty}$ is a sequence of random variables with $x_T \xrightarrow{L} x$ and if g: $\mathbb{R}^1 \to \mathbb{R}^1$ is a continuous function, then $g(x_T) \xrightarrow{L} g(x)$. A similar result holds for sequences of random functions. Here, the analog to the function $g(\cdot)$ is a continuous *functional*, which could associate a real random variable y with the stochastic function $S(\cdot)$. For example, $y = \int_0^1 S(r)\, dr$ and $y = \int_0^1 [S(r)]^2\, dr$ represent continuous functionals.[5] The *continuous mapping theorem*[6] states that if $S_T(\cdot) \xrightarrow{L} S(\cdot)$ and $g(\cdot)$ is a continuous functional, then $g(S_T(\cdot)) \xrightarrow{L} g(S(\cdot))$.

[5]Continuity of a functional $g(\cdot)$ in this context means that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $h(r)$ and $k(r)$ are any continuous bounded functions on $[0, 1]$, h: $[0, 1] \to \mathbb{R}^1$ and k: $[0, 1] \to \mathbb{R}^1$, such that $|h(r) - k(r)| < \delta$ for all $r \in [0, 1]$, then

$$|g[h(\cdot)] - g[k(\cdot)]| < \varepsilon.$$

[6]See, for example, Theorem A.3 on p. 276 in Hall and Heyde (1980).

The continuous mapping theorem also applies to a continuous functional $g(\cdot)$ that maps a continuous bounded function on [0, 1] into another continuous bounded function on [0, 1]. For example, the function whose value at r is a positive constant σ times $h(r)$ represents the result of applying the continuous functional $g[h(\cdot)] \equiv \sigma \cdot h(\cdot)$ to $h(\cdot)$.[7] Thus, it follows from [17.3.6] that

$$\sqrt{T} \cdot X_T(\cdot) \xrightarrow{L} \sigma \cdot W(\cdot). \qquad [17.3.8]$$

Recalling that $W(r) \sim N(0, r)$, result [17.3.8] implies that $\sqrt{T} \cdot X_T(r) \approx N(0, \sigma^2 r)$.

As another example, consider the function $S_T(\cdot)$ whose value at r is given by

$$S_T(r) \equiv [\sqrt{T} \cdot X_T(r)]^2. \qquad [17.3.9]$$

Since $\sqrt{T} \cdot X_T(\cdot) \xrightarrow{L} \sigma \cdot W(\cdot)$, it follows that

$$S_T(\cdot) \xrightarrow{L} \sigma^2 [W(\cdot)]^2. \qquad [17.3.10]$$

In other words, if the value $W(r)$ from a realization of standard Brownian motion at every date r is squared and then multiplied by σ^2, the resulting continuous-time process would follow essentially the same probability law as does the continuous-time process defined by $S_T(r)$ in [17.3.9] for T sufficiently large.

Applications to Unit Root Processes

The use of the functional central limit theorem to calculate the asymptotic distribution of statistics constructed from unit root processes was pioneered by Phillips (1986, 1987).[8] The simplest illustration of Phillips's approach is provided by a random walk,

$$y_t = y_{t-1} + u_t, \qquad [17.3.11]$$

where $\{u_t\}$ is an i.i.d. sequence with mean zero and variance σ^2. If $y_0 = 0$, then [17.3.11] implies that

$$y_t = u_1 + u_2 + \cdots + u_t. \qquad [17.3.12]$$

Equation [17.3.12] can be used to express the stochastic function $X_T(r)$ defined in [17.3.3] as

$$X_T(r) = \begin{cases} 0 & \text{for } 0 \le r < 1/T \\ y_1/T & \text{for } 1/T \le r < 2/T \\ y_2/T & \text{for } 2/T \le r < 3/T \\ \vdots & \\ y_T/T & \text{for } r = 1. \end{cases} \qquad [17.3.13]$$

Figure 17.1 plots $X_T(r)$ as a function of r. Note that the area under this step function

[7]Here continuity of the functional $g(\cdot)$ means that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that if $h(r)$ and $k(r)$ are any continuous bounded functions on [0, 1], h: [0, 1] → $\mathbb{R}^1$ and k: [0, 1] → $\mathbb{R}^1$, such that $|h(r) - k(r)| < \delta$ for all $r \in [0, 1]$, then

$$|g[h(r)] - g[k(r)]| < \varepsilon$$

for all $r \in [0, 1]$.

[8]Result [17.4.7] in the next section for the case with i.i.d. errors was first derived by White (1958). Phillips (1986, 1987) developed the general derivation presented here based on the functional central limit theorem and the continuous mapping theorem. Other important contributions include Dickey and Fuller (1979), Chan and Wei (1988), Park and Phillips (1988, 1989), Sims, Stock, and Watson (1990), and Phillips and Solo (1992).

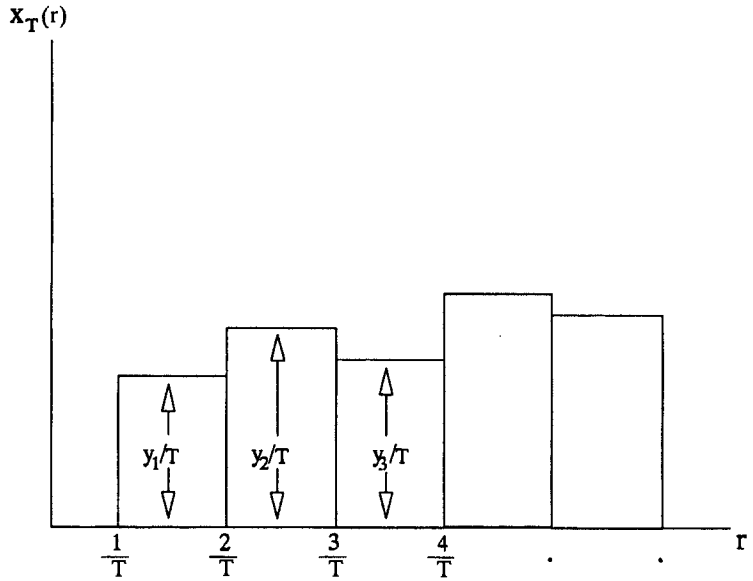


FIGURE 17.1 Plot of $X_T(r)$ as a function of r.

is the sum of T rectangles. The tth rectangle has width $1/T$ and height y_{t-1}/T, and therefore has area y_{t-1}/T^2. The integral of $X_T(r)$ is thus equivalent to

$$\int_0^1 X_T(r)\, dr = y_1/T^2 + y_2/T^2 + \cdots + y_{T-1}/T^2. \qquad [17.3.14]$$

Multiplying both sides of [17.3.14] by $\sqrt{T}$ establishes that

$$\int_0^1 \sqrt{T}\cdot X_T(r)\, dr = T^{-3/2} \sum_{t=1}^{T} y_{t-1}. \qquad [17.3.15]$$

But we know from [17.3.8] and the continuous mapping theorem that as $T \to \infty$,

$$\int_0^1 \sqrt{T}\cdot X_T(r)\, dr \xrightarrow{L} \sigma\cdot \int_0^1 W(r)\, dr,$$

implying from [17.3.15] that

$$T^{-3/2} \sum_{t=1}^{T} y_{t-1} \xrightarrow{L} \sigma\cdot \int_0^1 W(r)\, dr. \qquad [17.3.16]$$

It is also instructive to derive [17.3.16] from first principles. From [17.3.12], we can write

$$\begin{aligned} T^{-3/2} \sum_{t=1}^{T} y_{t-1} &= T^{-3/2}[u_1 + (u_1 + u_2) + (u_1 + u_2 + u_3) + \cdots \\ &\quad + (u_1 + u_2 + u_3 + \cdots + u_{T-1})] \\ &= T^{-3/2}[(T-1)u_1 + (T-2)u_2 + (T-3)u_3 + \cdots \\ &\quad + [T - (T-1)]u_{T-1}] \qquad [17.3.17] \\ &= T^{-3/2} \sum_{t=1}^{T} (T-t)u_t \\ &= T^{-1/2} \sum_{t=1}^{T} u_t - T^{-3/2} \sum_{t=1}^{T} tu_t. \end{aligned}$$

Recall from [16.1.24] that

$$\begin{bmatrix} T^{-1/2}\sum_{t=1}^{T} u_t \\ T^{-3/2}\sum_{t=1}^{T} tu_t \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{3} \end{bmatrix}\right). \qquad [17.3.18]$$

Thus, [17.3.17] implies that $T^{-3/2}\sum_{t=1}^{T} y_{t-1}$ is asymptotically Gaussian with mean zero and variance equal to

$$\sigma^2\{1 - 2\cdot(1/2) + 1/3\} = \sigma^2/3.$$

Evidently, $\sigma\cdot\int_0^1 W(r)\,dr$ in [17.3.16] describes a random variable that has a $N(0, \sigma^2/3)$ distribution.

Thus, if y_t is a driftless random walk, the sample mean $T^{-1}\sum_{t=1}^{T} y_t$ diverges but $T^{-3/2}\sum_{t=1}^{T} y_t$ converges to a Gaussian random variable whose distribution can be described as the integral of the realization of Brownian motion with variance σ^2.

Expression [17.3.17] also gives us a way to describe the asymptotic distribution of $T^{-3/2}\sum_{t=1}^{T} tu_t$ in terms of functionals on Brownian motion:

$$\begin{aligned} T^{-3/2}\sum_{t=1}^{T} tu_t &= T^{-1/2}\sum_{t=1}^{T} u_t - T^{-3/2}\sum_{t=1}^{T} y_{t-1} \\ &\xrightarrow{L} \sigma\cdot W(1) - \sigma\cdot\int_0^1 W(r)\,dr, \end{aligned} \qquad [17.3.19]$$

with the last line following from [17.3.7] and [17.3.16]. Recalling [17.3.18], the random variable on the right side of [17.3.19] evidently has a $N(0, \sigma^2/3)$ distribution.

A similar argument to that in [17.3.15] can be used to describe the asymptotic distribution of the sum of squares of a random walk. The statistic $S_T(r)$ defined in [17.3.9],

$$S_T(r) \equiv T\cdot[X_T(r)]^2, \qquad [17.3.20]$$

can be written using [17.3.13] as

$$S_T(r) = \begin{cases} 0 & \text{for } 0 \le r < 1/T \\ y_1^2/T & \text{for } 1/T \le r < 2/T \\ y_2^2/T & \text{for } 2/T \le r < 3/T \\ \vdots & \\ y_T^2/T & \text{for } r = 1. \end{cases} \qquad [17.3.21]$$

It follows that

$$\int_0^1 S_T(r)\,dr = y_1^2/T^2 + y_2^2/T^2 + \cdots + y_{T-1}^2/T^2.$$

Thus, from [17.3.10] and the continuous mapping theorem,

$$T^{-2}\sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{L} \sigma^2\cdot\int_0^1 [W(r)]^2\,dr. \qquad [17.3.22]$$

Two other useful results are

$$T^{-5/2}\sum_{t=1}^{T} ty_{t-1} = T^{-3/2}\sum_{t=1}^{T} (t/T)y_{t-1} \xrightarrow{L} \sigma\cdot\int_0^1 rW(r)\,dr \qquad [17.3.23]$$

for $r = t/T$ and

$$T^{-3}\sum_{t=1}^{T} ty_{t-1}^2 = T^{-2}\sum_{t=1}^{T} (t/T)y_{t-1}^2 \xrightarrow{L} \sigma^2\cdot\int_0^1 r\cdot[W(r)]^2\,dr. \qquad [17.3.24]$$

As yet another useful application, consider the statistic in [17.1.11]:

$$T^{-1}\sum_{t=1}^{T} y_{t-1}u_t = (1/2)\cdot(1/T)y_T^2 - (1/2)\cdot(1/T)\sum_{t=1}^{T} u_t^2.$$

Recalling [17.3.21], this can be written

$$T^{-1}\sum_{t=1}^{T} y_{t-1}u_t = (1/2)S_T(1) - (1/2)(1/T)\sum_{t=1}^{T} u_t^2. \qquad [17.3.25]$$

But $(1/T)\Sigma_{t=1}^{T} u_t^2 \xrightarrow{p} \sigma^2$, by the law of large numbers, and $S_T(1) \xrightarrow{L} \sigma^2[W(1)]^2$, by [17.3.10]. It thus follows from [17.3.25] that

$$T^{-1}\sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{L} (1/2)\sigma^2[W(1)]^2 - (1/2)\sigma^2. \qquad [17.3.26]$$

Recall that $W(1)$, the value of standard Brownian motion at date $r = 1$, has a $N(0, 1)$ distribution, meaning that $[W(1)]^2$ has a $\chi^2(1)$ distribution. Result [17.3.26] is therefore just another way to express the earlier result [17.1.15] using a functional on Brownian motion instead of the χ^2 distribution.

17.4. *Asymptotic Properties of a First-Order Autoregression when the True Coefficient Is Unity*

We are now in a position to calculate the asymptotic distribution of some simple regressions involving unit roots. For convenience, the results from Section 17.3 are collected in the form of a proposition.

Proposition 17.1: *Suppose that ξ_t follows a random walk without drift,*

$$\xi_t = \xi_{t-1} + u_t,$$

where $\xi_0 = 0$ and $\{u_t\}$ is an i.i.d. sequence with mean zero and variance σ^2. Then

(*a*) $T^{-1/2}\sum_{t=1}^{T} u_t \xrightarrow{L} \sigma\cdot W(1)$ [17.3.7];

(*b*) $T^{-1}\sum_{t=1}^{T} \xi_{t-1}u_t \xrightarrow{L} (1/2)\sigma^2\{[W(1)]^2 - 1\}$ [17.3.26];

(*c*) $T^{-3/2}\sum_{t=1}^{T} tu_t \xrightarrow{L} \sigma\cdot W(1) - \sigma\cdot\int_0^1 W(r)\,dr$ [17.3.19];

(*d*) $T^{-3/2}\sum_{t=1}^{T} \xi_{t-1} \xrightarrow{L} \sigma\cdot\int_0^1 W(r)\,dr$ [17.3.16];

(*e*) $T^{-2}\sum_{t=1}^{T} \xi_{t-1}^2 \xrightarrow{L} \sigma^2\cdot\int_0^1 [W(r)]^2\,dr$ [17.3.22];

(*f*) $T^{-5/2}\sum_{t=1}^{T} t\xi_{t-1} \xrightarrow{L} \sigma\cdot\int_0^1 rW(r)\,dr$ [17.3.23];

(*g*) $T^{-3}\sum_{t=1}^{T} t\xi_{t-1}^2 \xrightarrow{L} \sigma^2\cdot\int_0^1 r\cdot[W(r)]^2\,dr$ [17.3.24];

(*h*) $T^{-(\nu+1)}\sum_{t=1}^{T} t^{\nu} \to 1/(\nu + 1)$ *for* $\nu = 0, 1, \ldots$ [16.1.15].

The expressions in brackets indicate where the stated result was earlier derived. Though the earlier derivations assumed that the initial value ξ_0 was equal to zero, the same results are obtained when ξ_0 is any fixed value or drawn from a specified distribution as in Phillips (1987).

The asymptotic distributions in Proposition 17.1 are all written in terms of functionals on standard Brownian motion, denoted $W(r)$. Note that this is the *same* Brownian motion $W(r)$ in each result (a) through (g), so that in general the magnitudes in Proposition 17.1 are all correlated. If we are not interested in capturing these correlations, then there are simpler ways to describe the asymptotic distributions. For example, we have seen that (a) is just a $N(0, \sigma^2)$ distribution, (b) is $(1/2)\sigma^2 \cdot [\chi^2(1) - 1]$, and (c) and (d) are $N(0, \sigma^2/3)$. Exercise 17.1 gives an example of one approach to calculating the covariances among random variables described by these functionals on Brownian motion.

Proposition 17.1 can be used to calculate the asymptotic distributions of statistics from a number of simple regressions involving unit roots. This section discusses several key cases.

Case 1. No Constant Term or Time Trend in the Regression; True Process Is a Random Walk

Consider first *OLS* estimation of ρ based on an $AR(1)$ regression,

$$y_t = \rho y_{t-1} + u_t, \qquad [17.4.1]$$

where u_t is i.i.d. with mean zero and variance σ^2. We are interested in the properties of the *OLS* estimate

$$\hat{\rho}_T = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2} \qquad [17.4.2]$$

when the true value of ρ is unity. From [17.1.6], the deviation of the *OLS* estimate from the true value is characterized by

$$T(\hat{\rho}_T - 1) = \frac{T^{-1} \sum_{t=1}^{T} y_{t-1} u_t}{T^{-2} \sum_{t=1}^{T} y_{t-1}^2}. \qquad [17.4.3]$$

If the true value of ρ is unity, then

$$y_t = y_0 + u_1 + u_2 + \cdots + u_t. \qquad [17.4.4]$$

Apart from the initial term y_0 (which does not affect any of the asymptotic distributions), the variable y_t is the same as the quantity labeled ξ_t in Proposition 17.1. From result (b) of that proposition,

$$T^{-1} \sum_{t=1}^{T} y_{t-1} u_t \xrightarrow{L} (1/2)\sigma^2\{[W(1)]^2 - 1\}, \qquad [17.4.5]$$

while from result (e),

$$T^{-2} \sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{L} \sigma^2 \cdot \int_0^1 [W(r)]^2 \, dr. \qquad [17.4.6]$$

Since [17.4.3] is a continuous function of [17.4.5] and [17.4.6], it follows from Proposition 7.3(c) that under the null hypothesis that $\rho = 1$, the *OLS* estimate

$\hat{\rho}_T$ is characterized by

$$T(\hat{\rho}_T - 1) \xrightarrow{L} \frac{(1/2)\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2\, dr}. \qquad [17.4.7]$$

Recall that $[W(1)]^2$ is a $\chi^2(1)$ variable. The probability that a $\chi^2(1)$ variable is less than unity is 0.68, and since the denominator of [17.4.7] must be positive, the probability that $\hat{\rho}_T - 1$ is negative approaches 0.68 as T becomes large. In other words, in two-thirds of the samples generated by a random walk, the estimate $\hat{\rho}_T$ will be less than the true value of unity. Moreover, in those samples for which $[W(1)]^2$ is large, the denominator of [17.4.7] will be large as well. The result is that the limiting distribution of $T(\hat{\rho}_T - 1)$ is skewed to the left.

Recall that in the stationary case when $|\rho| < 1$, the estimate $\hat{\rho}_T$ is downward-biased in small samples. Even so, in the stationary case the limiting distribution of $\sqrt{T}(\hat{\rho}_T - \rho)$ is symmetric around zero. By contrast, when the true value of ρ is unity, even the limiting distribution of $T(\hat{\rho}_T - 1)$ is asymmetric, with negative values twice as likely as positive values.

In practice, critical values for the random variable in [17.4.7] are found by calculating the exact small-sample distribution of $T(\hat{\rho}_T - 1)$ for given T, assuming that the innovations $\{u_t\}$ are Gaussian. This can be done either by Monte Carlo, as in the critical values reported in Fuller (1976), or by using exact numerical procedures described in Evans and Savin (1981). Sample percentiles for $T(\hat{\rho}_T - 1)$ are reported in the section labeled Case 1 in Table B.5 of Appendix B. For finite T, these are exact only under the assumption of Gaussian innovations. As T becomes large, these values also describe the asymptotic distribution for non-Gaussian innovations.

It follows from [17.4.7] that $\hat{\rho}_T$ is a superconsistent estimate of the true value ($\rho = 1$). This is easily seen by dividing [17.4.3] by $\sqrt{T}$:

$$\sqrt{T}(\hat{\rho}_T - 1) = \frac{T^{-3/2} \sum_{t=1}^{T} y_{t-1} u_t}{T^{-2} \sum_{t=1}^{T} y_{t-1}^2}. \qquad [17.4.8]$$

From Proposition 17.1(b), the numerator in [17.4.8] converges to $T^{-1/2}(1/2)\sigma^2$ times $(X - 1)$, where X is a $\chi^2(1)$ random variable. Since a $\chi^2(1)$ variable has finite variance, the variance of the numerator in [17.4.8] is of order $1/T$, meaning that the numerator converges in probability to zero. Hence,

$$\sqrt{T}(\hat{\rho}_T - 1) \xrightarrow{p} 0.$$

Result [17.4.7] allows the point estimate $\hat{\rho}_T$ to be used by itself to test the null hypothesis of a unit root, without needing to calculate its standard error. Another popular statistic for testing the null hypothesis that $\rho = 1$ is based on the usual *OLS* t test of this hypothesis,

$$t_T = \frac{(\hat{\rho}_T - 1)}{\hat{\sigma}_{\hat{\rho}_T}} = \frac{(\hat{\rho}_T - 1)}{\left\{ s_T^2 \div \sum_{t=1}^{T} y_{t-1}^2 \right\}^{1/2}}, \qquad [17.4.9]$$

where $\hat{\sigma}_{\hat{\rho}_T}$ is the usual *OLS* standard error for the estimated coefficient,

$$\hat{\sigma}_{\hat{\rho}_T} = \left\{ s_T^2 \div \sum_{t=1}^{T} y_{t-1}^2 \right\}^{1/2},$$

and s_T^2 denotes the OLS estimate of the residual variance:

$$s_T^2 = \sum_{t=1}^{T} (y_t - \hat{\rho}_T y_{t-1})^2/(T - 1).$$

Although the t statistic [17.4.9] is calculated in the usual way, it does not have a limiting Gaussian distribution when the true process is characterized by $\rho = 1$. To find the appropriate limiting distribution, note that [17.4.9] can equivalently be expressed as

$$t_T = T(\hat{\rho}_T - 1)\left\{T^{-2} \sum_{t=1}^{T} y_{t-1}^2\right\}^{1/2} \div \{s_T^2\}^{1/2}, \qquad [17.4.10]$$

or, substituting from [17.4.3],

$$t_T = \frac{T^{-1} \sum_{t=1}^{T} y_{t-1} u_t}{\left\{T^{-2} \sum_{t=1}^{T} y_{t-1}^2\right\}^{1/2} \{s_T^2\}^{1/2}}. \qquad [17.4.11]$$

As in Section 8.2, consistency of $\hat{\rho}_T$ implies $s_T^2 \xrightarrow{p} \sigma^2$. It follows from [17.4.5] and [17.4.6] that as $T \to \infty$,

$$t_T \xrightarrow{L} \frac{(1/2)\sigma^2\{[W(1)]^2 - 1\}}{\left\{\sigma^2 \int_0^1 [W(r)]^2\, dr\right\}^{1/2} \{\sigma^2\}^{1/2}} = \frac{(1/2)\{[W(1)]^2 - 1\}}{\left\{\int_0^1 [W(r)]^2\, dr\right\}^{1/2}}. \qquad [17.4.12]$$

Statistical tables for the distribution of [17.4.11] for various sample sizes T are reported in the section labeled Case 1 in Table B.6; again, the small-sample results assume Gaussian innovations.

