

Time Series Analysis

Class Notes by Prof. H. D. Vinod, Economics Dept. Fordham University, Bronx, New York 10458.
Copyright Prof. Vinod, All rights reserved.

Stochastic Process

If y_t is an observed time series, it is useful to regard it as one realization a stochastic process. The ORDERED sequence of random variables $[\dots, y_{-2}, y_{-1}, y_0, y_1, y_2, \dots]$ defined on an appropriate multidimensional probability space (Ω, \mathcal{E}, P) is called a stochastic process. The set of subscripts $T = \{\dots, -2, -1, 0, 1, 2, \dots\}$ is called an index set. Suppose a die is tossed at times 1, 2 and 3 the set Ω is now a three dimensional space with $\{1, 2, \dots, 6\}$ points along each dimension. Now the stochastic process may be defined by $y_t = \omega$, the outcome at time t where ω is an element of Ω . More generally $y_t = t \omega$, may be the value of t times the outcome. We may define a stochastic process $y(t, \omega)$ as a function of time as well as elements of Ω . The underlying distributions can be complicated multivariate entities. We often assume multivariate normality for convenience. Some stochastic processes are purely random. For example,

White noise stochastic process a_t is simply a normal random variable with zero mean and variance σ^2 , that is

$$a_t \sim N(0, \sigma^2) \text{ for all } t. \text{ Note that } E(a_t a_{t-j}) = 0, \text{ for all } j \neq 0. \quad (1)$$

Why is it called white noise? This name comes from engineering literature where certain gadgets are available which analyze the light waves into components of various frequencies. The spectrum of light for the colors in a rainbow can be seen by these machines. The white light has the property that all frequencies enter equally. It turns out that with truly iid (independent and identically distributed) realizations of random variables $a_t \sim N(0,1)$ as input into these gadgets one obtains a flat spectrum for the light waves which is similar to that of white light. Just as all rainbow colors can be obtained from white color, many stationary processes can be written as linear combinations of white noise processes. White noise process is stationary with zero mean and is uncorrelated over time. If normality is present it is strictly stationary, otherwise it is covariance (second-order) stationary or weakly stationary.

Autocovariance of a stochastic process and Ergodicity

The autocovariance γ_k of y_t is defined for lag k as $E(y_t, y_{t-k})$. Clearly this expectation will be different for different lags k . For economic times series like the GNP we can expect the autocovariance to be large when k is small. When $k=0$, γ_0 becomes the variance. In practice, we have a single realization of the time series and we wish to make consistent estimates of autocovariances and unknown parameters of a joint probability distribution. In traditional probability theory, by consistent we mean that as the sample size increases, the amount of useful information increases. The assumption of ergodicity rigorously adapts the familiar notion of consistency for time series analysis. Further strengthening of convergence is called uniform convergence which lets the sequences of functions inherit important and useful properties of continuity and integrability.

Hamilton (1994, p.46) whether time averages eventually converge to ensemble averages has to do with ergodicity. If autocovariances satisfy $\sum_{i=1}^* |\gamma_j| < \infty$.

Exercise: From the definition of ordinary correlation verify that the autocorrelation coefficient is

$$\rho_k = E(y_t y_{t-k}) / \sqrt{E(y_t^2)E(y_{t-k}^2)} = \gamma_k / \sqrt{\gamma_0 \gamma_0} = \gamma_k / \gamma_0 \quad (2)$$

What modification do we need if the assumption $E y_t = 0$ is replaced by $E y_t = \mu$?

Sample autocorrelation function (SACF) is obtained from a set of sample autocorrelation coefficients r_k based on n observations is defined by:

$$r_k = \frac{\sum_{t=k+1}^n (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^n (y_t - \bar{y})^2} \quad \text{where } k=1,2,\dots$$

If the observations come from a population having $\rho_k=0$, the approximate sampling variance of r_k is n^{-1} , Box and Jenkins (1976, ch.2). Bartlett (1946) proved that under the null hypothesis of $\rho_k=0$ for $k>q$ the variance $V(r_k)=n^{-1}(1+2\rho_1^2 + \dots + 2\rho_k^2)$. It is customary to plot r_k against k along with confidence band based on $\pm 2\sqrt{V(r_k)}$ where $V(r_k)=n^{-1}$ based on Box-Jenkins approximation or upon replacing the ρ_k by r_k in the above formula for $V(r_k)$.

Stationary process

The plot of a time series over a time interval $[t, t+h]$ may sometimes closely resemble a plot at another interval $[s, s+h]$. This implies that there is a temporal homogeneity in the behavior of the series, which is called stationarity. For example, the number of personal bankruptcies may be stationary in monthly data. Intuitively this means for example, that the time series between January to March in one year may resemble June to August of another year. A stationary series should have no discernible trends. By contrast non-stationary series have trends. GNP is a good example of a non-stationary series. For a precise definition of stationarity, some concepts from probability theory, which are developed later, are needed. An imprecise operational definition of stationary time series is as follows: When the mean: $E(X_t)$, the variance: $\text{Var}(X_t)$, and all autocovariances of specified lags (say h): $\text{cov}(X_t, X_{t+h})$ do not depend on t , the time at which they are measured, we have a stationary series.

Strict Stationarity

Let $I^* = [-T, T]$ be a set of integers. Let $I_1 = [t_1, \dots, t_T]$ and $I_2 = [t_1 + h, \dots, t_T + h]$ be two subsets of I , where h is an integer. A stochastic process X_t defined over I_1 has T random variables (without requiring that the integers in I_1 be in any sequential order), has a joint density function $f(X_t \text{ for } t \in I_1)$. A time series is strictly stationary if $f(X_t \text{ for } t \in I_1) = f(X_t \text{ for } t \in I_2)$ for every T , every h , and every subset t_1, \dots, t_T .

If $I^\infty = [\dots, -2, -1, 0, 1, \dots]$ is a set of all integers the set $[X_t \text{ } t \in I^\infty]$ is infinite dimensional. The definition of strict stationarity may be applied to such discrete time series, or even to the continuous counterparts defined over the real line. To characterize the joint density function $f(X_t)$ for $t \in I^\infty$ it is better to use the characteristic function rather than moments because the latter may not exist.

Weak Stationarity or Covariance Stationarity

A weaker concept of stationarity allows the joint distributions to change somewhat over time, but requires that $E(X_t)$, $\text{Var}(X_t)$ do not change. Also, $\text{cov}(X_t, X_{t+h})$ is required to be a function of lag length h only, not depend on time t at which it is measured. One should distinguish between stationarity of a stochastic process and that of a realized time series. A strictly stationary stochastic process with first two finite moments is also weakly stationary, but a strictly stationary time series may not be weakly stationary because its moments need not exist (i.e. be finite). Many stationary series are also called covariance stationary, wide-sense stationary, or second order stationary in the literature. For the multivariate normal joint distribution, weak and strict stationarity are equivalent.

Most economic time series in their original form are usually non stationary, and are sometimes “differenced”, i.e., replaced by $Y_t = X_{t+1} - X_t$, or detrended with $Y_t = X_t - f(t)$, where $f(t)$ is an appropriate function of time representing the deterministic evolutionary trend in the series. The concepts of differencing and detrending will be discussed later.

For a stochastic process to be weakly stationary – also called covariance stationary – there are three requirements:

- (i) $E(y_t) = \mu < \infty$ for all t , and is not a function of t .
- (ii) $E(y_t - \mu)^2 < \infty$ for all t , and is not a function of t .
- (iii) $E(y_t - \mu)(y_{t-k} - \mu) = \gamma_k < \infty$ for all t , and is not a function of t

Strict stationarity is a stronger concept where the properties are unaffected by a change of time origin. It requires that the joint distribution function F

$$F[y(t_1), y(t_2), \dots, y(t_n)] = F[y(t_1+k), y(t_2+k), \dots, y(t_n+k)] \quad (3)$$

for all choices of time points y_1 to y_n and for all k . If the first two moments of a strictly stationary process exist, (i.e., they are finite) it is also covariance stationary and it satisfies the three conditions (2). For the white noise process of (1) covariance stationarity and strict stationarity are equivalent. This follows from the property of the normal distribution that it is completely described by only mean and variance.

Nonstationary time series or stochastic process does not satisfy (2) or (3). It usually evolves over time and critically depends on time. It may have trends in its mean or variance. Many economic time series are nonstationary, and some transformation such as differencing or detrending is often needed to make them stationary. Stationarity implies stable relations and the object of any theory is to obtain stable relationships among variables. Economic equilibria are often characterized by stable long term relations, even though there may fluctuations around the equilibria in the short run.

If one wishes to deal with nonstationary sequences, certain so-called mixing properties are important. They are uniform mixing or ϕ -mixing and α -mixing discussed in White (1984) and Spanos (1986, ch. 8) which define various forms of asymptotic independence.

Exercise Show that $y_t = a+bt+\epsilon_t$ is nonstationary. Hint: compute the $E(y_t)$ and covariance $E(y_t, y_{t-k})$ to show that they are dependent on time t .

Certain Dichotomies

1) Deterministic, non-deterministic (stochastic)

A deterministic function of time, say $f(t)$, is such that its value at specified time t is known exactly. For example, a person's salary may be determined according to the number of years worked, perhaps in certain civil service jobs. Non deterministic functions are more common where certain random (stochastic) influences are present.

2) Discrete and Continuous Stochastic Process

Consider the time variable t defined on an infinite interval I from the real line ($-\infty$ to ∞). For example $I^* = [-T, T]$, a closed interval, is a subset of I . The time is said to be continuous if it is defined at each real number in the interval and discrete if it is defined only at specified integral time values (e.g. distinct integers). The statistical theory of time series is based on random variables X_t having probability distribution at each $t \in I$. The observed x_t value is regarded as one realization from the collection of all possible realizations (so-called ensemble). A stochastic process has as many random variables as there are elements in the interval over which it is defined, and a joint distribution of all these random variables is considered. Since the form of the distribution function of the random variable may be unknown, it is customary to summarize it by its first two moments, or its characteristic function. For both discrete and random processes we may need infinite dimensional objects, and spaces of functions defined over them. This requires the use of Measure Theory.

3) iid and Non-iid Observations

The reader is assumed to be familiar with independent and identically distributed (iid) random variables. For example, the irregular component or error term of a time series is often said to iid normal with zero mean and variance σ^2 or $N(0, \sigma^2)$ in standard notation. The adjacent values in a time series are often correlated with each other, i.e. $\text{cov}(X_t, X_{t+h})$ is generally non-zero. This implies that X_t is not iid.

Variance of Product of Independent Random Variables

$$\text{Var}(X_1 X_2) = [E(X_2)]^2 \text{Var}(X_1) + [E(X_1)]^2 \text{Var}(X_2) + \text{Var}(X_1) \text{Var}(X_2)$$

Conditional Expectation

We define the conditional probability (density) function of X_1 given that the r.v. X_2 has the particular value x_2 to be

$$p(X_1 = x_1 | X_2 = x_2) = \frac{f_{12}(x_1, x_2)}{f_2(x_2)}$$

where Pr denotes probability, f_{12} is the joint probability function of (X_1, X_2) , f_2 is the (marginal) probability function of X_2 , and $f_2(x_2) > 0$.

Conditional expectation is simply the expectation over the conditional probability density function.

Law of Iterated Expectations

In Sargent (1979, p. 208) macroeconomics text a proof of a similar result is given. In simple cases macro economists write $E_{t-2}(E_{t-1}X_t) = E_{t-2}X_t$, where the subscript on E indicates the time at which the expectation is made.

Two period expectation (Exp of Exp) at time zero of one-period expectation regarding time 2 is simply the original two period expectation regarding time 2 made at time 0.

Hamilton (1994, p. 742) notes that conditional expectation $E(Y|X)$ depends on the random variable X. If we view $E(Y|X)$ itself as a random variable and take its expectation w.r.t. the distribution of X, we have iterated expectations:

$$E_X[E_{Y|X}(Y|X)] = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right] f_X(x) dx$$

conditional density equals joint divided by marginal, or: $f_{Y|X} = f_{XY} / f_X$. Hence iterated expectation is

$$\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f_{XY} dy dx \right] = \int_{-\infty}^{\infty} y f_Y dy = E_Y(Y). \text{ Thus the random variable } E(Y|X) \text{ has the same expectation as the random variable } Y.$$

Law of Iterated Projections

Hamilton (1994, p. 100) notes that if we project the (projection of Y_3 given Y_1 and Y_2) itself on Y_1 is simply the projection of Y_3 on Y_1 . This is also used in Sargent.

Concepts Related to Probability Theory

1) White Noise

This name comes from engineering literature where certain gadgets are available which analyze the light waves into components of various frequencies. The spectrum of light for the colors in a rainbow can be seen by these machines. The white light has the property that all frequencies enter equally. It turns out that with truly iid realizations of random variables $a_t \sim N(0,1)$ as input into these gadgets one obtains a flat spectrum for the light waves which is similar to that of white light. Just as all rainbow colors can be obtained from white color, many stationary processes can be written as linear combinations of white noise processes. Recall that Kronecker and Dirac delta functions are used in defining white noise.

2) Random Walk

Consider the iid model $X_t \sim N(0, \sigma^2)$ and consider the cumulative sum based on realizations x_t of the r.v. X_t . A random walker takes steps of size x_t at time t, where the movement may be in positive or negative direction. The walker starts at zero to take a walk based on x_1 at the first time unit.

$$y_1 = x_1$$

$$y_2 = x_1 + x_2$$

.

.

