Estimating Time-Series Models

The Box-Jenkins methodology for fitting a model to a scalar time series $\{x_t\}$ consists of five steps:

1. Decide on the order of differencing d that is needed to produce a stationary series $y_t = (1-L)^d x_t$ that can be approximated by an ARMA(p,q) model (with intercept if the mean of y_t is not zero).
2. By inspecting the sample autocorrelations and partial autocorrelations of $\{y_t\}$, determine tentative values for p and q.
3. Estimate the lag coefficients by approximate maximum likelihood assuming normally distributed innovations.
4. Compute approximate standard errors and confidence intervals for the unknown coefficients.
5. Using various diagnostics, check if the tentative model was indeed appropriate.

Box and Jenkins suggest informal ways to perform the first two steps. A formal treatment of step 1 will appear in the second half of the course. A formal approach to step 2 can be developed using the theory of nonnested model selection; see, for example, the book *Time Series: Theory and Methods* by Brockwell and Davis. Here we shall discuss steps 3-5. Finally, in Section 6 we extend the discussion to ARMAX models.

1 Some Alternative Estimators

Consider the stationary ARMA(p,q) model $A_p(L)y_t = B_q(L)\varepsilon_t$ where the ε_t are white noise with zero mean and variance σ^2. The lag polynomials

$$A_p(L) = 1 - \alpha_1 L - \cdots - \alpha_p L^p \qquad \text{and} \qquad B_q(L) = 1 + \beta_1 L + \cdots + \beta_q L^q$$

are assumed to be invertible and have no common factors. Let θ be the column vector of the $m = p + q$ unknown lag coefficients. Suppose we want to estimate the unknown parameters θ and σ^2 using the vector of observations $\mathbf{y} = (y_1, ..., y_T)'$. If the ε_t's are jointly normal, so are the y_t's. Let $\Omega(\theta)$ be a $T \times T$ matrix defined by $E\mathbf{y}\mathbf{y}' = \sigma^2\Omega(\theta)$. Except for an additive constant, the Gaussian log likelihood function is

$$L(\theta, \sigma^2) = -\frac{1}{2}\ln|\Omega(\theta)| - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}.$$

Note that, for given θ, L is maximized when $\sigma^2 = T^{-1}\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}$. Thus the *concentrated* log likelihood function is

$$L^*(\theta) = \max_{\sigma^2} L(\theta, \sigma^2) = -\frac{1}{2}\ln|\Omega(\theta)| - \frac{T}{2}\ln\frac{\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}}{T} - \frac{T}{2}. \tag{1}$$

Three alternative methods are commonly used in practice for estimating θ:

1. Quasi Maximum Likelihood maximizes L^* in the parameter space of θ.
2. Unconditional Least Squares uses the fact that $|\Omega(\theta)|$ is negligible in large samples and minimizes $\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}$. This method is described in more detail in Section 4 below.

3. Conditional Least Squares: If one conditions on $y_1, ...y_p$ and sets $\varepsilon_p, \varepsilon_{p-1}, \cdots, \varepsilon_{p-q+1}$ to zero, the likelihood for the remaining y's has the same form as (1) but $|\Omega(\theta)|$ becomes a constant and $\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}$ becomes $e'e$, where $e = (\varepsilon_{p+1}, \cdots, \varepsilon_T)$. As discussed in Section 2 below, mimimizing $e'e$ is a problem in nonlinear least squares and can be accomplished using the Gauss-Newton algorithm.

Because it is computationally the simplest, conditional least squares is commonly used in practice. It can be shown that all three estimators have the same limiting distribution even if the errors are not normal. We have the following basic result:

Theorem 1 *Assume*

i. the parameter space for θ is such that the roots of $A_p(z) = 0$ and $B_q(z) = 0$ are outside the unit circle and there are no common roots,

ii. the true value θ_0 is in the interior of the parameter space,

iii. the ε_t are i.i.d. with mean zero and variance σ^2.

Then all three estimators of θ_0 and σ^2 converge in probability to the true value. Furthermore, for all three estimators, $\sqrt{T}(\hat{\theta} - \theta_0)$ converges in distribution to a normal random variable with mean zero and covariance matrix

$$V = \sigma^2 \left[E\left(\frac{\partial \varepsilon_t}{\partial \theta} \frac{\partial \varepsilon_t}{\partial \theta'} \right) \right]^{-1} .$$

If the ε_t are normally distributed, the estimators are asymptotically efficient.

