1. Univariate Time Series Analysis and Forecasting

1.1 *Expectations, stationarity, Ergodicity and White Noise Process*

A time series y_t is a process observed in sequence over time, $t = 1, \ldots, T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations. Because of the sequential nature of time series, we expect that y_t and y_{t-1} are not independent, so classical assumptions are not valid. We can separate time series into two categories: univariate $\left(y_t \in \mathbb{R} \text{ is scalar}\right)$; and multivariate $\left(\boldsymbol{y}_t \in \mathbb{R}^n \text{ is vector-valid}\right)$. The primary model for univariate time series is autoregressions (*ARs*). The primary model for multivariate time series is vector autoregressions (VARs).

Suppose we observed a sample of size *T* of some random variable Y$_t$: $\{y_t\}_{t=1}^{T}$. The observed sample represents *T* particular number, but this *T* numbers is only one possible outcome of the underlying stochastic process that generated the data. Indeed, even if we were to imagine having observed the process $\{y_t\}_{t=-\infty}^{\infty}$, this infinite sequence would still be viewed as a single realization from a time series process. Imagine we have *N* computers generating sequences $\{y_t^{(1)}\}_{t=-\infty}^{\infty}, \{y_t^{(2)}\}_{t=-\infty}^{\infty}, \ldots, \{y_t^{(N)}\}_{t=-\infty}^{\infty}$, and consider selecting the observation associated with date t from each sequence: $\left\{y_t^{(1)}, y_t^{(2)}, \cdots, y_t^{(N)}\right\}$. This would be described as a sample of *N* realizations of a random variable Y$_t$. This random variable has some density function, denoted by $f_{Y_t}(y_t)$, which is the unconditional density of Y$_t$.

The expectation of the tth observation of the time series refers to the mean of this probability distribution provided it exists:

$$E(Y_t) = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t)\, dy_t \qquad [1.1.1]$$

We might view this as the probability limit of the ensemble average:

$$E(Y_t) = \operatorname*{plim}_{N\to\infty}(1/\ N)\sum_{i=1}^{N} y_t^{(i)} \qquad [1.1.2]$$

Sometimes for emphasis the expectation *E*(Y$_t$) is called the *unconditional mean* of Y$_t$, denoted by μ_t. Note that this notation allows the general possibility that the mean can be a function of the date of the observation *t*.

The variance of the random variable Y$_t$ (denoted by γ_0) is similarly defined as:

$$\gamma_{0t} = E(Y_t - \mu_t)^2 = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t) dy_t \qquad [1.1.3]$$

Autocovariance at lag k of Y$_t$ (denoted by □$_k$) is defined as:

$$\begin{aligned}\gamma_{kt} &= E\left[(Y_t - \mu_t)(Y_{t-k} - \mu_{t-k})\right] \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (y_t - \mu_t)(y_{t-k} - \mu_{t-k}) f_{Y_t, Y_{t-1},\ldots,Y_{t-k}}(y_t) dy_t dy_{t-1} \ldots dy_{t-k}\end{aligned} \qquad [1.1.4]$$

Again it may also be helpful to think of the autocovariance at lag *k* as the probability limit of an ensemble average:

$$\gamma_{kt} = \operatorname*{plim}_{N\to\infty} (1/\ N) \sum_{i=1}^{N} \left(y_t^{(i)} - \mu_t\right)\left(y_{t-k}^{(i)} - \mu_{t-k}\right) \qquad [1.1.5]$$

Stationarity and Ergodicity

Definition 1.1.1 $\{y_t\}$ *is* ***covariance (weakly) stationary*** *if* $E(y_t) = \mu$ *is independent of t, and*

$\operatorname{cov}(y_t, y_{t-k}) = \gamma(\kappa)$ *is independent of t for all k.* $\gamma(\kappa)$ *is the* ***autocovariance function*** *at lag k.*

$\rho(k) = \gamma(k)/\gamma(0) = \operatorname{corr}(y_t, y_{t-k})$ *is the* ***autocorrelation function*** *at lag k.*

Definition 1.1.2 $\{y_t\}$ *is* ***strictly stationary*** *if the joint distribution of* $(y_t, \ldots, y_{t-k})$ *is independent of t for all k.*

We have viewed expectations of a time series in terms of ensemble averages such as [1.1.2] and [1.1.5]. These definitions may seem a bit contrived, since usually all one has available is a single realization of size *T* from the process, which may be denoted by $\left\{y_1^{(1)}, y_2^{(1)}, \cdots, y_T^{(1)}\right\}$. From these observations, we would calculate the sample mean $\bar{y}$. This, of course, is not an ensemble average but rather a time average:

$$\bar{y} = (1/\ T) \sum_{t=1}^{T} y_t^{(1)} \qquad [1.1.6]$$

Whether time averages such as [1.1.6] eventually converges to the ensemble average concept $E(y_t)$ for a stationary process has to do with ergodicity

Definition 1.1.3 ***A stationary*** *time series is ergodic if* $\gamma(\kappa) \uparrow 0$ *as* $k \uparrow \infty$.

The following two theorems are essential to the analysis of stationary time series. There proofs are rather difficult, however.

Theorem 1.1.1 *If y_t is strictly stationary and ergodic and $x_t = f(y_t, y_{t-1}, \ldots)$ is a random variable, then x_t is strictly stationary and ergodic.*

Theorem 1.1.2 *(Ergodic Theorem). If y_t is strictly stationary and ergodic and $E|y_t| < \infty$, then as $T\uparrow\infty$, $\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} E(y_t)$.*

This allows us to consistently estimate parameters using time-series moments:

The sample mean:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The sample autocovariance

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu})(y_{t-k} - \hat{\mu})$$

The sample autocorrelation

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$$

Theorem 1.1.3 *If y_t is strictly stationary and ergodic and $E(y_t^2) < \infty$, then as $T\uparrow\infty$*

1. $\hat{\mu} \xrightarrow{p} \mu$
2. $\hat{\gamma}(k) \xrightarrow{p} \gamma(k)$
3. $\hat{\rho}(k) \xrightarrow{p} \rho(k)$

Proof of Theorem 1.1.3. Part (1) is a direct consequence of the Ergodic theorem. For part (2), note that

$$\begin{aligned}\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu})(y_{t-k} - \hat{\mu}) \\ &= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} - \frac{1}{T}\sum_{t=1}^{T} y_t \hat{\mu} - \frac{1}{T}\sum_{t=1}^{T} y_{t-k}\hat{\mu} + \hat{\mu}^2\end{aligned}$$

By Theorem 1.1.1 above, the sequence $y_t y_{t-k}$ is strictly stationary and ergodic and it has a finite mean by the assumption that $E(y_t^2) < \infty$. Thus an application of the Ergodic Theorem yields

$$\frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} \xrightarrow{p} E\left(y_t y_{t-k}\right)$$

Hence,

$$\hat{\gamma}(k) \xrightarrow{p} E\left(y_t y_{t-k}\right) - \mu^2 - \mu^2 + \mu^2 = E\left(y_t y_{t-k}\right) - \mu^2 = \gamma(k)$$

A covariance stationary process is said to be ergodic for the mean if [1.1.6] converges in probability to $E(y_t)$ as $T \to \infty$. Alternatively, if the autocovariance for a covariance-stationary process satisfy $\sum_{j=1}^{\infty}\left|\gamma_j\right| < \infty$, then $\{y_t\}$ is ergodic for the mean.

Lp can be defined as the set of random variables **X** with $E|X|^p < \infty$. When p is two, the set of random variables X satisfying $E\left(X^2\right) < \infty$ are said to be square integrable. The set of square integrable real random variable X is a normed linear space $\mathbb{R}$ with norm $\|X\| = \left[E\left(X^2\right)\right]^{1/2}$.

Definition 1.1.4 *Two variables* x and y *such that* $E\left(x^2\right) < \infty$ *and* $E\left(y^2\right) < \infty$ *are said to be orthogonal with respect L2 if* $E(xy) = 0$.

Definition 1.1.5 *A sequence of variables* x_n, *satisfying* $E\left(x^2\right) < \infty$ *converges to a variable* x *with respect L2 (or converges in mean square) if* $\|x_n - x\| \to 0$ as $n \uparrow \infty$.

The function $\gamma : k \to \gamma(k)$, *k being integer, is called autocovariance function. This function is*

i. Even: that is, $\forall k, \quad \gamma(-k) = \gamma(k);$

ii. Positive since $\sum_{j=1}^{n}\sum_{k=1}^{n}\phi_j\phi_k\gamma\left(t_j - t_k\right) = Var\left(\sum_{j=1}^{n}\phi_j x_{t_j}\right) > 0$, $\forall$ positive integer n, $\forall\phi_j, \phi_k$ real, and $\forall t_j$ integer.

Theorem 1.1.4 *If* x_t *is a stationary process, and if* $(\phi_i, i \text{ integer})$ *forms a sequence of absolutely summable real numbers with* $\sum_{i=-\infty}^{\infty}\left|\phi_i\right| < \infty$, *the new variable obtained by* $y_t = \sum_{i=-\infty}^{\infty}\phi_i x_{t-i}$ *defines a new stationary process.*

Proof: The series formed by $\phi_i x_{t-i}$ is convergent in mean square since

$$\sum_{i=-\infty}^{\infty} \left\| \phi_i x_{t-i} \right\| = \sum_{i=-\infty}^{\infty} \left| \phi_i \right| \left\| x_{t-i} \right\| = \left(\gamma(0) + \mu^2 \right)^{1/\,2} \sum_{i=-\infty}^{\infty} \left| \phi_i \right| < \infty$$

The expression $\sum_{i=-\infty}^{\infty} \phi_i x_{t-i}$ is an element of L$_2$, y_{t} so defined is square integrable. The moment of $\{y_t\}$ can be written as

$$E(y_t) = E\left(\sum_{i=-\infty}^{\infty} \phi_i x_{t-i} \right) = \sum_{i=-\infty}^{\infty} \phi_i E(x_{t-i}) = \mu_x \sum_{i=-\infty}^{\infty} \phi_i = \mu_y \quad \text{(Independent of } t\text{).}$$

$$Cov(y_t, y_{t+k}) = \operatorname{cov}\left(\sum_{i=-\infty}^{\infty} \phi_i x_{t-i}, \sum_{i=-\infty}^{\infty} \phi_i x_{t+k-i} \right) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \phi_i \phi_j \operatorname{cov}(x_{t-i}, x_{t+k-j})$$

$$= \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \phi_i \phi_j \gamma_x(k+i-j) = \gamma_y(k)$$

(Independent of *t*).

White Noise Process:

The building block for all the process considered in this chapter is a sequence $\{\epsilon_t\}_{t=-\infty}^{\infty}$ whose elements have mean zero and variance σ^2,

$$\left.\begin{aligned} E(\epsilon_t) &= 0 \\ E(\epsilon_t^2) &= \sigma^2 \end{aligned}\right\} \qquad [1.1.7]$$

And for which the ϵ 's are uncorrelated across time:

$$E(\epsilon_t \epsilon_\tau) = 0 \text{ for } t \neq \tau \qquad [1.1.8]$$

A process satisfying [1.1.7] and [1.1.8] is described as a white noise process. One may on occasion wish to replace [1.1.8] with slightly stronger condition that the ϵ 's are independent across time:

$$\epsilon_t \text{ and } \epsilon_\tau \text{ are independent for } t \neq \tau. \qquad 1.1.9]$$

A process satisfying [1.1.7] and [1.1.9] is called an independent white noise process. Finally, if [1.1.7] and [1.1.9] hold along with

$$\epsilon_t \sim N(0, \sigma^2), \qquad [1.1.10]$$

then we have the Gaussian white noise process.

1.2 *Autoregressive and Moving Average Processes*

In time series, the series $\{\ldots,y_1,y_2,\ldots,y_T,\ldots\}$ are jointly random. We consider the conditional expectation $E\left(y_t\,|\mathcal{F}_{t-1}\right)$ where $\mathcal{F}_{t-1}=\{y_{t-1},y_{t-2},\ldots\}$is the history of the series or the information set at time *t-1.* An autoregressive (AR) model specifies that only a finite number of past lags matter.

$$E\left(y_t\,|\mathcal{F}_{t-1}\right)=E\left(y_t\,|\,y_{t-1},y_{t-2,}\ldots,y_{t-k}\right)$$

A linear AR model (the most common type used in practice) specifies linearity:

$$E\left(y_t\,|\mathcal{F}_{t-1}\right)=\rho_0+\rho_1 y_{t-1}+\rho_2 y_{t-2}+\cdots+\rho_k y_{t-k}$$

Letting $e_t=y_t-E\left(y_t\,|\mathcal{F}_{t-1}\right)$, then we have the autoregressive model

$$y_t=\rho_0+\rho_1 y_{t-1}+\rho_2 y_{t-2}+\cdots+\rho_k y_{t-k}+e_t$$
$$E\left(e_t\,|\mathcal{F}_{t-1}\right)=0$$

The last property defines a special time-series process.

Definition 1.1.6 e_t *is a martingale difference sequence (MDS) if* $E\left(e_t\,|\mathcal{F}_{t-1}\right)=0$.

Regression errors are naturally a *MDS*. Some time-series processes may be a *MDS* as a consequence of optimizing behavior. For example, some versions of the life-cycle hypothesis imply that either changes in consumption or consumption growth rates should be a *MDS*. Most asset pricing models imply that asset returns should be the sum of a constant plus a *MDS*.

The *MDS* property for the regression error plays the same role in a time-series regression as does the conditional mean-zero property for the regression error in a cross-section regression. In fact, it is even more important in the time-series context, as it is difficult to derive distribution theories without this property. A useful property of a *MDS* is that e_t is uncorrelated with any function of the lagged information $\mathcal{F}_{t-1}$. Thus for *k>0,* $E\left(e_t\,|\mathcal{F}_{t-k}\right)=0$.

1.2.1 *Stationary AR(1) Process*

A *first-order autoregressive process*, denoted by *AR(1)*, satisfy the following difference equation:

$$y_t=\alpha+\phi y_{t-1}+\epsilon_t \qquad [1.2.1]$$

where, $\{\epsilon_t\}$ is white noise sequence. If $|\phi|\geq 1$, the consequences of the ϵ 's for y accumulate rather than die out over time and hence there does not exist a covariance stationary process for y_t with finite variance that satisfy [1.2.1]. In the case where $|\phi| < 1$, there is a covariance stationary process for y_t satisfying [1.2.1]. By repeated back-substitution, [1.2.1] can be expressed as

$$y_t = \alpha/(1-\phi) + \epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \phi^3\epsilon_{t-3}\cdots \qquad [1.2.2]$$

This can be viewed as an *MA(∞)* process. When $|\phi| < 1, \sum_{j=0}^{\infty} |\phi|^j = 1/(1-|\phi\,|)$. The remainder of this discussion of first-order autoregressive process assumes that $|\phi| < 1$. This ensures that the *MA(∞)* representation exists and that the *AR(1)* process is ergodic for the mean.

Theorem 1.2.1 *If* $|\phi| < 1$ *then* y_t *is strictly stationary and ergodic.*

Taking expectation of both sides of [1.2.2] yields the mean of *AR(1)* process given by:

$$E(y_t) = \mu = \alpha/(1-\phi) \qquad [1.2.3]$$

Variance of a stationary AR(1) process is

$$\begin{aligned}\gamma(0) = E(y_t-\mu)^2 &= E\left(\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \phi^3\epsilon_{t-3}\cdots\right)^2 \\ &= \left(1+\phi^2+\phi^4+\phi^6+\cdots\right)\sigma^2 = \sigma^2/\left(1-\phi^2\right)\end{aligned} \qquad [1.2.4]$$

Covariance at lag *k* of a stationary AR(1) process is

$$\begin{aligned}\gamma(k) &= E(y_t-\mu)(y_{t-k}-\mu) = \\ &= E\begin{bmatrix}\left(\epsilon_t + \phi\epsilon_{t-1} + \phi^2\epsilon_{t-2} + \phi^3\epsilon_{t-3}\cdots + \phi^{\kappa}\epsilon_{t-k} + \phi^{\kappa+1}\epsilon_{t-k-1} + \phi^{\kappa+2}\epsilon_{t-k-2} + \cdots\right) \\ \left(\epsilon_{t-k} + \phi\epsilon_{t-k-1} + \phi^2\epsilon_{t-k-2} + \cdots\right)\end{bmatrix} \\ &= \left(\phi^{\kappa} + \phi^{\kappa+2} + \phi^{\kappa+4} + \cdots\right)\sigma^2 = \left[\phi^k/\left(1-\phi^2\right)\right]\sigma^2\end{aligned} \qquad [1.2.5]$$

From [1.2.4] and [1.2.5] it follows that the autocorrelation function at lag *k* of a stationary AR(1) is

$$\rho(k) = \gamma(k)/\gamma(0) = \phi^{\kappa} \qquad [1.2.6]$$

The moments for a stationary *AR(1)* process were derived above by viewing it as an *MA(∞)* process. The second way to arrive at the same results is to assume that the process is covariance- stationary and calculate the moments directly from [1.2.1]. Taking expectations of both sides of [1.2.1],

$$E(y_t) = \alpha + \phi E(y_{t-1}) + E(\epsilon_t) \qquad [1.2.7]$$

Assuming that the process is covariance stationary,

$$E(y_t) = E(y_{t-1}) = \mu \qquad [1.2.8]$$

Substituting [1.2.8] into [1.2.7]

$$\mu = \alpha + \phi\mu + 0 \Rightarrow \mu = \alpha/(1-\phi) \qquad [1.2.9]$$

To find the second moment of y_t in analogous manner, use [1.2.9] to rewrite [1.2.1] as

$$\left.\begin{array}{l} y_t = \mu(1-\phi) + \phi y_{t-1} + \epsilon_t \\ (y_t - \mu) = \phi(y_{t-1} - \mu) + \epsilon_t \end{array}\right\} \qquad [1.2.10]$$

Now square both sides of [1.2.10] and take expectations

$$E(y_t - \mu)^2 = \phi^2 E(y_{t-1} - \mu)^2 + 2\phi E\left[(y_{t-1} - \mu)\epsilon_t\right] + E(\epsilon_t^2) \qquad [1.2.11]$$

Recall from [1.2.2] that $(y_{t-1} - \mu)$ is a linear function of $\epsilon_{t-1}, \epsilon_{t-2}, \cdots$:

$$(y_{t-1} - \mu) = \epsilon_{t-1} + \phi\epsilon_{t-2} + \phi^2\epsilon_{t-3} + \cdots$$

But ϵ_t is uncorrelated with $\epsilon_{t-1}, \epsilon_{t-2}, \cdots$, so ϵ_t must be uncorrelated with $(y_{t-1} - \mu)$. Hence, the second term on the right hand side of [1.2.11] is zero:

$$E\left[(y_{t-1} - \mu)\epsilon_t\right] = 0 \qquad [1.2.12]$$

Again, assuming covariance-stationarity, we have

$$E(y_t - \mu)^2 = E(y_{t-1} - \mu)^2 = \gamma_0 \qquad [1.2.13]$$

Substituting [1.2.12] and [1.2.13] into [1.2.11],

$$\gamma_0 = \phi^2\gamma_0 + \sigma^2 \Rightarrow \gamma_0 = \sigma^2/(1-\phi^2)$$