.

$$y_t = x_1 + x_2 + \dots + x_t$$

which can be written in the alternative form:

$$y_t = y_{t-1} + x_t \text{ for } t > 1.$$

The random variable Y_t represents the summation of random variables and indicates the position of the walker at time t . Its expectation is

$$E(Y_t) = E \sum_{i=1}^t x_i = 0, \text{ for all } t$$

and the variance for sum of independent r.v.'s is

$$\text{Var}(Y_t) = \sum \text{Var}(X_i)$$

$= t \sigma^2$. Note that if x_t is $N(0,1)$, y_t is $N(0,t)$, where var. is t not unity as in Wiener process (Brownian motion).

$$\text{Cov}(Y_t, Y_s) = E(x_1 + x_2 + \dots + x_s)(x_1 + x_2 + \dots + x_t)$$

$$= E x_1^2 + E x_1 x_2 + \dots$$

$$= t \sigma^2 \text{ for } 1 < t < s$$

From our definition of stationarity, this is obviously a non-stationary process because two indicators, the variance and covariance, depend on t . If we consider a confidence interval for Y_t , its width will be seen to increase linearly with t .

Now let us consider a r.v. defined by the first differences of Y_t series $Z_t = Y_t - Y_{t-1}$ for $t > 2$, and note that equation for the alternative form of the random walk given above implies

$$y_t - y_{t-1} = x_t \text{ for } t > 1.$$

Hence Z_t is the same as X_t except that it starts at $t=2$, that is, $Z_t \sim N(0, \sigma^2)$, which is of course stationary.

Random Walk with Drifts

If the r.v. comprising the random walk has a nonzero mean we have $x_t \sim N(\mu, \sigma^2)$ $E X_t = \sum \mu = t\mu$ will be nonzero. The mean will be non-stationary also, and the walker will drift toward it.

Data Transformation to Achieve Stationarity

Recall that a covariance stationary time series is defined by the following properties: $E(y_t) = \mu$ is the mean, $\text{Var}(y_t) = \sigma_y^2$ is the variance for all t , and $\text{Cov}(y_t y_{t-k}) = \sigma_y^2 \rho_k$ is the covariance for all t and k where ρ_k is the serial correlation coefficient of order k . The point is that the covariance does not change with time. Sometimes the shorter expression "stationary" is used instead of the more precise "covariance stationary".

Trend Removal by Regression or Differencing Compared

Nonstationary time-series usually have trends in them which make their means and covariances time dependent. They need to be de-trended by one of two procedures before further analysis based on the statistical properties of stationary time series can be performed :

1. Performing regressions on time variable to make the residuals stationary.
2. Repeated differencing.

We discuss the regression approach first. Let us assume that the series y_t is generated by

$$y_t = f(t) + u_t \tag{1}$$

where $f(t)$ is called the trend and the error term u_t is $STN(0, \sigma_u^2)$, that is, a stationary series with zero mean and variance σ_u^2 . The following types of trends have been used in the literature: polynomial, exponential, logistic, explosive growth, and S-curve trend in (1) to (5) of §3.1 above, among others.

Let us suppose that $f(t)$ is linear and let

$$y_t = b_0 + \beta t + u_t \text{ (also called LTSP model below)} \quad (2)$$

be a nonstationary linear trend model. Note that the de-trended series is simply \hat{u}_t the least squares regression residuals, that add up to zero $\sum \hat{u}_t = 0$, and are orthogonal to the time regressor t in the usual sense. Since there are no lagged values on the right hand side of (2), it is intuitively apparent that the shocks do not last a long time in such trend stationary models.

The second method of de-trending uses successive differencing. It exploits a well-known mathematical fact that a polynomial of p -th degree is flattened by considering p -th successive differences. Following exercises checks this result.

Exercise using a Computer: It is convenient to state the problem in the notation of GAUSS computing language. Define $t = \text{seqa}(1, 100, 1)$; a sequence of numbers from 1 to 100 and denote by $\text{rndn}(100, 1)$ a call to a unit normal random number generator to create a vector of 100 such numbers. In PCGIVE the notation is $x = \text{rannormal}(0, 1)$ Next define 3 vectors:

$$y1 = 1 + 2 * t + .001 * \text{rndn}(100, 1); \text{ (equation (2) with } p=1 \text{ here)}$$

$$y2 = 1 + 2 * t + 3 * t^2 + .001 * \text{rndn}(100, 1); \text{ (} p=2 \text{ here)}$$

$$y3 = 1 + 2 * t + 3 * t^2 + 4 * t^3 + .001 * \text{rndn}(100, 1); \text{ (} p=3 \text{ here)}$$

These are polynomials of degree $p=1, 2, 3$ respectively with a small (.001 times) normal random number (denoted rndn) added to each. Compute the first difference of $y1$ ($\Delta y1$) and note that it is detrended (flattened). Similarly second successive first difference of $y2$ ($\Delta^2 y2$) detrends $y2$. For detrending the cubic defined by $y3$ above, the differencing operation Δ is applied three times.

If the data are generated by the logistic, explosive growth, or S-curve trend, verify that differencing can make strange transformation of the data, and certainly does not detrend it. Also, verify that over-differencing can be misleading. • (Bullet symbol suggests the end of a long exercise).

Exercise Practice the differencing operator. Start with $y1$ to $y6$, say and evaluate the higher order differences.

The y_t from (2), which is linear (polynomial order $p=1$) is flattened by considering the first differences to eliminate the trend. We get

$$\Delta y_t = y_t - y_{t-1} = \beta + u_t - u_{t-1} \quad (3)$$

where β is now called a drift parameter.

Exercise: Verify that (3) is covariance stationary. Hint: write $E(\Delta y_t) = \beta$ and the variance, $\text{Var}(\Delta y_t) = 2\sigma_u^2$ and similarly the covariance $\text{Cov}(\Delta y_t, \Delta y_{t-k})$ to verify that they are no longer functions of time. •

To eliminate the drift parameter β in (3) we can either re-center the data y or take a first difference once again to yield

$$\Delta^2 y_t = \Delta^2 u_t = u_t - 2u_{t-1} + u_{t-2} \quad (4)$$

as a de-trended series.

Exercise: Prove that the process in (4) is covariance stationary by showing that γ_k defined from the right hand side of (4) does not depend on time t . •

Let us rewrite (3) as a random walk model with drift:

$$y_t - y_{t-1} = \beta + \epsilon_t \quad (\text{also called the FDSP model below}) \quad (5)$$

where $\epsilon_t = u_t - u_{t-1}$ is $\text{STN}(0, \sigma_\epsilon^2)$, a stationary series with mean zero and variance σ_ϵ^2 . Imagine a person starting a walk at the origin on the real line at time 0. At time 1 he walks to $y_1 = \beta + \epsilon_1$, and at time $t=2$ he ends up at $y_2 = 2\beta + \epsilon_1 + \epsilon_2$ and so on to $y_t = t\beta + \epsilon_1 + \epsilon_2 + \dots + \epsilon_t$. In this case the first difference of y_t on the left hand side of (5) is stationary with mean β because the covariance of the right hand side of (5) can be shown to be independent of time t . More generally, we consider a random walker starting with an initial value y_0 rather than 0.

Unlike the trend stationary model (2) having a single error term ϵ_t , observe that Random Walk (RW) models based on (5) have lagged values present, which lead to the presence of partial sums of error terms:

$$y_t = y_0 + \beta t + \sum_{j=0}^{t-1} \epsilon_j \quad \text{where } \epsilon_0 \equiv 0 \quad (6)$$

which has the same form as the nonstationary linear trend model of (2) except that the disturbance is now a summation to t , hence it depends on time t . Of course, this disturbance is not stationary, its variance $t\sigma_\epsilon^2$ increases over time. Nelson and Plosser have specific names for the two linear trend models of (2) and (5). The model in (2) is called (linear) trend-stationary processes (LTSP) and the model in (5) is called (first) difference stationary processes (FDSP). Observe that the appropriate method of eliminating the trend in (2) is regression and for (5) it is differencing. Next, we consider the practical problem is choosing between the two methods by means of a statistical test. The issues related to possible over differencing are discussed by Nelson and Kang(1981,84).

Statistical Test for Comparing Trend Removal by Regression with Differencing

We can test the null hypothesis that a time series belongs to the LTSP class (2) against the alternative that it belongs to the FDSP class (5) as follows. Nelson and Plosser suggest a test based on an auxiliary regression which encompasses the two choices:

$$y_t = b_0 + \rho y_{t-1} + \beta t + \epsilon_t \quad (7)$$

which belongs to the FDSP of (5) when $\rho = 1$, $\beta = 0$ and the LTSP class (2) when $|\rho|=0$. It is stationary if $|\rho| < 1$. If $|\rho| > 1$ neither of the two de-trending methods can induce stationarity and we have the so-called explosive case. In the auxiliary regression (7) it is tempting to test the joint null hypothesis $\rho = 1, \beta = 0$ against $\rho < 1$ by the usual F test. Unfortunately, we cannot use the usual least squares distribution theory because of the presence of the lagged dependent variable y_{t-1} in (7) whose coefficient $|\rho| = 1$.

Dickey and Fuller(1979,1981) show that the least squares estimate of ρ is distributed around a value which is smaller than unity under the FDSP hypothesis, although the negative bias reduces asymptotically, as the number of observations $T \rightarrow \infty$. They suggest a likelihood ratio test and tabulate the significance points for testing the joint null hypothesis $\beta=0$ and $\rho= 1$ in (7). A nonstandard distribution is needed because the limiting distribution is nonnormal which can be expressed as integrals of Brownian motion, Phillips (1987). The presence of intercept leads to practical problems in these tests and the other practical issue is that ϵ_t are often not white noise. Dickey and Fuller (1979) suggest a solution to the latter problem in the form of an augmented Dickey Fuller (ADF) test. This involves a t statistic on y_{t-1} in the augmented model:

$$\Delta y_t = (\rho - 1) y_{t-1} + \sum_i \theta_i \Delta y_{t-i} + \epsilon_t \quad (7a)$$

The asymptotic distribution is the same as in the original Dickey-Fuller case and does not depend on θ_i . However, the researcher must pick the order of autoregression by some data based criterion (e.g. Akaike Information Criterion) or by searching over a set of possible choices. Schwert (1989) notes that this method works better than a nonparametric alternative suggested by Phillips and Perron (1988).

Nelson and Plosser applied the Dickey-Fuller test to a wide range of historical time series for the U.S. economy and found that the FDSP hypothesis was accepted in all cases, with the exception of the unemployment rate. They conclude that for most economic time series the FDSP model is more appropriate, and that the LTSP model would be the relevant one only if we assume that the errors u_t in (2) are highly autocorrelated. In other words differencing does a better job of transforming the nonstationary economic time series into a stationary series for further analysis.

Dickey Fuller Test for the AR(1) Model

In this subsection we explain briefly the idea behind the Dickey-Fuller test mentioned above. We consider the simpler problem of testing the hypothesis $|\rho|= 1$ in the first order autoregressive AR(1) model

$$y_t = b_0 + \rho y_{t-1} + u_t \quad (8)$$

where $u_t \sim N(0,1)$ is a white noise process. This test is also called "testing for unit roots," because the characteristic polynomial in z has only one root which equals ρ . To obtain the characteristic polynomial use the lag operator L and replace it by z^{-1} , where the notation z is standard in systems literature as the variable of the "z-transform". From (8) first write $0=1 - \rho L$ and then $0 = z - \rho$. There is considerable literature on the unit root problem, Dickey, Bell and Miller(1986) provide a survey, and one of the popular tests is the Dickey-Fuller(1981) test. From the solution of (8) as a

stochastic difference equation discussed later we will see that when $\rho=1$ the starting value and the shocks at distant past get weighted by 1 as are the recent shocks.

The standard expression for the asymptotic variance of the least squares estimator $\hat{\rho}$ of the AR(1) model is $(1 - \rho^2)/T$ for $|\rho|<1$ according to Mann and Wald (1943). Otherwise, if our null hypothesis $\rho=1$ is true, this expression for the variance becomes zero. Intuitively, the problem is that it may not make sense to use $\hat{\rho}$ whose asymptotic variance is zero under the null hypothesis $\rho=1$. Hence, one needs to derive the limiting distribution of $\hat{\rho}$ under $H_0, \rho=1$ to apply the test. When the sample size tends to infinity ($T \rightarrow \infty$) the sampling distribution of $\hat{\rho}$ under the null hypothesis ($\rho=1$) is not degenerate, but complicated, Rao (1978). Hence the estimate $\hat{\rho}$ is more accurate when $\rho=1$ than when $|\hat{\rho}|<1$. Regarding the size of T, further work indicates that $T>70$ is needed to get a reliable five percent test, Dickey et. al (1986, p.16).

For testing the joint hypotheses $\rho=1, \beta=0$ in the random walk with drift in (5) Dickey and Fuller(1981) suggest a Likelihood Ratio (LR) test, derive the limiting distribution and present tables for the test. The F-values in Dickey and Fuller are much higher than those in the usual F-tables. For instance, the 5% significance values from the tables presented in Dickey and Fuller, and the corresponding F-values from the standard F-tables (when the numerator d.f. is 2 and the d.f. for denominator is $n - 3$, as in this test) are as follows:

Sample Size n	F-ratio from Dickey-Fuller	F-ratios from Standard F-tables	Dickey et.al's t_μ
25	7.24	3.42	3.33
50	6.73	3.20	3.22
100	6.49	3.10	3.17
∞	6.25	3.00	3.12

 The basic message from this table and underlying theory is that one should be weary about over confidence in the significance of estimates if a model was not differenced, when it should have been differenced. Dickey, et.al.'s(1986) t_μ values are reported in the last column, and are applicable when $\Delta y_t=y_t - y_{t-1}$ is regressed on a vector of ones (for the intercept) and y_{t-1} . The usual t statistic for the coefficient of y_{t-1} is not distributed as Student's t but as t_μ in the above table.

