3.1 Limits

Definition:

Let $\{b_n\}$ be a sequence of real numbers. If there exists a real number b and if for every real $\delta < 0$ there exists an integer $N(\delta)$ such that for all $n \geq N(\delta)$, $|b_n - b| < \delta$, then b is the **limit** of the sequence $\{b_n\}$.

In this definition the constant δ can take on any real value, but its is the very small values of δ that provide the definition with its impact. By choosing δ very small, we ensure that b_n gets arbitrarily close to its limit b for all n sufficiently large. When a limit exists we say that the sequence $\{b_n\}$ *converges to* b as n tends to infinity, written $b_n \to b$ as $n \to \infty$.

Some examples of limits:

1. Let $b_n = 1 - \frac{1}{n}$. Then $b_n \to 1$. What happens if $b_n = 1 - \frac{1}{n^2}$?
2. Let $b_n = (1 + \frac{a}{n})^n$. then $b_n \to e^a$.
3. Let $b_n = n^2$. Then $b_n \to \infty$.
4. Let $b_n = (-1)^n$. Then no limit exists.

The concept of a limit extends directly to sequences of real vectors. Let b_n be a $k \times 1$ vector with real elements $b_{ni},\ i = 1, \cdots, k$. If $b_n \to b_i,\ 1, \cdots, k$, then $b_n \to b$, where b has elements $b_i,\ i = 1, \cdots, k$. An analogous extension applies to matrices.

3.2 Almost Sure Convergence

Sequences that converge almost surely can be manipulated in almost exactly the same ways as nonrandom sequences.

let ω represent the entire random sequence $\{Z_t\}$. Interest typically centers on averages such as

$$b_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} Z_t.$$

DEFINITION:

Let $\{b_n(\omega)\}$ be a sequence of real-valued random variables. We say that $b_n(\omega)$ converges *almost surely* to b, written $b_n(\omega) \xrightarrow{\text{a.s.}} b$ if and only if there exists a real number b such that $P[\omega : b_n(\omega) \to b] = 1$.

The probability measure P describes the distribution of ω and determines the joint distribution function of the entire sequence $\{Z_t\}$. A sequence $b_n(\omega)$ converges almost surely if the probability of obtaining a realization of the sequence $\{Z_t\}$ for which convergence to b occurs is unity. Equivalently, the probability of observing a realization of $\{Z_t\}$ for which convergence to b does not occur is zero. Failure to converge is possible but will almost never happen under this definition. Obviously, then, nonstochastic convergence implies almost sure convergence.

Since a realization of ω can be regarded as a point in an infinite-dimensional space, $b_n(\omega)$ is sometimes said to converge *almost everywhere* (a.e) in that space. Other common terminology is that $b_n(\omega)$ converges to b *with probability* 1, (w.p.1) or that $b_n(\omega)$ is *strongly consistent* for b. When no ambiguity is possible, we drop ω and simply write $b_n \xrightarrow{\text{a.s.}} b$.

EXAMPLE:

Let $\bar{Z}_n = n^{-1}\sum_{t=1}^{n} Z_t$, where $\{Z_t\}$ is a sequence of independent identically distributed

(i.i.d.) random variables with $E(Z_t) = \mu \leq \infty$. Then $\bar{Z}_n \xrightarrow{\text{a.s.}} \mu$, by Komolgorov strong law of large numbers—we will discuss this next week.

The almost sure convergence of the sample mean illustrated by this example occurs under a wide variety of conditions on the sequence $\{Z_t\}$.

As with nonstochastic limits, the almost sure convergence concept extends immediately to vectors and matrices of finite dimension. almost sure convergence element by element suffices for almost sure convergence of vectors and matrices.

IMPORTANT PROPOSITION (Slutzky's Theorem):

Given $g : \Re^k \to \Re^l$ $(k, l \leq \infty)$. and any sequence $\{b_n\}$ such that $\bar{b}_n \xrightarrow{\text{a.s.}} b$, where b_n and b are $k \times 1$ vectors, if g is continuous at b, then $g(b_n) \xrightarrow{\text{a.s.}} g(b)$.