Similarly, we could multiply both sides of [1.2.10] by $(y_{t-k} - \mu)$ and take expectations:

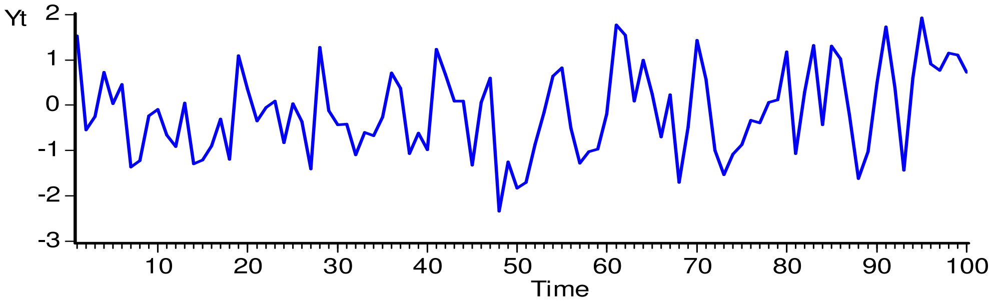
$$E\left[(y_t - \mu)(y_{t-k} - \mu)\right] = \phi E\left[(y_{t-1} - \mu)(y_{t-k} - \mu)\right] + E\left[(y_{t-k} - \mu)\epsilon_t\right] \qquad [1.2.14]$$

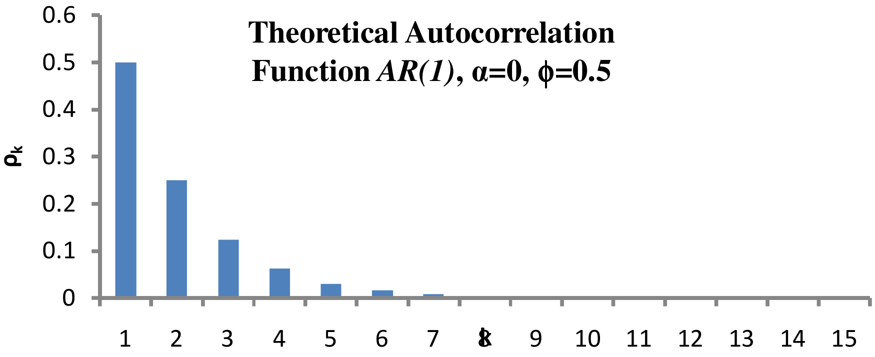
But the term $(y_{t-k} - \mu)$ is a linear function of $\epsilon_{t-k}, \epsilon_{t-k-1}, \epsilon_{t-k-2}, \cdots$, which, for *k>0*, will be uncorrelated with ϵ_t. Thus, for k>0, the last term on the right hand side of [1.2.14] is zero. Hence, for *k>0*, [1.2.14] becomes:

$$\gamma(k) = \phi\gamma(k-1) \Rightarrow \gamma(k) = \phi^{\kappa}\gamma(0) \qquad [1.2.15]$$

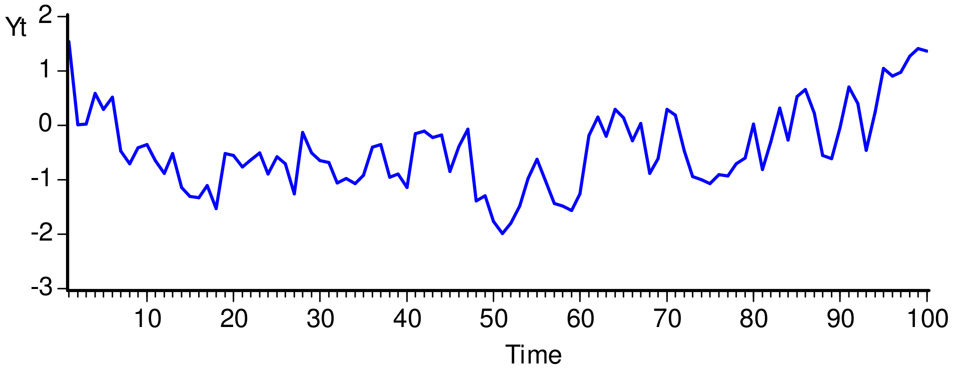
The data for the simulated *AR(1)* processes with parameter ϕ equal to 0.5 and 0.9 are depicted in the following figures, combined with their autocorrelation function. All series are standardized to have unit variance and zero mean. If we compare the *AR* series with ϕ=0.5 and ϕ=0.9, it appears that the latter process is smoother, that is, a higher degree of persistence. The autocorrelation function show an exponential decay in both cases, although it takes large lags for the *ACF* of the ϕ=0.9 series to become close to zero. For example, after 15 periods, the effect of a shock is still $0.9^{15} = 0.21$ of its original effect. For ϕ=0.5 series, the effect at lag 15 is virtually zero.

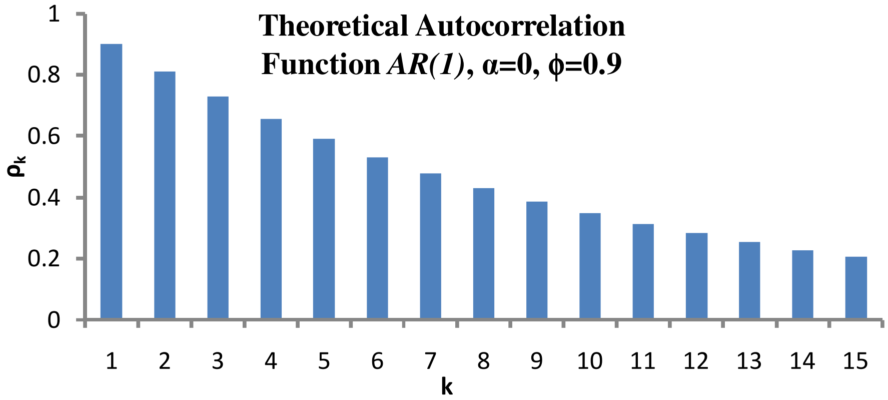
_AR(1)_ with α=0 and ϕ=0.5




Theoretical Autocorrelation Function _AR(1)_, α=0, ϕ=0.5

_AR(1)_ with α=0 and ϕ=0.9





1.2.2 *The Second-Order Autoregressive Process*

A *second-order Autoregression*, denoted by *AR(2)*, satisfies:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t \qquad [1.2.16]$$

Or, in the lag operator notation,

$$\left(1 - \phi_1 L - \phi_2 L^2\right) y_t = \alpha + \epsilon_t \qquad [1.2.17]$$

The difference equation [1.2.16] is stable provided that the roots of

$$\left(1 - \phi_1 z - \phi_2 z^2\right) = 0 \qquad [1.2.18]$$

lie outside the unit circle. When this condition is satisfied, the *AR(2)* process turns out to be covariance-stationary, and the inverse of the autoregressive operator in [1.2.17] is given by

$$\psi(L) = \left(1 - \phi_1 L - \phi_2 L^2\right)^{-1} = \psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \cdots. \qquad [1.2.19]$$

The value of $\psi_j{}'s$ can be found from the fact that

$$\left(1 - \phi_1 L - \phi_2 L^2\right)\left(\psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \cdots.\right) = 1.$$

From this it follows that $\psi_0 = 1; \psi_1 = \phi_1; \psi_j = \phi_1 \psi_{j-1} + \phi_2 \psi_{j-2}$ for $j \geq 2$.

Multiplying both sides of [1.2.17] by ψ(L) gives

$$y_t = \psi(L)\alpha + \psi(L)\epsilon_t \qquad [1.2.20]$$

One can easily show that

$$\psi(L)\alpha = \alpha / \left(1 - \phi_1 - \phi_2\right) \qquad [1.2.21]$$

$$\text{and } \sum_{j=0}^{\infty} \left|\psi_j\right| < \infty \qquad [1.2.22]$$

Proof: taking expectation of both sides of [1.2.16] yields

$$E\left(y_t\right) = \alpha + \phi_1 E\left(y_{t-1}\right) + \phi_2 E\left(y_{t-2}\right)$$

Assuming that the process is covariance stationary, we obtain

$$E(y_t) = \mu = \alpha/(1-\phi_1-\phi_2)$$

Taking expectation of both sides of [1.2.20] yields

$$E(y_t) = \psi(L)\alpha$$

The preceding two expression imply that

$$\psi(L)\alpha = \alpha/(1-\phi_1-\phi_2)$$

This proves the result in [1.2.21] for a covariance-stationary second-order autoregressive process.

Proof of [1.2.22] proceeds as follows:

Factorizing a second-order polynomial in the lag operator,

$$(1-\phi_1 L-\phi_2 L^2)^{-1} = \left[(1-\lambda_1 L)(1-\lambda_2 L)\right]^{-1} = \frac{c_1}{1-\lambda_1 L} + \frac{c_2}{1-\lambda_2 L}$$

Where $c_1 = \dfrac{\lambda_1}{\lambda_1-\lambda_2}$ and $c_2 = \dfrac{\lambda_2}{\lambda_2-\lambda_1}$

Assuming that the eigenvalues are inside the unit circle or the roots of the polynomial in the lag operator lie outside the unit circle, one would express the second order polynomial in the lag operator as

$$\begin{aligned}(1-\phi_1 L-\phi_2 L^2)^{-1} &= \left[(1-\lambda_1 L)(1-\lambda_2 L)\right]^{-1} = \frac{c_1}{1-\lambda_1 L} + \frac{c_2}{1-\lambda_2 L}\\ &= c_1\sum_{j=0}^{\infty}(\lambda_1 L)^j + c_2\sum_{j=0}^{\infty}(\lambda_2 L)^j\\ &= \sum_{j=0}^{\infty}\left(c_1\lambda_1^j + c_2\lambda_2^j\right)L^j\end{aligned}$$

From this last expression it follows that $\psi_j = c_1\lambda_1^j + c_2\lambda_2^j$. Hence,

$$\sum_{j=0}^{\infty}|\psi_j| = \sum_{j=0}^{\infty}\left|c_1\lambda_1^j + c_2\lambda_2^j\right| \le \sum_{j=0}^{\infty}|c_1||\lambda_1|^j + \sum_{j=0}^{\infty}|c_2||\lambda_2|^j = \frac{|c_1|}{1-|\lambda_1|} + \frac{|c_2|}{1-|\lambda_2|} < \infty$$

This proves the expression in [1.2.22].

To find the second moments, write [1.2.16] as

$$y_t = \mu(1-\phi_1-\phi_2) + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

Or $(y_t-\mu) = \phi_1(y_{t-1}-\mu) + \phi_2(y_{t-2}-\mu) + \epsilon_t$ [1.2.23]

Multiplying both sides of [1.2.23] by $(y_{t-k}-\mu)$ and taking expectations produces

$$\gamma(k) = \phi_1\gamma(k-1) + \phi_2\gamma(k-2) \quad \text{for } k = 1,2,3,\cdots \qquad [1.2.24]$$

Hence, the autocovariances follow the same second-order difference equation as does the process for y_t. An *AR(2)* process is covariance stationary if the roots of the second-order polynomial in the lag operator lie outside the unit circle. If both roots are real and lie outside the unit circle, the autocovariance function $\gamma(k)$ is the sum of two decaying exponential functions of *k*. When both roots are complex and their modulus lie outside the unit circle, the autocovariance function $\gamma(k)$ is a damped sinusoidal function.

The autocorrelations are found by dividing both sides of [1.2.24] by $\gamma(0)$:

$$\rho(k) = \phi_1 \rho(k-1) + \phi_2 \rho(k-2) \quad \text{for } k = 1, 2, 3, \cdots \qquad [1.2.25]$$

In particular, setting *k=1* yields

$$\rho(1) = \phi_1 + \phi_2 \rho(1)$$

$$\text{Or } \rho(1) = \phi_1 / (1 - \phi_2) \qquad [1.2.26]$$

For *k=2*, $\rho(2) = \phi_1 \rho(1) + \phi_2$ [1.2.27]

Variance of a covariance stationary second-order autoregressive process can be found by multiplying both sides of [1.2.23] by $(y_t - \mu)$ and taking expectations:

$$E(y_t - \mu)^2 = \phi_1 E\left[(y_{t-1} - \mu)(y_t - \mu)\right] + \phi_2 E\left[(y_{t-2} - \mu)(y_t - \mu)\right] + E\left[\epsilon_t (y_t - \mu)\right]$$

$$\text{Or } \gamma(0) = \phi_1 \gamma(1) + \phi_2 \gamma(2) + \sigma^2 \qquad [1.2.28]$$

Equation [1.2.28] can be written as:

$$\gamma(0) = \phi_1 \rho(1)\gamma(0) + \phi_2 \rho(2)\gamma(0) + \sigma^2 \qquad [1.2.29]$$

Substituting [1.2.26] and [1.2.27] into [1.2.29] gives

$$\gamma(0) = \frac{(1 - \phi_2)\sigma^2}{(1 + \phi_2)\left[(1 - \phi_2)^2 - \phi_1^2\right]}$$

1.2.3 *The pth-order Autoregressive process*

A p^{th}-*order Autoregression*, denoted by *AR(p)*, satisfies

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \cdots + \phi_p y_{t-p} + \epsilon_t \qquad [1.2.30]$$

Provided that the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \phi_3 z^3 - \cdots - \phi_p z^p = 0 \qquad [1.2.31]$$

all lie outside the unit circle, one can easily verify that a covariance stationary representation of the form

$$y_t = \mu + \psi(L)\epsilon_t \qquad [1.2.32]$$

exits where

$$\psi(L) = \left(1 - \phi_1 L - \phi_2 L^2 - \phi_3 L^3 - \cdots - \phi_p L^p\right)^{-1} \text{ and } \sum_{\mathrm{j}=0}^{\infty} |\psi_{\mathrm{j}}| < \infty.$$

Theorem 1.3.1 *The AR(p) process is strictly stationary and ergodic if and only if* $|\lambda_k| > 1$*for all k.*

Assuming that the stationarity condition is satisfied, one alternative way to find the mean is to take expectations of [1.2.30]:

$$\mu = \alpha/\left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right) \qquad [1.2.33]$$

Using [1.2.33], one can easily rewrite equation [1.2.30] as

$$\left(y_t - \mu\right) = \phi_1\left(y_{t-1} - \mu\right) + \phi_2\left(y_{t-2} - \mu\right) + \cdots + \phi_p\left(y_{t-p} - \mu\right) + \epsilon_t \qquad [1.2.34]$$

One can easily find autocovariances by multiplying both sides of [1.2.34] by $\left(y_{t-k} - \mu\right)$and taking expectations:

$$\gamma(k) = \begin{cases} \phi_1\gamma(k-1) + \phi_2\gamma(k-2) + \cdots + \phi_p\gamma(k-p) & \text{for } k = 1,2,3,\cdots \\ \phi_1\gamma(1) + \phi_2\gamma(2) + \cdots + \phi_p\gamma(p) + \sigma^2 & \text{for } k = 0 \end{cases} \qquad [1.2.35]$$

Using the fact that $\gamma(-k) = \gamma(k)$, the system of equations in [1.2.35] can be solved for γ_0, γ_1, … , γ_p as functions of σ^2, ϕ_1, ϕ_2, … , ϕ_p. It can be shown that the $(p\times 1)$vector $\left(\gamma(0), \gamma(1), \cdots, \gamma(p-1)\right)'$is given by the first p elements of the first column of the $\left(p^2\times p^2\right)$ matrix $\sigma^2\left[I_{p^2} - (F\otimes F)\right]^{-1}$where F is $(p\times p)$matrix defined as

$$F = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

and $\otimes$ indicates the Kronecker product.

Dividing [1.2.35] by $\gamma(0)$produces the *Yule-Walker* equations:

$$\rho(k) = \phi_1\rho(k-1) + \phi_2\rho(k-2) + \cdots + \phi_p\rho(k-p) \quad \text{for } k = 1,2,3,\cdots \qquad [1.2.36]$$

Hence, the autocovariances and the autocorrelations follow the same p^{th}-order difference equation as does the process [1.2.30] itself. For district roots, their solutions take the form

$$\gamma(k) = g_1\lambda_1^k + g_2\lambda_2^k + \cdots + g_p\lambda_p^k \qquad [1.2.37]$$

Where the eigenvalues $(\lambda_1, \lambda_2, \cdots, \lambda_p)$ are the solutions to

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_p = 0 .$$

1.2.4 *Stationary MA(1) process*

Let $\{\epsilon_t\}$ be a white noise process as in [1.1.7] and [1.1.8], and consider the process

$$y_t = \mu + \epsilon_t + \theta\epsilon_{t-1} \qquad [1.2.38]$$

where μ and θ could be any constants. This time series process is called a *first-order moving average process*, denoted by *MA(1)*. The term "moving average" comes from the fact that y_t is constructed from a weighted sum of two most recent values of ϵ . The expectation of y_t is given by

$$E(y_t) = E(\mu + \epsilon_t + \theta\epsilon_{t-1}) = \mu + E(\epsilon_t) + \theta E(\epsilon_{t-1}) = \mu \qquad [1.2.39]$$

We used the symbol μ for the constant term in [1.2.38] in anticipation of the result that this constant term turns out to be the mean of the process. The variance of y_t is:

$$\begin{aligned} E(y_t - \mu)^2 = E(\epsilon_t + \theta\epsilon_{t-1})^2 &= E(\epsilon_t^2) + 2\theta E(\epsilon_t\epsilon_{t-1}) + \theta^2 E(\epsilon_{t-1}^2) \\ &= (1+\theta^2)\sigma^2 \end{aligned} \qquad [1.2.40]$$

The autocovariance at lag 1 is

$$E\left[(y_t - \mu)(y_{t-1} - \mu)\right] = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2})\right] = \theta\sigma^2 \qquad [1.2.41]$$

Autocovariance at lags larger than one are all zero:

$$E\left[(y_t - \mu)(y_{t-k} - \mu)\right] = E\left[(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-k} + \theta\epsilon_{t-k-1})\right] = 0 \text{ for all } k > 1. \qquad [1.2.42]$$

Since the mean and autocovariances are not functions of time, an *MA(1)* process is covariance-stationary regardless of the magnitude of θ. Furthermore, an *MA(1)* process satisfies the condition that $\sum_{k=0}^{\infty} |\gamma(k)| = (1+\theta^2)\sigma^2 + |\theta\sigma^2| < \infty$.

Hence, if $\{\epsilon_t\}$ is Gaussian white noise process, then the *MA(1)* process in [1.2.38] is ergodic for all moments.

