Time Series Models

James L. Powell
Department of Economics
University of California, Berkeley

Overview

In contrast to the classical linear regression model, in which the components of the dependent variable vector $\mathbf{y}$ are not identically distributed (because its mean vector varies with the regressors) but may be independently distributed, *time series models* have dependent variables which may be identically distributed, but are typically not independent across ovbservations. Such models are applicable for data that are collected over time. A leading example, to be discussed more fully below, is the *first order autoregression* model, for which the dependent variable y_t for time period t satisfies

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

for ε_t satisfying the assumptions of the error terms in a classical linear regression model (i.e., mean zero, constant variance, and uncorrelated across t). For some values of the parameters, y_t will have constant mean and variances over t, but the covariance between y_t and y_s will generally be nonzero when $t \neq s$. The first-order autoregression model can be viewed as a special case of a *dynamic regression model*, with

$$y_t = \alpha + \phi y_{t-1} + \mathbf{x}_t'\beta + \varepsilon_t,$$

with $\mathbf{x}_t$ a vector of regressors.

The usual purpose of these models is *prediction*; given the recursive structure of such models, realizations of the dependent variable today are useful in forecasting its value in the future. In much of time series modeling, the values of the parameters themselves are not the objects of interest, but rather the ability of the specified model to forecast out of the observed sample; thus, much of the statistical methodology is devoted to finding a "good" model for the data rather than the "right" model.

Stationarity and Ergodicity

The statistical theory for time series data views the sequence of dependent variables $\{y_t\}$ as a stochastic process, i.e., a realization of a random function whose argument is the time index t. (Unless stated otherwise, the discussion here will assume y_t is scalar.) Without restrictions on the parameters of

the joint distribution of the values of y_t over t – so that, for example, the means and variances of y_t are allowed to vary freely over t – it would clearly be impossible to construct consistent estimators of those parameters with a single realization of history. The concept of *stationarity* imposes such restrictions.

The process $\{y_t\}$ is said to be *weakly stationary* (or *covariance stationary*) if the second moments of y_t exist, and the first and second moments satisfy

$$\begin{aligned} E(y_t) &= \mu, \\ Var(y_t) &= \sigma^2 \equiv \gamma_y(0) \\ Cov(y_t, y_s) &= \gamma_y(t-s) = \gamma_y(|t-s|). \end{aligned}$$

That is, the mean values of y_t are constant, and the covariance between any pair y_t and y_s of observations depends only on the (absolute) difference of their indices $|t-s|$. By reducing the means, variances, and covariances between pairs of observations to a single time-invariant parameter, there is some hope of consistently estimating those parameters with a single realization of the process $\{y_t\}$. The function $\gamma_y(s)$ is called the *autocovariance function* of the y_t process.

A "stronger" definition of stationarity, suggestively titled *strong stationarity*, restricts the joint distribution of any finite collection of consecutive realizations of y_t to be invariant across t, in the sense that

$$\Pr\{(y_t, y_{t+1}, ...y_{t+K}) \in B\} = \Pr\{(y_0, y_1, ...y_K) \in B\}$$

for any integer K and corresponding event B. This is not, strictly speaking, stronger than weak stationarity without the additional conditions that the second moment of y_t is finite, with which it does indeed imply covariance stationarity. For the theoretical development, when deriving the mean-squared error of forecasts, etc., the assumption of weak stationarity usually suffices; when deriving probability limits and asymptotic distributions for statistics, typically strong stationarity (or a similar strengthening of weak stationarity) is assumed.

Since econometric modeling typically involves characterization of relationships for several variables, it is useful to extend the notion of stationarity to *vector processes,* where $\mathbf{y}_t \in R^M$ for some $M > 1$. Such a

process is covariance stationary if

$$\begin{aligned} E(\mathbf{y}_t) &= \boldsymbol{\mu}, \\ Var(\mathbf{y}_t) &= \boldsymbol{\Sigma} \equiv \boldsymbol{\Gamma}_y(0) \\ \mathbf{C}(\mathbf{y}_t, \mathbf{y}_s) &= E\left[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'\right] \\ &= \boldsymbol{\Gamma}_y(t-s) \\ &= \left[\boldsymbol{\Gamma}_y(s-t)\right]'. \end{aligned}$$

Extension of the concept of strong stationarity to vector processes is similarly straightforward.