Autoregressive Model of p-th order AR(p)

In this section we define the p-th order autoregressive model which is a generalization of (8). The "auto" in autoregression suggests that the variable is regressed on its own (past) values. First we define $\phi(L)$ as an autoregressive polynomial in the lag operator L to be $1 - \phi_1L - \phi_2L^2 - \dots - \phi_pL^p$. Now the AR(p) model is given by

$$\phi(L) y_t = b_0 + u_t \tag{9}$$

Exercise: Verify that (8) is a special case of (9). For the unit root case show that the root of the polynomial in z and of the polynomial in L (as is) is unity.

Moving Average Model of q-th order MA(q) and Wold Decomposition

The notion of moving average involves averaging over a set of consecutive values of a random variable. It is this set that moves similar to a moving window. For example, one thinks of a weighted averages of $\{u_1, u_2 \text{ and } u_3\}$, $\{u_2, u_3 \text{ and } u_4\}$, $\{u_3, u_4 \text{ and } u_5\}$, \dots , etc.. To define MA(q) let us first define as the moving average polynomial $\theta(L) = 1 - \theta_1L - \theta_2L^2 - \dots - \theta_qL^q$. Now, the MA(q) model is given by

$$y_t = b_0 + \theta(L) u_t \quad (10)$$

where $u_t \sim N(0,1)$ is the white noise process.

Properties of MA(1):

Hamilton (1994, p.48) $y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}$.

$$E(y_t) = \mu, \gamma_0 = \text{Var}(y_t) = (1 + \theta^2)\sigma^2$$

First autocov.

$$\gamma_1 = E(y_t - \mu)(y_{t-1} - \mu) = E(\epsilon_t + \theta\epsilon_{t-1})(\epsilon_{t-1} + \theta\epsilon_{t-2}) = \theta\sigma^2$$

γ_j for $j > 1$ are all zero.

$$|\gamma_0| + |\gamma_1| + |\gamma_2| + \dots = (1 + \theta^2)\sigma^2 + \theta\sigma^2.$$

This is finite so MA(1) is ergodic.

autocorrelation ρ_1 for MA(1) is γ_1/γ_0 or $\theta/(1 + \theta^2)$. Hamilton's fig 3.2 page 51 plots ρ_1 for different values of θ . Positive θ leads to positive ρ_1 . Even if $\theta = 2$, ρ_1 is finite.

Exercise: Construct an MA(∞) model.

Wold (1938) showed that after removing nonstationary components and deterministic components, we have a purely nondeterministic series $(y_t - \mu)$ which can be written as a linear combination of a sequence of uncorrelated random variables. Wold's remarkable theorem implies that almost any time series (after appropriate transformations to achieve stationarity) can be written as a special case of an MA(∞) process. Any MA(q) model is obviously a special case of MA(∞). If we start with an AR(1) model, we can formally invert the polynomial $1 - \phi L$ and write it as an infinite series $1 + w + w^2 + \dots$, with $w = \phi L$. The infinite order polynomial can be interpreted as an MA(∞) model. Then $y_t = b_0(1 - \phi)^{-1} + \sum_{i=0}^{\infty} \phi^i u_{t-i}$. We shall see that this expression for y_t may be viewed as a solution of a stochastic difference equation.

If we start with a mixed AR(1) and MA(1) model denoted by ARMA(1,1) can this also be written as MA(∞) model? ARMA(1,1) is given by

$$(1 - \phi_1 L) y_t = b_0 + (1 - \theta_1 L) u_t \quad (11)$$

It can be verified that the polynomial expression $(1 - \theta_1 L)(1 - \phi_1 L)^{-1}$ can indeed be expressed as an infinite polynomial in the lag operator. Thus the Wold decomposition may be written as:

$$y_t = \sum_{j=0}^{\infty} G_j u_{t-j}, \text{ where } G_0 = 1$$

Invertibility of MA(1): Hamilton(1994, p. 64). If invertible, $|\theta| < 1$, then MA(1) can be written as AR(∞). MA processes have the interesting property that invertible and noninvertible representations

have the same autocovariances. Same autocovariance generating functions (AGF's) are obtained by replacing θ by $(1/\theta)$. If one is larger than unity the other is smaller! Why do we need invertible representation? As a practical matter, for an invertible representation, the past data are used to generate future. If it is noninvertible, one would need future data to predict past, a ridiculous strategy!

Autocorrelation and Partial Autocorrelation Function

Autocorrelation coefficient is the simple correlation coefficient ρ_j between y_t and y_{t-j} for $j=1,2,\dots,M$, where M =maximum lags. Correlation coefficients may be useful for testing seasonality or other forms of periodicity, and as a precursor to choosing the parametric model for the data. We can compute large lag (asymptotic?) standard errors (SE's), and plot autocorrelations at zero plus or minus twice standard errors, $(0 \pm 2 \text{ SE})$. If an individual autocorrelation falls outside these bounds, this suggests that it is significantly different from zero.

In partial autocorrelation plots $(0 \pm 2/\sqrt{T})$ are used as the bounds for determining whether they are significantly different from zero. Partial autocorrelations are defined and discussed in chapter 6. These are useful for determining the number of terms in an AR(p) autoregressive model that are needed to adequately represent the data. For example, if the correct model is AR(2) only the first two partial autocorrelations are significantly different from zero.

Pre processing by Box-Cox Transformation and Data Tapering.

Data processing of time series is often made by some preliminary steps to obtain data sets that conform closely to the assumptions of time series models. Some examples are trend removal, aggregation or stationarity transformation. One danger in this is that spurious relationships and nonexistent dynamics may be introduced, or important existing relations or dynamics may be hidden/removed during preprocessing. Working(1960) gives an example where averaging over successive periods introduces a spurious autocorrelation.

$$\text{Box-Cox Transformation : } y_t^* = \begin{cases} \lambda_1^{-1} [(y_t + \lambda_2)^{\lambda_1} - 1] & \text{if } \lambda_1 > 0 \\ g \ln(y_t + \lambda_2) & \text{if } \lambda_1 = 0 \end{cases}$$

where λ_2 is a number added to the data before the transformation (usually $\lambda_2 = 0$), g is the sample geometric mean of $(y_t + \lambda_2)$, λ_1 governs the strength (power) of the transformation such that when $\lambda_1 = 1$ we simply use the original data without any transformation. When $\lambda_1 = 0$, this is the usual logarithmic transformation. In general, the Box-Cox transformation is well defined for $y_t > 0$ only, because we may have problems with negative y_t since the logarithm of a negative number is not defined. Then a modification is to use $[\text{sign}(y_t)|y_t|^\lambda - 1]/\lambda$, which does not have attractive properties. For further discussion see Davidson and MacKinnon (1993, ch.14)

Data Tapering: The end points of a time series vector are adjusted by using the Bell-shaped part of the cosine curve.

Keynes argued that all sorts of considerations enter the stock market valuations which may have little to do with the underlying yields. Of course, if the markets are working efficiently, one would expect that the yields not depart too much from the fundamental values. Constant returns is an implication of some theories of market efficiency. Early studies surveyed by Fama(1970) favor the random walk in stock markets, and hence the efficient markets hypothesis. However, due to low power of statistical tests there are difficulties in distinguishing the random walk model from some alternative specifications. Fama and French(1988) and others have found negative autocorrelations at 20-year horizons.

Porterba and Summers(1988) advocate the use of a non-parametric test based on variance ratios suggested by Osborne(1959). The idea is that if the returns are uncorrelated through time the return variance should be proportional to the return horizon. The test statistic is that the following ratio should tend to 1

$$\text{Var. Ratio of horizon } k = \text{VR}(k) = \frac{\text{Var.}(R_t^k)}{k} / \frac{\text{Var.}(R_t^{12})}{12} \rightarrow 1 \quad (1)$$

where R_t denotes the aggregated return in t-th month and $R_t^k = \sum_{i=0}^{k-1} R_{t-i}$. The statistic $\text{VR}(k)$ of (1) is designed for monthly data, and uses one year as the reference horizon in the denominator. Poterba and Summers express $\text{VR}(k) = \sum_{i=0}^{k-1} w_i \hat{\rho}_i$ in terms of a weighted sum of sample autocorrelation coefficients $\hat{\rho}_i$. Is the small sample estimate of $\text{VR}(k) = 1$? Kendall and Stuart's(1983) result is that $E\hat{\rho}_j = -1/(T-j)$, where T is the size of the sample in the entire data. Hence the expected value $E[\text{VR}(k)]$ can be approximated by substituting $E\hat{\rho}_i$ in the weighted sum. Since this does not equal 1, the $\text{VR}(k)$ statistic is biased in small samples. However, the bias can be simply removed by dividing the estimated $\text{VR}(k)$ by its approximate $E[\text{VR}(k)]$.

The following variance ratio test is devised by Cochrane(1988). If y_t is a pure random walk, the variance of the k-th difference, $\text{var}(y_t - y_{t-k})$ grows linearly with the difference k . By contrast, if y_t is trend stationary, $\text{var}(y_t - y_{t-1})$ will approach a constant. Define the notation σ_k^2 for $(1/k)$ times $\text{var}(y_t - y_{t-k})$. One computes an unbiased estimate \hat{V} of σ_k^2 / σ_1^2 for various values of k

$$\hat{V} = \frac{\text{var}(y_t - y_{t-k})}{k \text{var}(y_t - y_{t-1})} \cdot \frac{T}{(T-k+1)} \quad \text{and} \quad \text{SE}(\hat{V}) = \hat{V} \left[\frac{4k}{3T} \right]^{1/2} \quad (3.1)$$

where the $T/(T-k+1)$ term corrects for a small sample bias, and where SE denotes the asymptotic standard error of the statistic. In our case, $k=1$ to 30, since we have $T=100$ observations for most countries. Using certain well known results due to Bartlett from the statistical theory of spectral estimation, Cochrane shows that if $k/T \rightarrow 0$ as $T \rightarrow \infty$, the σ_k^2 is a consistent estimator of the spectral density at zero frequency. Hence, he derives the SE of (3.1) using the known asymptotic variance of σ_k^2 . Cochrane suggests plotting \hat{V} , and the confidence bounds $\hat{V} - \text{SE}$, and $\hat{V} + \text{SE}$ against k . These plots eventually go to zero as k increases, if the underlying processes are mean reverting. Unfortunately, most \hat{V} plots show rather large confidence intervals as k becomes large.

When k gets large, the number of observations for whom k -th difference can be defined becomes very small. Hence, it is intuitively stands to reason that the variance should increase.

Persistence of Shocks to GNP

The notion of “potential GNP” in macroeconomics literature is based on the assumption that there is a trend in the GNP which evolves smoothly over time. Detrending of GNP was common until recently when researchers have found evidence that shocks to GNP may last forever. If GNP does revert to a trend it is called mean reversion. More formally, mean reversion is defined as negative serial correlation at all leads and lags. The empirical questions are the existence of mean reversion and the time it takes for it to occur. Cochrane(1988) finds that the mean reversion in log GNP takes longer than the time it takes to complete a business cycle, i.e., several years. Let the errors satisfy $a_t \sim N(0, \sigma^2)$. A Linear trend model is defined by

$$y_t = b_0 + \beta t + u_t \text{ where } u_t = \sum_j \theta_j a_{t-j} \quad (1)$$

It is trend-stationary if the summation in (1) is a stationary stochastic process.

Exercise: Compute the expectation of both sides of (1) to verify that $E(y) = b_0 + \beta t$ is a function of time, and is obviously nonstationary. Verify that upon de-trending what remains is u_t which is stationary by assumption. Generalize this to nonlinear trends. Show that the variance of u_t is $\sum_0^\infty \theta_j^2 \sigma^2$.

By contrast, a random walk model with a drift is defined as:

$$y_t = \beta + y_{t-1} + a_t \quad (2)$$

This is a nonstationary model, and β is called the constant drift. Note that we cannot define an unconditional expectation $E(y)$ for (2), since each y_t depends on y_{t-1} . However, if we assume that some value of y is known, we can condition on that value and compute conditional expectations and k -step ahead forecasts. The conditional expectation

$$E(y_{t+k} | y_{t-1}) = k\beta + y_{t-1} + a_t + a_{t+1} + \dots + a_{t+k}, \text{ and } a_t \sim N(0, \sigma^2) \quad (3)$$

Note that the shock to GNP is measured by the error term a_t . If the shock is negative and given by $a_{t+1} = -1$, the GNP at time $t+1$ will be $y_t - 1$. At time $t+2$ it is $y_t - 1 + a_{t+2}$ and so on. Thus, for the random walk model, the -1 term will remain in the forecast of y_{t+k} forever into the future. By contrast, the model (1) is trend-stationary, and the θ coefficients will be such that the effect of a negative shock in GNP at time $t+1$ will eventually completely die down. For (1), y_{t+k} will completely revert to the trend line. Thus the models (1) and (2) are two extremes according to Campbell and Mankiw's(1987) interpretation.

One unit shock to the GNP affects the long term forecasts of GNP by exactly one unit according to the random walk model (2), and by zero according to the trend-stationary model (1). Of course, the empirical reality may be somewhere in between the two extremes of zero and unity, and may mean a mixture of both (1) and (2). Statistical testing seems to be unable to resolve the issue, due to the limited power of the asymptotic unit root tests. One can model a given shock to a time series as consisting of two components: (i) permanent component arising from the random walk,

and (ii) a transitory component represented by a stationary process. If the first component is small relative to the second, one may use the trend-stationary model (1) in practical work.

