Overview

In contrast to the classical linear regression model, in which the components of the dependent variable vector $y$ 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 observations. 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

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

for $\varepsilon_t$ 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

$$y_t = \alpha + \beta y_{t-1} + x_t' \delta + \varepsilon_t,$$

with $x_t$ 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 discrete time stochastic process, i.e., a realization of a random function whose argument is the (integer) time index $t$ (Unless stated otherwise, the discussion here will assume $y_t$ is scalar; the index $t$ is sometimes
restricted to be nonnegative or positive.) More formally, if \((\Omega, \mathcal{F}, P)\) is a probability space with sample space \(\Omega\), set of events \(\mathcal{F}\), and probability measure \(P\), then a discrete time stochastic process \(y_t = y_t(\omega)\) is a mapping from the integers \(\mathbb{N}\) and sample space into the real line \(\mathbb{R}\):

\[
y : \mathbb{N} \times \Omega \to \mathbb{R}.
\]

Unless stated otherwise, it will be assumed that moments of \(y_t\) exist, with

\[
\mu_t = E[y_t], \\
\sigma_t^2 = E[(y_t - \mu_t)^2].
\]

If we could observe a sequence of \(i.i.d.\) realizations of the random variable \(y_t(\omega_i)\) ability space, then by the strong law of large numbers, the \textit{ensemble averages}

\[
y_t = \frac{1}{N} \sum_{i=1}^{N} y_t(\omega_i)
\]

would converge to \(\mu_t\) with probability one as \(N \to \infty\), as would the corresponding sample variance. However, without restrictions on the variation of the joint distribution of \(y_t\) over \(t\) – so that, for example, the means and variances of \(y_t\) could vary freely over \(t\) – it would clearly be impossible to construct consistent estimators of those parameters with a single realization \(\omega = \omega_1\) of history. The concept of \textit{stationarity} imposes such restrictions.

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

\[
E(y_t) = \mu, \\
Var(y_t) = \sigma^2 \equiv \gamma_y(0) \\
Cov(y_t, y_s) = \gamma_y(t-s) = \gamma_y(|t-s|).
\]

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 \textit{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

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

for any nonnegative 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 $y_t \in \mathbb{R}^M$ for some $M > 1$. Such a process is covariance stationary if

$$E(y_t) = \mu,$$

$$\text{Var}(y_t) = \Sigma = \Gamma_y(0),$$

$$C(y_t, y_s) = E \left[ (y_t - \mu)(y_t - \mu)' \right]$$

$$= \Gamma_y(t - s)$$

$$= [\Gamma_y(s - t)]'.
$$

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

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

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 \textit{deterministic processes}, which can be perfectly predicted by a linear combination of past values:

\[
y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots
\]

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 \( \beta \) 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. For weakly stationary processes, a sufficient condition for \( \bar{y}_T \overset{p}{\to} \mu \) is that the autocovariance function declines to zero for large arguments, i.e., \( \gamma_y(s) \to 0 \) as \( s \to \infty \); the proof of this result is a useful exercise. However, 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 be considered; suffice it to say that, for all of the nondeterministic processes considered below, it is possible to find sufficient regularity conditions to ensure that a central limit theorem applies:

\[
\sqrt{T}(\bar{y}_T - \mu) \overset{d}{\to} N(0, V_0),
\]

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

\[
V_0 = \lim_{T \to \infty} Var(\sqrt{T}(\bar{y}_T - \mu)) = \sum_{s=-\infty}^{\infty} \gamma_y(s).
\]

For i.i.d. observations, \( V_0 \) reduces to the usual \( \gamma_y(0) = Var(y_t) \equiv \sigma_y^2 \). All of these results will extend to the case when \( y_t \) is a random vector; in this case, a dependent CLT for vector stochastic processes would yield

\[
\sqrt{T}(\bar{y}_T - \mu) \overset{d}{\to} N(0, V_0),
\]

4
with the asymptotic covariance matrix $V_0$ defined as

$$V_0 \equiv \sum_{s=-\infty}^{\infty} \Gamma_y(s),$$

for $\Gamma_y(s)$ the autocovariance matrix defined above.

**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 – which will be discussed more fully below. 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 $\varepsilon_t$ 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 $\sigma^2$, denoted $\varepsilon_t \sim WN(\sigma^2)$, if it is weakly stationary with

$$E(\varepsilon_t) = 0,$$
$$Var(\varepsilon_t) = \sigma^2,$$
$$Cov(\varepsilon_t, \varepsilon_s) = 0 \quad \text{if} \quad t \neq s.$$

From this simplest example of a weakly stationary and weakly ergodic process (which is strongly stationary if $\varepsilon_t$ 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 $\varepsilon_t$. First are the *moving average processes*, which are linear combinations of white noise "error terms" across different time periods.