Even if a scalar dependent variable y_t is stationary, it need not be true that a law of large numbers applies, i.e., stationarity does not imply that

$$\bar{y}_T \equiv \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} E(y_t) = \mu.$$

If this condition is satisfied, y_t is said to be *(weakly) ergodic*; It is said to be *strongly ergodic* if

$$\frac{1}{T}\sum_{t=1}^{T} f(y_t, y_{t+1}, ..., y_{t+K}) \xrightarrow{a.s.} E(f(y_t, y_{t+1}, ..., y_{t+K}))$$

whenever the latter moment exists. It is easy to construct examples of stationary processes which are not ergodic; for example, if

$$y_t \equiv z \sim N(\mu, 1),$$

then y_t is clearly (weakly and strongly) stationary, but $\bar{y}_T \equiv z \neq \mu$ with probability one. Another example is

$$y_t = \begin{cases} z_1 \sim N(\mu, 1) & \text{when } t \text{ is even,} \\ z_2 \sim N(\mu, 1) & \text{when } t \text{ is odd,} \end{cases}$$

where z_1 and z_2 are independent. Such processes are special cases of *deterministic processes*, which can be perfectly predicted by a linear combination of past values:

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + ...$$

For the first process, $\beta_1 = 1$ and the rest are zero, while for the second, only $\beta_2 = 1$ is nonzero; in general, the β coefficients must sum to one to ensure stationarity. In practice, seasonal factors (which are periodically-recurring "constant terms") are good examples of deterministic processes; for these processes,

it is usually possible to consistently estimate the realized values (from noisy data), but not the parameters of the distributions which generated them

For a process to be ergodic, the some measure of the dependence between observations y_t and y_s must vanish as $|t-s|$ increases. If so, a law of large numbers should be applicable, as in the following result:

Weak Ergodic Theorem: If y_t is covariance stationary with $E(y_t) = \mu$, $Cov(y_t, y_{t-s}) = \gamma_y(s)$, and if

$$\sum_{s=-\infty}^{\infty} |\gamma_y(s)| < \infty,$$

then

$$\bar{y}_T \equiv \frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} \mu.$$

Proof: Since $E(\bar{y}_T) = \mu$, quadratic mean convergence of $\bar{y}_T$ to μ will follow if $Var(\bar{y}_T) \to 0$ as $T \to \infty$. But

$$\begin{aligned}
Var(\bar{y}_T) &= \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T} \gamma_y(t-s) \\
&\leq \frac{1}{T^2}\sum_{s=1}^{T}\sum_{t=1}^{T} |\gamma_y(t-s)| \\
&= \frac{1}{T^2}\sum_{s=-(T-1)}^{T-1} (T-|s|)\cdot|\gamma_y(s)| \\
&\leq \frac{1}{T}\sum_{s=-\infty}^{\infty} |\gamma_y(s)| \\
&\to 0 \qquad \text{as} \qquad T \to \infty.
\end{aligned}$$

The middle equality in this proof, which rewrites the double sum as a single sum, is easiest to understand by considering two ways to add up all elements in a $T \times T$ matrix with element $|\gamma_y(t-s)|$ in row t, column s; the double sum adds across columns and rows, while the single sum adds along the diagonals (whose elements are constant by stationarity).

The condition $\sum_{s=-\infty}^{\infty} |\gamma_y(s)| < \infty$, known as *absolute summability of the autocovariance function,* is obviously satisfied for i.i.d. processes (since the doubly-infinite sum reduces to $Var(y_t)$ in that case); it suffices for weak ergodicity but is by no means necessary for it. It is easy to see how to modify the proof while imposing only the weaker condition

$$\frac{1}{T}\sum_{s=0}^{T-1} |\gamma_y(s)| \to 0;$$

that is, it suffices that the (sample) average covariance between y_t and all values of y_t in the sample (including y_t itself) converges to zero as T increases. This is one way that declining dependence between observations in a sample implies ergodicity; there are a number of other ergodic theorems (weak and strong) that restrict other measures of dependence across observations to obtain a law of large numbers for dependent data.