The *autocorrelation* at lag *k* of a covariance-stationary process (denoted by ρ(*k*)) is defined as its autocovariance at lag k divided by the variance:

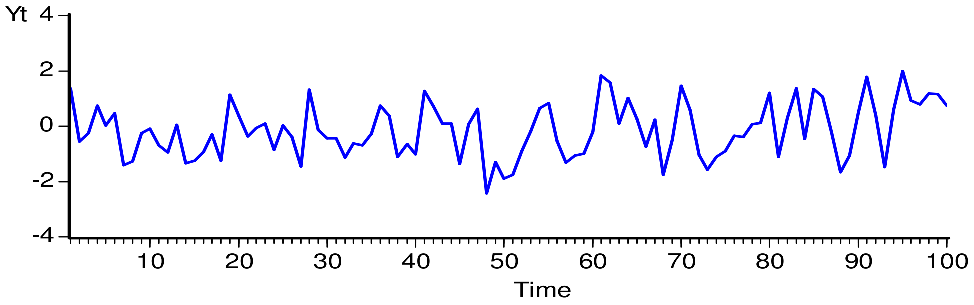
$$\rho(k) = \gamma(k)/\gamma(0) \qquad [1.2.43]$$

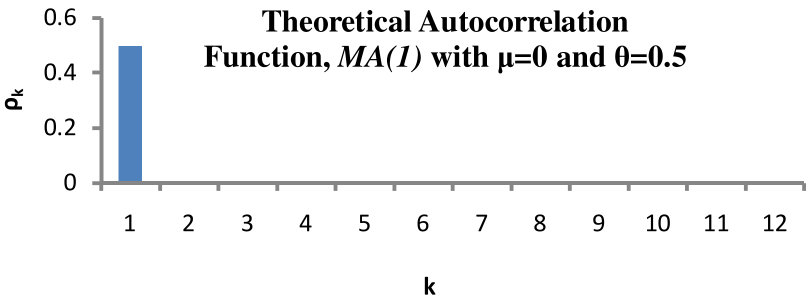
Notice also that the autocorrelation at lag 0 is equal to unity for any covariance-stationary process by definition. From [1.2.40] and [1.2.41], the first autocorrelation for an *MA(1)* process is given by:

$$\rho(1) = \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{(1+\theta^2)} \qquad [1.2.44]$$

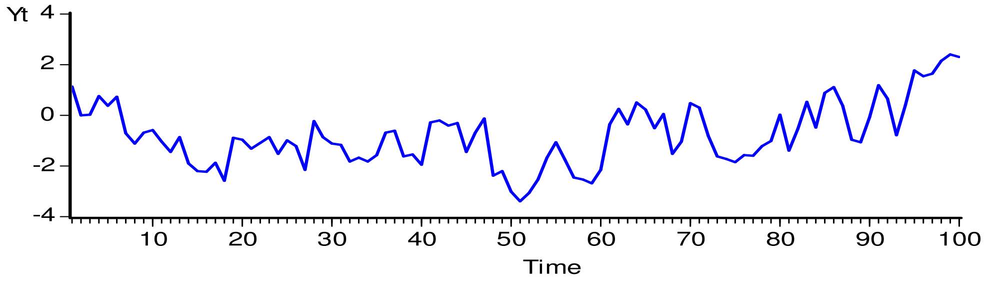
Higher order autocorrelations are all zero.

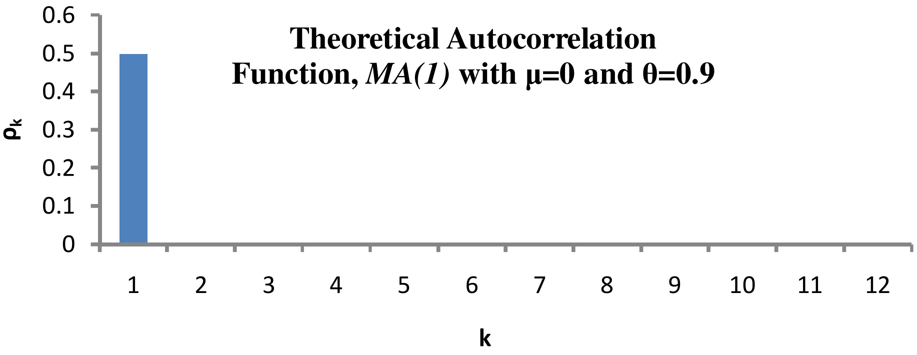
***MA(1)* with μ=0 and θ=0.5**




Theoretical Autocorrelation Function, *MA(1)* with μ=0 and θ=0.5

***MA(1)* with μ=0 and θ=0.9**





1.2.5 *Stationary MA(q) process*

A q^{th}-*order moving average process*, denoted by *MA(q)*, is characterized by

$$y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q} \qquad [1.2.45]$$

where $\{\epsilon_t\}$ satisfy [1.1.7] and [1.1.8] and (θ_1, θ_2, …, θ_q) could be any real numbers. The mean of the process in [1.2.45] is given by:

$$E(y_t) = \mu + E(\epsilon_t) + \theta_1 E(\epsilon_{t-1}) + \theta_2 E(\epsilon_{t-2}) + \cdots + \theta_q E(\epsilon_{t-q}) = \mu \qquad [1.2.46]$$

The variance of an *MA(q)* process is:

$$\begin{aligned}\gamma(0) &= E(y_t - \mu)^2 = E(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q})^2 \\ &= (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2\end{aligned} \qquad [1.2.47]$$

For k=1,2, …, q,

$$\begin{aligned}\gamma(k) &= E\left[(\epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + \cdots + \theta_q\epsilon_{t-q})(\epsilon_{t-k} + \theta_1\epsilon_{t-k-1} + \theta_2\epsilon_{t-k-2} + \cdots + \theta_q\epsilon_{t-k-q})\right] \\ &= E\left[\theta_k\epsilon_{t-k}^2 + \theta_{k+1}\theta_1\epsilon_{t-k-1}^2 + \theta_{k+2}\theta_2\epsilon_{t-k-2}^2 + \cdots + \theta_q\theta_{q-k}\epsilon_{t-q}^2\right]\end{aligned}$$

[1.2.48]

Terms involving ϵ 's at different dates have been dropped because their product has expectation zero, and θ_0 is defined to be unity. For $k>q$, there are no ϵ 's with common dates in the definition of $\gamma(k)$, and so the expectation is zero. Hence,

$$\gamma(k) = \begin{cases}\left[\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 + \cdots + \theta_q\theta_{q-k}\right]\sigma^2 & \text{for } k = 1,2,\cdots,q \\ 0 & \text{for } k > q\end{cases} \qquad [1.2.49]$$

For any values of (θ_1, θ_2, …, θ_q), the MA(q) process is thus covariance-stationary as $\sum_{k=0}^{\infty}|\gamma(k)| < \infty$ and furthermore, for Gaussian ϵ_t the *MA(q)* process is covariance-stationary and also ergodic for all moments. The autocorrelation function is zero after q lags.

1.2.6 *The Infinite-Order Moving Average Process*

The *MA(q)* process can be written as

$$y_t = \mu + \sum_{j=0}^{q} \theta_j \epsilon_{t-j} \text{ with } \theta_0 = 1 .$$

Consider the process that result as $q\uparrow\infty$:

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} = \mu + \psi_0\epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \cdots \quad \text{with } \psi_0 = 1 \qquad [1.2.50]$$

This is described as an *MA(∞)* process. The infinite sequence in [1.2.50] generates a well defined covariance-stationary process provided that

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty \qquad [1.2.51]$$

It is often convenient to work with a slightly stronger condition than [1.2.51] given by:

$$\sum_{j=0}^{\infty} \left|\psi_j\right| < \infty \qquad [1.2.52]$$

A sequence of numbers $\left\{\psi_j\right\}_{j=0}^{\infty}$ satisfying [1.2.51] is said to be *square summable*, whereas a sequence satisfying [1.2.52] is said to be *absolutely summable*. Absolute summability implies square summability, but the converse does not hold.

Proof of the result in [1.2.51]:

We need to show that square-summability of a moving average coefficients implies that the *MA*(∞) representation in [1.2.50] generates a mean square convergent random variable. For a stochastic process such as [1.2.50], the question is whether $\sum_{j=0}^{T} \psi_j \epsilon_{t-j}$ converges in mean square to some random variable $y_t$as T↑∞. To prove this we use the Cauchy Criterion for a stochastic process which states that $\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$ converges if and if, for any $\varepsilon > 0$, there exists a suitably large integer *N* such that for any integer $M > N$

$$E\left[\sum_{j=0}^{M} \psi_j \epsilon_{t-j} - \sum_{j=0}^{N} \psi_j \epsilon_{t-j}\right]^2 < \varepsilon \qquad [1.2.53]$$

Now, the expression on the left hand side of [1.2.53] can written

$$E\left[\psi_M \epsilon_{t-M} + \psi_{M-1}\epsilon_{t-M+1} + \cdots + \psi_{N+1}\epsilon_{t-N-1}\right]^2 = \left(\psi_M^2 + \psi_{M-1}^2 + \cdots + \psi_{N+1}^2\right)\sigma^2$$

$$= \left[\sum_{\mathrm{j}=0}^{M} \psi_{\mathrm{j}}^2 - \sum_{\mathrm{j}=0}^{N} \psi_{\mathrm{j}}^2\right]\sigma^2 \qquad [1.2.54]$$

But if $\sum_{j=0}^{\infty}\psi_j^2$ converges as required by [1.2.51], then by the Cauchy Criterion the right hand side of [1.2.54] may be made as small as desired by choice of a suitably large *N*. Hence, the infinite series in [1.2.50] converges in mean square provided that [1.2.51] is satisfied.

Proof of the result that absolute summability implies square summability:

Absolute summability of the moving average coefficients implies square summability. Suppose that $\{\psi_j\}_{j=0}^{\infty}$ is absolutely summable. Then there exists an N <∞ such that $|\psi_j| < 1$for all $j \geq N$. Then

$$\sum_{j=0}^{\infty}\psi_j^2 = \sum_{j=0}^{N-1}\psi_j^2 + \sum_{j=N}^{\infty}\psi_j^2 < \sum_{j=0}^{N-1}\psi_j^2 + \sum_{j=N}^{\infty}|\psi_j|$$

But $\sum_{j=0}^{N-1}\psi_j^2$ is finite, since N is finite, and$\sum_{j=N}^{\infty}|\psi_j|$is finite since $\{\psi_j\}$ is absolutely summable. Hence, $\sum_{j=0}^{\infty}\psi_j^2 < \infty$, establishing the result that [1.2.52] implies [1.2.51].

The mean and variance of an *MA*(∞) process with absolutely summable coefficients can be calculated from a simple extrapolation of the results for an *MA(q)* process:

$$E(y_t) = \lim_{T\to\infty} E(\mu + \psi_0\epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \cdots + \psi_T\epsilon_{t-T}) = \mu \qquad [1.2.55]$$

$$\begin{aligned}\gamma(0) &= E(y_t - \mu)^2 \\ &= \lim_{T\to\infty} E(\psi_0\epsilon_t + \psi_1\epsilon_{t-1} + \psi_2\epsilon_{t-2} + \cdots + \psi_T\epsilon_{t-T})^2 \\ &= \lim_{T\to\infty}(\psi_0^2 + \psi_1^2 + \psi_2^2 + \cdots + \psi_T^2)\sigma^2\end{aligned} \qquad [1.2.56]$$

$$\begin{aligned}\gamma(k) &= E(y_t - \mu)(y_{t-k} - \mu) \\ &= (\psi_k\psi_0 + \psi_{k+1}\psi_1 + \psi_{k+2}\psi_2 + \psi_{k+3}\psi_3 + \cdots)\sigma^2\end{aligned} \qquad [1.2.57]$$

Moreover, an *MA*(∞) process with absolutely summable coefficients has absolutely summable autocovariance:

$$\sum_{k=0}^{\infty}|\gamma(k)| < \infty \qquad [1.2.58]$$

Hence an *MA*(∞) process satisfying [1.2.52] is ergodic for the mean. Proof of this result is as follows:

Write [1.2.57] as $\gamma(k) = \sigma^2\sum_{j=0}^{\infty}\psi_{j+k}\psi_j$.

Then $|\gamma(k)| = \sigma^2 \left| \sum_{j=0}^{\infty} \psi_{j+k}\psi_j \right| \leq \sigma^2 \sum_{j=0}^{\infty} |\psi_{j+k}\psi_j|$

Hence, $\sum_{k=0}^{\infty} |\gamma(k)| \leq \sigma^2 \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} |\psi_{j+k}\psi_j| = \sigma^2 \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} |\psi_{j+k}||\psi_j| = \sigma^2 \sum_{j=0}^{\infty} |\psi_j| \sum_{k=0}^{\infty} |\psi_{j+k}|$

But there exists an $M<\infty$ such that $\sum_{j=0}^{\infty}|\psi_j| < M$, and therefore $\sum_{k=0}^{\infty}|\psi_{j+k}| < M$ for j=0, 1, 2, …, meaning that $\sum_{k=0}^{\infty}|\gamma(k)| < \sigma^2 \sum_{j=0}^{\infty}|\psi_j| \cdot M < \sigma^2 M^2 < \infty$. Hence [1.2.51] holds and the process is ergodic for the mean. If the ϵ 's are Gaussian, then the process is ergodic for all moments.

1.2.7 *Stationary ARMA(p,q) Process*

An *ARMA(p,q)* process includes both autoregressive and moving average terms defined as:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \qquad [1.2.59]$$

Or, in lag operator form,

$$\left(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right) y_t = \alpha + \left(1 + \theta_1 L + \theta_2 L \ + \cdots + \theta_q L^q\right)\epsilon_t \qquad [1.2.60]$$

Provided that the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \qquad [1.2.61]$$

lie outside the unit circle, both sides of [1.2.60] can be dived by $\left(1-\phi_1 L-\phi_2 L^2 - \cdots - \phi_p L^p\right)$ to obtain

$$y_t = \mu + \psi(L)\epsilon_t$$

Where $\psi(L) = \dfrac{\left(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\right)}{\left(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p\right)}$

$\sum_{j=0}^{\infty} |\psi_j| < \infty$ and $\mu = \alpha / \left(1 - \phi_1 - \phi_2 - \cdots - \phi_p\right)$

Hence, stationarity of an *ARMA(p,q)* process depends entirely on the autoregressive parameters $\left(\phi_1, \phi_2, \cdots, \phi_p\right)$ and not on the moving average parameters $\left(\theta_1, \theta_2, \cdots, \theta_q\right)$. It is often convenient to write the *ARMA* process [1.2.59] in terms of deviations from the mean:

$$(y_t - \mu) = \phi_1 (y_{t-1} - \mu) + \phi_2 (y_{t-2} - \mu) + \cdots + \phi_p (y_{t-p} - \mu) + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

[1.2. 62]

Autocovariances are found by multiplying both sides of [1.2.62] by $(y_{t-k} - \mu)$ and taking expectations. For $k > q$, the resulting equations take the form

$$\gamma(k) = \phi_1 \gamma(k-1) + \phi_2 \gamma(k-2) + \cdots + \phi_p \gamma(k-p), \ \text{for } k = q+1, q+2, \cdots \qquad [1.2.63]$$

Hence, after q lags the autocovariance function $\gamma(k)$ (and the autocorrelation function ρ(*k*)) follow the pth-order difference equation governed by the autoregressive parameters. Note that [1.2.63] does not hold for $k \leq q$, owing to the correlation between $\theta_k \epsilon_{t-k}$ and y_{t-k}. Therefore, an *ARMA(p,q)* process will have more complicated autocovariances for lags 1 through q than would the corresponding *AR(p)* process. For $k > q$ with distinct autoregressive roots, the autocovariances will be given by:

$$\gamma(k) = h_1 \lambda_1^k + h_2 \lambda_2^k + \cdots + h_p \lambda_p^k \qquad [1.2.64]$$

This takes the same form as the autocovariances for an *AR(p)* process [1.2.37], though because the initial conditions $(\gamma(0), \gamma(1), \cdots, \gamma(q))$ differ for the *ARMA* and *AR* processes, the parameters h$_k$ in [1.2.64] will not be the same as the parameters g$_k$ in [1.2.37].

There is a potential for overparameterization with *ARMA* processes. Consider, for instance, a simple white noise process

$$y_t = \epsilon_t \qquad [1.2.65]$$

Suppose we multiply both sides of [1.2.65] by $(1-\rho L)$:

$$(1-\rho L) y_t = (1-\rho L) \epsilon_t \qquad [1.2.66]$$

Clearly, if [1.2.65] is a valid parameterization, then so is [1.2.66] for any value of ρ. Hence, [1.2.66] might be described as an *ARMA(1,1)* process with $\phi_1 = \rho$ and $\theta = -\rho$. It is important to avoid such a parameterization. Since any value of ρ in [1.2.66] describes the data equally well, we will obviously get into trouble trying to estimate the parameter ρ in [1.2.66] by maximum likelihood. Moreover, theoretical manipulation based on a representation such as [1.2.66] may overlook key cancellations. If we are using an *ARMA(1,1)* model in which θ_1 is close to $-\phi_1$, then the data might better be modeled as simple white noise process.

A related overparameterization can arise with an *ARM(p,q)* model. Consider factoring the lag polynomial in [1.2.60] as

$$(1-\lambda_1 L)(1-\lambda_2 L)\cdots(1-\lambda_p L)(y_t-\mu)=(1-\eta_1 L)(1-\eta_2 L)\cdots(1-\eta_q L)\epsilon_t \qquad [1.2.67]$$

Assume $|\lambda_i|<1$ for all *i*, so that the process is covariance-stationary. If the autoregressive operator $\left(1-\phi_1 L-\cdots-\phi_p L^p\right)$ and the moving average operator $\left(1+\theta_1 L+\theta_2 L^2+\cdots+\theta_q L^q\right)$ have any roots in common, say $\lambda_i=\eta_j$ for some *i* and *j*, then both sides of [1.2.67] can be divided by (1-λ*_iL*):

$$\prod_{\substack{k=1\\k\neq i}}^{p}(1-\lambda_k L)(y_t-\mu)=\prod_{\substack{k=1\\k\neq j}}^{q}(1-\eta_k L)\epsilon_t \qquad [1.2.68]$$

Or,

$$\left(1-\phi_1^* L-\phi_2^* L^2-\cdots-\phi_{p-1}^* L^{p-1}\right)(y_t-\mu)=\left(1+\theta_1^* L+\theta_2^* L^2+\cdots+\theta_{q-1}^* L^{q-1}\right)\epsilon_t$$

Where $\left(1-\phi_1^* L-\phi_2^* L^2-\cdots-\phi_{p-1}^* L^{p-1}\right)=(1-\lambda_1 L)(1-\lambda_2 L)\cdots(1-\lambda_{i-1} L)(1-\lambda_{i+1} L)\cdots(1-\lambda_p L)$

$$\left(1+\theta_1^* L+\theta_2^* L^2+\cdots+\theta_{q-1}^* L^{q-1}\right)=(1-\eta_1 L)(1-\eta_2 L)\cdots(1-\eta_{j-1} L)(1-\eta_{j+1} L)\cdots(1-\eta_q L)$$

The stationary *ARMA(p,q)* process satisfying [1.2.60] is clearly identical to the stationary *ARMA(p-1, q-1)* process satisfying [1.2.68].