In another interpretation, one considers the variance of the random walk component. From (3) it is clear that the variance of the k-th difference

$$\text{Var}(y_{t+k} - y_t) = \text{Var}(a_{t+1}) + \text{Var}(a_{t+2}) + \dots + \text{Var}(a_{t+k}) = k \sigma^2 \quad (4)$$

The variance of y_t around the linear trend in (1) is constant which equals $\sigma^2 \sum_0^\infty \theta_j^2$. From (1)

$$y_{t+k} = b_0 + \beta (t+k) + u_{t+k}, \text{ where } u_t = \sum_0^\infty \theta_j a_{t+k-j} \quad (5)$$

From (1) and (5)

$$y_{t+k} - y_t = \beta k + u_{t+k} - u_t \text{ where } u_{t+k} - u_t = \sum_0^\infty \theta_j (a_{t+k-j} - a_t) \quad (6)$$

Hence the variance of the k-th difference for the model (1) can be obtained from (6) as:

$$\text{Var}(y_{t+k} - y_t) = \text{Var}(u_{t+k} - u_t) = \sum_0^\infty \theta_j^2 E(a_{t+k-j} - a_t)^2 = 2 \sum_0^\infty \theta_j^2 \sigma^2 \quad (7)$$

which also tends to a constant. We have noted earlier that the empirical reality may be a mixture model having both (1) and (2). We are interested in the long term behavior of the mixture model. Hence we consider the limits of (4) and (7) as $k \rightarrow \infty$. From (4) it is obvious that we must divide by k to have a finite limit. The limit of the ratio $\text{Var}(y_{t+k} - y_t)/k$ from the random walk model of (4) is σ^2 . A comparable limit from the trend-stationary model based on (7) tends to zero, because of the division by the k . The limiting constant in (4) is σ^2 , which is also the variance of the first difference, $\text{Var}(y_{t+1} - y_t)$. Thus, from (4)

$$\text{Limit}_{k \rightarrow \infty} (1/k) [\text{Var}(y_{t+k} - y_t)/\text{Var}(y_{t+1} - y_t)] = 1 \text{ (for Random Walk)} \quad (8)$$

Fortunately, the final right hand side of (7) can also be written as $\text{Var}(y_{t+1} - y_t)$. Hence,

$$\text{Limit}_{k \rightarrow \infty} (1/k) [\text{Var}(y_{t+k} - y_t)/\text{Var}(y_{t+1} - y_t)] = 0 \text{ (for Trend-stationary)} \quad (9)$$

In a mixture of the two models, the limit may be somewhere in between 0 and 1. Cochrane (1988) suggests a plot of $[\text{Var}(y_{t+k} - y_t)/k] / \text{Var}(y_{t+1} - y_t)$ against k to see if it tends to 1 or zero. More specifically, he suggests plotting the statistic

$$\hat{V}_k = \frac{\text{var}(y_t - y_{t-k})}{k \text{ var}(y_t - y_{t-1})} \cdot \frac{T}{(T-k+1)} \text{ and } SE = \hat{V}_k \left[\frac{4k}{3T} \right]^{1/2} \quad (10)$$

where the $T/(T-k+1)$ term corrects for a small sample bias, and where SE denotes the asymptotic standard error of the statistic.

Difference Equations and Stochastic Difference Equations.

Difference equations are similar to differential equations and can be written in a form involving the difference operator defined by $\Delta y_t = y_t - y_{t-1}$, where a time series y_t has t defined on the index set

$\mathcal{T} = \{\dots, -2, -1, 0, 1, 2, \dots\}$. The Δy_t is called the "first" difference. Second difference is defined by $\Delta^2 y_t = (1 - L)^2 y_t = (1 - 2L + L^2)y_t = y_t - 2y_{t-1} + y_{t-2}$, which uses the identity $\Delta \equiv (1 - L)$ between the difference operator and the lag operator. In general, $\Delta^p = (1 - L)^p$ will obviously involve the Binomial coefficients. Since the presence of Δ also leads to the presence of lagged terms, typical difference equations are stated in terms of lagged rather than Δ terms.

Linear Difference Equations:

Definition LDE(p) : It is convenient to define the p^{th} order (ordinary) linear difference equation as follows.

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \dots - \phi_p y_{t-p} = a_t \quad (1)$$

where $p \leq t < \infty$, where the coefficients ϕ_i are assumed to be constant, and where a_t is called an input function or a forcing function.

The homogeneous form of the difference equation is defined by setting $a_t \equiv 0$, yielding

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \dots - \phi_p y_{t-p} = 0. \quad (2)$$

The solution of the homogeneous form: This is obtained by first writing (2) in terms of the so-called characteristic polynomial in L , the lag operator written with $(L = z^{-1})$:

$$(1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p) y_t = 0 = (z^p - \phi_1 z^{p-1} - \phi_2 z^{p-2} - \dots - \phi_p) y_t \quad (3)$$

We define the roots $\lambda_i, i=1,2,\dots, p$ of the (autoregressive) polynomial $\phi(z)$ or $\phi(L)$ by writing $(1 - \lambda_1 L)(1 - \lambda_2 L)(1 - \lambda_3 L) \dots (1 - \lambda_p L) y_t = 0 = (z - \lambda_1)(z - \lambda_2)(z - \lambda_3) \dots (z - \lambda_p) y_t \quad (4)$

The solution will contain the λ_i values and the starting value at $t=0$ to be y_0 . For clarity of exposition let us first assume $p=1$, yielding a first order homogeneous difference equation: $y_t - \phi_1 y_{t-1} = 0$. At $t=1$, we have $y_1 = \phi_1 y_0$. At $t=2$, we have $y_2 = \phi_1 y_1 = \phi_1(\phi_1 y_0) = \phi_1^2 y_0$. Similarly at $t=3$ we have $y_3 = \phi_1^3 y_0$. In general the solution of the first order homogeneous difference equation is $y_t = \phi_1^t y_0$.

What is a solution? This is called a solution because it evaluates y_t from the two assumed known quantities ϕ_1 and y_0 . The solutions are a way of learning about the dynamics of the system, since they give a value of y_t for any t . In the case of stochastic difference equations discussed later, the solution involves Green's function values G_j as weights on lagged errors.

First Order Homogeneous Difference Equation: Note that the polynomial (3) with $p=1$ for our first order equation is: $(1 - \phi_1 L) y_t = 0$. This makes it obvious that the root of the autoregressive (AR) characteristic polynomial in $z = L^{-1}$ is $\lambda_1 = \phi_1$. In the solution $y_t = \phi_1^t y_0$ we raise the root ϕ_1 to the power t . If t is large and $|\phi_1| > 1$, it is clear that when such a ϕ_1 is raised to a large power, ϕ_1^t term may become a very large positive or negative number depending on the sign of ϕ_1 . Thus, if $|\phi_1| > 1$, the solution $|y_t| \rightarrow \infty$ as $t \rightarrow \infty$, where we have used the absolute value of y_t because its sign will depend on the sign of the starting value y_0 , which may be negative.

Exercise: Verify that any negative starting value of y_0 will lead to oscillatory behavior with changing signs of successive y_t .

By contrast, if $|\phi_1| < 1$, a fraction, it will become smaller and smaller as it is raised to a higher power. Then the solution $|y_t| \rightarrow 0$ as $t \rightarrow \infty$ if $|\phi_1| < 1$. Finally, if $|\phi_1| \equiv 1$, raising it to any power makes no difference, and the solution is $y_t = y_0$ as $t \rightarrow \infty$ if $\phi_1 = 1$ and $y_t = -y_0$ as $t \rightarrow \infty$ if

$\phi_1 = -1$. It is customary to describe the roots larger than 1 as being outside the unit circle, roots smaller than 1 as being inside the unit circle, and the roots equal to 1 as being on the unit circle. The dynamic system whose solution $|y_t| \rightarrow \infty$ as $t \rightarrow \infty$ is called unstable or non-stationary, and if $|y_t| \rightarrow 0$, it is called stable or stationary. When $\phi_1 = 1$, it is called a random walk model. Since this has considerable importance in econometrics it is discussed separately. For the random walk model the solution $y_t = a_t + a_{t-1} + a_{t-2} \dots$, is such that the random shocks accumulate indefinitely. It can be shown that its Green's function is unity, that is the memory is long and constant, and the process never dies out. The solutions of higher order homogeneous difference equations are analogous, having somewhat complicated expressions involving all λ_i roots of the characteristic polynomial in z defined above in (3).

Now, let us consider the first order non-homogeneous case: $y_t - \phi_1 y_{t-1} = A$, where $a_t \equiv A$ for all t . This is called non-homogeneous because $A \neq 0$. At $t=1$, we have $y_1 = \phi_1 y_0 + A$. At $t=2$, we have $y_2 = \phi_1 y_1 + A = \phi_1(\phi_1 y_0 + A) + A = \phi_1^2 y_0 + \phi_1 A + A$. Similarly at $t=3$ we have $y_3 = \phi_1^3 y_0 + \phi_1^2 A + \phi_1 A + A$. In general the solution of the first order non-homogeneous difference equation is

$$y_t = \phi_1^t y_0 + A \left(\frac{1 - \phi_1^t}{1 - \phi_1} \right) \quad \text{if } \phi_1 \neq 1, \text{ and} \quad (5)$$

$$y_t = y_0 + A t \quad \text{if } \phi_1 = 1. \quad (6)$$

This is called a solution because it evaluates y_t from the three assumed known quantities ϕ_1 , A and y_0 . In deriving the above formula we have assumed that the input $a_t \equiv A$ for all t . If we let a_t be a function of time t , and substitute a_t values at $t=1, 2, 3, \dots$ for A , we get the following solution.

$$y_t = \phi_1^t y_0 + \sum_{j=0}^{t-1} \phi_1^j a_{t-j}. \quad (7)$$

The contribution of y_0 to the solution remains bounded as long as $|\phi_1| < 1$.

Second Order Non-homogeneous Difference Equation and Business Cycles:

Samuelson (1939) applied second order difference equations to explain the business cycles as follows. The accelerator principle states that the investment (I_t) is dependent on the changes in income rather than the level of income: $I_t = b_1(Y_{t-1} - Y_{t-2})$. Consumption (C_t) depends on the past income by the first order non homogeneous equation $C_t = b_2 Y_{t-1} + b_3$. and the accounting identity: $Y_t = C_t + I_t$. Substituting the first two equations in this identity, we have $Y_t = b_1(Y_{t-1} - Y_{t-2}) + b_2 Y_{t-1} + b_3$. For certain values of b_i when the roots of the characteristic polynomial are imaginary, this second order equation can be shown to lead to oscillation in income.

Stochastic difference equations:

So far, we have considered the ordinary (nonstochastic) difference equations. The stochastic difference equations are characterized by the fact that the input function a_t of the difference equation is a white noise random variable with zero mean and variance σ^2 , that is

$$a_t \sim N(0, \sigma^2) \text{ for all } t. \text{ Note that } E(a_t a_{t-j}) = 0, \text{ for all } j \neq 0.$$

First Order AR(1) Stochastic difference equations (homog. driftless):

Consider the first order autoregressive model denoted by AR(1) and defined by the stochastic difference equation for y_t measured from the mean, i.e., $E y_t = 0$.

$$y_t - \phi y_{t-1} = (1 - \phi L) y_t = a_t, \quad \text{where } a_t \sim N(0, \sigma^2) \text{ and } |\phi| < 1 \quad (1)$$

In a higher order autoregressive model ϕ has subscripts 1, 2 etc. Now the solution may be obtained by the brute force method of substitution similar to the one above, and it is also obtained by a formal division of both sides of the equation by the characteristic polynomial

$$y_t = (1 - \phi L)^{-1} a_t = \sum_{j=0}^{\infty} \phi^j a_{t-j} = \sum_{j=0}^{\infty} G_j a_{t-j} \quad (\text{where } G_0=1) \quad (2)$$

and where the coefficients G_j are called Green's functions, Miller (1968). A random shock at time $t - j$ gets multiplied by the weight G_j suggesting that the Green's function measures the extent to which the system remembers a random shock. Hence, G_j may be used to represent the memory of the dynamic system. It is interesting to note that Green's function can be regarded as an "orthogonal decomposition" of y_t first proved by the econometrician H. Wold (1938). To visualize the orthogonality imagine infinite dimensional space with axes marked a_t, a_{t-1}, \dots along which we have the orthogonal (independent) random inputs, and y_t is a vector through the origin.

For future reference we note that the solution of the first order difference equation (1) arising from AR(1) model having one root ($=\phi$) of the characteristic polynomial is given by $G_j = \phi^j$. Note that (2) may be viewed as an MA(∞) representation of the AR(1) model. This avoids explicit statement of initial conditions. The coefficients G_j are also interpreted as impulse response coefficients, $\partial y_t / \partial a_{t-j}$, measuring the effect of a unit perturbation (impulse) a_{t-j} on y_t or equivalently the effect of a_t on y_{t+j} . (See Hamilton, pp. 5-10). The impulse is shown by a vertical bar of height 1, a purely transitory (short lived) change and the path of output is given by G_j which will show a decay when $|\phi| < 1$. The cumulative impulse response measures the long-run effect of the change in a_t as $\phi + \phi^2 + \dots = 1 / (1 - \phi)$. This should not be confused with step-response, where the change or the perturbation a_t is not transitory, but long-lived.