Even stronger restrictions on dependence between observations and existence of moments are needed to obtain central limit theorems; some of the restrictions on dependence have the headings "mixing conditions" or "martingale difference sequences." Such conditions will not be presented here; suffice it to say that, for all the processes considered below, it is possible to find sufficient regularity conditions to ensure that a central limit theorem applies:

$$\sqrt{T}(\bar{y}_T - \mu) \xrightarrow{d} N(0, V_0),$$

where V_0 is the limit of the variance of the normalized average $\sqrt{T}(\bar{y}_T - \mu)$,

$$\begin{aligned} V_0 &= \lim_{T\to\infty} Var(\sqrt{T}(\bar{y}_T - \mu)) \\ &= \sum_{s=-\infty}^{\infty} \gamma_y(s). \end{aligned}$$

For i.i.d. data, V_0 reduces to the usual $\gamma_y(0) = Var(y_t) \equiv \sigma_y^2$.

All of the results discussed above extend to the case when $\mathbf{y}_t$ is a random vector; in this case, weak stationarity is defined in terms of *autocovariance matrices*

$$\mathbf{\Gamma}_s \equiv \mathbf{C}(\mathbf{y}_t, \mathbf{y}_{t-s}),$$

and, for example, a dependent CLT for vector stochastic processes would yield

$$\sqrt{T}(\bar{\mathbf{y}}_T - \boldsymbol{\mu}) \xrightarrow{d} N(0, \mathbf{V}_0),$$

with the asymptotic covariance matrix $\mathbf{V}_0$ defined as

$$\mathbf{V}_0 \equiv \sum_{s=-\infty}^{\infty} \mathbf{\Gamma}_s.$$

Autoregressive and Moving Average Processes

A flexible class of models for (possibly) stationary univariate time series, proposed by Box and Jenkins in the mid-1960s, are *autoregressive moving average* models – ARMA models for short. The

fundamental building block for ARMA models is a *white noise process*, which is just a colorful mixed metaphor (light and sound) for a stochastic process ε_t which satisfies the properties imposed upon error terms in the standard linear model.

White Noise Process: The process $\{\varepsilon_t\}$ is called a *white noise process* with parameter σ^2, denoted $\varepsilon_t \sim WN(\sigma^2)$, if it is weakly stationary with

$$\begin{aligned} E(\varepsilon_t) &= 0, \\ Var(\varepsilon_t) &= \sigma^2, \\ Cov(\varepsilon_t, \varepsilon_s) &= 0 \qquad \text{if} \qquad t \neq s. \end{aligned}$$

From this simplest example of a weakly stationary and weakly ergodic process (which is strongly stationary if ε_t is assumed to be i.i.d.), it is possible to build other processes y_t with more interesting autocovariance patterns by assuming y_t is generated by a linear combination of its past values plus a linear combination of current and past values of a white noise error term ε_t. First are the *purely autoregressive* processes, which only involve a single, contemporaneous white noise term.

First-order Autoregressive Process: The process y_t is *first-order autoregressive,* denoted $y_t \sim AR(1)$, if it satisfies

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

where $\varepsilon_t \sim WN(\sigma^2)$ and $Cov(\varepsilon_t, y_{t-s}) = 0$ if $s \geq 1$.

Not all $AR(1)$ processes are stationary; if the process is stationary, then $E(y_t) = E(y_{t-1})$, implying

$$\begin{aligned} E(y_t) &= \alpha + \phi E(y_t) \\ &= \frac{\alpha}{1-\phi}, \end{aligned}$$

which requires $\phi \neq 1$. Furthermore $Var(y_t) = Var(y_{t-1})$, which requires

$$\begin{aligned} Var(y_t) &= \phi^2 Var(y_t) + Var(\varepsilon_t) + 2 \cdot Cov(y_{t-1}, \varepsilon_t) \\ &= \phi^2 Var(y_t) + \sigma^2 \\ &= \frac{\sigma^2}{1-\phi^2}, \end{aligned}$$

which is only well-defined and nonnegative if $|\phi| < 1$. This latter condition is sufficient for weak stationarity of y_t; calculations analogous to those for the variance yield

$$\gamma_y(s) = Cov(y_t, y_{t-s}) = \phi^s \frac{\sigma^2}{1-\phi^2} = \phi^s Var(y_t),$$

so the covariance between y_t and y_{t-s} declines geometrically as s increases; if ϕ is negative, the autocovariance function oscillates between positive and negative values.