*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 $\varepsilon_t \sim WN(\sigma^2)$ (and $\mu$, $\theta$ and $\sigma^2 > 0$ are constants, possibly unknown).
It is easy to see that an $MA(1)$ process is covariance stationary, with mean $E(y_t) = \mu$ and autocovariance function
\[
\begin{align*}
\gamma(0) &= \text{Var}(y_t) = \sigma^2(1 + \theta^2), \\
\gamma(1) &= \text{Cov}(y_t, y_{t-1}) = \sigma^2\theta, \\
\gamma(s) &= \text{Cov}(y_t, y_{t-s}) = 0 \quad \text{if} \quad s > 1.
\end{align*}
\]

Generalizations of the $MA(1)$ process include more lagged white noise components on the right-hand side of the equation for $y_t$.

$q^{th}$-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
\[
y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \ldots + \theta_q\varepsilon_{t-q},
\]
where $\varepsilon_t \sim WN(\sigma^2)$.

Straightforward calculations show that the moments of this $MA(q)$ process are
\[
\begin{align*}
E(y_t) &= \mu, \\
\gamma(0) &= \text{Var}(y_t) = \gamma(0) = \sigma^2(1 + \theta_1^2 + \ldots + \theta_q^2), \\
&= \gamma(1) = \text{Cov}(y_t, y_{t-1}) = \sigma^2(\theta_1\theta_2 + \ldots + \theta_{q-1}\theta_q), \\
& \ldots \\
\gamma(s) &= \text{Cov}(y_t, y_{t-s}) = 0 \quad \text{if} \quad s > q.
\end{align*}
\]

So the autocovariance function $\gamma(0)$ of an $MA(q)$ process “cuts off” after $q$ lags; conversely, if a weakly stationary process has an autocovariance sequence that is zero after $q$ lags, then it can be represented as a $MA(q)$ process for some white noise sequence $\varepsilon_t$. If the white noise components \( \{\varepsilon_t\} \) are mutually independent, and not just serially uncorrelated, then the corresponding $MA(m)$ process is said to be $m$-dependent, since $y_t$ and $y_{t-s}$ are independent if $t$ exceeds $m$.

A more compact notation for the $MA(q)$ equation involves lag polynomials. Defining the lag operator \( L \) as

\[Ly_t = y_{t-1}\]
(i.e., application the lag operator shifts the subscript backward one period), lags of \( y_t \) can be expressed in terms of powers of \( L \):

\[
L^2 y_t \equiv L(Ly_t) = L(y_{t-1}) = y_{t-2},
\]

\[
\ldots
\]

\[
L^s y_t \equiv y_{t-s}.
\]

Similarly, leads of \( y_t \) can be expressed in terms of negative integer powers of \( L \):

\[
L^{-s} y_t \equiv y_{t+s}.
\]

With this notation, a \( MA(q) \) process can be expressed as

\[
y_t = \mu + \theta(L)\varepsilon_t,
\]

where \( \theta(L) \) is the “lag polynomial”

\[
\theta(L) \equiv 1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q,
\]

with the interpretation that

\[
\theta(L)\varepsilon_t \equiv (1 + \theta_1 L + \theta_2 L^2 + \ldots + \theta_q L^q) \varepsilon_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}.
\]

In the jargon of time series analysis, a linear combination of a time series \( y_t \) at different leads and lags is said to be a \textit{linear filter} of \( y_t \), and the lag polynomial \( \theta(L) \) is called a \textit{one-sided linear filter}, since it involves only nonnegative powers of the lag operator \( L \), meaning that it only involves current and past values of the process it premultiplies.

**Autoregressive Processes**

Another simple class of time series models are \textit{purely autoregressive} processes, which only involve a single, contemporaneous white noise term, but instead involve linear combinations of lagged values of the output process (i.e., a one-sided linear filter of \( y_t \)). Unlike moving average processes, these are not weakly stationary for all possible values of the parameters.
First-order Autoregressive Process: The process $y_t$ is first-order autoregressive, denoted $y_t \sim AR(1)$, if it satisfies