Example 17.3

The following $AR(1)$ process for the nominal three-month U.S. Treasury bill rate was fitted by OLS regression to quarterly data, t = 1947:II to 1989:I:

$$i_t = \underset{(0.010592)}{0.99694}\, i_{t-1}, \qquad [17.4.13]$$

with the standard error of $\hat{\rho}$ in parentheses. Here $T = 168$ and

$$T(\hat{\rho} - 1) = (168)(0.99694 - 1) = -0.51.$$

The distribution of this statistic was calculated in [17.4.7] under the assumption that the true value of ρ is unity. The null hypothesis is therefore that $\rho = 1$, and the alternative is that $\rho < 1$. From Table B.5, in a sample of this size, 95% of the time when there really is a unit root, the statistic $T(\hat{\rho} - 1)$ will be above -7.9. The observed value (-0.51) is well above this, and so the null hypothesis is accepted at the 5% level and we should conclude that these data might well be described by a random walk.

In order to have rejected the null hypothesis for a sample of this size, the estimated autoregressive coefficient $\hat{\rho}$ would have to be less than 0.95:

$$168(0.95 - 1) = -8.4.$$

The OLS t test of H_0: $\rho = 1$ is

$$t = (0.99694 - 1)/0.010592 = -0.29.$$

This is well above the 5% critical value from Table B.6 of -1.95, so the null

hypothesis that the Treasury bill rate follows a random walk is also accepted by this test.

The test statistics [17.4.7] and [17.4.12] are examples of the *Dickey-Fuller test* for unit roots, named for the general battery of tests proposed by Dickey and Fuller (1979).

Case 2. Constant Term but No Time Trend Included in the Regression; True Process Is a Random Walk

For case 2, we continue to assume, as in case 1, that the data are generated by a random walk:

$$y_t = y_{t-1} + u_t,$$

with u_t i.i.d. with mean zero and variance σ^2. Although the true model is the same as in case 1, suppose now that a constant term is included in the $AR(1)$ specification that is to be estimated by OLS:

$$y_t = \alpha + \rho y_{t-1} + u_t. \qquad [17.4.14]$$

The task now is to describe the properties of the OLS estimates,

$$\begin{bmatrix} \hat{\alpha}_T \\ \hat{\rho}_T \end{bmatrix} = \begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y_t \\ \Sigma y_{t-1} y_t \end{bmatrix}, \qquad [17.4.15]$$

under the null hypothesis that $\alpha = 0$ and $\rho = 1$ (here Σ indicates summation over $t = 1, 2, \ldots, T$). Recall the familiar characterization in [8.2.3] of the deviation of an estimated OLS coefficient vector ($\mathbf{b}_T$) from the true value ($\boldsymbol{\beta}$),

$$\mathbf{b}_T - \boldsymbol{\beta} = \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t u_t\right], \qquad [17.4.16]$$

or, in this case,

$$\begin{bmatrix} \hat{\alpha}_T \\ \hat{\rho}_T - 1 \end{bmatrix} = \begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma u_t \\ \Sigma y_{t-1} u_t \end{bmatrix}. \qquad [17.4.17]$$

As in case 1, y_t has the same properties as the variable ξ_t described in Proposition 17.1 under the maintained hypothesis. Thus, result (d) of that proposition establishes that the sum Σy_{t-1} must be divided by $T^{3/2}$ before obtaining a random variable that converges in distribution:

$$T^{-3/2} \Sigma y_{t-1} \xrightarrow{L} \sigma \cdot \int_0^1 W(r)\, dr. \qquad [17.4.18]$$

In other words,

$$\Sigma y_{t-1} = O_p(T^{3/2}).$$

Similarly, results [17.4.5] and [17.4.6] establish that

$$\Sigma y_{t-1} u_t = O_p(T)$$
$$\Sigma y_{t-1}^2 = O_p(T^2),$$

and from Proposition 17.1(a),

$$\Sigma u_t = O_p(T^{1/2}).$$

Thus, the order in probability of the individual terms in [17.4.17] is as follows:

$$\begin{bmatrix} \hat{\alpha}_T \\ \hat{\rho}_T - 1 \end{bmatrix} = \begin{bmatrix} O_p(T) & O_p(T^{3/2}) \\ O_p(T^{3/2}) & O_p(T^2) \end{bmatrix}^{-1} \begin{bmatrix} O_p(T^{1/2}) \\ O_p(T) \end{bmatrix}. \qquad [17.4.19]$$

It is clear from [17.4.19] that the estimates $\hat{\alpha}_T$ and $\hat{\rho}_T$ have different rates of convergence, and as in the previous chapter, a scaling matrix $\mathbf{Y}_T$ is helpful in describing their limiting distributions. Recall from [16.1.18] that this rescaling is achieved by premultiplying [17.4.16] by $\mathbf{Y}_T$ and writing the result as

$$\begin{aligned} \mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}) &= \mathbf{Y}_T\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \mathbf{Y}_T\mathbf{Y}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t u_t\right] \\ &= \left\{\mathbf{Y}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]\mathbf{Y}_T^{-1}\right\}^{-1}\left\{\mathbf{Y}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t u_t\right]\right\}. \end{aligned} \qquad [17.4.20]$$

From [17.4.19], for this application $\mathbf{Y}_T$ should be specified to be the following matrix:

$$\mathbf{Y}_T \equiv \begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}, \qquad [17.4.21]$$

for which [17.4.20] becomes

$$\begin{aligned} \begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}\begin{bmatrix} \hat{\alpha}_T \\ \hat{\rho}_T - 1 \end{bmatrix} &= \left\{\begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-1} \end{bmatrix}\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}\begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-1} \end{bmatrix}\right\}^{-1} \\ &\quad \times \left\{\begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-1} \end{bmatrix}\begin{bmatrix} \Sigma u_t \\ \Sigma y_{t-1}u_t \end{bmatrix}\right\} \end{aligned}$$

or

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T \\ T(\hat{\rho}_T - 1) \end{bmatrix} = \begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-1}\Sigma y_{t-1}u_t \end{bmatrix}. \qquad [17.4.22]$$

Consider the first term on the right side of [17.4.22]. Results [17.4.6] and [17.4.18] establish that

$$\begin{aligned} &\begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y_{t-1}^2 \end{bmatrix} \\ &\xrightarrow{L} \begin{bmatrix} 1 & \sigma\cdot\int W(r)\,dr \\ \sigma\cdot\int W(r)\,dr & \sigma^2\cdot\int [W(r)]^2\,dr \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}\begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}, \end{aligned} \qquad [17.4.23]$$

where the integral sign denotes integration over r from 0 to 1. Similarly, result (a) of Proposition 17.1 along with [17.4.5] determines the asymptotic distribution of

the second term in [17.4.22]:

$$\begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-1}\Sigma y_{t-1}u_t \end{bmatrix} \xrightarrow{L} \begin{bmatrix} \sigma \cdot W(1) \\ (1/2)\sigma^2\{[W(1)]^2 - 1\} \end{bmatrix}$$
$$= \sigma \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} W(1) \\ (1/2)\{[W(1)]^2 - 1\} \end{bmatrix}. \qquad [17.4.24]$$

Substituting [17.4.23] and [17.4.24] into [17.4.22] establishes

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T \\ T(\hat{\rho}_T - 1) \end{bmatrix} \xrightarrow{L} \sigma \cdot \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1}$$
$$\times \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix} \begin{bmatrix} W(1) \\ (1/2)\{[W(1)]^2 - 1\} \end{bmatrix}$$
$$= \begin{bmatrix} \sigma & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \qquad [17.4.25]$$
$$\times \begin{bmatrix} W(1) \\ (1/2)\{[W(1)]^2 - 1\} \end{bmatrix}.$$

Notice that

$$\begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} = \Delta^{-1} \begin{bmatrix} \int [W(r)]^2\,dr & -\int W(r)\,dr \\ -\int W(r)\,dr & 1 \end{bmatrix}, \qquad [17.4.26]$$

where

$$\Delta \equiv \int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2. \qquad [17.4.27]$$

Thus, the second element in the vector expression in [17.4.25] states that

$$T(\hat{\rho}_T - 1) \xrightarrow{L} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1)\cdot\int W(r)\,dr}{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2}. \qquad [17.4.28]$$

Neither estimate $\hat{\alpha}_T$ nor $\hat{\rho}_T$ has a limiting Gaussian distribution. Moreover, the asymptotic distribution of the estimate of ρ in [17.4.28] is not the same as the asymptotic distribution in [17.4.7]—when a constant term is included in the distribution, a different table of critical values must be used.

The second section of Table B.5 records percentiles for the distribution of $T(\hat{\rho}_T - 1)$ for case 2. As in case 1, the calculations assume Gaussian innovations, though as T becomes large, these are valid for non-Gaussian innovations as well.

Notice that this distribution is even more strongly skewed than that for case 1, so that when a constant term is included in the regression, the estimated coefficient on y_{t-1} must be farther from unity in order to reject the null hypothesis of a unit root. Indeed, for $T > 25$, 95% of the time the estimated value $\hat{\rho}_T$ will be less than unity. For example, if the estimated value $\hat{\rho}_T$ is 0.999 in a sample of size $T = 100$, the null hypothesis of $\rho = 1$ would be rejected in favor of the alternative that $\rho > 1$! If the true value of ρ is unity, we would not expect to obtain an estimate as large as 0.999.

Dickey and Fuller also proposed an alternative test based on the *OLS t* test of the null hypothesis that $\rho = 1$:

$$t_T = \frac{\hat{\rho}_T - 1}{\hat{\sigma}_{\hat{\rho}_T}}, \qquad [17.4.29]$$

where

$$\hat{\sigma}^2_{\hat{\rho}_T} = s_T^2[0 \quad 1]\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad [17.4.30]$$

$$s_T^2 = (T-2)^{-1}\sum_{t=1}^{T}(y_t - \hat{\alpha}_T - \hat{\rho}_T y_{t-1})^2.$$

Notice that if both sides of [17.4.30] are multiplied by T^2, the result can be written as

$$\begin{aligned} T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} &= s_T^2[0 \quad T]\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ T \end{bmatrix} \\ &= s_T^2[0 \quad 1]\mathbf{Y}_T\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1}\mathbf{Y}_T\begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{aligned} \qquad [17.4.31]$$

for $\mathbf{Y}_T$ the matrix in [17.4.21]. Recall from [17.4.23] that

$$\begin{aligned} &\mathbf{Y}_T\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1}\mathbf{Y}_T \\ &= \left\{\mathbf{Y}_T^{-1}\begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}\mathbf{Y}_T^{-1}\right\}^{-1} \\ &= \begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y_{t-1}^2 \end{bmatrix}^{-1} \\ &\xrightarrow{L} \begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}^{-1}\begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1}\begin{bmatrix} 1 & 0 \\ 0 & \sigma \end{bmatrix}^{-1}. \end{aligned} \qquad [17.4.32]$$

Thus, from [17.4.31],

$$T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \xrightarrow{p} s_T^2[0 \quad \sigma^{-1}]\begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ \sigma^{-1} \end{bmatrix}. \qquad [17.4.33]$$

It is also easy to show that

$$s_T^2 \xrightarrow{p} \sigma^2, \qquad [17.4.34]$$

from which [17.4.33] becomes

$$T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \xrightarrow{L} [0 \quad 1] \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad [17.4.35]$$

$$= \frac{1}{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2}.$$

Thus, the asymptotic distribution of the *OLS* t test in [17.4.29] is

$$t_T = \frac{T(\hat{\rho}_T - 1)}{\{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T}\}^{1/2}} \xrightarrow{p} T(\hat{\rho}_T - 1) \times \left\{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2\right\}^{1/2}$$

$$\xrightarrow{L} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1) \cdot \int W(r)\,dr}{\left\{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2\right\}^{1/2}}. \qquad [17.4.36]$$

Sample percentiles for the *OLS* t test of $\rho = 1$ are reported for case 2 in the second section of Table B.6. As T grows large, these approach the distribution in the last line of [17.4.36].

Example 17.4

When a constant term is included in the estimated autoregression for the interest rate data from Example 17.3, the result is

$$i_t = \underset{(0.112)}{0.211} + \underset{(0.019133)}{0.96691}\, i_{t-1}, \qquad [17.4.37]$$

with standard errors reported in parentheses. The Dickey-Fuller test based on the estimated value of ρ for this specification is

$$T(\hat{\rho} - 1) = (168)(0.96691 - 1) = -5.56.$$

From Table B.5, the 5% critical value is found by interpolation to be -13.8. Since $-5.56 > -13.8$, the null hypothesis of a unit root ($\rho = 1$) is accepted at the 5% level based on the Dickey-Fuller $\hat{\rho}$ test. The *OLS* t statistic is

$$(0.96691 - 1)/0.019133 = -1.73,$$

which from Table B.6 is to be compared with -2.89. Since $-1.73 > -2.89$, the null hypothesis of a unit root is again accepted.

These statistics test the null hypothesis that $\rho = 1$. However, a maintained assumption on which the derivation of [17.4.25] was based is that the true value of α is zero. Thus, it might seem more natural to test for a unit root in this specification by testing the joint hypothesis that $\alpha = 0$ and $\rho = 1$. Dickey and Fuller (1981) used Monte Carlo to calculate the distribution of the Wald form of the *OLS* F test of this hypothesis (expression [8.1.32] or [8.1.37]). Their values are reported under the heading Case 2 in Table B.7.

Example 17.5

The *OLS* Wald F statistic for testing the joint hypothesis that $\alpha = 0$ and $\rho = 1$ for the regression in [17.4.37] is 1.81. Under the classical regression assump-

tions, this would have an $F(2, 166)$ distribution. In this case, however, the usual statistic is to be compared with the values under Case 2 in Table B.7, for which the 5% critical value is found by interpolation to be 4.67. Since $1.81 < 4.67$, the joint null hypothesis that $\alpha = 0$ and $\rho = 1$ is accepted at the 5% level.

Case 3. Constant Term but No Time Trend Included in the Regression; True Process Is Random Walk with Drift

In case 3, the same regression [17.4.14] is estimated as in case 2, though now it is supposed that the true process is a random walk with drift:

$$y_t = \alpha + y_{t-1} + u_t, \qquad [17.4.38]$$

where the true value of α is not zero. Although this might seem like a minor change, it has a radical effect on the asymptotic distribution of $\hat{\alpha}$ and $\hat{\rho}$. To see why, note that [17.4.38] implies that

$$y_t = y_0 + \alpha t + (u_1 + u_2 + \cdots + u_t) = y_0 + \alpha t + \xi_t, \qquad [17.4.39]$$

where

$$\xi_t \equiv u_1 + u_2 + \cdots + u_t \qquad \text{for } t = 1, 2, \ldots, T$$

with $\xi_0 \equiv 0$.

Consider the behavior of the sum

$$\sum_{t=1}^{T} y_{t-1} = \sum_{t=1}^{T} [y_0 + \alpha(t-1) + \xi_{t-1}]. \qquad [17.4.40]$$

The first term in [17.4.40] is just Ty_0, and if this is divided by T, the result will be a fixed value. The second term, $\Sigma\alpha(t-1)$, must be divided by T^2 in order to converge:

$$T^{-2} \sum_{t=1}^{T} \alpha(t-1) \to \alpha/2,$$

by virtue of Proposition 17.1(h). The third term converges when divided by $T^{3/2}$:

$$T^{-3/2} \sum_{t=1}^{T} \xi_{t-1} \xrightarrow{L} \sigma \cdot \int_0^1 W(r)\, dr,$$

from Proposition 17.1(d). The order in probability of the three individual terms in [17.4.40] is thus

$$\sum_{t=1}^{T} y_{t-1} = \underbrace{\sum_{t=1}^{T} y_0}_{O_p(T)} + \underbrace{\sum_{t=1}^{T} \alpha(t-1)}_{O_p(T^2)} + \underbrace{\sum_{t=1}^{T} \xi_{t-1}}_{O_p(T^{3/2})}.$$

The time trend $\alpha(t-1)$ asymptotically dominates the other two components:

$$T^{-2} \sum_{t=1}^{T} y_{t-1} = T^{-1}y_0 + T^{-2} \sum_{t=1}^{T} \alpha(t-1) + T^{-1/2}\left\{T^{-3/2} \sum_{t=1}^{T} \xi_{t-1}\right\} \qquad [17.4.41]$$
$$\xrightarrow{p} 0 + \alpha/2 + 0.$$

Similarly, we have that

$$\sum_{t=1}^{T} y_{t-1}^2 = \sum_{t=1}^{T} [y_0 + \alpha(t-1) + \xi_{t-1}]^2$$

$$= \underbrace{\sum_{t=1}^{T} y_0^2}_{O_p(T)} + \underbrace{\sum_{t=1}^{T} \alpha^2(t-1)^2}_{O_p(T^3)} + \underbrace{\sum_{t=1}^{T} \xi_{t-1}^2}_{O_p(T^2)}$$

$$+ \underbrace{\sum_{t=1}^{T} 2y_0\alpha(t-1)}_{O_p(T^2)} + \underbrace{\sum_{t=1}^{T} 2y_0\xi_{t-1}}_{O_p(T^{3/2})} + \underbrace{\sum_{t=1}^{T} 2\alpha(t-1)\xi_{t-1}}_{O_p(T^{5/2})}.$$

When divided by T^3, the only term that does not vanish asymptotically is that due to the time trend $\alpha^2(t-1)^2$:

$$T^{-3} \sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} \alpha^2/3. \qquad [17.4.42]$$

Finally, observe that

$$\sum_{t=1}^{T} y_{t-1}u_t = \sum_{t=1}^{T} [y_0 + \alpha(t-1) + \xi_{t-1}]u_t$$

$$= \underbrace{y_0 \sum_{t=1}^{T} u_t}_{O_p(T^{1/2})} + \underbrace{\sum_{t=1}^{T} \alpha(t-1)u_t}_{O_p(T^{3/2})} + \underbrace{\sum_{t=1}^{T} \xi_{t-1}u_t}_{O_p(T)},$$

from which

$$T^{-3/2} \sum_{t=1}^{T} y_{t-1}u_t \xrightarrow{p} T^{-3/2} \sum_{t=1}^{T} \alpha(t-1)u_t. \qquad [17.4.43]$$

Results [17.4.41] through [17.4.43] imply that when the true process is a random walk with drift, the estimated *OLS* coefficients in [17.4.15] satisfy

$$\begin{bmatrix} \hat{\alpha}_T - \alpha \\ \hat{\rho}_T - 1 \end{bmatrix} = \begin{bmatrix} O_p(T) & O_p(T^2) \\ O_p(T^2) & O_p(T^3) \end{bmatrix}^{-1} \begin{bmatrix} O_p(T^{1/2}) \\ O_p(T^{3/2}) \end{bmatrix}.$$

Thus, for this case, the Sims, Stock, and Watson scaling matrix would be

$$\mathbf{Y}_T \equiv \begin{bmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{bmatrix},$$

for which [17.4.20] becomes

$$\begin{bmatrix} T^{1/2} & 0 \\ 0 & T^{3/2} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_T - \alpha \\ \hat{\rho}_T - 1 \end{bmatrix}$$

$$= \left\{ \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix} \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \right\}^{-1}$$

$$\times \left\{ \begin{bmatrix} T^{-1/2} & 0 \\ 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} \Sigma u_t \\ \Sigma y_{t-1}u_t \end{bmatrix} \right\}$$

or

$$\begin{bmatrix} T^{1/2}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\rho}_T - 1) \end{bmatrix} = \begin{bmatrix} 1 & T^{-2}\Sigma y_{t-1} \\ T^{-2}\Sigma y_{t-1} & T^{-3}\Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-3/2}\Sigma y_{t-1} u_t \end{bmatrix}. \quad [17.4.44]$$

From [17.4.41] and [17.4.42], the first term in [17.4.44] converges to

$$\begin{bmatrix} 1 & T^{-2}\Sigma y_{t-1} \\ T^{-2}\Sigma y_{t-1} & T^{-3}\Sigma y_{t-1}^2 \end{bmatrix} \xrightarrow{p} \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/3 \end{bmatrix} \equiv \mathbf{Q}. \quad [17.4.45]$$

From [17.4.43] and [17.3.18], the second term in [17.4.44] satisfies

$$\begin{aligned} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-3/2}\Sigma y_{t-1} u_t \end{bmatrix} &\xrightarrow{p} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-3/2}\Sigma \alpha(t-1) u_t \end{bmatrix} \\ &\xrightarrow{L} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/3 \end{bmatrix}\right) \quad [17.4.46] \\ &= N(\mathbf{0}, \sigma^2\mathbf{Q}). \end{aligned}$$

Combining [17.4.44] through [17.4.46], it follows that

$$\begin{bmatrix} T^{1/2}(\hat{\alpha}_T - \alpha) \\ T^{3/2}(\hat{\rho}_T - 1) \end{bmatrix} \xrightarrow{L} N(\mathbf{0}, \mathbf{Q}^{-1} \cdot \sigma^2\mathbf{Q} \cdot \mathbf{Q}^{-1}) = N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}). \quad [17.4.47]$$

Thus, for case 3, both estimated coefficients are asymptotically Gaussian. In fact, the asymptotic properties of $\hat{\alpha}_T$ and $\hat{\rho}_T$ are exactly the same as those for $\hat{\alpha}_T$ and $\hat{\delta}_T$ in the deterministic time trend regression analyzed in Chapter 16. The reason for this correspondence is very simple: the regressor y_{t-1} is asymptotically dominated by the time trend $\alpha \cdot (t - 1)$. In large samples, it is as if the explanatory variable y_{t-1} were replaced by the time trend $\alpha \cdot (t - 1)$. Recalling the analysis of Section 16.2, it follows that for case 3, the standard *OLS* t and F statistics can be calculated in the usual way and compared with the standard tables (Tables B.3 and B.4, respectively).