Note that, except for the efficiency result, we do not need to assume the innovations are in fact normally distributed. Although the asymptotic variance expression does require that the innovations be conditionally homoskedastic, the i.i.d. assumption can be weakened. For example, Assumption iii can be replaced by the following without affecting the result:

iii'. both ε_t and $\varepsilon_t^2 - \sigma^2$ are martingale difference sequences (with respect to the past history of ε_t); that is, $\mathrm{E}(\varepsilon_t |\ \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = 0$ and $\mathrm{E}(\varepsilon_t^2 - \sigma^2 | \varepsilon_{t-1}, \varepsilon_{t-2}, ...) = 0$. In addition, $\mathrm{E}\varepsilon_t^{2+\delta} < \infty$ for some $\delta > 0$.

For a proof of the theorem, see W. Fuller, *Introduction to Statistical Time Series*, 2nd edition, pp. 429-43.

2 Conditional Least Squares

Maximum likelihood estimation is computationally demanding when the sample size T is large since the calculation involves inverting the $T \times T$ matrix Ω and computing its determinant. Later we shall show how a computational algorithm known as the Kalman filter can be employed to obtain exact maximum likelihood estimates by doing these calculations recursively. In practice, most empirical economists maximize an approximation to the likelihood based on a slight change in the initial conditions. For example, in the $AR(p)$ model, if we condition on the first p values $y_1, ..., y_p$ and examine the process starting at $t = p + 1$, maximum likelihood is equivalent to a least

squares regression of y_t on its p lagged values. We now show that a similar modification of initial conditions in the general ARMA case leads to a *nonlinear* least squares regression.

In every invertible ARMA(p,q) model, the observed y's have a moving average representation expressing them linearly to current and lagged ε's. Hence the likelihood function (the joint density of the observed y's) can be derived from the density of the ε's using familiar change-of-variable techniques. Unfortunately, the mapping y to ε is typically not one-to-one, so the calculation is nontrivial. However, conditioning on the observed values $y_1, ..., y_p$ and on $\varepsilon_p = \varepsilon_{p-1} = ... = \varepsilon_{p-q+1} = 0$ necessarily leads to a one-to-one mapping. Let $\mathbf{e}$ be the vector of white noise innovations $(\varepsilon_{p+1}, ..., \varepsilon_T)'$ and let $\mathbf{y}$ be the vector of observations $(y_{p+1}, ..., y_T)'$. Then, by successive substitution, we can write $\mathbf{e} = \mathbf{D}\mathbf{y} + \mathbf{d}$ where $\mathbf{D}$ is a triangular matrix with ones on the diagonal; the vector $\mathbf{d}$ depends on $y_1, ..., y_p$ and is nonrandom because of our conditioning. Thus we can write $\mathbf{e} = \mathbf{e}(\mathbf{y}, \theta)$ where the Jacobian $|\partial \mathbf{e}/\partial \mathbf{y}|$ does not depend on θ. If $\mathbf{e}/\sigma$ is standard normal, the change-of-variable rule gives the density for $\mathbf{y}$ having the form

$$f(\mathbf{y}) = (2\pi\sigma^2)^{-(T-p)/2} \exp\{-\frac{1}{2}\mathbf{e}(\mathbf{y}, \theta)'\mathbf{e}(\mathbf{y}, \theta)/\sigma^2\}$$

Maximum likelihood estimates of θ can be obtained by minimizing $\mathbf{e}'\mathbf{e}$. Although $\mathbf{e}$ is linear in $\mathbf{y}$, it is nonlinear in θ as long as q (the number of moving average coefficients to be estimated) is greater than zero. This nonlinear least squares problem can be solved using the Gauss-Newton algorithm. Let $m = p + q$ be the number of unknown parameters in θ and let $n = T - p$ be the number of observations after conditioning. (If there is an intercept, then $m = p + q + 1$.) Then, defining the $n \times m$ matrix $Z(\theta, \mathbf{y}) \equiv \partial \boldsymbol{e}/\partial \theta'$ and starting with some initial guess θ^0 , we use the recursion

$$\theta^{r+1} = \theta^r - (Z_r'Z_r)^{-1}Z_r'e^r$$

where $Z_r = Z(\theta^r, \mathbf{y})$ and $e^r = e(\theta^r, \mathbf{y})$. When $Z_r'\theta^r \simeq 0$, one stops.