Comments on Mean Reversion in the context of First order difference equation:

How can we use the solution of first order difference equation in Economics? From the underlying economic variable (GNP) y_t it is customary to consider the first difference of logs of y_t as the time series for growth. The growth rate $u_t = \ln y_t - \ln y_{t-1}$, or $(y_t / y_{t-1}) = \exp(u_t)$.

Definition 1: Mean reversion is defined by $y_t \rightarrow \bar{y}$, a constant in equilibrium, as $t \rightarrow \infty$. Then, $(y_t / y_{t-1}) = (\bar{y} / \bar{y}) = 1 = \exp(u_t)$, and mean reversion implies zero growth rate, that is: $u_t \rightarrow 0$.

Definition 2: Trend reversion is defined by the series reverting to a (long term) trend (function of time and constants) $y_t \rightarrow g(t)$ in equilibrium, as $t \rightarrow \infty$. To study the implications of trend reversion on the growth rate u_t we consider two cases:

(i) Assuming that $g^j(t)$, the j -th derivative of $g(t)$ evaluated at the limiting value $t=\infty$ is finite, we can apply L'Hopital's rule to evaluate the limit $(y_t/y_{t-1}) = [g^j(t)/g^j(t-1)] = 1 = \exp(u_t)$. For example, let y_t exhibit a linear trend in the limit, $g(t)=a+bt$. Since $(y_t/y_{t-1}) \rightarrow (\infty/\infty)$, we evaluate this by using the fact that the first derivative $g^1(t)=b$ is a constant for all t . Hence $(y_t/y_{t-1}) \rightarrow (b/b)=1$. Thus, trend reversion of y_t to a linear trend implies that the asymptotic (exponential) growth is zero, $u_t \rightarrow 0$. More generally, limiting value of (y_t/y_{t-1}) for polynomial and/or sinusoidal trends also experience a similar cancellation.

(ii) If the trend function $g(t)$ is exponential, then trend reversion can have nonzero u_t as $t \rightarrow \infty$. For example, if $g(t)=\exp(a+bt)$, then $(y_t/y_{t-1})=\exp(b)$ and $u_t \rightarrow b$.

Trend reversion of GNP is sometimes thought in terms of the GNP returning to a linear, polynomial or sinusoidal growth path after assimilating all shocks. We have shown that such paths asymptotically imply zero asymptotic growth rate. Our figures in the following section will show that $u_t \rightarrow 0$ is a very strong condition, which is difficult to satisfy in practice. When the macro economists refer to mean reversion, they do not mean reversion to \bar{y} in the sense of $y_t \rightarrow \bar{y}$ of definition 1. Our definition 2 of trend reversion, and the discussion of the case (i) above shows that by mean reversion the macro economists cannot imply reversion to a linear, polynomial or certain sinusoidal trends. This reveals a second ambiguity in the literature. When various authors discuss mean reversion for y_t (GNP in equilibrium) we shall show that they are almost never thinking of mean reversion in the sense of definition 1 or definition 2(i). The model of the following section discusses AR(1) and AR(2) models for growth rates u_t to derive limiting values for finite horizons. It shows with graphics that "exponential trend reversion" not "mean reversion" is a more accurate description of the underlying growth process.

Definition 3: Poterba and Summers (JFinEco, 1988,p27)

Mean reversion requires negative serial correlation at some frequency. This is perhaps the most reliable definition.

The Model and Results for Finite Horizons

Let us consider the AR(1) model defined by:

$$u_t = \alpha + \phi u_{t-1} + \epsilon_t \quad (1)$$

which can be written in terms of the lag operator L as: $(1-\phi L) u_t = \alpha + \epsilon_t$, leading to

$$u_t = [\alpha/(1-\phi L)] + [1/(1-\phi L)] \epsilon_t \quad (2)$$

or sum of two components:

$$u_t = \alpha \sum_{i=0}^{t-1} \phi^i + \sum_{i=0}^{t-1} \phi^i \epsilon_{t-i} \quad (3)$$

Given that $|\phi|<1$, when $t \rightarrow \infty$ the first summation tends to $\alpha/(1-\phi)$.

Result 1: For finite samples, given that the shocks ϵ_t are independent and identically distributed (iid) with zero mean, we have state-dependent growth of y_t since the contribution of the second term of (3) is not always zero. In the unit root case, $\phi=1$ and $u_t=\alpha t + \sum_{i=0}^{t-1} \epsilon_{t-i}$, and the growth rates u_t are not only state-dependent, but can increase indefinitely with t when the drift $\alpha \neq 0$.

Proof: If the shocks ϵ_t are zero mean iid, we know from the familiar statistical results (laws of large numbers and central limit theorems) that the sample mean $(1/T)\sum_{i=0}^{T-1} \epsilon_{t-i}$ will have zero mean as long as $T > 30$. However, since (3) has a weighted sum of shocks with weights depending on the powers of the root ϕ , we cannot invoke the LLN. In fact, if $\phi \geq 1$ it is clear that the second term will remain nonzero even if the drift is zero.

Since AR(1) model is important in econometrics, we discuss further properties of the model.

Variance and Autocovariance for the AR(1):

Now we turn to the autocovariance $\gamma_k = E(y_t y_{t-k})$ of order k , which is the numerator of the autocorrelation coefficient ρ_j . When $k=0$ we simply have the variance γ_0 defined as follows.

$$\begin{aligned} \gamma_0 &= E(y_t^2) = E\left[\sum_{i=0}^{\infty} \phi^i a_{t-i} \sum_{j=0}^{\infty} \phi^j a_{t-j}\right] \\ &= E\left[\left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right) \left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right)\right] \\ \gamma_0 &= E\left[\left(a\right) \left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right) + \phi a_{t-1} \left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right) + \phi^2 a_{t-2} \left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right)\right] \\ &= \sigma^2 + \phi^2 \sigma^2 + \phi^4 \sigma^2 + \dots \quad \text{[Using } E(a_t a_{t-j}) = 0 \text{ when } j \neq 0; \text{ and that } E(a_t^2) = \sigma^2 \text{ for all } t \text{]} \end{aligned}$$

Using the geometric series $\frac{1}{(1-z)} = 1+z+z^2+\dots$ for $z = \phi^2$ we have

$$\gamma_0 = \left[\frac{1}{1-\phi^2}\right] \sigma^2 \quad (3)$$

To evaluate γ_k , we use (2) to write

$$y_{t-k} = (1 - \phi L)^{-1} a_{t-k} = \sum_{j=0}^{\infty} \phi^j a_{t-j-k} = \sum_{j=0}^{\infty} G_j a_{t-j-k} \quad (\text{where } G_0=1) \quad (4)$$

From (2) and (4) we write

$$\begin{aligned} \gamma_k &= E(y_t y_{t-k}) = E\left[\sum_{i=0}^{\infty} \phi^i a_{t-i} \sum_{j=0}^{\infty} \phi^j a_{t-j-k}\right] \\ \text{Hence,} \\ \gamma_k &= E\left[\left(a + \phi a_{t-1} + \phi^2 a_{t-2} \dots\right) \left(a_{t-k} + \phi a_{t-k-1} + \phi^2 a_{t-k-2} \dots\right)\right] \\ &= E\left\{\left(a\right) \left(a_{t-k} + \phi a_{t-k-1} + \phi^2 a_{t-k-2} \dots\right) + \phi a_{t-1} \left(a_{t-k} + \phi a_{t-k-1} + \phi^2 a_{t-k-2} \dots\right) + \phi^2 a_{t-2} \left(a_{t-k} + \phi a_{t-k-1} + \phi^2 a_{t-k-2} \dots\right)\right\} \end{aligned}$$

For $k=1$ we have:

$$\gamma_1 = 0 + \phi \sigma^2 + \phi^3 \sigma^2 + \dots = \phi \gamma_0, \text{ using (3), and the geometric series used there.}$$

To derive an expression for a larger $k (>1)$ it is convenient to write these results in terms of the Kronecker delta function – not to be confused with the Kronecker product of matrices – defined below.

$$E(a_t a_{t'}) = \delta_{t,t'} \sigma^2, \quad \text{with } \delta_{t,t'} = \begin{cases} 1, & \text{if } t=t' \\ 0 & \text{if } t \neq t' \end{cases} \quad (5)$$

Delta function can be used when multiplying two infinite sums above, because the nonzero terms occur only when the subscripts of a are equal to each other.

$$\gamma_k = E(y_t y_{t-k}) = E \left[\sum_{i=0}^{\infty} \phi^i a_{t-i} \sum_{j=0}^{\infty} \phi^j a_{t-j-k} \right] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \phi^{i+j} \delta_{t-i,t-j-k} \sigma^2$$

Now one infinite sum can be eliminated by choosing the subscripts as follows. We require $t - i = t - j - k$, which implies that $-i = -j - k$, that is $i = j+k$, which allows us to replace $\delta_{t-i,t-j-k} \equiv 1$ when $i = j+k$. Thus

$$\gamma_k = \sum_{j=0}^{\infty} \phi^{2j+k} \sigma^2 = \sigma^2 \phi^k \sum_{j=0}^{\infty} \phi^{2j} = \sigma^2 \phi^k \left[\frac{1}{1-\phi^2} \right] \quad (6)$$

using the geometric series. The autocorrelation coefficient is

$$\rho_k = \gamma_k / \sqrt{\gamma_0 \gamma_0} = \gamma_k / \gamma_0 \quad (7)$$

Exercise: Use (3) and (5) to evaluate ρ_k for the AR(1) model, where $\sigma^2(1 - \phi^2)^{-1}$ cancels, to show that

$$\rho_k = \phi^k \text{ for AR(1)} \quad (8)$$

Exercise: Show that $\gamma_k = \gamma_{-k}$, the autocovariance function is also meaningful for negative values of k . Show that the k in the last equation of (6) should be replaced by $|k|$ to accommodate the negative k values.

Autocovariance Generating Function (AGF)

Since macroeconomics texts (e.g. Sargent) refer to AGFs, it is useful to review them in the present context, since they give us an opportunity to offer a like with the Green's function coefficients G_j noted above. This concept requires some familiarity with complex numbers and Laurent expansions. Nerlove, et al (1979, p.38) and Sargent(1979, p.221) are some of the econometric references. Box and Jenkins(1970, p.81) is a statistical reference. Hamilton (1994, p.61) describes AGF's as a function constructed by taking k -th autocovariance and multiplying it by k -th power of some (complex) number z and summing it over all possible values of k . Define:

$$AGF(z) = \sum_{k=-\infty}^{\infty} \gamma_k z^k$$

where negative powers and subscripts are involved. Thus AGF is simply a doubly infinite sum of a covariance times z^k , where z is a complex variable and k is the order of the lag. One may also think of z itself as the familiar L the lag operator. In some contexts, one uses $z=L^{-1}$ and this should not cause confusion. Using (6) defined for both negative and positive values of k we try to get rid of the

infinite sum by using a geometric progression in four steps (9a) to (9d). The following infinite sums converge on the unit circle and in an annulus region around it : $|\phi| < |z| < 1/|\phi|$.

$$AGF(z) = \sum_{k=-\infty}^{\infty} \gamma_k z^k = \sum_{k=-\infty}^{\infty} \sigma^2 \phi^{|k|} \left[\frac{1}{1-\phi^2} \right] z^k = \frac{\sigma^2}{1-\phi^2} \left[\frac{1}{1-\phi z} + \sum_{j=1}^{\infty} \phi^j z^{-j} \right] \quad (9a)$$

To verify the last equality use $\sum(-\infty \text{ to } \infty) = \sum(-\infty \text{ to } 1) + \sum(1 \text{ to } \infty)$ and $j = -k$. To remove the remaining infinite sum we take out the common term with $j=1$ as follows.

$$AGF(z) = \frac{\sigma^2}{1-\phi^2} \left[\frac{1}{1-\phi z} + \frac{\phi z^{-1}}{1-\phi z^{-1}} \right] = \sigma^2 \left[\frac{1}{(1-\phi z)(1-\phi z^{-1})} \right] \quad (9b)$$

where the bracketed first term gets a $(1-\phi^2)$ in the numerator which cancels.

For a one-sided MA(∞) process, $y_t = \sum_{j=0}^{\infty} G_j a_{t-j}$, the autocovariances are:

$$\gamma_k = E(y_t y_{t+k}) = E \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} G_j G_l a_{t-j} a_{t+k-l},$$

where the expectation $E(a_i a_j) = \sigma^2$ only when the subscripts $i \equiv j$, and is zero otherwise. Thus

$\gamma_k = \sigma^2 \sum_{j=0}^{\infty} G_j G_{j+k}$ may be substituted in the definition of AGF to yield

$$AGF(z) = \sigma^2 \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} G_j G_{j+k} z^k = \sigma^2 \sum_{k=-\infty}^{\infty} \sum_{j=0}^{\infty} G_j G_{j+k} z^k = \sigma^2 \sum_{h=0}^{\infty} \sum_{j=0}^{\infty} G_j G_h z^{h-j} \quad (9c)$$

To verify the second equality use $G_{-j} = 0$ when $j < 0$. To verify the third equality use $h=j+k$, $k=h-j$.

$$AGF(z) = \sigma^2 \sum_{h=0}^{\infty} \sum_{j=0}^{\infty} G_j G_h z^{h-j} = \sigma^2 \sum_{j=0}^{\infty} G_j z^{-j} \sum_{h=0}^{\infty} G_h z^h = \sigma^2 G(z) G(z^{-1}) \quad (9d)$$

where the particular case (9b) reveals what is meant by the function $G(z^{-1})$. The last equality is sometimes called canonical factorization. Thus $z = \cos(\omega) - i \sin(\omega) = \exp(-i\omega)$ where $i = \sqrt{-1}$ and ω is the radian angle that z makes with the real axis.