PROOF:

Since $b_n(\omega) \xrightarrow{\text{a.s.}} b$ implies $g(b_n(\omega)) \to g(b)$, $[\omega : b_n(\omega) \to b] \subset [\omega : g(b_n(\omega)) \to g(b)]$. Hence $1 = P[\omega : b_n(\omega) \to b] \leq P[\omega : g(b_n(\omega)) \to g(b)] \leq 1$, so that $g(b_n) \xrightarrow{\text{a.s.}} g(b)$.

This result is one of the most important of the class, since consistency results for many of our estimators follow by simply applying this proposition.

3.3 Proof of Existence (Asymptotic Conditions)

Suppose:

1. $Y = X\beta_0 + \epsilon$;

2. $X'\epsilon/n \xrightarrow{\text{a.s.}} 0$;

3. $X'X/n \xrightarrow{\text{a.s.}} M$, finite and positive definite.

Then $\hat{\beta}_n$ exists a.s. for all n sufficiently large, and $\hat{\beta}_n \xrightarrow{\text{a.s.}} \beta_0$.

PROOF:

Since $X'X/n \xrightarrow{\text{a.s.}} M$, it follows from Slutzky's Theorem that $det(X'X/n) \xrightarrow{\text{a.s.}} det(M)$. Because M is positive definite, $det(M) > 0$. It follows that $det(X'X/n) > 0$ a.s. for all n sufficiently large, so $(X'N/n)^{-1}$ exists a.s. for all n sufficiently large. Hence $\hat{\beta}_n = (X'X/n)^{-1}X'Y/n$ exists a.s. for all n sufficiently large.

Now $\hat{\beta}_{n'} = \beta_0 + (X'X/n)^{-1}X'\epsilon/n$ by assumption 1. It follows, by Slutzky's Theorem, that $\hat{\beta}_n \xrightarrow{\text{a.s.}} \beta_0 + M^{-1} \times 0 = \beta_0$, given assumptions 2 and 3.

This is a fundamental consistency result for least squares estimation in many commonly encountered situations. Whether this result applies in a given situation depends on the nature of the data. For example, if our observations are randomly drawn from a population, as in a pure cross section, they may be taken to be i.i.d. The conditions of this consistency theorem hold for i.i.d. observations provided $E(X'X) = M$, finite and positive definite, and $E(X'\epsilon) = 0$. Komolgorov's strong law of large numbers ensures that these conditions lead to our theorem. If the observations are dependent (as in a time series), different laws of large

numbers must be applied to guarantee that the appropriate conditions hold—e.g., Markov's Theorem.

4 Basic Time Series Issues

4.1 Example: The nonaccelerating inflation rate of unemployment (NAIRU)

The NAIRU model, the standard model for inflation, is usually based on the expectations-augmented Phillips relation

$$\pi_t - \pi_t^e = \beta\left(u_t - u_t^*\right) + \delta X_t + v_t, \tag{95}$$

where π_t is an estimate of the actual inflation rate, π_t^e is the expected inflation rate, u_t is the unemployment rate, u_t^* is the NAIRU, X_t contains additional regressors intended to control for supply shocks, and v_t is an error term.

We use a "random walk" model for inflationary expectations (Congressional Budget Office 1994; Fuhrer 1995; Gordon 1990; Staiger, Stock and Watson (SSW) 1997a, 1997b; Tootell 1994; Weiner 1993). According to the "random walk model," $\pi_t^e = \pi_{t-1}$, so that $\pi_t - \pi_t^e = \Delta\pi_t$. Hence,

$$\Delta\pi_t = \beta\left(u_t - u_t^*\right) + \delta X_t + v_t. \tag{96}$$

Equation 96 neglects the possibility of serial correlation in the error term. It is therefore conventional to estimate an autoregressive specification:

$$\Delta\pi_t = \beta(L)\left(u_t - u_t^*\right) + \delta(L)\Delta\pi_{t-1} + \gamma(L)X_t + \epsilon_t, \tag{97}$$

where L is the lag operator, $\beta(L)$, $\delta(L)$, and $\gamma(L)$ are lag polynomials and ϵ_t is a serially uncorrelated error term.