For example, suppose $y_t = \alpha y_{t-1} + \varepsilon_t + \beta\varepsilon_{t-1}$ and we treat y_1 as fixed and assume $\varepsilon_1 = 0$. Then, using the equation $\varepsilon_t = y_t - \alpha y_{t-1} - \beta\varepsilon_{t-1}$ with $\varepsilon_1 = 0$, we find

$$\begin{aligned}
\varepsilon_2 &= y_2 - \alpha y_1, \\
\varepsilon_3 &= y_3 - (\alpha + \beta)y_2 + \alpha\beta y_1, \\
\varepsilon_4 &= y_4 - (\alpha + \beta)y_3 + \beta(\alpha + \beta)y_2 - \alpha\beta^2 y_1, \text{ etc.}
\end{aligned}$$

Clearly, the Jacobian matrix $\partial\varepsilon/\partial y$ is triangular with determinant equal to one. However, we do not have to explicitly solve for the ε's in terms of the y's. Under normality, the MLE which minimizes the sum of squared innovations can be computed using the following G-N algorithm:

1. For some initial $\theta^0 = (\alpha^0, \beta^0)'$ and starting with the initial condition

$$\varepsilon_1 = \partial\varepsilon_1/\partial\alpha = \partial\varepsilon_1/\partial\beta = 0 \ ,$$

build up the vector e^0 and the $(T-1) \times 2$ matrix $Z_0 = [\partial e/\partial\alpha, \partial e/\partial\beta]$ using the recursions

$$\begin{aligned}
\varepsilon_t &= y_t - \alpha^0 y_{t-1} - \beta^0\varepsilon_{t-1} \\
\partial\varepsilon_t/\partial\alpha &= -y_{t-1} - \beta^0\partial\varepsilon_{t-1}/\partial\alpha \\
\partial\varepsilon_t/\partial\beta &= -\varepsilon_{t-1} - \beta^0\partial\varepsilon_{t-1}/\partial\beta
\end{aligned}$$

2. Regress e^0 on Z_0 and compute $\theta^1 = \theta^0 - (Z_0'Z_0)^{-1}Z_0'e^0$.

3. Redo steps one and two using θ^1 in place of θ^0 so $\theta^2 = \theta^1 - (Z_1'Z_1)^{-1}Z_1'e^1$.

4. Continue to convergence.

Of course, this algorithm fails if, at iteration r, Z_r has rank less than 2. So, for example, one should not start with the initial choice $\theta^0 = (0,0)'$.

Note that exogenous regressors can be easily handled by the Gauss-Newton algorithm. For example, if the model is

$$y_t = \alpha y_{t-1} + \gamma x_t + \varepsilon_t + \beta\varepsilon_{t-1}$$

then Z is the $(T-1)\times 3$ matrix $[\partial e/\partial\alpha, \partial\varepsilon/\partial\beta, \partial\varepsilon/\partial\gamma]$ and the recursions are

$$\begin{aligned}
\varepsilon_t &= y_t - \alpha^0 y_{t-1} - \gamma^0 x_t - \beta^0\varepsilon_{t-1} \\
\partial\varepsilon_t/\partial\alpha &= -y_{t-1} - \beta^0\partial\varepsilon_{t-1}/\partial\alpha \\
\partial\varepsilon_t/\partial\beta &= -\varepsilon_{t-1} - \beta^0\partial\varepsilon_{t-1}/\partial\beta \\
\partial\varepsilon_t/\partial\gamma &= -x_t - \beta^0\partial\varepsilon_{t-1}/\partial\gamma
\end{aligned}$$

Of course, to insure that the algorithm converges to a global (and not just a local) minimum, alternative starting values should be used. Or at least the starting values should be chosen as some consistent method-of-moments estimator that has high probability of being near the true parameter value.