Generalizations of the AR(1) process include more lagged dependent variables on the right-hand side of the equation for y_t :

p^{th}-order Autoregressive Process: The process y_t is *p^{th}-order autoregressive,* denoted $y_t \sim AR(p)$, if it satisfies

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t,$$

where $\varepsilon_t \sim WN(\sigma^2)$ and $Cov(\varepsilon_t, y_{t-s}) = 0$ if $s \geq 1$.

The conditions for stationarity of this process are related to the conditions for stability of the corresponding deterministic difference equation

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p};$$

specifically, the AR(p) process is stationary if any (real or complex) root z^* of the associated polynomial equation

$$0 = \tilde{\phi}(z) \equiv z^p - \phi_1 z^{p-1} - ... - \phi_{p-1} z - \phi_p$$

is inside the unit circle, i.e., $|z^*| < 1$.

Another simple class of time series models are *moving average processes.* Unlike autoregressive processes, these are weakly stationary by construction.

First-order Moving Average Process: The process y_t is a *first-order moving average* process, denoted $y_t \sim MA(1)$, if it can be written as

$$y_t = \mu + \varepsilon_t + \theta \varepsilon_{t-1},$$

where $\varepsilon_t \sim WN(\sigma^2)$. This process is covariance stationary with

$$\begin{aligned} E(y_t) &= \mu, \\ Var(y_t) &= \sigma^2(1+\theta^2), \\ Cov(y_t, y_{t-1}) &= \sigma^2\theta, \\ Cov(y_t, y_{t-s}) &= 0 \qquad \text{if} \qquad s > 1. \end{aligned}$$

q^{th}-order Moving Average Process: The process y_t is a *q^{th}-order moving average* process, denoted $y_t \sim MA(q)$, if it can be written as

$$y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q},$$

where $\varepsilon_t \sim WN(\sigma^2)$. Here

$$\begin{aligned} E(y_t) &= \mu, \\ Var(y_t) &= \sigma^2(1+\theta_1^2 + ... + \theta_q^2), \\ Cov(y_t, y_{t-1}) &= \sigma^2(\theta_1\theta_2 + ... + \theta_{q-1}\theta_q), \\ &... \\ Cov(y_t, y_{t-s}) &= 0 \qquad \text{if} \qquad s > q. \end{aligned}$$

The autoregressive and moving average processes can be combined to obtain a very flexible class of univariate processes (proposed by Box and Jenkins), known as *ARMA* processes.

ARMA(p,q) Process: The time series y_t is an *ARMA(p,q)* process, written $y_t \sim ARMA(p,q)$, if

$$y_t = \alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q},$$

where $\varepsilon_t \sim WN(\sigma^2)$ and $Cov(\varepsilon_t, y_{t-s}) = 0$ if $s \geq 1$. The requirements for stationarity of this process are the same as for stationarity of the corresponding $AR(p)$ process.

The Wold Decomposition

If we permit the order q of a MA(q) process to increase to infinity – that is, if we write

$$y_t = \mu + \sum_{\sigma=0}^{\infty} \theta_s \varepsilon_{t-s}$$

with $\varepsilon_t \sim WN(\sigma^2)$ and $\theta_0 \equiv 1$, we obtain what is known as a *linearly indeterministic* process, denoted $y_t \sim MA(\infty)$. This process is well-defined (in a mean-squared error sense) if the sequence of moving average coefficients $\{\theta_s\}$ is square-summable,

$$\sum_{\sigma=0}^{\infty} \theta_s^2 < \infty.$$

By recursion, stationary ARMA processes can be written as linearly deterministic processes; for example, a stationary AR(1) process $y_t = \alpha + \phi y_{t-1} + \varepsilon_t$ has $\theta_s \equiv \phi^s$. Conversely, the MA coefficients for any linearly indeterministic process can be arbitrarily closely approximated by the corresponding coefficients of a suitable ARMA process of sufficiently high order.

Wold showed that *all* covariance stationary stochastic processes could be written as the sum of deterministic and linearly indeterministic processes which were uncorrelated at all leads and lags; that is, if y_t is covariance stationary, then

$$y_t = x_t + z_t,$$

where x_t is a covariance stationary deterministic process (as defined above) and z_t is linearly indeterministic, with $Cov(x_t, z_s) = 0$ for all t and s. This result gives a theoretical underpinning to Box and Jenkins' proposal to model (seasonally-adjusted) scalar covariance stationary processes as ARMA processes.