Case 4. Constant Term and Time Trend Included in the Regression; True Process Is Random Walk With or Without Drift

Suppose, as in the previous case, that the true model is

$$y_t = \alpha + y_{t-1} + u_t,$$

where u_t is i.i.d. with mean zero and variance σ^2. For this case, the true value of α turns out not to matter for the asymptotic distribution. In contrast to the previous case, we now assume that a time trend is included in the regression that is actually estimated by *OLS*:

$$y_t = \alpha + \rho y_{t-1} + \delta t + u_t. \quad [17.4.48]$$

If $\alpha \neq 0$, y_{t-1} would be asymptotically equivalent to a time trend. Since a time trend is already included as a separate variable in the regression, this would make the explanatory variables collinear in large samples. Describing the asymptotic

distribution of the estimates therefore requires not just a rescaling of variables but also a rotation of the kind introduced in Section 16.3.

Note that the regression model of [17.4.48] can equivalently be written as

$$\begin{aligned} y_t &= (1-\rho)\alpha + \rho[y_{t-1} - \alpha(t-1)] + (\delta + \rho\alpha)t + u_t \\ &\equiv \alpha^* + \rho^*\xi_{t-1} + \delta^* t + u_t, \end{aligned} \qquad [17.4.49]$$

where $\alpha^* \equiv (1-\rho)\alpha$, $\rho^* \equiv \rho$, $\delta^* \equiv (\delta + \rho\alpha)$, and $\xi_t \equiv y_t - \alpha t$. Moreover, under the null hypothesis that $\rho = 1$ and $\delta = 0$,

$$\xi_t = y_0 + u_1 + u_2 + \cdots + u_t;$$

that is, ξ_t is the random walk described in Proposition 17.1. Consider, as in Section 16.3, a hypothetical regression of y_t on a constant, ξ_{t-1}, and a time trend, producing the *OLS* estimates

$$\begin{bmatrix} \hat{\alpha}_T^* \\ \hat{\rho}_T^* \\ \hat{\delta}_T^* \end{bmatrix} = \begin{bmatrix} T & \Sigma\xi_{t-1} & \Sigma t \\ \Sigma\xi_{t-1} & \Sigma\xi_{t-1}^2 & \Sigma\xi_{t-1}t \\ \Sigma t & \Sigma t\xi_{t-1} & \Sigma t^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y_t \\ \Sigma\xi_{t-1}y_t \\ \Sigma t y_t \end{bmatrix}. \qquad [17.4.50]$$

The maintained hypothesis is that $\alpha = \alpha_0$, $\rho = 1$, and $\delta = 0$, which in the transformed system would mean $\alpha^* = 0$, $\rho^* = 1$, and $\delta^* = \alpha_0$. The deviations of the *OLS* estimates from these true values are given by

$$\begin{bmatrix} \hat{\alpha}_T^* \\ \hat{\rho}_T^* - 1 \\ \hat{\delta}_T^* - \alpha_0 \end{bmatrix} = \begin{bmatrix} T & \Sigma\xi_{t-1} & \Sigma t \\ \Sigma\xi_{t-1} & \Sigma\xi_{t-1}^2 & \Sigma\xi_{t-1}t \\ \Sigma t & \Sigma t\xi_{t-1} & \Sigma t^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma u_t \\ \Sigma\xi_{t-1}u_t \\ \Sigma t u_t \end{bmatrix}. \qquad [17.4.51]$$

Consulting the rates of convergence in Proposition 17.1, in this case the scaling matrix should be

$$\Upsilon_T = \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix},$$

and [17.4.20] would be

$$\begin{aligned} &\begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_T^* \\ \hat{\rho}_T^* - 1 \\ \hat{\delta}_T^* - \alpha_0 \end{bmatrix} \\ &= \left\{ \begin{bmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} T & \Sigma\xi_{t-1} & \Sigma t \\ \Sigma\xi_{t-1} & \Sigma\xi_{t-1}^2 & \Sigma\xi_{t-1}t \\ \Sigma t & \Sigma t\xi_{t-1} & \Sigma t^2 \end{bmatrix} \right. \\ &\quad \left. \times \begin{bmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & T^{-3/2} \end{bmatrix} \right\}^{-1} \\ &\quad \times \left\{ \begin{bmatrix} T^{-1/2} & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & T^{-3/2} \end{bmatrix} \begin{bmatrix} \Sigma u_t \\ \Sigma\xi_{t-1}u_t \\ \Sigma t u_t \end{bmatrix} \right\} \end{aligned}$$

or

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T^* \\ T(\hat{\rho}_T^* - 1) \\ T^{3/2}(\hat{\delta}_T^* - \alpha_0) \end{bmatrix} = \begin{bmatrix} 1 & T^{-3/2}\Sigma\xi_{t-1} & T^{-2}\Sigma t \\ T^{-3/2}\Sigma\xi_{t-1} & T^{-2}\Sigma\xi_{t-1}^2 & T^{-5/2}\Sigma\xi_{t-1}t \\ T^{-2}\Sigma t & T^{-5/2}\Sigma t\xi_{t-1} & T^{-3}\Sigma t^2 \end{bmatrix}^{-1} \times \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-1}\Sigma\xi_{t-1}u_t \\ T^{-3/2}\Sigma t u_t \end{bmatrix} \qquad [17.4.52]$$

The asymptotic distribution can then be found from Proposition 17.1:

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T^* \\ T(\hat{\rho}_T^* - 1) \\ T^{3/2}(\hat{\delta}_T^* - \alpha_0) \end{bmatrix}$$

$$\xrightarrow{L} \begin{bmatrix} 1 & \sigma\cdot\int W(r)\,dr & \frac{1}{2} \\ \sigma\cdot\int W(r)\,dr & \sigma^2\cdot\int [W(r)]^2\,dr & \sigma\cdot\int rW(r)\,dr \\ \frac{1}{2} & \sigma\cdot\int rW(r)\,dr & \frac{1}{3} \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} \sigma\cdot W(1) \\ \frac{1}{2}\sigma^2\{[W(1)]^2 - 1\} \\ \sigma\cdot\{W(1) - \int W(r)\,dr\} \end{bmatrix}$$

$$= \sigma\begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\,dr & \frac{1}{2} \\ \int W(r)\,dr & \int [W(r)]^2\,dr & \int rW(r)\,dr \\ \frac{1}{2} & \int rW(r)\,dr & \frac{1}{3} \end{bmatrix}^{-1} \qquad [17.4.53]$$

$$\times \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \\ W(1) - \int W(r)\,dr \end{bmatrix}$$

$$= \begin{bmatrix} \sigma & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sigma \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\,dr & \frac{1}{2} \\ \int W(r)\,dr & \int [W(r)]^2\,dr & \int rW(r)\,dr \\ \frac{1}{2} & \int rW(r)\,dr & \frac{1}{3} \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \\ W(1) - \int W(r)\,dr \end{bmatrix}.$$

Note that $\hat{\rho}_T^*$, the *OLS* estimate of ρ based on [17.4.49], is identical to $\hat{\rho}_T$, the *OLS* estimate of ρ based on [17.4.48]. Thus, the asymptotic distribution of $T(\hat{\rho}_T - 1)$ is given by the middle row of [17.4.53]. Note that this distribution does not depend on either σ or α; in particular, it does not matter whether or not the true value of α is zero.

The asymptotic distribution of $\hat{\sigma}_{\hat{\rho}_T}$, the *OLS* standard error for $\hat{\rho}_T$, can be found using similar calculations to those in [17.4.31] and [17.4.32]. Notice that

$$
\begin{aligned}
T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} &= T^2 \cdot s_T^2 [0 \quad 1 \quad 0] \begin{bmatrix} T & \Sigma\xi_{t-1} & \Sigma t \\ \Sigma\xi_{t-1} & \Sigma\xi_{t-1}^2 & \Sigma\xi_{t-1}t \\ \Sigma t & \Sigma t\xi_{t-1} & \Sigma t^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= s_T^2 [0 \quad 1 \quad 0] \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix} \\
&\quad \times \begin{bmatrix} T & \Sigma\xi_{t-1} & \Sigma t \\ \Sigma\xi_{t-1} & \Sigma\xi_{t-1}^2 & \Sigma\xi_{t-1}t \\ \Sigma t & \Sigma t\xi_{t-1} & \Sigma t^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{1/2} & 0 & 0 \\ 0 & T & 0 \\ 0 & 0 & T^{3/2} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= s_T^2 [0 \quad 1 \quad 0] \\
&\quad \times \begin{bmatrix} 1 & T^{-3/2}\Sigma\xi_{t-1} & T^{-2}\Sigma t \\ T^{-3/2}\Sigma\xi_{t-1} & T^{-2}\Sigma\xi_{t-1}^2 & T^{-5/2}\Sigma\xi_{t-1}t \\ T^{-2}\Sigma t & T^{-5/2}\Sigma t\xi_{t-1} & T^{-3}\Sigma t^2 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad [17.4.54] \\
&\xrightarrow{L} \sigma^2 [0 \quad 1 \quad 0] \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \\
&\quad \times \begin{bmatrix} 1 & \int W(r)\,dr & \frac{1}{2} \\ \int W(r)\,dr & \int [W(r)]^2\,dr & \int rW(r)\,dr \\ \frac{1}{2} & \int rW(r)\,dr & \frac{1}{3} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sigma & 0 \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&= [0 \quad 1 \quad 0] \begin{bmatrix} 1 & \int W(r)\,dr & \frac{1}{2} \\ \int W(r)\,dr & \int [W(r)]^2\,dr & \int rW(r)\,dr \\ \frac{1}{2} & \int rW(r)\,dr & \frac{1}{3} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \\
&\equiv Q.
\end{aligned}
$$

From this result it follows that the asymptotic distribution of the *OLS* t test of the hypothesis that $\rho = 1$ is given by

$$t_T = T(\hat{\rho}_T - 1) \div (T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T})^{1/2} \xrightarrow{p} T(\hat{\rho}_T - 1) \div \sqrt{Q}. \qquad [17.4.55]$$

Again, this distribution does not depend on α or σ. The small-sample distribution of the *OLS* t statistic under the assumption of Gaussian disturbances is presented under case 4 in Table B.6. If this distribution were truly t, then a value below -2.0 would be sufficient to reject the null hypothesis. However, Table B.6 reveals that, because of the nonstandard distribution, the t statistic must be below -3.4 before the null hypothesis of a unit root could be rejected.

The assumption that the true value of δ is equal to zero is again an auxiliary hypothesis upon which the asymptotic properties of the test depend. Thus, as in case 2, it is natural to consider the *OLS* F test of the joint null hypothesis that $\delta = 0$ and $\rho = 1$. Though this F test is calculated in the usual way, its asymptotic distribution is nonstandard, and the calculated F statistic should be compared with the value under case 4 in Table B.7.

Summary of Dickey-Fuller Tests in the Absence of Serial Correlation

We have seen that the asymptotic properties of the *OLS* estimate $\hat{\rho}_T$ when the true value of ρ is unity depend on whether or not a constant term or a time trend is included in the regression that is estimated and on whether or not the random walk that describes the true process for y_t includes a drift term. These results are summarized in Table 17.1.

Which is the "correct" case to use to test the null hypothesis of a unit root? The answer depends on why we are interested in testing for a unit root. If the analyst has a specific null hypothesis about the process that generated the data, then obviously this would guide the choice of test. In the absence of such guidance, one general principle would be to fit a specification that is a plausible description of the data under both the null hypothesis and the alternative. This principle would suggest using the case 4 test for a series with an obvious trend and the case 2 test for series without a significant trend.

For example, Figure 17.2 plots the nominal interest rate series used in the examples in this section. Although this series has tended upward over this sample period, there is nothing in economic theory to suggest that nominal interest rates should exhibit a deterministic time trend, and so a natural null hypothesis is that the true process is a random walk without trend. In terms of framing a plausible alternative, it is difficult to maintain that these data could have been generated by $i_t = \rho i_{t-1} + u_t$ with $|\rho|$ significantly less than 1. If these data were to be described by a stationary process, surely the process would have a positive mean. This argues for including a constant term in the estimated regression, even though under the null hypothesis the true process does not contain a constant term. Thus, case 2 is a sensible approach for these data, as analyzed in Examples 17.4 and 17.5.

As a second example, Figure 17.3 plots quarterly real GNP for the United States from 1947:I to 1989:I. Given a growing population and technological improvements, such a series would certainly be expected to exhibit a persistent upward trend, and this trend is unmistakable in the figure. The question is whether this trend arises from the positive drift term of a random walk:

$$H_0\colon y_t = \alpha + y_{t-1} + u_t \qquad \alpha > 0,$$

or from a deterministic time trend added to a stationary $AR(1)$:

$$H_A\colon y_t = \alpha + \delta t + \rho y_{t-1} + u_t \qquad |\rho| < 1.$$

Thus, the recommended test statistics for this case are those described in case 4.

The following model for 100 times the log of real GNP (denoted y_t) was estimated by *OLS* regression:

$$y_t = \underset{(13.53)}{27.24} + \underset{(0.019304)}{0.96252}\, y_{t-1} + \underset{(0.01521)}{0.02753}\, t. \qquad [17.4.56]$$

(standard errors in parentheses). The sample size is $T = 168$. The Dickey-Fuller

ρ test is

$$T(\hat{\rho} - 1) = 168(0.96252 - 1.0) = -6.3.$$

Since $-6.3 > -21.0$, the null hypothesis that GNP is characterized by a random walk with possible drift is accepted at the 5% level. The Dickey-Fuller t test,

$$t = \frac{0.96252 - 1.0}{0.019304} = -1.94,$$

exceeds the 5% critical value of -3.44, so that the null hypothesis of a unit root is accepted by this test as well. Finally, the F test of the joint null hypothesis that $\delta = 0$ and $\rho = 1$ is 2.44. Since this is less than the 5% critical value of 6.42 from Table B.7, this null hypothesis is again accepted.

TABLE 17.1
Summary of Dickey-Fuller Tests for Unit Roots in the Abseuce of Serial Correlation

Case 1:

Estimated regression: $y_t = \rho y_{t-1} + u_t$
True process: $y_t = y_{t-1} + u_t \qquad u_t \sim$ i.i.d. $N(0, \sigma^2)$
$T(\hat{\rho}_T - 1)$ has the distribution described under the heading Case 1 in Table B.5.
$(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T}$ has the distribution described under Case 1 in Table B.6.

Case 2:

Estimated regression: $y_t = \alpha + \rho y_{t-1} + u_t$
True process: $y_t = y_{t-1} + u_t \qquad u_t \sim$ i.i.d. $N(0, \sigma^2)$
$T(\hat{\rho}_T - 1)$ has the distribution described under Case 2 in Table B.5.
$(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T}$ has the distribution described under Case 2 in Table B.6.
OLS F test of joint hypothesis that $\alpha = 0$ and $\rho = 1$ has the distribution described under Case 2 in Table B.7.

Case 3:

Estimated regression: $y_t = \alpha + \rho y_{t-1} + u_t$
True process: $y_t = \alpha + y_{t-1} + u_t \qquad \alpha \neq 0, u_t \sim$ i.i.d. $(0, \sigma^2)$
$(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T} \xrightarrow{L} N(0, 1)$

Case 4:

Estimated regression: $y_t = \alpha + \rho y_{t-1} + \delta t + u_t$
True process: $y_t = \alpha + y_{t-1} + u_t \qquad \alpha$ any, $u_t \sim$ i.i.d. $N(0, \sigma^2)$
$T(\hat{\rho}_T - 1)$ has the distribution described under Case 4 in Table B.5.
$(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T}$ has the distribution described under Case 4 in Table B.6.
OLS F test of joint hypothesis that $\rho = 1$ and $\delta = 0$ has the distribution described under Case 4 in Table B.7.

Notes to Table 17.1

Estimated regression indicates the form in which the regression is estimated, using observations $t = 1, 2, \ldots, T$ and conditioning on observation $t = 0$.

True process describes the null hypothesis under which the distribution is calculated.

$\hat{\rho}_T$ is the *OLS* estimate of ρ from the indicated regression based on a sample of size T.

$(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T}$ is the *OLS t* test of $\rho = 1$.

OLS F test of a hypothesis involving two restrictions is given by expression [17.7.39].

If $u_t \sim$ i.i.d. $N(0, \sigma^2)$, then Tables B.5 through B.7 give Monte Carlo estimates of the exact small-sample distribution. The tables are also valid for large T when u_t is non-Gaussian i.i.d. as well as for certain heterogeneously distributed serially uncorrelated processes. For serially correlated u_t, see Table 17.2 or 17.3.

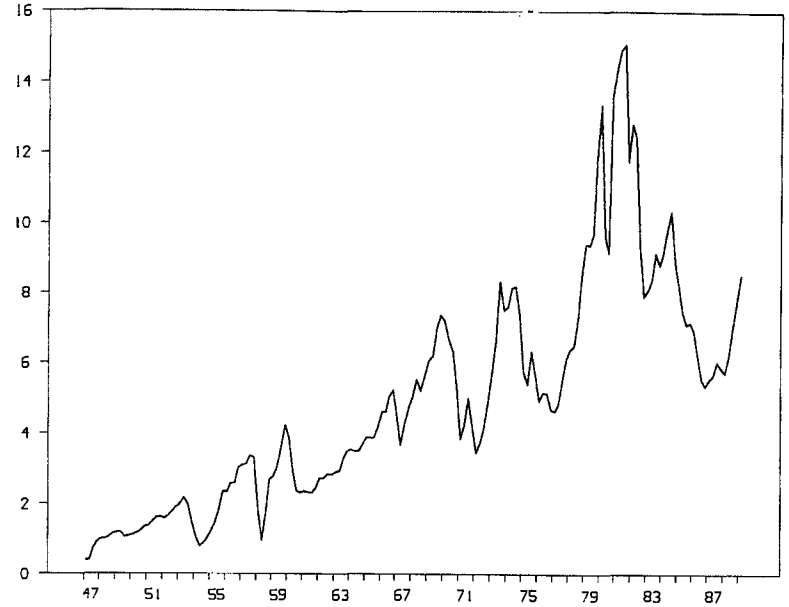


FIGURE 17.2 U.S. nominal interest rate on 3-month Treasury bills, data sampled quarterly but quoted at an annual rate, 1947:I to 1989:I.

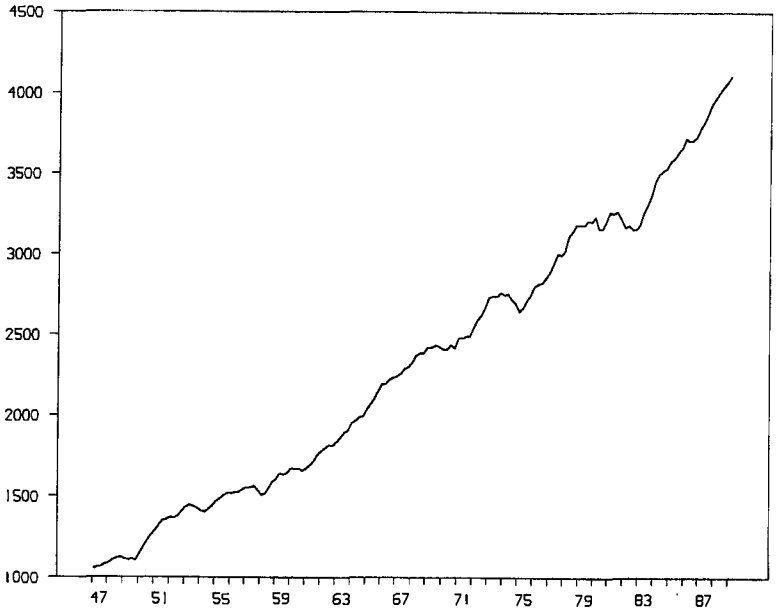


FIGURE 17.3 U.S. real GNP, data sampled quarterly but quoted at an annual rate in billions of 1982 dollars, 1947:I to 1989:I.

Of the tests discussed so far, those developed for case 2 seem appropriate for the interest rate data and the tests developed for case 4 seem best for the GNP data. However, more general tests presented in Sections 17.6 and 17.7 are to be preferred for describing either of these series. This is because the maintained assumption throughout this section has been that the disturbance term u_t in the regression is i.i.d. There is no strong reason to expect this for either of these time series. The next section develops results that can be used to test for unit roots in serially correlated processes.

17.5. *Asymptotic Results for Unit Root Processes with General Serial Correlation*

This section generalizes Proposition 17.1 to allow for serial correlation. The following preliminary result is quite helpful.

Proposition 17.2: *Let*

$$u_t = \psi(L)\varepsilon_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad [17.5.1]$$

where

$$E(\varepsilon_t) = 0$$

$$E(\varepsilon_t\varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise} \end{cases}$$

$$\sum_{j=0}^{\infty} j \cdot |\psi_j| < \infty. \qquad [17.5.2]$$

Then

$$u_1 + u_2 + \cdots + u_t = \psi(1)\cdot(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t) + \eta_t - \eta_0, \qquad [17.5.3]$$

where $\psi(1) \equiv \Sigma_{j=0}^{\infty}\psi_j$, $\eta_t = \Sigma_{j=0}^{\infty}\alpha_j\varepsilon_{t-j}$, $\alpha_j = -(\psi_{j+1} + \psi_{j+2} + \psi_{j+3} + \cdots)$, *and* $\Sigma_{j=0}^{\infty}|\alpha_j| < \infty$.

The condition in [17.5.2] is slightly stronger than absolute summability, though it is satisfied by any stationary *ARMA* process.

