Introduction to Statistical Time Series
WAYNE A. FULLER

CHAPTER 8

Parameter Estimation

8.1. FIRST ORDER AUTOREGRESSIVE TIME SERIES

The stationary time series defined by

$$Y_t - \mu = \rho(Y_{t-1} - \mu) + e_t, \tag{8.1.1}$$

where the e_t are normal independent $(0, \sigma^2)$ random variables and $|\rho| < 1$, is one of the simplest and most heavily used models in time series analysis. It is often a satisfactory representation of the error time series in economic models. This model also underlies many tests of the hypothesis that the observed time series is a sequence of independently and identically distributed random variables.

One estimator of ρ is the first order autocorrelation coefficient discussed in Chapter 6,

$$\hat{r}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\sum_{t=1}^{n-1}(Y_t - \bar{y}_n)(Y_{t+1} - \bar{y}_n)}{\sum_{t=1}^{n}(Y_t - \bar{y}_n)^2}, \tag{8.1.2}$$

where $\bar{y}_n = n^{-1}\sum_{t=1}^{n} Y_t$. To introduce some other estimators of the parameter ρ, let us consider the distribution of the Y_t for the normal time series defined by (8.1.1).

The expected value of Y_t is μ, and the expected value of $Y_t - \rho Y_{t-1}$ is $\lambda = (1-\rho)\mu$. For a sample of n observations we can write

$$\begin{aligned} Y_1 &= \mu + v_1, \\ Y_t - \mu &= \rho(Y_{t-1} - \mu) + e_t, \qquad t = 2, 3, \ldots, n, \end{aligned} \tag{8.1.3}$$

or

$$\begin{aligned} Y_1 &= \mu + v_1 \\ Y_t &= \lambda + \rho Y_{t-1} + e_t, \qquad t = 2, 3, \ldots, n, \end{aligned}$$

where the vector $(v_1, e_2, e_3, \ldots, e_n)$ is distributed as a multivariate normal with zero mean and covariance matrix

$$\Sigma = \operatorname{diag}\{(1-\rho^2)^{-1}\sigma^2, \sigma^2, \sigma^2, \ldots, \sigma^2\}. \tag{8.1.4}$$

It follows that twice the logarithm of the likelihood of a sample of n observations is

$$2\log L(y\colon \mu, \rho, \sigma^2) = -n\log 2\pi - n\log\sigma^2 + \log(1-\rho^2) - \sigma^{-2}\Big\{(Y_1-\mu)^2(1-\rho^2) + \sum_{t=2}^{n}[(Y_t-\mu)-\rho(Y_{t-1}-\mu)]^2\Big\}. \tag{8.1.5}$$

The computation of the maximum likelihood estimators is greatly simplified if we treat Y_1 as fixed and investigate the conditional likelihood. This is also an appropriate model in some experimental situations. For example, if we initiate an experiment at time 1 with an initial input of Y_1, it is very reasonable to condition on this initial input.

To construct the conditional likelihood, we consider the last $n-1$ equations of (8.1.3). Maximizing twice the logarithm of the likelihood,

$$2\log L(y\colon \mu, \rho, \sigma^2 \mid Y_1) = -(n-1)\log 2\pi - (n-1)\log\sigma^2 - \sigma^{-2}\sum_{t=2}^{n}(Y_t-\lambda-\rho Y_{t-1})^2, \tag{8.1.6}$$

leads to the estimators

$$\begin{aligned}\hat{\rho} &= \Big[\sum_{t=2}^{n}(Y_{t-1}-\bar{y}_{(-1)})^2\Big]^{-1}\sum_{t=2}^{n}(Y_t-\bar{y}_{(0)})(Y_{t-1}-\bar{y}_{(-1)}),\\ \hat{\lambda} &= \bar{y}_{(0)} - \hat{\rho}\bar{y}_{(-1)},\\ \hat{\sigma}^2 &= (n-1)^{-1}\sum_{t=2}^{n}[(Y_t-\bar{y}_{(0)})-\hat{\rho}(Y_{t-1}-\bar{y}_{(-1)})]^2,\end{aligned} \tag{8.1.7}$$

where $(\bar{y}_{(-1)}, \bar{y}_{(0)}) = (n-1)^{-1}\sum_{t=2}^{n}(Y_{t-1}, Y_t)$. These estimators are, strictly speaking, not the maximum likelihood estimators for the model stated in (8.1.1). The estimator $\hat{\rho}$ can take on values outside of $(-1, 1)$, while the maximum likelihood estimator is constrained to the parameter space.

The estimators of λ and ρ are those that would be obtained by applying least squares to the last $n-1$ equations. The least squares estimator of σ^2,

$$s^2 = (n-3)^{-1}\sum_{t=2}^{n}[(Y_t-\bar{y}_{(0)})-\hat{\rho}(Y_{t-1}-\bar{y}_{(-1)})]^2, \tag{8.1.8}$$

is typically used in place of $\hat{\sigma}^2$.

The least squares estimator for ρ differs from (8.1.2) by terms whose order in probability is n^{-1}. Therefore, by (6.2.9) and Corollary 6.3.6.1, $n^{1/2}(\hat{\rho}-\rho)$ is approximately normally distributed with mean zero and variance equal to $1-\rho^2$. The limiting distribution is also derived in the next section. Note that the estimator

of ρ defined by (8.1.7) can be greater than one in absolute value, while that defined in (8.1.2) cannot.

Let us now return to a consideration of the unconditional likelihood as given by equation (8.1.5). Differentiating the log likelihood with respect to μ, ρ, and σ^2 and setting the derivatives equal to zero, we obtain

$$\mu = [2 + (n-2)(1-\rho)]^{-1}\left[Y_1 + (1-\rho)\sum_{t=2}^{n-1} Y_t + Y_n\right],$$

$$[(Y_1-\mu)^2 - (1-\rho^2)^{-1}\sigma^2]\rho + \sum_{t=2}^{n}[(Y_t-\mu)-\rho(Y_{t-1}-\mu)](Y_{t-1}-\mu) = 0,$$

$$\sigma^2 = n^{-1}\left\{(Y_1-\mu)^2(1-\rho^2) + \sum_{t=2}^{n}[(Y_t-\mu)-\rho(Y_{t-1}-\mu)]^2\right\}. \quad (8.1.9)$$

If μ is known, Anderson (1971, p. 354) shows that the maximum likelihood estimator of ρ is a root of the cubic equation

$$f(\rho) = \rho^3 + c_1\rho^2 + c_2\rho + c_3 = 0, \quad (8.1.10)$$

where $c_3 = -(n-2)^{-1}nc_1$,

$$c_1 = -(n-2)(n-1)^{-1}\left[\sum_{t=2}^{n-1} y_t^2\right]^{-1}\sum_{t=2}^{n} y_t y_{t-1},$$

$$c_2 = -(n-1)^{-1}\left[n + \left[\sum_{t=2}^{n-1} y_t^2\right]^{-1}\sum_{t=1}^{n} y_t^2\right],$$

and $y_t = Y_t - \mu$. Hasza (1980) gives explicit expressions for the three roots of (8.1.10) and shows that there is a root in each of the intervals $(-\infty, -1)$, $(-1, 1)$ and $(1, \infty)$. If y_t is stationary, then

$$[c_1, c_2] = -[\hat{\rho}_l, (n-2)^{-1}n] + O_p(n^{-1}),$$

where $\hat{\rho}_l = [\Sigma_{t=2}^{n} y_{t-1}^2]^{-1}\Sigma_{t=2}^{n} y_{t-1}y_t$ is the least squares estimator. We show in the next section that $\hat{\rho}_l - \rho_0 = O_p(n^{-1/2})$, where ρ_0 is the true value. Hence,

$$f(\rho) \xrightarrow{p} \rho^3 - \rho_0\rho^2 - \rho + \rho_0 = (\rho^2-1)(\rho-\rho_0).$$

It follows from the results of Section 5.8 that the three roots of $f(\rho) = 0$ converge in probability to -1, ρ_0, and 1, respectively. Therefore, the root in $(-1, 1)$ is consistent for ρ_0. We show in Section 8.4 that the least squares estimator (8.1.7) and the maximum likelihood estimator have the same limiting distribution for stationary processes.

If μ is unknown, Gonzalez–Farias (1992) showed that the unconditional

maximum likelihood estimator is a solution of a fifth degree polynomial. A numerical solution can be obtained by iterating equation (8.1.10) and the estimator for μ in (8.1.9), beginning with $\hat{\mu} = \bar{y}_n$.

8.2. HIGHER ORDER AUTOREGRESSIVE TIME SERIES

8.2.1. Least Squares Estimation for Univariate Processes

In this section we study the ordinary least squares estimators of the pth order autoregressive process. Consider the time series

$$Y_t + \sum_{i=1}^{p} \alpha_i Y_{t-i} = \theta_0 + e_t, \tag{8.2.1}$$

where the roots of

$$m^p + \sum_{i=1}^{p} \alpha_i m^{p-i} = 0 \tag{8.2.2}$$

are less than one in absolute value and the e_t are uncorrelated $(0, \sigma^2)$ random variables with additional properties to be specified. Because the procedures of this section are closely related to multiple regression, it is convenient to write (8.2.1) as

$$Y_t = \theta_0 + \theta_1 Y_{t-1} + \theta_2 Y_{t-2} + \cdots + \theta_p Y_{t-p} + e_t, \tag{8.2.3}$$

where $\theta_i = -\alpha_i$, $i = 1, 2, \ldots, p$. If the process is stationary, the expression

$$Y_t - \mu = \sum_{i=1}^{p} \theta_i (Y_{t-i} - \mu) + e_t, \tag{8.2.4}$$

where $E\{Y_t\} = \mu = (1 + \Sigma_{i=1}^{p} \alpha_i)^{-1} \theta_0$, is also useful.

The ordinary least squares estimator of $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)'$ is

$$\hat{\boldsymbol{\theta}} = \left[\sum_{t=p+1}^{n} \mathbf{X}_t' \mathbf{X}_t \right]^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t' Y_t, \tag{8.2.5}$$

where $\mathbf{X}_t = (1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$. The error in this estimator is

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left[(n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t' \mathbf{X}_t \right]^{-1} (n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t' e_t. \tag{8.2.6}$$

By the properties of autoregressive processes, $E\{Y_{t-j} e_t\} = 0$ for $j > 0$ and $E\{\mathbf{X}_t' e_t\} = \mathbf{0}$. Also, if the process is stationary with autocovariance function $\gamma(h)$, then

$$E\left\{(n-p)^{-1}\sum_{t=p+1}^{n} Y_{t-i}Y_{t-j}\right\} = \gamma(|i-j|) + \mu^2,$$

and under the conditions of Theorem 6.2.1,

$$(n-p)^{-1}\sum_{t=p+1}^{n} Y_{t-i}Y_{t-j} = \gamma(|i-j|) + \mu^2 + O_p(n^{-1/2}). \tag{8.2.7}$$

It follows that the error in $\hat{\boldsymbol{\theta}}$ is $O_p(n^{-1/2})$. Also, by (8.2.7) the least squares estimator is asymptotically equivalent to the estimator

$$\begin{pmatrix} \hat{\theta}_1^{\dagger} \\ \hat{\theta}_2^{\dagger} \\ \vdots \\ \hat{\theta}_p^{\dagger} \end{pmatrix} = \begin{pmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \cdots & \hat{\gamma}(p-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \cdots & \hat{\gamma}(p-2) \\ \vdots & \vdots & & \vdots \\ \hat{\gamma}(p-1) & \hat{\gamma}(p-2) & \cdots & \hat{\gamma}(0) \end{pmatrix}^{-1} \begin{pmatrix} \hat{\gamma}(1) \\ \hat{\gamma}(2) \\ \vdots \\ \hat{\gamma}(p) \end{pmatrix}, \tag{8.2.8}$$

where $\hat{\gamma}(h)$ is defined in (6.2.3). The estimator (8.2.8) is sometimes called the Yule–Walker estimator.

The least squares estimator of σ^2 is

$$\hat{\sigma}^2 = (n-2p-1)^{-1}\sum_{t=p+1}^{n} \hat{e}_t^2, \tag{8.2.9}$$

where $\hat{e}_t = Y_t - \mathbf{X}_t\hat{\boldsymbol{\theta}}$. The divisor for $\hat{\sigma}^2$ is defined by analogy to ordinary regression theory. There are $n-p$ observations in the regression, and $p+1$ parameters are estimated.

The asymptotic properties of $\hat{\boldsymbol{\theta}}$ and $\hat{\sigma}^2$ are given in the following theorem. The theorem is stated and proven for martingale difference errors, but the result also holds for e_t that are iid$(0, \sigma^2)$ random variables.

Theorem 8.2.1. Let Y_t satisfy

$$Y_t = \theta_0 + \sum_{j=1}^{p} \theta_j Y_{t-j} + e_t, \qquad t = p+1, p+2, \ldots. \tag{8.2.10}$$

Assume $(Y_1, Y_2, \ldots, Y_p)$ is fixed, or assume $(Y_1, Y_2, \ldots, Y_p)$ is independent of e_t for $t > p$ and $E\{|Y_i|^{2+\nu}\} < \infty$ for $i = 1, 2, \ldots, p$ and some $\nu > 0$. Let the roots of

$$m^p - \sum_{j=1}^{p} \theta_j m^{p-j} = 0 \tag{8.2.11}$$

be less than one in absolute value. Suppose $\{e_t\}_{t=1}^{\infty}$ is a sequence of $(0, \sigma^2)$ random variables with

$$E\{(e_t, e_t^2) \mid \mathscr{A}_{t-1}\} = (0, \sigma^2) \qquad \text{a.s.}$$

and

$$E\{|e_t|^{2+\nu} \mid \mathscr{A}_{t-1}\} < L < \infty \qquad \text{a.s.}$$

for all t and some $\nu > 0$, where $\mathscr{A}_{t-1}$ is the sigma-field generated by $\{e_j : j \leq t-1\}$. If $(Y_1, Y_2, \ldots, Y_p)$ is random, the sigma-field is generated by $\{e_j : j \leq t-1\}$ and $(Y_1, Y_2, \ldots, Y_p)$. Let

$$\hat{\boldsymbol{\theta}} = \left[\sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t\right]^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t' Y_t,$$

where $\mathbf{X}_t = (1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$. Then

(a) $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathscr{L}} \mathrm{N}(\mathbf{0}, \mathbf{A}^{-1}\sigma^2)$, and

(b) $\hat{\sigma}^2 \xrightarrow{P} \sigma^2$,

where

$$\mathbf{A} = \lim_{n\to\infty} n^{-1} \sum_{t=p+1}^{n} E(\mathbf{X}_t'\mathbf{X}_t),$$

and $\hat{\sigma}^2$ is defined in (8.2.9).

Proof. We have

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left[\sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t\right]^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t' e_t.$$

Given $\epsilon > 0$, there is some N_0 such that $\Sigma_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t$ is nonsingular with probability greater than $1-\epsilon$ for $n > N_0$. Let $\boldsymbol{\eta}$ be a column vector of arbitrary real numbers such that $\boldsymbol{\eta}'\boldsymbol{\eta} \neq 0$, and consider those samples for which $\Sigma_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t$ is nonsingular. Let

$$n^{-1/2}\boldsymbol{\eta}' \sum_{t=p+1}^{n} \mathbf{X}_t' e_t = \sum_{t=p+1}^{n} Z_{tn} = S_{nn},$$

where $Z_{tn} = n^{-1/2}\boldsymbol{\eta}'\mathbf{X}_t' e_t$. Then $E\{Z_{tn} \mid \mathscr{A}_{t-1}\} = 0$ a.s.,

$$\delta_{tn}^2 = E\{Z_{tn}^2 \mid \mathscr{A}_{t-1}\} = n^{-1}\boldsymbol{\eta}'\mathbf{X}_t'\mathbf{X}_t\boldsymbol{\eta}\sigma^2,$$

and

$$V_{nn}^2 = n^{-1} \sum_{t=p+1}^{n} \boldsymbol{\eta}'\mathbf{X}_t'\mathbf{X}_t\boldsymbol{\eta}\sigma^2.$$

By Theorem 6.3.5,

$$\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{t=p+1}^{n} \boldsymbol{\eta}'\mathbf{X}_t'\mathbf{X}_t\boldsymbol{\eta}\sigma^2 = \boldsymbol{\eta}'\mathbf{A}\boldsymbol{\eta}\sigma^2 .$$

Now

$$\begin{aligned} E\{S_{nn}^2\} &= V\left\{n^{-1/2}\boldsymbol{\eta}' \sum_{t=p+1}^{n} \mathbf{X}_t' e_t\right\} \\ &= n^{-1} \sum_{t=p+1}^{n} E\{\boldsymbol{\eta}'\mathbf{X}_t'\mathbf{X}_t\boldsymbol{\eta}\}\sigma^2 \end{aligned}$$

and

$$\lim_{n\to\infty} s_{nn}^2 = \lim_{n\to\infty} E\{S_{nn}^2\} = \boldsymbol{\eta}'\mathbf{A}\boldsymbol{\eta}\sigma^2 ,$$

and hence condition ii of Theorem 5.3.4 is satisfied.

We now investigate

$$\begin{aligned} & s_{nn}^{-2} \sum_{t=p+1}^{n} E\{Z_{tn}^2 I(|Z_{tn}| \geq \epsilon s_{nn})\} \\ & \quad \leq s_{nn}^{-2} \sum_{t=p+1}^{n} E\{(\epsilon s_{nn})^{-\nu} |Z_{tn}|^{2+\nu} I(|Z_{tn}| \geq \epsilon s_{nn})\} \\ & \quad \leq s_{nn}^{-2-\nu}\epsilon^{-\nu} \sum_{t=p+1}^{n} E\{|n^{-1/2}\boldsymbol{\eta}'\mathbf{X}_t' e_t|^{2+\nu}\} \\ & \quad < s_{nn}^{-2-\nu}\epsilon^{-\nu} n^{-1-\nu/2} L \sum_{t=p+1}^{n} E\{|\boldsymbol{\eta}'\mathbf{X}_t'|^{2+\nu}\} . \end{aligned}$$

Let w_j be the weights of Theorem 2.6.1. By Holder's inequality,

$$\begin{aligned} \left|\sum_{j=0}^{t-1} w_j e_{t-j}\right|^{2+\nu} &\leq \left[\sum_{j=0}^{t-1} |w_j|\,|e_{t-j}|\right]^{2+\nu} \\ &\leq \left[\sum_{j=0}^{t-1} |w_j|\right]^{1+\nu} \sum_{j=0}^{t-1} |w_j|\,|e_{t-j}|^{2+\nu} . \end{aligned}$$

Therefore, $E\{|\boldsymbol{\eta}'\mathbf{X}_t'|^{2+\nu}\}$ is bounded, condition iii of Theorem 5.3.4 is satisfied, and result a follows.

To prove part b, we observe that $\hat{e}_t = e_t - \mathbf{X}_t(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ and

$$\begin{aligned} \sum_{t=p+1}^{n} \hat{e}_t^2 &= \sum_{t=p+1}^{n} e_t^2 - (n-p)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\mathbf{A}_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ &= \sum_{t=p+1}^{n} e_t^2 + O_p(1) \end{aligned}$$

where $\mathbf{A}_n = (n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t$. The conclusion of part b follows because $(n-2p-1)^{-1} \sum_{t=p+1}^{n} e_t^2$ converges to σ^2 a.s. by Corollary 5.3.8. ▲

Because the matrix $\mathbf{A}_n = (n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t$ converges to $\mathbf{A}$ in probability, the usual regression distribution theory holds, approximately, for the autoregressive model. For example, $v_{ii}^{-1/2}(\hat{\theta}_i - \theta_i)$, where v_{ii} is the ith diagonal element of $(\sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t)^{-1}\hat{\sigma}^2$, is approximately a $N(0, 1)$ random variable.

While we have obtained pleasant asymptotic results, the behavior of the estimators in small samples can deviate considerably from that based on asymptotic theory. If the roots of the autoregressive equation are near zero, the approach to normality is quite rapid. For example, if $\rho = 0$ in the first order autoregressive process, the normal or t-distribution approximations will perform very well for $n > 30$. On the other hand, if the roots are near one in absolute value, very large samples may be required before the distribution is well approximated by the normal.

In Figure 8.2.1 we present the empirical density of the least squares estimator $\hat{\rho}$

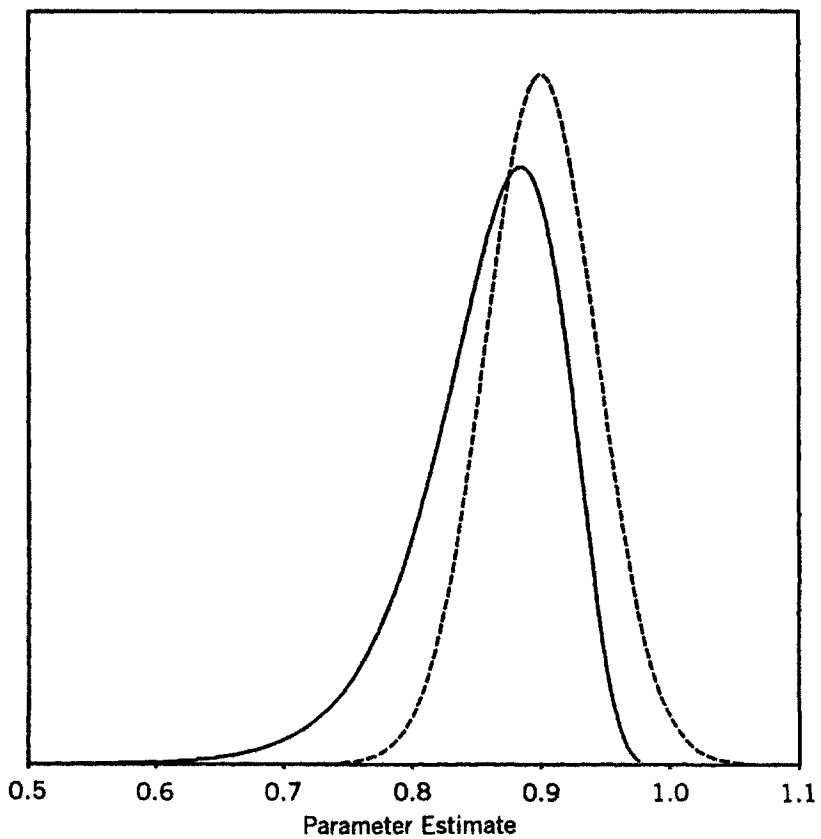

FIGURE 8.2.1. Estimated density of $\hat{\rho}$ compared with normal approximation for $\rho = 0.9$ and $n = 100$. (Dashed line is normal density.)

defined in (8.1.7). The density was estimated from 20,000 samples of size 100. The observations were generated by the autoregressive equation

$$Y_t = 0.9Y_{t-1} + e_t,$$

where the e_t are normal independent (0, 1) random variables. The empirical distribution displays a skewness similar to that which we would encounter in sampling from the binomial distribution. The mean of the empirical distribution is 0.861, and the variance is 0.0032. The distribution obtained from the normal approximation has a mean of 0.90 and a variance of 0.0019.

The mean of the empirical distribution agrees fairly well with the approximation obtained by the methods of Section 5.4. The bias approximated from a first order Taylor series is $E\{\hat{\rho} - \rho\} \doteq -n^{-1}(1 + 3\rho)$. This approximation to the expectation has been discussed by Mariott and Pope (1954) and Kendall (1954). Also see Pantula and Fuller (1985) and Shaman and Stine (1988).

Often the practitioner must determine the degree of autoregressive process as well as estimate the parameters. If it is possible to specify a maximum for the degree of the process, a process of that degree can first be estimated and high order terms discarded using the standard regression statistics. Anderson (1962) gives a procedure for this decision problem. Various model building methods based on regression theory can be used. Several such procedures are described in Draper and Smith (1981). In Section 8.4, we discuss other order determination procedures.

Often one inspects the residuals from the fit and perhaps computes the autocorrelations of these residuals. If the model is correct, the sample autocorrelations estimate zero with an error that is $O_p(n^{-1/2})$, but the variances of these estimators are generally smaller than the variances of estimators computed from a time series of independent random variables. See Box and Pierce (1970) and Ljung and Box (1978). Thus, while it is good practice to inspect the residuals, it is suggested that final tests of model adequacy be constructed by adding terms to the model and testing the hypothesis that the true value of the added coefficients is zero.

Example 8.2.1. To illustrate the regression estimation of the autoregressive process, we use the unemployment time series investigated in Section 6.3. The second order process, estimated by regressing $Y_t - \bar{y}_n$ on $Y_{t-1} - \bar{y}_n$ and $Y_{t-2} - \bar{y}_n$, is

$$\hat{Y}_t - 4.77 = \underset{(0.073)}{1.568}\,(Y_{t-1} - 4.77) - \underset{(0.073)}{0.699}\,(Y_{t-2} - 4.77),$$

where $\bar{y}_{100} = 4.77$ and the numbers below the coefficients are the estimated standard errors obtained from the regression analysis. The residual mean square is 0.105.

If we condition the analysis on the first two observations and regress Y_t on Y_{t-1} and Y_{t-2} including an intercept term in the regression, we obtain

$$\hat{Y}_t = \underset{(0.13)}{0.63} + \underset{(0.073)}{1.568}\, Y_{t-1} - \underset{(0.073)}{0.699}\, Y_{t-2}.$$

The coefficients are slightly different from those in Section 6.4, since the coefficients in Section 6.4 were obtained from equation (8.2.8).

To check on the adequacy of the second order representation, we fit a fifth order process. The results are summarized in Table 8.2.1. The F-test for the hypothesis that the time series is second order autoregressive against the alternative that it is fifth order is

$$F_{89}^{3} = 0.478[3(0.101)]^{-1} = 1.578.$$

The tabular 0.10 point for Snedecor's F with 3 and 89 degrees of freedom is 2.15, and so the null hypothesis is accepted at that level. ▲▲

8.2.2. Alternative Estimators for Autoregressive Time Series

The regression method of estimation is simple, easily understood, and asymptotically efficient for the parameters of stationary autoregressive processes. However, given the power of modern computing equipment, other procedures that are more efficient in small samples and (or) appropriate for certain models can be considered.