If the model has been estimated by the Gauss-Newton algorithm, then a natural estimate of the asymptotic covariance matrix V is $s^2(Z'Z/T)^{-1}$, where $Z = de/d\theta$ and $s^2 = e'e/(n-m)$; both e and Z are evaluated at the estimate $\hat\theta$ from the final iteration. That is, we behave as though $\hat\theta$ is normal with mean θ^0 and variance matrix $s^2(Z'Z)^{-1}$. The usual tests and confidence regions based on the t and F-distributions are asymptotically valid.

The G-N algorithm maximizes an approximation to the likelihood since ε_1 is not really 0 and y_1 is not really constant. However, the G-N estimate is usually close to the actual MLE as long as the parameters are far away from the noninvertibility boundaries $|\alpha| = 1$ and $|\beta| = 1$.

3 Score Tests for ARMA Parameters

Score test statistics are typically much easer to compute than Wald test statistics but are asymptotically equivalent. If the null hypothesis puts r independent constraints on θ, we need only find the constrained MLE $\widetilde\theta$ which involves a lower dimensional nonlinear estimation problem. Using the output of a Gauss-Newton iteration, we would evaluate e and the full $n\times m$ matrix Z at $\widetilde\theta$ so the estimated score is $\widetilde Z'\widetilde e/\widetilde s^2$ and the estimated information matrix is $\widetilde Z'\widetilde Z/\widetilde s^2$. If

$$\frac{\widetilde e'\widetilde Z(\widetilde Z'\widetilde Z)^{-1}\widetilde Z'\widetilde e}{\widetilde s^2}$$

is large compared to a chi-square(r) critical value, one rejects the hypothesis. If θ is composed of two subvectors (θ_1,θ_2) and the null hypothesis is $\theta_2 = a$, the score statistic takes a simpler form. Partitioning the matrix Z as (Z_1, Z_2), we note that $\widetilde Z_1'\widetilde e = 0$. Thus the score statistic becomes

$$\frac{\widetilde e'\widetilde Z_2'(\widetilde Z_2'\widetilde Z_2 - \widetilde Z_2'\widetilde Z_1(\widetilde Z_1'\widetilde Z_1)^{-1}\widetilde Z_1'\widetilde Z_2)^{-1}\widetilde Z_2'\widetilde e}{\widetilde s^2}\,.$$

Score statistics are particularly useful for diagnostic checking. After fitting an ARMA(p,q) model, one might want to see whether an ARMA(p+1,q) or an ARMA(p,q+1) provides a significantly better fit. Since $\widetilde Z_1$ and $\widetilde e$ have already been computed, one need only compute the additional

vector $\widetilde{Z}_2$. There is no need to compute another Gauss-Newton set of iterations. Note that a score test of the hypothesis that an ARMA(p,q) is appropriate (in a maintained ARMA(p+1,q+1) model) would fail since $\widetilde{Z}_2$ would have rank 1. This reflects the fact that, under the null hypothesis, the AR and the MA polynomials have a common factor.

An alternative diagnostic test is also commonly used. If an ARMA(p,q) model is adequate, then the estimated innovations should behave like white noise. The low order sample autocorrelations of a white-noise series are asymptotically independent $N(0, T^{-1})$ variables. So T times the sum of the first k squared sample autocorrelations would be approximately distributed as chi-square-k. If the autocorrelations are computed from the $\widetilde{e}$, the estimated innovations after fitting an ARMA(p,q) model, the asymptotic theory is affected by the estimation. For large k and T, the *Box-Pierce* portmanteau statistic

$$T(\widetilde{\rho}_1^2 + \widetilde{\rho}_2^2 + \cdots \widetilde{\rho}_k^2)$$

is approximately distributed as chi square with $k-p-q$ degrees of freedom if the model is correctly specified.