Notice that if y_t is an $I(1)$ process y_t whose first difference is given by u_t, or

$$\Delta y_t = u_t,$$

then

$$y_t = u_1 + u_2 + \cdots + u_t + y_0 = \psi(1)\cdot(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t) + \eta_t - \eta_0 + y_0.$$

Proposition 17.2 thus states that any $I(1)$ process whose first difference satisfies [17.5.1] and [17.5.2] can be written as the sum of a random walk $(\psi(1)\cdot(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t))$, initial conditions $(y_0 - \eta_0)$, and a stationary process (η_t). This observation was first made by Beveridge and Nelson (1981), and [17.5.3] is sometimes referred to as the *Beveridge-Nelson decomposition*.

Notice that η_t is a stationary process. An important implication of this is that if [17.5.3] is divided by $\sqrt{t}$, only the first term $(1/\sqrt{t})\psi(1)\cdot(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t)$ should matter for the distribution of $(1/\sqrt{t})\cdot(u_1 + u_2 + \cdots + u_t)$ as $t \to \infty$.

As an example of how this result can be used, suppose that $X_T(r)$ is defined

as in [17.3.2]:

$$X_T(r) \equiv (1/T) \sum_{t=1}^{[Tr]^*} u_t, \qquad [17.5.4]$$

where u_t satisfies the conditions of Proposition 17.2 with ε_t i.i.d. and $E(\varepsilon_t^4) < \infty$. Then the continuous-time process $\sqrt{T} \cdot X_T(r)$ converges to $\sigma \cdot \psi(1)$ times standard Brownian motion:

$$\sqrt{T} \cdot X_T(\cdot) \xrightarrow{L} \sigma \cdot \psi(1) \cdot W(\cdot). \qquad [17.5.5]$$

To derive [17.5.5], note from Proposition 17.2 that

$$\begin{aligned} \sqrt{T} \cdot X_T(r) &= (1/\sqrt{T}) \cdot \sum_{t=1}^{[Tr]^*} u_t \\ &= \psi(1) \cdot (1/\sqrt{T}) \cdot \sum_{t=1}^{[Tr]^*} \varepsilon_t + (1/\sqrt{T}) \cdot (\eta_{[Tr]^*} - \eta_0) \qquad [17.5.6] \\ &\equiv \psi(1) \cdot (1/\sqrt{T}) \cdot \sum_{t=1}^{[Tr]^*} \varepsilon_t + S_T(r), \end{aligned}$$

where we have defined $S_T(r) \equiv (1/\sqrt{T}) \cdot (\eta_{[Tr]^*} - \eta_0)$. Notice as in Example 17.2 that

$$S_T(\cdot) \xrightarrow{p} 0 \qquad [17.5.7]$$

as $T \to \infty$. Furthermore, from [17.3.8],

$$(1/\sqrt{T}) \cdot \sum_{t=1}^{[Tr]^*} \varepsilon_t \xrightarrow{L} \sigma \cdot W(r). \qquad [17.5.8]$$

Substituting [17.5.7] and [17.5.8] into [17.5.6] produces [17.5.5].

Another implication is found by evaluating the functions in [17.5.5] at $r = 1$:

$$(1/\sqrt{T}) \sum_{t=1}^{T} u_t \xrightarrow{L} \sigma \cdot \psi(1) \cdot W(1). \qquad [17.5.9]$$

Since $W(1)$ is distributed $N(0, 1)$, result [17.5.9] states that

$$(1/\sqrt{T}) \sum_{t=1}^{T} u_t \xrightarrow{L} N(0, \sigma^2[\psi(1)]^2),$$

which is the usual central limit theorem of Proposition 7.11.

The following proposition uses this basic idea to generalize the other results from Proposition 17.1; for details on the proofs, see Appendix 17.A.

Proposition 17.3: *Let $u_t = \psi(L)\varepsilon_t = \Sigma_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$, where $\Sigma_{j=0}^{\infty} j \cdot |\psi_j| < \infty$ and $\{\varepsilon_t\}$ is an i.i.d. sequence with mean zero, variance σ^2, and finite fourth moment. Define*

$$\gamma_j \equiv E(u_t u_{t-j}) = \sigma^2 \sum_{s=0}^{\infty} \psi_s \psi_{s+j} \qquad \text{for } j = 0, 1, 2, \ldots \qquad [17.5.10]$$

$$\lambda \equiv \sigma \sum_{j=0}^{\infty} \psi_j = \sigma \cdot \psi(1)$$

$$\xi_t \equiv u_1 + u_2 + \cdots + u_t \qquad \text{for } t = 1, 2, \ldots, T \qquad [17.5.11]$$

with $\xi_0 \equiv 0$. Then

(a) $T^{-1/2} \sum_{t=1}^{T} u_t \xrightarrow{L} \lambda \cdot W(1);$

(b) $T^{-1/2} \sum_{t=1}^{T} u_{t-j}\varepsilon_t \xrightarrow{L} N(0, \sigma^2\gamma_0) \qquad for\ j = 1, 2, \ldots;$

(c) $T^{-1} \sum_{t=1}^{T} u_t u_{t-j} \xrightarrow{p} \gamma_j \qquad for\ j = 0, 1, 2, \ldots;$

(d) $T^{-1} \sum_{t=1}^{T} \xi_{t-1}\varepsilon_t \xrightarrow{L} (1/2)\sigma \cdot \lambda \cdot \{[W(1)]^2 - 1\};$

(e) $T^{-1} \sum_{t=1}^{T} \xi_{t-1}u_{t-j}$

$$\xrightarrow{L} \begin{cases} (1/2)\{\lambda^2 \cdot [W(1)]^2 - \gamma_0\} & for\ j = 0 \\ (1/2)\{\lambda^2 \cdot [W(1)]^2 - \gamma_0\} + \gamma_0 + \gamma_1 + \gamma_2 + \cdots + \gamma_{j-1} & \\ & for\ j = 1, 2, \ldots; \end{cases}$$

(f) $T^{-3/2} \sum_{t=1}^{T} \xi_{t-1} \xrightarrow{L} \lambda \cdot \int_0^1 W(r)\, dr;$

(g) $T^{-3/2} \sum_{t=1}^{T} tu_{t-j} \xrightarrow{L} \lambda \cdot \left\{ W(1) - \int_0^1 W(r)\, dr \right\} \qquad for\ j = 0, 1, 2, \ldots;$

(h) $T^{-2} \sum_{t=1}^{T} \xi_{t-1}^2 \xrightarrow{L} \lambda^2 \cdot \int_0^1 [W(r)]^2\, dr;$

(i) $T^{-5/2} \sum_{t=1}^{T} t\xi_{t-1} \xrightarrow{L} \lambda \cdot \int_0^1 rW(r)\, dr;$

(j) $T^{-3} \sum_{t=1}^{T} t\xi_{t-1}^2 \xrightarrow{L} \lambda^2 \cdot \int_0^1 r \cdot [W(r)]^2\, dr;$

(k) $T^{-(v+1)} \sum_{t=1}^{T} t^v \rightarrow 1/(v + 1) \qquad for\ v = 0, 1, \ldots.$

Again, there are simpler ways to describe individual results; for example, (a) is a $N(0, \lambda^2)$ distribution, (d) is $(1/2)\sigma\lambda \cdot [\chi^2(1) - 1]$, and (f) and (g) are both $N(0, \lambda^2/3)$ distributions.

These results can be used to construct unit root tests for serially correlated observations in two ways. One approach, due to Phillips (1987) and Phillips and Perron (1988), is to continue to estimate the regressions in exactly the form indicated in Table 17.1, but to adjust the test statistics to take account of serial correlation and potential heteroskedasticity in the disturbances. This approach is described in Section 17.6. The second approach, due to Dickey and Fuller (1979), is to add lagged changes of y as explanatory variables in the regressions in Table 17.1. This is described in Section 17.7.

17.6. *Phillips-Perron Tests for Unit Roots*

Asymptotic Distribution for Case 2 Assumptions with Serially Correlated Disturbances

To illustrate the basic idea behind the Phillips (1987) and Phillips and Perron (1988) tests for unit roots, we will discuss in detail the treatment they propose for the analog of case 2 of Section 17.4. After this case has been reviewed, similar

results will be stated for case 1 and case 4, with details developed in exercises at the end of the chapter.

Case 2 of Section 17.4 considered *OLS* estimation of α and ρ in the regression model

$$y_t = \alpha + \rho y_{t-1} + u_t \qquad [17.6.1]$$

under the assumption that the true $\alpha = 0$, $\rho = 1$, and u_t is i.i.d. Phillips and Perron (1988) generalized these results to the case when u_t is serially correlated and possibly heteroskedastic as well. For now we will assume that the true process is

$$y_t - y_{t-1} = u_t = \psi(L)\varepsilon_t,$$

where $\psi(L)$ and ε_t satisfy the conditions of Proposition 17.3. More general conditions under which the same techniques are valid will be discussed at the end of this section.

If [17.6.1] were a stationary autoregression with $|\rho| < 1$, the *OLS* estimate $\hat{\rho}_T$ in [17.4.15] would not give a consistent estimate of ρ when u_t is serially correlated. However, if ρ is equal to 1, the rate T convergence of $\hat{\rho}_T$ turns out to ensure that $\hat{\rho}_T \xrightarrow{p} 1$ even when u_t is serially correlated. Phillips and Perron (1988) therefore proposed estimating [17.6.1] by *OLS* even when u_t is serially correlated and then modifying the statistics in Section 17.4 to take account of the serial correlation.

Let $\hat{\alpha}_T$ and $\hat{\rho}_T$ be the *OLS* estimates based on [17.6.1] without any correction for serial correlation; that is, $\hat{\alpha}_T$ and $\hat{\rho}_T$ are the magnitudes defined in [17.4.15]. If the true values are $\alpha = 0$ and $\rho = 1$, then, as in [17.4.22],

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T \\ T(\hat{\rho}_T - 1) \end{bmatrix} = \begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-1}\Sigma y_{t-1}u_t \end{bmatrix}, \qquad [17.6.2]$$

where Σ denotes summation over t from 1 to T. Also, under the null hypothesis that $\alpha = 0$ and $\rho = 1$, it follows as in [17.4.4] that

$$y_t = y_0 + u_1 + u_2 + \cdots + u_t.$$

If $u_t = \psi(L)\varepsilon_t$ as in Proposition 17.3, then y_t is the variable labeled ξ_t in Proposition 17.3, plus the inconsequential value y_0. Using results (f) and (h) of that proposition,

$$\begin{aligned} &\begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y_{t-1}^2 \end{bmatrix}^{-1} \\ &\xrightarrow{L} \begin{bmatrix} 1 & \lambda \cdot \int W(r)\,dr \\ \lambda \cdot \int W(r)\,dr & \lambda^2 \cdot \int [W(r)]^2\,dr \end{bmatrix}^{-1} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1}, \end{aligned} \qquad [17.6.3]$$

where the integral sign denotes integration over r from 0 to 1. Similarly, results (a) and (e) of Proposition 17.3 give

$$\begin{aligned} \begin{bmatrix} T^{-1/2}\Sigma u_t \\ T^{-1}\Sigma y_{t-1}u_t \end{bmatrix} &\xrightarrow{L} \begin{bmatrix} \lambda \cdot W(1) \\ \frac{1}{2}\{\lambda^2[W(1)]^2 - \gamma_0\} \end{bmatrix} \\ &= \begin{bmatrix} \lambda \cdot W(1) \\ \frac{1}{2}\lambda^2\{[W(1)]^2 - 1\} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2}\{\lambda^2 - \gamma_0\} \end{bmatrix} \\ &= \lambda \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2}\{\lambda^2 - \gamma_0\} \end{bmatrix}. \end{aligned} \qquad [17.6.4]$$

Substituting [17.6.3] and [17.6.4] into [17.6.2] produces

$$\begin{bmatrix} T^{1/2}\hat{\alpha}_T \\ T(\hat{\rho}_T - 1) \end{bmatrix} \xrightarrow{L} \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1}$$

$$\times \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \left\{ \lambda \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{2}\{\lambda^2 - \gamma_0\} \end{bmatrix} \right\}$$

$$= \begin{bmatrix} \lambda & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix}$$

$$+ \left\{ \begin{bmatrix} 1 & 0 \\ 0 & \lambda^{-1} \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ \frac{1}{2}\{\lambda^2 - \gamma_0\}/\lambda \end{bmatrix} \right\}. \qquad [17.6.5]$$

The second element of this vector states that

$$T(\hat{\rho}_T - 1)$$

$$\xrightarrow{L} [0 \quad 1] \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix}$$

$$+ (1/2)\frac{\{\lambda^2 - \gamma_0\}}{\lambda^2} [0 \quad 1] \begin{bmatrix} 1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad [17.6.6]$$

$$= \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1)\int W(r)\,dr}{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2} + \frac{(1/2)\cdot(\lambda^2 - \gamma_0)}{\lambda^2\left\{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2\right\}}.$$

The first term of the last equality in [17.6.6] is the same as [17.4.28], which described the asymptotic distribution that $T(\hat{\rho}_T - 1)$ would have if u_t were i.i.d. The final term in [17.6.6] is a correction for serial correlation. Notice that if u_t is serially uncorrelated, then $\psi_0 = 1$ and $\psi_j = 0$ for $j = 1, 2, \ldots$. Thus, if u_t is serially uncorrelated, then $\lambda^2 = \sigma^2 \cdot [\psi(1)]^2 = \sigma^2$ and $\gamma_0 = E(u_t^2) = \sigma^2$. Hence [17.6.6] includes the earlier result [17.4.28] as a special case when u_t is serially uncorrelated.

It is easy to use $\hat{\sigma}_{\hat{\rho}_T}$, the *OLS* standard error for $\hat{\rho}_T$, to construct a sample statistic that can be used to estimate the correction for serial correlation. Let $\mathbf{Y}_T$ be the matrix defined in [17.4.21], and let s_T^2 be the *OLS* estimate of the variance of u_t:

$$s_T^2 = (T - 2)^{-1} \sum_{t=1}^{T} (y_t - \hat{\alpha}_T - \hat{\rho}_T y_{t-1})^2.$$

Then the asymptotic distribution of $T^2 \cdot \hat{\sigma}_{\hat{\rho}_T}^2$ can be found using the same approach as in [17.4.31] through [17.4.33]:

$$
\begin{aligned}
&T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \\
&= s_T^2 [0 \quad 1] \mathbf{Y}_T \begin{bmatrix} T & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1} \mathbf{Y}_T \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&\xrightarrow{p} s_T^2 [0 \quad 1] \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\, dr \\ \int W(r)\, dr & \int [W(r)]^2\, dr \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= (s_T^2/\lambda^2)[0 \quad 1] \begin{bmatrix} 1 & \int W(r)\, dr \\ \int W(r)\, dr & \int [W(r)]^2\, dr \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= (s_T^2/\lambda^2) \frac{1}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2}.
\end{aligned}
\tag{17.6.7}
$$

It follows from [17.6.6] that

$$
\begin{aligned}
&T(\hat{\rho}_T - 1) - \tfrac{1}{2}(T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \div s_T^2)(\lambda^2 - \gamma_0) \\
&\xrightarrow{p} T(\hat{\rho}_T - 1) - \frac{1}{2}\left(\frac{1}{\lambda^2}\right) \frac{1}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2} (\lambda^2 - \gamma_0) \\
&\xrightarrow{L} \frac{\tfrac{1}{2}\{[W(1)]^2 - 1\} - W(1) \int W(r)\, dr}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2}.
\end{aligned}
\tag{17.6.8}
$$

Thus, the statistic in [17.6.8] has the same asymptotic distribution [17.4.28] as the variable tabulated under the heading Case 2 in Table B.5.

Result [17.6.8] can also be used to find the asymptotic distribution of the *OLS* t test of $\rho = 1$:

$$
\begin{aligned}
t_T &= \frac{(\hat{\rho}_T - 1)}{\hat{\sigma}_{\hat{\rho}_T}} \\
&= \frac{T(\hat{\rho}_T - 1)}{\{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T}\}^{1/2}} \\
&\xrightarrow{p} \left\{ \frac{\tfrac{1}{2}\{[W(1)]^2 - 1\} - W(1) \int W(r)\, dr}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2} + \tfrac{1}{2}(T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T}/s_T^2) \cdot (\lambda^2 - \gamma_0) \right\} \\
&\quad \div \{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T}\}^{1/2} \\
&= \frac{\tfrac{1}{2}\{[W(1)]^2 - 1\} - W(1) \int W(r)\, dr}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2} \div \{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T}\}^{1/2} \\
&\quad + \{(1/2)(1/s_T)(\lambda^2 - \gamma_0)\} \times \{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \div s_T^2\}^{1/2}
\end{aligned}
$$

$$\xrightarrow{p} \left\{ \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1)\int W(r)\,dr}{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2} \right\} \left(\frac{\lambda^2}{s_T^2}\right)^{1/2}$$

$$\times \left\{ \int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2 \right\}^{1/2}$$

$$+ \{(1/2)(1/s_T)(\lambda^2 - \gamma_0)\} \times \{T^2 \cdot \hat{\sigma}^2_{\hat{\beta}_T} \div s_T^2\}^{1/2} \qquad [17.6.9]$$

with the last convergence following from [17.6.7]. Moreover,

$$s_T^2 = (T-2)^{-1} \sum_{t=1}^{T} (y_t - \hat{\alpha}_T - \hat{\rho}_T y_{t-1})^2 \xrightarrow{p} E(u_t^2) = \gamma_0. \qquad [17.6.10]$$

Hence, [17.6.9] implies that

$$(\gamma_0/\lambda^2)^{1/2} \cdot t_T \xrightarrow{p} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1)\int W(r)\,dr}{\left\{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2\right\}^{1/2}} \qquad [17.6.11]$$

$$+ \{\tfrac{1}{2}(\lambda^2 - \gamma_0)/\lambda\} \times \{T \cdot \hat{\sigma}_{\hat{\beta}_T} \div s_T\}.$$

Thus,

$$(\gamma_0/\lambda^2)^{1/2} \cdot t_T - \{\tfrac{1}{2}(\lambda^2 - \gamma_0)/\lambda\} \times \{T \cdot \hat{\sigma}_{\hat{\beta}_T} \div s_T\}$$

$$\xrightarrow{L} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1)\int W(r)\,dr}{\left\{\int [W(r)]^2\,dr - \left[\int W(r)\,dr\right]^2\right\}^{1/2}}, \qquad [17.6.12]$$

which is the same limiting distribution [17.4.36] obtained for the random variable tabulated for case 2 in Table B.6.

The statistics in [17.6.8] and [17.6.12] require knowledge of the population parameters γ_0 and λ^2. Although these moments are unknown, they are easy to estimate consistently. Since $\gamma_0 = E(u_t^2)$, one consistent estimate is given by

$$\hat{\gamma}_0 = T^{-1} \sum_{t=1}^{T} \hat{u}_t^2, \qquad [17.6.13]$$

where $\hat{u}_t = y_t - \hat{\alpha}_T - \hat{\rho}_T y_{t-1}$ is the *OLS* sample residual. Phillips and Perron (1988) instead used the standard *OLS* estimate $\hat{\gamma}_0 = (T-2)^{-1}\Sigma\hat{u}_t^2 = s_T^2$. Similarly, from result (a) of Proposition 17.3, λ^2 is the asymptotic variance of the sample mean of u:

$$\sqrt{T} \cdot \bar{u} = T^{-1/2} \sum_{t=1}^{T} u_t \xrightarrow{L} N(0, \lambda^2). \qquad [17.6.14]$$

Recalling the discussion of the variance of the sample mean in Sections 7.2 and 10.5, this magnitude can equivalently be described as

$$\lambda^2 = \sigma^2 \cdot [\psi(1)]^2 = \gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j = 2\pi s_u(0), \qquad [17.6.15]$$

where γ_j is the jth autocovariance of u_t and $s_u(0)$ is the population spectrum of u_t at frequency zero. Thus, any of the estimates of this magnitude proposed in Section

10.5 might be used. For example, if only the first q autocovariances are deemed relevant, the Newey-West estimator could be used:

$$\hat{\lambda}^2 = \hat{\gamma}_0 + 2\sum_{j=1}^{q}[1 - j/(q+1)]\hat{\gamma}_j, \qquad [17.6.16]$$

where

$$\hat{\gamma}_j = T^{-1}\sum_{t=j+1}^{T}\hat{u}_t\hat{u}_{t-j} \qquad [17.6.17]$$

and $\hat{u}_t = y_t - \hat{\alpha}_T - \hat{\rho}_T y_{t-1}$.

To summarize, under the null hypothesis that the first difference of y_t is a zero-mean covariance-stationary process, the Phillips and Perron[9] approach is to estimate equation [17.6.1] by *OLS* and use the standard *OLS* formulas to calculate $\hat{\rho}$ and its standard error $\hat{\sigma}_{\hat{\rho}}$ along with the standard error of the regression s. The jth autocovariance of $\hat{u}_t = y_t - \hat{\alpha} - \hat{\rho}y_{t-1}$ is then calculated from [17.6.17]. The resulting estimates $\hat{\gamma}_0$ and $\hat{\lambda}^2$ are then used in [17.6.8] to construct a statistic that has the same asymptotic distribution as does the variable tabulated in the case 2 section of Table B.5. The analogous adjustments to the standard *OLS* t test of $\rho = 1$ described in [17.6.12] produce a statistic that can be compared with the case 2 section of Table B.6

Example 17.6

Let $\hat{u}_t$ denote the *OLS* sample residual for the interest rate regression [17.4.37] of Example 17.4:

$$\hat{u}_t \equiv i_t - \underset{(0.112)}{0.211} - \underset{(0.019133)}{0.96691}\, i_{t-1} \qquad \text{for } t = 1, 2, \ldots, 168.$$

The estimated autocovariances of these *OLS* residuals are

$$\hat{\gamma}_0 = (1/T)\sum_{t=1}^{T}\hat{u}_t^2 = 0.630 \qquad \hat{\gamma}_1 = (1/T)\sum_{t=2}^{T}\hat{u}_t\hat{u}_{t-1} = 0.114$$

$$\hat{\gamma}_2 = (1/T)\sum_{t=3}^{T}\hat{u}_t\hat{u}_{t-2} = -0.162 \qquad \hat{\gamma}_3 = (1/T)\sum_{t=4}^{T}\hat{u}_t\hat{u}_{t-3} = 0.064$$

$$\hat{\gamma}_4 = (1/T)\sum_{t=5}^{T}\hat{u}_t\hat{u}_{t-4} = 0.047.$$

[9]The procedure recommended by Phillips and Perron differs slightly from that in the text. To see the relation, write the first line of [17.6.7] as

$$T^2\cdot\hat{\sigma}^2_{\hat{\rho}_T} \div s^2_T = [0 \quad 1]\begin{bmatrix} 1 & T^{-3/2}\Sigma y_{t-1} \\ T^{-3/2}\Sigma y_{t-1} & T^{-2}\Sigma y^2_{t-1}\end{bmatrix}^{-1}\begin{bmatrix}0\\1\end{bmatrix}$$

$$= \frac{1}{T^{-2}\Sigma y^2_{t-1} - T^{-3}(\Sigma y_{t-1})^2} = \frac{1}{T^{-1}[T^{-1}\Sigma y^2_{t-1} - (T^{-1}\Sigma y_{t-1})^2]}$$

$$= \frac{1}{T^{-1}[T^{-1}\Sigma(y_{t-1} - \bar{y}_{-1})^2]},$$

where $\bar{y}_{-1} \equiv T^{-1}\Sigma y_{t-1}$ and the last equality follows from [4.A.5]. Instead of this expression, Phillips and Perron used

$$\frac{1}{T^{-2}\Sigma(y_t - \bar{y})^2}.$$

The advantage of the formula in the text is that it is trivial to calculate from the output produced by standard regression packages and the identical formula can be used for cases 1, 2, and 4.