If the Y_t are normally distributed, the logarithm of the likelihood of a sample of n observations from a stationary pth order autoregressive process is the generalization of (8.1.5),

$$\log L(y:\boldsymbol{\theta}) = -0.5n \log 2\pi - 0.5n \log|\boldsymbol{\Sigma}_{YY}| - 0.5(\mathbf{Y} - \mathbf{J}\mu)'\boldsymbol{\Sigma}_{YY}^{-1}(\mathbf{Y} - \mathbf{J}\mu), \tag{8.2.12}$$

where $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$, $\mathbf{J}' = (1, 1, \ldots, 1)$, $\boldsymbol{\Sigma}_{YY}$ is the covariance matrix of $\mathbf{Y}$ expressed as a function of $(\sigma^2, \theta_1, \theta_2, \ldots, \theta_p)$, and

Table 8.2.1. Analysis of Variance for Quarterly Seasonally Adjusted Unemployment rate, 1948 to 1972

Source	Degrees of Freedom	Mean Square
Y_{t-1}	1	112.949
Y_{t-2} after Y_{t-1}	1	9.481
Y_{t-3} after Y_{t-1}, Y_{t-2}	1	0.305
Y_{t-4} after $Y_{t-1}, Y_{t-2}, Y_{t-3}$	1	0.159
Y_{t-5} after $Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}$	1	0.014
Error	89	0.101

$$\mu = \left[1 - \sum_{i=1}^{p} \theta_i\right]^{-1} \theta_0 . \tag{8.2.13}$$

Several computer packages contain algorithms that compute the maximum likelihood estimator. In Section 8.4, it is proven that the limiting distribution of the maximum likelihood estimator is the same as the limiting distribution of the ordinary least squares estimator.

By Corollary 2.6.1.3, a stationary autoregressive process can be given a forward representation

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j} = e_t$$

or a backward representation

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t+j} = v_t ,$$

where $\{e_t\}$ and $\{v_t\}$ are sequences of serially uncorrelated $(0, \sigma^2)$ random variables. The ordinary least squares estimator of $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ is the value of $\boldsymbol{\alpha}$ that minimizes the sum of squares of the estimated e_t. One could also construct an estimator that minimizes the sum of squares of the estimated v_t. This suggests a class of estimators, where the estimator of $\boldsymbol{\alpha}$ is the $\boldsymbol{\alpha}$ that minimizes

$$Q(\boldsymbol{\alpha}) = \sum_{t=p+1}^{n} w_t \left[Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j}\right]^2 + \sum_{t=1}^{n-p} (1 - w_{t+1}) \left[Y_t + \sum_{j=1}^{p} \alpha_j Y_{t+j}\right]^2 . \tag{8.2.14}$$

The ordinary least squares estimator is obtained by setting $w_t \equiv 1$. The estimator obtained by setting $w_t \equiv 0.5$ was studied by Dickey, Hasza, and Fuller (1984). We call the estimator with $w_t \equiv 0.5$ the simple symmetric estimator. For the zero mean first order process, the simple symmetric estimator is

$$\hat{\alpha}_{1s} = -\left[\sum_{t=2}^{n-1} Y_t^2 + 0.5(Y_1^2 + Y_n^2)\right]^{-1} \sum_{t=2}^{n} Y_{t-1} Y_t .$$

Because $|Y_{t-1} Y_t| \leq 0.5(Y_{t-1}^2 + Y_t^2)$, $\hat{\alpha}_{1s}$ for the first order process is always less than or equal to one in absolute value.

We call the estimator constructed with

$$w_t = \begin{cases} 0 , & t = 1, 2, \ldots, p , \\ (n - 2p + 2)^{-1}(t - p) , & t = p + 1, p + 2, \ldots, n - p + 1 , \\ 1 , & t = n - p + 2, n - p + 3, \ldots, n , \end{cases} \tag{8.2.15}$$

the weighted symmetric estimator. The weights (8.2.15) assume $n \geq 2p$. The

weighted symmetric estimator is nearly identical to the maximum likelihood estimator unless one of the estimated roots is close to one. The roots of the weighted symmetric estimator with weights (8.2.15) are not restricted to be less than one in absolute value. For the zero mean first order process, the weighted symmetric estimator with weights (8.2.15) is

$$\hat{\alpha}_{1w} = -\left[\sum_{t=2}^{n-1} Y_t^2 + n^{-1}\sum_{t=1}^{n} Y_t^2\right]^{-1} \sum_{t=2}^{n} Y_{t-1}Y_t .$$

Thus, the least squares estimator, $\hat{\alpha}_{1s}$, and $\hat{\alpha}_{1w}$ differ in the weights given to Y_1 and Y_n in the divisor.

Table 8.2.2 contains the variables required for the symmetric estimation of the pth order process. The estimator is

$$\hat{\boldsymbol{\alpha}} = -(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} ,$$

where $\mathbf{X}$ is the $(2n-2p)\times p$ matrix below the headings $-\alpha_1, -\alpha_2, \ldots, -\alpha_p$, $\mathbf{Y}$ is the $(2n-2p)$-dimensional column vector called the dependent variable, and $\mathbf{W}$ is the $(2n-2p)$ diagonal matrix whose elements are given in the "Weight" column. An estimator of the covariance matrix of $\hat{\boldsymbol{\alpha}}$ is

$$(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\hat{\sigma}^2 ,$$

where

$$\hat{\sigma}^2 = (n-p-1)^{-1}(\mathbf{Y}'\mathbf{W}\mathbf{Y} - \hat{\boldsymbol{\alpha}}'\mathbf{X}'\mathbf{W}\mathbf{Y}) .$$

Table 8.2.2. Data Arrangement for Regression Estimation of Autoregressive Parameters by the Weighted Symmetric Procedure

Weight	Dependent Variable	Parameter			
		$-\alpha_1$	$-\alpha_2$	$\cdots$	$-\alpha_p$
w_{p+1}	Y_{p+1}	Y_p	Y_{p-1}	$\cdots$	Y_1
w_{p+2}	Y_{p+2}	Y_{p+1}	Y_p	$\cdots$	Y_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
w_n	Y_n	Y_{n-1}	Y_{n-2}	$\cdots$	Y_{n-p}
$1-w_{n-p+1}$	Y_{n-p}	Y_{n-p+1}	Y_{n-p+2}	$\cdots$	Y_n
$1-w_{n-p}$	Y_{n-p-1}	Y_{n-p}	Y_{n-p+1}	$\cdots$	Y_{n-1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$1-w_2$	Y_1	Y_2	Y_3	$\cdots$	Y_{p+1}

If the mean is unknown, there are several ways to proceed. One option is to replace each Y_t in the table with $Y_t - \bar{y}$, where $\bar{y} = n^{-1}\sum_{t=1}^{n} Y_t$. The second is to replace the elements in each column of the table with $Y_t - \bar{y}_{(i)}$, where $\bar{y}_{(i)}$ is the mean of the elements in the ith column. The third is to add a column of ones to the table and use the ordinary least squares formulas. The procedures are asymptotically equivalent, but the work of Park (1990) indicates that the use of separate means, or the column of ones, is preferred in small samples when the process has a root close to one in absolute value. If the estimator of the mean is of interest, the mean can be estimated by $\bar{y}$, or one can use estimated generalized least squares where the covariance matrix is based on the estimated autoregressive parameters.

If a regression program has a missing value option that omits an observation if any variable is missing, one can obtain the estimators by creating a data set composed of the original observations followed by a missing value, followed by the original data in reverse order. The created data set contains $2n + 1$ "observations." Lagging the created vector p times gives the p explanatory variables required for the regression. Then calling the regression option that deletes observations with a missing value enables one to compute the simple symmetric estimator for any autoregression up to order p. The addition of weights is required to compute the weighted symmetric estimator.

The ideas of partial autocorrelation introduced in Section 1.4 can be used in a sequential estimation scheme for autoregressive models. Let a sample of n observations $(Y_1, Y_2, \ldots, Y_n)$ be available, and define $X_t = Y_t - \bar{y}_n$, where $\bar{y}_n = n^{-1}\sum_{t=1}^{n} Y_t$. Let

$$\hat{\theta}_{11} = \left[\sum_{t=2}^{n} X_t^2 \sum_{t=2}^{n} X_{t-1}^2\right]^{-1/2} \sum_{t=2}^{n} X_t X_{t-1} \tag{8.2.16}$$

be an estimator of the first autocorrelation. Then an estimator of the first order autoregressive equation is

$$\hat{Y}_t = \bar{y}_n(1 - \hat{\theta}_{11}) + \hat{\theta}_{11} Y_{t-1}\,, \tag{8.2.17}$$

and an estimator of the residual mean square for the autoregression is

$$\hat{\sigma}^2_{(11)} = n(n-2)^{-1}(1 - \hat{\theta}_{11}^2)\hat{\gamma}(0)\,, \tag{8.2.18}$$

where

$$\hat{\gamma}(0) = n^{-1}\sum_{t=1}^{n} (Y_t - \bar{y}_n)^2\,.$$

A test that the first order autocorrelation is zero under the assumption that higher order partial autocorrelations are zero is

$$t_1 = [(n-2)^{-1}(1 - \hat{\theta}_{11}^2)]^{-1/2}\hat{\theta}_{11}\,. \tag{8.2.19}$$

Under the null hypothesis, this statistic is approximately distributed as a $N(0, 1)$ random variable in large samples.

Higher order partial autocorrelations, higher order autoregressions, higher order autocorrelations, and tests can be computed with the following formulas:

$$\begin{aligned}
\hat{\theta}_{i+1,i+1} &= \left[\sum_{t=i+2}^{n} W_{ti}^2 \sum_{t=i+2}^{n} Z_{t-i-1,i}^2\right]^{-1/2} \sum_{t=i+2}^{n} W_{ti} Z_{t-i-1,i}\,, \\
\hat{\theta}_{j,i+1} &= \hat{\theta}_{ji} - \hat{\theta}_{i+1,i+1}\hat{\theta}_{i+1-j,i}\,, \qquad j = 1, 2, \ldots, i\,, \\
\hat{\sigma}_{(ii)}^2 &= n(n-1-i)^{-1}\hat{\gamma}(0)[1 - \overset{*}{R}{}_{(ii)}^2]\,, \\
\overset{*}{R}{}_{(ii)}^2 &= 1 - \prod_{j=1}^{i}(1 - \hat{\theta}_{jj}^2)\,, \\
\overset{*}{\rho}(i) &= \sum_{j=1}^{i} \hat{\theta}_{ji}\overset{*}{\rho}(i-j)\,, \\
\hat{\theta}_{0,i+1} &= \bar{y}_n\left[1 - \sum_{j=1}^{i+1} \hat{\theta}_{j,i+1}\right], \\
t_i &= [(n-1-i)^{-1}(1 - \hat{\theta}_{ii}^2)]^{-1/2}\hat{\theta}_{ii}\,,
\end{aligned} \tag{8.2.20}$$

where

$$(W_{ti}, Z_{ti}) = \left[X_t - \sum_{j=1}^{i} \hat{\theta}_{ji}X_{t-j},\ X_t - \sum_{j=1}^{i} \hat{\theta}_{ji}X_{t+j}\right],$$

and $\overset{*}{\rho}(0) = 1$. The estimated partial autocorrelations $\hat{\theta}_{ii}$ are defined for $i = 1, 2, \ldots, n-1$, and t_i and $\hat{\sigma}_{(ii)}^2$ are defined for $i = 1, 2, \ldots, n-2$. A test that $\theta_{ii} = 0$, under the assumption that $\theta_{jj} = 0$, $j > i$, is t_i, which is the generalization of t_1 of (8.2.19).

If a pth order autoregression is selected as the representation for the time series, the covariance matrix of the vector of coefficients can be estimated with

$$\hat{\mathbf{V}}\{\hat{\boldsymbol{\theta}}_p\} = (n-1-p)^{-1}\hat{\mathbf{P}}^{-1}\hat{\sigma}_{(pp)}^2\,, \tag{8.2.21}$$

where

$$\hat{\mathbf{P}} = \begin{bmatrix}
1 & \overset{*}{\rho}(1) & \overset{*}{\rho}(2) & \cdots & \overset{*}{\rho}(p-1) \\
\overset{*}{\rho}(1) & 1 & \overset{*}{\rho}(1) & \cdots & \overset{*}{\rho}(p-2) \\
\vdots & \vdots & \vdots & & \vdots \\
\overset{*}{\rho}(p-1) & \overset{*}{\rho}(p-2) & \overset{*}{\rho}(p-3) & \cdots & 1
\end{bmatrix},$$

$\hat{\boldsymbol{\theta}}_p = (\hat{\theta}_{1p}, \hat{\theta}_{2p}, \ldots, \hat{\theta}_{pp})'$, and $\overset{*}{\rho}(h)$ is defined in (8.2.20). The matrix $\hat{\mathbf{P}}^{-1}$ is also given by

$$\mathbf{B}'\mathbf{S}^{-1}\mathbf{B}\,, \tag{8.2.22}$$

where

$$\mathbf{S} = \operatorname{diag}\{1, 1 - \overset{*}{R}^2_{(11)}, 1 - \overset{*}{R}^2_{(22)}, \ldots, 1 - \overset{*}{R}^2_{(p-1,p-1)}\},$$

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\hat{\theta}_{11} & 1 & 0 & \cdots & 0 \\ -\hat{\theta}_{22} & -\hat{\theta}_{12} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -\hat{\theta}_{p-1,p-1} & -\hat{\theta}_{p-2,p-1} & -\hat{\theta}_{p-3,p-1} & \cdots & 1 \end{bmatrix}.$$

An alternative estimator of the partial autocorrelation is

$$\tilde{\theta}_{ii} = \left(\sum_{t=i+2}^{n} W_{ti}^2 + \sum_{t=i+2}^{n} Z_{t-i,i-1}^2 \right)^{-1} 2 \sum_{t=i+2}^{n} W_{ti} Z_{t-i,i-1}. \tag{8.2.23}$$

It can be verified that $\tilde{\theta}_{ii}$ is also always less than one in absolute value. The estimator (8.2.23) in combination with (8.2.20) was suggested by Burg (1975).

The sequential method of computing the autoregression has the advantage for stationary time series that all roots of the estimated autoregression are less than one in absolute value. The estimators obtained by the sequential methods are very similar to the simple symmetric estimators. The sequential procedure also uses all available observations at each step. If regressions are only computed by regression in one direction, moving from an equation of order $p-1$ to an equation of order p involves dropping an observation and adding an explanatory variable. Hence, the maximum possible order for the one-direction regression procedure is $\frac{1}{2}n$. The sequential procedure defines the autoregression up to order $n-1$.

The alternative estimators are asymptotically equivalent for stationary autoregressive processes.

Theorem 8.2.2. Let the assumptions of Theorem 8.2.1 hold. Then the limiting distribution of $n^{1/2}(\overset{*}{\boldsymbol{\theta}} - \boldsymbol{\theta})$, where $\overset{*}{\boldsymbol{\theta}}$ is the maximum likelihood estimator, the simple symmetric estimator, the partial correlation estimator, or the weighted symmetric estimator, is the same as that given for the ordinary least squares estimator in Theorem 8.2.1.

Proof. Omitted. ▲

The maximum likelihood estimator is available in many computer packages and performs well in simulation studies for the correct model. The simple symmetric estimator, the weighted symmetric estimator, the Burg estimator, and the maximum likelihood estimator have similar efficiencies for stationary processes. The maximum likelihood estimator and the weighted symmetric estimator perform better than other estimators for processes with roots close to one in absolute value. The maximum normal likelihood estimator and the partial correlation methods

produce estimated equations such that the roots of the estimated characteristic equation are all less than one in absolute value. It is possible for the roots associated with ordinary least squares or with the weighted symmetric estimator to be greater than one in absolute value. The ordinary least squares estimator performs well for forward prediction and is recommended as a preliminary estimator if it is possible that the process is nonstationary with a root greater than one.

8.2.3. Multivariate Autoregressive Time Series

In this subsection, we extend the autoregressive estimation procedures to vector valued processes. Let $\mathbf{Y}_t$ be a k-dimensional stationary process that satisfies the equation

$$\mathbf{Y}_t - \boldsymbol{\mu} + \sum_{i=1}^{p} \mathbf{A}_i(\mathbf{Y}_{t-i} - \boldsymbol{\mu}) = \mathbf{e}_t \tag{8.2.24}$$

for $t = p+1, p+2, \ldots$, where $\mathbf{e}_t$ are independent $(\mathbf{0}, \boldsymbol{\Sigma})$ random variables or martingale differences. If the process is stationary, $E\{\mathbf{Y}_t\} = \boldsymbol{\mu}$. The equation can also be written

$$\mathbf{Y}_t + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{Y}_{t-i} = \boldsymbol{\theta}_0 + \mathbf{e}_t, \tag{8.2.25}$$

where $\boldsymbol{\theta}_0$ is a k-dimensional column vector. Let $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$, be observed. Discussion of estimation for vector autoregressive processes can proceed by analogy to the univariate case. Each of the equations in (8.2.25) can be considered as a regression equation. We write the ith equation of (8.2.24) as

$$Y_{it} = \theta_{0i} - \sum_{j=1}^{p} \mathbf{A}_{ji.} \mathbf{Y}_{t-j} + e_{it}, \tag{8.2.26}$$

where Y_{it} is the ith element of $\mathbf{Y}_t$, θ_{0i} is the ith element of $\boldsymbol{\theta}_0$, $\mathbf{A}_{ji.}$ is the ith row of $\mathbf{A}_j$, and e_{it} is the ith element of $\mathbf{e}_t$.

Defining $\boldsymbol{\theta}_i = (\theta_{0i}, -\mathbf{A}_{1i.}, -\mathbf{A}_{2i.}, \ldots, -\mathbf{A}_{pi.})'$, $i = 1, 2, \ldots, k$, and $\mathbf{X}_t = (1, \mathbf{Y}'_{t-1}, \mathbf{Y}'_{t-2}, \ldots, \mathbf{Y}'_{t-p})$, we can write equation (8.2.26) as

$$Y_{it} = \mathbf{X}_t \boldsymbol{\theta}_i + e_{it}. \tag{8.2.27}$$

On the basis of our scalar autoregressive results, we are led to consider the estimators

$$\hat{\boldsymbol{\theta}}_i = \left[\sum_{t=p+1}^{n} \mathbf{X}'_t \mathbf{X}_t \right]^{-1} \sum_{t=p+1}^{n} \mathbf{X}'_t \mathbf{Y}_{it}, \qquad i = 1, 2, \ldots, k. \tag{8.2.28}$$

If we let $\mathbf{A}' = (-\boldsymbol{\theta}_0, \mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_p)'$, then the system (8.2.25) can be written

$$\mathbf{Y}_t' = -\mathbf{X}_t\mathbf{A}' + \mathbf{e}_t', \qquad t = p+1, p+2, \ldots, n, \tag{8.2.29}$$

and the ordinary least squares estimator of $\mathbf{A}'$ is

$$\hat{\mathbf{A}}' = -\left[\sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t\right]^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{Y}_t', \tag{8.2.30}$$

where $\hat{\boldsymbol{\theta}}_i$ is the ith column of $-\hat{\mathbf{A}}'$ and $\boldsymbol{\theta}_i$ is the ith column of $-\mathbf{A}$. The least squares estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = [n - (k+1)p - 1]^{-1} \sum_{t=p+1}^{n} \hat{\mathbf{e}}_t\hat{\mathbf{e}}_t', \tag{8.2.31}$$

where

$$\hat{\mathbf{e}}_t = \mathbf{Y}_t - \hat{\boldsymbol{\theta}}_0 + \sum_{i=1}^{p} \hat{\mathbf{A}}_i\mathbf{Y}_{t-p}$$

and $\hat{\mathbf{A}}$ is defined in (8.2.30).

An alternative method of computing the estimator of $(\mathbf{A}_1, \ldots, \mathbf{A}_p) = \mathbf{A}_{[2]}$ is as

$$\hat{\mathbf{A}}_{[2]}' = -\left[\sum_{t=p+1}^{n} \mathbf{U}_t'\mathbf{U}_t\right]^{-1} \sum_{t=p+1}^{n} \mathbf{U}_t'(\mathbf{Y}_t - \bar{\mathbf{y}})', \tag{8.2.32}$$

where $\mathbf{U}_t = (\mathbf{Y}_{t-1}' - \bar{\mathbf{y}}', \mathbf{Y}_{t-2}' - \bar{\mathbf{y}}', \ldots, \mathbf{Y}_{t-p}' - \bar{\mathbf{y}}')$ and $\bar{\mathbf{y}} = n^{-1}\Sigma_{t=1}^{n}\mathbf{Y}_t$. The distributions of the estimators are the vector analogs of the distributions of Theorem 8.2.1.

Theorem 8.2.3. Let the vector time series $\mathbf{Y}_t$ satisfy

$$\sum_{j=0}^{p} \mathbf{A}_j\mathbf{Y}_{t-j} = \boldsymbol{\theta}_0 + \mathbf{e}_t$$

for $t = p+1, p+2, \ldots$, where $\{\mathbf{e}_t\}$ is a sequence of k-dimensional random variables and the $\mathbf{A}_j$ are fixed $k \times k$ matrices such that $\mathbf{A}_0 = \mathbf{I}$ and the roots of

$$\left|\sum_{j=0}^{p} \mathbf{A}_j m^{p-j}\right| = 0$$

are less than one in absolute value. Suppose $\{\mathbf{e}_t\}_{t=1}^{\infty}$ is a sequence of $(\mathbf{0}, \boldsymbol{\Sigma})$ random vectors with

$$E\{(\mathbf{e}_t, \mathbf{e}_t\mathbf{e}_t') \mid \mathscr{A}_{t-1}\} = (\mathbf{0}, \boldsymbol{\Sigma}) \qquad \text{a.s.}$$

and

$$E\{|\mathbf{e}_t|^{2+\nu} \mid \mathscr{A}_{t-1}\} < L < \infty \qquad \text{a.s.}$$

for all t and some $\nu > 0$, where $\mathcal{A}_{t-1}$ is the sigma-field generated by $\{\mathbf{e}_j : j \leqslant t-1\}$. Assume

$$\mathbf{G} = \underset{n\to\infty}{p\lim}(n-p)^{-1} \sum_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t$$

is positive definite. If $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p)$ are random, it is assumed that $E\{|\mathbf{Y}_i|^{2+\nu}\} < \infty$ for $i = 1, 2, \ldots, p$ and the sigma-field is generated by $\{\mathbf{e}_j : j \leqslant t-1\}$ and $(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_p)$. Then

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{G}^{-1}),$$

and $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} \boldsymbol{\Sigma}$, where $\hat{\boldsymbol{\theta}}_i$ is defined in (8.2.28), $\hat{\boldsymbol{\Sigma}}$ is defined in (8.2.31),

$$\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2', \ldots, \boldsymbol{\theta}_k'),$$

and $\boldsymbol{\Sigma} \otimes \mathbf{G}^{-1}$ is the Kronecker product of the matrices $\boldsymbol{\Sigma}$ and $\mathbf{G}^{-1}$.

Proof. We only outline the proof, because it differs in no substantive way from that of Theorem 8.2.1. Since $\mathbf{Y}_t$ is converging to a stationary time series, $p\lim(n-p)^{-1} \Sigma_{t=p+1}^{n} \mathbf{X}_t'\mathbf{X}_t = \mathbf{G}$ by Corollary 6.3.5. Therefore, the limiting distribution of $\hat{\boldsymbol{\theta}}$ follows from that of $n^{1/2}(n-p)^{-1} \Sigma_{t=p+1}^{n} \mathbf{X}_t' e_{it}$, $i = 1, 2, \ldots, k$. By arguments parallel to those of Theorem 8.2.1, we obtain the limiting distribution. If $\boldsymbol{\Sigma}$ is singular, then the limiting distribution contains singular components. ▲

An alternative estimator for the vector process is the maximum likelihood estimator which can be computed by numerical procedures. The limiting distribution of the maximum likelihood estimator is the same as that of the least squares estimator for stationary vector processes.

8.3. MOVING AVERAGE TIME SERIES

We have seen that ordinary least squares regression procedures can be used to obtain efficient estimators for the parameters of autoregressive time series. Unfortunately, the estimation for moving average processes is less simple.

By the results of Chapter 2 we know that there is a relationship between the correlation function of a moving average time series and the parameters of the time series. For example, for the first order process, $\rho(1) = \beta(1+\beta^2)^{-1}$, where β is the parameter of the process. On this basis, we might be led to estimate β from an estimator of $\rho(1)$. Since we demonstrated in Chapter 6 that

$$\hat{r}(1) = \left[\sum_{t=1}^{n} (Y_t - \bar{y}_n)^2\right]^{-1} \sum_{t=2}^{n} (Y_t - \bar{y}_n)(Y_{t-1} - \bar{y}_n)$$

estimates $\rho(1)$ with an error which is $O_p(n^{-1/2})$, it follows that

$$\hat{\beta}_r = \begin{cases} [2\hat{r}(1)]^{-1}\{1 - [1 - 4\hat{r}^2(1)]^{1/2}\}, & 0 < |\hat{r}(1)| \leq 0.5, \\ -1, & \hat{r}(1) < -0.5, \\ 1, & \hat{r}(1) > 0.5, \\ 0, & \hat{r}(1) = 0, \end{cases} \tag{8.3.1}$$

estimates β with an error of the same order. Obviously, if $\hat{r}(1)$ lies outside the range $(-0.5, 0.5)$ by a significant amount, the model of a first order moving average is suspect.

It follows from equation (6.2.8) that

$$\text{Var}\{\hat{r}(1)\} = n^{-1}(1+\beta^2)^{-4}(1+\beta^2+4\beta^4+\beta^6+\beta^8) + O(n^{-2}). \tag{8.3.2}$$

Because the derivative of $\rho(1)$ with respect to β is $(1+\beta^2)^{-2}(1-\beta^2)$, the approximate variance of $\hat{\beta}_r$ is, for $|\beta| < 1$,

$$\text{Var}\{\hat{\beta}_r\} \doteq n^{-1}(1-\beta^2)^{-2}(1+\beta^2+4\beta^4+\beta^6+\beta^8). \tag{8.3.3}$$

While the estimator (8.3.1) is consistent, we shall see that it is inefficient for $\beta \neq 0$. Roughly, the other sample autocorrelations contain information about β.

More efficient estimators can be obtained by least squares or by likelihood methods. We first present an estimation procedure based on the Gauss–Newton method of estimating the parameters of a nonlinear function discussed in Chapter 5. One purpose for describing the procedure is to demonstrate the nature of the derivatives that define the limiting distribution of the estimator. In practice, there are a number of computer programs available to perform the numerical estimation. Maximum likelihood estimation is discussed in Section 8.4.

