A Primer on Asymptotics

Eric Zivot
Department of Economics
University of Washington

September 30, 2003
Revised: January 7, 2013

1 Introduction

The two main concepts in asymptotic theory covered in these notes are

- Consistency
- Asymptotic Normality

Intuition

- consistency: as we get more and more data, we eventually know the truth
- asymptotic normality: as we get more and more data, averages of random variables behave like normally distributed random variables

1.1 Motivating Example

Let $X_1, \ldots, X_n$ denote an independent and identically distributed (iid) random sample with $E[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2$. We don't know the probability density function (pdf) $f(X_i, \boldsymbol{\theta})$, but we know the value of σ^2. The goal is to estimate the mean value μ from the random sample of data. A natural estimate is the sample mean

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Using the iid assumption, straightforward calculations show that

$$E[\hat{\mu}] = \mu, \ \text{var}(\hat{\mu}) = \frac{\sigma^2}{n}.$$

Since we don't know $f(X_i, \boldsymbol{\theta})$ we don't know the pdf of $\hat{\mu}$. All we know about the pdf of μ is that $E[\hat{\mu}] = \mu$ and $\text{var}(\hat{\mu}) = \frac{\sigma^2}{n}$. However, as $n \to \infty$, $\text{var}(\hat{\mu}) = \frac{\sigma^2}{n} \to 0$ and the pdf of $\hat{\mu}$ collapses at μ. Intuitively, as $n \to \infty$, $\hat{\mu}$ converges in some sense to μ. In other words, the estimator $\hat{\mu}$ is consistent for μ.

Furthermore, consider the standardized random variable

$$Z = \frac{\hat{\mu} - \mu}{\sqrt{\text{var}(\hat{\mu})}} = \frac{\hat{\mu} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right).$$

For any value of n, $E[Z] = 0$ and $\text{var}(Z) = 1$, but we don't know the pdf of Z since we don't know $f(X_i,\theta)$. Asymptotic normality says that as n gets large, the pdf of Z is well approximated by the standard normal density. We use the short-hand notation

$$Z = \sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right) \overset{A}{\sim} N(0,1) \tag{1}$$

to represent this approximation. The symbol " $\overset{A}{\sim}$ " denotes "asymptotically distributed as", and represents the asymptotic normality approximation. Dividing both sides of (1) by $\sqrt{n}/\sigma$ and adding μ, the asymptotic approximation may be re-written as

$$\hat{\mu} = \mu + \frac{Z\sigma}{\sqrt{n}} \overset{A}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right). \tag{2}$$

The above is interpreted as follows: the pdf of the estimate $\hat{\mu}$ is asymptotically distributed as a normal random variable with mean μ and variance $\frac{\sigma^2}{n}$. The quantity $\frac{\sigma^2}{n}$ is often referred to as the *asymptotic variance* of $\hat{\mu}$, and is denoted $\text{avar}(\hat{\mu})$. The square root of $\text{avar}(\hat{\mu})$ is called the *asymptotic standard error* of $\hat{\mu}$ and is denoted $\text{ASE}(\hat{\mu})$. With this notation, (2) may be re-expressed as

$$\hat{\mu} \overset{A}{\sim} N\left(\mu, \text{avar}(\hat{\mu})\right) \text{ or } \hat{\mu} \overset{A}{\sim} N\left(\mu, \text{ASE}(\hat{\mu})^2\right). \tag{3}$$

The quantity σ^2 in (2) is sometimes referred to as the asymptotic variance of $\sqrt{n}(\hat{\mu} - \mu)$.

The asymptotic normality result (2) is commonly used to construct a confidence interval for μ. For example, an asymptotic 95% confidence interval for μ has the form

$$\hat{\mu} \pm 1.96 \times \sqrt{\text{avar}(\hat{\mu})} = 1.96 \times \text{ASE}(\hat{\mu}).$$

This confidence interval is asymptotically valid in that, for large enough samples, the probability that the interval covers μ is approximately 95%.

The asymptotic normality result (3) is also commonly used for hypothesis testing. For example, consider testing the hypothesis $H_0 : \mu = \mu_0$ against the hypothesis $H_1 : \mu \neq \mu_0$. A commonly used test statistic is the t-ratio

$$t_{\mu=\mu_0} = \frac{\hat{\mu} - \mu_0}{\text{ASE}(\hat{\mu})} = \sqrt{n}\left(\frac{\hat{\mu} - \mu_0}{\sigma}\right). \tag{4}$$

If the null hypothesis $H_0 : \mu = \mu_0$ is true then the asymptotic normality result (1) shows that the t-ratio (4) is asymptotically distributed as a standard normal random variable. Hence, for $\alpha \in (0, 1)$ we can reject $H_0 : \mu = \mu_0$ at the $\alpha \times 100\%$ level if

$$\left| t_{\mu=\mu_0} \right| > z_{1-\alpha/2},$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2) \times 100\%$ quantile of the standard normal distribution. For example, if $\alpha = 0.05$ then $z_{1-\alpha/2} = z_{0.975} = 1.96$.

Remarks

1. A natural question is: how large does n have to be in order for the asymptotic distribution to be accurate? Unfortunately, there is no general answer. In some cases (e.g., $X_i \sim iid$ Bernoulli), the asymptotic approximation is accurate for n as small as 15. In other cases (e.g., X_i follows a near unit root process), n may have to be over 1000. If we know the exact finite sample distribution of $\hat{\mu}$, then, for example, we can evaluate the accuracy of the asymptotic normal approximation for a given n by comparing the quantiles of the exact distribution with those from the asymptotic approximation.