Wold's Decomposition Theorem

All of the covariance-stationary processes considered in this section can be written in the form

$$y_t=\mu+\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j} \qquad [1.2.69]$$

where ϵ_t is the white noise error one would make in forecasting y_t as a linear function of lagged y and where $\sum_{j=0}^{\infty}\psi_j^2<\infty$ with $\psi_0=1$.

Proposition: *(Wold's Decomposition).* Any zero-mean covariance-stationary process y_t can be represented in the form

$$y_t=\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j}+\kappa_t \qquad [1.2.70]$$

where $\psi_0=1$ and $\sum_{j=0}^{\infty}\psi_j^2<\infty$. The term ϵ_t is a white noise and represents the error made in forecasting y_t on the basis of a linear function of lagged y:

$$\epsilon_t=y_t-E\left(y_t\,|\,y_{t-1},y_{t-2},\cdots\right) \qquad [1.2.71]$$

The value of κ_t is uncorrelated with ϵ_{t-j} for any j, though κ_t can be predicted arbitrarily well from a linear function of past values of y:

$$\kappa_t = E\left(\kappa_t \left| y_{t-1}, y_{t-2}, \cdots\right.\right)$$

The term κ_t is called the *linearly deterministic* component of y_t, while $\sum_{j=0}^{\infty}\psi_j\epsilon_{t-j}$ is called the linearly indeterministic component. If $\kappa_t \equiv 0$, then the process is called purely linearly indeterministic.

This proposition was first proved by Wold(1938). The proposition relies on stable second moments of y but makes no use of higher moments. It thus describes only optimal linear forecasts of y.

Finding the Wold representation requires fitting an infinite number of parameters $(\psi_1, \psi_2, \cdots)$ to the data. With a finite number of observations on $(y_1, y_2, \cdots, y_T)$, this will never be possible. As a practical matter, we therefore, need to make some additional assumptions about the nature of $(\psi_1, \psi_2, \cdots)$. A typical assumption is that $\psi(L)$ can be expressed as the ratio of two finite-order polynomials:

$$\sum_{j=0}^{\infty}\psi_j L^j = \frac{\theta(L)}{\phi(L)} = \frac{1+\theta_1 L+\theta_2 L^2+\cdots+\theta_q L^q}{1-\phi_1 L-\phi_2 L^2-\cdots-\phi_p L^p}$$

1.3 *Use of Lag operator*

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 1.3.1 *The* ***lag operator*** *L satisfies* $Ly_t = y_{t-1}$.

Defining $L^2 = LL$, we see that $L^2 y_t = Ly_{t-1} = y_{t-2}$. In general, $L^k y_t = y_{t-k}$. The *AR(1)* model can be written in the format

$$y_t = \alpha + \phi y_{t-1} + \epsilon_t$$

or $$(1 - \phi L) y_t = \alpha + \epsilon_t$$

The operator $\Phi(L) = (1 - \phi L)$ is a polynomial in the lag operator *L*. We say that the *root* of the polynomial is $1/\phi$, since $\Phi(z) = 0$ when $z = 1/\phi$. We call $\Phi(L)$ the autoregressive polynomial of y_t. From Theorem 1.2.1, an *AR(1)* is stationary iff $|\phi| < 1$. Note that an equivalent way to say this is that an *AR(1)* is stationary iff the root of the autoregressive polynomial lies outside the unit circle or larger than one (in absolute value).

We can write a general *AR(p)* model as

$$\Phi(L) y_t = \epsilon_t$$

Where $\Phi(\mathrm{L})$ is a polynomial of order *p* in the lag operator *L*, usually referred to as a ***lag polynomial***, given by

$$\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$$

We can interpret the lag polynomial as a filter that, if applied to a time series, produces a new series. So the filter $\Phi(L)$ applied to an *AR(p)* process y_t produces a white noise process ϵ_t. It is relatively easy to manipulate lag polynomials. For example, transforming a series by two such polynomials one after the other is the same as transforming the series once by a polynomial that is the product of the two original ones. This way we can define the inverse of a filer, which is naturally given by the inverse of the polynomial. Thus the inverse of $\Phi(L)$, denoted by $\Phi^{-1}(L)$, is defined so as to satisfy $\Phi^{-1}(L)\Phi(L) = 1$. If $\Phi(\mathrm{L})$ is a finite-order polynomial in *L*, its inverse will be one of infinite order. For the *AR(1)* case we find

$$(1 - \phi L)^{-1} = \sum_{j=0}^{\infty} \phi^j L^j \qquad \text{provided that } |\phi| < 1. \qquad [1.3.1]$$

This is similar to the result that the infinite sum $\sum_{j=0}^{\infty}\phi^{j}$ equals to $(1-\phi)^{-1}$ if $|\phi|<1$, while it does not converge for $|\phi|\geq 1$. In general, the inverse of a polynomial $\Phi(\mathrm{L})$ exists if it satisfies certain conditions on its parameters, in which case we call $\Phi(\mathrm{L})$ invertible. For example the *AR(1)* model can be written as

$$(1-\phi L)^{-1}(1-\phi L)y_t=(1-\phi L)^{-1}\epsilon_t$$

Or

$$y_t=\sum_{j=0}^{\infty}\phi^{j}L^{j}\epsilon_t=\sum_{j=0}^{\infty}\phi^{j}\epsilon_{t-j} \qquad [1.3.2]$$

which corresponds to (1.2.2) when α=0.

Under appropriate conditions, the converse is also possible and we can write a moving average model in autoregressive form. Using the lag operator, we can write the *MA(1)* process as

$$y_t=(1+\theta L)\epsilon_t$$

and the general MA(q) process as

$$y_t=\theta(L)\epsilon_t,$$

Where

$$\theta(L)=1+\theta_1 L+\theta_2 L^2+\cdots+\theta_q L^q$$

Now, if $\theta^{-1}(L)$ exists, we can write

$$\theta^{-1}(L)y_t=\epsilon_t,$$

which in general, will be an *AR* model with infinite order. For the *MA(1)* case, we use,

$$(1+\theta L)^{-1}=\sum_{j=0}^{\infty}(-\theta)^{j}L^{j},\quad \text{provided that } |\theta|<1. \qquad [1.3.3]$$

Consequently, an *MA(1)* model can be written as

$$y_t=\theta\sum_{j=0}^{\infty}(-\theta)^{j}y_{t-j-1}+\epsilon_t \qquad [1.3.4]$$

A necessary condition for the infinite *AR* representation *(AR(∞))* to exist is that the *MA* polynomial is invertible, which, in the *MA(1)* case, requires that $|\theta|<1$. Particularly for making predictions conditional upon an observed past, the *AR* representations are very convenient. The *MA* representations are often convenience to determine variances and covariances.

For a more parsimonious representation, we may want to work with an *ARMA* model that contains both an autoregressive and moving average part. The general *ARMA* model can be written as

$$\Phi(L)y_t = \theta(L)\epsilon_t,$$

Which (if the AR lag polynomial is invertible) can be written in *MA(∞)* representation as

$$y_t = \Phi^{-1}(L)\theta(L)\epsilon_t$$

or (if the *MA* lag polynomial is invertible) can be written in AR(∞) form as

$$\theta^{-1}(L)\Phi(L)y_t = \epsilon_t$$

Both $\Phi^{-1}(L)\theta(L)$and $\theta^{-1}(L)\Phi(L)$are lag polynomials of infinite lag length, with restrictions on the coefficients.

1.4 *Invertibility of Lag Polynomials*

As we have seen before, the first order lag polynomial $1-\phi L$is invertible if $|\phi|<1$. This condition can be generalized to higher-order lag polynomials. Let us consider the case of a second-order lag polynomial, given by $1-\phi_1 L-\phi_2 L^2$. Generally we can find values λ_1 and λ_2 such that the polynomial can be written as

$$1-\phi_1 L-\phi_2 L^2 = (1-\lambda_1 L)(1-\lambda_2 L) \qquad [1.4.1]$$

It can easily be verified that λ_1 and $\lambda_2$can be solved from $\lambda_1+\lambda_2=\phi_1$ and $\lambda_1\lambda_2=-\phi_2$. The conditions for invertibility of the second-order polynomial are jus the conditions that both the first-order polynomials $1-\lambda_1 L$and $1-\lambda_2 L$are invertible. Thus, the requirement for invertibility is that both $|\lambda_1|<1$ and $|\lambda_2|<1$. These requirements can be formulated in terms of the so-called **characteristic equation**.

$$(1-\lambda_1 z)(1-\lambda_2 z)=0 \qquad [1.4.2]$$

This equation has two solutions, z_1 and z_2, referred to as the **characteristic roots**. The requirement $|\lambda_i|<1$corresponds to$|z_i|>1$. If any solution satisfies$|z_i|\le 1$, the corresponding polynomial is noninvertible. A solution that equals unity is referred to as a **unit root**.

The presence of a unit root in the lag polynomial $\Phi(L)$ can be detected relatively easily, without solving the characteristic equation, by nothing that the polynomial $\Phi(z)$ evaluated at $z = 1$ is zero if $\sum_{j=1}^{p} \phi_j = 1$. Thus, the presence of a first unit root can be verified by checking whether the polynomial coefficients sum to one. If the sum exceeds one, the polynomial is not invertible.

As an example, consider the following *AR(2)* model

$$y_t = 1.2y_{t-1} - 0.32y_{t-2} + \epsilon_t \qquad [1.4.3]$$

The *AR(2)* model in equation [1.4.3] can be written as

$$(1-0.8L)(1-0.4L)y_t = \epsilon_t \qquad [1.4.4]$$

with characteristic equation

$$1-1.2z-0.32z^2 = (1-0.8z)(1-0.4z) = 0 \qquad [1.4.5]$$

The solution (characteristic roots) are $z_1 = 1/0.8 = 1.25$ and $z_2 = 1/0.4 = 2.5$, which are both larger than one. Consequently, the *AR* polynomial in [1.4.3] is invertible. Note that the following AR(1) model

$$y_t = 1.2y_{t-1} + \epsilon_t \qquad [1.4.6]$$

describes a noninvertible AR process.

Invertibility of a lag polynomial is very important for several reasons. For moving average models, or more generally, models with moving average component, invertibility of the MA is important for estimation and prediction. For models with an autoregressive part, the AR polynomial is invertible if and only if the process is stationary.

1.5 *Problem of Common Roots of Lag Polynomials*

Decomposing the moving average and autoregressive polynomials into products of linear functions in *L* also shows the problem of common roots or cancelling roots. This means that the *AR* and *MA* parts of the model have roots that are identical and the corresponding linear functions in *L* cancel out. To illustrate this, consider the model described by

$$\left(1-\phi_1 L-\phi_2 L^2\right)y_t = (1+\theta L)\epsilon_t \qquad [1.5.1]$$

Thus we can write this as

$$(1-\lambda_1 L)(1-\lambda_2 L)y_t = (1+\theta L)\epsilon_t \qquad [1.5.2]$$

Now, if $\theta = -\lambda_1$, we can divide both sides by $(1+\theta L)$ to obtain

$$(1-\lambda_2 L)y_t = \epsilon_t \qquad [1.5.3]$$

which is exactly the same as [1.5.2]. Thus, in the case of one cancelling root, an *ARMA(p,q)* model can be written equivalently as an *ARMA(p-1, q-1)* model.

As an example, the following *ARMA(2,1)* model

$$y_t = y_{t-1} - 0.25y_{t-2} + \epsilon_t - 0.5\epsilon_{t-1} \qquad [1.5.4]$$

which can be rewritten as

$$(1-0.5L)(1-0.5L)y_t = (1-0.5L)\epsilon_t \qquad [1.5.5]$$

This reduces to an *AR(1)* model given by

$$(1-0.5L)y_t = \epsilon_t \qquad [1.5.6]$$

which describes the same process as [1.5.4].

The problem of common roots illustrates why it may be problematic, in practice to estimate an *ARMA* model with an *AR* and an *MA* part of a high order. The reason is that identification and estimation are hard if roots of the *MA* and *AR* polynomials are almost identical. In this case, a simplified *ARMA(p-1, q-1)* model will yield an almost equivalent representation.

1.6 *Autocovariance Generating Function*

For each covariance-stationary process y_t we calculated the sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$. If this sequence is absolutely summable, then one way of summarizing the autocovariances is through a scalar-valued function called the autocovariance-generating function:

$$g_y(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j \qquad [1.6.1]$$

This function is constructed by taking the jth autocovariance and multiplying it by some number z raised to the jth power, and then summing over all the possible values of *j*. The argument of this function is taken to be a complex scalar. Of particular interest as an argument for the autocovariance-generating function is any value of z that lies on the complex unit circle,

$$z = \cos(\omega) - i\sin(\omega) = e^{-i\omega}$$

Where $i = \sqrt{-1}$ and ω is the radian angle that z makes with the real axis. If the autocovariance-generating function is evaluated at $z = e^{-i\omega}$ and divided by 2π, the resulting function of ω,

$$S_y(\omega) = \frac{1}{2\pi} g_y(z) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}$$

is called the population spectrum of y. For a process with absolutely summable autocovariances, the function $S_y(\omega)$ exits and can be used to calculate all of the autocovariances. This means that if two different processes share the same autocovariance-generating function, then the two processes exhibit the identical sequence of autocovariances.

As an example of calculating an autocovariance-generating function, consider the MA(1) process. From equations [1.2.40] to [1.2.42], its autocovariance-generating function is

$$g_y(z) = [\theta\sigma^2]z^{-1} + [(1+\theta^2)\sigma^2]z^0 + [\theta\sigma^2]z^1 = \sigma^2[\theta z^{-1} + (1+\theta^2) + \theta z]$$

This expression could alternatively be written as

$$g_y(z) = \sigma^2(1+\theta z)(1+\theta z^{-1}) \qquad [1.6.2]$$

The form of expression [1.6.2] suggests that for the MA(q) process,

$$y_t = \mu + \left(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q\right)\epsilon_t$$

The autocovariance generating function might be calculated as

$$g_y(z) = \sigma^2\left(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q\right)\left(1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \cdots + \theta_q z^{-q}\right) \qquad [1.6.3]$$

This conjecture can be verified by carrying out the multiplication in [1.6.3] and collecting terms by power of z:

$$\begin{aligned}
&\left(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q\right)\left(1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \cdots + \theta_q z^{-q}\right) \\
&= (\theta_q) z^q + (\theta_{q-1} + \theta_q\theta_1) z^{(q-1)} + (\theta_{q-2} + \theta_{q-1}\theta_1 + \theta_q\theta_2) z^{(q-2)} \\
&\quad + \cdots + (\theta_1 + \theta_2\theta_1 + \theta_3\theta_2 + \cdots + \theta_q\theta_{q-1}) z^1 + (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2) z^0 \\
&\quad + (\theta_1 + \theta_2\theta_1 + \theta_3\theta_2 + \cdots + \theta_q\theta_{q-1}) z^{-1} + \cdots + (\theta_q) z^{-q}
\end{aligned} \qquad [1.6.4]$$

Comparison of [1.6.4] with [1.2.47] or [1.2.49] confirms that the coefficient on z^k in [1.6.3] is indeed the k^th autocovariance.

This method for finding $g_y(z)$ extends to the MA(∞) case. If

$$y_t = \mu + \theta(L)\epsilon_t \qquad [1.6.5]$$

with

$$\theta(L) = \theta_0 + \theta_1 L + \theta_2 L^2 + \cdots \qquad [1.6.6]$$

and

$$\sum_{j=0}^{\infty} |\theta_j| < \infty, \qquad [1.6.7]$$

then

$$g_y(z) = \sigma^2 \theta(z)\theta(z^{-1}) \qquad [1.6.8]$$

For example, the stationary *AR(1)* process can be written as

$$y_t - \mu = (1 - \phi L)^{-1} \epsilon_t$$

which is in the form of [1.6.5] with $\theta(L) = (1 - \phi L)^{-1}$. The autocovariance-generating function for *AR(1)* process could therefore be calculated from

$$g_y(z) = \frac{\sigma^2}{(1 - \phi z)(1 - \phi z^{-1})} \qquad [1.6.9]$$

To verify this claim, expand out the terms in [1.6.9]:

$$\frac{\sigma^2}{(1 - \phi z)(1 - \phi z^{-1})} = \sigma^2 (1 + \phi z + \phi^2 z^2 + \phi^3 z^3 + \cdots)(1 + \phi z^{-1} + \phi^2 z^{-2} + \phi^3 z^{-3} + \cdots),$$

From which the coefficient on z^k is

$$\sigma^2 (\phi^k + \phi^{k+1}\phi + \phi^{k+2}\phi^2 + \cdots) = \sigma^2 \phi^k / (1 - \phi^2).$$

This indeed yields the k$^{\text{th}}$ autocovariance as earlier calculated in equation [1.2.5].

The autocovariance-generating function for a stationary *ARMA(p,q)* process can be written as:

$$g_y(z) = \frac{\sigma^2 (1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q)(1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \cdots + \theta_q z^{-q})}{(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_q z^q)(1 - \phi_1 z^{-1} - \phi_2 z^{-2} - \cdots - \phi_q z^{-q})} \qquad [1.6.10]$$

Filters

Sometimes the data are filtered, or treated in a particular way before they are analyzed, and we would like to summarize the effects of this treatment on the autocovariances. This calculation is particularly simple using the autocovariance-generating function. For example, suppose that the original data y_t were generated from an *MA(1)* process,

$$y_t = (1+\theta L)\epsilon_t \tag{1.6.11}$$

with autocovariance-generating function given by [1.6.2]. Let's say that the data as actually analyzed, x_t, represent the change in y_t over its value the previous period:

$$x_t = y_t - y_{t-1} = (1-L)\,y_t \tag{1.6.12}$$

Substituting [1.6.11] into [1.6.12], the observed data can be characterized as the following *MA(2)* process,

$$x_t = (1-L)(1+\theta L)\epsilon_t = \left[1+(\theta-1)L - \theta L^{-2}\right]\epsilon_t = \left[1+\theta_1 L + \theta_2 L^2\right]\epsilon_t \tag{1.6.13}$$

with $\theta_1 \equiv (\theta-1)$ and $\theta_2 \equiv -\theta$. The autocovariance-generating function of the observed series x_t can be calculated by direct application of [1.6.3]:

$$g_x(z) = \sigma^2\left(1+\theta_1 z + \theta_2 z^2\right)\left(1+\theta_1 z^{-1} + \theta_2 z^{-2}\right) \tag{1.6.14}$$

It is often instructive, however, to keep the polynomial $\left(1+\theta_1 z + \theta_2 z^2\right)$ in its factored form of the first line of [1.6.13],

$$\left(1+\theta_1 z + \theta_2 z^2\right) = (1-z)(1+\theta z),$$

in which case [1.6.14] could be written as

$$\begin{aligned} g_x(z) &= \sigma^2 (1-z)(1+\theta z)\left(1-z^{-1}\right)\left(1+\theta z^{-1}\right) \\ &= (1-z)\left(1-z^{-1}\right) g_y(z) \end{aligned} \tag{1.6.15}$$

Of course, [1.6.14] and [1.6.15] represent the identical function of z, and which way we choose to write it is simply a matter of convenience. Applying the filter $(1-L)$ to y_t thus results in multiplying its autocovariance-generating function by $(1-z)\left(1-z^{-1}\right)$.