Thus, if the serial correlation of u_t is to be described with $q = 4$ autocovariances,

$$\hat{\lambda}^2 = 0.630 + 2(\tfrac{4}{5})(0.114) + 2(\tfrac{3}{5})(-0.162) + 2(\tfrac{2}{5})(0.064) + 2(\tfrac{1}{5})(0.047)$$
$$= 0.688.$$

The usual *OLS* formula for the variance of the residuals from this regression is

$$s^2 = (T - 2)^{-1} \sum_{t=1}^{T} \hat{u}_t^2 = 0.63760.$$

Hence, the Phillips-Perron ρ statistic is

$$T(\hat{\rho} - 1) - (1/2)\cdot(T^2\cdot\hat{\sigma}_{\hat{\rho}}^2/s^2)\cdot(\hat{\lambda}^2 - \hat{\gamma}_0)$$
$$= 168(0.96691 - 1) - \tfrac{1}{2}\{[(168)(0.019133)]^2/(0.63760)\}(0.688 - 0.630)\cdot$$
$$= -6.03.$$

Comparing this with the 5% critical value for case 2 of Table B.5, we see that $-6.03 > -13.8$. We thus accept the null hypothesis that the interest rate data could plausibly have been generated by a simple unit root process.

Similarly, the adjustment to the t statistic from Example 17.4 described in [17.6.12] is

$$(\hat{\gamma}_0/\hat{\lambda}^2)^{1/2}t - \{\tfrac{1}{2}(\hat{\lambda}^2 - \hat{\gamma}_0)(T\cdot\hat{\sigma}_{\hat{\rho}}/s) \div \hat{\lambda}\}$$
$$= \{(0.630)/(0.688)\}^{1/2}(0.96691 - 1)/0.019133$$
$$-\{(1/2)(0.688 - 0.630)[(168)(0.019133)/\sqrt{(0.63760)}] \div \sqrt{(0.688)}\}$$
$$= -1.80.$$

Since $-1.80 > -2.89$, the null hypothesis of a unit root is again accepted at the 5% level.

Phillips-Perron Tests for Cases 1 and 4

The asymptotic distributions in [17.6.8] and [17.6.12] were derived under the assumption that the true process for the first difference of y_t is serially correlated with mean zero. Even though the true unit root process exhibited no drift, it was assumed that the estimated *OLS* regression included a constant term as in case 2 of Section 17.4.

The same ideas can be used to generalize case 1 or case 4 of Section 17.4, and the statistics [17.6.8] and [17.6.12] can be compared in each case with the corresponding values in Tables B.5 and B.6. These results are summarized in Table 17.2. The reader is invited to confirm these claims in exercises at the end of the chapter.

Example 17.7

The residuals from the GNP regression [17.4.56] have the following estimated autocovariances:

$$\hat{\gamma}_0 = 1.136 \qquad \hat{\gamma}_1 = 0.424 \qquad \hat{\gamma}_2 = 0.285$$
$$\hat{\gamma}_3 = 0.006 \qquad \hat{\gamma}_4 = -0.110,$$

from which

$$\hat{\lambda}^2 = 1.136 + 2\{\tfrac{4}{5}(0.424) + \tfrac{3}{5}(0.285) + \tfrac{2}{5}(0.006) - \tfrac{1}{5}(0.110)\} = 2.117.$$

Also, $s^2 = 1.15627$. Thus, for these data the Phillips-Perron ρ test is

$$\begin{aligned} &T(\hat{\rho} - 1) - \tfrac{1}{2}(T^2 \cdot \hat{\sigma}_{\hat{\rho}}^2/s^2)(\hat{\lambda}^2 - \hat{\gamma}_0) \\ &= 168(0.96252 - 1) - \tfrac{1}{2}\left\{\frac{[(168)(0.019304)]^2}{1.15627}\right\}(2.117 - 1.136) \\ &= -10.76. \end{aligned}$$

Since $-10.76 > -21.0$, the null hypothesis that log GNP follows a unit root process with or without drift is accepted at the 5% level.

The Phillips-Perron t test is

$$\begin{aligned} &(\hat{\gamma}_0/\hat{\lambda}^2)^{1/2}t - \{\tfrac{1}{2}(\hat{\lambda}^2 - \hat{\gamma}_0)(T \cdot \hat{\sigma}_{\hat{\rho}}/s) \div \hat{\lambda}\} \\ &= \{(1.136)/(2.117)\}^{1/2}(0.96252 - 1)/0.019304 \\ &\quad - \{\tfrac{1}{2}(2.117 - 1.136)[[(168)(0.019304)]/\sqrt{1.15627}] \div \sqrt{(2.117)}\} \\ &= -2.44. \end{aligned}$$

Since $-2.44 > -3.44$, the null hypothesis of a unit root is again accepted.

More General Processes for u_t

The Newey-West estimator $\hat{\lambda}^2$ in [17.6.16] can provide a consistent estimate of λ^2 for an $MA(\infty)$ process, provided that q, the lag truncation parameter, goes to infinity as the sample size T grows, and provided that q grows sufficiently slowly relative to T. Phillips (1987) established such consistency assuming that $q_T \to \infty$ and $q_T/T^{1/4} \to 0$; for example, $q_T = A \cdot T^{1/5}$ satisfies this requirement. Phillips's results warrant using a larger value of q with a larger data set, though they do not tell us exactly how large to choose q in practice. Monte Carlo investigations have been provided by Phillips and Perron (1988), Schwert (1989), and Kim and Schmidt (1990), though no simple rule emerges from these studies. Andrews's (1991) procedures might be used in this context.

Asymptotic results can also be obtained under weaker assumptions about u_t than those in Proposition 17.3. For example, the reader may note from the proof of result 17.3(e) that the parameter γ_0 appears because it is the plim of $T^{-1} \times \Sigma_{t=1}^T u_t^2$. Under the conditions of the proposition, the law of large numbers ensures that this plim is just the expected value of u_t^2, which expected value was denoted γ_0. However, even if the data are heterogeneously distributed with $E(u_t^2) = \gamma_{0,t}$, it may still be the case that $T^{-1}\Sigma_{t=1}^T \gamma_{0,t}$ converges to some constant. If $T^{-1}\Sigma_{t=1}^T u_t^2$ also converges to this constant, then this constant plays the role of γ_0 in a generalization of result 17.3(e).

Similarly, let $\bar{u}_T$ denote the sample mean from some heterogeneously distributed process with population mean zero:

$$\bar{u}_T \equiv T^{-1} \sum_{t=1}^{T} u_t;$$

and let λ_T^2 denote T times the variance of $\bar{u}_T$:

$$\lambda_T^2 \equiv T \cdot \mathrm{Var}(\bar{u}_T) = T^{-1} \cdot E(u_1 + u_2 + \cdots + u_T)^2.$$

The sample mean $\bar{u}_T$ may still satisfy the central limit theorem:

$$T^{-1/2} \sum_{t=1}^{T} u_t \xrightarrow{L} N(0, \lambda^2)$$

TABLE 17.2
Summary of Phillips-Perron Tests for Unit Roots

Case 1:

Estimated regression: $y_t = \rho y_{t-1} + u_t$
True process: $y_t = y_{t-1} + u_t$

Z_ρ has the same asymptotic distribution as the variable described under the heading Case 1 in Table B.5.

Z_t has the same asymptotic distribution as the variable described under Case 1 in Table B.6.

Case 2:

Estimated regression: $y_t = \alpha + \rho y_{t-1} + u_t$
True process: $y_t = y_{t-1} + u_t$

Z_ρ has the same asymptotic distribution as the variable described under Case 2 in Table B.5.

Z_t has the same asymptotic distribution as the variable described under Case 2 in Table B.6.

Case 4:

Estimated regression: $y_t = \alpha + \rho y_{t-1} + \delta t + u_t$
True process: $y_t = \alpha + y_{t-1} + u_t \qquad \alpha$ any

Z_ρ has the same asymptotic distribution as the variable described under Case 4 in Table B.5.

Z_t has the same asymptotic distribution as the variable described under Case 4 in Table B.6.

Notes to Table 17.2

Estimated regression indicates the form in which the regression is estimated, using observations $t = 1, 2, \ldots, T$ and conditioning on observation $t = 0$.

True process describes the null hypothesis under which the distribution is calculated. In each case, u_t is assumed to have mean zero but can be heterogeneously distributed and serially correlated with

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} E(u_t^2) = \gamma_0$$

$$\lim_{T\to\infty} T^{-1} E(u_1 + u_2 + \cdots + u_T)^2 = \lambda^2.$$

Z_ρ is the following statistic:

$$Z_\rho \equiv T(\hat{\rho}_T - 1) - (1/2)\{T^2 \cdot \hat{\sigma}^2_{\hat{\rho}_T} \div s_T^2\}(\hat{\lambda}_T^2 - \hat{\gamma}_{0,T}),$$

where

$$\hat{\gamma}_{j,T} = T^{-1} \sum_{t=j+1}^{T} \hat{u}_t \hat{u}_{t-j}$$

$\hat{u}_t$ = *OLS* sample residual from the estimated regression

$$\hat{\lambda}_T^2 = \hat{\gamma}_{0,T} + 2 \cdot \sum_{j=1}^{q} [1 - j/(q+1)]\hat{\gamma}_{j,T}$$

$$s_T^2 = (T - k)^{-1} \sum_{t=1}^{T} \hat{u}_t^2$$

k = number of parameters in estimated regression

$\hat{\sigma}_{\hat{\rho}_T}$ = *OLS* standard error for $\hat{\rho}$.

Z_t is the following statistic:

$$Z_t \equiv (\hat{\gamma}_{0,T}/\hat{\lambda}_T^2)^{1/2} \cdot (\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T} - (1/2)(\hat{\lambda}_T^2 - \hat{\gamma}_{0,T})(1/\hat{\lambda}_T)\{T \cdot \hat{\sigma}_{\hat{\rho}_T} \div s_T\}.$$

or

$$T^{-1/2} \sum_{t=1}^{T} u_t \xrightarrow{L} \lambda \cdot W(1),$$

where

$$\lambda^2 \equiv \lim_{T \to \infty} \lambda_T^2, \qquad [17.6.18]$$

providing a basis for generalizing result 17.3(a).

If u_t were a covariance-stationary process with absolutely summable autocovariances, then Proposition 7.5(b) would imply that $\lim_{T\to\infty} \lambda_T^2 = \sum_{j=-\infty}^{\infty} \gamma_j$. Recalling [7.2.8], expression [17.6.18] would in this case just be another way to describe the parameter λ^2 in Proposition 17.3.

Thus, the parameters γ_0 and λ^2 in [17.6.8] and [17.6.12] can more generally be defined as

$$\gamma_0 \equiv \lim_{T \to \infty} T^{-1} \sum_{t=1}^{T} E(u_t^2) \qquad [17.6.19]$$

$$\lambda^2 \equiv \lim_{T \to \infty} T^{-1} \cdot E(u_1 + u_2 + \cdots + u_T)^2. \qquad [17.6.20]$$

Phillips (1987) and Perron and Phillips (1988) derived [17.6.8] and [17.6.12] assuming that u_t is a zero-mean but otherwise heterogeneously distributed process satisfying certain restrictions on the serial dependence and higher moments. From this perspective, expressions [17.6.19] and [17.6.20] can be used as the definitions of the parameters γ_0 and λ^2. Clearly, the estimators [17.6.13] and [17.6.16] continue to be appropriate for this alternative interpretation.

On the Observational Equivalence of Unit Root and Covariance-Stationary Processes

We saw in Section 15.4 that given any $I(0)$ process for y_t and any finite sample size T, there exists an $I(1)$ process that will be impossible to distinguish from the $I(0)$ representation on the basis of the first and second sample moments of y. Yet the Phillips and Perron procedures seem to offer a way to test the null hypothesis that the sample was generated from an arbitrary $I(1)$ process. What does it mean if the test leads us to reject the null hypothesis that y_t is $I(1)$ when we know that there exists an $I(1)$ process that describes the sample arbitrarily well?

Some insight into this question can be gained by considering the example in equation [15.4.8],

$$(1 - L)y_t = (1 + \theta L)\varepsilon_t, \qquad [17.6.21]$$

where θ is slightly larger than -1 and ε_t is i.i.d. with mean zero and variance σ^2. The model [17.6.21] implies that

$$\begin{aligned} y_t &= (\varepsilon_t + \theta\varepsilon_{t-1}) + (\varepsilon_{t-1} + \theta\varepsilon_{t-2}) + \cdots + (\varepsilon_1 + \theta\varepsilon_0) + y_0 \\ &= \varepsilon_t + (1 + \theta)\varepsilon_{t-1} + (1 + \theta)\varepsilon_{t-2} + \cdots + (1 + \theta)\varepsilon_1 + \theta\varepsilon_0 + y_0 \\ &= \varepsilon_t + (1 + \theta)\xi_{t-1} + \theta\varepsilon_0 + y_0, \end{aligned}$$

where

$$\xi_{t-1} \equiv \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-1}.$$

For large t, the variable y_t is dominated by the unit root component, $(1 + \theta)\xi_{t-1}$, and the asymptotic results are all governed by this term. However, if θ is close to -1, then in a finite sample y_t would behave essentially like the white noise series ε_t plus a constant $(\theta\varepsilon_0 + y_0)$. In such a case the Phillips-Perron test is likely to reject the null hypothesis of a unit root in finite samples even though it is true.[10] For example, Schwert (1989) generated Monte Carlo samples of size $T = 1{,}000$ according to the unit root model [17.6.21] with $\theta = -0.8$. The Phillips-Perron test that is supposed to reject only 5% of the time actually rejected the null hypothesis in virtually every sample, even though the null hypothesis is true! Similar results were reported by Phillips and Perron (1988) and Kim and Schmidt (1990).

Campbell and Perron (1991) argued that such false rejections are not necessarily a bad thing. If θ is near -1, then for many purposes an $I(0)$ model may provide a more useful description of the process in [17.6.21] than does the true $I(1)$ model. In support of this claim, they generated samples from the process [17.6.21] and estimated by *OLS* both an autoregressive process in levels,

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

and an autoregressive process in differences,

$$\Delta y_t = \alpha + \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_p \Delta y_{t-p} + \varepsilon_t.$$

They found that for θ close to -1, forecasts based on the levels y_t tended to perform better than those based on the differences Δy_t, even though the true data-generating process was $I(1)$.

A related issue, of course, arises with false acceptances. Clearly, if the true model is

$$y_t = \rho y_{t-1} + \varepsilon_t \qquad [17.6.22]$$

with ρ slightly below 1, then the null hypothesis that $\rho = 1$ is likely to be accepted in small samples, even though it is false. The value of accepting a false null hypothesis in this case is that imposing the condition $\rho = 1$ may produce a better forecast than one based on an estimated $\hat{\rho}_T$, particularly given the small-sample downward bias of $\hat{\rho}_T$. Furthermore, when ρ is close to 1, the values in Table B.6 might give a better small-sample approximation to the distribution of $(\hat{\rho}_T - 1) \div \hat{\sigma}_{\hat{\rho}_T}$ than the traditional t tables.[11]

This discussion underscores that the goal of unit root tests is to find a parsimonious representation that gives a reasonable approximation to the true process, as opposed to determining whether or not the true process is literally $I(1)$.

17.7. *Asymptotic Properties of a* pth-*Order Autoregression and the Augmented Dickey-Fuller Tests for Unit Roots*

The Phillips-Perron tests were based on simple *OLS* regressions of y_t on its own lagged value and possibly a constant or time trend as well. Corrections for serial correlation were then made to the standard *OLS* coefficient and t statistics. This section discusses an alternative approach, due to Dickey and Fuller (1979), which controls for serial correlation by including higher-order autoregressive terms in the regression.

[10]For more detailed discussion, see Phillips and Perron (1988, p. 344).

[11]See Evans and Savin (1981, 1984) for a description of the small-sample distributions.

An Alternative Representation of an AR(p) *Process*

Suppose that the data were really generated from an $AR(p)$ process,

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) y_t = \varepsilon_t, \qquad [17.7.1]$$

where $\{\varepsilon_t\}$ is an i.i.d. sequence with mean zero, variance σ^2, and finite fourth moment. It is helpful to write the autoregression [17.7.1] in a slightly different form. To do so, define

$$\rho \equiv \phi_1 + \phi_2 + \cdots + \phi_p \qquad [17.7.2]$$

$$\zeta_j \equiv -[\phi_{j+1} + \phi_{j+2} + \cdots + \phi_p] \qquad \text{for } j = 1, 2, \ldots, p - 1. \qquad [17.7.3]$$

Notice that for any values of $\phi_1, \phi_2, \ldots, \phi_p$, the following polynomials in L are equivalent:

$$\begin{aligned}
&(1 - \rho L) - (\zeta_1 L + \zeta_2 L^2 + \cdots + \zeta_{p-1} L^{p-1})(1 - L) \\
&= 1 - \rho L - \zeta_1 L + \zeta_1 L^2 - \zeta_2 L^2 + \zeta_2 L^3 - \cdots - \zeta_{p-1} L^{p-1} + \zeta_{p-1} L^p \\
&= 1 - (\rho + \zeta_1) L - (\zeta_2 - \zeta_1) L^2 - (\zeta_3 - \zeta_2) L^3 - \cdots \\
&\quad - (\zeta_{p-1} - \zeta_{p-2}) L^{p-1} - (-\zeta_{p-1}) L^p \\
&= 1 - [(\phi_1 + \phi_2 + \cdots + \phi_p) - (\phi_2 + \phi_3 + \cdots + \phi_p)] L \\
&\quad - [-(\phi_3 + \phi_4 + \cdots + \phi_p) + (\phi_2 + \phi_3 + \cdots + \phi_p)] L^2 - \cdots \\
&\quad - [-(\phi_p) + (\phi_{p-1} + \phi_p)] L^{p-1} - (\phi_p) L^p \\
&= 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_{p-1} L^{p-1} - \phi_p L^p. \qquad [17.7.4]
\end{aligned}$$

Thus, the autoregression [17.7.1] can equivalently be written

$$\{(1 - \rho L) - (\zeta_1 L + \zeta_2 L^2 + \cdots + \zeta_{p-1} L^{p-1})(1 - L)\} y_t = \varepsilon_t \qquad [17.7.5]$$

or

$$y_t = \rho y_{t-1} + \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \varepsilon_t. \qquad [17.7.6]$$

Suppose that the process that generated y_t contains a single unit root; that is, suppose one root of

$$(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0 \qquad [17.7.7]$$

is unity,

$$1 - \phi_1 - \phi_2 - \cdots - \phi_p = 0, \qquad [17.7.8]$$

and all other roots of [17.7.7] are outside the unit circle. Notice that [17.7.8] implies that the coefficient ρ in [17.7.2] is unity. Moreover, when $\rho = 1$, expression [17.7.4] would imply

$$\begin{aligned}
&(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) \\
&\qquad = (1 - \zeta_1 z - \zeta_2 z^2 - \cdots - \zeta_{p-1} z^{p-1})(1 - z).
\end{aligned} \qquad [17.7.9]$$

Of the p values of z that make the left side of [17.7.9] zero, one is $z = 1$ and all other roots are presumed to be outside the unit circle. The same must be true of the right side as well, meaning that all roots of

$$(1 - \zeta_1 z - \zeta_2 z^2 - \cdots - \zeta_{p-1} z^{p-1}) = 0$$

lie outside the unit circle. Under the null hypothesis that $\rho = 1$, expression [17.7.5] could then be written as

$$(1 - \zeta_1 L - \zeta_2 L^2 - \cdots - \zeta_{p-1} L^{p-1})\,\Delta y_t = \varepsilon_t$$

or

$$\Delta y_t = u_t, \qquad [17.7.10]$$

where

$$u_t = (1 - \zeta_1 L - \zeta_2 L^2 - \cdots - \zeta_{p-1} L^{p-1})^{-1} \varepsilon_t.$$

Equation [17.7.10] indicates that y_t behaves like the variable ξ_t described in Proposition 17.3, with

$$\psi(L) = (1 - \zeta_1 L - \zeta_2 L^2 - \cdots - \zeta_{p-1} L^{p-1})^{-1}.$$

One of the advantages of writing the autoregression of [17.7.1] in the equivalent form of [17.7.6] is that only one of the regressors in [17.7.6], namely, y_{t-1}, is $I(1)$, whereas all of the other regressors $(\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-p+1})$ are stationary. Thus, [17.7.6] is the Sims, Stock, and Watson (1990) canonical form, originally proposed for this problem by Fuller (1976). Since no knowledge of any population parameters is needed to write the model in this canonical form, in this case it is convenient to estimate the parameters by direct *OLS* estimation of [17.7.6].