Common Factors and Identification

In a sense, ARMA processes are *too* flexible, in the sense that low-order processes (i.e., those with p and q small) are nested in higher-order processes with certain parameter restrictions. In general, if $y_t \sim ARMA(p, q)$, then it can always be rewritten as an $ARMA(p + r, q + r)$ process for arbitrary positive integer r by suitable "generalized differencing". For example, suppose $y_t = \varepsilon_t \sim WN\left(\sigma^2\right)$, so that $y_t \sim ARMA(0, 0)$. Then for any ρ with $|\rho| < 1$,

$$y_t - \rho y_{t-1} = \varepsilon_t - \rho \varepsilon_{t-1},$$

or

$$y_t = \rho y_{t-1} + \varepsilon_t - \rho \varepsilon_{t-1},$$

so $y_t \sim ARMA(1, 1)$ with a particular restriction on the parameters (i.e., the sum of the first-order autoregressive and moving average coefficients is zero). For this example this redundancy is easy to find, but for more complicated $ARMA$ processes the restrictions on the parameters may be difficult to find in the population, and even harder to detect in estimation.

Box and Jenkins' proposed solution to this *common factors* problem, which they called their "principle of parsimony", is simple enough – just pick p and q to be small enough to do the job (of forecasting, etc.). To implement this general idea, however, they proposed a methodology for model selection which they termed *time series identification* procedures. In econometric applications, the tradition has been to consider only *purely autoregressive* processes, i.e., assume that $y_t \sim AR(p)$ for some value of p (chosen in practice by a suitable model selection procedure). Purely autoregressive processes, while typically requiring a higher number of parameters to approximate complicated dynamic patterns, do not suffer from the common factor problem, since a redundant generalized difference in the autoregressive component is accompanied by an error term which is not white noise (i.e., $q = r > 0$). Furthermore, as will be discussed later, purely autoregressive processes are simpler to estimate, requiring only linear (not nonlinear) LS estimation.

Vector Autoregressions

The definition of $ARMA$ processes is straightforward to extend a vector $\mathbf{y}_t \in R^M$ of dependent variables; here, though, restriction to purely autoregressive models for the vector process can yield complex dynamic patterns for the individual components of $\mathbf{y}_t$. A *p^{th}-order vector autoregression,* denoted $\mathbf{y}_t \sim VAR(p)$, is a vector process of the form

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}_1\mathbf{y}_{t-1} + ... + \mathbf{B}_p y_{t-p} + \boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\alpha}$ is an M-dimensional vector of intercept terms, $\mathbf{B}_1$ through $\mathbf{B}_p$ are $(M \times M)$ matrices of unknown coefficients, and $\boldsymbol{\varepsilon}_t$ is a *vector white noise* process, denoted $\boldsymbol{\varepsilon}_t \sim VWN(\boldsymbol{\Sigma})$ and defined to satisfy

$$\begin{aligned} E(\boldsymbol{\varepsilon}_t) &= \mathbf{0}, \\ \mathbf{V}(\boldsymbol{\varepsilon}_t) &= \boldsymbol{\Sigma}, \qquad \text{and} \\ \mathbf{C}(\mathbf{y}_t, \mathbf{y}_s) &= \mathbf{0} \qquad \text{if } t \neq s. \end{aligned}$$

If $\mathbf{y}_t \sim VAR(p)$, then each component of $\mathbf{y}_t$ can be shown to have a univariate $ARIMA(Mp, (M-1)p)$ representation, with the same AR parameters for each component. Consider, for example, the special case $M = 2$ and $p = 1$, where $\mathbf{y}_t \equiv (y_t, x_t)'$, $\boldsymbol{\varepsilon}_t = (u_t, v_t)'$, $\boldsymbol{\alpha} = \mathbf{0}$ and

$$\mathbf{B}_1 = \begin{bmatrix} \beta_1 & \boldsymbol{\beta}_2 \\ \phi_1 & \phi_2 \end{bmatrix},$$

i.e.,

$$\begin{aligned} y_t &= \beta_1 y_{t-1} + \beta_2 x_{t-1} + u_t, \\ x_t &= \phi_1 y_{t-1} + \phi_2 x_{t-1} + v_t. \end{aligned}$$