This principle readily generalizes. Suppose that the original data series $\{y_t\}$ satisfies [1.6.5] through [1.6.7]. Let's say the data are filtered according to

$$x_t = h(L)\,y_t \tag{1.6.16}$$

with

$$h(L)=\sum_{j=-\infty}^{\infty} h_j L^j \text{ and } \sum_{j=-\infty}^{\infty} |h_j| < \infty.$$

Substituting [1.6.5] into [1.6.16] , the observed data x_t are then generated by

$$x_t = h(1)\mu + h(L)\theta(L)\epsilon_t = \mu^* + \theta^*(L)\epsilon_t,$$

where $\mu^* = h(1)\mu$ and $\theta^*(L) = h(L)\theta(L)$. The sequence of coefficients associated with the compound operator $\{\theta_j^*\}_{j=-\infty}^{\infty}$ turns out to be absolutely summable and the autocovariance-generating function of x_t can accordingly be calculated as

$$g_x(z) = \sigma^2\theta^*(z)\theta^*(z^{-1}) = \sigma^2 h(z)\theta(z)\theta(z^{-1})h(z^{-1}) = h(z)h(z^{-1})g_y(z) \qquad [1.6.17]$$

Applying the filter $h(L)$ to a series thus results in multiplying its autocovariance-generating function by $h(z)h(z^{-1})$.

1.7 *Trend Stationarity*

$$y_t = \mu_0 + \mu_1 t + S_t \qquad [1.7.\ 1]$$

$$S_t = \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-k} + \epsilon_t \qquad [1.7.\ 2]$$

Or

$$y_t = \alpha_0 + \alpha_1 t + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \cdots + \rho_k y_{t-k} + \epsilon_t \qquad [1.7.\ 3]$$

There are two essentially equivalent ways to estimate the autoregressive parameters $(\rho_1, \cdots, \rho_k)$.

- You can estimate (1.7.3) by OLS.
- You can estimate (1.7.1)-(1.7.2) sequentially by OLS. That is, first estimate (1.7.1), get the residual $\hat{S}_t$, and then perform regression (1.7.2) replacing S_t with $\hat{S}_t$. This procedure is sometimes called Detrending.

Seasonal Effects

There are three popular methods to deal with seasonal data.

- Include dummy variables for each season. This presumes that "seasonality" does not change over the sample.
- Use "seasonally adjusted" data. The seasonal factor is typically estimated by a two-sided weighted average of the data for that season in neighboring years. Thus the

seasonally adjusted data is a “filtered” series. This is a flexible approach which can extract a wide range of seasonal factors. The seasonal adjustment, however, also alters the time-series correlations of the data.

- First apply a seasonal differencing operator. If s is the number of seasons (typically $s= 4$ or $s = 12$);

$$\Delta_s y_t = y_t - y_{t-s},$$

or the season-to-season change. The series $\Delta_s y_t$ is clearly free of seasonality. But the long-run trend is also eliminated, and perhaps this was of relevance.

1.8 *Testing for Omitted Serial Correlation*

For simplicity, let the null hypothesis be an $AR(1)$:

$$y_t = \alpha + \rho y_{t-1} + u_t \qquad [1.8.\ 1]$$

We are interested in the question if the error u_t is serially correlated. We model this as an $AR(1)$:

$$u_t = \theta u_{t-1} + e_t \qquad [1.8.\ 2]$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$H_0 : \theta = 0$$

$$H_1 : \theta \neq 0$$

We want to test H_0 against H_1:

To combine (1.8.1) and (1.8.2), we take (1.8.1) and lag the equation once:

$$y_{t-1} = \alpha + \rho y_{t-2} + u_{t-1}$$

We then multiply this by θ and subtract from (1.8.1), to find

$$y_t - \theta y_{t-1} = \alpha - \alpha\theta + \rho y_{t-1} - \theta\rho y_{t-2} + u_t - \theta u_{t-1},$$

Or

$$y_t = \alpha(1-\theta) + (\rho+\theta) y_{t-1} - \theta\rho y_{t-2} + e_t = AR(2)$$

Thus under H_0, y_t is an *AR(1),* and under H_1 it is an *AR(2).* H_0 may be expressed as the restriction that the coefficient on y_{t-2} is zero. An appropriate test of H_0 against H_1 is therefore a Wald test that the coefficient on y_{t-2} is zero (a simple exclusion test).

In general, if the null hypothesis is that y_t is an *AR(k),* and the alternative is that the error is an

AR(m), this is the same as saying that under the alternative y_t is an *AR(k+m),* and this is equivalent to the restriction that the coefficients on $y_{t-k-1},\cdots,y_{t-k-m}$ are jointly zero. An appropriate test is the Wald test of this restriction.

1.9 *Stationarity and Unit Roots*

Stationarity of a stochastic process requires that the variances and autocovariances are finite and independent of time. It is easily verified that finite-order *MA* process are stationary by construction as they correspond to a weighted sum of white noise processes. Of course this result breaks down if we allow the *MA* coefficients to vary over time, as in

$$y_t = \epsilon_t + g(t)\epsilon_{t-1} \qquad [1.9.1]$$

where $g(t)$ is some deterministic function of *t*. Now, we have

$$E\left(y_t^2\right) = \sigma^2 + g^2(t)\sigma^2,$$

which is not independent of *t*. Consequently, the process in [1.9.1] is non-stationary.

Stationarity of autoregressive or *ARMA* processes is less trivial. Consider, for example, the following *AR(1)* process

$$y_t = \phi y_{t-1} + \epsilon_t, \qquad [1.9.2]$$

with $\phi = 1$. Taking variances on both sides gives $\text{var}(y_t) = \text{var}(y_{t-1}) + \sigma^2$, which has no solution for the variance of the process consistent with stationarity, unless $\sigma^2 = 0,$ in which case infinity of solutions exists. The process in [1.9.2] is a first-order autoregressive process with a unit root$(\phi = 1)$, usually referred to as a **random walk without a drift**. The unconditional variance of y_t does not exist, i.e., is infinite and the process is non-stationary. In fact, for any value of ϕ with $|\phi| \geq 1$, [1.9.2] describes a non-stationary process.

We can formalize the above result as follows. The *AR(1)* process is stationary if and only if the polynomial $1-\phi L$ is invertible, i.e., if the roots of the characteristic equation $1-\phi z = 0$ lies outside the unit circle. This result is straightforwardly generalized to arbitrary *ARMA* models. The *ARMA(p,q)* model

$$\psi(L)y_t = \theta(L)\epsilon_t \qquad [1.9.3]$$

corresponds to a stationary process if and only if the solutions $z_1,\cdots,z_p$ to $\psi(z) = 0$ lie outside the unit circle, that is, when the *AR* polynomial is invertible.

For example, the *ARMA(2,1)* process given by

$$y_t = 1.2y_{t-1} - 0.2y_{t-2} + \epsilon_t - 0.5\epsilon_{t-1}, \qquad [1.9.4]$$

is non-stationary because $z = 1$ is a solution to $1 - 1.2z + 0.2z^2 = 0$. A special case that is of particular interest arises when one root is exactly equal to one, while the other roots are larger than one. If this arises, we can write the process for y_t as

$$\psi^*(L)(1-L)y_t = \psi^*(L)\Delta y_t = \theta(L)\epsilon_t, \qquad [1.9.5]$$

where $\psi^*(L)$ is an invertible polynomial in *L* of order *p-1*, and $\Delta \equiv 1 - L$ is the first-difference operator. Because the roots of the *AR* polynomial are the solutions to $\psi^*(z)(1-z) = 0$, there is one solution z=1, or in other words a single unit root. Equation [1.9.5] thus shows that Δy_t can be described as a stationary *ARMA* model if the process for y_t has one unit root. Consequently, we can eliminate the non-stationarity by transforming the series into first-differences. Writing the process in [1.9.4] as

$$(1 - 0.2L)(1 - L)y_t = (1 - 0.5L)\epsilon_t$$

shows that Δy_t is described as a stationary *ARMA(1,1)* process given by

$$\Delta y_t = 0.2\Delta y_{t-1} + \epsilon_t - 0.5\epsilon_{t-1}$$

A series that becomes stationary after first-differencing is said to be **integrated of order one**, denoted by $I(1)$. If Δy_t is described by stationary *ARMA(p,q)* model, we say that y_t is described by an autoregressive integrated moving average (*ARIMA*) model of order *p, 1, q*, or in short an *ARIMA(p,1,q)* model.

First differencing quite often transforms a non-stationary series into a stationary series. In particular this may be the case for aggregate economic series or their natural logarithms. In some cases, taking first-differences is insufficient to produce stationary series and another differencing step is required. In this case the stationary series is given by $\Delta(\Delta y_t) = \Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$. If the series must be differenced twice before it becomes stationary, then it is said to be integrated of order two, denoted by $I(2)$, and it must have two unit roots. Thus, a series y_t is $I(2)$ if Δy_t is non-stationary but $\Delta^2 y_t$ is stationary.

In general, the main difference between $I(0)$ and $I(1)$ processes can be summarized as follows. An $I(0)$ series fluctuates around its mean with a finite variance that does not depend on time, while an $I(1)$ series wanders widely. Typically, it is said that an $I(0)$ series is **mean reverting**, as there is a tendency in the long run to return to its mean. Furthermore, an $I(0)$ series has a limited memory of its past behavior (implying that the effects of a particular random innovation are only transitory), while an $I(1)$ series has infinitely long memory (implying that an innovation will permanently affect the process). This last aspect becomes clear from the autocorrelation functions: for an $I(0)$ series the autocorrelation function declines rapidly as the lag length increases, while for the $I(1)$ process the estimated autocorrelation coefficients decay to zero very slowly.

The last property makes the presence of a unit root an interesting question from an economic point of view. In models with unit roots, shocks (which may be due to policy interventions) have persistent effects that last forever, while, in the case of stationary models, shocks can only have a temporary effect. Of course, the long-run effect of a shock is not necessarily of the same magnitude as the short-run effect. The fact that the autocorrelations of a stationary series die out rapidly may help in determining the number of times differencing is needed to achieve stationarity. In addition, a number of formal unit root test has been proposed in the literature.

1.9.1 Testing for Unit Roots in a First-order Autoregressive Model

Consider the $AR(1)$ process

$$y_t = \mu + \rho y_{t-1} + \epsilon_t \qquad [1.9.6]$$

where $\rho=1$ corresponds to a unit root. It seems obvious to use the estimate $\hat{\rho}$ for ρ from an OLS procedure (which is consistent, irrespective of the true value of ρ) and the corresponding standard error to test the null hypothesis of a unit root. However, as shown in the seminal paper of Dickey and Fuller (1979), under the null hypothesis that $\rho = 1$ the standard *t*-ratio does not have a *t*-distribution, not even asymptotically. The reason for this is the non-stationarity of the process invalidating standard results on the distribution of the OLS estimator $\hat{\rho}$. To test for the presence of a unit root in $AR(1)$ the null and the alternative hypothesis are given by

$$H_0 : \rho = 1$$

$$H_1 : \rho < 1$$

It is possible to use the standard *t*-statistic to test the above hypothesis, where the *t*-statistic is given by

$$\hat{\tau}_{\mu} = \frac{\hat{\rho} - 1}{se(\hat{\rho})} \qquad [1.9.7]$$

Where $se(\hat{\rho})$ denotes the usual OLS standard error. Critical Values, however, have to be taken from the appropriate distribution, which under the null hypothesis of non-stationarity is nonstandard. In particular, the distribution is skewed to the left (with a long left-hand tail) so that critical values are smaller than those for (the normal approximation of) the *t*-distribution. Using a 5% significance level in a one-tailed test of $H_0 : \rho = 1$ (a unit root) against $H_1 : \rho < 1$ (stationary), the correct critical value in large samples is -2.86 rather than -1.65 for the normal approximation.

Usually, a slightly more convenient regression procedure is used, in which case, the model is written as:

$$\Delta y_t = \mu + (\rho - 1) y_{t-1} + \epsilon_t \qquad [1.9.8]$$

From which the $\hat{\tau}_{\mu}$-statistic for $\rho - 1 = 0$ is identical to the $\hat{\tau}_{\mu}$ above. The reason for this is that the least squares method is invariant to linear transformations of the model. Under the null hypothesis of a unit root the above model turns out to be

$$\Delta y_t = \mu + \epsilon_t \qquad [1.9.9]$$

which is known as **a random walk with drift**, where μ is the drift parameter. In the model for the level variable y_t, μ corresponds to a linear time trend as [1.9.9] implies that $E(\Delta y_t) = \mu$. Hence for a given initial value y_0, $E(y_t) = y_0 + \mu t$. This shows that the interpretation of the intercept term in [1.9.9] depends on the presence of a unit root. In the stationary case, μ reflects the nonzero mean of the series, while in the unit root case it reflects a deterministic trend in y_t. Because in the latter case first-differencing produces a stationary time series, the process for y_t is referred to as **difference stationary**.

It is also possible that non-stationarity is caused by the presence of a deterministic time trend in the process, rather than by the presence of a unit root. This happens when the *AR*(1) model is extended to

$$y_t = \mu + \gamma t + \rho y_{t-1} + \epsilon_t \qquad [1.9.10]$$

with $|\rho| < 1$ and $\gamma \neq 0$. In this case we have a non-stationary process because of the linear trend γt. This non-stationarity can be removed by regressing y_t upon a constant and *t*, and then considering the residuals of this regression, or by simply including *t* as an additional variable in the model. The process for y_t in this case is referred to as being **trend stationary**. Non-stationary process may thus be characterized by the presence of a deterministic trend or a stochastic trend implied by the presence of a unit root or both.

It is possible to test whether y_t follows a random walk against the alternative that it follows the trend stationary process as in [1.9.10]. This can be tested by running the regression

$$\Delta y_t = \mu + \gamma t + (\rho - 1) y_{t-1} + \epsilon_t \qquad [1.9.11]$$

The null hypothesis one would like to test is that the process is a random walk given by $H_0 : \mu = \gamma = \rho - 1 = 0$. Instead of testing this joint hypothesis, it is quite common to use the *t*-ratio on $\hat{\rho} - 1$, denoted by $\hat{\tau}_\tau$, assuming that the other restrictions in the null hypothesis are satisfied. Although the null hypothesis is still the same as in the previous unit root test, the testing regression is different and thus we have different distribution of the test statistic. It should be noted that if the unit root hypothesis $\rho - 1 = 0$ is rejected, we cannot conclude that the process for y_t is likely to be stationary. Under the alternative hypothesis γ may be nonzero so that the process for y_t is not stationary (but only trend stationary).

If a graphical inspection of the series indicates a clear positive or negative trend, it is most appropriate to perform the Dickey-Fuller test with a trend. This implies that the alternative hypothesis allows the process to exhibit a linear deterministic trend. If we are unable to reject the presence of a unit root, it does not necessarily mean that it is true. It could just be that there is insufficient information in the data to reject it.

Kwiatkowski, Phillips, Schmidt and Shin (1992) propose an alternative test where stationarity is the null hypothesis and the existence of a unit root is the alternative. This test is usually referred to as the **KPSS** test. The basic idea is that a time series is decomposed into the sum of a deterministic time trend, a random walk and a stationary error term (typically not a white noise). The null hypothesis (of trend stationarity) specifies that the variance of the random walk component is zero. The test is actually a Lagrange multiplier test, and computation of the test statistic is fairly simple. First run an auxiliary regression of y_t upon an intercept and a time trend *t*. Next, save the OLS residuals e_t and compute the partial sums $S_t = \sum_{s=1}^{t} e_s$ for all *t*. Then the test statistic is given by

$$KPSS = T^{-2}\sum_{t=1}^{T} S_t^2 \big/ \hat{\sigma}^2, \qquad [1.9.12]$$

where $\hat{\sigma}^2$ is an estimator of the error variance. This latter estimator $\hat{\sigma}^2$ may involve corrections for autocorrelations based on the Newey-West formula. The asymptotic distribution is nonstandard, and Kwiatkowski, Phillips, Schmidt and Shin (1992) report a 5% critical value of 0.146. If the null hypothesis is stationary rather than trend stationary, the trend term should be dropped from the auxiliary regression. The test statistic is computed in the same fashion, but the 5% critical value is 0.463.

1.9.2 Testing for Unit Roots in Higher-order Autoregressive Models

A test for a unit root in higher-order AR processes can easily be obtained by extending the Dickey-Fuller test procedures. The general strategy is that lagged differences, such as $\Delta y_{t-1}, \Delta y_{t-2}, \cdots,$ are included in the regression, such that its error term corresponds to white noise. This leads to the so called **augmented Dickey-Fuller tests** (ADF tests), for which the same asymptotic critical values hold as those of the Dickey-Fuller tests.

Consider the *AR(k)* model

$$\begin{aligned} \rho(L)y_t &= \mu + \epsilon_t \\ \rho(L) \quad &= 1 - \rho_1 L - \rho_2 L^2 - \cdots - -\rho_k L^k. \end{aligned} \qquad [1.9.13]$$

As we discussed before, y_t has a unit root when $\rho(1) = 0,$ or $\rho_1 + \rho_2 + \cdots + \rho_{\text{k}} = 1.$

In this case, y_t is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the above equation is using the so-called Dickey-Fuller (DF) reparametrization:

$$\Delta y_t = \mu + \alpha_0 y_{t-1} + \alpha_1 \Delta y_{t-2} + \cdots + \alpha_{k-1} \Delta y_{t-(k-1)} + \epsilon_t \qquad [1.9.14]$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter α_0 summarizes the information about the unit root, since $\rho(1) = -\alpha_0$. To see this, observe that the lag polynomial for the y_t computed from (1.9.14) is

$$(1-L)-\alpha_0 L-\alpha_1\left(L-L^2\right)-\cdots-\alpha_{k-1}\left(L^{k\text{-}1}-L^k\right)$$

But this must equal to $\rho(L)$, as the models are equivalent. Thus

$$\rho(1)=(1-1)-\alpha_0-\alpha_1(1-1)-\cdots-\alpha_{k-1}(1-1)=-\alpha_0.$$

Hence, the hypothesis of a unit root in y_t can be stated as

$$H_0:\alpha_0=0$$

Note that the model is stationary if $\alpha_0<0$. So the natural alternative is

$$H_1:\alpha_0<0$$

Under H_0, the model for y_t is

$$\Delta y_t=\mu+\alpha_1\Delta y_{t-1}+\cdots+\alpha_{k-1}\Delta y_{t-(k-1)}+\epsilon_t$$

which is an *AR(k-1)* in the first-difference Δy_t. Thus if y_t has a (single) unit root, then Δy_t is a stationary *AR* process. Because of this property, we say that if y_t is non-stationary but $\Delta^d y_t$ is stationary, then y_t is "integrated of order d", or I (d). Thus a time series with unit root is I (1).