Consider the first order moving average

$$Y_t = e_t + \beta e_{t-1}, \tag{8.3.4}$$

where $|\beta| < 1$ and the e_t are independent $(0, \sigma^2)$ random variables. Equation (8.3.4) may also be written

$$e_t = -\beta e_{t-1} + Y_t, \tag{8.3.5}$$

and, using our previous difference equation results, we have

$$\begin{aligned} Y_t &= -\sum_{j=1}^{t-1} (-\beta)^j Y_{t-j} - (-\beta)^t e_0 + e_t \\ &= f_t(Y; \beta, e_0) + e_t, \end{aligned} \tag{8.3.6}$$

where

$$f_1(Y; \beta, e_0) = \beta e_0$$

and

$$f_t(Y;\beta,e_0) = -\sum_{j=1}^{t-1}(-\beta)^j Y_{t-j} - (-\beta)^t e_0 \tag{8.3.7}$$

for $t > 1$. The expression (8.3.6) places the estimation problem in the nonlinear estimation format of Section 5.5.2.

We assume initial estimators $\tilde{\beta}$ and $\tilde{e}_0$ satisfying $(\tilde{\beta}-\beta) = o_p(n^{-1/4})$ and $\tilde{e}_0 = O_p(1)$ are available. These requirements are satisfied if one uses $\tilde{e}_0 = 0$ and the estimator $\hat{\beta}_r$ of (8.3.1). The one-step Gauss–Newton estimator of β is obtained by regressing the deviations

$$e_t(Y;\tilde{\beta}) = Y_t - f_t(Y;\tilde{\beta},\tilde{e}_0) = \sum_{j=0}^{t-1}(-\tilde{\beta})^j Y_{t-j} + (-\tilde{\beta})^t \tilde{e}_0 \tag{8.3.8}$$

on the first derivative of $f_t(Y;\beta,e_0)$ evaluated at $\beta = \tilde{\beta}$; that derivative is

$$W_t(Y;\tilde{\beta}) = \begin{cases} \tilde{e}_0, & t = 1, \\ \sum_{j=1}^{t-1} j(-\tilde{\beta})^{j-1} Y_{t-j} + t(-\tilde{\beta})^{t-1}\tilde{e}_0, & t = 2,3,\ldots,n. \end{cases} \tag{8.3.9}$$

We could also include e_0 as a "random parameter" to be estimated. The inclusion of the derivative for a change in $\tilde{e}_0$ does not affect the limiting distribution of the estimator of β for invertible moving averages. Therefore, we simplify our discussion by considering only the derivative with respect to β.

The computation of $e_t(Y;\tilde{\beta})$ and $W_t(Y;\tilde{\beta})$ is simplified by noting that both satisfy difference equations:

$$e_t(Y;\tilde{\beta}) = \begin{cases} Y_1 - \tilde{\beta}\tilde{e}_0, & t = 1, \\ Y_t - \tilde{\beta}e_{t-1}(Y;\tilde{\beta}), & t = 2,3,\ldots,n, \end{cases} \tag{8.3.10}$$

and

$$W_t(Y;\tilde{\beta}) = \begin{cases} \tilde{e}_0, & t = 1, \\ e_{t-1}(Y;\tilde{\beta}) - \tilde{\beta}W_{t-1}(Y;\tilde{\beta}), & t = 2,3,\ldots,n. \end{cases} \tag{8.3.11}$$

The difference equation for $e_t(Y;\tilde{\beta})$ follows directly from (8.3.5). Equation (8.3.11) can be obtained by differentiating both sides of

$$e_t(Y;\beta) = Y_t - \beta e_{t-1}(Y;\beta)$$

with respect to β and evaluating the resulting expression at $\beta = \tilde{\beta}$.

Regressing $e_t(Y;\tilde{\beta})$ on $W_t(Y;\tilde{\beta})$, we obtain an estimator of $\beta - \tilde{\beta}$. The improved estimator of β is then

$$\hat{\beta} = \tilde{\beta} + \Delta\hat{\beta},$$

where

$$\Delta\hat{\beta} = \left[\sum_{t=1}^{n} [W_t(Y; \tilde{\beta})]^2\right]^{-1} \sum_{t=1}^{n} e_t(Y; \tilde{\beta}) W_t(Y; \tilde{\beta}) . \tag{8.3.12}$$

The asymptotic properties of $\hat{\beta}$ are developed in Theorem 8.3.1. An interesting result is that the large sample behavior of the estimator of the moving average parameter β is the same as that of the estimator of the autoregressive process with parameter $-\beta$. The limiting distribution follows from the fact that the derivative (8.3.11) evaluated at the true β is an autoregressive process.

Theorem 8.3.1. Let Y_t satisfy (8.3.4), where $|\beta^0| < 1$, β^0 is the true value, and the e_t are independent $(0, \sigma^2)$ random variables with $E\{|e_t|^{2+\nu}\} < L < \infty$ for some $\nu > 0$. Let $\tilde{e}_0$ and $\tilde{\beta}$ be initial estimators satisfying $\tilde{e}_0 = O_p(1)$, $\tilde{\beta} - \beta = o_p(n^{-1/4})$, and $|\tilde{\beta}| < 1$. Then

$$n^{1/2}(\hat{\beta} - \beta^0) \xrightarrow{\mathcal{L}} N[0, 1 - (\beta^0)^2] ,$$

where $\hat{\beta}$ is defined in (8.3.12). Also, $\hat{\sigma}^2 \xrightarrow{P} (\sigma^0)^2$, where σ^0 is the true value of σ and

$$\hat{\sigma}^2 = n^{-1} \sum_{t=1}^{n} e_t^2(Y; \hat{\beta}) .$$

Proof. The model (8.3.6) is of the form (5.5.52) discussed in Section 5.5.2. The first derivative of $f_t(Y; \beta)$ is given in (8.3.9). The next two derivatives are

$$\frac{\partial^2 f_t(Y; \beta)}{\partial \beta^2} = H_t(Y; \beta) = -\sum_{j=2}^{t-1} j(j-1)(-\beta)^{j-2} Y_{t-j} - t(t-1)(-\beta)^{t-2} e_0 ,$$

$$\frac{\partial^3 f_t(Y; \beta)}{\partial \beta^3} = G_t(Y; \beta)$$

$$= \sum_{j=3}^{t-1} j(j-1)(j-2)(-\beta)^{j-3} Y_{t-j} + t(t-1)(t-2)(-\beta)^{t-3} e_0 ,$$

where it is understood that the summation is defined as if $Y_t = 0$ for $t \leq 0$. Let $\bar{S}$ be a closed interval containing β as an interior point and such that $\max_{\beta \in \bar{S}} |\beta| < \lambda < 1$. By Corollary 2.2.2.3, $f_t(Y; \beta)$ and the derivatives are moving averages with exponentially declining weights for all β in $\bar{S}$. Hence, for β in $\bar{S}$ they converge to stationary infinite moving average time series, the effect of e_0 being transient. It follows from Theorem 6.3.5 that the sample covariances and autocovariances of the four time series converge in probability for all β in $\bar{S}$ and the limits are continuous functions of β. Convergence is uniform on the compact set $\bar{S}$. Hence, conditions 1 and 2 of Theorem 5.5.4 are satisfied and

$$\hat{\beta} - \beta = \left[\sum_{t=1}^{n} W_t^2(Y; \beta^0)\right]^{-1} \sum_{t=1}^{n} W_t(Y; \beta^0) e_t + o_p(n^{-1/2}) .$$

If $\beta = \beta^0$,

$$W_t(Y; \beta^0) = \sum_{j=1}^{t-1} j(-\beta^0)^{j-1}(e_{t-j} + \beta^0 e_{t-j-1}) + t(-\beta^0)^{t-1}\tilde{e}_0$$
$$= \sum_{j=1}^{t-1} (-\beta^0)^{j-1} e_{t-j} + (-\beta^0)^{t-1}[t\tilde{e}_0 - (t-1)e_0]$$

and $W_t(Y; \beta^0)$ is converging to a stationary first order autoregressive process with parameter $-\beta^0$. Hence,

$$n^{-1} \sum_{t=1}^{n} W_t^2(Y; \beta^0) \xrightarrow{P} [1 - (\beta^0)^2]^{-1}\sigma^2$$

by Theorem 6.3.5, and

$$n^{-1/2} \sum_{t=1}^{n} W_t(Y; \beta^0)e_t \xrightarrow{\mathcal{L}} N(0, [1 - (\beta^0)^2]^{-1}\sigma^4)$$

by the arguments used in the proof of Theorem 8.2.1. Thus, the limiting distribution for $n^{1/2}(\hat{\beta} - \beta^0)$ is established.

Because $\hat{\sigma}^2$ is a continuous function of $\hat{\beta}$ and because $n^{-1}\sum_{t=1}^{n} e_t^2(Y; \beta)$ converges uniformly on $\bar{S}$, it follows that $\hat{\sigma}^2$ converges to $(\sigma^0)^2$ in probability. ▲

Comparison of the result of Theorem 8.3.1 and equation (8.3.3) establishes the large sample inefficiency of the estimator constructed from the first order autocorrelation. By the results of Theorem 8.3.1, we can use the regular regression statistics as approximations when drawing inferences about β.

In our discussion we have assumed that the mean of the time series was known and taken to be zero. It can be demonstrated that the asymptotic results hold for Y_t replaced by $Y_t - \bar{y}_n$, where $\bar{y}_n$ is the sample mean.

The procedure can be iterated using $\hat{\beta}$ as the initial estimator. In the next section we discuss estimators that minimize a sum of squares criterion or maximize a likelihood criterion. The asymptotic distribution of those estimators is the same as that obtained in Theorem 8.3.1. However, the estimators of the next section generally perform better in small samples.

A method of obtaining initial estimators that is applicable to higher order processes is an estimation procedure suggested by Durbin (1959). By Theorem 2.6.2, any qth order moving average

$$Y_t = \sum_{s=1}^{q} \beta_s e_{t-s} + e_t \tag{8.3.13}$$

for which the roots of the characteristic equation are less than one can be represented in the form

$$Y_t = -\sum_{j=1}^{\infty} c_j Y_{t-j} + e_t ,$$

where the weights satisfy the difference equation $c_1 = -\beta_1, c_2 = -\beta_2 - \beta_1 c_1$,

$$\begin{aligned} c_3 &= -\beta_3 - \beta_1 c_2 - \beta_2 c_1 , \\ &\vdots \\ c_q &= -\beta_q - \beta_1 c_{q-1} - \beta_2 c_{q-2} - \cdots - \beta_{q-1} c_1 , \\ c_j &= -\sum_{m=1}^{q} \beta_m c_{j-m} , \qquad j = q+1, q+2, \ldots . \end{aligned} \tag{8.3.14}$$

Since the weights c_j are sums of powers of the roots, they decline in absolute value, and one can terminate the sum at a convenient finite number, say k. Then we can write

$$Y_t \doteq -\sum_{j=1}^{k} c_j Y_{t-j} + e_t . \tag{8.3.15}$$

On the basis of this approximation, we treat Y_t as a finite autoregressive process and estimate c_j, $j = 1, 2, \ldots, k$, by the regression procedures of Section 8.2. As the true weights satisfy equation (8.3.14), we treat the estimated c_j's as a finite autoregressive process satisfying (8.3.14) and estimate the β's. That is, we treat

$$\hat{c}_j = -\sum_{s=1}^{q} \hat{c}_{j-s} \beta_s \tag{8.3.16}$$

as a regression equation and estimate the β's by regressing $-\hat{c}_j$ on $\hat{c}_{j-1}$, $\hat{c}_{j-2}, \ldots, \hat{c}_{j-q}$, where the appropriate modifications must be made for $j = 1, 2, \ldots, q$, as per (8.3.14).

If we let $\{k_n\}$ be a sequence such that $k_n = o(n^{1/3})$ and $k_n \to \infty$ as $n \to \infty$, it is possible to use the results of Berk (1974) to demonstrate that

$$\sum_{j=1}^{k_n} (\hat{c}_j - c_j)^2 = O_p(k_n n^{-1}) .$$

It follows that the preliminary estimators constructed from the $\hat{c}_j$ will have an error that is $o_p(n^{-1/3})$.

In carrying out the Gauss–Newton procedure, initial values $\tilde{e}_{1-q}, \tilde{e}_{2-q}, \ldots, \tilde{e}_0$ are required. The simplest procedure is to set them equal to zero. Alternatively, the autoregressive equation (8.3.15) can be used to estimate the Y-values preceding the sample period. Recall that a stationary autoregressive process can be written in either a forward or a backward manner (Corollary 2.6.1.2) and, as a result, the extrapolation formula for Y_0 is of the same form as that for Y_{n+1}. If the true process is a qth order moving average, one uses the autoregressive equation to

estimate q values, since the best predictors for Y_t, $t \leq -q$, are zero. Thus, one predicts $Y_0, Y_{-1}, \ldots, Y_{1-q}$, sets $Y_t = 0$ for $t < 1 - q$, and uses equation (8.3.13) and the predicted Y-values to estimate $e_{1-q}, e_{2-1}, \ldots, e_0$.

Example 8.3.1. We illustrate the Gauss–Newton procedure by fitting a first order moving average to an artificially generated time series. Table 8.3.1 contains 100 observations on X_t defined by $X_t = 0.7e_{t-1} + e_t$, where the e_t are computer generated normal independent (0, 1) random variables. We assume that we know the mean of the time series is zero.

As the first step in the analysis, we fit a seventh order autoregressive model to the data. This yields

$$\begin{aligned}\hat{Y}_t = {} & \underset{(0.108)}{0.685}\, Y_{t-1} - \underset{(0.131)}{0.584}\, Y_{t-2} + \underset{(0.149)}{0.400}\, Y_{t-3} - \underset{(0.152)}{0.198}\, Y_{t-4} \\ & + \underset{(0.149)}{0.020}\, Y_{t-5} + \underset{(0.133)}{0.014}\, Y_{t-6} + \underset{(0.108)}{0.017}\, Y_{t-7},\end{aligned}$$

Table 8.3.1. One Hundred Observations from a First Order Moving Average Time Series with $\beta = 0.7$

	First 25	Second 25	Third 25	Fourth 25
1	1.432	1.176	−1.311	2.607
2	−0.343	0.846	−0.105	1.572
3	−1.759	0.079	0.313	−0.261
4	−2.537	0.815	−0.890	−0.686
5	−0.295	2.566	−1.778	−2.079
6	0.689	1.675	−0.202	−2.569
7	−0.633	0.933	0.450	−0.524
8	−0.662	0.284	−0.127	0.044
9	−0.229	0.568	−0.463	−0.088
10	−0.851	0.515	0.344	−1.333
11	−3.361	−0.436	−1.412	−1.977
12	−0.912	0.567	−1.525	0.120
13	1.594	1.040	−0.017	1.558
14	1.618	0.064	−0.525	0.904
15	−1.260	−1.051	−2.689	−1.437
16	0.288	−1.845	−0.211	0.427
17	0.858	0.281	2.145	0.061
18	−1.752	−0.136	0.787	0.120
19	−0.960	−0.992	−0.452	1.460
20	1.738	0.321	1.267	−0.493
21	−1.008	2.621	2.316	−0.888
22	−1.589	2.804	0.258	−0.530
23	0.289	2.174	−1.645	−2.757
24	−0.580	1.897	−1.552	−1.452
25	1.213	−0.781	−0.213	0.158

where the numbers in parentheses are the estimated standard errors obtained from a standard regression program. The residual mean square for the regression is 1.22. The first few regression coefficients are declining in absolute magnitude with alternating signs. Since the third coefficient exceeds twice its standard error, a second order autoregressive process would be judged an inadequate representation for this realization. Thus, even if one did not know the nature of the process generating the data, the moving average representation would be suggested as a possibility by the regression coefficients. The regression coefficients estimate the negatives of the c_j, $j \geq 1$, of Theorem 2.6.2. By that theorem, $\beta = -c_1$ and $c_j = -\beta c_{j-1}$, $j = 2, 3, \ldots$. Therefore, we arrange the regression coefficients as in Table 8.3.2. The initial estimator of β is obtained by regressing the first row on the second row of that table. This regression yields $\tilde{\beta} = 0.697$.

Our initial estimator for e_0 is

$$\begin{aligned}\tilde{e}_0 &= 0.685Y_1 - 0.584Y_2 + 0.400Y_3 - 0.198Y_4 + 0.020Y_5 + 0.014Y_6 + 0.017Y_7 \\ &= 0.974 .\end{aligned}$$

The values of $e_t(Y; 0.697)$ are calculated using (8.3.10), and the values of $W_t(Y; 0.697)$ are calculated using (8.3.11). We have

$$\begin{aligned}&e_t(Y; 0.697) \\ &\quad = \begin{cases} Y_1 - 0.697\tilde{e}_0 = 1.432 - 0.697(0.974) = 0.753 , & t = 1 , \\ Y_t - 0.697e_{t-1}(Y; 0.697) , & t = 2, 3, \ldots, 100 , \end{cases}\end{aligned}$$

$$\begin{aligned}&W_t(Y; 0.697) \\ &\quad = \begin{cases} \tilde{e}_0 = 0.974 , & t = 1 , \\ e_{t-1}(Y; 0.697) - 0.697W_{t-1}(Y; 0.697) , & t = 2, 3, \ldots, 100 . \end{cases}\end{aligned}$$

The first five observations are displayed in Table 8.3.3. Regressing $e_t(Y; \tilde{\beta})$ on $W_t(Y; \tilde{\beta})$ gives a coefficient of 0.037 and an estimate of $\hat{\beta} = 0.734$. The estimated standard error is 0.076, and the residual mean square is 1.16.

Table 8.3.2. Observations for Regression Estimation of an Initial Estimate for β

j	Regression Coefficient $-c_j$	Multiplier of β
1	0.685	1
2	−0.584	−0.685
3	0.400	0.584
4	−0.198	−0.400
5	0.020	0.198
6	0.014	−0.020
7	0.017	−0.014

Table 8.3.3. First Five Observations Used in Gauss–Newton Computation

t	$e_t(Y; 0.697)$	$W_t(Y; 0.697)$
1	0.753	0.974
2	−0.868	0.074
3	−1.154	−0.920
4	−1.732	−0.513
5	0.913	−1.375

Maximum likelihood estimation of the parameters of moving average processes is discussed in Section 8.4. Most computer programs offer the user the option of specifying initial values for the likelihood maximization routine or of permitting the program to find initial values. Because of our preliminary analysis, we use 0.697 as our initial value in the likelihood computations. The maximum likelihood estimator computed in SAS/ETS® is 0.725 with an estimated standard error of 0.071. The maximum likelihood estimator of the error variance adjusted for degrees of freedom is $\hat{\sigma}^2 = 1.17$. The similarity of the two sets of estimates is a reflection of the fact that the estimators have the same limiting distribution. ▲▲

The theory we have presented is all for large samples. There is some evidence that samples must be fairly large before the results are applicable. For example, Macpherson (1981) and Nelson (1974) have conducted Monte Carlo studies in which the empirical variances for the estimated first order moving average parameter are about 1.1 to 1.2 times that based on large sample theory for samples of size 100 and $0 \leq \beta \leq 0.7$. Unlike the estimator for the autoregressive process, the distribution of the nonlinear estimator of β for β near zero differs considerably from that suggested by asymptotic theory for n as large as 100. In the Macpherson study the variance of $\hat{\beta}$ for $\beta = 0.1$ and $n = 100$ was 1.17 times that suggested by asymptotic theory.

In Figure 8.3.1 we compare an estimate of the density of $\hat{\beta}$ for $\beta = 0.7$ with the normal density suggested by the asymptotic theory. The empirical density is based on 15,000 samples of size 100. The estimator of β is the maximum likelihood estimator determined by a grid search procedure.

The estimated density for $\hat{\beta}$ is fairly symmetric about 0.7 with a mean of 0.706. However, the empirical density is considerably flatter than the normal approximation. The variance of the empirical distribution is 0.0064, compared to 0.0051 for the normal approximation.

8.4. AUTOREGRESSIVE MOVING AVERAGE TIME SERIES

In this section, we treat estimation of the parameters of time series with representation

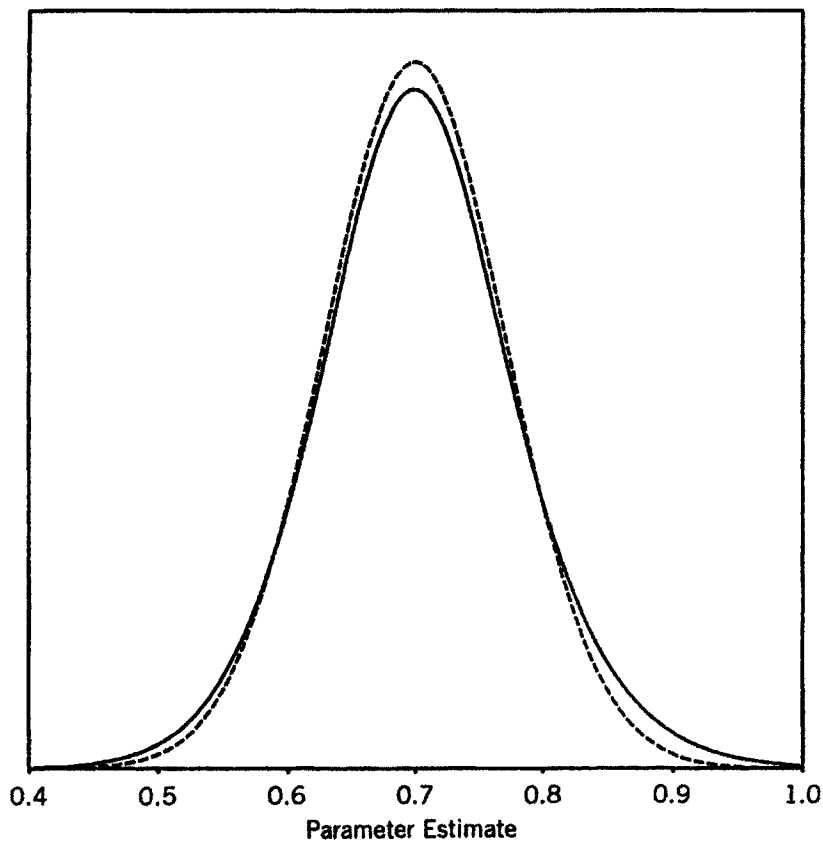

FIGURE 8.3.1. Estimated density of maximum likelihood estimator of β compared with normal approximation for $\beta = 0.7$ and $n = 1.00$. (Dashed line is normal density.)

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j} = e_t + \sum_{i=1}^{q} \beta_i e_{t-i}, \tag{8.4.1}$$

where the e_t are independent $(0, \sigma^2)$ random variables, and the roots of

$$A(m; \boldsymbol{\alpha}) = m^p + \sum_{j=1}^{p} \alpha_j m^{p-j} = 0, \tag{8.4.2}$$

and of

$$B(s; \boldsymbol{\beta}) = s^q + \sum_{i=1}^{q} \beta_i s^{q-i} = 0 \tag{8.4.3}$$

are less than one in absolute value.

Estimators of the parameters of the process can be defined in a number of different ways. We consider three estimators obtained by minimizing three

different functions. If Y_t is a stationary normal time series, the logarithm of the likelihood is

$$L_n(\zeta) = -0.5n \log 2\pi - 0.5 \log|\Sigma_{YY}(\zeta)| - 0.5\mathbf{Y}'\Sigma_{YY}^{-1}(\zeta)\mathbf{Y} \tag{8.4.4}$$

where $\zeta' = (\theta', \sigma^2)$, $\theta' = (\alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q)$, $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$, and $\Sigma_{YY} = \Sigma_{YY}(\zeta) = E\{\mathbf{Y}\mathbf{Y}'\}$.

The estimator that maximizes $L_n(\zeta)$ is often called the maximum likelihood estimator or Gaussian likelihood estimator even if Y_t is not normal. Let $\sigma^{-2}\Sigma_{YY}(\zeta) = \mathbf{M}_{YY}(\theta)$. Then the maximum likelihood estimator of θ can be obtained by minimizing

$$l_n(\theta) = n^{-1}|\mathbf{M}_{YY}(\theta)|^{1/n}\mathbf{Y}'\mathbf{M}_{YY}^{-1}(\theta)\mathbf{Y}\,. \tag{8.4.5}$$

The estimator of θ obtained by minimizing

$$Q_{1n}(\theta) = n^{-1}\mathbf{Y}'\mathbf{M}_{YY}^{-1}(\theta)\mathbf{Y} \tag{8.4.6}$$

is called the least squares estimator or the unconditional least squares estimator. An approximation to the least squares estimator is the estimator that minimizes

$$Q_{2n}(\theta) = n^{-1}\sum_{t=p+1}^{n}\left[Y_t + \sum_{j=1}^{t-1} d_j(\theta)Y_{t-j}\right]^2, \tag{8.4.7}$$

where the $d_j(\theta)$ are defined in Theorem 2.7.2. Observe that $Q_{2n}(\theta)$ is the average of the squares obtained by truncating the infinite autoregressive representation for e_t. In Corollary 8.4.1 of this section, we show that the estimators that minimize $l_n(\theta)$, $Q_{1n}(\theta)$, and $Q_{2n}(\theta)$ have the same limiting distribution for stationary invertible time series.

To obtain the partial derivatives associated with the estimation, we express (8.4.1) as

$$e_t(\mathbf{Y}; \theta) = Y_t + \sum_{j=1}^{p}\alpha_j Y_{t-j} - \sum_{i=1}^{q}\beta_i e_{t-i}(\mathbf{Y}; \theta)\,. \tag{8.4.8}$$

Differentiating both sides of (8.4.8), we have

$$\begin{aligned}
-\frac{\partial e_t(\mathbf{Y}; \theta)}{\partial \alpha_j} &= W_{\alpha_j, t}(\mathbf{Y}; \theta) \\
&= -Y_{t-j} - \sum_{s=1}^{q}\beta_s W_{\alpha_j, t-s}(\mathbf{Y}; \theta)\,, \qquad j = 1, 2, \ldots, p\,, \\
-\frac{\partial e_t(\mathbf{Y}; \theta)}{\partial \beta_i} &= W_{\beta_i, t}(\mathbf{Y}; \theta) \\
&= e_{t-i}(\mathbf{Y}; \theta) - \sum_{s=1}^{q}\beta_s W_{\beta_i, t-s}(\mathbf{Y}; \theta)\,, \qquad i = 1, 2, \ldots, q\,.
\end{aligned} \tag{8.4.9}$$

Using the initial conditions

$$W_{\alpha_j,t}(\mathbf{Y};\boldsymbol{\theta}) = 0\,, \qquad j = 1, 2, \ldots, p\,,$$
$$W_{\beta_i,t}(\mathbf{Y};\boldsymbol{\theta}) = 0\,, \qquad i = 1, 2, \ldots, q\,,$$

for $t \leq p$ and $e_{t-i}(\mathbf{Y};\boldsymbol{\theta}) = 0$ for $t - i \leq p$, the derivatives of $Q_{2n}(\boldsymbol{\theta})$ are defined recursively. The $W_{\alpha_j,t}(\mathbf{Y};\boldsymbol{\theta})$ and $W_{\beta_i,t}(\mathbf{Y};\boldsymbol{\theta})$ are autoregressive moving averages of Y_t. Therefore, if the roots of (8.4.2) and (8.4.3) associated with the $\boldsymbol{\theta}$ at which the derivatives are evaluated are less than one in absolute value, the effect of the initial conditions dies out and the derivatives converge to autoregressive moving average time series as t increases.