Results generalizing those for case 1 in Section 17.4 are obtained when the regression is estimated as written in [17.7.6] without a constant term. Cases 2 and 3 are generalized by including a constant term in [17.7.6], while case 4 is generalized by including a constant term and a time trend in [17.7.6]. For illustration, the case 2 regression is discussed in detail. Comparable results for case 1, case 3, and case 4 will be summarized in Table 17.3 later in this section, with details developed in exercises at the end of the chapter.

Case 2. The Estimated Autoregression Includes a Constant Term, but the Data Were Really Generated by a Unit Root Autoregression with No Drift

Following the usual notational convention for *OLS* estimation of autoregressions, we assume that the initial sample is of size $T + p$, with observations numbered $\{y_{-p+1}, y_{-p+2}, \ldots, y_T\}$, and condition on the first p observations. We are interested in the properties of *OLS* estimation of

$$\begin{aligned} y_t &= \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \varepsilon_t \\ &\equiv \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t, \end{aligned} \qquad [17.7.11]$$

where $\boldsymbol{\beta} \equiv (\zeta_1, \zeta_2, \ldots, \zeta_{p-1}, \alpha, \rho)'$ and $\mathbf{x}_t \equiv (\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-p+1}, 1, y_{t-1})'$. The deviation of the *OLS* estimate $\mathbf{b}_T$ from the true value $\boldsymbol{\beta}$ is given by

$$\mathbf{b}_T - \boldsymbol{\beta} = \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t\right]. \qquad [17.7.12]$$

Letting $u_t \equiv y_t - y_{t-1}$, the individual terms in [17.7.12] are

$$\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' = \qquad [17.7.13]$$

$$\begin{bmatrix} \Sigma u_{t-1}^2 & \Sigma u_{t-1}u_{t-2} & \cdots & \Sigma u_{t-1}u_{t-p+1} & \Sigma u_{t-1} & \Sigma u_{t-1}y_{t-1} \\ \Sigma u_{t-2}u_{t-1} & \Sigma u_{t-2}^2 & \cdots & \Sigma u_{t-1}u_{t-p+1} & \Sigma u_{t-2} & \Sigma u_{t-2}y_{t-1} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \Sigma u_{t-p+1}u_{t-1} & \Sigma u_{t-p+1}u_{t-2} & \cdots & \Sigma u_{t-p+1}^2 & \Sigma u_{t-p+1} & \Sigma u_{t-p+1}y_{t-1} \\ \Sigma u_{t-1} & \Sigma u_{t-2} & \cdots & \Sigma u_{t-p+1} & T & \Sigma y_{t-1} \\ \Sigma y_{t-1}u_{t-1} & \Sigma y_{t-1}u_{t-2} & \cdots & \Sigma y_{t-1}u_{t-p+1} & \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}$$

$$\sum_{t=1}^{T} \mathbf{x}_t\varepsilon_t = \begin{bmatrix} \Sigma u_{t-1}\varepsilon_t \\ \Sigma u_{t-2}\varepsilon_t \\ \vdots \\ \Sigma u_{t-p+1}\varepsilon_t \\ \Sigma \varepsilon_t \\ \Sigma y_{t-1}\varepsilon_t \end{bmatrix} \qquad [17.7.14]$$

with Σ denoting summation over $t = 1, 2, \ldots, T$.

Under the null hypothesis that $\alpha = 0$ and $\rho = 1$, we saw in [17.7.10] that y_t behaves like $\xi_t = u_1 + u_2 + \cdots + u_t$ in Proposition 17.3. Consulting the rates of convergence in Proposition 17.3, for this case the scaling matrix should be

$$\underset{(p+1\times p+1)}{\mathbf{Y}_T} \equiv \begin{bmatrix} \sqrt{T} & 0 & \cdots & 0 & 0 \\ 0 & \sqrt{T} & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{T} & 0 \\ 0 & 0 & \cdots & 0 & T \end{bmatrix}. \qquad [17.7.15]$$

Premultiplying [17.7.12] by $\mathbf{Y}_T$ as in [17.4.20] results in

$$\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}) = \left\{\mathbf{Y}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]\mathbf{Y}_T^{-1}\right\}^{-1}\left\{\mathbf{Y}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\varepsilon_t\right]\right\}. \qquad [17.7.16]$$

Consider the matrix $\mathbf{Y}_T^{-1}\Sigma\mathbf{x}_t\mathbf{x}_t'\mathbf{Y}_T^{-1}$. Elements in the upper left $(p \times p)$ block of $\Sigma\mathbf{x}_t\mathbf{x}_t'$ are divided by T, the first p elements of the $(p + 1)$th row or $(p + 1)$th column are divided by $T^{3/2}$, and the row $(p + 1)$, column $(p + 1)$ element of $\Sigma\mathbf{x}_t\mathbf{x}_t'$ is divided by T^2. Moreover,

$$T^{-1}\Sigma u_{t-i}u_{t-j} \xrightarrow{p} \gamma_{|i-j|} \qquad \text{from result (c) of Proposition 17.3}$$

$$T^{-1}\Sigma u_{t-j} \xrightarrow{p} E(u_{t-j}) = 0 \qquad \text{from the law of large numbers}$$

$$T^{-3/2}\Sigma y_{t-1}u_{t-j} \xrightarrow{p} 0 \qquad \text{from Proposition 17.3(e)}$$

$$T^{-3/2}\Sigma y_{t-1} \xrightarrow{L} \lambda \cdot \int W(r)\,dr \qquad \text{from Proposition 17.3(f)}$$

$$T^{-2}\Sigma y_{t-1}^2 \xrightarrow{L} \lambda^2 \cdot \int [W(r)]^2\,dr \qquad \text{from Proposition 17.3(h),}$$

where

$$\begin{aligned} \gamma_j &= E\{(\Delta y_t)(\Delta y_{t-j})\} \\ \lambda &\equiv \sigma \cdot \psi(1) = \sigma/(1 - \zeta_1 - \zeta_2 - \cdots - \zeta_{p-1}) \qquad [17.7.17] \\ \sigma^2 &= E(\varepsilon_t^2) \end{aligned}$$

and the integral sign denotes integration over r from 0 to 1. Thus,

$$\mathbf{Y}_T^{-1}[\Sigma \mathbf{x}_t \mathbf{x}_t']\mathbf{Y}_T^{-1}$$

$$\xrightarrow{L} \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} & 0 & 0 \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & \lambda \cdot \int W(r)\, dr \\ 0 & 0 & \cdots & 0 & \lambda \cdot \int W(r)\, dr & \lambda^2 \cdot \int [W(r)]^2\, dr \end{bmatrix}$$

$$\equiv \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}, \qquad [17.7.18]$$

where

$$\mathbf{V} \equiv \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix} \qquad [17.7.19]$$

$$\mathbf{Q} \equiv \begin{bmatrix} 1 & \lambda \cdot \int W(r)\, dr \\ \lambda \cdot \int W(r)\, dr & \lambda^2 \cdot \int [W(r)]^2\, dr \end{bmatrix} \qquad [17.7.20]$$

Next, consider the second term in [17.7.16],

$$\mathbf{Y}_T^{-1}[\Sigma \mathbf{x}_t \varepsilon_t] = \begin{bmatrix} T^{-1/2}\Sigma u_{t-1}\varepsilon_t \\ T^{-1/2}\Sigma u_{t-2}\varepsilon_t \\ \vdots \\ T^{-1/2}\Sigma u_{t-p+1}\varepsilon_t \\ T^{-1/2}\Sigma \varepsilon_t \\ T^{-1}\Sigma y_{t-1}\varepsilon_t \end{bmatrix} \qquad [17.7.21]$$

The first $p - 1$ elements of this vector are $\sqrt{T}$ times the sample mean of a martingale difference sequence whose variance-covariance matrix is

$$E\begin{bmatrix} u_{t-1}\varepsilon_t \\ u_{t-2}\varepsilon_t \\ \vdots \\ u_{t-p+1}\varepsilon_t \end{bmatrix}[u_{t-1}\varepsilon_t \quad u_{t-2}\varepsilon_t \quad \cdots \quad u_{t-p+1}\varepsilon_t]$$

$$= \sigma^2 \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix} \qquad [17.7.22]$$

$$= \sigma^2 \mathbf{V}.$$

Thus, the first $p - 1$ terms in [17.7.21] satisfy the usual central limit theorem,

$$\begin{bmatrix} T^{-1/2}\Sigma u_{t-1}\varepsilon_t \\ T^{-1/2}\Sigma u_{t-2}\varepsilon_t \\ \vdots \\ T^{-1/2}\Sigma u_{t-p+1}\varepsilon_t \end{bmatrix} \xrightarrow{L} \mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{V}). \qquad [17.7.23]$$

The distribution of the last two elements in [17.7.21] can be obtained from results (a) and (d) of Proposition 17.3:

$$\begin{bmatrix} T^{-1/2}\Sigma\varepsilon_t \\ T^{-1}\Sigma y_{t-1}\varepsilon_t \end{bmatrix} \xrightarrow{L} \mathbf{h}_2 \sim \begin{bmatrix} \sigma\cdot W(1) \\ \frac{1}{2}\sigma\lambda\cdot\{[W(1)]^2 - 1\} \end{bmatrix}. \qquad [17.7.24]$$

Substituting [17.7.18] through [17.7.24] into [17.7.16] results in

$$\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{V}^{-1}\mathbf{h}_1 \\ \mathbf{Q}^{-1}\mathbf{h}_2 \end{bmatrix}. \qquad [17.7.25]$$

Coefficients on Δy_{t-j}

The first $p - 1$ elements of $\boldsymbol{\beta}$ are $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$, which are the coefficients on zero-mean stationary regressors $(\Delta y_{t-1}, \Delta y_{t-2}, \ldots, \Delta y_{t-p+1})$. The block consisting of the first $p - 1$ elements in [17.7.25] states that

$$\sqrt{T}\begin{bmatrix} \hat{\zeta}_{1,T} - \zeta_1 \\ \hat{\zeta}_{2,T} - \zeta_2 \\ \vdots \\ \hat{\zeta}_{p-1,T} - \zeta_{p-1} \end{bmatrix} \xrightarrow{L} \mathbf{V}^{-1}\mathbf{h}_1. \qquad [17.7.26]$$

Recalling from [17.7.23] that $\mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, it follows that $\mathbf{V}^{-1}\mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{V}^{-1})$, or

$$\sqrt{T}\begin{bmatrix} \hat{\zeta}_{1,T} - \zeta_1 \\ \hat{\zeta}_{2,T} - \zeta_2 \\ \vdots \\ \hat{\zeta}_{p-1,T} - \zeta_{p-1} \end{bmatrix} \xrightarrow{L} N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \sigma^2\begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix}^{-1}\right), \qquad [17.7.27]$$

where $\gamma_j = E\{(\Delta y_t)(\Delta y_{t-j})\}$.

This means that a null hypothesis involving the coefficients on the stationary regressors $(\zeta_1, \zeta_2, \ldots, \zeta_{p-1})$ in [17.7.11] can be tested in the usual way, with the standard t and F statistics asymptotically valid. To see this, suppose that the null hypothesis is H_0: $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ for $\mathbf{R}$ a known $[m \times (p + 1)]$ matrix where m is the number of restrictions. The Wald form of the *OLS* χ^2 test [8.2.23] is given by

$$\begin{aligned} \chi^2_T &= (\mathbf{R}\mathbf{b}_T - \mathbf{r})'\left\{s_T^2\mathbf{R}\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\mathbf{R}'\right\}^{-1}(\mathbf{R}\mathbf{b}_T - \mathbf{r}) \\ &= [\mathbf{R}\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})]'\left\{s_T^2\mathbf{R}\cdot\sqrt{T}\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\sqrt{T}\cdot\mathbf{R}'\right\}^{-1} \\ &\quad \times [\mathbf{R}\cdot\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})], \end{aligned} \qquad [17.7.28]$$

where

$$s_T^2 = [T - (p + 1)]^{-1} \sum_{t=1}^{T} (y_t - \hat{\zeta}_{1,T}\Delta y_{t-1} - \hat{\zeta}_{2,T}\Delta y_{t-2} - \cdots - \hat{\zeta}_{p-1,T}\Delta y_{t-p+1} - \hat{\alpha}_T - \hat{\rho}_T y_{t-1})^2 \xrightarrow{p} E(\varepsilon_t^2) = \sigma^2. \qquad [17.7.29]$$

If none of the restrictions involves α or ρ, then the last two columns of $\mathbf{R}$ contain all zeros:

$$\underset{[m\times(p+1)]}{\mathbf{R}} = \begin{bmatrix} \underset{[m\times(p-1)]}{\mathbf{R}_1} & \underset{(m\times 2)}{\mathbf{0}} \end{bmatrix}. \qquad [17.7.30]$$

In this case, $\mathbf{R}\sqrt{T} = \mathbf{R}\mathbf{Y}_T$ for $\mathbf{Y}_T$ the matrix in [17.7.15], so that [17.7.28] can be written as

$$\chi_T^2 = [\mathbf{R}\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta})]' \left\{ s_T^2 \mathbf{R}\mathbf{Y}_T \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \right]^{-1} \mathbf{Y}_T\mathbf{R}' \right\}^{-1} [\mathbf{R}\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta})].$$

From [17.7.18], [17.7.25], [17.7.29], and [17.7.30], this converges to

$$\begin{aligned} \chi_T^2 &\xrightarrow{L} \left\{ [\mathbf{R}_1 \quad \mathbf{0}] \begin{bmatrix} \mathbf{V}^{-1}\mathbf{h}_1 \\ \mathbf{Q}^{-1}\mathbf{h}_2 \end{bmatrix} \right\}' \\ &\quad \times \left\{ \sigma^2 [\mathbf{R}_1 \quad \mathbf{0}] \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{R}_1' \\ \mathbf{0} \end{bmatrix} \right\}^{-1} \left\{ [\mathbf{R}_1 \quad \mathbf{0}] \begin{bmatrix} \mathbf{V}^{-1}\mathbf{h}_1 \\ \mathbf{Q}^{-1}\mathbf{h}_2 \end{bmatrix} \right\} \\ &= [\mathbf{R}_1\mathbf{V}^{-1}\mathbf{h}_1]'[\sigma^2\mathbf{R}_1\mathbf{V}^{-1}\mathbf{R}_1']^{-1}[\mathbf{R}_1\mathbf{V}^{-1}\mathbf{h}_1]. \end{aligned} \qquad [17.7.31]$$

But since $\mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, it follows that the $(m \times 1)$ vector $[\mathbf{R}_1\mathbf{V}^{-1}\mathbf{h}_1]$ is distributed $N(\mathbf{0}, [\sigma^2\mathbf{R}_1\mathbf{V}^{-1}\mathbf{R}_1'])$. Hence, expression [17.7.31] is a quadratic form in a Gaussian vector that satisfies the conditions of Proposition 8.1:

$$\chi_T^2 \xrightarrow{L} \chi^2(m).$$

This verifies that the usual t or F tests applied to any subset of the coefficients $\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_{p-1}$ have the standard limiting distributions.

Note, moreover, that [17.7.27] is exactly the same asymptotic distribution that would be obtained if the data were differenced before estimating the autoregression:

$$\Delta y_t = \zeta_1\Delta y_{t-1} + \zeta_2\Delta y_{t-2} + \cdots + \zeta_{p-1}\Delta y_{t-p+1} + \alpha + \varepsilon_t.$$

Thus, if the goal is to estimate $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$ or test hypotheses about these coefficients, there is no need based on asymptotic distribution theory for differencing the data before estimating the autoregression. Many researchers do recommend differencing the data first, but the reason is to reduce the small-sample bias and small-sample mean squared errors of the estimates, not to change the asymptotic distribution.

Coefficients on Constant Term and y_{t-1}

The last two elements of $\boldsymbol{\beta}$ are α and ρ, which are coefficients on the constant term and the $I(1)$ regressor, y_{t-1}. From [17.7.25], [17.7.20], and [17.7.24], their

limiting distribution is given by

$$\begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix} \begin{bmatrix} \hat{\alpha}_T \\ \hat{\rho}_T - 1 \end{bmatrix}$$

$$\xrightarrow{L} \begin{bmatrix} 1 & \lambda \cdot \int W(r)\, dr \\ \lambda \cdot \int W(r)\, dr & \lambda^2 \cdot \int [W(r)]^2\, dr \end{bmatrix}^{-1} \begin{bmatrix} \sigma \cdot W(1) \\ \frac{1}{2}\sigma\lambda \cdot \{[W(1)]^2 - 1\} \end{bmatrix}$$

$$= \sigma \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \begin{bmatrix} 1 & \int W(r)\, dr \\ \int W(r)\, dr & \int [W(r)]^2\, dr \end{bmatrix}^{-1} \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix}^{-1} \qquad [17.7.32]$$

$$\times \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma & 0 \\ 0 & \sigma/\lambda \end{bmatrix} \begin{bmatrix} 1 & \int W(r)\, dr \\ \int W(r)\, dr & \int [W(r)]^2\, dr \end{bmatrix}^{-1} \begin{bmatrix} W(1) \\ \frac{1}{2}\{[W(1)]^2 - 1\} \end{bmatrix}.$$

The second element of this vector implies that (λ/σ) times $T(\hat{\rho}_T - 1)$ has the same asymptotic distribution as [17.4.28], which described the estimate of ρ in a regression without lagged Δy and with serially uncorrelated disturbances:

$$T \cdot (\lambda/\sigma) \cdot (\hat{\rho}_T - 1) \xrightarrow{L} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1) \cdot \int W(r)\, dr}{\left\{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2\right\}}. \qquad [17.7.33]$$

Recall from [17.7.17] that

$$\lambda/\sigma = (1 - \zeta_1 - \zeta_2 - \cdots - \zeta_{p-1})^{-1}. \qquad [17.7.34]$$

This magnitude is clearly estimated consistently by

$$(1 - \hat{\zeta}_{1,T} - \hat{\zeta}_{2,T} - \cdots - \hat{\zeta}_{p-1,T})^{-1},$$

where $\hat{\zeta}_{j,T}$ denotes the estimate of ζ_j based on the *OLS* regression [17.7.11]. Thus, the generalization of the Dickey-Fuller ρ test when lagged changes in y are included in the regression is

$$\frac{T \cdot (\hat{\rho}_T - 1)}{1 - \hat{\zeta}_{1,T} - \hat{\zeta}_{2,T} - \cdots - \hat{\zeta}_{p-1,T}} \xrightarrow{L} \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1) \cdot \int W(r)\, dr}{\int [W(r)]^2\, dr - \left[\int W(r)\, dr\right]^2}. \qquad [17.7.35]$$

This is to be compared with the case 2 section of Table B.5.

Consider next an *OLS* t test of the null hypothesis that $\rho = 1$:

$$t_T = \frac{(\hat{\rho}_T - 1)}{\{s_T^2 \cdot \mathbf{e}'_{p+1} \cdot (\Sigma \mathbf{x}_t \mathbf{x}'_t)^{-1} \cdot \mathbf{e}_{p+1}\}^{1/2}}, \qquad [17.7.36]$$

where $\mathbf{e}_{p+1}$ denotes a $[(p + 1) \times 1]$ vector with unity in the last position and zeros elsewhere. Multiplying the numerator and denominator of [17.7.36] by T re-

sults in

$$t_T = \frac{T(\hat{\rho}_T - 1)}{\left\{ s_T^2 \cdot \mathbf{e}'_{p+1} \mathbf{Y}_T (\Sigma \mathbf{x}_t \mathbf{x}'_t)^{-1} \mathbf{Y}_T \mathbf{e}_{p+1} \right\}^{1/2}}. \qquad [17.7.37]$$

But

$$\begin{aligned} \mathbf{e}'_{p+1} \mathbf{Y}_T (\Sigma \mathbf{x}_t \mathbf{x}'_t)^{-1} \mathbf{Y}_T \mathbf{e}_{p+1} &= \mathbf{e}'_{p+1} \left\{ \mathbf{Y}_T^{-1} (\Sigma \mathbf{x}_t \mathbf{x}'_t) \mathbf{Y}_T^{-1} \right\}^{-1} \mathbf{e}_{p+1} \\ &\xrightarrow{L} \mathbf{e}'_{p+1} \begin{bmatrix} \mathbf{V}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}^{-1} \end{bmatrix} \mathbf{e}_{p+1} \\ &= \frac{1}{\lambda^2 \cdot \left\{ \int [W(r)]^2 \, dr - \left[\int W(r) \, dr \right]^2 \right\}} \end{aligned}$$

by virtue of [17.7.18] and [17.7.20]. Hence, from [17.7.37] and [17.7.33],

$$\begin{aligned} t_T &\xrightarrow{L} (\sigma/\lambda) \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1) \cdot \int W(r) \, dr}{\int [W(r)]^2 \, dr - \left[\int W(r) \, dr \right]^2} \\ &\quad \div \left\{ \frac{\sigma^2}{\lambda^2 \left\{ \int [W(r)]^2 \, dr - \left[\int W(r) \, dr \right]^2 \right\}} \right\}^{1/2} \qquad [17.7.38] \\ &= \frac{\frac{1}{2}\{[W(1)]^2 - 1\} - W(1) \cdot \int W(r) \, dr}{\left\{ \int [W(r)]^2 \, dr - \left[\int W(r) \, dr \right]^2 \right\}^{1/2}}. \end{aligned}$$

This is the same distribution as in [17.4.36]. Thus, the usual t test of $\rho = 1$ for *OLS* estimation of [17.7.11] can be compared with the case 2 section of Table B.6 without any corrections for the fact that lagged values of Δy are included in the regression.