Equation 97 is difficult to estimate because the model is nonlinear in the parameters. When the NAIRU, u_t^*, does not vary with t, Equation 97 can be rewritten in a form which can be conveniently estimated by ordinary least squares (OLS):

$$\Delta\pi_t = \mu + \beta(L)u_t + \delta(L)\Delta\pi_{t-1} + \gamma(L)X_t + \epsilon_t. \tag{98}$$

The estimate of the NAIRU is then

$$\hat{u}^* = \frac{-\hat{\mu}}{\hat{\beta}(1)}, \tag{99}$$

where $\beta(1) = \sum_{i=1}^{p} \beta_i$, with p being the order of the lag polynomial $\beta(L)$. Notice that the NAIRU is a nonlinear function of the coefficients μ and $\beta(1)$.

The estimation approach outlined in Equations 98–99 can be altered for the case where the NAIRU varies with time. One way to accomplish this is to replace $\hat{\mu}$ with $\sum_{j=1}^{l} \alpha_j \mathrm{H}^{j-1}(t)$ where H^i is a Hermite polynomial of order i and where t is the time argument centered around 0. Therefore,

$$\Delta\pi_t = \sum_{j=1}^{l} \alpha_j \mathrm{H}^{j-1}(t) + \beta(L)u_t + \delta(L)\Delta\pi_{t-1} + \gamma(L)X_t + \epsilon_t, \tag{100}$$

and the estimate of the time-varying NAIRU is

$$\hat{u}_t^* = \frac{-\sum_{j=1}^{l} \hat{\alpha}_j \mathrm{H}^{j-1}(t)}{\hat{\beta}(1)}. \tag{101}$$

SSW (1997a, 1997b) use several approaches to estimate the time-varying NAIRU. One of them is similar to the one specified in Equation 100. They do not allow contemporaneous unemployment and supply shocks to affect the current period's inflation rate. SSW (1997a, 1997b) approximate the NAIRU by a cubic spline in time, written as $\bar{\phi}' S_t$, where S_t is a vector of deterministic functions of time. The SSW estimation equation for the spline

approximated time-varying NAIRU is

$$\Delta\pi_t = \phi' S_{t-1} + \beta(L)u_{t-1} + \delta(L)\Delta\pi_{t-1} + \gamma(L)X_{t-1} + \epsilon_t, \tag{102}$$

where $\phi = -\beta(1)\bar{\phi}$ and $\hat{u}_t^* = -\hat{\phi}' S_t / \hat{\beta}(1)$. SSW estimate Equation 102 by OLS.

Following SSW (1997b) the lag operators begin with u_{t-1} and X_{t-1} instead of u_t and X_t—see Equation 102. Also following SSW (1997b), We use monthly lags over one year for both u_t and $\Delta\pi_t$ and lags over one quarter for the supply shocks, X_t. The supply shocks are specified as the difference between food and energy inflation and overall CPI inflation.

4.2 Hermite Polynomial Series

Hermite polynomials form one of the families of classical orthogonal polynomials (Szegö 1975). They are often used to implement cubic splines (de Boor 1978). The variance-covariance matrix of orthogonal polynomials is often well-conditioned when compared to the family of nonorthogonal polynomials (Hinich and Roll 1981). Other orthogonal polynomials (e.g., Legendre polynomials) also have this nice property. An alternative approach is cubic splines.

Hinich and Roll (1981) show that it is useful to use orthogonal polynomials which are a function of time to estimate time-varying parameters. They advocate such an estimation strategy in place of a variety of usual estimation strategies such as creating special event variables to take care of the time trends.