The large sample properties of the estimator associated with $Q_{2n}(\boldsymbol{\theta})$ are given in Theorem 8.4.1. We prove the theorem for iid$(0, \sigma^2)$ errors, but the result also holds for martingale difference errors satisfying the conditions of Theorem 8.2.1.

Theorem 8.4.1. Let the stationary time series Y_t satisfy (8.4.1) where the e_t are iid$(0, \sigma^2)$ random variables. The parameter space Θ is such that all roots of (8.4.2) and (8.4.3) are less than one in absolute value, and (8.4.2) and (8.4.3) have no common roots. Let $\boldsymbol{\theta}^0$ denote the true parameter, which is in the interior of the parameter space Θ. Let $\hat{\boldsymbol{\theta}}_2$ be the value of $\boldsymbol{\theta}$ in the closure of Θ that minimizes $Q_{2n}(\boldsymbol{\theta})$. Then

$$[\hat{\boldsymbol{\theta}}_2', Q_{2n}(\hat{\boldsymbol{\theta}}_2)] \xrightarrow{P} [\boldsymbol{\theta}^{0\prime}, (\sigma^0)^2]$$

and

$$n^{1/2}(\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^0) \xrightarrow{\mathcal{L}} N[0, \mathbf{V}_{00}^{-1}(\sigma^0)^2]\,, \tag{8.4.10}$$

where

$$\mathbf{V}_{00} = \lim_{t\to\infty} E\{\mathbf{W}_{0t}'\mathbf{W}_{0t}\}\,,$$

$$\mathbf{W}_{0t} = [W_{\alpha_1,t}(\mathbf{Y};\boldsymbol{\theta}^0), W_{\alpha_2,t}(\mathbf{Y};\boldsymbol{\theta}^0), \ldots, W_{\alpha_p,t}(\mathbf{Y};\boldsymbol{\theta}^0),$$
$$W_{\beta_1,t}(\mathbf{Y};\boldsymbol{\theta}^0), W_{\beta_2,t}(\mathbf{Y};\boldsymbol{\theta}^0), \ldots, W_{\beta_q,t}(\mathbf{Y};\boldsymbol{\theta}^0)]\,,$$

the elements of $\mathbf{W}_{\theta t}$ are defined in (8.4.9), and $\mathbf{W}_{0t}$ is $\mathbf{W}_{\theta t}$ evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$.

Proof. Let Θ_ϵ be a compact space such that, for any $\boldsymbol{\theta} \in \Theta_\epsilon$, the roots of (8.4.2) are less than or equal to one in absolute value, the roots of (8.4.3) are less than or equal to $1 - \epsilon$ in absolute value for some $\epsilon > 0$, and $\boldsymbol{\theta}^0$ is in the interior of Θ_ϵ. For any $\boldsymbol{\theta}$ in Θ_ϵ,

$$Q_{2n}(\boldsymbol{\theta}) \xrightarrow{P} V\{Z_t(\boldsymbol{\theta})\}\,,$$

where

$$Z_t(\boldsymbol{\theta}) = [\mathscr{B}^p A(\mathscr{B}^{-1}; \boldsymbol{\alpha})][\mathscr{B}^q B(\mathscr{B}^{-1}; \boldsymbol{\beta})]^{-1} Y_t$$
$$= \frac{[\mathscr{B}^p A(\mathscr{B}^{-1}; \boldsymbol{\alpha})][\mathscr{B}^q B(\mathscr{B}^{-1}; \boldsymbol{\beta}^0)]}{[\mathscr{B}^q B(\mathscr{B}^{-1}; \boldsymbol{\beta})][\mathscr{B}^p A(\mathscr{B}^{-1}; \boldsymbol{\alpha}^0)]} e_t$$

and the polynomials are defined in (8.4.2) and (8.4.3). The time series $Z_t(\boldsymbol{\theta})$ is a stationary autoregressive moving average because the roots of $B(s; \boldsymbol{\beta})$ are less than one in absolute value. See Corollary 2.2.2.3, equation (2.7.14), and Theorem 6.3.5. The convergence is uniform by Lemma 5.5.5, because $Z_t(\boldsymbol{\theta})$ and its derivatives are infinite moving averages with exponentially declining coefficients.

Now

$$[\mathscr{B}^p A(\mathscr{B}^{-1}; \boldsymbol{\alpha}^0)][\mathscr{B}^q B(\mathscr{B}^{-1}; \boldsymbol{\beta}^0)]^{-1} Y_t = e_t$$

defines the unique minimum variance prediction error for the predictor of Y_t based on $Y_{t-1}, Y_{t-2}, \ldots$. See Section 2.9. Therefore, if $\boldsymbol{\theta} \neq \boldsymbol{\theta}^0$ and $\boldsymbol{\theta} \in \Theta_\epsilon$,

$$V\{Z_t(\boldsymbol{\theta})\} > V\{Z_t(\boldsymbol{\theta}^0)\}.$$

If any of the roots of $B(\mathscr{B}; \boldsymbol{\beta})$ are equal to one in absolute value, $Q_{2n}(\boldsymbol{\theta})$ increases without bound as $n \to \infty$. It follows that the condition (5.5.5) of Lemma 5.5.1 is satisfied for $\hat{\boldsymbol{\theta}}_2$ defined by the minimum of $Q_{2n}(\boldsymbol{\theta})$ over the closure of Θ. Hence, $\hat{\boldsymbol{\theta}}_2$ converges to $\boldsymbol{\theta}^0$ in probability as $n \to \infty$. Because $Q_{2n}(\boldsymbol{\theta})$ is a continuous function of $\boldsymbol{\theta}$ that converges uniformly to $V\{Z_t(\boldsymbol{\theta})\}$ on Θ_ϵ, and $\hat{\boldsymbol{\theta}}_2 \xrightarrow{p} \boldsymbol{\theta}^0$, it follows that $Q_{2n}(\hat{\boldsymbol{\theta}}_2) \xrightarrow{p} (\sigma^0)^2$.

The first derivatives of $e_t(\mathbf{Y}; \boldsymbol{\theta})$ are defined in (8.4.9). The second derivatives can be defined in a similar manner. For example,

$$\frac{\partial^2 e_t(\mathbf{Y}; \boldsymbol{\theta})}{\partial \alpha_j \, \partial \alpha_i} = -\frac{\partial W_{\alpha_j, t}(\mathbf{Y}; \boldsymbol{\theta})}{\partial \alpha_i} = \sum_{s=1}^{q} \beta_s \frac{\partial W_{\alpha_j, t-s}(\mathbf{Y}; \boldsymbol{\theta})}{\partial \alpha_i}.$$

Therefore, the second derivatives are also autoregressive moving averages. It follows that the matrix

$$\mathbf{B}_n(\boldsymbol{\theta}) = 0.5 \frac{\partial^2 Q_{2n}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} = 0.5 n^{-1} \sum_{t=1}^{n} \left[e_t(\mathbf{Y}; \boldsymbol{\theta}) \frac{\partial^2 e_t(\mathbf{Y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \, \partial \boldsymbol{\theta}'} + \mathbf{W}'_{\theta t} \mathbf{W}_{\theta t} \right]$$

converges uniformly to $\mathbf{B}(\boldsymbol{\theta}) = \lim_{t \to \infty} 0.5 E\{\mathbf{W}'_{\theta t} \mathbf{W}_{\theta t}\}$, which is a continuous function of $\boldsymbol{\theta}$, on some convex compact neighborhood S of $\boldsymbol{\theta}^0$ containing $\boldsymbol{\theta}^0$ as an interior point.

By the Taylor series arguments used in the proof of Theorem 5.5.1,

$$\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^0 = \mathbf{B}^{-1}(\boldsymbol{\theta}^0) n^{-1} \sum_{t=1}^{n} \mathbf{W}'_{\theta t}(\mathbf{Y}; \boldsymbol{\theta}^0) \left[Y_t + \sum_{j=1}^{t-1} d_j(\boldsymbol{\theta}^0) Y_{t-j} \right] + \mathbf{r}_n,$$

where $\mathbf{B}(\boldsymbol{\theta}^0)$ is defined in that theorem and $\mathbf{r}_n$ is of smaller order than $\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}^0$.

Now,

$$n^{-1/2}\sum_{t=1}^{n}\mathbf{W}'_{\theta t}(\mathbf{Y};\boldsymbol{\theta}^0)\left\{e_t-\left[Y_t+\sum_{j=1}^{t-1}d_j(\boldsymbol{\theta}^0)Y_{t-j}\right]\right\}\xrightarrow{P}0,$$

because the $d_j(\boldsymbol{\theta}^0)$ decline exponentially. The vector $\mathbf{W}'_{\theta t}(\mathbf{Y};\boldsymbol{\theta}^0)$ is a function of $(Y_1, Y_2, \ldots, Y_{t-1})$ and hence is independent of e_t. Following the arguments in the proof of Theorem 5.5.1, we obtain the asymptotic normality of

$$n^{-1/2}\sum_{t=1}^{n}\mathbf{W}'_{\theta t}(\mathbf{Y};\boldsymbol{\theta}^0)e_t,$$

and hence of $n^{1/2}(\hat{\boldsymbol{\theta}}_2-\boldsymbol{\theta}^0)$. Theorem 5.5.1 does not apply directly, because $Y_t+\Sigma_{j=1}^{t-1}d_j(\boldsymbol{\theta})Y_{t-j}$ are not independent and not identically distributed. However, as t increases, every element of $\mathbf{W}'_{\theta t}(\mathbf{Y};\boldsymbol{\theta}^0)$ converges to an autoregressive moving average time series in the e_t, and we obtain $\mathbf{V}_{00}^{-1}(\sigma^0)^2$ as the covariance matrix of the limiting distribution. ▲

The three estimators defined by (8.4.5), (8.4.6), and (8.4.7) have the same limiting behavior.

Theorem 8.4.2. Let $\hat{\boldsymbol{\theta}}_{ml}$ and $\hat{\boldsymbol{\theta}}_1$ be the estimators obtained by minimizing (8.4.5) and (8.4.6), respectively. Then under the assumptions of Theorem 8.4.1, the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_{ml}-\boldsymbol{\theta}^0)$ is equal to the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}^0)$ and is that given in (8.4.10).

Proof. Let $\overset{*}{Y}_t=\overset{*}{Y}_t(\boldsymbol{\theta})$ be an autoregressive moving average (p, q) with parameter $\boldsymbol{\theta}\in\Theta_\epsilon$ and uncorrelated $(0, \sigma^2)$ errors $\overset{*}{e}_t$, where Θ_ϵ is defined in the proof of Theorem 8.4.1. Let

$$\overset{*}{\mathbf{v}}=(\overset{*}{Y}_{1-p}, \overset{*}{Y}_{2-p}, \ldots, \overset{*}{Y}_0, \overset{*}{e}_{1-p}, \overset{*}{e}_{2-p}, \ldots, \overset{*}{e}_0),$$

$\overset{*}{\mathbf{Y}}'_n=(\overset{*}{Y}_1, \overset{*}{Y}_2, \ldots, \overset{*}{Y}_n)$, and $\overset{*}{\mathbf{e}}'_n=(\overset{*}{e}_1, \overset{*}{e}_2, \ldots, \overset{*}{e}_n)$. Then

$$\begin{bmatrix}\overset{*}{\mathbf{v}}\\ \overset{*}{\mathbf{e}}_n\end{bmatrix}=\begin{bmatrix}\mathbf{0}\\ \mathbf{D}\end{bmatrix}\overset{*}{\mathbf{Y}}_n+\begin{pmatrix}\mathbf{I}\\ \mathbf{K}\end{pmatrix}\overset{*}{\mathbf{v}},$$

where $\mathbf{D}=\mathbf{D}(\boldsymbol{\theta})$ is the $n\times n$ lower triangular matrix with $d_{ij}=d_{ij}(\boldsymbol{\theta})=d_{|i-j|}(\boldsymbol{\theta})$, $d_j(\boldsymbol{\theta})$ is defined in Theorem 2.7.2, $\mathbf{0}$ is a $(p+q)\times n$ matrix of zeros, $\mathbf{I}$ is $(p+q)\times(p+q)$, $\overset{*}{\mathbf{K}}=(\mathbf{I}, \mathbf{K}')'$ is a matrix whose elements satisfy

$$\overset{*}{k}_{ij}=\begin{cases}\delta_{ij}, & 1\le i\le p+q, \quad 1\le j\le p+q,\\ -\sum_{r=1}^{q}\beta_r\overset{*}{k}_{i-r,j}-\alpha_{i-j-q}, & p+q+1\le i\le n+p+q, \quad 1\le j\le p,\\ -\sum_{r=1}^{q}\beta_r\overset{*}{k}_{i-r,j}, & p+q+1\le i\le n+p+q, \quad p+1\le j\le p+q,\end{cases}$$

and δ_{ij} is the Kronecker delta. Note that the $\overset{*}{k}_{ij}=\overset{*}{k}_{ij}(\boldsymbol{\theta})$ satisfy the same difference

equation as the $d_{ij}(\boldsymbol{\theta})$, but with different initial values. Galbraith and Galbraith (1974) give the expressions

$$\mathbf{M}_{YY}^{-1} = \mathbf{D}'\mathbf{D} - \mathbf{D}'\mathbf{K}(\boldsymbol{\Lambda}^{-1} + \mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{D} \tag{8.4.11}$$

and

$$|\mathbf{M}_{YY}| = |\boldsymbol{\Lambda}|\,|\boldsymbol{\Lambda}^{-1} + \mathbf{K}'\mathbf{K}|\,,$$

where

$$\boldsymbol{\Lambda} = \sigma^{-2}\mathbf{V}\{(\overset{*}{Y}_{1-p}, \overset{*}{Y}_{2-p}, \ldots, \overset{*}{Y}_{0}, \overset{*}{e}_{1-q}, \overset{*}{e}_{2-q}, \ldots, \overset{*}{e}_{0})'\}\,.$$

The dependence of all matrices on $\boldsymbol{\theta}$ has been suppressed to simplify the expressions. Observe that $Q_{2n} = n^{-1}\mathbf{Y}'\mathbf{D}'\mathbf{D}\mathbf{Y}$ and

$$Q_{2n} - Q_{1n} = n^{-1}\mathbf{Y}'\mathbf{D}'\mathbf{K}(\boldsymbol{\Lambda}^{-1} + \mathbf{K}'\mathbf{K})^{-1}\mathbf{K}'\mathbf{D}\mathbf{Y}\,.$$

To show that $Q_{2n} - Q_{1n}$ converges to zero, we look at the difference on the set $\boldsymbol{\Theta}_\epsilon$. Now

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} |\mathbf{Y}'\mathbf{D}'\mathbf{K}|^2 &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} \sum_{j=1}^{p+q} \left(\sum_{i=1}^{n} Y_i \sum_{r=0}^{n-i} d_r k_{i+r,j}\right) \\ &\leqslant \sum_{j=1}^{p+q} \left(\sum_{i=1}^{n} |Y_i| M\lambda^i\right)^2 \\ &= (p+q)M^2\left(\sum_{i=1}^{n} |Y_i|\lambda^i\right)^2\end{aligned}$$

for some $M < \infty$ and some $0 < \lambda < 1$, because the $d_r(\boldsymbol{\theta})$ and the elements of K decline exponentially. Also, $\boldsymbol{\Lambda}$ is a positive definite matrix for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\epsilon$, and the determinant is uniformly bounded above and below by positive numbers for all n and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. It follows that

$$\sup_{n,\boldsymbol{\theta}\in\boldsymbol{\Theta}_\epsilon} |Q_{2n}(\boldsymbol{\theta}) - Q_{1n}(\boldsymbol{\theta})| = O(n^{-1}) \qquad \text{a.s.}$$

We now investigate the ratio $l_n(\boldsymbol{\theta})[Q_1(\boldsymbol{\theta})]^{-1}$. Let the $n \times n$ lower triangular matrix $\mathbf{T} = \mathbf{T}(\boldsymbol{\theta})$ define the prediction error made in using the best predictor based on $(Y_{t-1}, Y_{t-2}, \ldots, Y_1)$ to predict Y_t. See Theorem 2.9.1 and Theorem 2.10.1. Then,

$$E\{(\mathbf{T}\mathbf{Y})(\mathbf{T}\mathbf{Y})'\} = \mathbf{H} = \operatorname{diag}(h_{11}, h_{22}, \ldots, h_{nn})$$

and

$$\mathbf{M}_{YY}^{-1} = \mathbf{T}'\mathbf{H}^{-1}\mathbf{T}\sigma^2\,,$$

where $h_{ii} \geq \sigma^2$ and $h_{ii} \to \sigma^2$ as $i \to \infty$. Also, letting $\sum_{j=1}^{i-1} b_j Y_{i-j}$ be the best predictor of Y_i,

$$\begin{aligned} h_{ii} &= V\left\{Y_i - \sum_{j=1}^{i-1} b_j Y_{i-j}\right\} \leq V\left\{Y_i - \sum_{j=1}^{i-1} d_j Y_{i-j}\right\} \\ &= V\left\{e_i - \sum_{j=i}^{\infty} d_j Y_{i-j}\right\} \leq \sigma^2 + \left[\sum_{j=i}^{\infty} |d_j|\right]^2 V\{Y_t\} \\ &\leq \sigma^2 + M^2(1-\lambda)^{-2}\lambda^{2i} V\{Y_t\} \end{aligned}$$

for some $M < \infty$ and $0 < \lambda < 1$. Thus,

$$|\mathbf{M}_{YY}| \leq \prod_{t=1}^{n} (1 + \overset{*}{M}\lambda^{2t}) \leq \exp\left\{\overset{*}{M} \sum_{t=1}^{n} \lambda^{2t}\right\} \leq \overset{*}{k}$$

for some $\overset{*}{M} < \infty$ and $\overset{*}{k} < \infty$. It follows that

$$l_n(\boldsymbol{\theta})[Q_{1n}(\boldsymbol{\theta})]^{-1} = |\mathbf{M}_{YY}(\boldsymbol{\theta})|^{1/n} < \overset{*}{k}^{1/n} \to 1.$$

Let the $(p+q)$-vector $\overset{*}{\mathbf{Z}} = \overset{*}{\mathbf{Z}}(\boldsymbol{\theta}) = \mathbf{K}'\mathbf{D}\mathbf{Y}$, where $\overset{*}{Z}_j(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i g_{ij}(\boldsymbol{\theta})$ and

$$g_{ij}(\boldsymbol{\theta}) = \sum_{r=0}^{n-i} d_r k_{i+r,j}.$$

Now the $g_{ij}(\boldsymbol{\theta})$, their first derivatives, and their second derivatives are exponentially declining in i for $\boldsymbol{\theta} \in \Theta_\epsilon$. Also, the first and second derivatives of $k_{ij}(\boldsymbol{\theta})$ are exponentially declining in i. The first and second derivatives of $\boldsymbol{\Lambda}$ are bounded on $\boldsymbol{\theta}_\epsilon$. Therefore, $l_n(\boldsymbol{\theta}) - Q_{1n}(\boldsymbol{\theta})$, $Q_{1n}(\boldsymbol{\theta}) - Q_{2n}(\boldsymbol{\theta})$, and $Q_{2n}(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta})$ and their first and second derivatives converge uniformly to zero in probability.

By the stated derivative properties of $\mathbf{K}(\boldsymbol{\theta})$ and $\boldsymbol{\Lambda}(\boldsymbol{\theta})$, the first and second derivatives of $n^{-1}\log|\mathbf{M}_{YY}(\boldsymbol{\theta})|$ converge uniformly to zero in probability for $\boldsymbol{\theta} \in \Theta_\epsilon$. Therefore, the limits of $l_n(\boldsymbol{\theta})$, of $Q_{1n}(\boldsymbol{\theta})$, and of $Q_{2n}(\boldsymbol{\theta})$ are the same, and the limits of the first and second derivatives of the three quantities are also the same. It follows that the limiting distributions of the three estimators are the same. ▲

We derived the limiting distribution of the estimators for the time series with known mean. Because the limiting behavior of sample autocovariances computed with mean adjusted data is the same as that for autocovariances computed with known mean, the results extend to the unknown mean case.

The mean squares and products of the estimated derivatives converge to the mean squares and products of the derivatives based on the true $\boldsymbol{\theta}^0$. Therefore, the usual nonlinear least squares estimated covariance matrix can be used for inference.

There are a number of computer programs available that compute the Gaussian

maximum likelihood estimates or approximations to them. It is good practice to use the autoregressive representation technique introduced in Section 8.3 to obtain initial estimates for these programs. By Theorem 2.7.2, we can represent the invertible autoregressive moving average by the infinite autoregressive process

$$Y_t = -\sum_{j=1}^{\infty} d_j Y_{t-j} + e_t, \tag{8.4.12}$$

where the d_j are defined in Theorem 2.7.2. Hence, by terminating the sum at a convenient finite number, say k, and estimating the autoregressive parameters, d_1, $d_2, \ldots, d_k$, we can use the definitions of the d_j to obtain initial estimates of the autoregressive moving average parameters.

In using a nonlinear method such as the Gauss–Newton procedure, one must beware of certain degeneracies that can occur. For example, consider the autoregressive moving average (1, 1) time series. If we specify zero initial estimates for both parameters, the derivative with respect to α_1 evaluated at $\alpha_1 = \beta_1 = 0$ is $-Y_{t-1}$. Likewise, the derivative with respect to β_1 evaluated at $\alpha_1 = \beta_1 = 0$ is Y_{t-1}, and the matrix of partial derivatives to be inverted is clearly singular. At second glance, this is not a particularly startling result. It means that a first order autoregressive process with small α_1 behaves very much like a first order moving average process with small β_1 and that both behave much like an autoregressive moving average (1, 1) time series where both α_1 and β_1 are small. Therefore, one should consider the autoregressive moving average (1, 1) representation only if at least one of the trial parameter values is well away from zero. Because the autoregressive moving average (1, 1) time series with $\alpha_1 = \beta_1$ is a sequence of uncorrelated random variables, the singularity occurs whenever the initial values are taken to be equal.

In developing our estimation theory for autoregressive moving averages, we have assumed that estimation is carried out for a correctly specified model. In practice, one often is involved in specifying the model at the same time that one is constructing estimates. A number of criteria have been suggested for use in model selection, some developed with time series applications in mind. Because the estimated model parameters are asymptotically normally distributed and because the estimation procedures are closely related to regression, model selection procedures developed for regression models are also applicable to the autoregressive moving average problem. A test based on such statistics was used in Example 8.2.1.

Another selection procedure is based on the variance of regression prediction. See Mallows (1973) and Akaike (1969a). Assume one fits the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{8.4.13}$$

where $\mathbf{y}$ is $n \times 1$, $\mathbf{X}$ is $n \times k$, $\boldsymbol{\beta}$ is $k \times 1$, and $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$. If one predicts the y-value for each of the observed $\mathbf{X}_t$ rows of $\mathbf{X}$, the average of the n prediction variances is

$$n^{-1}\sum_{t=1}^{n} V\{\hat{Y}_t - Y_t\} = n^{-1}\sigma^2 \sum_{t=1}^{n} [1 + \mathbf{X}_t(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_t'] = n^{-1}(n+k)\sigma^2 , \tag{8.4.14}$$

where $\hat{Y}_t = \mathbf{X}_t\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and k is the dimension of $\mathbf{X}_t$. Thus, one might choose the model that minimizes an estimator of the mean square prediction error. To estimate (8.4.14), one requires an estimator of σ^2. The estimator of σ^2 that is most often suggested in the regression literature is

$$\hat{\sigma}^2 = (n - k_M)^{-1}(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}_M'\mathbf{X}_M'\mathbf{X}_M\hat{\boldsymbol{\beta}}_M) , \tag{8.4.15}$$

where $\mathbf{X}_M$ is the model of highest dimension and that dimension is k_M. Thus, the model is chosen to minimize

$$\text{MPE} = n^{-1}(n+k)\hat{\sigma}^2 . \tag{8.4.16}$$

The use of the regression residual mean square for the particular set of regression variables being evaluated as the estimator of σ^2 generally leads to the same model selection.

A criterion closely related to the mean square prediction error is the criterion called AIC, introduced by Akaike (1973). This criterion is

$$\text{AIC} = -2\log L(\hat{\boldsymbol{\theta}}) + 2k , \tag{8.4.17}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood function evaluated at the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$, and k is the dimension of $\boldsymbol{\theta}$. For normal autoregressive moving average models,

$$-2\log L(\hat{\boldsymbol{\theta}}) = \log|\boldsymbol{\Sigma}_{YY}(\hat{\boldsymbol{\theta}})| + n + n\log 2\pi ,$$

where $\boldsymbol{\Sigma}_{YY}(\hat{\boldsymbol{\theta}})$ is the covariance matrix of $(Y_1, Y_2, \ldots, Y_n)$ evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The determinant of $\boldsymbol{\Sigma}_{YY}(\boldsymbol{\theta})$ can be expressed as a product of the prediction error variances (see Theorem 2.9.3). Also, the variances of the prediction errors converge to the error variance σ^2 of the process as t increases. Therefore, $\log|\boldsymbol{\Sigma}_{YY}(\hat{\boldsymbol{\theta}})|$ is close to $n\log\hat{\sigma}_m^2$, where $\hat{\sigma}_m^2 = \sigma_m^2(\hat{\boldsymbol{\theta}})$ is the maximum likelihood estimator of σ^2. It follows that AIC and MPE are closely related, because

$$n^{-1}\log|\boldsymbol{\Sigma}_{YY}(\hat{\boldsymbol{\theta}})| + 2n^{-1}k \doteq \log[(1 + 2n^{-1}k)\hat{\sigma}_m^2] \doteq \log \text{MPE} ,$$

when the $\hat{\sigma}^2$ of MPE is close to $(n-k)^{-1}n\hat{\sigma}_m^2$. The AIC criterion is widely used, although it is known that this criterion tends to select rather high order models and will overestimate the true order of a finite order autoregressive model. Because of the tendency to overestimate the order of the model, a number of related criteria have been developed, and considerable research has been conducted on the use of model selection criteria. See Parzen (1974, 1977), Shibata (1976, 1986), Hannan

and Quinn (1979), Hannan (1980), Hannan and Rissanen (1982), Bhansali (1991), Findley (1985), Schwartz (1978), and Findley and Wei (1989).