4 Unconditional Least Squares and Backcasting

Although no longer used much in practice, *unconditional least squares* is another way to avoid the computational difficulties involved in maximizing the exact Gaussian likelihood function. It avoids calculating $|\Omega|$ by simply ignoring it. It simplifies the calculation of $\mathbf{y}'\Omega(\theta)^{-1}\mathbf{y}$, by the trick of "backcasting." Let $\boldsymbol{\varepsilon}$ be the vector $(\varepsilon_{-m}, \varepsilon_{-m+1}, \cdots, \varepsilon_T)'$ where m is some large positive integer. Then the vector of observations $\mathbf{y} = (y_1, \cdots y_T)'$ can be approximated by $\mathbf{y} = D\boldsymbol{\varepsilon}$ where D is a triangular matrix of moving average coefficients. It follows that $E\mathbf{y}\mathbf{y}' \approx \sigma^2 DD'$ and $\mathbf{y}'\Omega^{-1}\mathbf{y} \approx \boldsymbol{\varepsilon}' D'(DD')^{-1}D\boldsymbol{\varepsilon}$. It is easy to verify that, under normality, the conditional expectation of $\boldsymbol{\varepsilon}$ given $\mathbf{y}$ is

$$\boldsymbol{\varepsilon}_c \equiv E(\boldsymbol{\varepsilon}|\mathbf{y}) = (E\boldsymbol{\varepsilon}\mathbf{y}')(E\mathbf{y}\mathbf{y}')^{-1}\mathbf{y} \approx D'(DD')^{-1}\mathbf{y} \approx D'(DD')^{-1}D\boldsymbol{\varepsilon}.$$

Since $D'(DD')^{-1}D$ is idempotent, it follows that

$$\mathbf{y}'\Omega^{-1}\mathbf{y} \approx \boldsymbol{\varepsilon}' D'(DD')^{-1}D\boldsymbol{\varepsilon} \approx \boldsymbol{\varepsilon}_c'\boldsymbol{\varepsilon}_c.$$

Thus we find that estimation method 2 is equivalent to minimizing the sum of squares of the *expected* current and past innovations.

Box and Jenkins suggested a very clever way of doing this calculation. They noted that the backwards ARMA model $A(L^{-1})y_t = B(L^{-1})\eta_t$ has the same autocorrelation properties as the original model. Thus, given tentative estimates of the parameters, one can "backcast" the observations $y_0, y_{-1}, \cdots, y_{-m}$. From these values and using the original difference equation, one can compute best linear predictors of $\boldsymbol{\varepsilon}$ given $\mathbf{y}$. Hence the Gauss-Newton algorithm can be employed to minimize $\mathbf{y}'\Omega^{-1}\mathbf{y}$ just as it is employed to minimize $e'e$. Details of the calculation can be found in the Granger and Newbold text.

5 Asymptotic Theory

Proofs of the large-sample results given in the previous sections are nontrivial and rely on a general asymptotic theory for sums of dependent variables, a theory that is well beyond the scope of these notes. Here we shall present just a few of the basic ideas.

Suppose $\{y_t\}$ is a stationary AR(1) process generated by $y_t = \alpha y_{t-1} + \varepsilon_t$ with $|\alpha| < 1$. If α is estimated by least squares from the observations $y_0, y_1, ..., y_T$, then the standardized estimator can be written as

$$\sqrt{T}(\alpha - \alpha) = \frac{T^{-1/2} \sum y_{t-1}\varepsilon_t}{T^{-1} \sum y_{t-1}^2} = \frac{N_T}{D_T}$$

where the summations are from 1 to T. If N_T converges in distribution to a N(0, A) variable and if D_T converges in probability to the constant B, then the standardized estimator has a limiting N(0, A/B^2) distribution. If $\widehat{A}$ and $\widehat{B}$ are consistent estimates, then we might approximate the distribution of $\widehat{\alpha}$ by a normal with mean α and variance $\widehat{A}/T\widehat{B}^2$.

Note that $\mathrm{E}D_T = \sigma^2/(1-\alpha^2)$. To show that D_T converges in probability to its expectation, it is sufficient to show that its variance tends to zero as $T \to \infty$. But

$$var(D_T) = \frac{1}{T^2} \sum_{t=0}^{T-1} \sum_{s=0}^{T-1} cov(y_t^2, y_s^2) = \frac{1}{T^2} \sum_{r=-T+1}^{T-1} \sum_{t=0}^{T-1-r} cov(y_t^2, y_{t+r}^2)$$