Since α_0 is the parameter of a linear regression, the natural test statistic is the *t*-statistic for H_0 from OLS estimation of (1.9.14). Indeed, this is the most popular unit root test, and is called the **Augmented Dickey-Fuller (ADF)** test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under H_0, y_t is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

Theorem 1.9.1 Dickey-Fuller Theorem

Assume $\alpha_0=0$. As $T\rightarrow\infty$,

$$T\hat{\alpha}_0\xrightarrow{d}\left(1-\alpha_1-\alpha_2-\cdots-\alpha_{k-1}\right)\rho_\mu$$

$$ADF=\frac{\hat{\alpha}_0}{se\left(\hat{\alpha}_0\right)}\xrightarrow{d}\tau_\mu$$

The limiting distributions of ρ_μ and τ_μ are non-normal. They are skewed to the left, and have negative means.

The first result states that $\hat{\alpha}_0$ converges to its true value (of zero) at rate T, rather than the conventional rate of $T^{\frac{1}{2}}$. This is called a "**super-consistent**" rate of convergence.

The second result states that the *t*-statistic for $\hat{\alpha}_0$ converges to a limiting distribution which is non-normal, but does not depend on the parameters. This distribution has been extensively tabulated, and may be used for testing the hypothesis H_0. Note: The standard error $se(\hat{\alpha}_0)$ is the conventional ("homoscedastic") standard error. But the theorem does not require an assumption of homoscedasticity. Thus the Dickey-Fuller test is robust to heteroscedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects H_0 in favor of H_1 when $ADF < c$, where c is the critical value from the ADF table. If the test rejects H_0, this means that the evidence points to y_t being stationary. If the test does not reject H_0, a common conclusion is that the data suggests that y_t is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data are stationary or not.

We have described the test for the setting with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta y_t = \mu_1 + \mu_2 t + \alpha_0 y_{t-1} + \alpha_1 \Delta y_{t-1} + \cdots + \alpha_{k-1} \Delta y_{t-(k-1)} + \epsilon_t \qquad [1.9.15]$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for α_0 from OLS estimation of (1.9.15).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

If too many lags are included, this will somewhat reduce the power of the tests, but, if too few lags are included, the asymptotic distributions from the table are simply invalid (because of autocorrelation in the residuals), and the tests may lead to seriously biased conclusions. It is possible to use the model selection criterion or statistical significance of the additional variables to select the lag length in the ADF tests.

A regression of the form [1.9.14] can also be used to test for a unit root in a general (invertible) *ARMA* model. Said and Dickey (1984) argue that when, theoretically, one lets the number of lags in the regression grow with the sample size, the same asymptotic distributions hold and the *ADF* tests are also valid for an *ARMA* model. The argument essentially is that an *ARMA* model (with invertible *MA* polynomial) can be written as an infinite order autoregressive process. This explains why, when testing for unit roots, people usually do not worry about *MA* components.

Phillips and Perron (1988) have suggested an alternative to the augmented Dickey-Fuller tests. Instead of adding additional lags in the regressions to obtain an error term that has no autocorrelation, they stick to the original Dickey-Fuller regressions but make nonparametric adjustments to the Dickey-Fuller statistics to take into account of potential autocorrelation pattern in the errors. Phillips-Perron tests are nonparametric in nature and applicable to a wide class of weakly dependent and heterogeneously distributed innovations.

If the *ADF* test does not reject the null hypothesis of one unit root, the presence of a second unit root may be tested by estimating the regression of $\Delta^2 y_t$ on $\Delta y_{t-1}, \Delta^2 y_{t-1}, \cdots,$ $\Delta^2 y_{t-p+1},$ and comparing the *t*-ratio of the coefficient on Δy_{t-1} with the appropriate critical value. Alternatively, the presence of two unit roots may be tested jointly by estimating the regression of $\Delta^2 y_t$ on $y_{t-1}, \Delta y_{t-1}, \Delta^2 y_{t-1}, \cdots,$ and computing the usual F-statistic for testing the joint significance of y_{t-1} and Δy_{t-1}. This test statistic, under the null hypothesis of a double unit root, has a distribution that is not the usual F-distribution. Critical values of this distribution are tabulated by Hasza and Fuller (1979).

A stochastic process may be non-stationary for other reasons than the presence of unit roots. A linear deterministic trend is one example and structural breaks in the series may also mimic the presence of unit root non-stationarity. Without going into details, it may be mentioned that the recent literature on unit roots also includes discussions on stochastic unit roots, seasonal unit roots, fractional integration and panel data unit roots. A stochastic unit root implies that a process is characterized by a root that is not constant but stochastic and vary around unity. Such a process can be stationary for some periods and explosive for others (see Granger and Swanson, 1977). A seasonal unit root arises if a series becomes stationary after seasonal differencing. Fractional integration starts from the idea that a series may be integrated of order *d*, where *d* is not an integer. If $d \geq 1/2$, the process is non-stationary and said to be fractionally integrated. Finally, panel data unit root tests involve tests for unit roots in multiple series, for example GDP in ten different countries.

1.10 *Estimation of ARMA Models*

Suppose we know that the data series $y_1, y_2, \cdots, y_T$ is generated by an ARMA process of order *p, q*. Depending upon the specification of the model, and the distributional assumptions we are willing to make, we can estimate the unknown parameters by ordinary or nonlinear least squares or by maximum likelihood.

1.10.1 *Least Squares*

The least squares approach chooses the model parameters such that the residual sum of squares is minimal. This is particularly easy for models in autoregressive form. Consider the *AR(p)* model

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t \qquad [1.10.1]$$

where ϵ_t is a white noise error term that is uncorrelated with anything dated *t-1* or before. Consequently, we have

$$E\left(y_{t-j}\epsilon_t\right) = 0 \text{ for } j = 1, 2, 3, \cdots, p,$$

that is, error terms and explanatory variables are contemporaneously uncorrelated and OLS applied to [1.10.1] provides consistent estimates. Estimation of an autoregressive model is thus not different from a linear regression model with lagged dependent variable.

Estimation of the AR(p) process

$$\text{Let } \boldsymbol{x}_t = \left(1, y_{t-1}, y_{t-2}, \cdots, y_{t-p}\right)'$$

$$\boldsymbol{\beta} = \left(\boldsymbol{\alpha}, \phi_1, \phi_2, \cdots, \phi_p\right)'$$

Then the *AR(p)* model can be written as $y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + e_t$. The OLS estimator of the model is then given by $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$.To study properties of $\hat{\boldsymbol{\beta}}$, it helpful to define the process $\boldsymbol{u}_t = \boldsymbol{x}_t e_t$.

Note that u_t is a *MDS*, since

$$E\left(\boldsymbol{u}_t \middle| \mathcal{F}_{t-1}\right) = E\left(\boldsymbol{x}_t e_t \middle| \mathcal{F}_{t-1}\right) = \boldsymbol{x}_t E\left(e_t \middle| \mathcal{F}_{t-1}\right) = 0 .$$

By Theorem 1.1.1, it is also strictly stationary and ergodic. Hence,

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t e_t = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{u}_t \xrightarrow{p} E\left(\boldsymbol{u}_t\right) = \mathbf{0} \qquad [1.10.2]$$

The vector $\boldsymbol{x}_t$ is strictly stationary and ergodic, and by Theorem 1.1.1, so is $\boldsymbol{x}_t\boldsymbol{x}_t'$. Therefore by the Ergodic Theorem,

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t' \xrightarrow{p} E\left(\boldsymbol{x}_t\boldsymbol{x}_t'\right) = \mathbf{Q}$$

Combined with [1.10.2] and the continuous mapping theorem, we see that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right) \xrightarrow{p} \boldsymbol{\beta} + \mathbf{Q}^{-1}\mathbf{0} = \boldsymbol{\beta} .$$

We have shown the following:

Theorem 1.10.1 *If the AR(p) process* y_t *is strictly stationary and ergodic and* $E\left(y_t^2\right) < \infty$, then $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ as $T \uparrow \infty$.

Asymptotic Distribution

Theorem 1.10.2 *MDS CLT. If* $\boldsymbol{u}_t$ *is strictly stationary and ergodic MDS and* $\mathbf{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right) = \Omega < \infty$ *, then as* $T \uparrow \infty$,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t \xrightarrow{d} N\left(\mathbf{0},\Omega\right).$$

Since $\boldsymbol{x}_t e_t$ is a MDS, we can apply theorem 1.10.2 to see that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t \xrightarrow{d} N\left(\mathbf{0},\Omega\right), \qquad \text{Where } \Omega = E\left(\boldsymbol{x}_t\boldsymbol{x}_t' e_t^2\right).$$

Theorem 1.10.3 *If the AR(p) process* y_t *is strictly stationary and ergodic and* $E\left(y_t^4\right) < \infty$, t*hen as* $T \uparrow \infty$, $\sqrt{T}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{Q}^{-1}\boldsymbol{\Omega}\,\mathbf{Q}^{-1}\right)$.

This is identical in form to the asymptotic distribution of OLS estimator in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

Estimation of MA Process

For moving average models, estimation is somewhat more complicated. Suppose that we have an *MA(1)* model

$$y_t = \alpha + \epsilon_t + \theta\epsilon_{t-1}$$

Because $\epsilon_{t-1}$is not observed, we cannot apply regression techniques here. In theory, ordinary least squares would minimize

$$S(\theta,\alpha) = \sum_{t=2}^{T}\left(y_t - \alpha - \theta\epsilon_{t-1}\right)^2$$

A possible solution arises if we write ϵ_{t-1}in this expression as a function of observed $y_t s$. This is only possible if the MA polynomial is invertible. In this case we can use

$$\epsilon_{t-1} = \sum_{j=0}^{\infty}(-\theta)^j\left(y_{t-j-1} - \alpha\right)$$

and write

$$S(\theta,\alpha) = \sum_{t=2}^{T}\left(y_t - \alpha - \theta\sum_{j=0}^{\infty}(-\theta)^j\left(y_{t-j-1} - \alpha\right)\right)^2$$

In practice, y_t is not observed for $t = 0, -1, \cdots,$so we have to cut off the infinite sum in the above expression to obtain an approximate sum of squares

$$\tilde{S}(\theta,\alpha) = \sum_{t=2}^{T}\left(y_t - \alpha - \theta\sum_{j=0}^{t-2}(-\theta)^j\left(y_{t-j-1} - \alpha\right)\right)^2 \qquad [1.10.3]$$

Because, asymptotically, if T goes to infinity the difference between $S(\theta,\alpha)$and $\tilde{S}(\theta,\alpha)$ disappears and minimizing [1.10.3] with respect to α and θ gives consistent estimators $\hat{\alpha}$ and $\hat{\theta}$. Unfortunately, [1.10.3] is a higher-order polynomial in θ and thus has very many local minima. Therefore, analytically [1.10.3] is complicated. However, as we know that $-1 < \theta < 1$, a grid search can be performed. The resulting nonlinear least squares estimators for α and θ are consistent and asymptotically normal.

1.10.2 *Maximum Likelihood Estimation*

Consider an ARMA model of the form

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \qquad [1.10.4]$$

with ϵ_t white noise:

$$E(\epsilon_t) = 0 \qquad [1.10.5]$$

$$E(\epsilon_t \epsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise} \end{cases} \qquad [1.10.6]$$

This section explores how to estimate the values of $(\alpha, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, \sigma^2)$ on the basis of observations on y. The primary principle on which estimation will be based is *maximum likelihood*. Let $\boldsymbol{\theta} \equiv (\alpha, \phi_1, \cdots, \phi_p, \theta_1, \cdots, \theta_q, \sigma^2)'$ denote the vector of population parameters. Suppose we have observed a sample of size T $(y_1, y_2, \cdots, y_T)$. The approach will be to calculate the probability density

$$f_{y_T, y_{T-1}, \cdots, y_1}(y_T, y_{T-1}, \cdots, y_1; \boldsymbol{\theta}), \qquad [1.10.7]$$

which might be viewed as the probability of having observed this particular sample. The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is the value for which this sample is most likely to have been observed; that is, it is the value of $\boldsymbol{\theta}$ that maximizes [1.10.7]. This approach requires specifying a particular distribution for the white noise process ϵ_t. Typically we will assume that ϵ_t is Gaussian white noise:

$$\epsilon_t \sim \text{i.i.d. } N(0, \sigma^2). \qquad [1.10.8]$$

Although this assumption is strong, the estimates of $\boldsymbol{\theta}$ that results from it will often turn out to be a sensible for non-Gaussian processes as well. Finding the maximum likelihood estimates conceptually involves two steps. First, the likelihood function [1.10.7] must be calculated. Second, values of $\boldsymbol{\theta}$ must be found that maximize this function.

The Likelihood Function for a Gaussian AR(1) Process

A Gaussian AR(1) process takes the form

$$y_t = \alpha + \phi y_{t-1} + \epsilon_t, \qquad [1.10.9]$$

with $\epsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. For this case , the vector of parameters to be estimated consists of

$\boldsymbol{\theta} \equiv (\alpha, \phi, \sigma^2)'$.

Consider the probability distribution of y_1, the first observation in the sample. From equations [1.2.3] and [1.2.4] this is a random variable with mean

$E(Y_1)=\mu=\alpha/(1-\phi)$ and variance $E(Y_1-\mu)^2=\sigma^2/(1-\phi^2)$

Since $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is Gaussian, Y_1 is also Gaussian. Hence, the density of the first observation takes the form

$$f_{Y_1}(y_1;\boldsymbol{\theta})=f_{Y_1}(y_1;\alpha,\phi,\sigma^2)$$

$$=\frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}}\exp\left\{\frac{-[y_1-\alpha/(1-\phi)]^2}{2\sigma^2/(1-\phi^2)}\right\} \qquad [1.10.10]$$

Next consider the distribution of the second observation Y_2 conditional on observing $Y_1=y_1$. From [1.10.9],

$$Y_2=\alpha+\phi Y_1+\epsilon_2, \qquad [1.10.11]$$

Conditional on $Y_1=y_1$, $(Y_2|Y_1=y_1)\sim N(\alpha+\phi y_1,\sigma^2)$. Hence,

$$f_{Y_2|Y_1}(y_2|y_1;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y_2-\alpha-\theta y_1)^2}{2\sigma^2}\right\} \qquad [1.10.12]$$

The joint density of observations 1 and 2 is then just the product of [1.10.10] and [1.10.12]:

$$f_{Y_2,Y_1}(y_2,y_1;\boldsymbol{\theta})=f_{Y_2|Y_1}(y_2|y_1;\boldsymbol{\theta}).f_{Y_1}(y_1;\boldsymbol{\theta})$$

Similarly, the distribution of the third observation conditional on the first two is

$$f_{Y_3|Y_2,Y_1}(y_3|y_2,y_1;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y_3-\alpha-\theta y_2)^2}{2\sigma^2}\right\}$$

from which,

$$f_{Y_2,Y_2,Y_1}(y_3,y_2,y_1;\boldsymbol{\theta})=f_{Y_3|Y_2,Y_1}(y_3|y_2,y_1;\boldsymbol{\theta}).f_{Y_2|Y_1}(y_2|y_1;\boldsymbol{\theta}).f_{Y_1}(y_1;\boldsymbol{\theta})$$

In general, the values of $Y_1,Y_2,\cdots,Y_{t-1}$ matter for Y_t only through the value of Y_{t-1}, and the density of observation *t* conditional on the preceding *t-1* observations is given by

$$f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_1}(y_t|y_{t-1},y_{t-2},\cdots,y_1;\boldsymbol{\theta})=f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta})$$

$$=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-(y_t-\alpha-\theta y_{t-1})^2}{2\sigma^2}\right\} \qquad [1.10.13]$$

The joint density of the first *t* observations is then

$$f_{Y_t,Y_{t-1},\cdots,Y_1}(y_t,y_{t-1},\cdots,y_1;\boldsymbol{\theta})=f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta}).f_{Y_{t-1},Y_{t-2},\cdots,Y_1}(y_{t-1},y_{t-2},\cdots,y_1;\boldsymbol{\theta}) \qquad [1.10.14]$$

The likelihood of the complete sample can thus be calculated as

$$f_{Y_T,Y_{T-1},\cdots,Y_1}(y_T,y_{T-1},\cdots,y_1;\boldsymbol{\theta}) = f_{Y_1}(y_1;\boldsymbol{\theta}).\prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta}) \qquad [1.10.15]$$

The log likelihood function (denoted by $\mathscr{L}(\boldsymbol{\theta})$) can be found by taking logs of [1.10.15]

$$\mathscr{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1;\boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta}) \qquad [1.10.16]$$

Substituting [1.10.10] and [1.10.13] into [1.10.16], the log likelihood for a sample of size T from a Gaussian AR(1) process is seen to be

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta}) = &-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left[\sigma^2/(1-\phi^2)\right] - \frac{\left[y_1 - \alpha/(1-\phi)\right]^2}{2\sigma^2/(1-\phi^2)} \\ &-\frac{(T-1)}{2}\log(2\pi) - \frac{(T-1)}{2}\log\sigma^2 - \sum_{t=2}^{T}\left[\frac{(y_t-\alpha-\phi y_{t-1})^2}{2\sigma^2}\right]\end{aligned} \qquad [1.10.17]$$

The MLE $\hat{\boldsymbol{\theta}}$ is the value for which [1.10.17] is maximized. In principle, this requires differentiating [1.10.17] and setting the result equal to zero. In practice, when an attempt is made to carry this out, the result is a system of nonlinear equations in $\boldsymbol{\theta}$ and $(y_1,y_2,\cdots,y_T)$ for which there is no simple solution for $\boldsymbol{\theta}$ in terms of $(y_1,y_2,\cdots,y_T)$. Maximization of [1.10.17] thus requires iterative or numerical procedures.