Example 8.4.1. The data of Table 10.B.2 of Appendix 10.B are artificially created data generated by an autoregressive moving average model. As a first step in the analysis, we fit a pure autoregressive model of order 10 by ordinary least squares. The fitted model is

$$\begin{aligned}\hat{Y}_t = &- \underset{(0.104)}{0.099} + \underset{(0.105)}{2.153}\, Y_{t-1} - \underset{(0.250)}{2.032}\, Y_{t-2} + \underset{(0.332)}{0.778}\, Y_{t-3} \\ &+ \underset{(0.352)}{0.072}\, Y_{t-4} - \underset{(0.361)}{0.440}\, Y_{t-5} + \underset{(0.366)}{0.363}\, Y_{t-6} \\ &- \underset{(0.368)}{0.100}\, Y_{t-7} - \underset{(0.355)}{0.144}\, Y_{t-8} + \underset{(0.268)}{0.150}\, Y_{t-9} - \underset{(0.113)}{0.082}\, Y_{t-10},\end{aligned}$$

where the numbers in parentheses are the estimated standard errors of an ordinary regression program. We added seven observations equal to the sample mean to the beginning of the data set and then created ten lags of the data. This is a compromise between the estimators (8.2.5) and (8.2.8). Also, we keep the same number of observations when we fit reduced autoregressive models. There are 97 observations in the regression. The residual mean square with 86 degrees of freedom is 1.020. The fourth order autoregressive process estimated by ordinary least squares based on 97 observations is

$$\hat{Y}_t = - \underset{(0.102)}{0.086} + \underset{(0.099)}{2.170}\, Y_{t-1} - \underset{(0.224)}{2.074}\, Y_{t-2} + \underset{(0.225)}{0.913}\, Y_{t-3} - \underset{(0.100)}{0.209}\, Y_{t-4}$$

with a residual mean square of 0.996. This model might be judged acceptable if one were restricting oneself to autoregressive processes because the test for the fourth order versus the tenth order gives $F = 0.63$, where the distribution of F can be approximated by Snedecor's F with 6 and 86 degrees of freedom.

We consider several alternative models estimated by Gaussian maximum likelihood in Table 8.4.1. All calculations were done in SAS/ETS®. The estimate

Table 8.4.1. Comparison of Alternative Models Estimated by Maximum Likelihood

Model	$\hat{\sigma}^2$	AIC	MPE
AR(10)	1.005	301.4	1.115
AR(4)	0.996	294.5	1.046
ARMA(2,1)	1.032	296.9	1.074
ARMA(2,2)	0.991	294.0	1.041
ARMA(2,3)	0.990	295.0	1.050
ARMA(3,1)	1.010	295.8	1.061
ARMA(3,2)	0.992	295.2	1.052

of σ^2 given in the table is the maximum likelihood estimator adjusted for degrees of freedom,

$$\hat{\sigma}^2 = (n-r)^{-1}\,\mathrm{tr}\{\mathbf{Y}'\hat{\mathbf{M}}_{YY}^{-1}\mathbf{Y}\},$$

where $\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_n)$, $\hat{\mathbf{M}}_{YY}$ is the estimate of $\mathbf{M}_{YY}$, $\boldsymbol{\Sigma}_{YY} = \sigma^2\mathbf{M}_{YY}$, $\boldsymbol{\Sigma}_{YY}$ is the $n \times n$ covariance matrix of $\mathbf{Y}$, and r is the total number of parameters estimated. The mean prediction error was calculated as

$$\mathrm{MPE} = n^{-1}(n+r)\hat{\sigma}^2,$$

where $n = 100$, and the AIC is defined in (8.4.17).

Several of the models have similar properties. On the basis of the AIC criterion, one would choose the autoregressive moving average (2, 2). For this model, the estimated mean is -0.26 with a standard error of 0.56, and the other estimated parameters are

$$\begin{array}{rcccc}(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2) = (- & 1.502, & 0.843, & 0.715, & 0.287).\\ & (0.109) & (0.110) & (0.061) & (0.059)\end{array}$$ ▲▲

Example 8.4.2. As an example of estimation for a process containing several parameters, we fit an autoregressive moving average to the United States monthly unemployment rate from October 1949 to September 1974 (300 observations). The periodogram of this time series was discussed in Section 7.2. Because of the very large contribution to the total sum of squares from the seasonal frequencies, it seems reasonable to treat the time series as if there were a different mean for each month. Therefore, we analyze the deviations from monthly means, which we denote by Y_t. The fact that the periodogram ordinates close to the seasonal frequencies are large relative to those separated from the seasonal frequencies leads us to expect a seasonal component in a representation for Y_t.

As the first step in the analysis, we regress Y_t on Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-12}, Y_{t-13}, Y_{t-14}, Y_{t-15}, Y_{t-24}, Y_{t-25}, Y_{t-26}, Y_{t-27}, Y_{t-36}, Y_{t-37}, Y_{t-38}, Y_{t-39}, Y_{t-48}, Y_{t-49}, Y_{t-50}, and Y_{t-51}. Notice that we are anticipating a model of the "component" or "multiplicative" type, so that when we include a variable of lag 12, we also include the next three lags. That is, we are anticipating a model of the form

$$\begin{aligned}(1 - \theta_1\mathcal{B} - \theta_2\mathcal{B}^2 - \theta_3\mathcal{B}^3)(1 - \theta_4\mathcal{B}^{12} - \theta_5\mathcal{B}^{24} - \theta_6\mathcal{B}^{36} - \theta_7\mathcal{B}^{48})Y_t\\ = Y_t - \theta_1 Y_{t-1} - \theta_2 Y_{t-2} - \theta_3 Y_{t-3} - \theta_4 Y_{t-12}\\ + \theta_1\theta_4 Y_{t-13} + \theta_2\theta_4 Y_{t-14} + \cdots + \theta_3\theta_7 Y_{t-51} = e_t.\end{aligned}$$

In calculating the regression equation we added 36 zeros to the beginning of the data set, lagged Y_t the requisite number of times, and used the last 285 observations in the regression. This is a compromise between the forms (8.2.5) and (8.2.8) for the estimation of the autoregressive parameters. The regression vectors for the explanatory variables Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-12}, Y_{t-13}, Y_{t-14}, and Y_{t-15}

Table 8.4.2. Regression Coefficients Obtained in Preliminary Autoregressive Fit to United States Monthly Unemployment Rate

Variable	Coefficient	Standard Error of Coefficient
Y_{t-1}	1.08	0.061
Y_{t-2}	0.06	0.091
Y_{t-3}	−0.16	0.063
Y_{t-12}	0.14	0.060
Y_{t-13}	−0.22	0.086
Y_{t-14}	0.00	0.086
Y_{t-15}	0.06	0.059
Y_{t-24}	0.18	0.053
Y_{t-25}	−0.17	0.075
Y_{t-26}	−0.09	0.075
Y_{t-27}	0.09	0.054
Y_{t-36}	0.08	0.054
Y_{t-37}	−0.13	0.075
Y_{t-38}	0.04	0.075
Y_{t-39}	0.00	0.054
Y_{t-48}	0.11	0.055
Y_{t-49}	−0.01	0.075
Y_{t-50}	−0.01	0.075
Y_{t-51}	−0.09	0.053

contain all observed values, but the vectors for longer lags contain zeros for some of the initial observations.

The regression coefficients and standard errors are given in Table 8.4.2. The data seem to be consistent with the component type of model. The coefficients for $Y_{t-12i-1}$ are approximately the negatives of the coefficients on Y_{t-12i} for $i = 1, 2, 3$. Even more consistently, the sum of the three coefficients for $Y_{t-12i-j}$ for $j = 1, 2, 3$ is approximately the negative of the coefficient for Y_{t-12i}, $i = 1, 2, 3, 4$. The individual coefficients show variation about the anticipated relationships, but they give us no reason to reject the component model. The residual mean square for this regression is 0.0634 with 254 degrees of freedom. There are 285 observations in the regression and 19 regression variables. We deduct an additional 12 degrees of freedom for the 12 means previously estimated. We also fit the model with Y_{t-1}, Y_{t-2}, and the corresponding lags of 12 as well as the model with Y_{t-1}, Y_{t-2}, Y_{t-3}, Y_{t-4}, Y_{t-5}, and the corresponding lags of 12. Since the coefficient on Y_{t-3} is almost twice its standard error, while the coefficients on Y_{t-4} and Y_{t-5} were small, we take the third order autoregressive process as our tentative model for the nonseasonal component.

The coefficients for Y_{t-12}, Y_{t-24}, Y_{t-36}, Y_{t-48} are of the same sign, are of small magnitude relative to one, and are declining slowly. The autoregressive co-

efficients for an autoregressive moving average can display this behavior. Therefore, we consider

$$(1 - \theta_1 \mathcal{B} - \theta_2 \mathcal{B}^2 - \theta_3 \mathcal{B}^3)(1 - \delta \mathcal{B}^{12}) Y_t = e_t + \beta e_{t-12}, \qquad (8.4.18)$$

as a potential model. On the basis of Theorem 2.7.2, the regression coefficients for lags of multiples of 12 should satisfy, approximately, the relations of Table 8.4.3. Regressing the first column of that table on the second two columns, we obtain the initial estimates $\tilde{\delta} = 0.97$, $\tilde{\beta} = -0.83$. The estimate of β is of fairly large absolute value to be estimated from only four coefficients, but we are only interested in obtaining crude values that can be used as start values for the nonlinear estimation. The maximum likelihood estimate of the parameter vector of model (8.4.18) using (1.08, 0.06, −0.16, 0.97, −0.83) as the initial vector is

$$(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\delta}, \hat{\beta}) = (1.152, -0.002, -0.195, 0.817, -0.651)$$

with estimated standard errors of (0.055, 0.085, 0.053, 0.067, 0.093). The residual mean square error is 0.0626 with 268 degrees of freedom. Since this residual mean square is smaller than that associated with the previous regressions, the hypothesis that the restrictions associated with the autoregressive moving average representation are valid is easily accepted. As a check on model adequacy beyond that given by our initial regressions, we estimated four alternative models with the additional terms Y_{t-4}, e_{t-1}, Y_{t-24}, e_{t-24}. In no case was the added term significant at the 5% level using the approximate tests based on the regression statistics.

One interpretation of our final model is of some interest. Define $X_t = Y_t - 1.152Y_{t-1} + 0.002Y_{t-2} + 0.195Y_{t-3}$. Then X_t has the autoregressive moving average representation

$$X_t = 0.817X_{t-12} - 0.651e_{t-12} + e_t,$$

where the e_t are uncorrelated (0, 0.0626) random variables. Now X_t would have this representation if it were the sum of two independent time series $X_t = S_t + v_t$, where v_t is a sequence of uncorrelated (0, 0.0499) random variables,

$$S_t = 0.817S_{t-12} + u_t,$$

Table 8.4.3. Calculation of Initial Estimates of δ and β

j	Regression Coefficients $-d_j$	Multipliers for δ	Multipliers for β
1	0.14	1	1
2	0.18	0	−0.14
3	0.08	0	−0.18
4	0.11	0	−0.08

and u_t is a sequence of $(0, 0.0059)$ random variables. In such a representation, S_t can be viewed as the "seasonal component" and the methods of Section 4.5 could be used to construct a filter to estimate S_t. ▲▲

8.5. PREDICTION WITH ESTIMATED PARAMETERS

We now investigate the use of the estimated parameters of autoregressive moving average time series in prediction. Prediction was introduced in Section 2.9 assuming the parameters to be known. The estimators of the parameters of stationary finite order autoregressive invertible moving average time series discussed in this chapter possess errors whose order in probability is $n^{-1/2}$. For such time series, the use of the estimated parameters in prediction increases the prediction error by a quantity of $O_p(n^{-1/2})$.

Let the time series Y_t be defined by

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j} = \sum_{i=1}^{q} \beta_i e_{t-i} + e_t , \tag{8.5.1}$$

where the roots of

$$m^p + \sum_{j=1}^{p} \alpha_j m^{p-j} = 0$$

and of

$$r^q + \sum_{i=1}^{q} \beta_i r^{q-i} = 0$$

are less than one in absolute value and the e_t are independent $(0, \sigma^2)$ random variables with $E\{e_t^4\} = \eta\sigma^4$. Let

$$\boldsymbol{\theta}' = (-\alpha_1, -\alpha_2, \ldots, -\alpha_p, \beta_1, \beta_2, \ldots, \beta_q)$$

denote the vector of parameters of the process.

When $\boldsymbol{\theta}$ is known, the best one-period-ahead predictor for Y_t is given by Theorems 2.9.1 and 2.9.3. The predictor for large n is given in (2.9.25). The large n predictor obtained by replacing α_j and β_i in (2.9.25) by $\hat{\alpha}_j$ and $\hat{\beta}_i$ is

$$\tilde{Y}_{n+1}(Y_1, \ldots, Y_n) = -\sum_{j=1}^{p} \hat{\alpha}_j Y_{n+1-j} + \sum_{i=1}^{q} \hat{\beta}_i \tilde{e}_{n+1-i}(Y; \hat{\boldsymbol{\theta}}) , \tag{8.5.2}$$

where

$$\tilde{e}_t(Y;\boldsymbol{\theta}) = \begin{cases} 0, & t = p-q+1, p-q+2, \ldots, p \\ Y_t + \sum_{j=1}^{p} \hat{\alpha}_j Y_{t-j} - \sum_{i=1}^{q} \hat{\beta}_i \tilde{e}_{t-i}(Y;\boldsymbol{\theta}), & t = p+1, p+2, \ldots, n. \end{cases} \tag{8.5.3}$$

Theorem 8.5.1. Let Y_t be the time series defined in (8.5.1). Let $\hat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta} = (-\alpha_1, \ldots, -\alpha_p, \beta_1, \ldots, \beta_q)'$ such that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$. Then

$$\hat{Y}_{n+1}(Y_1, \ldots, Y_n) - \tilde{Y}_{n+1}(Y_1, \ldots, Y_n) = O_p(n^{-1/2}),$$

where $\hat{Y}_{n+1}(Y_1, \ldots, Y_n)$ is defined in (2.9.25) and $\tilde{Y}_{n+1}(Y_1, \ldots, Y_n)$ in (8.5.2).

Proof. We write

$$\begin{aligned} \tilde{Y}_{n+1}(Y_1, \ldots, Y_n) = &- \sum_{j=1}^{p} \alpha_j Y_{n+1-j} - \sum_{j=1}^{p} (\hat{\alpha}_j - \alpha_j) Y_{n+1-j} \\ &+ \sum_{i=1}^{q} \beta_i \tilde{e}_{n+1-i}(Y;\boldsymbol{\theta}) \\ &+ \sum_{k=1}^{p+q} \frac{\partial \sum_{i=1}^{q} \beta_i^{\dagger} \tilde{e}_{n+1-i}(Y;\boldsymbol{\theta}^{\dagger})}{\partial \theta_k} (\theta_k - \hat{\theta}_k), \end{aligned}$$

where $\boldsymbol{\theta}^{\dagger}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ and, for example,

$$\frac{\partial \beta_1^{\dagger} \tilde{e}_n(Y;\boldsymbol{\theta}^{\dagger})}{\partial \beta_1} = \tilde{e}_n(Y;\boldsymbol{\theta}^{\dagger}) - \beta_1^{\dagger} \left[\tilde{e}_{n-1}(Y;\boldsymbol{\theta}^{\dagger}) + \sum_{i=1}^{q} \beta_i^{\dagger} \frac{\partial \tilde{e}_{n-i}(Y;\boldsymbol{\theta}^{\dagger})}{\partial \beta_1} \right]$$

and

$$\tilde{e}_n(Y;\boldsymbol{\theta}^{\dagger}) = Y_n + \sum_{j=1}^{p} \alpha_j^{\dagger} Y_{n-j} - \sum_{i=1}^{q} \beta_i^{\dagger} \tilde{e}_{n-i}(Y;\boldsymbol{\theta}^{\dagger}).$$

For $\boldsymbol{\theta}$ such that the roots of the characteristic equations are less than one in absolute value, the derivatives multiplying $\theta_k - \hat{\theta}_k$ are, except for the initial effects, stationary time series. Since $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$, the result follows. ▲

Theorem 8.5.1 generalizes immediately to predictions s periods ahead. On the basis of this result, the prediction variance formulas of Section 2.9 can be used as approximations for the predictor $\tilde{Y}_{n+s}(Y_1, \ldots, Y_n)$.

Additional results are available for the error in predictions for the stationary p-th order autoregressive process. Let Y_t be the stationary process satisfying

$$Y_t + \sum_{j=1}^{p} \alpha_j Y_{t-j} = \alpha_0 + e_t, \qquad t = 1, 2, \ldots, \tag{8.5.4}$$

where the roots of

$$m^p + \sum_{j=1}^{p} \alpha_j m^{p-j} = 0 \tag{8.5.5}$$

are less than one in absolute value and the e_t are independent $(0, \sigma^2)$ random variables. We can also write

$$\mathbf{Y}_t = \mathbf{A}\mathbf{Y}_{t-1} + \mathbf{e}_t, \tag{8.5.6}$$

where $\mathbf{Y}_t = (Y_t, Y_{t-1}, \ldots, Y_{t-p+1}, 1)'$, $\mathbf{e}_t = (e_t, 0, \ldots, 0)'$, and

$$\mathbf{A} = \begin{bmatrix} -\alpha_1 & -\alpha_2 & -\alpha_3 & \cdots & -\alpha_{p-1} & -\alpha_p & \alpha_0 \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}.$$

The least squares estimator of $\boldsymbol{\alpha} = (-\alpha_1, -\alpha_2, \ldots, -\alpha_p, \alpha_0)'$ is

$$\hat{\boldsymbol{\alpha}} = \left[\sum_{t=p+1}^{n} \mathbf{Y}_{t-1}\mathbf{Y}_{t-1}' \right]^{-1} \sum_{t=p+1}^{n} \mathbf{Y}_{t-1} Y_t, \tag{8.5.7}$$

and the estimator of $\mathbf{A}$, denoted by $\hat{\mathbf{A}}$, is obtained from $\mathbf{A}$ by replacing the α_i with the estimator $\hat{\alpha}_i$. Let the least squares predictor of $\mathbf{Y}_{n+s}$, based upon (8.5.7), be

$$\hat{\mathbf{Y}}_{n+s} = \hat{\mathbf{A}}^s \mathbf{Y}_n \tag{8.5.8}$$

for $s \geq 1$.

Theorem 8.5.2. Let the model (8.5.4) hold with the e_t independent identically distributed random variables with a symmetric distribution function. Let Y_1, $Y_2, \ldots, Y_p$ be symmetrically distributed with finite variance, independent of e_{p+1}, $e_{p+2}, \ldots$. Assume the distributions are such that

$$E\{|\hat{\mathbf{A}}^s \mathbf{Y}_n|\} < \infty$$

for $n > p$. Then, for $n > p$,

$$E\{\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s}\} = \mathbf{0}.$$

Proof. The least squares estimators $\hat{\alpha}_i$, $i = 1, 2, \ldots, p$, can be expressed as functions of

$$\sum_{t=p+1}^{n} Y_{t-i}Y_{t-j} - n\bar{y}_{(-i)}\bar{y}_{(-j)}, \qquad i, j = 0, 1, 2, \ldots, p,$$

where $\bar{y}_{(-i)} = (n-p)^{-1}\sum_{t=p+1}^{n} Y_{t-i}$. Also,

$$\hat{\alpha}_0 = \bar{y}_{(0)} + \sum_{i=1}^{p} \hat{\alpha}_i \bar{y}_{(-i)}.$$

Therefore, the $\hat{\alpha}_i$, $i = 1, 2, \ldots, p$, are even functions of $(Y_1, Y_2, \ldots, Y_n)$, and $\hat{\alpha}_0$ is an odd function of $(Y_1, Y_2, \ldots, Y_n)$. Note that adding a constant to each Y_t will change the predictor by the same amount. Therefore, there is no loss of generality in the assumption that $\alpha_0 = 0$. The prediction error is

$$\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s} = \sum_{i=0}^{s-1} \mathbf{A}^i \mathbf{e}_{n+s-i} + (\mathbf{A}^s - \hat{\mathbf{A}}^s)\mathbf{Y}_n.$$

The last element of the last column of $\hat{\mathbf{A}}^s$ is one. All other entries of the last column are multiples of $\hat{\alpha}_0$, where the multiples are functions of $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p$ and the functions may be zero for small s. Let $(Y_1, Y_2, \ldots, Y_n) = \mathcal{S}$ be a realization, and let $(-Y_1, -Y_2, \ldots, -Y_n) = \overset{*}{\mathcal{S}}$. Then the value of $(\mathbf{A}^s - \hat{\mathbf{A}}^s)\mathbf{Y}_n$ for $\mathcal{S}$ is the negative of the value for $\overset{*}{\mathcal{S}}$. The result follows because $E\{e_t\} = 0$, and because the distribution of $(Y_1, Y_2, \ldots, Y_n)$ is symmetric. ▲

We now obtain an approximation for the variance of the prediction error, conditional on $(\mathbf{Y}_n, \mathbf{Y}_{n-1}, \ldots, \mathbf{Y}_{n-p+1})$.

Theorem 8.5.3. Let Y_t be a stationary normal time series satisfying the model (8.5.4) with the roots of (8.5.5) less than one in absolute value. Then the mean square error of the predictor $\hat{\mathbf{Y}}_{n+s}$ of (8.5.8), conditional on $\mathbf{Y}_n$, is

$$\begin{aligned} E\{(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})' \,|\, \mathbf{Y}_n\} \\ = \sigma^2 \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^j + n^{-1}\sigma^2 \sum_{j=0}^{s-1}\sum_{k=0}^{s-1} \bar{\mathbf{Y}}'_{n+s-j-1} \boldsymbol{\Gamma}^{-1} \bar{\mathbf{Y}}_{n+s-k-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^k + \mathbf{R}_n, \\ E\{\|\mathbf{R}_n\| \,|\, \mathbf{Y}_n\} = O_p(n^{-3/2}), \\ E\{\|\mathbf{R}_n\|\} = O(n^{-3/2}), \end{aligned} \tag{8.5.9}$$

where $\boldsymbol{\Gamma} = E\{\mathbf{Y}_t \mathbf{Y}_t'\}$, $\|\mathbf{R}_n\|^2 = \operatorname{tr} \mathbf{R}_n' \mathbf{R}_n$, $\bar{\mathbf{Y}}_{n+j} = \mathbf{A}^j \mathbf{Y}_n$ for $j = 0, 1, \ldots, s$, and $\mathbf{M}$ is a matrix with one as the upper left element and zeros elsewhere.

Proof. We have

$$\mathbf{Y}_{n+s} = \mathbf{A}^s \mathbf{Y}_n + \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{e}_{n+s-j},$$

where $\mathbf{A}^0 = \mathbf{I}$. Therefore, the error in the predictor is

$$\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s} = \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{e}_{n+s-j} - (\hat{\mathbf{A}}^s - \mathbf{A}^s)\mathbf{Y}_n .$$

It follows that

$$E\{(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})' \mid \mathbf{Y}_n\}$$
$$= \sigma^2 \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^j + E\{(\hat{\mathbf{A}}^s - \mathbf{A}^s)\mathbf{Y}_n \mathbf{Y}_n' (\hat{\mathbf{A}}^s - \mathbf{A}^s)' \mid \mathbf{Y}_n\} .$$

Because $\hat{A} = A + O_p(n^{-1/2})$, we may expand $\hat{\mathbf{A}}^s$ in a first order Taylor series about $\mathbf{A}$ to obtain

$$\hat{\mathbf{A}}^s = \mathbf{A}^s + \sum_{j=0}^{s-1} \mathbf{A}^j (\hat{\mathbf{A}} - \mathbf{A}) \mathbf{A}^{s-j-1} + O_p(n^{-1}) .$$

By the normality, $E\{|\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}|^8\} = O(n^{-4})$, and we obtain

$$E\{(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})' \mid \mathbf{Y}_n\} = \sigma^2 \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^j$$
$$+ E\left\{ \left[\sum_{j=0}^{s-1} \mathbf{A}^j (\hat{\mathbf{A}} - \mathbf{A}) \bar{\mathbf{Y}}_{n+s-j-1} \right] \left[\sum_{j=0}^{s-1} \mathbf{A}^j (\hat{\mathbf{A}} - \mathbf{A}) \bar{\mathbf{Y}}_{n+s-j-1} \right]' \,\middle|\, \mathbf{Y}_n \right\} + \mathbf{R}_{1n} , \tag{8.5.10}$$

where $E\{\|\mathbf{R}_{1n}\|\} = O(n^{-3/2})$ and $E\{\|\mathbf{R}_{1n}\| \mid \mathbf{Y}_n\} = O_p(n^{-3/2})$.

Note that for $r > n$, $q > n$ the upper left element of $(\hat{\mathbf{A}} - \mathbf{A})\bar{\mathbf{Y}}_r \bar{\mathbf{Y}}_q' (\hat{\mathbf{A}} - \mathbf{A})'$ is $(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})' \bar{\mathbf{Y}}_r \bar{\mathbf{Y}}_q' (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})$ and all other elements are zero. Therefore, because $\bar{\mathbf{Y}}_q$ and $\bar{\mathbf{Y}}_r$ are functions of $\mathbf{Y}_n$,

$$E\{\bar{\mathbf{Y}}_q' (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})' \bar{\mathbf{Y}}_r \mid \mathbf{Y}_n\} = n^{-1} \sigma^2 \bar{\mathbf{Y}}_q' \boldsymbol{\Gamma}^{-1} \bar{\mathbf{Y}}_r + \mathbf{R}_{2n} ,$$

where

$$\mathbf{R}_{2n} = \bar{\mathbf{Y}}_q' E\{[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})' - n^{-1}\sigma^2 \boldsymbol{\Gamma}^{-1}] \mid \mathbf{Y}_n\} \bar{\mathbf{Y}}_r .$$

Under our assumptions,

$$E\{[(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})' - n^{-1}\boldsymbol{\Gamma}^{-1}\sigma^2] \mid \mathbf{Y}_n\} = O_p(n^{-3/2})$$

and

$$E\{\|\mathbf{R}_{2n}\|\} = O(n^{-3/2}) . \qquad \blacktriangle$$

Note that the upper left element of $\mathbf{A}^j \mathbf{M} \mathbf{A}'^j$ is w_j^2, where the w_j satisfy the

difference equation (2.6.4). For the first order process ($p = 1$), the expression for the conditional mean square error of $\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s}$ reduces to

$$\sigma^2 \sum_{j=0}^{s-1} \alpha_1^{2j} + n^{-1}\sigma^2 \left[s^2 \alpha_1^{2s-2} + \left(\sum_{j=0}^{s-1} (-\alpha_1)^j \right)^2 \gamma_Y^{-1}(0)(Y_n - \mu)^2 \right] + R_n ,$$

where $\mu = E\{Y_t\}$ and $\gamma_Y(0) = E\{(Y_t - \mu)^2\}$. If the mean is known to be zero and not estimated, we obtain for $p = 1$

$$E\{(Y_{n+1} - \hat{Y}_{n+1})^2 \,|\, \mathbf{Y}_n\} = \sigma^2 + Y_n^2 \gamma_Y^{-1}(0) + R_n .$$

The unconditional mean square error of $\hat{\mathbf{Y}}_{n+s}$ is the expected value of (8.5.9).