Relation between AGF and spectral analysis:

The AGF is fundamental in time series analysis for various reasons. If the AGF is evaluated at the $z = \exp(-i\omega)$ and divided by 2π , the resulting function of ω is called the population spectrum $s_Y(\omega) = (1/2\pi)AGF$. If autocovariances are absolutely summable, and if two process have common AGF, then they will also have identical autocovariances. This is why it is a generating function.

Examples of AGF for MA processes: Recall that the only nonzero autocovariances are $\gamma_{-1}, \gamma_0, \gamma_1$. Hence the infinite sum only has 3 terms with powers of z equal to the subscripts of γ .

$AGF(\text{for MA}(1)) = \theta \sigma^2 z^{-1} + (1+\theta^2)\sigma^2 z^0 + \theta \sigma^2 z = \sigma^2 [(1+\theta z)(1+\theta z^{-1})]$. More generally, for the MA(∞)

process, $y_t = \mu + G(L)\epsilon_t$, where $G(L) = G_0 + G_1L + G_2z^2 + \dots$ is an infinite polynomial. Now, $AGF(\text{for } MA(\infty)) = \sigma^2 G(z)G(z^{-1})$. This is a powerful result because ARMA processes can often be written as $MA(\infty)$ processes by Wold's theorem. $AGF(\text{for } AR(1))$ is $\sigma^2 / [(1-\phi z)(1-\phi z^{-1})]$

Stationarity of AR(1) :

Now we turn to checking the stationarity of the AR(1) model. For a stochastic process to be weakly stationary – also called covariance stationary – recall that there are three requirements: (i) $E(y_t) = \mu < \infty$ for all t , and is not a function of t . (ii) $E(y_t - \mu)^2 < \infty$ for all t , and is not a function of t , and (iii) $E(y_t - \mu)(y_{t-k} - \mu) = \gamma_k < \infty$ for all t , and is not a function of t . For the AR(1) model the variance γ_0 from (3) is $\sigma^2(1 + \phi^2 + \phi^4 + \dots)$, where the infinite series in ϕ^2 will be finite, if and only if $|\phi| < 1$. This stationarity condition will also ensure the finiteness of γ_k .

If the parameters of an ARMA process are known and if the polynomial $\Phi(L)$ on the AR side and the polynomial $\Theta(L)$ on the MA side have no roots in common, one can easily check if the ARMA process is stationary as follows. The roots of the $\Phi(L)$ must lie outside the unit circle. e.g. the root of $1 - \phi L = 0$ is $L = 1/\phi$. Hence requiring ϕ to be fractional is the same as requiring the root to be larger than unity, or outside the unit circle.

AGF and the Power Spectrum

Some econometricians introduce the idea of a power spectrum by using the relationship between the complex number z and the familiar lag operator. Nerlove(1979, App. C) gives a thorough discussion of the related material. Here we state some facts for convenience of the reader, and not attempt a complete discussion. When we substitute $z = e^{-i2\pi f}$ we obtain half the power spectrum at frequency $0 \leq f \leq 1/2$. Thus the power spectrum of the general moving average process $MA(\infty)$ is

$$p(f) = 2\sigma^2 G(z)G(z^{-1}) = 2\sigma^2 G(e^{-i2\pi f})G(e^{i2\pi f}) = 2\sigma^2 |G(e^{i2\pi f})|^2 \quad (9d)$$

The spectral density function $sdf(\omega) = (\sigma^2/2\pi) G(e^{i\omega})G(e^{-i\omega})$ is a Fourier transform of the autocovariance function. It is proportional to the AGF defined on the unit circle. It is always real even though AGF involves complex numbers. The integral of sdf from $-\pi$ to π is the variance of the process. In a usual density function the area under the density represents a probability. By analogy, the area under a spectral density represents the part of the total variance attributed to the range represented by the corresponding angular frequencies.

Autocorrelations satisfy the Homogeneous form of the relevant Difference Equation

We note an interesting fact that the autocorrelation coefficients ρ_k satisfy the homogeneous form of the difference equation (1), viz., $y_t = \phi y_{t-1}$ if we replace y_t by ρ_k and define $\rho_0 = 1$. We have shown in (8) above that $\rho_k = \phi^k$. Hence all we have to do is to verify

$$\rho_k = \phi \rho_{k-1} = \phi(\phi^{k-1})\rho_0 = \phi^k \quad (10)$$

We shall see that this generalizes to higher order processes, and equations similar to (10) are called Yule-Walker equations.

Business Cycles and the AR(2) Framework

Since macroeconomic business cycles are related to the AR(2) model, it is useful to study the AR(2) model in detail. We shall see that the complex number solution is particularly relevant.

Stochastic Difference Equation of the AR(2) Model:

Consider the AR(2) model

$$y_t - \phi_1 y_{t-1} - \phi_2 y_{t-2} = a_t \quad (1)$$

Its characteristic polynomial

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L) = \left[-[\lambda_1 + \lambda_2] L + \lambda_1 \lambda_2 L^2 \right] \quad (2)$$

where the roots satisfy two relations:

$$\lambda_1 + \lambda_2 = \phi_1, \text{ and } \lambda_1 \lambda_2 = -\phi_2 \quad (3)$$

The roots of the quadratic are known explicitly by the formula

$$\lambda_1, \lambda_2 = \frac{1}{2} \left[\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2} \right] \quad (4)$$

which are both real if the square root term in the above expression $(\phi_1^2 + 4\phi_2) \geq 0$, otherwise they are complex conjugate to be discussed later. When they are real, the solution of AR(2) is obtained by dividing both sides of (1) by the characteristic polynomial (2). Then we have:

$$y_t = \frac{1}{(1-\lambda_1 L)(1-\lambda_2 L)} a_t = \frac{A}{(1-\lambda_1 L)} a_t + \frac{B}{(1-\lambda_2 L)} a_t \quad (5)$$

where A and B are obtained below by the method of partial fractions. When λ_1 and λ_2 are real numbers, the method of partial fractions involves following steps. To get A we solve the denominator $(1 - \lambda_1 L) = 0$, to yield $L = 1/\lambda_1$. Now substitute this L in the long denominator of (5). That is, $(1 - \lambda_1 L)(1 - \lambda_2 L) = (1 - \lambda_1 \lambda_1^{-1})(1 - \lambda_2 \lambda_1^{-1}) = (0)(1 - \lambda_2 \lambda_1^{-1})$. which is taken to be $(1 - \lambda_2 \lambda_1^{-1})$ upon ignoring the zero multiplier. Similarly for B. We have:

$$A = [1 - (\lambda_2/\lambda_1)]^{-1}, \text{ and } B = [1 - (\lambda_1/\lambda_2)]^{-1}. \quad (5a)$$

To verify (5a) obtained by ignoring zeros it suffices to see that the numerators of expressions on the left hand and the right hand side of the last equal sign in (5) are identical. Upon simplification the question becomes:

$$\begin{aligned} 1 &\stackrel{?}{=} A(1 - \lambda_2 L) + B(1 - \lambda_1 L) = (\lambda_1 - \lambda_2)^{-1} \lambda_1 (1 - \lambda_2 L) + (\lambda_2 - \lambda_1)^{-1} \lambda_2 (1 - \lambda_1 L) \\ &= (\lambda_1 - \lambda_2)^{-1} \left[(1 - \lambda_2 L) \lambda_1 - (1 - \lambda_1 L) \lambda_2 \right] = (\lambda_1 - \lambda_2)^{-1} (\lambda_1 - \lambda_2) = 1, \text{ hence (5a) holds. } \square \end{aligned}$$

In numerical problems it is easy to write the partial fractions. For example.

$$\frac{1}{(x-4)(x-5)} = \frac{1}{(5-4)(x-5)} + \frac{1}{(x-4)(4-5)} = \frac{1}{(x-5)} + \frac{-1}{(x-4)} \quad (6)$$

where all we have done is to substitute the solution $x = 5$ in the first denominator, except when it becomes zero; and $x = 4$ in the second denominator, except when the term becomes zero.

Real Number Solution of the Stochastic AR(2) Difference Equation:

We use the partial fractions to write the second order problem as a sum of two AR(1) problems in (5). Hence to write the solution for the second order dynamics represented by the stochastic difference equation arising from the AR(2) model, we recall the solution (2) of §6.2.1 for the AR(1) in terms of its root ϕ_1 . Now, the solution for y_t is given by a sum of two infinite series from (5).

$$y_t = A \left[+\lambda_1 L + \lambda_1^2 L^2 + \dots \right] a_t + B \left[+\lambda_2 L + \lambda_2^2 L^2 + \dots \right] a_t = \sum_{j=0}^{\infty} G_j a_{t-j} \tag{7}$$

Hence in terms of Green's function, $G_j = A \lambda_1^j + B \lambda_2^j$, $A = [1 - (\lambda_2/\lambda_1)]^{-1}$, and $B = [1 - (\lambda_1/\lambda_2)]^{-1}$ when the roots λ_1 and λ_2 are real numbers. If the roots satisfy $|\lambda_1| < 1$, and $|\lambda_2| < 1$ the solution is stable, since the effect of past errors a_{t-j} is weighted by a smaller and smaller number as j increases. A plot of G against j will show a sum of two declining curves, representing the dynamic memory of the system.

Complex Number Solution of the Stochastic AR(2) Difference Equation:

When the roots given in (4) of the AR(2) polynomial in (2) are imaginary, because $(\phi_1^2 + 4\phi_2) < 0$, they are complex conjugate, and the coefficients A and B in the expression for G are also complex conjugate. If the reader is familiar with equation (10) below, the elementary discussion between here and (10) may be skipped.

Let us review the use of the imaginary number $\sqrt{-1} = i$ for convenience of students. Any complex number can be written as $C = X + Yi$, which is usually represented by a vector (X, Y) in a two dimensional space. The horizontal coordinate X represents the real part and the vertical coordinate Y represents the imaginary part. The modulus of a complex number $|C|$ is its length. A companion of a complex number $\bar{C} = X - Yi$, is called its complex conjugate.

Exercise: What is the length of the vector C ? Verify that $|C|^2 = X^2 + Y^2$ by the Pythagorous Theorem. Plot C and \bar{C} vectors for $X = 2$, $Y = 1$ and verify that \bar{C} is directly under C . Geometrically show the product of two vectors using the parallelogram, and note that it will land on the horizontal axis at the squared distance

$$C \bar{C} = (X + Yi)(X - Yi) = X^2 + XYi - XYi - Y^2 i^2 = X^2 + Y^2 = |C|^2.$$

The complex roots, especially in their polar form are useful in Time Series analysis for representing cyclical behavior of the series. The polar coordinates have two parts: the amplitude (α) and the phase angle (ψ), and there is a simple, unique, one-to-one mapping of the coordinates (X, Y) to (α, ψ) , when the angle is restricted to the half open interval $[0, 2\pi)$ in radians or to $[0^\circ, 360^\circ)$ in degrees.

$$C = \alpha(\cos\psi + i \sin\psi), \alpha = |C| \text{ and } \psi = \tan^{-1} \left[\frac{Y}{X} \right] = \arg(C) \tag{8}$$

where the notation \arg represents the argument of the complex number or the phase angle. In the mapping $(X, Y) \rightarrow (\alpha, \psi)$, we also have $X = \alpha \cos\psi$ and $Y = \alpha \sin\psi$, respectively the real and

imaginary coordinates of the complex number C. Recall that the sine function is periodic in the sense that it has the same value after every interval of 360° . In other words, $\sin 0 = \sin 360^\circ = \sin k2\pi$, for any integer k. Similarly the cosine function is periodic. Hence, it is intuitively clear that the complex numbers may be useful for representing periodic or cyclical behavior of time series. If we permit $\psi=2\pi=360^\circ$ we lose uniqueness of the mapping $(X,Y) \rightarrow (\alpha,\psi)$, because $\psi=360^\circ$ cannot be distinguished from $\psi=0^\circ$. Losing uniqueness is actually a good thing, because the same complex number C represents a new cycle starting at $\psi=360^\circ$ and ending at 720° , that is in the interval $\psi \in [2\pi,4\pi)$. Yet another cycle has $\psi \in [4\pi,6\pi)$. In general, $\psi \in [k\pi, k\pi+2\pi)$ for an arbitrary integer multiple k, which can be negative.

The trigonometric and exponential functions are also represented as power series:

$$\sin \psi = \psi - \frac{\psi^3}{3!} + \frac{\psi^5}{5!} - \dots + \dots, \quad \cos \psi = 1 - \frac{\psi^2}{2!} + \frac{\psi^4}{4!} - \dots + \dots, \quad \text{and} \quad \exp \psi = 1 + \psi + \frac{\psi^2}{2!} + \dots + (9)$$

Exercise Expand $\exp(\psi i)$ by the above formula, substitute $i = \sqrt{-1}$, and use the power series for sine and cosine to write $\exp(\psi i) = \cos \psi + i \sin \psi$. Hence verify that $C = X + Yi = \alpha \exp(\psi i)$, with $X = \alpha \cos \psi$, and $Y = \alpha \sin \psi$ as before. Similarly show that the complex conjugate $\bar{C} = X - Yi = \alpha \exp(-\psi i)$.