Since the first equation for y_t implies

$$\phi_2 y_{t-1} = \beta_1 \phi_2 y_{t-2} + \beta_2 \phi_2 x_{t-2} + \phi_2 u_{t-1},$$

subtracting this from the original equation for y_t yields

$$y_t - \phi_2 y_{t-1} = \beta_1 \left(y_{t-1} - \phi_2 y_{t-2}\right) + \beta_2 \left(x_{t-1} - \phi_2 x_{t-2}\right) + \left(u_t - \phi_2 u_{t-1}\right). \qquad (*)$$

But from the equation for x_t,

$$x_t - \phi_2 x_{t-1} = \phi_1 y_{t-1} + v_t,$$

which, when substituted into (*), yields

$$y_t - \phi_2 y_{t-1} = \beta_1 \left(y_{t-1} - \phi_2 y_{t-2}\right) + \beta_2 \left(\phi_1 y_{t-1} + v_t\right) + \left(u_t - \phi_2 u_{t-1}\right),$$

or

$$y_t = \left(\phi_2 + \beta_1 + \beta_2 \phi_1\right) y_{t-1} - \left(\beta_1 \phi_2\right) y_{t-2} + w_t,$$

where

$$\begin{aligned} w_t &= u_t + \beta_2 v_t - \phi_2 u_{t-1} \\ &\sim MA(1), \end{aligned}$$

since the autocovariances of w_t are zero after the first lag. So $y_t \sim ARMA(2,1)$, and the same algebra can be used to show that the univariate process for x_t is also $ARMA(2,1)$, with the same autoregressive coefficients but a different $MA(1)$ component.

Nonstationarity, Detrending, and Differencing

Most observed time series for levels (or logarithms) of economic variables do not appear to be stationary; generally the mean value of the process appears to increase over time. The traditional means to accommodate such "drift" in the level of the process over time was by use of a *trend-stationary dynamic*

model, in which a linear (or polynomial) function of the time index t is appended to the right-hand side of a stationary ARMA process. For example, a trend-stationary AR(1) model

$$y_t = \alpha + \delta \cdot t + \phi y_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim WN(\sigma^2)$ could be rewritten as

$$\begin{aligned} y_t^* &\equiv y_t - (1-\phi)^{-1}\left[\left(\alpha + \frac{\delta}{1-\phi}\right) + \delta \cdot t\right] \\ &= \phi y_{t-1}^* + \varepsilon_t, \end{aligned}$$

so the detrended series y_t^* follows a stationary AR(1) process. Traditional practice would be to apply time series estimation methods to the residuals of a regression of y_t on a constant term and time.

However, for many economic time series, the variability of the process also appears to increase over time, just as for its level. This led Box and Jenkins to propose *differencing* rather than detrending to transform nonstationary series into (hopefully) stationary versions. The *first difference operator* Δ is defined as

$$\Delta y_t \equiv y_t - y_{t-1};$$

a d^{th}*-order difference* is defined by the recursion relation

$$\Delta^d y_t \equiv \Delta\left(\Delta^{d-1} y_t\right)$$

for any $d > 1$. A series y_t for which $\Delta^d y_t$ is covariance stationary is called an *integrated process of order* d, and denoted $y_t \sim I(d)$.

A simple example of a first-order integrated process is a *random walk with drift*, for which

$$\Delta y_t = \gamma + \varepsilon_t, \qquad \varepsilon_t \sim WN(\sigma^2).$$

Assuming this representation is correct for all positive integers t, with y_0 taken to be a fixed (nonrandom) initial condition, the level of the process y_t has the form

$$y_t = y_0 + \gamma t + \sum_{s=1}^{t} \varepsilon_s.$$

Like the trend-stationary process, the mean of y_t is linear in t,

$$E(y_t) = y_0 + \gamma t,$$

but the variance of y_t also increases linearly (actually, proportionally) in t,

$$Var(y_t) = \sigma^2 t,$$

which is often a more realistic model for observed series.

Generalizations of the random walk model replace the white noise error term ε_t with a more general covariance stationary process u_t and the first difference with a d^{th} difference. When the d^{th} difference of y_t is an ARMA process, i.e., when $\Delta^d y_t \sim ARMA(p,q)$, then the level y_t of the process is said to satisfy an ARIMA model, denoted $y_t \sim ARIMA(p,d,q)$. Much of Box and Jenkins' statistical methodology was devoted to procedures to select the "best" ARIMA model – i.e., the best values of p, d, and q – for a given process, an objective that they termed "ARIMA identification."