Corollary 8.5.3. Let the assumptions of Theorem 8.5.3 hold. Then

$$\begin{aligned} &E\{(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})'\} \\ &\quad = \sigma^2 \sum_{j=0}^{s-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^j + n^{-1}\sigma^2 \sum_{j=0}^{s-1} \sum_{k=0}^{s-1} \mathbf{A}^j \mathbf{M} \mathbf{A}'^k \\ &\qquad \times \operatorname{tr}[(\mathbf{A}^{s-j-1}\boldsymbol{\Gamma})'(\boldsymbol{\Gamma}^{-1}\mathbf{A}^{s-k-1})] + O(n^{-3/2}) . \end{aligned} \tag{8.5.11}$$

Proof. Omitted ▲

The approximate mean square error of the predictor given in (8.5.9) can be expressed in scalar form as

$$E\{(Y_{n+s} - \hat{Y}_{n+s})^2 \,|\mathbf{Y}_n\} \doteq \sigma^2 \left\{ \sum_{j=0}^{s-1} w_j^2 + n^{-1} \sum_{j=0}^{s-1} \sum_{k=0}^{s-1} w_j w_k \bar{\mathbf{Y}}'_{n+s-j-1} \boldsymbol{\Gamma}^{-1} \bar{\mathbf{Y}}_{n+s-k-1} \right\}, \tag{8.5.12}$$

where the w_j are defined in Theorem 2.6.1. For $s = 1$, we have

$$E\{(Y_{n+1} - \hat{Y}_{n+1})^2 \,|\, \mathbf{Y}_n\} \doteq \sigma^2 + n^{-1}\sigma^2 \mathbf{Y}'_n \boldsymbol{\Gamma}^{-1} \mathbf{Y}_n$$

and

$$E\{(Y_{n+1} - \hat{Y}_{n+1})^2\} \doteq \sigma^2 [1 + n^{-1}(p+1)] , \tag{8.5.13}$$

an approximation that is often used.

Let

$$\hat{\sigma}^2 = (n - p - 1)^{-1} \sum_{t=1}^{n} (Y_t - \hat{\boldsymbol{\alpha}}' \mathbf{Y}_{t-1})^2$$

and $\hat{\boldsymbol{\Gamma}} = n^{-1} \Sigma_{t=1}^{n} \mathbf{Y}_{t-1} \mathbf{Y}'_{t-1}$. An estimator, $\hat{\mathbf{V}}(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s})$, of the mean square

error of prediction may be obtained from (8.5.9) by ignoring the remainder term and replacing $\mathbf{A}$, $\boldsymbol{\Gamma}$, and σ^2 by their estimators $\hat{\mathbf{A}}$, $\hat{\boldsymbol{\Gamma}}$, and $\hat{\sigma}^2$, respectively. Let

$$\hat{\mathbf{V}}(\mathbf{Y}_{n+s} - \hat{\mathbf{Y}}_{n+s}) = \hat{\sigma}^2 \sum_{j=0}^{s-1} \hat{\mathbf{A}}^j \mathbf{M} \hat{\mathbf{A}}'^j + n^{-1}\hat{\sigma}^2 \sum_{j=0}^{s-1}\sum_{k=0}^{s-1} [(\hat{\mathbf{A}}^{n+s-j-1}\mathbf{Y}_n)'\hat{\boldsymbol{\Gamma}}^{-1}\hat{\mathbf{A}}^{n+s-k-1}\mathbf{Y}_n]\hat{\mathbf{A}}^j \mathbf{M} \hat{\mathbf{A}}'^k . \tag{8.5.14}$$

For $s = 1$, equation (8.5.14) reduces to the estimated variance of the error in predicting $\mathbf{Y}_{n+1}$ obtained by treating the problem as a fixed-X regression prediction of $\mathbf{Y}_{n+1}$ given $\mathbf{X}_{n+1} = \mathbf{Y}_n$. Predictions more than one period ahead are nonlinear functions of the estimated parameters.

Example 8.5.1. A number of computer programs are available to compute predictions. Procedure ARIMA of SAS/ETS® was used to estimate models in Example 8.4.1. That program computes predictions that are essentially the estimated minimum mean square error predictor. The estimated prediction variance for the s-period prediction computed by the program is the formula that ignores the estimation error,

$$\hat{V}\{\hat{Y}_{n+s} - Y_{n+s}\} = \hat{\sigma}^2 \sum_{j=0}^{s-1} \hat{v}_j^2 , \tag{8.5.15}$$

where $\hat{v}_j$ are estimators of the coefficients of Theorem 2.7.1. Predictions for the next three periods using the estimated autoregressive moving average (2, 2) of Example 8.4.1 are -7.96, -5.17, and -1.14 for one, two, and three periods, respectively. The standard errors output by the program are 1.00, 2.42, and 3.67 for one, two, and three periods, respectively.

If we use Gaussian maximum likelihood to estimate the fourth order autoregressive process, we obtain the model

$$\begin{aligned} \hat{Y}_t - \hat{\mu} = {} & \underset{(0.098)}{2.188}\,(Y_{t-1} - \hat{\mu}) - \underset{(0.220)}{2.130}\,(Y_{t-2} - \hat{\mu}) \\ & + \underset{(0.220)}{0.977}\,(Y_{t-3} - \hat{\mu}) - \underset{(0.098)}{0.232}\,(Y_{t-4} - \hat{\mu}) , \end{aligned}$$

where $\hat{\mu} = -0.270$ (0.493). The predictions for the next three periods based on the fourth order autoregressive model are -8.11, -5.37, and -1.30 for one, two, and three periods, respectively. The standard errors that do not contain a component for estimation error, as output by the program, are 1.00, 2.40, and 3.58 for one, two, and three periods, respectively.

To illustrate the effect on the estimated standard errors of the component due to estimating parameters, we use the ordinary least squares estimates of the fourth

Table 8.5.1. Observations for Nonlinear Model to Construct Predictions

t	Dependent Variable	Intercept	X_{t1}	X_{t2}	X_{t3}	X_{t4}	X_{t5}	X_{t6}	X_{t7}
4	Y_4	1	Y_3	Y_2	Y_1	$\bar{y}$	0	0	0
5	Y_5	1	Y_4	Y_3	Y_2	Y_1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	Y_n	1	Y_{n-1}	Y_{n-2}	Y_{n-3}	Y_{n-4}	0	0	0
$n+1$	0	1	Y_n	Y_{n-1}	Y_{n-2}	Y_{n-3}	−1	0	0
$n+2$	0	1	0	Y_n	Y_{n-1}	Y_{n-2}	0	−1	0
$n+3$	0	1	0	0	Y_n	Y_{n-1}	0	0	−1

order autoregression. The estimated model is given in Example 8.4.1. To compute predictions, we use the method of indicator variables and a nonlinear regression program. To construct predictions for three periods, we write the regression model

$$Y_t = \theta_0 + \theta_1 X_{t1} + \theta_2 X_{t2} + \theta_3 X_{t3} + \theta_4 X_{t4} + \theta_5 X_{t5} + (\theta_6 - \theta_1\theta_5) X_{t6} + (\theta_7 - \theta_1\theta_6 - \theta_2\theta_5) X_{t7} + e_t ,$$

where the regression variables are defined in Table 8.5.1. The estimator of $(\theta_5, \theta_6, \theta_7)$ obtained by nonlinear least squares is the predictor of $(Y_{n+1}, Y_{n+2}, Y_{n+3})$. The estimated standard errors for $(\hat{\theta}_5, \hat{\theta}_6, \hat{\theta}_7)$ output by a nonlinear regression program are standard errors for the predictions containing a component for estimation error. The estimated predictions and the estimated standard errors for the nonlinear regression are given in the last column of Table 8.5.2.

The next to last column of Table 8.5.2 contains the standard errors computed with the formula (8.5.15) using the ordinary least squares coefficients. The differences between the standard errors of the last two columns are due to the

Table 8.5.2. Alternative Predictors for Data of Example 8.4.1

Prediction Period	Method and model			
	ML, ARMA(2,2) (8.5.15)	ML, AR(4) (8.5.15)	OLS, AR(4) (8.5.15)	OLS, AR(4) (Nonlinear)
$n+1$	−7.96	−8.11	−8.17	−8.17
	(1.00)	(1.00)	(1.00)	(1.01)
$n+2$	−5.17	−5.37	−5.51	−5.51
	(2.42)	(2.40)	(2.38)	(2.45)
$n+3$	−1.14	−1.30	−1.48	−1.48
	(3.67)	(3.58)	(3.55)	(3.68)

estimated effect of the estimation error. The estimation error makes a larger contribution to the estimated standard errors for the longer predictions.

Table 8.5.2 also contains predictions and standard errors for two models estimated by maximum likelihood. The maximum likelihood estimates of the autoregressive process differ slightly from the ordinary least squares estimates, and hence the predictions differ slightly. The autoregressive moving average predictions differ somewhat from the predictions based on the autoregressive model, but the differences are small relative to the standard errors. The differences among the predictions for the different models are larger than the differences among the estimated standard errors. ▲▲

8.6. NONLINEAR PROCESSES

In this section, we consider estimation for time series that fall outside the class of finite parameter autoregressive moving averages. As we have seen, estimation for autoregressive time series has many analogies to linear regression theory. Estimation for moving average time series is a particular nonlinear estimation problem, in that the model can be expressed as an autoregression in which the coefficients satisfy nonlinear restrictions. We now study extensions in which the conditional expectation of Y_t given $(Y_{t-1}, \ldots, Y_0)$ is a nonlinear function of $(Y_{t-1}, \ldots, Y_0)$.

Assume

$$Y_t = f(\mathbf{Z}_t; \boldsymbol{\theta}) + e_t, \qquad t = 1, 2, \ldots, \tag{8.6.1}$$

where $\mathbf{Z}_t = (Y_{t-1}, Y_{t-2}, \ldots, Y_0)$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)'$ is the parameter vector, and $\{e_t\}$ is a sequence of iid$(0, \sigma^2)$ random variables. Let $\hat{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ that minimizes

$$n^{-1} \sum_{t=1}^{n} [Y_t - f(\mathbf{Z}_t; \boldsymbol{\theta})]^2 . \tag{8.6.2}$$

Assume $f(\mathbf{Z}_t; \boldsymbol{\theta})$ has continuous first and second derivatives with respect to $\boldsymbol{\theta}$ for all t on a convex compact set S, $S \subset \Theta$, where the true value $\boldsymbol{\theta}^0$ is an interior point of S, and Θ is the parameter space. The model (8.6.1) is the same as the model (5.5.1) of Chapter 5. In Theorem 5.5.1, we gave conditions under which the nonlinear least squares estimator has a normal distribution in the limit. Basically, the derivatives must satisfy some convergence criteria.

In ordinary fitting problems, where Y_t is known to depend on a vector $\mathbf{Z}_t$, it is common practice to approximate an unknown functional relationship with a polynomial in $\mathbf{Z}_t$. The fitted function may be used directly as an approximation to the conditional expectation, or it may serve as an intermediate step in developing a nonlinear model with realistic properties. We demonstrate that similar approximations can be used in time series analysis.

Theorem 8.6.1. Let Y_t be a strictly stationary time series with finite $2(r+1)$ moments for some positive integer r. Let $\mathbf{X}_t$ be a vector of polynomials in lags of Y_t:

$$\mathbf{X}_t = (1, Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}, Y_{t-1}^2, Y_{t-1}Y_{t-2}, \ldots, Y_{t-p}^r)\,.$$

Assume

$$Y_t = f(\mathbf{Z}_t; \boldsymbol{\theta}) + e_t\,,$$

where $\{e_t\}$ is a sequence of iid$(0, \sigma^2)$ random variables, $\mathbf{Z}_t$ is the vector $(Y_{t-1}, Y_{t-2}, \ldots, Y_0)$, and $\boldsymbol{\theta}$ is a vector of parameters. Assume

$$n^{-1}\sum_{t=1}^{n}(Y_t, \mathbf{X}_t)'(Y_t, \mathbf{X}_t) = E\{(Y_t, \mathbf{X}_t)'(Y_t, \mathbf{X}_t)\} + O_p(n^{-1/2})\,, \tag{8.6.3}$$

where $E\{(Y_t, \mathbf{X}_t)'(Y_t, \mathbf{X}_t)\}$ is positive definite. Let

$$\boldsymbol{\xi} = [E\{\mathbf{X}_t'\mathbf{X}_t\}]^{-1}E\{\mathbf{X}_t' f(\mathbf{Z}_t; \boldsymbol{\theta})\} \tag{8.6.4}$$

and

$$\hat{\boldsymbol{\xi}} = \left[\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\right]^{-1}\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{Y}_t\,. \tag{8.6.5}$$

Then $\hat{\boldsymbol{\xi}} - \boldsymbol{\xi} = O_p(n^{-1/2})$ and

$$\operatorname*{plim}_{n\to\infty} n^{-1}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2 = E\{[f(\mathbf{Z}_t; \boldsymbol{\theta}) - \mathbf{X}_t\boldsymbol{\xi}]^2\} + \sigma^2\,,$$

where $\hat{Y}_t = \mathbf{X}_t\hat{\boldsymbol{\xi}}$.

Proof. We have

$$\hat{\boldsymbol{\xi}} - \boldsymbol{\xi} = \left[\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\right]^{-1}\sum_{t=1}^{n}\mathbf{X}_t'[f(\mathbf{Z}_t; \boldsymbol{\theta}) - \mathbf{X}_t\boldsymbol{\xi} + e_t]\,.$$

By construction, $E\{\mathbf{X}_t'[f(\mathbf{Z}_t; \boldsymbol{\theta}) - X_t\boldsymbol{\xi}]\} = 0$. Also,

$$n^{-1}\sum_{t=1}^{n}\mathbf{X}_t'[f(\mathbf{Z}_t; \boldsymbol{\theta}) - \mathbf{X}_t\boldsymbol{\xi}] = O_p(n^{-1/2})$$

by the assumption (8.6.3).

We demonstrate a stronger result in obtaining the order of $\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}$. By assumption, e_t is independent of Y_{t-h} for $h \geq 1$. Therefore, letting $\mathcal{A}_{t-1}$ denote the sigma-field generated by $(Y_{t-1}, Y_{t-2}, \ldots)$, $\mathbf{X}_t'e_t$ satisfies the conditions of Theorem 5.3.4 because

$$\left[E\left\{\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\right\}\right]^{-1}\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\xrightarrow{p}\mathbf{I}.$$

Therefore,

$$\left[\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\right]^{-1/2}\sum_{t=1}^{n}\mathbf{X}_t'e_t\xrightarrow{\mathscr{L}}N(\mathbf{0},\mathbf{I}\sigma^2)$$

and $\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}=O_p(n^{-1/2})$.

Because $\hat{\boldsymbol{\xi}}-\boldsymbol{\xi}=O_p(n^{-1/2})$,

$$\begin{aligned}n^{-1}\sum_{t=1}^{n}(Y_t-X_t\hat{\boldsymbol{\xi}})^2&=n^{-1}\sum_{t=1}^{n}[e_t+f(\mathbf{Z}_t;\boldsymbol{\theta})-\mathbf{X}_t\boldsymbol{\xi}-\mathbf{X}_t(\hat{\boldsymbol{\xi}}-\boldsymbol{\xi})]^2\\&=n^{-1}\sum_{t=1}^{n}[e_t+f(\mathbf{Z}_t;\boldsymbol{\theta})-\mathbf{X}_t\boldsymbol{\xi}]^2+o_p(n^{-1/2})\\&=\sigma^2+E\{[f(\mathbf{Z}_t;\boldsymbol{\theta})-\mathbf{X}_t\boldsymbol{\xi}]^2\}+O_p(n^{-1/2}),\end{aligned}$$

where we have used (8.6.3). ▲

It is a direct consequence of Theorem 8.6.1 that polynomial regressions can be used to test for nonlinearity in the conditional expectation.

Corollary 8.6.1. Let Y_t be the strictly stationary autoregressive process satisfying

$$Y_t=\alpha_0+\sum_{i=1}^{p}\theta_iY_{t-i}+e_t,$$

where $\{e_t\}$ is a sequence of iid$(0,\sigma^2)$ random variables. Assume Y_t satisfies the moment conditions of Theorem 8.6.1, including the assumption (8.6.3). Let $\hat{\boldsymbol{\xi}}$ be defined by (8.6.5), where $\hat{\boldsymbol{\xi}}=(\hat{\boldsymbol{\xi}}_1,\hat{\boldsymbol{\xi}}_2)$ and $\hat{\boldsymbol{\xi}}_1$ is the vector of coefficients of $(1,Y_{t-1},\ldots,Y_{t-p})$. Let

$$\hat{\sigma}^2=(n-k)^{-1}\sum_{t=1}^{n}(Y_t-\hat{Y}_t)^2,$$

where k is the dimension of $\mathbf{X}_t$ and $\hat{Y}_t=\mathbf{X}_t\hat{\boldsymbol{\xi}}$. Then

$$\hat{\sigma}^{-1}\hat{\mathbf{A}}_{22}^{-1/2}\hat{\boldsymbol{\xi}}_2\xrightarrow{\mathscr{L}}N(\mathbf{0},\mathbf{I}),$$

where

$$\left[\sum_{t=1}^{n}\mathbf{X}_t'\mathbf{X}_t\right]^{-1}=\hat{\mathbf{A}}=\begin{bmatrix}\hat{\mathbf{A}}_{11}&\hat{\mathbf{A}}_{12}\\\hat{\mathbf{A}}_{21}&\hat{\mathbf{A}}_{22}\end{bmatrix}$$

and the partition of $\hat{\mathbf{A}}$ conforms to the partition of $\hat{\boldsymbol{\xi}}$.

Proof. Under the assumptions,

$$\boldsymbol{\xi}' = (\alpha_0, \theta_1, \ldots, \theta_p, 0, 0, \ldots, 0)$$

and $f(\mathbf{Z}_t; \boldsymbol{\theta}) - \mathbf{X}_t\boldsymbol{\xi} = 0$. The limiting normal distribution follows from the proof of Theorem 8.6.1. ▲

It follows from Corollary 8.6.1 that the null distribution of the test statistic

$$F = \hat{\sigma}^{-2} k_2^{-1} \hat{\boldsymbol{\xi}}_2' \hat{\mathbf{A}}_{22}^{-1} \hat{\boldsymbol{\xi}}_2 ,$$

where k_2 is the dimension of $\hat{\boldsymbol{\xi}}_2$ and k is the dimension of $\hat{\boldsymbol{\xi}}$, is approximately that of Snedecor's F with k_2 and $n - k$ degrees of freedom. The distribution when $f(\mathbf{Z}_t; \boldsymbol{\theta})$ is a nonlinear function depends on the degree to which $f(\mathbf{Z}_t; \boldsymbol{\theta})$ is well approximated by a polynomial. If the approximation is good, the test statistic will have good power. Other tests for nonlinearity are discussed by Tong (1990, p. 221), Hinich (1982), and Tsay (1986).

There are several nonlinear models that have received special attention in the time series literature. One is the *threshold autoregressive model* that has been studied extensively by Tong (1983, 1990). A simple first order threshold model is

$$Y_t = \theta_1 Y_{t-1} + \theta_2 \delta(Y_{t-1}, A) Y_{t-1} + e_t , \tag{8.6.6}$$

where

$$\delta(Y_{t-1}, A) = \begin{cases} 1 & \text{if } Y_{t-1} \geq A , \\ 0 & \text{otherwise} . \end{cases} \tag{8.6.7}$$

The indicator function, $\delta(Y_{t-1}, A)$, divides the behavior of the process into two regimes. If $Y_{t-1} \geq A$, the conditional expected value of Y_t given Y_{t-1} is $(\theta_1 + \theta_2)Y_{t-1}$. If $Y_{t-1} < A$, the conditional expected value of Y_t given Y_{t-1} is $\theta_1 Y_{t-1}$. Notice that the conditional expected value of Y_t is not a continuous function of Y_{t-1} for $\theta_2 \neq 0$ and $A \neq 0$.

Threshold models with more than two regimes and models with more than one lag of Y entering the equation are easily constructed. Also, the indicator function $\delta(\cdot)$ can be a function of different lags of Y and (or) a function of other variables. Practitioners are often interested in whether or not a coefficient has changed at some point in time. In such a case, the indicator function can be a function of time.

Most threshold models do not satisfy the assumptions of our theorems when the parameters specifying the regimes are unknown. The conditional expected value of Y_t given Y_{t-1} defined by threshold model (8.6.6) is not a continuous function of Y_{t-1}, and the conditional expected value is not continuous in A. Models that are continuous and differentiable in the parameters can be obtained by replacing the function $\delta(\cdot)$ defining the regimes with a continuous differentiable function.

Candidate functions are continuous differentiable statistical cumulative distribution functions. An example of such a model is

$$Y_t = \theta_1 Y_{t-1} + \theta_2 \overset{*}{\delta}(Y_{t-2}, \kappa) Y_{t-1} + e_t , \tag{8.6.8}$$

$$\overset{*}{\delta}(x, \kappa) = [1 + \exp\{\kappa_1 (x - \kappa_2)\}]^{-1} , \tag{8.6.9}$$

where $\overset{*}{\delta}(x, \kappa)$ is the logistic function. The parameter κ_1 can be fixed or can be a parameter to be estimated. The conditional mean function of (8.6.8) is continuous and differentiable in $(\theta_1, \theta_2, \kappa_1, \kappa_2)$ for κ_1 in $(0, \infty)$. Tong (1990, p. 107) calls models with $\delta(\cdot)$ a smooth function, such as (8.6.9), *smoothed threshold autoregressive models*. See Jones (1978) and Ozaki (1980).

Example 8.6.1. One of the most analyzed realizations of a time series is the series on the number of lynx trapped in the Mackenzie River district of Canada based on the records of the Hudson Bay Company as compiled by Elton and Nicholson (1942). The data are annual records for the period 1821–1934. The biological interest in the time series arises from the fact that lynx are predators heavily dependent on the snowshoe hare. The first statistical model for the data is that of Moran (1953). Tong (1990, p. 360) contains a description of other analyses and a detailed investigation of the series. We present a few computations heavily influenced by Tong (1990). The observation of analysis is $\log_{10}$ of the original observations. The sample mean of the time series is 2.904, the smallest value of $Y_t - \bar{y}$ is -1.313, and the largest value of $Y_t - \bar{y}$ is 0.941. The second order autoregressive model estimated by ordinary least squares is

$$\underset{}{\hat{Y}_t} = \underset{(0.122)}{1.057} + \underset{(0.064)}{1.384}\, Y_{t-1} - \underset{(0.064)}{0.747}\, Y_{t-2} , \tag{8.6.10}$$

where $\hat{\sigma}^2 = 0.053$, and the numbers in parentheses are the ordinary least squares standard errors.

Let us assume that we are interested in a model based on Y_{t-1} and Y_{t-2} and are willing to consider a nonlinear model. For the moment, we ignore information from previously estimated nonlinear models. If we estimate a quadratic model as a first approximation, we obtain

$$\hat{y}_t = \underset{(0.034)}{0.079} + \underset{(0.073)}{1.366}\, y_{t-1} - \underset{(0.068)}{0.785}\, y_{t-2} + \underset{(0.159)}{0.063}\, y_{t-1}^2$$

$$+ \underset{(0.284)}{0.172}\, y_{t-1} y_{t-2} - \underset{(0.172)}{0.450}\, y_{t-2}^2 ,$$

where $y_t = Y_t - \bar{y} = Y_t - 2.904$ and $\hat{\sigma}^2 = 0.0442$. Also see Cox (1977). The ordinary regression F-test for the hypothesis that the three coefficients of the quadratic terms are zero is

$$F(3, 106) = (0.0442)^{-1} 0.3661 = 8.28 ,$$

and the hypothesis of zero values is strongly rejected. The coefficient on y_{t-1}^2 is small, and one might consider the estimated model

$$\hat{y}_t = \underset{(0.027)}{0.089} + \underset{(0.059)}{1.350}\, y_{t-1} - \underset{(0.059)}{0.772}\, y_{t-2} + \underset{(0.129)}{0.272}\, y_{t-1}y_{t-2} - \underset{(0.123)}{0.497}\, y_{t-2}^2, \tag{8.6.11}$$

where $\hat{\sigma}^2 = 0.0438$.

The estimated conditional expectation of Y_t, given (Y_{t-1}, Y_{t-2}), is changed very little from that in (8.6.11) if we replace y_{t-2} in the last two terms of (8.6.11) by a bounded function of y_{t-2} that is nearly linear in the interval $(-1.1, 1.1)$. Such a function is

$$g(y_{t-2}; \kappa_1, \kappa_2) = [1 + \exp\{\kappa_1(y_{t-2} - \kappa_2)\}]^{-1}. \tag{8.6.12}$$

If κ_1 is very small in absolute value, the function is approximately linear. As $\kappa_1(y_{t-2} - \kappa_2)$ moves from -2 to 2, $g(y_{t-2}; \kappa_1, \kappa_2)$ moves from 0.88 to 0.12, and as $\kappa_1(y_{t-2} - \kappa_2)$ moves from -3 to 3, $g(y_{t-2}; \kappa_1, \kappa_2)$ moves from 0.9526 to 0.0474. If κ_1 is very large, the function is essentially a step function. Also, as κ_1 increases, the derivative with respect to κ_1 approaches the zero function except for values very close to κ_2. The range of $y_t = Y_t - \bar{y}$ is from -1.313 to 0.941. Thus, $g(y_{t-2}; -2.5, 0)$ is nearly linear for the range of the data.