Suppose the covariances do not depend on t so

$$var(D_T) = \frac{1}{T} \sum_{r=-T+1}^{T-1} (1 - \frac{|r|}{T}) g_r$$

where $g_r = cov(y_t^2, y_{t+r}^2)$. If $g_r \to 0$ as $r \to \infty$, then this variance necessarily converges to zero. The g_r can be computed from the moving average representation for y_t. If the ε_t are i.i.d. and possess finite fourth moment, a little algebra shows that

$$g_r = [\frac{2\sigma^4}{(1-\alpha^2)^2} + \frac{\kappa_4}{1-\alpha^4}]\alpha^{2r}$$

where κ_4 is the fourth cumulant of ε_t. Since $|\alpha| < 1$,we conclude that indeed $D_T \to B = \sigma^2/(1-\alpha^2)$. Using a more delicate argument, the assumption of a finite fourth moment can be dropped. Indeed, the i.i.d. assumption can be replaced by the assumption that ε_t and $\varepsilon_t^2 - \sigma^2$ are martingale difference sequences, although the argument is somewhat more tedious.

Showing that the numerator N_T is asymptotically normal is a bit more complicated. Let $U_t = y_{t-1}\varepsilon_t$ so $N_T = T^{-1/2} \sum U_t$. Suppose that both ε_t and $\varepsilon_t^2 - \sigma^2$ are stationary martingale difference sequences. Then $\{U_t\}$ is also a stationary martingale difference sequence and $T^{-1} \sum U_t^2$ converges in probability to its expectation

$$\frac{1}{T} \sum E(\varepsilon_t^2 y_{t-1}^2) = \frac{\sigma^4}{1-\alpha^2} = \sigma^2 B$$

as long as sufficient moments exist. Then, using the Taylor series expansion of the exponential function, the characteristic function for N_T can be written as

$$\begin{aligned}
\psi(\lambda) &= E \prod_{t=1}^{T} e^{i\lambda U_t/\sqrt{T}} = E \prod_{t=1}^{T} \left(1 + i\lambda \frac{U_t}{\sqrt{T}} - \frac{\lambda^2 U_t^2}{2T} + O(T^{-3/2})\right) \\
&\approx E \prod_{t=1}^{T} \left(1 + i\lambda \frac{U_t}{\sqrt{T}}\right) \left(1 - \frac{\lambda^2 U_t^2}{2T}\right) \\
&\approx E \prod_{t=1}^{T} \left(1 + i\lambda \frac{U_t}{\sqrt{T}}\right) e^{-\sum \lambda^2 U_s^2/2T} \approx e^{-\sigma^2 B \lambda^2/2} E \prod_{t=1}^{T} \left(1 + i\lambda \frac{U_t}{\sqrt{T}}\right),
\end{aligned}$$

where some delicate arguments are needed to justify the indicated approximations. Since U_s is a martingale difference sequence, iterated expectations can be used to show that the final expected product equals one. Thus the limiting characteristic function of N_T is that of a N(0, σ^2B) and hence $\sqrt{T}(\hat{\alpha} - \alpha)$ has a limiting normal distribution with variance $\sigma^2/B = 1 - \alpha^2$. [Note: if E($\varepsilon_t^2$|past) is not a constant but depends on past variables, it may still be the case that N_T is asymptotically normal. However, its asymptotic variance will not equal B. Thus, in the presence of conditional heteroskedasticity, the "usual" standard errors are incorrect. Robust standard errors are available and are discussed in Hamilton's text.]

For general ARMA models, the nonlinear least squares estimator satisfies the first order conditions $Z(\widehat{\theta})'e(\widehat{\theta}) = 0$. Expanding $e(\widehat{\theta})$ around the true parameter value θ_0, we obtain

$$Z(\widehat{\theta})'[e(\theta_0) + Z(\theta^*)(\widehat{\theta} - \theta_0)] = 0.$$

Often we can show that

$$\begin{aligned} \sqrt{T}(\widehat{\theta} - \theta_0) &= \sqrt{T}[Z(\widehat{\theta})'Z(\theta^*)]^{-1}Z(\widehat{\theta})'e(\theta_0) \\ &= \left[\frac{Z(\theta_0)'Z(\theta_0)}{T}\right]^{-1} \frac{Z(\theta_0)'e(\theta_0)}{\sqrt{T}} + o_p(1). \\ &\equiv D_T^{-1}N_T + o_p(1). \end{aligned}$$