The Hermite polynomials may be defined for $j > 0$ by the recurrence relation

$$H^{j+1}(t) = 2tH^j(t) - 2jH^{j-1}(t).$$

where t is the argument (time), j is the order and where $H^0(t) = 1$ and $H^1(t) = 2t$. The argument, t, must be centered around 0 because the polynomials are orthogonal on the symmetric interval $(-\infty, \infty)$ with respect to the weighting function $w(t) = e^{-t^2}$,

$$\int_{-\infty}^{+\infty} H^j(t)H^k(t)e^{-t^2} \; dt = 2^k k! \pi^{\frac{1}{2}} \delta^{jk}, \qquad j, k \geq 0.$$

It is helpful if the time index for Hermite polynomials is centered around zero.

A few Hermite polynomials are listed below:

$$H^2(t) = 4\,t^2 - 2$$

$$H^3(t) = 8\,t^3 - 12\,t$$

$$H^4(t) = 16\,t^4 - 48\,t^2 + 12$$

$$H^{16}(t) = 65536\,t^{16} - 3932160\,t^{14} + 89456640\,t^{12} - 984023040\,t^{10} + 5535129600\,t^8$$
$$- 15498362880\,t^6 + 19372953600\,t^4 - 8302694400\,t^2 + 518918400.$$

Note how quickly the polynomials increase in size. This is one of the reasons why the parameters associated with the polynomials are extremely difficult to interpret directly.

It should be noted that Hermite polynomials are often used to approximate a wide variety of functions such as integrals (through the use of quadrature) and distributions.

4.3 Distributed Lags

It is often the case that a dependent variable y_t is thought to depend on several current and lagged values of an independent variable x_t. One way to model this type of dependence is to use a **distributed lag model** such as

$$y_t = \alpha + \sum_{j=0}^{q} \beta_j x_{t-j} + u_t, \quad u_t \sim IID(0, \sigma^2), \tag{103}$$

where the constant term α and the coefficients β_j are to be estimated. The integer q is the length of the longest lag.

An obvious problem with this type of model is that because the x_t's will usually be highly correlated with eachother, the least quares estimates of the coefficients β_j will tend to be quite imprecise.

However, if we are interested not in the individual coefficients but on their sum, which measures the long-run effect on y_t of a given change in x_t, we are in better shape. For even if the individuals β_j's are estrimatd very imprecisely, their sum may be estimated with precision.

Let $V(\hat{\beta})$ denote the covariance matrix of the vector of least squares estimates $\hat{\beta}$ with typical element $\hat{\beta}_j$. Then, if $\hat{\gamma}$ denotes the sum of the $\hat{\beta}_j$'s, the variance of $\hat{\gamma}$ is

$$V(\hat{\gamma}) = \sum_{j=0}^{q} V(\hat{\beta}_j) + 2 \sum_{j=0}^{q} \sum_{k=0}^{j-1} cov(\hat{\beta}_j, \hat{\beta}_k). \tag{104}$$

If x_{t-j} is positively correlated with x_{t-k} for all $j \neq k$, the covariance terms will usually be negative. When they are large and negative, as is often the case, $V(\hat{\gamma})$ may be much smaller than the sum of the $V(\hat{\beta}_j)$'s or, indeed, than any individual $V(\hat{\beta}_j)$.

If the parameter of interest if γ rather than the individual β_j's, the easiest approach is

simply to estimate a reparametrized version of this model by least squares:

$$y_t = \alpha + \gamma x_t + \sum_{j=1}^{q} \beta_j (x_{t-j} - x_t) + u_t. \tag{105}$$

I may be verified that the coefficient γ on x_t in Equation 105 is indeed equal to the sum of the β_j's in the original model. The advantage of this reparametrization is that the standard error of $\hat{\gamma}$ is immediately available from the regression output.

4.4 Serial Correlation AR(1)

The **first-order autoregressive** or **AR(1)** process:

$$u_t = \rho u_{t-1} + \epsilon_t, \quad \epsilon_t \sim IID(0, \omega^2), \quad |\rho| < 1. \tag{106}$$

This stochastic process says that the error at time t, u_t, is equal to some fraction ρ of the error at time $t-1$ (with the sign changed if $\rho < 0$, plus a new error term or **innovation** ϵ_t that is homoskedastic and independent of all past and future innovations. Thus in each period part of the error term is the last period's error term, shrunk somewhat toward zero and possibly changed in sign, and part is the innovation ϵ_t.