The estimated equation obtained by replacing y_{t-2} with $g(y_{t-2}; -2.5, 0)$ in the last two terms of (8.6.11) is

$$\begin{aligned}\hat{y}_t = {} & \underset{(0.029)}{0.107} + \underset{(0.144)}{1.092}\, y_{t-1} - \underset{(0.140)}{0.205}\, y_{t-2} + \underset{(0.251)}{0.512}\, g(y_{t-2}; -2.5, 0)y_{t-1} \\ & - \underset{(0.250)}{1.110}\, g(y_{t-2}; -2.5, 0)y_{t-2}\end{aligned} \tag{8.6.13}$$

with $\hat{\sigma}^2 = 0.0432$. We note that $g(y_{t-2}; \kappa_1, \kappa_2)$ converges to the indicator function with jump of height one at the point κ_2 as $\kappa_1 \to -\infty$. Therefore, the threshold model is the limit as $(\kappa_1, \kappa_2) \to (-\infty, \overset{*}{\kappa}_2)$, where $\overset{*}{\kappa}_2$ is the point dividing the space into two regimes.

We fit the nonlinear model obtained by letting κ_1 and κ_2 of (8.6.12) be parameters to be estimated. In the estimation we restricted κ_1 to be the interval $[-15, -0.5]$, and the minimum sum of squares occurred on the boundary $\kappa_1 = -15$. The estimated function is

$$\begin{aligned}\hat{y}_t = {} & \underset{(0.033)}{0.119} + \underset{(0.073)}{1.241}\, y_{t-1} - \underset{(0.092)}{0.395}\, y_{t-2} + \underset{(0.143)}{0.371}\, g(y_{t-2}; -15, 0.357)y_{t-1} \\ & - \underset{(0.162)}{0.870}\, g(y_{t-2}; -15, 0.357)y_{t-2},\end{aligned} \tag{8.6.14}$$

where $\hat{\sigma}^2 = 0.0420$ and the standard error of $\hat{\kappa}_2$ is 0.078. The standard errors are computed treating κ_1 as known. If the standard errors are computed treating κ_1 as

estimated, the standard error of $\hat{\kappa}_1$ is larger than the estimated value. This reflects the fact that the derivative with respect to κ_1 approaches zero as $\kappa_1 \to -\infty$.

The residual sum of squares for the model (8.6.14) is 4.448, while that for (8.6.13) is 4.620. The reduction due to fitting κ_1 and κ_2 is 0.172, and the F for a test against $(-2.5, 0)$ is 2.05, which is less than the 5% tabular value of 3.08. If we assumed that we were only estimating κ_2, then the improvement in the fit would be significant. The κ_1 of -15 means that about 76% of the shift occurs in an interval of length 0.27 centered at the estimated value of 0.357. The estimated variance of the original time series is 0.314. Therefore, the interval in which most of the estimated shift takes place is about one-half of one standard deviation of the original time series.

A threshold model fitted using $\hat{\kappa}_2$ to define the regimes is

$$\hat{y}_t = \begin{cases} \underset{(0.030)}{0.102} + \underset{(0.071)}{1.278}\, y_{t-1} - \underset{(0.086)}{0.456}\, y_{t-2} & \text{if } y_{t-2} < 0.36, \\ \underset{(0.152)}{0.166} + \underset{(0.108)}{1.527}\, y_{t-1} - \underset{(0.264)}{1.239}\, y_{t-2} & \text{if } y_{t-2} \geq 0.36, \end{cases}$$

where $\hat{\sigma}^2 = 0.0445$. Tong (1990) has given threshold models with smaller residual mean square.

It is reasonable to conclude that the process generating the data is nonlinear, but a number of nonlinear models are consistent with the data. Selection of a final model would rest heavily on subject matter considerations. ▲▲

A second nonlinear model that has been studied extensively is the bilinear model. See, for example, Granger and Andersen (1978), Subba Rao and Gabr (1984), and Tong (1990). The first order model can be written

$$Y_t = \alpha_1 Y_{t-1} + \beta_1 Y_{t-1} e_{t-1} + e_t, \tag{8.6.15}$$

where $e_t \sim \text{NI}(0, \sigma^2)$. Sufficient conditions for (8.6.15) to be a stationary process are $\alpha_1^2 + \beta_1^2 \sigma^2 < 1$ or $E\{|\alpha_1 + \beta_1 \epsilon_t|\} < 1$. Both conditions are intuitive in that they are analogous to a requirement that, on average, the coefficient of Y_{t-1} be less than one in absolute value. See Tong (1981) and Quinn (1982).

A third class of nonlinear models is the random coefficient autoregressive models. An example of such a model is

$$Y_t = \sum_{i=1}^{p} (\theta_i + a_{it}) Y_{t-i} + e_t,$$

where

$$\begin{bmatrix} \mathbf{a}_t \\ e_t \end{bmatrix} \sim \text{II}\left[\begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \end{pmatrix} \right],$$

$\mathbf{a}_t' = (a_{1t}, a_{2t}, \ldots, a_{pt})$, and e_t is independent of $\mathbf{a}_t$. Nichols and Quinn (1982) is a treatment of such models.

8.7. MISSING AND OUTLIER OBSERVATIONS

Estimation with missing observations requires a specification of the mechanism that is responsible for the observations being missing. The most common specification is that observations are *missing at random*. See Rubin (1976) and Little and Rubin (1987). Under the missing-at-random specification, the likelihood can be written as

$$L(\mathbf{Y}_B; \boldsymbol{\theta}) = \int L_C(\mathbf{Y}; \boldsymbol{\theta})\, dF(\mathbf{Y}_m)\,, \tag{8.7.1}$$

where $L_C(\mathbf{Y}; \boldsymbol{\theta})$ is the likelihood for a complete data set, and $F(\mathbf{Y}_m)$ is the distribution function of the missing observations,

$$\mathbf{Y} = (Y_{[1]}, Y_{[2]}, \ldots, Y_{[n-m]}, Y_{[n-m+1]}, \ldots, Y_{[n]})' = (\mathbf{Y}_B, \mathbf{Y}_m)'\,,$$

$\mathbf{Y}_B$ is the vector of Y_t actually observed, and $\mathbf{Y}_m$ is the vector of missing observations. The subscripts of $\mathbf{Y}$ are placed in square brackets because they do not necessarily correspond to the index t.

A simple missing mechanism specifies some probability $p_m = \eta$ that an observation is missing and that this probability is equal for all observations. A more general model would permit the probability to depend on the index t, but in order to estimate $\boldsymbol{\theta}$ by maximizing only $L(\mathbf{Y}_B; \boldsymbol{\theta})$, the probability that the observation is missing must not be a function of $\mathbf{Y}$.

Assume that Y_t is a normal stationary autoregressive moving average of order (p, q) and that m observations in a sample of size n are missing at random. Then

$$L(\mathbf{Y}_B, \boldsymbol{\theta}) = (2\pi\sigma^2)^{-0.5(n-m)} |\boldsymbol{\Sigma}_{BB}|^{-0.5} \exp\{-0.5\mathbf{Y}_B' \boldsymbol{\Sigma}_{BB}^{-1} \mathbf{Y}_B\}\,, \tag{8.7.2}$$

where $\mathbf{Y}_B$ is the $(n-m)$-dimensional vector of observations and $\boldsymbol{\Sigma}_{BB}$ is the covariance matrix of $\mathbf{Y}_B$. While it is relatively easy to write down the likelihood, there has been considerable research on computational methods.

Jones (1980) suggested using the Kalman filter representation of Section 4.6 to construct maximum likelihood estimators for autoregressive moving averages. Let

$$\mathbf{X}_{t|r} = (E\{Y_t \mid r\}, E\{Y_{t+1} \mid r\}, \ldots, E\{Y_{t+s-1} \mid r\})\,,$$

where $r \leq t$, $s = \max(p, q+1)$, and $E\{Y_{t+j} \mid r\} = E\{Y_{t+j} \mid (Y_1, Y_2, \ldots, Y_r)\}$. Let $X_{t,j|r}$ denote the jth element of $\mathbf{X}_{t|r}$, and abbreviate $\mathbf{X}_{t|t}$ to $\mathbf{X}_t$. The Kalman recursion is initiated at $t = 1$ with $\mathbf{X}_0$ of (4.6.43) equal to $\mathbf{0}$, and $\boldsymbol{\Sigma}_{vv00}$ equal to the covariance matrix of $\mathbf{X}_t$. Let the sample be complete. Then one can write the log-likelihood as

$$l(\mathbf{Y}; \boldsymbol{\theta}) = -0.5\left[n \log(2\pi\sigma^2) + \sum_{t=1}^{n} \log v_{tt} + \sum_{t=1}^{n} v_{tt}^{-1} (Y_t - \hat{X}_{t,1|t-1})^2 \right], \tag{8.7.3}$$

where $\hat{X}_{t,1|t-1}$ is the best linear predictor of Y_t given $(Y_{t-1}, \ldots, Y_1)$, and

$$v_{tt} = V\{Y_t - \hat{X}_{t,1|t-1}\}. \tag{8.7.4}$$

We have suppressed the dependence of v_{tt} and $\hat{X}_{t,1|t-1}$ on the parameters to simplify the notation. In Example 4.6.4, we illustrated how the usual Kalman recursion formulas can be used in the presence of missing data. Thus, the maximum likelihood estimators can be obtained by maximizing (8.7.3) with respect to the unknown parameters, using the formulas of Section 4.6, where the sum in (8.7.3) is over the elements of $\mathbf{Y}_B$.

If there are a modest number of missing values, the method of indicator variables can be used in conjunction with an ordinary maximum likelihood program to construct the maximum likelihood estimators. We illustrate the procedure in Example 8.7.1.

Example 8.7.1. The data of Table 8.7.1 are computer generated observations from a second order autoregressive process. We assume that observations 40, 57, and 58 are missing at random. One method of computing the maximum likelihood estimates under this assumption is the method of indicator variables illustrated in the construction of predictions in Example 8.5.1.

Table 8.7.1. Data for Example 8.7.1

1.467	2.722	−1.198	−0.107	4.156
7.181	0.201	−1.803	−4.723	4.224
8.939	−3.239	0.096	−7.357	2.605
5.810	−5.287	−0.113	−4.938	−0.561
−0.267	−5.491	−0.023	1.647	−1.842
−4.053	−3.210	−0.746	6.374	−1.580
−5.427	0.682	−1.072	5.993	0.266
−3.624	4.971	−1.235	3.705	1.526
−0.704	6.265	−0.943	−0.784	3.401
1.256	3.347	1.218	−3.036	3.450
2.162	−0.860	3.173	−3.516	1.361
3.232	−3.697	2.158	−1.510	−2.423
1.076	−4.030	1.859	0.294	−5.100
−2.572	−1.710	2.676	1.637	−5.433
−4.135	−0.217	3.011	0.633	−3.377
−3.836	−1.254	1.540	−1.447	−0.070
−1.185	−0.618	—	−4.060	2.924
0.787	0.087	—	−5.559	3.432
1.446	0.553	5.081	−3.814	3.209
2.425	—	3.900	2.067	1.745

To implement the procedure, we define

$$X_{t1} = \begin{cases} -1 & \text{if } t = 40, \\ 0 & \text{otherwise}, \end{cases}$$

$$X_{t2} = \begin{cases} -1 & \text{if } t = 57, \\ 0 & \text{otherwise}, \end{cases}$$

$$X_{t3} = \begin{cases} -1 & \text{if } t = 58, \\ 0 & \text{otherwise}. \end{cases}$$

Then the model

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t, \\ u_t &+ \alpha_1 u_{t-1} + \alpha_2 u_{t-2} = e_t, \end{aligned} \tag{8.7.5}$$

where $e_t \sim \text{NI}(0, \sigma^2)$, is estimated by maximum likelihood. Using the maximum likelihood option of the AUTOREG procedure of SAS®, we obtain

$$\begin{aligned} &(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2) \\ &\quad = (-0.001, \ -0.040, \ 1.442, \ 3.284, \ -1.377, \ -0.899, \ 1.183), \\ &\qquad (0.207) \quad (0.554) \ (0.787) \ (0.786) \quad (0.045) \quad (0.042) \ (0.173) \end{aligned}$$

where $\hat{\sigma}^2$ is the estimator adjusted for degrees of freedom. Observe that $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ is the best estimator of (Y_{40}, Y_{57}, Y_{58}), given the other observations and that the parameters are equal to $(\hat{\alpha}_1, \hat{\alpha}_2)$. The standard errors for Y_{57} and Y_{58} are larger than the standard error of the estimator of Y_{40} because Y_{57} and Y_{58} are adjacent missing observations. ▲▲

In any statistical analysis, it is wise practice to inspect the data at every stage of the analysis for extreme or unusual observations. Such observations are often called outlier observations. In time series analysis, the first step in the analysis is generally an inspection of plots of the original data. Such plots help one define the nature of the data. The plot of the data against time is often sufficient to identify very extreme observations, particularly extreme observations created by errors in recording data.

As with many important concepts, what defines an outlier is difficult to specify. This is because what is unusual requires a complete specification of the statistical model, a specification beyond that common in practice. Thus, in a sample of 100 normal (0, 1) random variables, a value of $s_{(n-1)}^{-1}(X_{(n)} - \bar{x}_{(n-1)}) = 5.00$, where $X_{(n)}$ is the largest observation, $s_{(n-1)}$ is the root mean square for the 99 smallest observations, and $\bar{x}_{(n-1)}$ is the mean of the 99 smallest observations, is unusual. In a sample of 100 from a Cauchy distribution, the same value is less unusual.

Fox (1972) defined two types of unusual observations in time series. The first type is called an *additive outlier*. As an example model for additive outliers, assume that we observe the process

$$Y_t = X_t + \delta_{tm} \zeta_m, \tag{8.7.6}$$

where X_t is a stationary autoregressive moving average, δ_{tm} is zero for $t \neq m$ and one for $t = m$, and ζ_m is the value of the perturbation at $t = m$. Under such a model, one might attempt to estimate ζ_m and to remove the effect of the ζ_m on the estimates of the autoregressive parameters. One can expand the specification by permitting more than one outlier.

The second type of outlier is called an *innovation outlier*. To illustrate, assume that Y_t is a stationary autoregressive process satisfying

$$\sum_{i=0}^{p} \alpha_i Y_{t-i} = e_t, \tag{8.7.7}$$

where $e_t = a_t + \delta_{tm}\zeta_m$, and (δ_{tm}, ζ_m) has the properties described following (8.7.6). Then ζ_m is called an innovation outlier.

If we assume that we know the point m at which an outlier may have occurred, treat the unknown ζ_m as a parameter to be estimated, and assume Y_t is a normal process, then one can use likelihood methods to estimate ζ_m and to test the hypothesis that $\zeta_m = 0$. If the point at which an outlier may have occurred is unknown, then a reasonable procedure is to search over the possible values. To construct the likelihood ratio for every value of t, or for every possible pair of values, is a large computational task. Therefore, it is common practice to fit a model to the original data and then inspect the residuals to see if any are unusual. The inspection is often done on the basis of plots. In an autoregressive process, an innovation outlier will generally appear as a single residual of large absolute value. This is because only the single observation fails to satisfy the autoregressive model. On the other hand, an additive outlier will affect the residual for the point at which it occurs and will also affect the p following residuals because they also fail to satisfy the basic autoregressive model. Statistics for testing for outliers were suggested by Fox (1972), and extensions have been considered by several authors, including Chang, Tiao, and Chen (1988) and Tsay (1988).

Assume we have an autoregressive process of order p with known parameters. The predicted value for Y_m, given that Y_m is not observed, is a function of $\mathbf{y}_m = (Y_{m-p}, \ldots, Y_{m-1}, Y_{m+1}, \ldots, Y_{m+p})'$. Therefore, one can construct the linear filter, say $\mathbf{H}$, which when applied to the $\mathbf{y}_m$ gives an estimator of Y_m. The difference between Y_m and $\mathbf{H}\mathbf{y}_m'$ divided by the estimated standard deviation of $Y_m - \mathbf{H}\mathbf{y}_m'$ provides evidence on whether or not Y_m is an additive outlier. For a first order process with parameter α_1,

$$\hat{\zeta}_m = Y_m - \hat{Y}_m(Y_{m-1}, Y_{m+1}) = Y_m + \alpha_1(1 + \alpha_1^2)^{-1}(Y_{m-1} + Y_{m+1}) \tag{8.7.8}$$

and the variance of the difference is $\sigma^2(1 + \alpha_1^2)^{-1}$. In practice, it is necessary to replace the unknown parameters with estimates.

If the parameters of the autoregressive process have been estimated and the residuals calculated, then the residuals at the points $m, m+1, \ldots, m+p$ are affected by the presence of an additive outlier. Thus, an estimate of ζ_m, calculated from the residuals, is

$$\tilde{\zeta}_m = \left[\sum_{i=0}^{p} \hat{\alpha}_i^2\right]^{-1} \sum_{i=0}^{p} \hat{\alpha}_i \hat{e}_{m+i}, \tag{8.7.9}$$

where $\hat{\alpha}_0 = 1$ and $\hat{e}_{m+i} = \sum_{j=0}^{p} \hat{\alpha}_j Y_{m+i-j}$, $i = 0, 1, \ldots, p$. A test statistic for the hypothesis that $\zeta_m = 0$ is

$$t_m = \hat{\sigma}^{-1}\left[\sum_{i=0}^{p} \hat{\alpha}_i^2\right]^{1/2} \tilde{\zeta}_m, \tag{8.7.10}$$

where $\hat{\sigma}^2$ is an estimator of σ^2, such as the regression residual mean square. If the process is normal, m is not determined by the data, and $\zeta_m = 0$, then t_m is, approximately, a $N(0, 1)$ random variable.

An innovation outlier at time m affects only the residual at time m. Thus, an estimate of ζ_m is the residuals at the point m. Fox (1972) and Chang, Tiao, and Chen (1988) have discussed choosing between the two types of outliers. A simple procedure is to choose the type of outlier that gives the largest absolute value of the estimate of ζ_m.

In Example 8.7.2, we use model fitting procedures similar to those used in Example 8.7.1 to identify unusual observations, to classify the unusual observations, and to construct estimates of the autoregressive parameters.

Example 8.7.2. Table 8.7.2 contains 100 computer generated observations from a second order autoregression. The basic model is

Table 8.7.2. Data for Example 8.7.2

25.267	23.722	19.802	20.893	26.266
27.081	21.201	19.197	16.277	27.653
26.819	17.761	21.096	13.643	25.962
25.110	15.713	20.887	16.062	21.455
20.733	15.509	20.977	25.712	22.665
16.947	17.790	20.254	31.666	17.355
15.573	21.682	19.928	30.120	19.120
17.376	25.971	19.765	25.048	21.462
20.296	27.265	20.057	17.757	24.929
22.256	24.347	22.218	14.199	26.190
23.162	20.140	24.173	14.525	24.300
24.232	17.303	23.158	18.885	19.655
22.076	16.970	22.859	23.230	15.589
18.428	19.290	23.676	25.915	14.117
16.865	20.783	24.011	24.403	15.886
17.164	19.746	22.540	20.349	19.861
19.815	20.382	22.818	15.452	24.060
21.787	21.087	24.914	12.609	25.628
22.446	21.553	26.081	14.620	25.755
23.425	22.063	24.900	22.137	23.786

$$Y_t - \mu + \alpha_1(Y_{t-1} - \mu) + \alpha_2(Y_{t-2} - \mu) = e_t, \tag{8.7.11}$$

where $e_t \sim \text{NI}(0, \sigma^2)$. The parameters of the second order autoregressive process estimated by maximum likelihood are

$$\begin{array}{llll} (\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2) = (21.150, & -1.308, & 0.840, & 2.20)\,. \\ \quad (0.276) & (0.053) & (0.053) & (0.32) \end{array} \tag{8.7.12}$$

Figure 8.7.1 is a plot of the standardized residuals from the second order autoregression, plotted against time. The standardized residuals are

$$\tilde{e}_1 = (Y_1 - \hat{\mu})\hat{\gamma}_Y^{-1/2}(0),$$

$$\tilde{e}_2 = [(Y_2 - \hat{\mu}) - \hat{\rho}(1)(Y_1 - \hat{\mu})]\{\hat{\gamma}_Y(0)[1 - \hat{\rho}^2(1)]\}^{-1/2},$$

$$\tilde{e}_t = [Y_t - \hat{\mu} + \hat{\alpha}_1(Y_{t-1} - \hat{\mu}) + \hat{\alpha}_2(Y_{t-2} - \hat{\mu})]\hat{\sigma}^{-1}, \qquad t = 2, 3, \ldots, n,$$

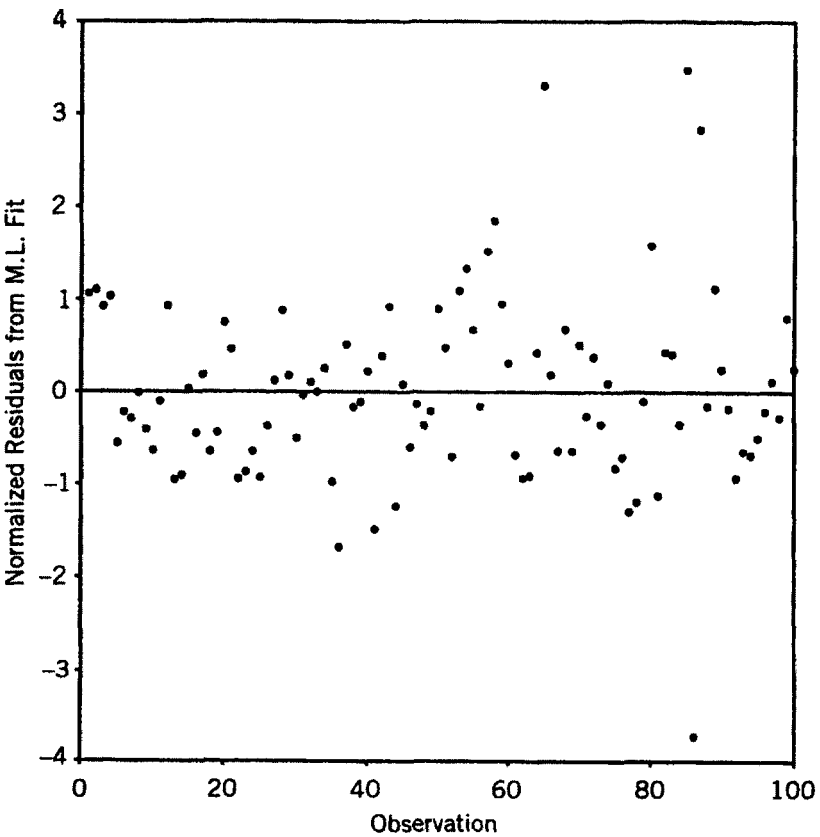

FIGURE 8.7.1. Residuals from second order autoregressive fit.

where $\hat{\sigma}^2 = 2.20$ is the estimated variance of e_t. There is a large deviation at $t = 65$ and three large deviations for $t = 85$, 86, and 87. In practice, one would attempt to determine if an error has been made in data recording or if there is a subject matter basis for the unusual observations. We proceed under the assumption that no explanation for the large deviations is available.

The plot suggests that there is an innovation outlier at $t = 65$, because there is a large deviation only at that point, the deviations at $t = 66$ and $t = 67$ appearing rather ordinary. The deviations associated with $t = 85$, 86, and 87 indicate that there may be an additive outlier at $t = 85$. There is a large positive deviation at $t = 85$, followed by a negative deviation at $t = 86$, followed by a positive deviation at $t = 87$, and these signs match the signs of the autoregressive coefficients. The value of t_m of (8.7.10) for $m = 85$ is 5.81, while the standardized residual is $\hat{\sigma}^{-1}\hat{e}_{85} = 3.48$. Thus, the data support the presence of an additive outlier over the presence of an innovation outlier at $t = 85$.

To estimate the parameters of the model treating Y_{65} as an innovation outlier and Y_{85} as an additive outlier, we write a regression model

$$\begin{aligned} Y_t &= \mathbf{X}_t\boldsymbol{\beta} + u_t, \\ u_t + \alpha_1 u_{t-1} + \alpha_2 u_{t-2} &= e_t, \end{aligned} \tag{8.7.13}$$

where u_t is the zero mean second order autoregressive process. In this representation, an additive outlier can be introduced by creating an indicator variable for the point. For our example, we let $X_{t1} \equiv 1$ for the parameter μ and let

$$\delta_{t,85} = X_{t2} = \begin{cases} 1 & \text{if } t = 85, \\ 0 & \text{otherwise}. \end{cases} \tag{8.7.14}$$

The process of maximizing the likelihood for the model (8.7.13) with $\mathbf{X}_t = (X_{t1}, X_{t2})$ can be visualized as transforming the data on the basis of the autoregressive model and then finding the $\boldsymbol{\beta}$ that minimizes the regression residual mean square. Thus, after transformation,

$$\begin{aligned} Y_{85} &= \mu + \zeta_{85} - \alpha_1(Y_{84} - \mu) - \alpha_2(Y_{83} - \mu) + e_{85}, \\ Y_{86} &= \mu - \alpha_1(Y_{85} - \mu - \zeta_{85}) - \alpha_2(Y_{84} - \mu) + e_{86}, \\ Y_{87} &= \mu - \alpha_1(Y_{86} - \mu) - \alpha_2(Y_{85} - \mu - \zeta_{85}) + e_{87}, \end{aligned}$$

where $(\beta_1, \beta_2) = (\mu, \zeta_{85})$.

A different X-variable is required for an innovation outlier in regression model (8.7.13). We require a variable which, after the autoregressive transformation, is one for $t = m$ and is zero for $t \neq m$. If we know (α_1, α_2), the variable

$$X_{t3} = \begin{cases} 0 & \text{if } t < 65, \\ 1 & \text{if } t = 65, \\ \alpha_1 X_{t-1,3} + \alpha_2 X_{t-2,3} & \text{if } t > 65 \end{cases} \tag{8.7.15}$$

will satisfy the requirements for $m = 65$, because X_{t3} satisfies the difference equation. We created the vector $\mathbf{X}_t = (X_{t1}, X_{t2}, X_{t3})$ using $(\hat{\alpha}_1, \hat{\alpha}_2) = (-1.308, 0.840)$ from our initial fit of the autoregressive model. The model (8.7.13), estimated by maximum likelihood using the ARIMA procedure in SAS®, gives the estimates

$$\begin{array}{rrrrrr} (\tilde{\mu}, \tilde{\zeta}_{65}, \tilde{\zeta}_{85}, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\sigma}^2) = (20.991, & 4.65, & 5.27, & -1.390, & 0.877, & 1.157). \\ & (0.220) & (0.55) & (1.07) & (0.048) & (0.048) & (0.170) \end{array}$$

If we iterate, using $(\tilde{\alpha}_1, \tilde{\alpha}_2) = (-1.390, 0.877)$ to create a new X_{t3} variable, we obtain

$$\begin{array}{rrrrrr} (\overset{*}{\mu}, \overset{*}{\zeta}_{65}, \overset{*}{\zeta}_{85}, \overset{*}{\alpha}_1, \overset{*}{\alpha}_2, \overset{*}{\sigma}^2) = (20.982, & 4.64, & 5.16, & -1.395, & 0.880, & 1.179). \\ & (0.223) & (0.55) & (1.10) & (0.048) & (0.047) & (0.173) \end{array} \tag{8.7.16}$$

Note that iteration changed the estimates very little.