A similar result applies to the Dickey-Fuller F test of the joint hypothesis that $\alpha = 0$ and $\rho = 1$. This null hypothesis can be represented as $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where

$$\underset{[2 \times (p+1)]}{\mathbf{R}} = \begin{bmatrix} \underset{[2 \times (p-1)]}{\mathbf{0}} & \underset{(2 \times 2)}{\mathbf{I}_2} \end{bmatrix}$$

and $\mathbf{r} = (0, 1)'$. The F test is then

$$F_T = (\mathbf{b}_T - \boldsymbol{\beta})' \mathbf{R}' \left\{ s_T^2 \cdot \mathbf{R} (\Sigma \mathbf{x}_t \mathbf{x}'_t)^{-1} \mathbf{R}' \right\}^{-1} \mathbf{R}(\mathbf{b}_T - \boldsymbol{\beta})/2. \qquad [17.7.39]$$

Define $\tilde{\mathbf{Y}}_T$ to be the following (2×2) matrix:

$$\tilde{\mathbf{Y}}_T \equiv \begin{bmatrix} T^{1/2} & 0 \\ 0 & T \end{bmatrix}. \qquad [17.7.40]$$

Notice that [17.7.39] can be written

$$F_T = (\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{R}'\tilde{\mathbf{Y}}_T\left\{s_T^2\cdot\tilde{\mathbf{Y}}_T\mathbf{R}(\Sigma\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{R}'\tilde{\mathbf{Y}}_T\right\}^{-1} \times \tilde{\mathbf{Y}}_T\mathbf{R}(\mathbf{b}_T - \boldsymbol{\beta})/2. \qquad [17.7.41]$$

The matrix in [17.7.40] has the property that

$$\tilde{\mathbf{Y}}_T\mathbf{R} = \mathbf{R}\mathbf{Y}_T$$

for $\mathbf{R} = [\mathbf{0} \quad \mathbf{I}_2]$ and $\mathbf{Y}_T$ the $(p+1)\times(p+1)$ matrix in [17.7.15]. From [17.7.25], $\mathbf{R}\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} \mathbf{Q}^{-1}\mathbf{h}_2$. Thus, [17.7.41] implies that

$$\begin{aligned}
F_T &= (\mathbf{b}_T - \boldsymbol{\beta})'(\mathbf{R}\mathbf{Y}_T)'\left\{s_T^2\cdot\mathbf{R}\mathbf{Y}_T(\Sigma\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{Y}_T\mathbf{R}'\right\}^{-1}\mathbf{R}\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta})/2 \\
&\xrightarrow{L} (\mathbf{Q}^{-1}\mathbf{h}_2)'\{\sigma^2\mathbf{Q}^{-1}\}^{-1}(\mathbf{Q}^{-1}\mathbf{h}_2)/2 \\
&= \mathbf{h}_2'\mathbf{Q}^{-1}\mathbf{h}_2/(2\sigma^2) \\
&= [1/(2\sigma^2)]\begin{bmatrix}\sigma\cdot W(1) & \tfrac{1}{2}\sigma\lambda\{[W(1)]^2 - 1\}\end{bmatrix} \\
&\quad\times\begin{bmatrix}1 & \lambda\cdot\int W(r)\,dr \\ \lambda\cdot\int W(r)\,dr & \lambda^2\cdot\int [W(r)]^2\,dr\end{bmatrix}^{-1}\begin{bmatrix}\sigma\cdot W(1) \\ \tfrac{1}{2}\sigma\lambda\{[W(1)]^2 - 1\}\end{bmatrix} \\
&= \left(\frac{1}{2\sigma^2}\right)\sigma^2\begin{bmatrix}W(1) & \frac{1}{2}\{[W(1)]^2 - 1\}\end{bmatrix}\begin{bmatrix}1 & 0 \\ 0 & \lambda\end{bmatrix} \qquad [17.7.42] \\
&\quad\times\begin{bmatrix}1 & 0 \\ 0 & \lambda\end{bmatrix}^{-1}\begin{bmatrix}1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr\end{bmatrix}^{-1}\begin{bmatrix}1 & 0 \\ 0 & \lambda\end{bmatrix}^{-1} \\
&\quad\times\begin{bmatrix}1 & 0 \\ 0 & \lambda\end{bmatrix}\begin{bmatrix}W(1) \\ \tfrac{1}{2}\{[W(1)]^2 - 1\}\end{bmatrix} \\
&= \tfrac{1}{2}\begin{bmatrix}W(1) & \tfrac{1}{2}\{[W(1)]^2 - 1\}\end{bmatrix} \\
&\quad\times\begin{bmatrix}1 & \int W(r)\,dr \\ \int W(r)\,dr & \int [W(r)]^2\,dr\end{bmatrix}^{-1}\begin{bmatrix}W(1) \\ \tfrac{1}{2}\{[W(1)]^2 - 1\}\end{bmatrix}.
\end{aligned}$$

This is identical to the asymptotic distribution of the F test when the regression does not include lagged Δy and the disturbances are i.i.d. Thus, the F statistic in [17.7.41] based on *OLS* estimation of [17.7.11] can be compared with the case 2 section of Table B.7 without corrections.

Finally, consider a hypothesis test involving a restriction[12] across $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$ and ρ,

$$H_0\colon r_1\zeta_1 + r_2\zeta_2 + \cdots + r_{p-1}\zeta_{p-1} + 0\cdot\alpha + r_{p+1}\rho = r$$

[12]Since the maintained assumption is that $\rho = 1$, this is a slightly unnatural way to write a hypothesis. Nevertheless, framing the hypothesis this way will shortly prove useful in deriving the asymptotic distribution of an autoregression estimated in the usual form without the Dickey-Fuller transformation.

or

$$\mathbf{r}'\boldsymbol{\beta} = r. \qquad [17.7.43]$$

The distribution of the t test of this hypothesis will be dominated asymptotically by the parameters with the slowest rate of convergence, namely, $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$. Since these are asymptotically Gaussian, the test statistic is asymptotically Gaussian and so can be compared with the usual t tables. To demonstrate this formally, note that the usual t statistic for testing this hypothesis is

$$t_T = \frac{\mathbf{r}'\mathbf{b}_T - r}{\left\{s_T^2\mathbf{r}'(\Sigma\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{r}\right\}^{1/2}} = \frac{T^{1/2}(\mathbf{r}'\mathbf{b}_T - r)}{\left\{s_T^2 T^{1/2}\mathbf{r}'(\Sigma\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{r}T^{1/2}\right\}^{1/2}}. \qquad [17.7.44]$$

Define $\tilde{\mathbf{r}}_T$ to be the vector that results when the last element of $\mathbf{r}$ is replaced by $r_{p+1}/\sqrt{T}$,

$$\tilde{\mathbf{r}}_T' \equiv [r_1 \quad r_2 \quad \cdots \quad r_{p-1} \quad 0 \quad r_{p+1}/\sqrt{T}], \qquad [17.7.45]$$

and notice that

$$T^{1/2}\mathbf{r} = \mathbf{Y}_T\tilde{\mathbf{r}}_T \qquad [17.7.46]$$

for $\mathbf{Y}_T$ the matrix in [17.7.15]. Using [17.7.46] and the null hypothesis that $r = \mathbf{r}'\boldsymbol{\beta}$, expression [17.7.44] can be written

$$t_T = \frac{\tilde{\mathbf{r}}_T'\mathbf{Y}_T(\mathbf{b}_T - \boldsymbol{\beta})}{\left\{s_T^2\tilde{\mathbf{r}}_T'\mathbf{Y}_T(\Sigma\mathbf{x}_t\mathbf{x}_t')^{-1}\mathbf{Y}_T\tilde{\mathbf{r}}_T\right\}^{1/2}}. \qquad [17.7.47]$$

Notice from [17.7.45] that

$$\tilde{\mathbf{r}}_T \to \tilde{\mathbf{r}},$$

where

$$\tilde{\mathbf{r}}' \equiv [r_1 \quad r_2 \quad \cdots \quad r_{p-1} \quad 0 \quad 0].$$

Using this result along with [17.7.18] and [17.7.25] in [17.7.47] produces

$$\begin{aligned} t_T &\xrightarrow{L} \frac{\tilde{\mathbf{r}}'\begin{bmatrix}\mathbf{V}^{-1}\mathbf{h}_1\\ \mathbf{Q}^{-1}\mathbf{h}_2\end{bmatrix}}{\left\{\sigma^2\tilde{\mathbf{r}}'\begin{bmatrix}\mathbf{V}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{Q}^{-1}\end{bmatrix}\tilde{\mathbf{r}}\right\}^{1/2}} \\ &= \frac{[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]\mathbf{V}^{-1}\mathbf{h}_1}{\{\sigma^2[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]\mathbf{V}^{-1}[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]'\}^{1/2}}. \end{aligned} \qquad [17.7.48]$$

Since $\mathbf{h}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, it follows that

$$[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]\mathbf{V}^{-1}\mathbf{h}_1 \sim N(0, h),$$

where

$$h = \sigma^2[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]\mathbf{V}^{-1}[r_1 \quad r_2 \quad \cdots \quad r_{p-1}]'.$$

Thus, the limiting distribution in [17.7.48] is that of a Gaussian scalar divided by its standard deviation and is therefore $N(0, 1)$. This confirms the claim that the t test of $\mathbf{r}'\boldsymbol{\beta} = r$ can be compared with the usual t tables.

One interesting implication of this last result concerns the asymptotic properties of the estimated coefficients if the autoregression is estimated in the usual levels form rather than the transformed regression [17.7.11]. Thus, suppose that the following specification is estimated by *OLS*:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t \qquad [17.7.49]$$

for some $p \geq 2$. Recalling [17.7.2] and [17.7.3], the relation between the estimates $(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_{p-1}, \hat{\rho})$ investigated previously and estimates $(\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)$ based on *OLS* estimation of [17.7.49] is

$$\begin{aligned} \hat{\phi}_p &= -\hat{\zeta}_{p-1} \\ \hat{\phi}_j &= \hat{\zeta}_j - \hat{\zeta}_{j-1} \qquad \text{for } j = 2, 3, \ldots, p-1 \\ \hat{\phi}_1 &= \hat{\rho} + \hat{\zeta}_1. \end{aligned}$$

Thus, each of the coefficients $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ is a linear combination of the elements of $(\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_{p-1}, \hat{\rho})$. The analysis of [17.7.43] establishes that any individual estimate $\hat{\phi}_j$ converges at rate $\sqrt{T}$ to a Gaussian random variable. Recalling the discussion of [16.3.20] and [16.3.21], an *OLS* t or F test based on [17.7.49] is numerically identical to the equivalent t or F test expressed in terms of the representation in [17.7.11]. Thus, the usual t tests associated with hypotheses about any individual coefficients $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ in [17.7.49] can be compared with standard t or $N(0, 1)$ tables. Indeed, any hypothesis about linear combinations of the ϕ's other than the sum $\phi_1 + \phi_2 + \cdots + \phi_p$ satisfies the standard conditions. The sum $\hat{\phi}_1 + \hat{\phi}_2 + \cdots + \hat{\phi}_p$, of course, has the nonstandard distribution of the estimate $\hat{\rho}$ described in [17.7.33].

Summary of Asymptotic Results for an Estimated Autoregression That Includes a Constant Term

The preceding analysis applies to *OLS* estimation of

$$y_t = \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \varepsilon_t$$

under the assumption that the true value of α is zero and the true value of ρ is 1. The other maintained assumptions were that ε_t is i.i.d. with mean zero, variance σ^2, and finite fourth moment and that roots of

$$(1 - \zeta_1 z - \zeta_2 z^2 - \cdots - \zeta_{p-1} z^{p-1}) = 0$$

are outside the unit circle. It was seen that the estimates $\hat{\zeta}_1, \hat{\zeta}_2, \ldots, \hat{\zeta}_{p-1}$ converge at rate $\sqrt{T}$ to Gaussian variates, and standard t or F tests for hypotheses about these coefficients have the usual limiting Gaussian or χ^2 distributions. The estimates $\hat{\alpha}$ and $\hat{\rho}$ converge at rates $\sqrt{T}$ and T, respectively, to nonstandard distributions. If the difference between the *OLS* estimate $\hat{\rho}$ and the hypothesized true value of unity is multiplied by the sample size and divided by $(1 - \hat{\zeta}_1 - \hat{\zeta}_2 - \cdots - \hat{\zeta}_{p-1})$, the resulting statistic has the same asymptotic distribution as the variable tabulated in the case 2 section of Table B.5. The usual t statistic of the hypothesis $\rho = 1$ does not need to be adjusted for sample size or serial correlation and has the same asymptotic distribution as the variable tabulated in the case 2 section of Table B.6. The usual F statistic of the joint hypothesis $\alpha = 0$ and $\rho = 1$ likewise does not have to be adjusted for sample size or serial correlation and has the same distribution as the variable tabulated in the case 2 section of Table B.7.

When the autoregression includes lagged changes as here, tests for a unit root based on the value of ρ, t tests, or F tests are described as *augmented Dickey-Fuller tests.*

Example 17.8

The following model was estimated by *OLS* for the interest rate data described in Example 17.3 (standard errors in parentheses):

$$i_t = \underset{(0.0788)}{0.335}\,\Delta i_{t-1} - \underset{(0.0808)}{0.388}\,\Delta i_{t-2} + \underset{(0.0800)}{0.276}\,\Delta i_{t-3} - \underset{(0.0794)}{0.107}\,\Delta i_{t-4} + \underset{(0.109)}{0.195} + \underset{(0.018604)}{0.96904}\,i_{t-1}.$$

Dates t = 1948:II through 1989:I were used for estimation, so in this case the sample size is $T = 164$. For these estimates, the augmented Dickey-Fuller ρ test [17.7.35] would be

$$\frac{164}{1 - 0.335 + 0.388 - 0.276 + 0.107}(0.96904 - 1) = -5.74.$$

Since $-5.74 > -13.8$, the null hypothesis that the Treasury bill rate follows a fifth-order autoregression with no constant term, and a single unit root, is accepted at the 5% level. The *OLS* t test for this same hypothesis is

$$(0.96904 - 1)/(0.018604) = -1.66.$$

Since $-1.66 > -2.89$, the null hypothesis of a unit root is accepted by the augmented Dickey-Fuller t test as well. Finally, the *OLS* F test of the joint null hypothesis that $\rho = 1$ and $\alpha = 0$ is 1.65. Since this is less than 4.68, the null hypothesis is again accepted.

The null hypothesis that the autoregression in levels requires only four lags is based on the *OLS* t test of $\zeta_4 = 0$:

$$-0.107/0.0794 = -1.35.$$

From Table B.3, the 5% two-sided critical value for a t variable with 158 degrees of freedom is -1.98. Since $-1.35 > -1.98$, the null hypothesis that only four lags are needed for the autoregression in levels is accepted.

Asymptotic Results for Other Autoregressions

Up to this point in this section, we have considered an autoregression that is a generalization of case 2 of Section 17.4—a constant is included in the estimated regression, though the population process is presumed to exhibit no drift. Parallel generalizations for cases 1, 3, and 4 can be obtained in the same fashion. The reader is invited to derive these generalizations in exercises at the end of the chapter. The key results are summarized in Table 17.3.

TABLE 17.3
Summary of Asymptotic Results for Autoregressions Containing a Unit Root

Case 1:

Estimated regression:

$$y_t = \zeta_1\Delta y_{t-1} + \zeta_2\Delta y_{t-2} + \cdots + \zeta_{p-1}\Delta y_{t-p+1} + \rho y_{t-1} + \varepsilon_t$$

True process: same specification as estimated regression with $\rho = 1$

Any t or F test involving $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$ can be compared with the usual t or F tables for an asymptotically valid test.

TABLE 17.3 (continued)

Z_{DF} has the same asymptotic distribution as the variable described under the heading Case 1 in Table B.5.

OLS t test of $\rho = 1$ has the same asymptotic distribution as the variable described under Case 1 in Table B.6.

Case 2:

Estimated regression:

$$y_t = \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \varepsilon_t$$

True process: same specification as estimated regression with $\alpha = 0$ and $\rho = 1$

Any t or F test involving $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$ can be compared with the usual t or F tables for an asymptotically valid test.

Z_{DF} has the same asymptotic distribution as the variable described under Case 2 in Table B.5.

OLS t test of $\rho = 1$ has the same asymptotic distribution as the variable described under Case 2 in Table B.6.

OLS F test of joint hypothesis that $\alpha = 0$ and $\rho = 1$ has the same asymptotic distribution as the variable described under Case 2 in Table B.7.

Case 3:

Estimated regression:

$$y_t = \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \varepsilon_t$$

True process: same specification as estimated regression with $\alpha \neq 0$ and $\rho = 1$

$\hat{\rho}_T$ converges at rate $T^{3/2}$ to a Gaussian variable; all other estimated coefficients converge at rate $T^{1/2}$ to Gaussian variables.

Any t or F test involving any coefficients from the regression can be compared with the usual t or F tables for an asymptotically valid test.

Case 4:

Estimated regression:

$$y_t = \zeta_1 \Delta y_{t-1} + \zeta_2 \Delta y_{t-2} + \cdots + \zeta_{p-1} \Delta y_{t-p+1} + \alpha + \rho y_{t-1} + \delta t + \varepsilon_t$$

True process: same specification as estimated regression with α any value, $\rho = 1$, and $\delta = 0$

Any t or F test involving $\zeta_1, \zeta_2, \ldots, \zeta_{p-1}$ can be compared with the usual t or F tables for an asymptotically valid test.

Z_{DF} has the same asymptotic distribution as the variable described under Case 4 in Table B.5.

OLS t test of $\rho = 1$ has the same asymptotic distribution as the variable described under Case 4 in Table B.6.

OLS F test of joint hypothesis that $\rho = 1$ and $\delta = 0$ has the same asymptotic distribution as the variable described under Case 4 in Table B.7.

Notes to Table 17.3

Estimated regression indicates the form in which the regression is estimated, using observations $t = 1, 2, \ldots, T$ and conditioning on observations $t = 0, -1, \ldots, -p + 1$.

True process describes the null hypothesis under which the distribution is calculated. In each case it is assumed that roots of

$$(1 - \zeta_1 z - \zeta_2 z^2 - \cdots - \zeta_{p-1} z^{p-1}) = 0$$

are all outside the unit circle and that ε_t is i.i.d. with mean zero, variance σ^2, and finite fourth moment.

Z_{DF} in each case is the following statistic:

$$Z_{DF} \equiv T(\hat{\rho}_T - 1)/(1 - \hat{\zeta}_{1,T} - \hat{\zeta}_{2,T} - \cdots - \hat{\zeta}_{p-1,T}),$$

where $\hat{\rho}_T, \hat{\zeta}_{1,T}, \hat{\zeta}_{2,T}, \ldots, \hat{\zeta}_{p-1,T}$ are the *OLS* estimates from the indicated regression.

OLS t test of $\rho = 1$ is $(\hat{\rho}_T - 1)/\hat{\sigma}_{\hat{\rho}_T}$, where $\hat{\sigma}_{\hat{\rho}_T}$ is the *OLS* standard error of $\hat{\rho}_T$.

OLS F test of a hypothesis involving two restrictions is given by expression [17.7.39].

Example 17.9

The following autoregression was estimated by *OLS* for the GNP data in Figure 17.3 (standard errors in parentheses):

$$y_t = \underset{(0.0777)}{0.329}\,\Delta y_{t-1} + \underset{(0.0813)}{0.209}\,\Delta y_{t-2} - \underset{(0.0818)}{0.084}\,\Delta y_{t-3} - \underset{(0.0788)}{0.075}\,\Delta y_{t-4} + \underset{(13.57)}{35.92} + \underset{(0.019386)}{0.94969}\,y_{t-1} + \underset{(0.0152)}{0.0378}\,t.$$

Here, $T = 164$ and the augmented Dickey-Fuller ρ test is

$$\frac{164}{1 - 0.329 - 0.209 + 0.084 + 0.075}(0.94969 - 1) = -13.3.$$

Since $-13.3 > -21.0$, the null hypothesis that the log of GNP is *ARIMA*(4, 1, 0) with possible drift is accepted at the 5% level. The augmented Dickey-Fuller t test also accepts this hypothesis:

$$(0.94969 - 1)/(0.019386) = -2.60 > -3.44.$$

The *OLS* F test of the joint null hypothesis that $\rho = 1$ and $\delta = 0$ is $3.74 < 6.42$, and so the augmented Dickey-Fuller F test is also consistent with the unit root specification.

Unit Root AR(p) *Processes with* p *Unknown*

Various suggestions have been proposed for how to proceed when the process is regarded as *ARIMA*(p, 1, 0) with p unknown but finite. One simple approach is to estimate [17.7.11] with p taken to be some prespecified upper bound $\bar{p}$. The *OLS* t test of $\zeta_{\bar{p}-1} = 0$ can then be compared with the usual critical value for a t statistic from Table B.3. If the null hypothesis is accepted, the *OLS* F test of the joint null hypothesis that both $\zeta_{\bar{p}-1} = 0$ and $\zeta_{\bar{p}-2} = 0$ can be compared with the usual $F(2, T - k)$ distribution in Table B.4. The procedure continues sequentially until the joint null hypothesis that $\zeta_{\bar{p}-1} = 0, \zeta_{\bar{p}-2} = 0, \ldots, \zeta_{\bar{p}-\ell} = 0$ is rejected for some ℓ. The recommended regression is then

$$y_t = \zeta_1\Delta y_{t-1} + \zeta_2\Delta y_{t-2} + \cdots + \zeta_{\bar{p}-\ell}\Delta y_{t-\bar{p}+\ell} + \alpha + \rho y_{t-1} + \delta t.$$

If no value of ℓ leads to rejection, the simple Dickey-Fuller test of Table 17.1 is used. Hall (1991) discussed a variety of alternative strategies for estimating p.

Just as in the Phillips-Perron consideration of the *MA*(∞) case, the researcher might want to choose bigger values for p, the autoregressive lag length, the larger is the sample size T. Said and Dickey (1984) showed that as long as p goes to infinity sufficiently slowly relative to T, then the *OLS* t test of $\rho = 1$ can continue to be compared with the Dickey-Fuller values in Table B.6.

Again, it is worthwhile to keep in mind that there always exists a p such that an *ARIMA*(p, 1, 0) representation can describe a stationary process arbitrarily well for a given sample. The Said-Dickey test of $\rho = 1$ might therefore best be viewed as follows. For a given fixed p, we can certainly ask whether an *ARIMA*($p - 1$, 1, 0) describes the data nearly as well as an *ARIMA*(p, 0, 0). Imposing $\rho = 1$ when the true value of ρ is close to unity may improve forecasts and small-sample estimates of the other parameters. The Said-Dickey result permits the researcher to use a larger value of p on which to base this comparison the larger is the sample size T.

17.8. *Other Approaches to Testing for Unit Roots*

This section briefly describes some alternative approaches to testing for unit roots.