Estimation of Time Series Models

Estimation of the parameters of $ARMA(p,q)$ models (and thus of $ARIMA(p,d,q)$ models, after appropriate differencing of the dependent variable) can be based upon *nonlinear least squares*, in which a sum of squared residuals is minimized over the possible values of the unknown coefficients. Letting $\boldsymbol{\beta} \equiv (\alpha, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q)'$ denote those coefficients for the $ARMA(p,q)$ specification, the error terms $\varepsilon_t \equiv \varepsilon_t(\boldsymbol{\beta})$ can be written recursively in terms of $p+1$ current and past values of y_t and q past values of ε_t:

$$\varepsilon_t(\boldsymbol{\beta}) = y_t - (\alpha + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1}(\boldsymbol{\beta}) + ... + \theta_q \varepsilon_{t-q}(\boldsymbol{\beta})).$$

To start this recursion for a particular choice of $\boldsymbol{\beta}$ values, we can treat the initial p values of y_t as fixed – essentially conditioning on these initial values – and set the corresponding ε_t values for these initial time periods equal to zero, their unconditional expectation. (More sophisticated procedures would exploit the relationship between these initial values of y_t and the unknown $\boldsymbol{\beta}$ parameters, but such refinements, involving only a fixed number p of observations, would not affect the asymptotic distribution of the estimators.) The nonlinear least squares procedure would then estimate the unknown parameters by minimizing the conditional sum of squares criterion

$$CSS(\boldsymbol{\beta}) \equiv \sum_{t=p+1}^{T} (\varepsilon_t(\boldsymbol{\beta}))^2$$

over $\boldsymbol{\beta}$, and would estimate the variance σ^2 of the white noise error terms ε_t by $\hat{\sigma}^2 = T^{-1} CSS(\hat{\boldsymbol{\beta}})$. The corresponding estimator $\hat{\boldsymbol{\beta}}$ cannot generally be written in closed form, since the $\boldsymbol{\beta}$ parameters enter the residuals nonlinearly through the lagged residual terms; while derivation of the asymptotic properties

of such estimators would require more asymptotic theory (for extremum estimation), their asymptotic distribution will be similar to that for a linear least squares estimator, with derivative vector

$$\tilde{\mathbf{x}}_t \equiv \frac{\partial \varepsilon_t(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

replacing the usual vector of regressors $\mathbf{x}_t$ in the asymptotic variance formulae. (Such a result would require stronger regularity conditions on the model, e.g., the assumption that ε_t is i.i.d., not just serially uncorrelated with common mean and variance.)

When there are no moving average components – i.e., when $q = 0$, so that the model is purely autoregressive – then the residuals $\varepsilon_t(\boldsymbol{\beta})$ are *linear* in the unknown coefficients $\boldsymbol{\beta} \equiv (\alpha, \phi_1, ..., \phi_p)'$ coefficients, and minimization of the $CSS(\boldsymbol{\beta})$ criterion reduces to *linear least squares* regression of the $(T-p)$ dimensional vector $\mathbf{y}$ on the $(T-p) \times (p+1)$ matrix $\mathbf{X}$, where

$$\mathbf{y} \equiv \begin{pmatrix} y_{p+1} \\ y_{p+2} \\ ... \\ y_T \end{pmatrix} \qquad \text{and} \qquad \mathbf{X} \equiv \begin{bmatrix} 1 & y_p & ... & y_1 \\ 1 & y_{p+1} & ... & y_2 \\ ... & ... & ... & ... \\ 1 & y_{T-1} & ... & y_{T-p} \end{bmatrix}.$$

The usual finite-sample properties of least squares for the classical regression model will clearly not apply. For example, even if ε_t is i.i.d., the mean of $\mathbf{y}$ will not be linear in $\mathbf{X}$,

$$E[\mathbf{y}|\mathbf{X}] \neq \mathbf{X}\boldsymbol{\beta},$$

because of the overlap of components of $\mathbf{X}$ and lagged components of $\mathbf{y}$, so the usual demonstration of unbiasedness of least squares fails. Still, assuming the ε_t are i.i.d. and other regularity conditions (higher order moments, etc.) hold, the appropriate ergodic theorems and central limit theorems can be invoked to show that $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal, with approximate distribution