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

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

$$E(y_t) = \alpha + \beta E(y_t) \implies E(y_t) = \frac{\alpha}{1 - \beta};$$

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

$$Var(y_t) = \beta^2 Var(y_t) + Var(\varepsilon_t) + 2\beta \cdot Cov(y_{t-1}, \varepsilon_t) \implies Var(y_t) = \beta^2 Var(y_t) + \sigma^2 \implies \underbrace{\frac{\sigma^2}{1 - \beta^2},}_{\text{which is only well-defined and nonnegative if } |\beta| < 1. \text{ This latter condition is sufficient for weak stationarity of } y_t}$$

of $y_t$; calculations analogous to those for the variance yield

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

so the covariance between $y_t$ and $y_{t-s}$ declines geometrically as $s$ increases; if $\beta$ 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

$$y_t = \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t \equiv \beta(L)y_{t-1} + \varepsilon_t,$$

where $\varepsilon_t \sim WN(\sigma^2)$,

$$\beta(L) \equiv \beta_1 + \beta_2 L + \ldots + \beta_p L^p.$$
and $\text{Cov}(\varepsilon_t, y_{t-s}) = 0$ if $s \geq 1$.

Another expression for the $AR(p)$ generating equation is

$$\phi(L)y_t = \alpha + \varepsilon_t,$$

with

$$\phi(L) \equiv 1 - \beta(L) \cdot L,$$

so that a process is autoregressive if a one-sided linear filter of the process is a white noise process (plus a constant term).

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

$$y_t = \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p};$$

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

$$0 = \tilde{\phi}(z) \equiv z^p \cdot \phi(z^{-1})$$

$$\equiv z^p - \beta_1 z^{p-1} - \ldots - \beta_{p-1} z - \beta_p$$

is inside the unit circle, i.e., $|z^*| < 1$. The derivation of this condition for stationarity is the subject of another lecture (as is the form of the autocovariance function for a stationary $AR(p)$ process). Nonetheless, it is useful to check the plausibility if the general stationarity condition by verifying it is the same condition $|\beta| < 1$ for the $AR(1)$ model; in this special case, $\tilde{\phi}(z) = z - \beta$, whose single root $z^* = \beta$ must be less than one in magnitude to ensure stationarity, for the reasons outlined above.

**ARMA Processes**

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

$$y_t = \alpha + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

$$= \alpha + \beta(L)y_{t-1} + \theta(L)\varepsilon_t,$$
or
\[
\phi(L)y_t = \alpha + \theta(L)\varepsilon_t
\]
where \(\varepsilon_t \sim WN(\sigma^2)\) and \(\text{Cov}(\varepsilon_t, y_{t-s}) = 0\) if \(s \geq 1\).

The requirements for stationarity of this process are the same as for stationarity of the corresponding \(AR(p)\) process.

**Linear Processes and the Wold Decomposition**

If we permit the order \(q\) of a \(MA(q)\) process to increase to infinity – that is, if we write
\[
y_t = \mu + \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j}
\]
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*,
\[
\sum_{j=0}^{\infty} \theta_j^2 < \infty,
\]
in which case it is easy to see that
\[
\gamma_y(0) = \text{Var}(y_t) = \sigma^2 \sum_{j=0}^{\infty} \theta_j^2,
\]
and, more generally,
\[
\gamma_y(s) = \sigma^2 \sum_{j=0}^{\infty} \theta_j \theta_{j+s}.
\]

By recursion, stationary ARMA processes can be written as linearly deterministic processes; for example, a stationary AR(1) process \(y_t = \alpha + \beta y_{t-1} + \varepsilon_t\) has \(\theta_s \equiv \beta^s\). 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 decomposed 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 \(\text{Cov}(x_t, z_s) = 0\) 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.
A linearly indeterministic process $y_t$ is said to be a generalized linear process if the white noise components $\{\varepsilon_t\}$ are independently and identically distributed over $t$; it is said to be a linear process if it satisfies the additional restriction that the moving average coefficients are absolutely summable, i.e.,

$$\sum_{j=0}^{\infty} |\theta_j| < \infty.$$  

Since the square-summability condition implies $\theta_j \to 0$ as $j \to \infty$, absolute summability is a stronger requirement than square summability; under this stronger condition, it can be shown that

$$V_0 = \lim_{T \to \infty} \text{Var}(\sqrt{T} \tilde{y}_T) = \sum_{s=-\infty}^{\infty} \gamma_y(s) < \infty,$$

so the stronger requirement that the $y_t$ process is often imposed to ensure applicability of a central limit theorem for $\tilde{y}_T$.

**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 $y_t = \varepsilon_t \sim WN(\sigma^2)$, so that $y_t \sim ARMA(0,0)$. Then for any $\rho$ 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.).
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.