Variance Ratio Tests

Let

$$\Delta y_t = \alpha + u_t,$$

where

$$u_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \equiv \psi(L)\varepsilon_t$$

for ε_t a white noise sequence with variance σ^2. Recall from expression [15.3.10] that the permanent effect of ε_t on the level of y_{t+s} is given by

$$\lim_{s\to\infty} \frac{\partial y_{t+s}}{\partial \varepsilon_t} = \psi(1).$$

If y_t is stationary or stationary around a deterministic time trend, an innovation ε_t has no permanent effect on y, requiring $\psi(1) = 0$.

Cochrane (1988) and Lo and MacKinlay (1988) proposed a test for unit roots that exploits this property. Consider the change in y over s periods,

$$y_{t+s} - y_t = \alpha s + u_{t+s} + u_{t+s-1} + \cdots + u_{t+1}, \qquad [17.8.1]$$

and notice that

$$(y_{t+s} - y_t)/s = \alpha + s^{-1}(u_{t+s} + u_{t+s-1} + \cdots + u_{t+1}). \qquad [17.8.2]$$

The second term in [17.8.2] could be viewed as the sample mean of s observations drawn from the process followed by u. Thus, Proposition 7.5(b) and result [7.2.8] imply that

$$\lim_{s\to\infty} s\cdot \mathrm{Var}[s^{-1}(u_{t+s} + u_{t+s-1} + \cdots + u_{t+1})] = \sigma^2\cdot[\psi(1)]^2. \qquad [17.8.3]$$

Let $\hat{\alpha}_T$ denote the average change in y in a sample of T observations:

$$\hat{\alpha}_T = T^{-1} \sum_{t=1}^{T} (y_t - y_{t-1}).$$

Consider the following estimate of the variance of the change in y over its value s periods earlier:

$$\hat{J}_T(s) = T^{-1} \sum_{t=0}^{T-s} (y_{t+s} - y_t - \hat{\alpha}_T s)^2. \qquad [17.8.4]$$

This should converge in probability to

$$J(s) = E(y_{t+s} - y_t - \alpha s)^2 = E(u_{t+s} + u_{t+s-1} + \cdots + u_{t+1})^2 \qquad [17.8.5]$$

as the sample size T becomes large. Comparing this expression with [17.8.3],

$$\lim_{s\to\infty} s^{-1}\cdot J(s) = \sigma^2\cdot[\psi(1)]^2.$$

Cochrane (1988) therefore proposed calculating [17.8.4] as a function of s. If the true process for y_t is stationary or stationary around a deterministic time trend,

this statistic should go to zero for large s. If the true process for y_t is $I(1)$, this statistic gives a measure of the quantitative importance of permanent effects of ε as reflected in the long-run multiplier $\psi(1)$. However, the statistic in [17.8.4] is not reliable unless s is much smaller than T.

If the data truly followed a random walk so that $\psi(L) = 1$, then $J(s)$ in [17.8.5] would equal $s \cdot \sigma^2$ for any s, where σ^2 is the variance of u_t. Lo and MacKinlay (1988) exploited this property to suggest tests of the random walk hypothesis based on alternative values of s. See Lo and MacKinlay (1989) and Cecchetti and Lam (1991) for evidence on the small-sample properties of these tests.

Other Tests for Unit Roots

The Phillips-Perron approach was based on an $MA(\infty)$ representation for Δy_t, while the Said-Dickey approach was based on an $AR(\infty)$ representation. Tests based on a finite $ARMA(p, q)$ representation for Δy_t have been explored by Said and Dickey (1985), Hall (1989), Said (1991), and Pantula and Hall (1991).

A number of other approaches to testing for unit roots have been proposed, including Sargan and Bhargava (1983), Solo (1984), Bhargava (1986), Dickey and Pantula (1987), Park and Choi (1988), Schmidt and Phillips (1992), Stock (1991), and Kwiatkowski, Phillips, Schmidt, and Shin (1992). See Stock (1993) for an excellent survey. Asymptotic inference for processes with near unit root behavior has been discussed by Chan and Wei (1987), Phillips (1988), and Sowell (1990).

17.9. *Bayesian Analysis and Unit Roots*

Up to this point in the chapter we have adopted a classical statistical perspective, calculating the distribution of $\hat{\rho}$ conditional on a particular value of ρ such as $\rho = 1$. This section considers the Bayesian perspective, in which the true value of ρ is regarded as a random variable and the goal is to describe the distribution of this random variable conditional on the data.

Recall from Proposition 12.3 that if the prior density for the vector of unknown coefficients $\boldsymbol{\beta}$ and innovation precision σ^{-2} is of the Normal-gamma form of [12.1.19] and [12.1.20], then the posterior distribution of $\boldsymbol{\beta}$ conditional on the data is multivariate t. This result holds exactly for any finite sample and holds regardless of whether the process is stationary. Thus, in the case of the diffuse prior distribution represented by $N = \lambda = 0$ and $\mathbf{M}^{-1} = \mathbf{0}$, a Bayesian would essentially use the usual t and F statistics in the standard way.

How can the classical distribution of $\hat{\rho}$ be strongly skewed while the Bayesian distribution of ρ is that of a symmetric t variable? Sims (1988) and Sims and Uhlig (1991) provided a detailed discussion of this question. The classical test of the null hypothesis $\rho = 1$ is based only on the distribution of $\hat{\rho}$ when the true value of ρ is unity. By contrast, the Bayesian inference is based on the distribution of $\hat{\rho}|\rho$ for all the possible values of ρ, with the distribution of $\hat{\rho}|\rho$ weighted according to the prior probability for ρ. If the distribution of $\hat{\rho}|\rho$ had the same skew and dispersion for every ρ as it does at $\rho = 1$, then we would conclude that, having observed any particular $\hat{\rho}$, the true value of ρ is probably somewhat higher. However, the distribution of $\hat{\rho}|\rho$ changes with ρ—the lower the true value of ρ, the smaller the skew and the greater the dispersion, since from [17.1.3] the variance of $\sqrt{T}(\hat{\rho} - \rho)$ is approximately $(1 - \rho^2)$. Because lower values of ρ imply greater dispersion for $\hat{\rho}$, in the absence of skew we would suspect that a given observation $\hat{\rho} = 0.95$ was more likely to have been generated by a distribution centered at $\rho = 0.90$ with

large dispersion than by a distribution centered at $\rho = 1$ with small dispersion. The effects of skew and dispersion turn out to cancel, so that with a uniform prior distribution for the value of ρ, having observed $\hat{\rho} = 0.95$, it is just as likely that the true value of ρ is greater than 0.95 as that the true value of ρ is less than 0.95.

Example 17.10

For the GNP regression in Example 17.9, the probability that $\rho \geq 1$ conditional on the data is the probability that a t variable with $T = 164$ degrees of freedom[13] exceeds $(1 - 0.94969)/0.019386 = 2.60$. From Table B.3, this probability is around 0.005. Hence, although the value of ρ must be large, it is unlikely to be as big as unity.

The contrast between the Bayesian inference in Example 17.10 and the classical inference in Example 17.9 is one of the reasons given by Sims (1988) and Sims and Uhlig (1991) for preferring Bayesian methods. Note that the probability calculated in Example 17.10 will be less than 0.025 if and only if a classical 95% confidence interval around the point estimate $\hat{\rho}$ does not contain unity. Thus, an alternative way of describing the finding of Example 17.10 is that the standard asymptotic classical confidence region around $\hat{\rho}$ does not include $\rho = 1$. Even so, Example 17.9 showed that the null hypothesis of a unit root is accepted by the augmented Dickey-Fuller test. The classical asymptotic confidence region centered at $\rho = \hat{\rho}$ seems inconsistent with a unit root, while the classical asymptotic confidence region centered at $\rho = 1$ supports a unit root. Such disconnected confidence regions resulting from the classical approach may seem somewhat troublesome and counterintuitive.[14] By contrast, the Bayesian has a single, consistent summary of the plausibility of different values of ρ, which is that implied by the posterior distribution of ρ conditional on the data.

One could, of course, use a prior distribution that reflected more confidence in the prior information about the value of ρ. As long as the prior distribution was in the Normal-gamma class, this would cause us to shift the point estimate 0.94969 in the direction of the prior mean and reduce the standard error and increase the degrees of freedom as warranted by the prior information, but a t distribution would still be used to interpret the resulting statistic.

Although the Normal-gamma class is convenient to work with, it might not be sufficiently flexible to reflect the researcher's true prior beliefs. Sims (1988, p. 470) discussed Bayesian inference in which a point mass with positive probability is placed on the possibility that $\rho = 1$. DeJong and Whiteman (1991) used numerical methods to calculate posterior distributions under a range of prior distributions defined numerically and concluded that the evidence for unit roots in many key economic time series is quite weak.

Phillips (1991a) noted that there is a prior distribution for which the Bayesian inference mimics the classical approach. He argued that the diffuse prior distribution of Proposition 12.3 is actually highly informative in a time series regression and suggested instead a prior distribution due to Jeffreys (1946). Although this prior distribution has some theoretical arguments on its behalf, it has the unusual property in this application that the prior distribution is a function of the sample size T—Phillips would propose using a different prior distribution for $f(\rho)$ when

[13]Recall from Proposition 12.3(b) that the degrees of freedom are given by $N^* = N + T$. Thus, the Bayesian interpretation is not quite identical to the classical t statistic, whose degrees of freedom would be $T - k$.

[14]Stock (1991) has recently proposed a solution to this problem from the classical perspective. Another approach is to rely on the exact small-sample distribution, as advocated by Andrews (1993).

the analyst is going to obtain a sample of size 50 than when the analyst is going to obtain a sample of size 100. This would not be appropriate if the prior distribution is intended to represent the actual information available to the analyst before seeing the data. Phillips (1991b, pp. 468–69) argued that, in order to be truly uninformative, a prior distribution in this context would have this property, since the larger the true value of ρ, the more rapidly information about ρ contained in the sample $\{y_1, y_2, \ldots, y_T\}$ is going to accumulate with the sample size T. Certainly the concept of what it means for a prior distribution to be "uninformative" can be difficult and controversial.[15]

The potential difficulty in persuading others of the validity of one's prior beliefs has always been the key weakness of Bayesian statistics, and it seems unavoidable here. The best a Bayesian can do may be to take an explicit stand on the nature and strength of prior information and defend it as best as possible. If the nature of the prior information is that all values of ρ are equally likely, then it is satisfactory to use the standard *OLS* t and F tests in the usual way. If one is unwilling to take such a stand, then Sims and Uhlig urged that investigators report both the classical hypothesis test of $\rho = 1$ and the classical confidence region around $\hat{\rho}$ and let the reader interpret the results as he or she sees fit.

APPENDIX 17.A. *Proofs of Chapter 17 Propositions*

■ **Proof of Proposition 17.2.** Observe that

$$\begin{aligned}
\sum_{s=1}^{t} u_s &= \sum_{s=1}^{t} \sum_{j=0}^{\infty} \psi_j \varepsilon_{s-j} \\
&= \{\psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots + \psi_t\varepsilon_0 + \psi_{t+1}\varepsilon_{-1} + \cdots\} \\
&\quad + \{\psi_0\varepsilon_{t-1} + \psi_1\varepsilon_{t-2} + \psi_2\varepsilon_{t-3} + \cdots + \psi_{t-1}\varepsilon_0 + \psi_t\varepsilon_{-1} + \cdots\} \\
&\quad + \{\psi_0\varepsilon_{t-2} + \psi_1\varepsilon_{t-3} + \psi_2\varepsilon_{t-4} + \cdots + \psi_{t-2}\varepsilon_0 + \psi_{t-1}\varepsilon_{-1} + \cdots\} \\
&\quad + \cdots + \{\psi_0\varepsilon_1 + \psi_1\varepsilon_0 + \psi_2\varepsilon_{-1} + \cdots\} \\
&= \psi_0\varepsilon_t + (\psi_0 + \psi_1)\varepsilon_{t-1} + (\psi_0 + \psi_1 + \psi_2)\varepsilon_{t-2} + \cdots \\
&\quad + (\psi_0 + \psi_1 + \psi_2 + \cdots + \psi_{t-1})\varepsilon_1 + (\psi_1 + \psi_2 + \cdots + \psi_t)\varepsilon_0 \\
&\quad + (\psi_2 + \psi_3 + \cdots + \psi_{t+1})\varepsilon_{-1} + \cdots \\
&= (\psi_0 + \psi_1 + \psi_2 + \cdots)\varepsilon_t - (\psi_1 + \psi_2 + \psi_3 + \cdots)\varepsilon_t \\
&\quad + (\psi_0 + \psi_1 + \psi_2 + \cdots)\varepsilon_{t-1} - (\psi_2 + \psi_3 + \cdots)\varepsilon_{t-1} \\
&\quad + (\psi_0 + \psi_1 + \psi_2 + \cdots)\varepsilon_{t-2} - (\psi_3 + \psi_4 + \cdots)\varepsilon_{t-2} + \cdots \\
&\quad + (\psi_0 + \psi_1 + \psi_2 + \cdots)\varepsilon_1 - (\psi_t + \psi_{t+1} + \cdots)\varepsilon_1 \\
&\quad + (\psi_1 + \psi_2 + \psi_3 + \cdots)\varepsilon_0 - (\psi_{t+1} + \psi_{t+2} + \cdots)\varepsilon_0 \\
&\quad + (\psi_2 + \psi_3 + \psi_4 + \cdots)\varepsilon_{-1} - (\psi_{t+2} + \psi_{t+3} + \cdots)\varepsilon_{-1} + \cdots
\end{aligned}$$

or

$$\sum_{s=1}^{t} u_s = \psi(1)\cdot\sum_{s=1}^{t} \varepsilon_s + \eta_t - \eta_0, \qquad [17.A.1]$$

where

$$\begin{aligned}
\eta_t &\equiv -(\psi_1 + \psi_2 + \psi_3 + \cdots)\varepsilon_t - (\psi_2 + \psi_3 + \psi_4 + \cdots)\varepsilon_{t-1} \\
&\quad - (\psi_3 + \psi_4 + \psi_5 + \cdots)\varepsilon_{t-2} - \cdots \\
\eta_0 &= -(\psi_1 + \psi_2 + \psi_3 + \cdots)\varepsilon_0 - (\psi_2 + \psi_3 + \psi_4 + \cdots)\varepsilon_{-1} \\
&\quad - (\psi_3 + \psi_4 + \psi_5 + \cdots)\varepsilon_{-2} - \cdots.
\end{aligned}$$

[15]See the many comments accompanying Phillips (1991a).

Notice that $\eta_t = \Sigma_{j=0}^{\infty}\alpha_j\varepsilon_{t-j}$, where $\alpha_j = -(\psi_{j+1} + \psi_{j+2} + \cdots)$, with $\{\alpha_j\}_{j=0}^{\infty}$ absolutely summable:

$$\begin{aligned}\sum_{j=0}^{\infty}|\alpha_j| &= |\psi_1 + \psi_2 + \psi_3 + \cdots| + |\psi_2 + \psi_3 + \psi_4 + \cdots| + |\psi_3 + \psi_4 + \psi_5 + \cdots| + \cdots \\ &\le \{|\psi_1| + |\psi_2| + |\psi_3| + \cdots\} + \{|\psi_2| + |\psi_3| + |\psi_4| + \cdots\} \\ &\quad + \{|\psi_3| + |\psi_4| + |\psi_5| + \cdots\} + \cdots \\ &= |\psi_1| + 2|\psi_2| + 3|\psi_3| + \cdots \\ &= \sum_{j=0}^{\infty} j\cdot|\psi_j|,\end{aligned}$$

which is bounded by the assumptions in Proposition 17.2. ■

■ **Proof of Proposition 17.3.**

(a) This was shown in [17.5.9].
(b) This follows from [7.2.17] and the fact that $E(u_t^2) = \gamma_0$.
(c) This is implied by [7.2.14].
(d) Since $\xi_t = \Sigma_{s=1}^{t} u_s$, Proposition 17.2 asserts that

$$\xi_t = \psi(1)\sum_{s=1}^{t}\varepsilon_s + \eta_t - \eta_0. \qquad [17.A.2]$$

Hence,

$$\begin{aligned}T^{-1}\sum_{t=1}^{T}\xi_{t-1}\varepsilon_t &= T^{-1}\sum_{t=2}^{T}\left(\psi(1)\sum_{s=1}^{t-1}\varepsilon_s + \eta_{t-1} - \eta_0\right)\varepsilon_t \\ &= \psi(1)\cdot T^{-1}\sum_{t=2}^{T}(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-1})\varepsilon_t \qquad [17.A.3] \\ &\quad + T^{-1}\sum_{t=2}^{T}(\eta_{t-1} - \eta_0)\varepsilon_t.\end{aligned}$$

But [17.3.26] established that

$$T^{-1}\sum_{t=2}^{T}(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t-1})\varepsilon_t \xrightarrow{L} (1/2)\sigma^2\cdot\{[W(1)]^2 - 1\}. \qquad [17.A.4]$$

Furthermore, Proposition 17.2 ensures that $\{(\eta_{t-1} - \eta_0)\varepsilon_t\}_{t=1}^{\infty}$ is a martingale difference sequence with finite variance, and so, from Example 7.11,

$$T^{-1}\sum_{t=2}^{T}(\eta_{t-1} - \eta_0)\varepsilon_t \xrightarrow{p} 0. \qquad [17.A.5]$$

Substituting [17.A.4] and [17.A.5] into [17.A.3] yields

$$T^{-1}\sum_{t=1}^{T}\xi_{t-1}\varepsilon_t \xrightarrow{L} (1/2)\sigma^2\cdot[\psi(1)]\cdot\{[W(1)]^2 - 1\}, \qquad [17.A.6]$$

as claimed in (d).

(e) For $j = 0$ we have from [17.1.11] that

$$T^{-1}\sum_{t=1}^{T}\xi_{t-1}u_t = (1/2)T^{-1}\xi_T^2 - (1/2)T^{-1}(u_1^2 + u_2^2 + \cdots + u_T^2). \qquad [17.A.7]$$

But

$$T^{-1}\xi_T^2 = [T^{-1/2}(u_1 + u_2 + \cdots + u_T)]^2 \xrightarrow{L} \lambda^2\cdot[W(1)]^2 \qquad [17.A.8]$$

from result (a). Also,

$$T^{-1}(u_1^2 + u_2^2 + \cdots + u_T^2) \xrightarrow{p} \gamma_0$$

from result (c). Thus, [17.A.7] converges to

$$T^{-1}\sum_{t=1}^{T}\xi_{t-1}u_t \xrightarrow{L} (1/2)\{\lambda^2\cdot[W(1)]^2 - \gamma_0\}, \qquad [17.A.9]$$

which establishes result (e) for $j = 0$.

For $j > 0$, observe that

$$\xi_{t-1} = \xi_{t-j-1} + u_{t-j} + u_{t-j+1} + \cdots + u_{t-1},$$

implying that

$$\begin{aligned} T^{-1}\sum_{t=j+1}^{T}\xi_{t-1}u_{t-j} &= T^{-1}\sum_{t=j+1}^{T}(\xi_{t-j-1} + u_{t-j} + u_{t-j+1} + \cdots + u_{t-1})u_{t-j} \\ &= T^{-1}\sum_{t=j+1}^{T}\xi_{t-j-1}u_{t-j} \qquad [17.A.10] \\ &\quad + T^{-1}\sum_{t=j+1}^{T}(u_{t-j} + u_{t-j+1} + \cdots + u_{t-1})u_{t-j}. \end{aligned}$$

But

$$T^{-1}\sum_{t=j+1}^{T}\xi_{t-j-1}u_{t-j} = [(T-j)/T]\cdot(T-j)^{-1}\sum_{t=1}^{T-j}\xi_{t-1}u_t \xrightarrow{L} (1/2)\{\lambda^2\cdot[W(1)]^2 - \gamma_0\},$$

as in [17.A.9]. Also,

$$T^{-1}\sum_{t=j+1}^{T}(u_{t-j} + u_{t-j+1} + \cdots + u_{t-1})u_{t-j} \xrightarrow{p} \gamma_0 + \gamma_1 + \gamma_2 + \cdots + \gamma_{j-1},$$

from result (c). This, [17.A.10] converges to

$$T^{-1}\sum_{t=j+1}^{T}\xi_{t-1}u_{t-j} \xrightarrow{L} (1/2)\{\lambda^2\cdot[W(1)]^2 - \gamma_0\} + \{\gamma_0 + \gamma_1 + \gamma_2 + \cdots + \gamma_{j-1}\}.$$

Clearly, $T^{-1}\Sigma_{t=1}^{T}\xi_{t-1}u_{t-j}$ has the same asymptotic distribution, since

$$T^{-1}\sum_{t=1}^{j}\xi_{t-1}u_{t-j} \xrightarrow{p} 0.$$

(f) From the definition of ξ_t in [17.5.11] and $X_T(r)$ in [17.5.4], it follows as in [17.3.15] that

$$\int_0^1 \sqrt{T}\cdot X_T(r)\,dr = T^{-3/2}\sum_{t=1}^{T}\xi_{t-1}.$$

Result (f) then follows immediately from [17.5.5].

(g) First notice that

$$T^{-3/2}\sum_{t=1}^{T}tu_{t-j} = T^{-3/2}\sum_{t=1}^{T}(t-j+j)u_{t-j},$$

where $j\cdot T^{-3/2}\Sigma_{t=1}^{T}u_{t-j} \xrightarrow{p} 0$. Hence,

$$T^{-3/2}\sum_{t=1}^{T}tu_{t-j} \xrightarrow{p} T^{-3/2}\sum_{t=1}^{T}(t-j)u_{t-j} \xrightarrow{p} T^{-3/2}\sum_{t=1}^{T}tu_t.$$

But from [17.3.19],

$$T^{-3/2}\sum_{t=1}^{T}tu_t = T^{-1/2}\sum_{t=1}^{T}u_t - T^{-3/2}\sum_{t=1}^{T}\xi_{t-1} \xrightarrow{L} \lambda\cdot W(1) - \lambda\cdot\int_0^1 W(r)\,dr,$$

by virtue of (a) and (f).

(h) Using the same analysis as in [17.3.20] through [17.3.22], for ξ_t defined in [17.5.11]