Now we turn to the roots of the quadratic polynomial $(1 - \phi_1 L - \phi_2 L^2)$

$$\lambda_1, \lambda_2 = \frac{1}{2} \left[\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2} \right] = X \pm Yi = \alpha \exp(\pm \psi i) \quad (10)$$

which are complex conjugate when $(\phi_1^2 + 4\phi_2) < 0$ in the present case. From the second equality of (10), note that

$$X = (1/2)\phi_1 \quad \text{and} \quad Y = (1/2)[-(\phi_1^2 + 4\phi_2)]^{1/2} \quad (10a)$$

which are the X and Y coordinates. In defining Y we have moved the negative sign inside the square root sign because $-(\phi_1^2 + 4\phi_2) > 0$, and its square root can be evaluated. Since $\lambda_1 = X + Yi$ and $\lambda_2 = X - Yi$ their sum $\lambda_1 + \lambda_2 = \phi_1$, and their product $\lambda_1 \lambda_2 = -\phi_2$ are still real numbers, as they should be for substitution in the expression for the polynomial. Now evaluate the absolute value of each of the two roots, and note that they are identical to each other. We have the squared amplitude $\alpha^2 = |\lambda_1|^2 = X^2 + Y^2$. In terms of the coefficients ϕ we have the following remarkable result:

$$\alpha = \phi_1^2/4 - (1/4)(\phi_1^2 + 4\phi_2) = -\phi_2 = \lambda_1 \lambda_2. \quad (10b)$$

Now what is the phase angle ψ in terms of ϕ_1 and ϕ_2 ? One way to find ψ might be to use the usual formulas for polar coordinates in (8), write $\psi = \tan^{-1}(Y/X)$ and use (10a). Since the expression for Y in (10a) is somewhat complicated, we can avoid using it by considering $\cos \psi = (\text{adjacent side}) / \text{hypotenuse} = X / (X^2 + Y^2)^{1/2} = X / \alpha$. Hence from (10a) and (10b) we have:

$$\psi = \cos^{-1} \left[\frac{\phi_1}{2\sqrt{-\phi_2}} \right] = \cos^{-1} \left[\frac{(\lambda_1 + \lambda_2)}{2\sqrt{(\lambda_1 \lambda_2)}} \right] \quad (11)$$

Exercise Plot the above roots (10) in polar coordinate system. First draw the vector $C=(X,Y)$ from the origin with $X=2$ and $Y=1$. Verify that the angle ψ of the vector C with the horizontal axis satisfies the above relation (11).

Now we turn to the solution of the stochastic difference equation. In terms of Green's function,

$$G_j = A \lambda_1^j + \bar{A} \lambda_2^j = \mu \exp(\delta i) \alpha^j \exp(j\psi i) + \mu \exp(-\delta i) \alpha^j \exp(-j\psi i)$$

$$= \mu \alpha^j \left[e^{i(\delta+j\psi)} + e^{-i(\delta+j\psi)} \right] = \mu \alpha^j \cos(\delta+j\psi) \quad (12)$$

where we have denoted the second coefficient by $\bar{A}=\mu \exp(-\delta i)$, because it is a complex conjugate of $A=\mu \exp(\delta i)$, having μ =amplitude and δ =phase angle for the polar form of the coefficients. Thus it can be verified that a plot of G_j against j is a damped cosine wave when AR(2) model has complex roots which satisfy: $|\lambda_1| < 1$, and $|\lambda_2| < 1$. The damping implies eventual dying of the memory, and the cosine wave captures the cyclical behavior. In the complex plane we require the roots to be inside the unit disk.

Stationarity conditions on the autoregressive parameters ϕ in the AR(2) Model:

The stationarity condition $|\lambda_1| < 1$, and $|\lambda_2| < 1$ and the relations (3) and (10) connecting them with ϕ_1 and ϕ_2 mean the following.

$$(a) \quad -2 \leq \phi_i \leq 2 \text{ for } i=1,2. \quad (b) \quad |\phi_2| < 1. \quad (c) \quad \phi_2 + \phi_1 < 1. \quad (d) \quad \phi_2 - \phi_1 < 1. \quad (13)$$

Exercise: Should we use the absolute values $|\phi_2|$ and $|\phi_1|$ in (c) and (d) in (13) above? Draw a triangular figure and indicate the stable regions satisfying the conditions (a) to (d) above.

Autocovariances for the Stochastic AR(2)

In the section for AR(1) model the equation ??(5) defined the $\gamma_k = E(y_t y_{t-k}) = \phi_1^k \gamma_0$, and ??(10) stated the Yule-Walker equations. Let us now find analogous results for the second order case. If the roots λ_1 and λ_2 of the autoregressive polynomial are real, recall that the solution is

$$y_t = \sum_{j=0}^{\infty} G_j a_{t-j} \text{ with } G_j = A \lambda_1^j + B \lambda_2^j, A = [1 - (\lambda_2/\lambda_1)]^{-1}, \text{ and } B = [1 - (\lambda_1/\lambda_2)]^{-1} \quad (14)$$

$$\gamma_k = E(y_t y_{t-k}) = E \left[\sum_{i=0}^{\infty} G_i a_{t-i} \right] \left[\sum_{j=0}^{\infty} G_j a_{t-j-k} \right] = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} G_i G_j \delta_{t-i, t-j-k} \sigma^2 \quad (15)$$

where the Kronecker delta mentioned earlier has the defining property that: $\delta_{k,m} = 1$ if $k=m$, and $\delta_{k,m} = 0$ otherwise. Hence the only nonzero terms are the ones with the two subscripts of δ equal to each other; that is only when:

$$t - i = t - j - k, \text{ or } -i = -j - k, \text{ or } i = j+k \quad (16)$$

Hence one of the summations can be eliminated if i is replaced by $j+k$. Thus

$$\gamma_k = \sigma^2 \sum_{j=0}^{\infty} G_{j+k} G_j \quad (17)$$

If $k=0$ we have

$$\gamma_0 = \sigma^2 \sum_{j=0}^{\infty} (A \lambda_1^j + B \lambda_2^j)^2 \quad (18)$$

for A and B in (14). In general,

$$\gamma_k = \sigma^2 \sum_{j=0}^{\infty} (A \lambda_1^j + B \lambda_2^j)(A \lambda_1^{j+k} + B \lambda_2^{j+k}) \quad (19)$$

Thus we note that an explicit expression for the autocovariances is available. There is an interesting alternative derivation of γ_k based on (1) written as $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + a_t$. Now multiply both sides by y_{t-k} to yield $y_{t-k} y_t = \phi_1 y_{t-k} y_{t-1} + \phi_2 y_{t-k} y_{t-2} + a_t y_{t-k}$. Now take expectation of both sides to give

$$\gamma_k = E(y_t y_{t-k}) = E(\phi_1 y_{t-1} y_{t-k} + \phi_2 y_{t-2} y_{t-k} + y_{t-k} a_t) = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} + E(a_t y_{t-k}) \quad (20)$$

where we use $a_t \sim N(0, \sigma^2)$, $E(a_k a_m) = \delta_{k,m} \sigma^2$, (with $\delta_{k,m} = 0$ if $k \neq m$, and $\delta_{k,m} = 1$ if $k=m$), and symmetric autocovariances

$$\gamma_k = E(y_t y_{t-k}) = E(y_{t-k} y_t) = E(y_t y_{t+k}) = \gamma_{-k}. \quad (21)$$

To evaluate the last term of (20) let us consider a_{t-1} instead of a_t *future reference, and evaluate expectations in terms of Green's function similar to (15).

$$E(a_{t-1} y_{t-k}) = E a_{t-1} \left[\sum_{j=0}^{\infty} G_j a_{t-j-k} \right] = \sum_{j=0}^{\infty} G_j E(a_{t-1} a_{t-j-k}) = \begin{cases} 0 & \text{if } k > 1 \\ -G_{1-k} \sigma^2 & \text{if } k \leq 1 \end{cases} \quad (22)$$

Exercise: Verify the last equality of (22). Note that only nonzero expectation ($= \sigma^2$) obtains when $t-1 = t-j-k$, that is when $1 = j+k$, that is when $j = 1-k$. Hence check the subscript of last G in (22). Next, apply (22) to verify that the last term of (20) is zero, i.e., $E(a_t y_{t-k}) = 0$. In (22) let $l=0$, and note that $k > 1$.

If $k=0$, we have

$$\gamma_0 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + \sigma^2 \quad (\text{since } G_0=1, \text{ by definition}) \quad (23)$$

For $k=1$,

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 \quad (24)$$

For $k=2$,

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0 \quad (25)$$

The relations (24) and (25) will be used below in deriving Yule-Walker relations of the AR(2) model.

For $k > 2$,

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} \quad (26)$$

Thus, for $k \geq 2$ the autocovariances are seen to satisfy the same difference equation (1) as y_t . This suggests that one can understand the behavior of the original series y_t by studying the behavior of the autocovariances γ_k . For example, the roots λ_1 and λ_2 of the AR(2) model studied above can then be directly used to solve the difference equation (26) for γ_k with a solution similar to (14) or (12).

$$\gamma_k = \sum_{j=0}^{\infty} G_j a_{k-j} \quad \text{with } G_j = A \lambda_1^j + B \lambda_2^j, \quad A = [1 - (\lambda_2/\lambda_1)]^{-1}, \quad \text{and } B = [1 - (\lambda_1/\lambda_2)]^{-1} \quad (27)$$

if the roots are real. If the roots are complex conjugate as in (10), we replace G_j by

$$G_j = A \lambda_1^j + \bar{A} \lambda_2^j = \mu \exp(\delta i) \alpha^j \exp(j\psi i) + \mu \exp(-\delta i) \alpha^j \exp(-j\psi i) \quad (28)$$

Yule Walker Relations to get ϕ_i from ρ_i :

Let us divide (24) and (25) by γ_0 to write them in terms of correlations instead of covariances.

$$\rho_1 = \phi_1 \rho_0 + \phi_2 \rho_1 \quad (29)$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2 \rho_0 \quad (30)$$

with $\rho_0=1$. In matrix notation (29) and (30) become:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} \quad (31)$$

which can be solved by the usual methods to obtain ϕ_1 and ϕ_2 – the autoregressive parameters – in terms of the autocorrelation coefficients ρ_1 and ρ_2 . It is somewhat instructive to use the Cramer's rule to obtain the so called Yule-Walker relations.

$$\phi_1 = \begin{vmatrix} \rho_1 & \rho_1 \\ \rho_2 & 1 \end{vmatrix} \div \begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} = (\rho_1 - \rho_1 \rho_2) / (1 - \rho_1^2) \quad (32)$$

$$\phi_2 = \begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix} \div \begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix} = (\rho_2 - \rho_1 \rho_2) / (1 - \rho_1^2) \quad (33)$$

For higher order AR(p) models Yule-Walker relations use ratios of higher order determinants.

Partial Autocorrelation Function (PACF)

In addition to autocovariances and autocorrelations, PACF are important in time series analysis for various reasons. For example, in the Box-Jenkins methodology noted later, ACF and PACF together help decide the order of the AR and MA polynomials of an ARMA process. PACF are also related to the fractionally integrated (long-memory) processes, similar to unit root processes. When we consider higher order autoregressive models AR(1), AR(2),..., AR(p) where do we stop ? This is similar to the problem of deciding when to stop when adding regressors in a multiple regression, and does not have a clear cut solution. There are statistical tests and other considerations including parsimony. In the methodology popularized by Box and Jenkins (1970) partial autocorrelation functions are used to help choose the p of the AR(p). The usual partial correlation coefficient of y on x_1 holding x_2 fixed is defined by the following formula:

$$r_{y1.2} = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)(1-r_{12}^2)}} = \frac{-C_{12}}{\sqrt{C_{11} C_{22}}} \quad (34)$$

where C_{ij} denotes the cofactor [determinant of the submatrix when i -th row and j -th column are moved out, multiplied by $(-1)^{i+j}$] of (i,j) th element of a 3×3 matrix of correlation coefficients of y , x_1 and x_2 respectively. We have

$$C = \begin{bmatrix} 1 & r_{y1} & r_{y2} \\ r_{y1} & 1 & r_{12} \\ r_{y2} & r_{12} & 1 \end{bmatrix} \quad (35)$$

The interpretation of $r_{y1.2}$ is that it measures the marginal contribution of x_1 when y is regressed on x_1 and x_2 . Roughly speaking, one may include x_1 in the regression if $r_{y1.2}$ is statistically significant. Now $r_{y2.1}$ is obtained by switching subscripts 1 and 2 in (34) and (35); and measures the marginal contribution of x_2 , helpful in deciding whether x_2 should be included. More generally, the partial correlation of y on x_1 , while holding both x_2 and x_3 fixed, will be based on a 4×4 matrix similar to (35) having the last column $(r_{y3}, r_{13}, r_{23}, 1)$.

In time series analysis, the partial autocorrelation coefficients (PACF) of order k are denoted by β_{kk} and defined in terms of regressions as follows.

$$y_t = -\beta_{11} y_{t-1} + a_t, \quad [\text{For AR(1), we have } \beta_{11} = \phi_1, \text{ from (1) of that section above}] \quad (36)$$

$$y_t = -\phi_1 y_{t-1} - \beta_{22} y_{t-2} + a_t \quad [\text{For AR(2), we have } \beta_{22} = \phi_2, \text{ from (1) above}] \quad (37)$$

$$y_t = -\phi_1 y_{t-1} - \phi_2 y_{t-2} - \beta_{33} y_{t-3} + a_t \quad [\text{For AR(3), we have } \beta_{33} = \phi_3] \quad (38)$$

If the AR(2) model of (1) above is valid, and we fitted AR(3) of (38) it is obvious that $\beta_{33} = 0$, and similarly $\beta_{kk} = 0$ for $k > 3$ if we fit an AR(k) model. This fact is used as a diagnostic in determining the order of the AR model, especially emphasized in the Box-Jenkins style of modeling. The numerical estimation of partial autocorrelation coefficients is based on the relationship between $\phi_k = \beta_{kk}$ and ρ_k discussed above in (33) for the AR(2) case of (37). The AR(3) model of (38) suggests the following interesting relation.