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, \hat{\sigma}^2 \mathbf{X}'\mathbf{X}),$$

where, again, $\hat{\sigma}^2 = T^{-1} CSS(\hat{\boldsymbol{\beta}})$. If, in addition, the ε_t is assumed to be normally distributed, then the least squares estimation is also a *conditional maximum likelihood* estimator (conditioning on the initial observations $y_1, ..., y_p$), and it thus inherits the efficiency properties of maximum likelihood estimators. The same general results apply for *vector autoregressions* – namely, that consistent estimation of their parameters can be based upon least squares regressions, equation by equation, and the estimators will be asymptotically normal, with asymptotic covariance matrices given from the usual LS formulae.

Asymptotic Normality of LS for AR(1)

To see how such asymptotic results might be obtained, it is useful to consider the simplest autoregressive model, a stationary $AR(1)$ model with zero intercept,

$$y_t = \beta y_{t-1} + \varepsilon_t,$$

where the initial value y_0 of y_t is assumed known, and where the error terms ε_t are assumed to be i.i.d. with expectation zero with plenty of well-behaved moments. Here the (linear) least squares estimator of β is clearly

$$\hat{\beta} = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Making the usual substitution of $\beta y_{t-1} + \varepsilon_t$ for y_t, rearranging terms, and multiplying by $\sqrt{T}$, we get

$$\sqrt{T}(\hat{\beta} - \beta) = \frac{\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1}\varepsilon_t}{\frac{1}{T} \sum_{t=1}^{T} y_{t-1}^2}.$$

The denominator is a sample analogue of the second (central) moment of y_{t-1}; applying a suitable law of large numbers for dependent data will yield

$$\frac{1}{T} \sum_{t=1}^{T} y_{t-1}^2 \xrightarrow{p} E[y_{t-1}^2] = Var(y_{t-1}) = \frac{\sigma_\varepsilon^2}{1-\beta^2}$$

by stationarity. And the numerator is a normalized average of mean-zero, serially-uncorrelated random variables $y_{t-1}\varepsilon_t$ with constant variance$\sqrt{T}(\hat{\beta} - \beta)$

$$\begin{aligned} V_0 &= Var(y_{t-1}\varepsilon_t) \\ &= E[y_{t-1}^2 \varepsilon_t^2] \\ &= E[y_{t-1}^2] E[\varepsilon_t^2] \\ &= \frac{\left[\sigma_\varepsilon^2\right]^2}{1-\beta^2} \end{aligned}$$

so a suitable central limit theorem can be invoked to show that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} y_{t-1}\varepsilon_t \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^4}{1-\beta^2}\right).$$

Application of Slutzky's Theorem gives an expression for the asymptotic (normal) distribution of $\hat{\beta}$

$$\sqrt{T}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}\left(0, 1-\beta^2\right),$$

and a consistent estimator of the asymptotic variance is obviously easy to obtain, particularly under the null hypothesis $H_0 : \beta = 0$, under which

$$\sqrt{T}\hat{\beta} \xrightarrow{d} \mathcal{N}(0, 1).$$

Note that, as $\beta \to 1$, the asymptotic distribution of the normalized estimator $\sqrt{T}(\hat{\beta} - \beta)$ approaches a degenerate distribution at zero. In fact, under some slightly different conditions (like $y_0 = 0$ and is nonrandom), when $\beta = 1$ the least squares estimator has asymptotic distribution

$$T(\hat{\beta} - 1) \xrightarrow{d} \mathcal{DF},$$

where $\mathcal{DF}$ is a non-normal limiting distribution, the "Dickey-Fuller coefficient" distribution. So the asymptotic distribution theory for the LS estimator $\hat{\beta}$ breaks down at $\beta = 1$ – the normalization changes from $\sqrt{T}$ to T and the right approximating distribution is a non-normal distribution that is skewed downward and more variable than a standard normal distribution. This discontinuity implies that the usual asymptotic normal approximation should not be trusted when β may be close to one, and a normal approximation is certainly inappropriate for testing the "unit root" hypothesis $H_0 : \beta = 1$.