The $|\rho| < 1$ is called a **stationarity condition**. It ensures that the variance of u_t tends to a limiting value, σ^2, rather than increasing without limit as t gets large. By substituting successively for $u_{t-1}, u_{t-2}, \cdots$, we see that

$$u_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{t-2}, \rho^3\epsilon_{t-3} + \cdots . \tag{107}$$

Thus, using the fact that the innovations ϵ_t are independent, the variance of u_t is seen

to be

$$\sigma^2 = V(u_t) = \omega^2 + \rho^2\omega^2 + \rho^4\omega^2 + \rho^6\omega^2 + \cdots \tag{108}$$

$$= \frac{\omega^2}{1-\rho^2} \tag{109}$$

This derivation is true only if the stationarity condition $|\rho| < 1$ holds. That condition is necessary for the infinite series $1+\rho^2+\rho^4+\rho^6+\cdots$ to converge. In conventional applications, where u_t is the error term appended to a regression model, this is a very reasonable condition to impose.

For a stationary AR(1) process which has been going on for a reasonable length of time, the error terms u_t will each have variance $\sigma^2 = \frac{\omega^2}{1-\rho^2}$. We can write:

$$u_t = \epsilon_t + \rho\epsilon_{t-1} + \cdots + \rho^{j-1}\epsilon_{t-j+1} + \rho^j u_{t-j}, \tag{110}$$

expressing u_t as a function of u_{t-j} and of all the innovation between periods $t-j+1$ and t. Hence the covariance of u_t and u_{t-j} may be calculated as

$$E[(\epsilon_t + \rho\epsilon_{t-1} + \cdots + \rho^{j-1}\epsilon_{t-j+1} + \rho^j u_{t-j})u_{t-j}]. \tag{111}$$

Since the innovations (ϵ) between periods $t-j+1$ and t are independent of u_{t-j}, the covariance is simply

$$E(\rho^j u_{t-j}^2) = \rho^j E(u_t^2) \tag{112}$$

$$= \rho^j E(u_t^2) \tag{113}$$

$$= \frac{\rho^j\omega^2}{1-\rho^2} \tag{114}$$

$$= \rho^j\sigma^2. \tag{115}$$

Thus we conclude that the covariance matrix of U is

$$\Omega = \frac{\omega^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}, \tag{116}$$

the matrix in brackets being the correlation matrix of U. It is evident from Equation 116 that every element of U is correlated with every other element of U, but except when $|\rho|$ is very close to 1, this correlation will tend to die out quite quickly as the time periods become further apart—this is the ergodicity assumption.

The ergodicity assumption accords well both with intuition and with the actual behavior of the residuals form many regression models estimated with time-series data. Thus it is not surprising that the AR(1) process is very frequently used in applied work.

The AR(1) process, is, however, a very special case. There are numerous other stochastic processes that error terms could reasonably follow.

4.5 The Consequences of Serial Correlation

What are the consequences if we use least squares to estimate a model in which the error terms are in fact serially correlated? Let us suppose we estimate the model

$$Y = X\beta + u, \quad E(uu') = \sigma^2 I, \tag{117}$$

when the data-generating process is actually

$$Y = X\beta_0 + u, \tag{118}$$

$$u_t = \rho u_{t-1} + \epsilon_t, \quad \epsilon \sim IID(0, \omega_0^2).$$

The OLS estimator is

$$\hat{\beta} = (X'X)^{-1}X'Y, \tag{119}$$

which under the GDP is equal to

$$\hat{\beta} = (X'X)^{-1}X'(X\beta_0 + u) \tag{120}$$

$$= \beta_0 + (X'X)^{-1}X'u. \tag{121}$$

Provided that X is exogenous, $\hat{\beta}$ will still be unbiased, because the fact that the u_t's are serially correlated does not prevent $E(X'u)$ from being zero. If X is not exogenous, $\hat{\beta}$ will be consistent as long as $plim(n^{-1}X'u)$ is equal to zero.