Using techniques similar to those used above, one can show that D_T converges in probability to a nonrandom nonsingular matrix B and that, under conditional homoskedasticity, N_T is a standardized sum whose summands are a martingale difference sequence and is asymptotically N(0, $\sigma^2 B$). Hence, $\sqrt{T}(\widehat{\theta} - \theta_0)$ is asymptotically normal with variance $\sigma^2 B^{-1}$. Since B can be estimated by $Z(\widehat{\theta})'Z(\widehat{\theta})/T$, one says that $\widehat{\theta}$ is approximately normal with mean θ_0 and variance $s^2[Z(\widehat{\theta})'Z(\widehat{\theta})]^{-1}$

General conditions under which sample moments constructed from a strictly stationary time series converge in probability (or almost surely) to population moments are often referred to as *ergodic theorems.* These rather deep results arise from statistical mechanics and make assumptions about the probabilities of events that are invariant with respect to a time shift. For example, the event $\{y_t > 0 \text{ for all } t\}$ is invariant since it holds if and only if the event $\{y_{t+1} > 0 \text{ for all } t\}$ holds; but the event $\{y_5 > 0\}$ is not invariant since it does not imply $\{y_6 > 0\}$. A stationary process is called *indecomposable* if all invariant events occur with probability one or zero. A stationary process is called *ergodic* if every sample moment converges almost surely to its expectation. The basic ergodic theorem then says that indecomposablity is a necessary and sufficient condition for ergodicity. It can be shown that stationary normally distributed processes whose autocovariances γ_r tend to zero (as r tends to infinity) are indecomposable and hence ergodic. Strong mixing conditions also can be used. In general, however, except for the case of discrete Markov chains, it is rather difficult to find economically meaningful conditions that imply indecomposability. In practice, we usually just make assumptions on the summability of autocorrelation or moving average coefficients to prove convergence of sample moments.

6 Regression with Autocorrelated Errors

As noted in section 2, ARMAX models of the form

$$A(L)y_t = \mu + F_1(L)x_{1t} + ... + F_k(L)x_{kt} + B(L)\varepsilon_t$$

can be estimated by conditional (nonlinear) least squares as long as the lag polynomials $A, B, F_1, ...,$ F_k are all of low order or are ratios of low order polynomials. This assumes that the orders of the polynomials are known or can be easily determined. Often we desire an alternative approach that is robust to specification error in polynomial order.

First we consider the special case where $A(L) = I$; that is, we have a linear regression model

$$y = X\beta + u$$

where the errors are assumed to be independent of the regressors. The errors are a stationary time series with mean zero and second-moment matrix Ω. If the errors are approximately normal and Ω is known, a natural estimator of β is the GLS estimator $(X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$. Of course, in practice, Ω is unknown so a multi-step procedure is often used. First one prewhitens the data by picking a nonrandom matrix D such that Du is thought to be roughly white noise. Second, one regresses Dy on DX obtaining residual vector e. (In practice, most empirical economists seem to use the identity matrix for D so e is the OLS residual.) Fit an ARMA model to the residuals and use this estimated model to compute the matrix $\hat{\Omega}$. The final estimate of β is $\hat{\beta} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$. Under regularity conditions on the exogenous variables and some assumptions on the rate at which the high-order autocovariances tend to zero, it can be shown that if the number of parameters in the ARMA model for $\{u_t\}$ are allowed to increase slowly to infinity as the sample size increases, $\hat{\beta}$ is asymptotically equivalent to the GLS estimator with known Ω.

A similar result is available for rational distributed lag models. For example, suppose

$$y_t = \frac{\beta}{1 - \lambda L}x_t + u_t,$$

where u_t is stationary with mean zero but covariance matrix Ω. If Ω were known one could estimate the remaining paameters by minimizing $u'\Omega^{-1}u$. Instead one can procede as follows. Find preliminary consistent estimates of β and λ, say by instrumental variables. Fit an ARMA model to the estimated $\hat{u}_t$ and use that model to compute $\hat{\Omega}$. Then estimate the parameters by minimizing $u'\hat{\Omega}^{-1}u$. Again, if the number of parameters in the ARMA model for $\{u_t\}$ are allowed to increase slowly to infinity as the sample size increases and some regularity conditions are satisfied, this procedure is asymptotically equivalent to the estimator using the true Ω.