Relation Between Yule-Walker Equations and Partial Autocorrelation Coefficients:

For the AR(1) model the Yule-Walker equation is simply $\rho_1 = \phi_1 \rho_0$ and we have $\beta_{11} = \phi_1$. For the AR(2) they are given by equations (29) and (30), and the solution for β_{22} is indicated in (33) as:

$$\beta_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} \quad (40)$$

where the determinant of the 3×3 matrix of autocorrelations in the denominator is directly estimated. The determinant in the numerator is the same as the one in the denominator except for the last column, which is (ρ_1, ρ_2, ρ_3) . This pattern holds for higher orders also, and the last column for β_{kk} becomes $(\rho_1, \rho_2, \rho_3, \dots, \rho_k)$ in the numerator. Computer programs routinely calculate the partial autocorrelation coefficients of all orders by this method without running any regressions similar to (38). The partial autocorrelation function β_{kk} has a sharp cutoff at the order p when the underlying true model is AR(p), so that all $\beta_{kk} = 0$ for all $k > p$. Durbin(1960) proposed a recursive method of finding the partial autocorrelations using the Yule-Walker relations.

Significance tests for Partial Autocorrelation Coefficients

Quenouille(1949) shows that the variance of the estimate $\hat{\beta}_{kk}$ of β_{kk} from a sample of size T is given by $\text{Var}[\hat{\beta}_{kk}] = \frac{1}{T}$, whence the standard error is

$$\text{SE}[\hat{\beta}_{kk}] = \frac{1}{\sqrt{T}} \text{ for } k \geq p+1, \text{ where the null hypothesis is the AR}(p) \text{ model} \quad (41)$$

Many computer packages plot the partial autocorrelations along with a band for standard errors based on (41). These plots are an important cornerstone in Box-Jenkins style identification of ARIMA models.

General Solution to ARMA($n, n-1$) Stochastic Difference Equations:

In general, consider the n -th order dynamics represented by the ARMA($n, n-1$) model:

$$\prod_{i=1}^n (1 - \phi_i L) y_t = b_0 + \prod_{j=1}^{n-1} (1 - \theta_j L) a_t \quad (1)$$

The solution is:

$$y_t = \sum_{j=1}^p G_j a_{t-j} \quad (2)$$

where

$$G_j = g_1 \lambda_1^j + g_2 \lambda_2^j + \dots + g_n \lambda_n^j \quad (3)$$

These $\lambda_i, i=1,2,\dots,n$ are the n solutions of the familiar n -th order characteristic polynomial in z (inverse of the lag operator L) on the autoregressive side involving the AR parameters ϕ . For example, in the AR(1) case the polynomial is simply $(1 - \phi_1 L)$, with the root $\lambda_1 = \phi_1$. In the ARMA(2,1) case the polynomial is $(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L)$, with the two roots λ_1 and λ_2 , which can be complex conjugate if the dynamics involves cyclical behavior. Furthermore, in (3) the coefficients g_i are explicitly known by the following remarkable formula, Pandit and Wu(1983, p.105).

$$g_i = \frac{(\lambda_i^{n-1} - \theta_1 \lambda_i^{n-2} - \dots - \theta_{n-1})}{(\lambda_i - \lambda_1)(\lambda_i - \lambda_2) \dots (\lambda_i - \lambda_{i-1})(\lambda_i - \lambda_{i+1}) \dots (\lambda_i - \lambda_{n-1})(\lambda_i - \lambda_n)} \quad (4)$$

In particular, when $n=2$, $g_1=(\lambda_1 - \theta_1)/(\lambda_1 - \lambda_2)$ and $g_2=(\lambda_2 - \theta_1)/(\lambda_2 - \lambda_1)$.

Box Jenkins ARIMA Modeling.

The general expression for ARIMA models should include seasonal factors as well, denoted by $ARIMA(p,d,q) \times (P,D,Q)_s$ is

$$(1 - L)^d y_t = \mu + \frac{\theta(L)\Theta_s(L)}{\phi(L)\Phi_s(L)} a_t \quad (1)$$

where y_t = original data, with t =time, the lag (backshift) operator L is defined by: $Ly_t = y_{t-1}$, $LLy_t = L^2y_t = y_{t-2}$ as before. The difference operator $\Delta y_t = y_t - y_{t-1}$ is subject to the identity: $\Delta \equiv (1 - L)$, and is usually used to induce stationarity into a time series.

μ = mean of the data.

p = order of the nonseasonal autoregressive term

q = order of the nonseasonal moving average term

P = order of the seasonal autoregressive term

Q = order of the seasonal moving average term

d = order of the nonseasonal differencing

D = order of the seasonal differencing

s = length of seasonality

a_t = random error or white noise term

$\theta(L)$ = moving average operator other than the seasonal

$$= 1 - \theta_1 L - \theta_2 L^2 - \dots - \theta_q L^q$$

$\phi(L)$ = autoregressive operator other than the seasonal

$$= 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$$

$\Theta_s(L)$ = moving average seasonal operator.

$$= 1 - \Theta_1 L - \Theta_2 L^2 - \dots - \Theta_Q L^Q \quad (\text{Upper case } Q \text{ and } \Theta)$$

$\Phi_s(L)$ = autoregressive seasonal operator.

$$= 1 - \Phi_1 L - \Phi_2 L^2 - \dots - \Phi_P L^P \quad (\text{Upper case } P \text{ and } \Phi)$$

The methodology of estimating these models is discussed in many books including Box and Jenkins(1970). We shall now review it briefly. The estimation usually uses Marquardt or similar nonlinear least squares algorithm. It may need backward forecasting (back casting) to estimate the model parameters.

Estimation

Estimation is generally carried out by the maximum likelihood (ML) or approximate maximum likelihood method. The exact likelihood function for a mixed ARMA(p,q) model is given by Newbold (1974). Denote the entire set of AR and MA parameters by β , the error by

$$\tilde{a}_t = E(a_t | y, \beta) \quad (2)$$

where the entire time series data on y_t is denoted by y . Also, denote error sum of squares by

$$ESS(\beta, y) = \sum_{t=-\infty}^T \tilde{a}_t^2 \quad (3)$$

Now, the likelihood is proportional to:

$$(1/\sigma)^T f(\beta) \exp \left[-\frac{1}{\sigma^2} ESS(\beta, y) \right] \quad (4)$$

There are three main methods for estimating (1). First, is ML, which maximizes (4). A practical transformation is suggested by Ansley (1979). The second method of estimating (1) is called exact least squares, which ignores the term $f(\beta)$ in (4), and minimizes the ESS. Observe that the lower limit of the summation in (3) is $-\infty$, and we would generally only have finite data. Box and Jenkins suggest back-casting (forecasting backwards in time) a number of pre sample values of \tilde{a}_t . The third method is even simpler and more approximate. It is called conditional least squares, and ignores the back-casting step. The word conditional refers to the fact that the conditional expectation of pre sample values is zero, and when we ignore pre sample forecasting, we are replacing them by their conditional expectations.

Diagnostic Checking

After estimation of (1) is completed, it is desirable to test the adequacy of the model. It is obvious that the residuals after the fit should not have any serial correlation. Instead of the Durbin-Watson type test, the time series literature uses tests based on autocorrelations of residuals, $\hat{\rho}_j$, their sums of squares, etc. A common graphical procedure plots the actual $\hat{\rho}_j$ against j along with the 95% confidence band based on asymptotic standard errors. If the observed $\hat{\rho}_j$ are beyond the 95% confidence interval, it indicates that one has to modify the model.

The goodness of fit is often tested by Akaike information criterion

(AIC) = $n \log \text{estimated}(\sigma^2) + 2p$, where p = total number of parameters estimated. Another similar criterion is:

Schwartz (Bayesian information criterion) BIC = $n \log \text{estimated}(\sigma^2) + p \log n$.

Instead of confidence intervals on $\hat{\rho}_j$ one may decide to use a single statistic. Hosking (1980) reviews a number of Lagrange Multiplier tests for this purpose. Note that in time series ARMA(p,q) modeling one may specify the orders p and q , which describes the null hypothesis. There is a variety of alternative hypotheses possible with other orders of p and q , possible interactions, non-linearities, etc. If the statistical test requires one to specify a specific alternative, it would be difficult to use. The Lagrange Multiplier tests are designed against general, non-specific alternatives. Hence, the popular tests are of this type.

Portmanteau Test by Box and Pierce (1970)

The word portmanteau means a large leather suitcase. Assume that certain conditions given by Box and Pierce (1970) are satisfied: (i) that the number of lags considered h_T increase as the

sample size T increase, (ii) that the Wold decomposition of y_t has coefficients of order $O(1/\sqrt{T})$ for $j \geq h_T$, and (iii) that h_T is of order $O(\sqrt{T})$. Then the statistic

$$Q = T \sum_{j=1}^{h'} \hat{\rho}_j^2 \quad \text{where no. of items added } h' = h_T \quad (11)$$

Since $\hat{\rho}_j$ are independently distributed in large samples, it can be shown that $Q \sim \chi^2(h' - p - q)$, a Chi-square random variable with the indicated degrees of freedom. If the observed Q is greater than the tabulated Chi-square value, one concludes that the estimated ARMA model is defective. This test is effective when the data are from a white noise process. If numerous $\hat{\rho}_j \neq 0$, Q gets inflated. The choice of h' is arbitrary and there may be a loss of power if Q is used for MA models. For the MA(1) model, the Durbin-Watson type (von Neumann ratio based) test based on $\hat{\rho}_1$ may have higher power than the Q test. To improve the performance of Q in small samples the following modification is suggested.

Ljung Box Modification of the Portmanteau Test

Ljung and Box (1978) argue that under the null hypothesis that the estimated model is adequate, the following statistic gives a closer approximation to the true critical region:

$$Q' = T(T+2) \sum_{j=1}^{h'} \hat{\rho}_j^2 / (T - j) \sim \chi^2(h' - p - q) \quad (12)$$

Davies et al (1977) point out that the variance of Q' statistic exceeds that of the Chi-square.

Autocorrelation of Squared Residuals

Since the publication of Engle (1982), econometricians are often concerned about the autoregressive conditional heteroscedasticity (ARCH) effects among the errors. In these models the error variance may be unconditionally constant ($=\sigma^2$), if it is conditioned on past values of the variable, it is not a constant:

$$E(a_{t+1} | y_t, t_{-1}, \dots, y_1) = h(y_t, t_{-1}, \dots, y_1). \quad (13)$$

Tests for ARCH effects are discussed by Engle(1982) and others. A simple portmanteau test for ARCH effects may be based on the autocorrelation of squared residuals ρ_{jj} . Granger and Anderson (1978b) report examples where there may be significant ρ_{jj} even though the ρ_j are uncorrelated. The following statistic suggested by McLeod and Li (1983)

$$Q'' = T(T+2) \sum_{j=1}^{h'} \hat{\rho}_j^2 / (T - j) \sim \chi^2(h') \quad (12)$$

Various portmanteau tests Q , Q' and Q'' considered above are found to be useful in rejecting inadequate models. They are not very robust in distinguishing between models for ultimate selection. Hosking (1980) views the Q tests as Lagrange Multiplier tests and shows that if the alternative hypothesis is an AR($p+1$) or an MA($q+1$) process, the statistic is the same.

Turning Point Tests

Brockwell and Davis (1987 p.302) discuss many of these and related tests. If there is a turning point at i if $1 < i < n$ and if $y_{i-1} < y_i$ and at the same time $y_i > y_{i+1}$. If y_1, y_2, \dots, y_n are random iid sequence, then the probability of a turning point at time i is $2/3$. The expected number of turning points (T) is $E(T) = 2(n-2)/3$ and variance $Var(T) = (16n-29)/90$ hence $(T - E(T))/\sqrt{Var(T)}$ is asymptotically unit normal.

Normality Tests:

The assumption of normality is present in maximum likelihood and many related estimation methods as well as in testing. Testing for normality is usually based on Pearson's measures of skewness and kurtosis

$$\text{Skew} = \sqrt{\beta_1} = \mu_3 / (\mu_2)^{3/2} \quad \text{and} \quad \text{Kurt} = \beta_2 = \mu_4 / (\mu_2)^2 .$$

where $\mu_j = \sum (y_t - \bar{y})^j / T$, the j -th central moment. Bowman and Shenton (1975) show that the estimates $\hat{\text{Skew}}$ are asymptotically normal, or $\hat{\text{Skew}} \sim AN(0, 6/T)$ and $\hat{\text{Kurt}} \sim AN(3, 24/T)$. Since the estimated Skew and Kurt are independent of each other, an asymptotic test statistic is $(T/6)\hat{\text{Skew}} + (T/24)(\hat{\text{Kurt}} - 3)^2$. If the distribution is normal (under the null), this statistic has χ^2 as the asymptotic distribution with 2 degrees of freedom. Granger and Newbold (1977, p. 315) note that the asymptotic variance of $\hat{\text{Skew}}$ is $(6/T)\sum_{j=-\infty}^{\infty} \rho_j^3$ and the asymptotic variance of $\hat{\text{Kurt}}$ is $(24/T)\sum_{j=-\infty}^{\infty} \rho_j^4$ where the ρ_j may be approximated by $\hat{\rho}_j$ and the infinite sums replaced by finite sums as an approximation. The approximation is reasonable for moderate sample sizes, but not for small samples.