The estimates of ζ_{65} and ζ_{85} are estimated differences between the observed values and the values expected under the autoregressive model, where the deviation is assumed to be an additive outlier for $t = 85$ and an innovation outlier for $t = 65$. Because we identified these points from the residual plots, we should not use ordinary critical values in making tests. Monte Carlo studies by Chang, Tiao, and Chen (1988) suggest that the largest t in a sample of 100 will exceed 4.00 about 1% of the time when the null model is true. Therefore, both observations are suspect.

Removing the outliers produces a large reduction in the estimated variance, where the estimated variance in (8.7.16) is only about 55% of that in (8.7.12). The changes in the autoregressive coefficients due to removing the outliers are between one and two standard errors, and the estimated mean is changed by about one-half the standard error.

If we use the estimates given in (8.7.12) to create predictions for the next three periods, we obtain

$$\begin{array}{rrrr} (\hat{Y}_{101}, \hat{Y}_{102}, \hat{Y}_{103}) = (20.73, & 18.38, & 17.89). \\ & (1.48) & (2.44) & (2.76) \end{array}$$

The corresponding predictions constructed with (8.7.16) are

$$\begin{array}{rrrr} (\hat{Y}_{101}, \hat{Y}_{102}, \hat{Y}_{103}) = (20.72, & 18.14, & 17.22). \\ & (1.08) & (1.85) & (2.18) \end{array}$$

The largest effect of removing the outliers is a reduction in the estimated standard error of the prediction errors. One should evaluate these effects when deciding to remove outliers. If the true model is one with long tailed errors, then removing

outliers and using the smaller estimate of σ^2 will lead to confidence intervals for predictions that are too narrow. ▲▲

The estimation procedure of Example 8.7.2 assigns zero weight to observations that are identified as outliers. Alternatively, the presence of unusual observations might lead to a modification of the working model and the application of estimation procedures to the modified model. One might specify a long tailed distribution or a mixture distribution for the errors and use maximum likelihood estimation under the specified model. See Kitagawa (1987), Peña and Guttman (1989), and Durbin and Cordero (1993).

Estimators that perform well under a wide range of statistical models are called robust estimators. See Hampel et al. (1986) and Huber (1981) for general discussions of robust procedures. Robust procedures that automatically down-weight large deviation observations have also been considered for time series. See, for example, Martin (1980), Kreiss (1987), and Bruce and Martin (1989).

8.8. LONG MEMORY PROCESSES

We introduced long memory time series in Section 2.11. The long memory time series that is called fractionally differenced noise satisfies

$$(1 - \mathcal{B})^d Y_t = \sum_{j=0}^{\infty} \kappa_j(d) Y_{t-j} = e_t \tag{8.8.1}$$

and

$$Y_t = \sum_{j=0}^{\infty} \upsilon_j(d) e_{t-j},$$

where

$$\kappa_j(d) = [\Gamma(j+1)\Gamma(-d)]^{-1}\Gamma(j-d),$$
$$\upsilon_j(d) = [\Gamma(j+1)\Gamma(d)]^{-1}\Gamma(j+d),$$

and $-0.5 < d < 0.5$. As might be expected, the estimation theory associated with such processes differs from that of short memory processes. For example, consider the variance of the sample mean. We have

$$V\{\bar{y}_n\} = n^{-2}\left[n\gamma_Y(0) + 2\sum_{h=1}^{n-1} (n-h)\gamma_Y(h) \right].$$

If $\gamma_Y(h)$ is proportional to h^{2d-1} and $d < 0$, then $\gamma_Y(h)$ is absolutely summable and $V\{\bar{y}_n\} = O(n^{-1})$. If $0 < d < 0.5$, $\gamma_Y(h)$ is not summable, but Yajima (1988) has shown that

$$\lim_{n\to\infty} n^{1-2d}V\{\bar{y}_n\} = \sigma^2\Gamma(1-2d)[\Gamma(d)\Gamma(1-d)d(1+2d)]^{-1}. \qquad (8.8.2)$$

The distributions of the sample autocovariances are difficult to obtain, but it is relatively easy to show that the sample autocovariances are consistent.

Lemma 8.8.1. Let Y_t be the infinite moving average time series

$$Y_t = \sum_{r=0}^{\infty} v_r e_{t-r},$$

where e_t are independent $(0, \sigma^2)$ random variables with bounded fourth moments and the v_r are square summable. Let

$$\hat{\gamma}_Y(h) = n^{-1}\sum_{t=1}^{n-h} Y_t Y_{t+h}.$$

Then $p\lim_{n\to\infty} \hat{\gamma}_Y(h) = \gamma_Y(h)$ for $h = 0, 1, 2, \ldots$.

Proof. Let

$$\hat{\gamma}_{X_j}(h) = n^{-1}\sum_{t=1}^{n} X_{tj}X_{t+h,j},$$

where $X_{tj} = \sum_{r=0}^{j} v_r e_{t-r}$. Because X_{tj} is a finite moving average,

$$\operatorname*{plim}_{n\to\infty} \hat{\gamma}_{X_j}(h) = \gamma_{X_j}(h) = E\{X_{tj}X_{t+h,j}\}$$

for every j. Also,

$$\operatorname*{plim}_{j\to\infty} \gamma_{X_j}(h) = E\{Y_t Y_{t+h}\}$$

because $\lim_{j\to\infty} E\{(X_{tj} - Y_t)^2\} = 0$. By Chebyshev's inequality,

$$P\{|\hat{\gamma}_Y(h) - \hat{\gamma}_{X_j}(h)| > \epsilon\} \leq (n\epsilon)^{-1}\sum_{t=1}^{n} E\{|X_{tj}X_{t+h,j} - Y_t Y_{t+h}|\}$$

and

$$\operatorname*{plim}_{j\to\infty} \{|\hat{\gamma}_Y(h) - \hat{\gamma}_{X_j}(h)|\} = 0$$

uniformly in n, because $\lim_{j\to\infty} E\{(X_{tj} - Y_t)^2\} = 0$, uniformly in t. Therefore, by Lemma 6.3.2,

$$\operatorname*{plim}_{n\to\infty} \hat{\gamma}_Y(h) = \gamma_Y(h). \qquad \blacktriangle$$

To introduce estimation for the parameters of long memory processes, assume

we have n observations from the Y_t of (8.8.1) and we desire an estimator of d. Consider the estimator of d obtained by minimizing

$$Q_n(d) = n^{-1} \sum_{t=1}^{n} \left[Y_t + \sum_{j=1}^{t-1} \kappa_j(d) Y_{t-j} \right]^2 \tag{8.8.3}$$

with respect to d. Note that the infinite autoregressive representation of Y_t is truncated at length $t-1$ in (8.8.3). The function (8.8.3) is of the same type as (8.4.7).

Lemma 8.8.2. Let the model (8.8.1) hold, let $d_0 \in \Theta$ be the true parameter, and let $\Theta = [d_{10}, d_{20}] \subset (0.0, 0.5)$. Assume the e_t are independent identically distributed $(0, \sigma^2)$ random variables with finite fourth moment $E\{e_t^4\} = \xi\sigma^4$. Let $\hat{d}$ be the value of $d \in \Theta$ that minimizes (8.8.3). Then $\operatorname{plim}_{n\to\infty} \hat{d} = d_0$.

Proof. We show that $\hat{d}$ is a consistent estimator of d_0 by showing that

$$\lim_{n\to\infty} P\{ \inf_{|d-d_0|\geq\eta} [Q_n(d) - Q_n(d_0)] > 0 \} = 1 \tag{8.8.4}$$

for all $\eta > 0$. Consider the difference

$$C_n(d) = n^{-1} \sum_{t=1}^{n} [S_t^2(d) - A_t^2(d)] = n^{-1} \sum_{t=1}^{n} \left\{ \left[A_t(d) - \sum_{j=t}^{\infty} \kappa_j(d) Y_{t-j} \right]^2 - A_t^2(d) \right\},$$

where $Q_n(d) = n^{-1} \Sigma_{t=1}^{n} S_t^2(d)$ and

$$[S_t(d), A_t(d)] = \left[\sum_{j=0}^{t-1} \kappa_j(d) Y_{t-j}, \sum_{j=0}^{\infty} \kappa_j(d) Y_{t-j} \right].$$

Now,

$$\begin{aligned} E\{[A_t(d) - S_t(d)]^2\} &= \lim_{n\to\infty} E\left\{ \left[\sum_{i=0}^{\infty} \kappa_{t+i}(d) \sum_{r=0}^{n} v_r(d_0) e_{-i-r} \right]^2 \right\} \\ &\leq \lim_{n\to\infty} \left[\sum_{i=0}^{\infty} |\kappa_{t+i}(d)| \right]^2 \sum_{r=0}^{n} v_r^2(d_0) \sigma^2 \\ &= O(t^{-2d}). \end{aligned}$$

Therefore

$$P\left\{ n^{-1} \sum_{t=1}^{n} [A_t(d) - S_t(d)]^2 > \epsilon \right\} = O(n^{-2d})$$

and

$$P\left\{\left|n^{-1}\sum_{t=1}^{n}A_t(d)[A_t(d)-S_t(d)]\right|>\epsilon\right\}=O(n^{-d}).$$

By the mean value theorem,

$$|C_n(d_1)-C_n(d_2)|\leq|d_1-d_2|\sup_{d\in\Theta}\left|\frac{\partial C_n(d)}{\partial d}\right| \tag{8.8.5}$$

for all d_1 and d_2 in Θ, where

$$\begin{aligned}\frac{\partial C_n(d)}{\partial d}&=n^{-1}\sum_{t=1}^{n}\left[2S_t(d)\frac{\partial S_t(d)}{\partial d}-2A_t(d)\frac{\partial A_t(d)}{\partial d}\right],\\ \frac{\partial A_t(d)}{\partial d}&=\sum_{j=1}^{\infty}\kappa_j(d)\left[d^{-1}-\sum_{i=2}^{j}(i-1-d)^{-1}\right]Y_{t-j}\\ &\overset{\text{say}}{=}\sum_{j=1}^{\infty}b_j(d)Y_{t-j}=U_t(d),\end{aligned} \tag{8.8.6}$$

and $|b_j(d)|\leq Mj^{-1-d}\log j$ as $j\to\infty$. Since the coefficients $\kappa_j(d)$ and $b_j(d)$ are absolutely summable and attain their supremum at some point in the set $[d_{10},d_{20}]$, it follows that the supremum of the derivative on the right of (8.8.5) is $O_p(1)$. Thus, by Lemma 5.5.5, $p\lim_{n\to\infty}C_n(d)=0$ uniformly in d. Therefore, we consider the d associated with the minimum of $\Sigma_{t=1}^{n}A_t^2(d)$.

We have

$$\operatorname*{plim}_{n\to\infty}n^{-1}\sum_{t=1}^{n}A_t^2(d)=E\{A_t^2(d)\}$$

because $A_t(d)=\Sigma_{i=0}^{\infty}g_i(d)e_{t-i}$, where $g_i(d)$ is square summable by Theorem 2.2.3,

$$g_i(d)=\sum_{r=-\infty}^{\infty}\kappa_{i-r}(d)v_r(d_0),$$

and it is understood that $\kappa_j(d)=0$ and $v_j(d_0)=0$ for $j<0$. Again, by the mean value theorem,

$$\left|n^{-1}\sum_{t=1}^{n}A_t^2(d_1)-n^{-1}\sum_{t=1}^{n}A_t^2(d_2)\right|\leq|d_1-d_2|\sup_{d\in\Theta}\left|2n^{-1}\sum_{t=1}^{n}A_t(d)U_t(d)\right|$$

for all d_1 and d_2 in Θ, and, by the properties of $\kappa_j(d)$ and $b_j(d)$, the supremum on the right is $O_p(1)$.

Now,

$$\begin{aligned}|E\{A_t^2(d_1)\}-E\{A_t^2(d_2)\}|&\leq|d_1-d_2|\sup_{d\in\Theta}\left|\frac{\partial E\{A_t^2(d)\}}{\partial d}\right|\\ &\leq 2|d_1-d_2|E\{\sup_{d\in\Theta}|A_t(d)U_t(d)|\}\\ &\leq M|d_1-d_2|\end{aligned}$$

for some $M < \infty$ because $\sup_{d\in\Theta} |\kappa_j(d)|$ and $\sup_{d\in\Theta}|b_j(d)|$ are absolutely summable; we have used the dominated convergence theorem. Hence, by Lemma 5.5.5,

$$\operatorname*{plim}_{n\to\infty} n^{-1} \sum_{t=1}^{n} A_t^2(d) = E\{A_t^2(d)\}$$

uniformly in d. Because $E\{A_t^2(d)\}$ reaches its minimum at d_0, the condition (8.8.4) is established. ▲

Results on estimators of d and of the parameters of the process

$$(1 - \mathcal{B})^d Y_t = Z_t, \tag{8.8.7}$$

where Z_t is a stationary autoregressive moving average, have been obtained by a number of authors. Maximum likelihood estimation for the normal distribution model has been studied by Fox and Taqqu (1986), Dahlhaus (1989), Haslett and Raftery (1989), and Beran (1992). Properties of Gaussian maximum likelihood estimators for linear processes have been investigated by Yajima (1985) and Giraitis and Surgailis (1990). Also see Robinson (1994a). The following theorem is due to Dahlhaus (1989). The result was extended to linear processes by Giraitis and Surgailis (1990).

Theorem 8.8.1. Let Y_t satisfy (8.8.7), where $d \in [d_{10}, d_{20}] \subset (0.0, 0.5)$, and Z_t is a stationary normal autoregressive moving average with $(k-1)$-dimensional parameter vector $\boldsymbol{\theta}_2$, where $\boldsymbol{\theta}_2$ is in a compact parameter space Θ_2. Let $\boldsymbol{\theta} = (d, \boldsymbol{\theta}_2')'$. Let $\hat{\boldsymbol{\theta}}$ be the value of $\boldsymbol{\theta}$ that maximizes the likelihood. Then

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{V}_{\theta\theta}),$$

where

$$\mathbf{V}_{\theta\theta}^{-1} = (4\pi)^{-1} \int_{-\pi}^{\pi} \mathbf{g}_\theta'(\omega)\mathbf{g}_\theta(\omega)\, d\omega,$$

$$\mathbf{g}_\theta(\omega) = \left[\frac{\partial \log f_Y(\omega)}{\partial\theta_1}, \ldots, \frac{\partial \log f_Y(\omega)}{\partial\theta_k}\right],$$

$f_Y(\omega)$ is the spectral density of Y_t, and the derivatives are evaluated at the true $\boldsymbol{\theta}$.

Proof. Omitted. See Dahlhaus (1989). ▲

The inverse of the covariance matrix of Theorem 8.8.1 is also the limit of the expected value of $n^{-1}\mathbf{h}_\theta'(\mathbf{Y})\mathbf{h}_\theta(\mathbf{Y})$, where

$$\mathbf{h}_\theta(\mathbf{Y}) = \left[\frac{\partial \log L(\mathbf{Y}:\boldsymbol{\theta})}{\partial\theta_1}, \ldots, \frac{\partial \log L(\mathbf{Y}:\boldsymbol{\theta})}{\partial\theta_k}\right],$$

$\log L(\mathbf{Y}: \boldsymbol{\theta})$ is the log-likelihood function, and the derivatives are evaluated at the true $\boldsymbol{\theta}$.

REFERENCES

Section 8.1. Anderson (1971), Gonzalez-Farias (1992), Hasza (1980), Jobson (1972), Koopmans (1942).

Section 8.2. Anderson (1959, 1962, 1971), Box and Pierce (1970), Draper and Smith (1981), Hannan (1970), Kendall (1954), Mann and Wald (1943a), Marriott and Pope (1954), Reinsel (1993), Salem (1971), Shaman and Stine (1988).

Sections 8.3, 8.4. Anderson and Takemura (1986), Berk (1974), Box, Jenkins and Reinsel (1994), Brockwell and Davis (1991), Cryer and Ledolter (1981), Durbin (1959), Eltinge (1991), Hannan (1979), Kendall and Stuart (1966), Macpherson (1975), Nelson (1974), Pierce (1970a), Sarkar (1990), Walker (1961), Wold (1938).

Section 8.5. Davisson (1965), Fuller (1980), Fuller and Hasza (1980, 1981), Hasza (1977), Phillips (1979), Yamamoto (1976).

Section 8.6. Granger and Andersen (1978), Priestley (1988), Quinn (1982), Subba Rao and Gabr (1984), Tong (1983, 1990).

Section 8.7. Chang, Tiao, and Chen (1988), Fox (1972), Jones (1980), Ljung (1993), Tsay (1988).

Section 8.8. Beran (1992), Dahlhaus (1989), Deo (1995), Fox and Taqqu (1986), Yajima (1985).

EXERCISES

1. Using the first 25 observations of Table 7.2.1, estimate ρ and λ using the equations (8.1.7). Compute the estimated standard errors of $\hat{\rho}$ and $\hat{\lambda}$ using the standard regression formulas. Estimate μ using the first equation of (8.1.9). Then compute the root of (8.1.10) using $y_t = Y_t - \hat{\mu}$. Iterate the computations.

2. Let Y_t be a stationary time series. Compare the limiting value of the coefficient obtained in the regression of $Y_t - \hat{r}(1)Y_{t-1}$ on $Y_{t-1} - \hat{r}(1)Y_{t-2}$ with the limiting value of the regression coefficient of Y_{t-2} in the multiple regression of Y_t on Y_{t-1} and Y_{t-2}.

3. Compare the variance of $\bar{y}_n$ for a first order autoregressive process with $\text{Var}\{\tilde{\mu}\}$ and $\text{Var}\{\hat{\mu}\}$, where $\tilde{\mu} = \hat{\lambda}(1-\rho)^{-1}$, and $\hat{\lambda}$ and $\hat{\mu}$ are defined in (8.1.7) and (8.1.9), respectively. In computing $\text{Var}\{\hat{\mu}\}$ and $\text{Var}\{\tilde{\mu}\}$, assume that ρ is known without error. What are the numerical values for $n = 10$ and $\rho = 0.7$? For $n = 10$ and $\rho = 0.9$?

4. Assume that 100 observations on a time series gave the following estimates:

$$\hat{\gamma}(0) = 200\,, \qquad \hat{r}(1) = 0.8\,, \qquad \hat{r}(2) = 0.7\,, \qquad \hat{r}(3) = 0.5\,.$$

Test the hypothesis that the time series is first order autoregressive against the alternative that it is second order autoregressive.

5. The estimated autocorrelations for a sample of 100 observations on the time series $\{X_t\}$ are $\hat{r}(1) = 0.8$, $\hat{r}(2) = 0.5$, and $\hat{r}(3) = 0.4$.
   (a) Assuming that the time series $\{X_t\}$ is defined by

$$X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + e_t,$$

   where the e_t are normal independent $(0, \sigma^2)$ random variables, estimate β_1 and β_2.
   (b) Test the hypothesis that the order of the autoregressive process is two against the alternative that the order is three.

6. Show that for a fixed $\tilde{\beta}$, $\tilde{\beta} \neq \beta^0$, the derivative $W_t(Y; \tilde{\beta})$ of (8.3.9) converges to an autoregressive moving average of order (2, 1). Give the parameters.

7. Fit a first order moving average to the first fifty observations in Table 8.3.1. Predict the next observation in the realization. Establish an approximate 95% confidence interval for your prediction.

8. Assume that the e_t of Theorem 8.3.1 are independent with $4 + \delta$ moments for some $\delta > 0$. Show that $n^{1/2}[\hat{\sigma}^2 - (\sigma^0)^2]$ converges in distribution to a normal random variable.

9. Prove Theorem 8.3.1 for the estimator constructed with $Y_t - \bar{y}_n$ replacing Y_t in all defining equations.

10. Fit an autoregressive moving average (1, 1) to the data of Table 8.3.1.

11. The sample variance of the Boone sediment time series discussed in Section 6.4 is 0.580, and the sample variance of the Saylorville sediment time series is 0.337. Let X_{1t} and X_{2t} represent the deviations from the sample mean of the Boone and Saylorville sediment, respectively. Using the correlations of Table 6.4.1, estimate the following models:

$$X_{2t} = \theta_1 X_{1,t-1} + \theta_2 X_{1,t-2},$$
$$X_{2t} = \theta_1 X_{1,t-1} + \theta_2 X_{1,t-2} + \theta_5 X_{2,t-1},$$
$$X_{2t} = \theta_1 X_{1,t-1} + \theta_2 X_{1,t-2} + \theta_3 X_{1,t-3} + \theta_5 X_{2,t-1} + \theta_6 X_{2,t-2},$$
$$X_{2t} = \sum_{i=1}^{4} \theta_i X_{i,t-i} + \sum_{i=5}^{7} \theta_i X_{2,t-i+4}.$$

   On the basis of these regressions, suggest a model for predicting Saylorville sediment, given previous observations on Boone and Saylorville sediment.

12. Fit autoregressive equations of order 1 through 7 to the data of Exercise 10 of Chapter 6. Choose a model for these data.

13. A test statistic for partial autocorrelations is given in (8.2.19). An alternative statistic is $t_i^* = n^{1/2}\hat{\beta}_{ii}$. Show that

$$t_i^* \xrightarrow{\mathcal{L}} N(0, 1)$$

under the assumptions of Theorem 8.2.1 and the assumption that $\beta_{ij} = 0$ for $j \geq i$.

14. A simple symmetric estimator with mean adjustment can be defined as a function of the deviations $Y_t - \bar{y}_n$. Show that

$$\tilde{\rho}_{s0} - 1 = -2\left\{\sum_{t=2}^{n} [(Y_t - \bar{y}_n)^2 + (Y_{t-1} - \bar{y}_n)^2]\right\}^{-1} \sum_{t=2}^{n} (Y_t - Y_{t-1})^2,$$

where $\tilde{\rho}_{s0} = \{\Sigma_{t=2}^{n}[(Y_t - \bar{y}_n)^2 + (Y_{t-1} - \bar{y}_n)^2]\}^{-1} \Sigma_{t=2}^{n}(Y_{t-1} - \bar{y}_n)(Y_t - \bar{y}_n)$.

15. We can express the $Q(\boldsymbol{\alpha})$ of (8.2.14) as

$$Q(\boldsymbol{\alpha}) = \sum_{t=1}^{n} \sum_{i=0}^{p} \sum_{j=0}^{p} c_{ijt} Y_{t-i} Y_{t-j} \alpha_i \alpha_j,$$

where $\alpha_0 = 1$ and $c_{ijt} = 0$ if $t - j < 1$ or $t - i < 1$. Show that $c_{01t} = 1$ for $t = 2, 3, \ldots, n$ for the weights as defined in (8.2.15). What are c_{00t} and c_{11t} for $w_t = 0.5$? What are c_{00t} and c_{11t} for the weights of (8.2.15)?

16. Compute and plot the estimated autocorrelation functions for the AR(4) and ARMA(2, 2) models of Table 8.4.2 for $h = 0, 1, \ldots, 12$. Does this help explain why it is difficult to choose between these models?

17. Consider a first order stationary autoregressive process $Y_t = \rho Y_{t-1} + e_t$, with $|\rho| < 1$ and e_t satisfying the conditions of Theorem 8.2.1. Prove Theorem 8.2.2 for the first order autoregressive process by showing that the difference between any two of the estimators of ρ is $o_p(n^{-1/2})$.

18. Consider a first order invertible moving average process $Y_t = \beta e_{t-1} + e_t$, where $|\beta| < 1$, and the e_t are iid$(0, \sigma^2)$ random variables with bounded fourth moments. Let $\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_k$ be the regression coefficients in (8.3.15). Let $\hat{\beta}_{D,k}$ denote the Durbin's estimator given by $-[\Sigma_{i=1}^{k} \hat{c}_{i-1}^2]^{-1} [\Sigma_{i=1}^{k} \hat{c}_{i-1}\hat{c}_i]$, where $\hat{c}_0 = -1$. Find the asymptotic distribution of $\hat{\beta}_{D,k}$ for a fixed k. Find the limit as $k \to \infty$ of the asymptotic mean and the variance of $\hat{\beta}_{D,k}$.

19. Let $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2)$ be the estimated parameters for the second order auto-

regressive process of Example 8.7.1. Use $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\sigma}^2)$ to construct estimates of the first five autocovariances of the process. Then use observations $(Y_{38}, Y_{39}, Y_{41}, Y_{42})$ and the estimated autocovariances to estimate Y_{40}. Use $(Y_{55}, Y_{56}, Y_{59}, Y_{60})$ to estimate (Y_{57}, Y_{58}).

20. Assume that Y_t is the first order normal stationary time series

$$Y_t = e_t + \beta e_{t-1}, e_t \sim \text{NI}(0, \sigma^2).$$

Show that the log-likelihood can be written

$$\log L(Y; \beta) = -0.5n \log 2\pi - 0.5 \sum_{t=1}^{n} \log V_t - 0.5n \log \sigma^2 - 0.5\sigma^{-2} \sum_{t=1}^{n} V_t^{-1} Z_t^2,$$

where $Z_1 = Y_1$, $V_1 = (1 + \beta^2)$, $Z_t = Y_t - V_{t-1}^{-1} \beta Z_{t-1}$ for $t = 2, 3, \ldots, n$, and

$$V_t = 1 + \beta^2 - V_{t-1}^{-1} \beta^2, \qquad t = 2, 3, \ldots, n.$$

21. Prove the following.

Lemma. Let Y_t and e_t satisfy the assumptions of Theorem 8.2.1. Let c_{nt}, $t = 1, 2, \ldots, n$ and $n = 1, 2, \ldots$, be a triangular array of constants with

$$\sum_{t=1}^{n} c_{nt}^2 = 1 \quad \text{and} \quad \lim_{n \to \infty} \sup_{1 \leq t \leq n} c_{nt}^2 = 0.$$

Then

$$\left[\sigma^{-1} \sum_{t=1}^{n} c_{nt} e_t, [n\gamma_Y(0)]^{-1/2} \sum_{t=p+1}^{n} Y_{t-j} e_t \right]' \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{I})$$

for $j = 1, 2, \ldots, p$.