Conditional maximum Likelihood Estimates of AR(1) Process

An alternative to numerical optimization of the exact likelihood function is to regard the value of y_1 as deterministic and maximize the likelihood conditioned on the first observation,

$$f_{Y_T,Y_{T-1},\cdots,Y_2|Y_1}(y_T,y_{T-1},\cdots,y_2|y_1;\boldsymbol{\theta}) = \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta}), \qquad [1.10.18]$$

the objective function then being to maximize

$$\begin{aligned}\log f_{Y_T,Y_{T-1},\cdots,Y_2|Y_1}(y_T,y_{T-1},\cdots,y_2|y_1;\boldsymbol{\theta}) = &-\frac{(T-1)}{2}\log(2\pi) - \frac{(T-1)}{2}\log\sigma^2 \\ &-\sum_{t=2}^{T}\left[\frac{(y_t-\alpha-\phi y_{t-1})^2}{2\sigma^2}\right]\end{aligned} \qquad [1.10.19]$$

Maximization of [1.10.19] with respect to α and ϕ is equivalent to minimization of

$$\sum_{t=2}^{T}(y_t-\alpha-\phi y_{t-1})^2, \qquad [1.10.20]$$

which is achieved by OLS regression of y_t on a constant and its own lagged value. The conditional maximum likelihood estimates of α and ϕ are therefore given by

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\phi}\end{pmatrix}=\begin{bmatrix}T-1 & \sum_{t=2}^{T} y_{t-1}\\ \sum_{t=2}^{T} y_{t-1} & \sum_{t=2}^{T} y_{t\text{-}1}^{2}\end{bmatrix}^{-1}\begin{bmatrix}\sum_{t=2}^{T} y_{t-1}\\ \sum_{t=2}^{T} y_{t-1}y_t\end{bmatrix},$$

The conditional maximum likelihood estimate of the innovation variance is found by differentiating [1.10.19] with respect to σ^2 and setting the result equal to zero:

$$-\frac{(T-1)}{2\hat{\sigma}^2}+\sum_{t=2}^{T}\left[\frac{\left(y_t-\hat{\alpha}-\hat{\phi}y_{t-1}\right)^2}{2\hat{\sigma}^4}\right]=0,$$

or

$$\hat{\sigma}^2=\frac{1}{T\text{ -}1}\sum_{t=2}^{T}\left(y_t-\hat{\alpha}-\hat{\phi}y_{t-1}\right)^2.$$

In contrast to the exact maximum likelihood estimates, the conditional maximum likelihood estimates are thus trivial to compute. Moreover, if the sample size *T* is sufficiently large, the first observation makes a negligible contribution to the total likelihood. The exact MLE and conditional MLE turns out to have the same large sample distribution, provided that $|\phi|<1.$ when $|\phi|>1,$ the conditional MLE continues to provide consistent estimates, whereas maximization of [1.10.17] does not. This is because [1.10.17] is derived from [1.10.10], which does not actually describe the density of Y_1 when $|\phi|>1.$ For these reasons, in most applications the parameters of an autoregression are estimated by OLS (conditional maximum likelihood) rather than exact maximum likelihood.

Conditional maximum Likelihood Estimates of AR(P) Process

A Gaussian *AR*(*p*) process takes the form

$$Y_t=\alpha+\phi_1Y_{t-1}+\phi_2Y_{t-2}+\cdots+\phi_pY_{t-p}+\epsilon_t, \qquad [1.10.21]$$

with $\epsilon_t\sim$ i.i.d. $N\left(0,\sigma^2\right)$. In this case , the vector of parameters to be estimated consists of

$$\boldsymbol{\theta}\equiv\left(\alpha,\phi_1,\phi_2,\cdots,\phi_p,\sigma^2\right)'.$$

Maximization of the exact log likelihood for *AR*(*p*) process must be accomplished numerically. In contrast, the log of the likelihood conditional on the first *p* observations assumes the simple form

$$\log f_{Y_T,Y_{T-1},\cdots,Y_{p+1}|Y_p,\cdots,Y_1}\left(y_T,y_{T-1},\cdots,y_{p+1}\middle|y_p,\cdots,y_1;\boldsymbol{\theta}\right)$$
$$=-\frac{(T-p)}{2}\log(2\pi)-\frac{(T-p)}{2}\log\sigma^2 \qquad [1.10.22]$$
$$-\sum_{t=p+1}^{T}\left[\frac{\left(y_t-\alpha-\phi_1 y_{t-1}-\phi_2 y_{t-2}-\cdots-\phi_p y_{t-p}\right)^2}{2\sigma^2}\right]$$

The values of $\alpha,\phi_1,\phi_2,\cdots,\phi_p$ that maximizes [1.10.22] are the same as those that minimize

$$\sum_{t=p+1}^{T}\left(y_t-\alpha-\phi_1 y_{t-1}-\phi_2 y_{t-2}-\cdots-\phi_p y_{t-p}\right)^2. \qquad [1.10.23]$$

Thus, the conditional maximum likelihood estimates of these parameters can be obtained from an OLS regression of y_t on a constant and p of its own lagged values. The conditional maximum likelihood estimate of σ^2 turns out to be

$$\hat{\sigma}^2=\frac{1}{T-p}\sum_{t=p+1}^{T}\left(y_t-\hat{\alpha}-\hat{\phi}_1 y_{t-1}-\hat{\phi}_2 y_{t-2}-\cdots-\hat{\phi}_p y_{t-p}\right)^2.$$

The exact maximum likelihood estimates and the conditional maximum likelihood estimates again have the same large sample distributions.

Conditional Likelihood Function for a Gaussian MA(1) Process

Calculation of the likelihood function for a moving average process is simple if we condition on initial values for the $\epsilon's$. Consider the Gaussian MA(1) process

$$Y_t=\mu+\epsilon_t+\theta\epsilon_{t-1} \qquad [1.10.24]$$

with $\epsilon_t\sim$ i.i.d. $N\left(0,\sigma^2\right)$. Let $\boldsymbol{\theta}\equiv\left(\mu,\theta,\sigma^2\right)'$ denote the polynomial parameters to be estimated. If the value of ϵ_{t-1} were known with certainty, then

$$Y_t|\epsilon_{t-1}\sim N\left(\mu+\theta\epsilon_{t-1},\sigma^2\right).$$

Or

$$f_{Y_t|\epsilon_{t-1}}\left(y_t|\epsilon_{t-1};\boldsymbol{\theta}\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-\left(y_t-\mu-\theta\epsilon_{t-1}\right)^2}{2\sigma^2}\right\} \qquad [1.10.25]$$

Suppose that we knew for certain that $\epsilon_0=0$. Then

$$\left(Y_1|\epsilon_0=0\right)\sim N\left(\mu,\sigma^2\right).$$

Moreover, given observation of y_1, the value of ϵ_1 is then known with certainty as well.

$$\epsilon_1 = y_1 - \mu$$

allowing application of [1.10.25] again:

$$f_{Y_2|Y_1,\,\epsilon_0=0}\left(y_2|y_1,\epsilon_0=0;\boldsymbol{\theta}\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-\left(y_2-\mu-\theta\epsilon_1\right)^2}{2\sigma^2}\right\}$$

Since ϵ_1 is known with certainty, ϵ_2 can be calculated from

$$\epsilon_2 = y_2 - \mu - \theta\epsilon_1$$

Proceeding in this fashion, it is clear that given knowledge that $\epsilon_0 = 0$, the full sequence $\{\epsilon_1,\epsilon_2,\cdots,\epsilon_T\}$ can be calculated from $\{y_1,y_2,\cdots,y_T\}$ by iterating on

$$\epsilon_t = y_t - \mu - \theta\epsilon_{t-1} \qquad [1.10.26]$$

for $t=1,2,\cdots,T$, starting from $\epsilon_0 = 0$. The conditional density of the t^{th} observation can then be calculated from [1.10.25] as

$$\begin{aligned} f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_1,\,\epsilon_0=0}\left(y_t|y_{t-1},y_{t-2},\cdots,y_1,\epsilon_0=0;\boldsymbol{\theta}\right) &= f_{Y_t|\epsilon_{t-1}}\left(y_t|\epsilon_{t-1};\boldsymbol{\theta}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{\epsilon_t^2}{2\sigma^2}\right\} \end{aligned} \qquad [1.10.27]$$

The sample likelihood would then be the product of these individual densities:

$$\begin{aligned} &f_{Y_T,Y_{T-1},\cdots,Y_1|\,\epsilon_0=0}\left(y_T,y_{T-1},\cdots,y_1|,\epsilon_0=0;\boldsymbol{\theta}\right) \\ &= f_{Y_1|\epsilon_0=0}\left(y_1|\epsilon_0=0;\boldsymbol{\theta}\right).\prod_{t=2}^{T} f_{Y_t|Y_{t-1},Y_{t-2},\cdots,Y_1,\,\epsilon_0=0}\left(y_t|y_{t-1},y_{t-2},\cdots,y_1,\epsilon_0=0;\boldsymbol{\theta}\right) \end{aligned}$$

The conditional log likelihood is

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},\cdots,Y_1|\,\epsilon_0=0}\left(y_T,y_{T-1},\cdots,y_1|\epsilon_0=0;\boldsymbol{\theta}\right) \\ &= -\frac{T}{2}\log(2\pi)-\frac{T}{2}\log\left(\sigma^2\right)-\sum_{t=1}^{T}\frac{\epsilon_t^2}{2\sigma^2} \end{aligned} \qquad [1.10.28]$$

For a particular numerical value of $\boldsymbol{\theta}$, we thus calculate the sequence of $\epsilon' s$ implied by the data from [1.10.26]. The conditional log likelihood function [1.10.28] is then a function of the sum of squares of these $\epsilon' s$. The log likelihood is a complicated function of μ and θ, so that an analytical expression for the maximum likelihood estimates of μ and θ is not readily calculated. Hence, even the conditional maximum likelihood estimates for MA(1) process must be found by numerical optimization.

Iteration on [1.10.26] from an arbitrary value of ϵ_0 will result in

$$\epsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \cdots + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t \theta^t \epsilon_0$$

If $|\theta|$ is substantially less than unity, the effect of imposing $\epsilon_0 = 0$ will quickly die out and the conditional likelihood [1.10.27] will give a good approximation to the unconditional likelihood for a reasonably large sample size. By contrast, if $|\theta| > 1$, the consequence of imposing $\epsilon_0 = 0$ accumulate over time. The conditional approach is not reasonable in this case. If numerical optimization of [1.10.28] results in a value of θ that exceeds one in absolute value, the results must be discarded. The numerical optimization should be attempted again with the reciprocal of $\hat{\theta}$ used as a starting value for the numerical search procedure.

Conditional Likelihood Function for a Gaussian MA(q) Process

For the *MA(q)* process

$$Y_t = \mu + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}, \qquad [1.10.29]$$

A simple approach is to condition on the assumption that the first q values for ϵ were all zero:

$$\epsilon_0 = \epsilon_{-1} = \cdots = \epsilon_{-q+1} = 0. \qquad [1.10.30]$$

From these starting values we can iterate on

$$\epsilon_t = y_t - \mu - \theta_1 \epsilon_{t-1} - \theta\ \epsilon_{t-2} - \cdots - \theta_q \epsilon_{t-q}, \qquad [1.10.31]$$

for $t = 1, 2, \cdots, T$. Let $\boldsymbol{\epsilon}_0$ denote the $(q \times 1)$ vector $(\epsilon_0 \quad \epsilon_{-1} \quad \cdots \quad \epsilon_{-q+1})'$. The conditional log likelihood is then

$$\begin{aligned} \mathscr{L}(\boldsymbol{\theta}) &= \log f_{Y_T, Y_{T-1}, \cdots, Y_1 | \boldsymbol{\epsilon}_0 = 0}\left(y_T, y_{T-1}, \cdots, y_1 | \boldsymbol{\epsilon}_0 = 0; \boldsymbol{\theta}\right) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\epsilon_t^2}{2\sigma^2} \end{aligned} \qquad [1.10.32]$$

where $\boldsymbol{\theta} \equiv (\mu, \theta_1, \theta_2, \cdots, \theta_q, \sigma^2)'$. Again, expression [1.10.32] is useful only if all values of z for which

$1+\theta_1 z+\theta_2 z^2+\cdots+\theta_q z^q=0$

lie outside the unit circle.

Conditional Likelihood Function for a Gaussian ARMA(p,q) Process

A Gaussian ARMA(*p,q*) process takes the form

$$Y_t=\alpha+\phi_1 Y_{t\text{-}1}+\phi_2 Y_{t\text{-}2}+\cdots+\phi_p Y_{t\text{-}p}+\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+\cdots+\theta_q\epsilon_{t-q}, \qquad [1.10.33]$$

where $\epsilon_t\sim$ i.i.d. $N(0,\sigma^2)$. The objective is to estimate the vector of population parameters $\boldsymbol{\theta}\equiv\begin{pmatrix}\alpha & \phi_1 & \phi_2 & \cdots & \phi_p & \theta_1 & \theta_2 & \cdots & \theta_q & \sigma^2\end{pmatrix}'$.

The approximation to the likelihood for an autoregression is conditioned on the initial values of the $y's$. The approximation to the likelihood function for a moving average process is conditioned on initial values for the $\epsilon's$. A common approach to the likelihood function for an *ARMA(p,q)* process conditions on both $y's$ and $\epsilon's$.

Taking initial values for $\boldsymbol{y}_0\equiv\begin{pmatrix}y_0 & y_{\text{-}1} & \cdots & y_{\text{-}p+1}\end{pmatrix}'$ and $\boldsymbol{\epsilon}_0\equiv\begin{pmatrix}\epsilon_0 & \epsilon_{\text{-}1} & \cdots & \epsilon_{\text{-}q+1}\end{pmatrix}'$ as given, the sequence $\{\epsilon_1,\ \epsilon_2,\ \cdots,\ \epsilon_T\}$ can be calculated from $\{y_1,\ y_2,\ \cdots,\ y_T\}$ by iterating on

$$\epsilon_t=y_t-\alpha-\phi_1 y_{t\text{-}1}-\phi_2 y_{t\text{-}2}-\cdots-\phi_p y_{t\text{-}p}-\theta_1\epsilon_{t-1}-\theta_2\epsilon_{t-2}-\cdots-\theta_q\epsilon_{t-q}, \qquad [1.10.34]$$

for $t=1,2,\cdots,T$. The conditional log likelihood is then

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta})&=\log f_{Y_T,Y_{T-1},\cdots,Y_1|\,\boldsymbol{Y}_0,\boldsymbol{\epsilon}_0}\left(y_T,y_{T-1},\cdots,y_1\middle|\boldsymbol{y}_0,\boldsymbol{\epsilon}_0;\boldsymbol{\theta}\right)\\&=-\frac{T}{2}\log(2\pi)-\frac{T}{2}\log(\sigma^2)-\sum_{t=1}^{T}\frac{\epsilon_t^2}{2\sigma^2}\end{aligned} \qquad [1.10.35]$$

One option is to set initial $y's$ and $\epsilon's$ equal to their expected values. That is, set $y_s=\alpha/(1-\phi_1-\phi_2-\cdots-\phi_p)$ for $s=0,-1,\cdots,-p+1$ and set $\epsilon_s=0$ for $s=0,-1,\cdots,-q+1$, and then proceed with the iteration in [1.10.34] for $t=1,2,\cdots,T$. Alternatively, Box and Jenkins (1976) recommended setting $\epsilon's$ to zero but $y's$ equal to their actual values. Thus, the iteration on [1.10.34] is started at date $t=p+1$ with $y_1\ \ y_2\ \ \cdots\ \ y_p$ set to the observed values and

$\epsilon_p=\epsilon_{p\text{-}1}=\cdots=\epsilon_{p-q+1}=0.$

Then the conditional log likelihood is calculated as

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta}) &= \log f\left(y_T, y_{T-1}, \cdots, y_{p+1} \middle| y_p, \cdots, y_1, \epsilon_p = 0, \cdots, \epsilon_{p-q+=0} = 0; \boldsymbol{\theta}\right) \\ &= -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log\left(\sigma^2\right) - \sum_{t=p+1}^{T}\frac{\epsilon_t^2}{2\sigma^2}\end{aligned}$$

As in the case for the moving average processes, these approximations should be used only if all values of z satisfying

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle.

1.11 *Model Selection and Diagnostic Checking*

Most of the time there are no economic reasons to choose a particular specification of the model. Consequently, to a large extent the data will determine which time series model is appropriate. Before estimating any model, it is common to estimate autocorrelation and partial autocorrelation coefficients directly from the data. Often it gives some idea about which models might be appropriate. After one or more models are estimated, their quality can be judged by checking whether the residuals are more or less white noise, and by comparing them with alternative specifications. These comparisons can be based on statistical significance tests or the use of particular model selection criteria.

The Autocorrelation Function

The autocorrelation function (*ACF*) describes the correlation between $y_t$and its lag y_{t-k} as a function of *k*. The k^{th}-order autocorrelation coefficient is defined as

$$\rho_k = \frac{\operatorname{cov}\left(y_t, y_{t-k}\right)}{\operatorname{var}\left(y_t\right)} = \frac{\gamma_k}{\gamma_0}$$

The sample autocorrelation function gives the estimated autocorrelation coefficient as a function of the lag *k* and estimated by

$$\hat{\rho}_k = \frac{\frac{1}{T-k}\sum_{t=k+1}^{T}\left(y_t - \bar{y}\right)\left(y_{t-k} - \bar{y}\right)}{\frac{1}{T}\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2} \qquad [1.11.1]$$

where $(1/T)\sum_{t=1}^{T} y_t$ denotes the sample average. Alternatively, it can be estimated by regressing $y_t$on a constant and y_{t-k}which will give a slightly different estimator as the summation in the numerator and denominator will be over the same set of observations. It will usually not to be the case that $\hat{\rho}_k$ is zero for an MA model of order $q < k$. But we can use $\hat{\rho}_k$ to test the hypothesis that $\rho_k = 0$ using the result that asymptotically

$$\sqrt{T}\left(\hat{\rho}_k - \rho_k\right) \xrightarrow{d} N\left(0, v_k\right)$$

where $v_k = 1 + 2\rho_1^2 + 2\rho_2^2 + \cdots + 2\rho_q^2$ if $q < k$.

Testing MA(k-1) versus MA(k) is done by testing $\rho_k = 0$ and comparing the test statistic

$$\sqrt{T}\frac{\hat{\rho}_k}{\sqrt{1 + 2\hat{\rho}_1^2 + 2\hat{\rho}_2^2 + \cdots + 2\hat{\rho}_{k-1}^2}} \qquad [1.11.2]$$

with critical values from the standard normal distribution. Typically, two-standard error bounds for $\hat{\rho}_k$ based on the estimated variance $1 + 2\hat{\rho}_1^2 + 2\hat{\rho}_2^2 + \cdots + 2\hat{\rho}_{k-1}^2$ are graphically displayed in the plot of the sample autocorrelation function. The order of a moving average can in this way be determined from an inspection of the sample *ACF*. At least it will give us a reasonable value for q to start with, and diagnostic checking should indicate whether it is appropriate or not.

For autoregressive models the *ACF* is less helpful. For *AR(1)* model we have seen that the autocorrelation coefficients do not cut off at a finite lag length. Instead they go to zero exponentially corresponding to $\rho_k = \phi^k$. For higher-order autoregressive models, the autocorrelation function is more complex. An alternative source of information that is helpful is provided by the partial autocorrelation function.