Inferences about β will not be correct, however. Assuming that X is exogenous, we see that

$$E(\hat{\beta} - \beta_0)(\hat{\beta} - \beta_0)' = E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \tag{122}$$

$$= (X'X)^{-1}X'\Omega_0 X(X'X)^{-1}, \tag{123}$$

where Ω_0 is the matrix defined by the GDP.

Except in special cases, it is not possible to say whether the incorrect standard error estimates obtained using OLS will be larger or smaller than the correct ones obtained by taking the square roots of the diagonal elements in the usual fashion. However, analysis of special cases suggests that for values of ρ greater than 0 (the most commonly encountered cases) the incorrect OLS standard errors are usually too small—see Nicholls and Pagan (1977), Sathe and Vinod (1974), and Vinod (1976).

4.6 Moving Average and ARMA Processes

Autoregressive processes are not the only way to model stationary time series. The other basic type of stochastic process is the **moving average**, or **MA**, process. The simplest moving average process is the **first-order moving average**, or **MA(1)**, process

$$u_t = \epsilon_t + \alpha_1 \epsilon_{t-1}, \quad \epsilon_t \sim IID(0, \omega^2), \tag{124}$$

in which the error u_t is literally a moving average of two successive innovations, ϵ_t and ϵ_{t-1}. Thus ϵ_t affects both u_t and u_{t+1} but does not affect u_{t+j} for $j > 1$.

Year-on-year inflation is an example of a MA process.

The more general MA(q) process may be written as

$$u_t = \epsilon_t + \alpha_1 \epsilon_{t-1} + \alpha_2 \epsilon_{t-2} + \cdots + \alpha_q \epsilon_{t-q}, \quad \epsilon_t \sim IID(0, \omega^2). \tag{125}$$

Finite-order MA processes are necessarily stationary, since each u_t is a weighted sum of a finite number of innovations $\epsilon_t, \epsilon_{t-1} \cdots$. Thus we do not have to impose stationarity conditions.

We do, however, have to impose an **invertibility condition** if we want α to be identifiable from data. In the MA(1) case, this condition is that $|\alpha_1| \leq 1$.

The reason we need an invertibility condition is that otherwise there will, in general, be more than one value of α that will yield any observed behavior pattern of the u_t's.

For example, the MA(1) process with $\alpha_1 = \gamma$, $-1 < \gamma < 1$, can be shown to be indistinguishable form an MA(1) process with $\alpha_1 = \dfrac{1}{\gamma}$.

It is straightforward to calculate the covariance matrix for a moving average process. for

example, in the MA(1) case the variance of u_t is evidently

$$\sigma^2 = E(e_t + \alpha_1 \epsilon_{t-1})^2 \tag{126}$$

$$= \omega^2 + \alpha_1^2 \omega^2 \tag{127}$$

$$= (1 + \alpha_1^2)\omega^2, \tag{128}$$

the covariance of u_t and u_{t-1} is

$$E(\epsilon_t + \alpha_1 \epsilon_{t-1})(\epsilon_{t-1} + \alpha_1 \epsilon_{t-2}) = \alpha_1 \omega^2, \tag{129}$$

and the covariance of u_t and u_{t-j} for $j > 1$ is zero. Thus the covariance matrix of U is

$$\omega^2 \begin{bmatrix} 1+\alpha_1^2 & \alpha_1 & 0 & \cdots & 0 & 0 & 0 \\ \alpha_1 & 1+\alpha_1^2 & \alpha_1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \alpha_1 & 1+\alpha_1^2 & \alpha_1 \\ 0 & 0 & 0 & \cdots & 0 & \alpha_1 & 1+\alpha_1^2 \end{bmatrix} \tag{130}$$

Notice that the correlation between successive error terms varies only between -0.5 and 0.5, since those are the smallest and largest possible values of $\dfrac{\alpha_1}{1+\alpha_1^2}$, achieved when $\alpha_1 = -1$ and $\alpha_1 = 1$, respectively. It is thus evident that an MA(1) process cannot be appropriate when the observed correlation between successive residuals is large in absolute value.