The Partial Autocorrelation Function

The k^{th} order sample partial autocorrelation coefficient is defined as an estimate for ϕ_k in an *AR(k)* model. We denote this by $\hat{\phi}_k$. So, estimating

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_k y_{t-k} + \epsilon_t$$

yields $\hat{\phi}_k$, the estimated coefficient for y_{t-k} in the *AR(k)* model. The partial autocorrelation coefficient $\hat{\phi}_k$ measures the additional correlation between y_t and y_{t-k} after controlling for the effects of other regressors $y_{t-1}, \cdots, y_{t-k+1}$.

Obviously if the true model is an *AR(p)* process, then estimating an *AR(k)* model by OLS gives consistent estimators for the model parameters if $k \geq p$. Consequently, we have

$$\text{plim}\ \hat{\phi}_k = 0 \ \text{ if } k > p. \qquad [1.11.3]$$

Moreover, it can be shown that the asymptotic distribution is normal, i.e.,

$$\sqrt{\mathrm{T}}\left(\hat{\phi}_k - 0\right) \xrightarrow{d} N(0,1) \ \text{ if } k > p. \qquad [1.11.4]$$

Consequently, the partial autocorrelation coefficients can be used to determine the order of AR process. Testing an *AR(k-1)* model versus and *AR(k)* model implies testing the null hypothesis that $\phi_k = 0$. Under the null hypothesis that the model is *AR(k-1),* the appropriate standard error of $\hat{\phi}_k$ based on [1.11.4] is $1/\sqrt{T}$, so that $\phi_k = 0$is rejected if $\left|\sqrt{T}\,\hat{\phi}_k\right| > 1.96$. This way one can look at the *PACF* and test for each lag whether the partial autocorrelation coefficient is zero. For a genuine *AR(p)* model the partial autocorrelation will be close to zero after the *p*th lag.

For a moving average model it can be shown that the partial autocorrelations do not have a cut-off point but tail off to zero just like the autocorrelations in an autoregressive model. In summary the *AR(p)* process is described by:

- an *ACF* that is infinite in extent (it tails off).
- a *PACF* that is (close to) zero for lags larger than *p*.

For an MA(*q*) process we have:

- an *ACF* that is (close to) zero for lags larger than *q*.
- a *PACF* that is infinite in extent (tails off).

In the absence of any of these two situations, a combined *ARMA* model may provide a parsimonious representation of the data.

Diagnostic Checking

As a last step in model-building cycle, some checks on the model adequacy are required. Possibilities are doing a residual analysis and overfitting the specified model. For example, if an *ARMA(p,q)* model is chosen, we could also estimate an *ARMA(p+1, q)* and an *ARMA(p, q+1)* models and test the significance of the additional parameters.

A residual analysis is usually based on the fact that the residuals of an adequate model should approximately be white noise. A plot of the residuals can be a useful tool in checking for outliers. Moreover, the estimated residual autocorrelation are usually examined. For a white noise series the autocorrelations are zero. Therefore the significance of the residuals autocorrelations is often checked by comparing with approximate two-standard error bounds $\pm 2/\sqrt{T}$. To check the overall acceptability of the residual autocorrelations, the Ljung-Box (1978) portmanteau test statistic

$$Q_k = T(T+2)\sum_{k=1}^{K} \tfrac{1}{T-k} r_k^2 \qquad [1.11.5]$$

is often used. Where r_k are the estimated autocorrelation coefficients of the residuals $\hat{\epsilon}_t$, and *K* is the number chosen by the researcher. Values of Q for different K may be computed in the residual analysis. For an *ARMA(p,q)* process, the statistic Q_k is approximately Chi-squared distributed with *K-p-q* $(\text{where } K > p+q)$ degrees of freedom under the null hypothesis that the *ARMA(p,q)* is correctly specified. If the model is rejected at this stage, the model-building cycle has to be repeated.

Criteria for Model Selection

Economic theory does not provide any guidance to the appropriate choice of models, some additional criteria can be used to choose from alternative models that are acceptable from statistical point of view. What is the appropriate choice of p in practice? This is a problem of model selection. One approach to model selection is to choose p based on a Wald test.

Another is to minimize the AIC or BIC information criterion assuming constant is included in the *AR(p)* model, e.g.

$$AIC(p) = \log\hat{\sigma}^2(p) + \frac{2(p+1)}{T} \qquad [1.11.6]$$

where $\hat{\sigma}^2(p)$ is the estimated residual variance from an AR(p).

One ambiguity in defining the AIC criterion is that the sample available for estimation changes as p changes. (If you increase p, you need more initial conditions.) This can induce strange behaviour into the *AIC*. The best remedy is to fix an upper value $\bar{p}$, and then reserve the first $\bar{p}$ as initial conditions, and then calculate the models *AR(1), AR(2), …, AR(*$\bar{p}$*)* on this unified sample. Alternatively one can use the *BIC* given by

$$BIC(p) = \log\hat{\sigma}^2(p) + \frac{2(p+1)}{T}\log T \qquad [1.11.7]$$

If one is to choose between alternative *ARMA(p,q)* models, the *AIC* and *BIC* information criteria are given by

$$AIC = \log\hat{\sigma}^2(p+q) + \frac{2(p+q+1)}{T}$$

$$BIC = \log\hat{\sigma}^2(p+q) + \frac{2(p+q+1)}{T}\log T$$

Both criteria are likelihood based and represent a different trade-off between ‘fit’, as measured by the loglikelihood value, and ‘parsimony’, as measured by the number of free parameters, $p+q+1$, (assuming the models include a constant). Usually the model with the smallest *AIC* or BC value is preferred, although one can choose to deviate from this if the differences in criterion values are small for a subset of the models.

While the two criterion differ in the trade-off between fit and parsimony, the *BIC* criterion can be preferred because it has the property that it will almost surely select the true model, if $T \to \infty$, provided that the true model is in the class of *ARMA(p,q)* models for relatively small values of p and q. The *AIC* criterion tends to result asymptotically in overparameterized models.

1.12 *Predicting with ARMA models*

The main goal of building a time series model is predicting the future path of economic variables. One can note that *ARMA* models usually perform quite well in prediction and often outperform more complicated structural models. Of course *ARMA* models do not provide any economic insight in one's predictions and are unable to forecast under alternative economic scenarios.

The Optimal Predictor

Suppose we are interested in predicting Y_{T+h} at time *T*, the value of Y_t *h*-periods ahead. A predictor for Y_{T+h} will based on an information set, denoted by $\mathcal{F}_T$, that contains the information that is available and potentially used at the time of making the forecast. Ideally it contains all the information that is observed and known at time *T*. In univariate time series modeling we will usually assume that the information set at any point *t* in time contains the value of Y_t and all its lags. Thus we have

$$\mathcal{F}_T = \{Y_{-\infty}, \cdots, Y_{T-1}, Y_T\} \qquad [1.12.1]$$

In general, the predictor $\hat{Y}_{T+h|T}$ (the predictor for Y_{T+h} as of time *T*) is a function of the information set $\mathcal{F}_T$. Our criterion for choosing a predictor from the many possible ones is to minimize the expected quadratic prediction error

$$E\left[\left(Y_{T+h} - \hat{Y}_{T+h|T}\right)^2 \middle| \mathcal{F}_T\right] \qquad [1.12.2]$$

Where $E(.|\mathcal{F}_T)$ denotes the conditional expectation given the information set $\mathcal{F}_T$. It can be shown that the best predictor for Y_{T+h} given the information set at time *T* is the conditional mean of Y_{T+h} given the information set. We denote this optimal predictor as

$$\hat{Y}_{T+h|T} \equiv E\left(Y_{T+h}|\mathcal{F}_T\right) \qquad [1.12.3]$$

The conditional expectation of Y_{T+h} given an information set $\mathcal{F}_T'$, where $\mathcal{F}_T'$ is a subset of $\mathcal{F}_T$, is at best as good as $\hat{Y}_{T+h|T}$ based on $\mathcal{F}_T$. In line with this intuition, it holds that, the more information one uses to determine the predictor, the better the predictor will be. For example, $E\left(Y_{T+h}|Y_T, Y_{T-1}, \cdots\right)$ will usually be a better predictor than $E\left(Y_{T+h}|Y_T\right)$ or $E\left(Y_{T+h}\right)$ (an empty information set.

To simplify things, assume that the parameters of the *ARMA* model for Y_T are known. In practice, one would simply replace the unknown parameters by their consistent estimates. Now, how do we determine these conditional expectations when Y_T follows an *ARMA* process? To simplify the notation, consider forecasting y_{T+h}, noting that $Y_{T+h|T} = \mu + y_{T+h|T}$. As a first example, consider an *AR(1)* process where

$$y_{T+1} = \phi y_T + \epsilon_{T+1}$$

Consequently,

$$\hat{y}_{T+1} = E\left(y_{T+1}|y_T, y_{T-1}, \cdots\right) = \phi y_T + E\left(\epsilon_{T+1}|y_T, y_{T-1}, \cdots\right) = \phi y_T \qquad [1.12.4]$$

To predict two periods ahead $(h = 2)$, we write

$$y_{T+2} = \phi y_{T+1} + \epsilon_{T+2}$$

from which it follows

$$\hat{y}_{T+2} = E\left(y_{T+2}|y_T, y_{T-1}, \cdots\right) = \phi E\left(y_{T+1}|y_T, y_{T-1}, \cdots\right) + E\left(\epsilon_{T+2}|y_T, y_{T-1}, \cdots\right) = \phi^2 y_T \qquad [1.12.5]$$

In general, we obtain $\hat{y}_{T+h} = \phi^h y_T$. Thus, the last observed value y_T contains all the information to determine the predictor for any future value. When h is large, the predictor $\hat{y}_{T+h}$ converges to 0 (the unconditional expectation of y_t), provided that $|\phi| < 1$. With a nonzero mean, the best predictor for Y_{T+h} is directly obtained as $\mu + \hat{y}_{T+h|T} = \mu + \phi^h\left(Y_T - \mu\right)$. Note that this differs from $\phi^h Y_T$.

As a second example, consider an *MA(1)* process where

$$y_t = \epsilon_t + \theta\epsilon_{t-1}$$

Then we have

$$E\left(y_{T+1}\middle|y_T,y_{T-1},\cdots\right)=\theta E\left(\epsilon_T\middle|y_T,y_{T-1},\cdots\right)=\theta\epsilon_T$$

Where implicitly, we assumed that ϵ_T is observed (contained in the information set $\mathcal{F}_T$). Assuming that the *MA process is invertible*, we can write

$$\epsilon_T=\sum_{j=0}^{\infty}(-\theta)^j\,y_{T-j}$$

and determine the one-period ahead predictor as

$$\hat{y}_{T+1|T}=\theta\sum_{j=0}^{\infty}(-\theta)^j\,y_{T-j} \qquad [1.12.6]$$

predicting two periods ahead gives

$$\hat{y}_{T+2|T}=E\left(y_{T+2}\middle|y_T,y_{T-1},\cdots\right)=E\left(\epsilon_{T+2}\middle|y_T,y_{T-1},\cdots\right)+\theta E\left(\epsilon_{T+1}\middle|y_T,y_{T-1},\cdots\right)=0 \qquad [1.12.7]$$

Which shows that the *MA(1)* model is uninformative for predicting two periods ahead: the best predictor is simply the (unconditional) expected value of y_t, normalized at zero. This also follows from the autocorrelation function of the process, because the *ACF* is zero after one lag. That is, the 'memory' of the process is only one period.

For the general *ARMA(p,q)* model

$$y_t=\phi_1 y_{t-1}+\phi_2 y_{t-2}+\cdots+\phi_p y_{t-p}+\epsilon_t+\theta_1\epsilon_{t-1}+\cdots+\theta_q\epsilon_{t-q}$$

We can derive the following recursive formula to determine the optimal predictors:

$$\hat{y}_{T+h|T}=\phi_1\hat{y}_{T+h-1|T}+\phi_2\hat{y}_{T+h-2|T}+\cdots+\phi_p\hat{y}_{T+h-p|T}+\hat{\epsilon}_{T+h|T}+\theta_1\hat{\epsilon}_{T+h-1|T}+\cdots+\theta_q\hat{\epsilon}_{T+h-q|T} \qquad [1.12.8]$$

where $\hat{\epsilon}_{T+k|T}$ is the optimal predictor for ϵ_{T+k} at time *T*, and

$$\begin{aligned}\hat{y}_{T+k|T}&=y_{T+k|T} && \text{if } k\le 0\\ \hat{\epsilon}_{T+k|T}&=0 && \text{if } k>0\\ \hat{\epsilon}_{T+k|T}&=\epsilon_{T+k} && \text{if } k\le 0\end{aligned}$$

where that latter innovation can be solved from the autoregressive representation of the model. For this we have used the fact that the process is stationary and invertible, in which case the information set $\{y_T,y_{T-1},\cdots\}$ is equivalent to $\{\epsilon_T,\epsilon_{T-1},\cdots\}$. That is, if all $\epsilon_t s$ are known from -∞ to *T*, then all $y_t s$ are known from -∞ to *T*, and vice versa.

To illustrate this, consider an *ARMA(1,1)* model that takes the form

$$y_t = \phi y_{t-1} + \epsilon_t + \theta \epsilon_{t-1}$$

The optimal predictor for y_{T+1} is then given by

$$\hat{y}_{T+1|T} = \phi y_T + \hat{\epsilon}_{T+1|T} + \theta \epsilon_T = \phi y_T + \theta \epsilon_T$$

Assuming invertibility

$$y_t - \phi y_{t-1} = (1+\theta L)\epsilon_t$$

Can be written as

$$\epsilon_t = (1+\theta L)^{-1}(y_t - \phi y_{t-1}) = \sum_{j=0}^{\infty}(-\theta)^j (y_t - \phi y_{t-1})$$

we can write for the one-period-ahead predictor

$$\hat{y}_{T+1|T} = \phi y_T + \theta \sum_{j=0}^{\infty}(-\theta)^j (y_T - \phi y_{T-1}) \qquad [1.12.9]$$

Predicting two period ahead gives

$$\hat{y}_{T+2|T} = \phi \hat{y}_{T+1|T} + \hat{\epsilon}_{T+2|T} + \theta \hat{\epsilon}_{T+1|T} = \phi \hat{y}_{T+1|T} \qquad 1.12.10]$$

Note that this does not equal to $\phi^2 y_T$.

Prediction Accuracy

In addition to prediction, it is important to know how accurate this prediction is. To judge forecasting precision, we define the **prediction error** as $Y_{T+h} - \hat{Y}_{T+h|T} = y_{T+h} - \hat{y}_{T+h|T}$ and the expected quadratic prediction error as

$$C_h = E\left(y_{T+h} - \hat{y}_{T+h|T}\right)^2 = \operatorname{var}\left(y_{T+h} \middle| \mathcal{F}_T\right) \qquad [1.12.11]$$

Where the latter step follows from the fact that $\hat{y}_{T+h|T} = E\left(y_{T+h} \middle| \mathcal{F}_T\right)$. Determining C_h, corresponding to the variance of the *h*-period-ahead prediction error, is relatively easy with the moving average representation. To start with the simplest case, consider an *MA(1)* model. Then we have

$$C_1 = \operatorname{var}\left(y_{T+1} \middle| y_T, y_{T-1}, \cdots\right) = \operatorname{var}\left(\epsilon_{T+1} + \theta\epsilon_T \middle| \epsilon_T, \epsilon_{T-1}, \cdots\right) = \operatorname{var}\left(\epsilon_{T+1}\right) = \sigma^2$$

Alternatively, we explicitly solve for the predictor, which is $\hat{y}_{T+1|T} = \theta \epsilon_T$, and determine the variance of y_{T+1} - $\hat{y}_{T+1|T} = \epsilon_{T+1}$, which gives the same result. For the two-period-ahead predictor we have

$$C_2 = \text{var}\left(y_{T+2} \middle| y_T, y_{T-1}, \cdots\right) = \text{var}\left(\epsilon_{T+2} + \theta\epsilon_{T+1} \middle| \epsilon_T, \epsilon_{T-1}, \cdots\right) = \left(1+\theta^2\right)\sigma^2.$$

As one would expect, the accuracy of the prediction decreases if we predict further into the future. It will not, however, increase any further if h is increased beyond 2. This becomes clear if we compare the expected quadratic prediction error with that of a simple unconditional predictor:

$$\hat{y}_{T+h|T} = E\left(y_{T+h}\right) = E\left(\epsilon_{T+h} + \theta\epsilon_{T+h-1}\right) = 0$$

For this predictor we have

$$C_h = E\left(y_{T+h} - 0\right)^2 = \text{var}\left(y_{T+h}\right) = \text{var}\left(\epsilon_{T+h} + \theta\epsilon_{T+h-1}\right) = \left(1+\theta^2\right)\sigma^2$$

Consequently, this gives an upper bound on the inaccuracy of the predictors. The *MA(1)* model thus gives more efficient predictors only if one predicts one period ahead. More general *ARMA* models, however, will yield efficiency gains also in further ahead predictors.

Suppose the general model is *ARMA(p,q)*, which we can write as an *MA(∞)* model, with α_j coefficients to be determined:

$$y_t = \sum_{j=0}^{\infty} \alpha_j \epsilon_{t-j} \qquad \text{with } \alpha_0 = 1$$

The h-period ahead predictor (in terms of $\epsilon_t s$) is given by

$$\hat{y}_{T+h|T} = E\left(y_{T+h} \middle| y_T, y_{T-1}, \cdots\right) = \sum_{j=0}^{\infty} \alpha_j E\left(\epsilon_{T+h-j} \middle| \epsilon_T, \epsilon_{T-1}, \cdots\right) = \sum_{j=h}^{\infty} \alpha_j \epsilon_{T+h-j}$$

such that

$$y_{T+h} - \hat{y}_{T+h|T} = \sum_{j=0}^{h-1} \alpha_j \epsilon_{T+h-j}$$

Consequently, we have

$$E\left[\left(y_{T+h} - \hat{y}_{T+h|T}\right)^2\right] = \sigma^2 \sum_{j=0}^{h-1} \alpha_j^2 \qquad [1.12.12]$$

This shows how the variance of the forecast errors can easily be determined from the coefficients of the moving average representation of the model. Recall that, for the computation of the predictor, the autoregressive representation was most convenient.

As an illustration, consider the *AR(1)* model where $\alpha_j = \phi^j$. The expected quadratic prediction error are given by

$$C_1 = \sigma^2, \quad C_2 = \sigma^2\left(1+\phi^2\right), \quad C_2 = \sigma^2\left(1+\phi^2+\phi^4\right), \text{etc.}$$

For $h = \infty$, we have $C_\infty = \sigma^2\left(1+\phi^2+\phi^4+\cdots\right) = \sigma^2/\left(1-\phi^2\right)$, which is the unconditional variance of y_t. Consequently, the informational value contained in *AR(1)* process slowly decays over time. In the long run the predictor equals the unconditional predictor, being the mean of the y_t series.

In practical cases, the parameters in ARMA models are unknown, and replaced by their estimated values. This introduces additional uncertainty in the predictors. Usually, however, this uncertainty is ignored. The motivation is that the additional variance that arises because of estimation error disappears asymptotically as the sample size becomes sufficiently large.