2. The asymptotic normality result is based on the *Central Limit Theorem*. This type of asymptotic result is called *first-order* because it can be derived from a first-order Taylor series type expansion. More accurate results can be derived, in certain circumstances, using so-called higher-order expansions or approximations. The most common higher-order expansion is the *Edgeworth expansion*. Another related approximation is called the *saddlepoint approximation*.

3. We can use *Monte Carlo simulation* experiments to evaluate the asymptotic approximations for particular cases. Monte Carlo simulation involves using the computer to generate pseudo random observations from $f(X_i, \boldsymbol{\theta})$ and using these observations to approximate the exact finite sample distribution of $\hat{\mu}$ for a given sample size n. The error in the Monte Carlo approximation depends on the number of simulations and can be made very small by using a very large number of simulations.

4. We can often use *bootstrap* techniques to provide numerical estimates for $\text{avar}(\hat{\mu})$ and asymptotic confidence intervals. These are alternatives to the analytic formulas derived from asymptotic theory. An advantage of the bootstrap is that under certain conditions, it can provide more accurate approximations than the asymptotic normal approximations. Bootstrapping, in contrast to Monte Carlo simulation, does not require specifying the distribution of X. In particular, *nonparametric bootstrapping* relies on resampling from the observed data. Parametric bootstrapping relies on using the computer to generate pseudo random observations from $f(X_i, \hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the sample estimate of $\boldsymbol{\theta}$.

5. If we don't know σ^2, we have to estimate $\text{avar}(\hat{\mu})$. If $\hat{\sigma}^2$ is a consistent estimate for σ^2, then we can compute a consistent estimate for the asymptotic variance of $\sqrt{n}(\hat{\mu} - \mu)$ by plugging in $\hat{\sigma}^2$ for σ^2 in (2) and compute an estimate for $\text{avar}(\hat{\mu})$

$$\widehat{\text{avar}}(\hat{\mu}) = \frac{\hat{\sigma}^2}{n}.$$

This gives rise to the asymptotic approximation

$$\hat{\mu} \overset{A}{\sim} N\left(\mu, \widehat{\text{avar}}(\hat{\mu})\right) \text{ or } \hat{\mu} \overset{A}{\sim} N\left(\mu, \widehat{\text{ASE}}(\hat{\mu})^2\right). \tag{5}$$

which is typically less accurate than the approximation (2) because of the estimate error in $\widehat{\text{avar}}(\hat{\mu})$.

2 Probability Theory Tools

The main statistical tool for establishing consistency of estimators is the *Law of Large Numbers* (LLN). The main tool for establishing asymptotic normality is the *Central Limit Theorem* (CLT). There are several versions of the LLN and CLT, that are based on various assumptions. In most textbooks, the simplest versions of these theorems are given to build intuition. However, these simple versions often do not technically cover the cases of interest. An excellent compilation of LLN and CLT results that are applicable for proving general results in econometrics is provided in White (1984).

2.1 Laws of Large Numbers

Let $X_1, \ldots, X_n$ be a iid random variables with pdf $f(X, \boldsymbol{\theta})$. For a given function g, define the sequence of random variables based on the sample

$$\begin{aligned} Y_1 &= g(X_1), \\ Y_2 &= g(X_1, X_2), \\ &\vdots \\ Y_n &= g(X_1, \ldots, X_n). \end{aligned}$$

For example, let $X \sim N(\mu, \sigma^2)$ so that $\boldsymbol{\theta} = (\mu, \sigma^2)$ and define $Y_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. This notation emphasizes that sample statistics are functions of the sample size and can be treated as a sequence of random variables.

Definition 1 *Convergence in Probability*

Let $Y_1, \ldots, Y_n$ be a sequence of random variables. We say that Y_n converges in probability to a constant, or random variable, c and write

$$Y_n \xrightarrow{p} c$$

or

$$p \lim_{n\to\infty} Y_n = c$$

if $\forall\ \varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|Y_n - c| > \varepsilon) = 0.$$

■

Remarks

1. $Y_n \xrightarrow{p} c$ is the same as $Y_n - c \xrightarrow{p} 0$.
2. For a vector process, $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nk})'$, $\mathbf{Y}_n \xrightarrow{p} \mathbf{c}$ if $Y_{ni} \xrightarrow{p} c_i$ for $i = i, \ldots, n$.

Definition 2 *Consistent Estimator*

If $\hat{\theta}$ is an estimator of the scalar parameter θ, then $\hat{\theta}$ is consistent for θ if

$$\hat{\theta} \xrightarrow{p} \theta.$$

If $\hat{\boldsymbol{\theta}}$ is an estimator of the $n \times 1$ vector $\boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ if $\hat{\theta}_i \xrightarrow{p} \theta_i$ for $i = 1, \ldots, n$. ■

All consistency proofs are based on a particular LLN. A LLN is a result that states the conditions under which a sample average of random variables converges to a population expectation. There are many LLN results. The most straightforward is the LLN due to Chebychev.

2.1.1 Chebychev's LLN

Theorem 3 *Chebychev's LLN*

Let $X_1, \ldots, X_n$ be iid random variables with $E[X_i] = \mu < \infty$ and $\text{var}(X_i) = \sigma^2 < \infty$. Then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} E[X_i] = \mu$$

■

The proof is based on the famous Chebychev's inequality.

Lemma 4 *Chebychev's Inequality*

Let X by any random variable with $E[X] = \mu < \infty$ and $\text{var}(X) = \sigma^2 < \infty$. Then for every $\varepsilon > 0$

$$\Pr(|X - \mu| \geq \varepsilon) \leq \frac{\text{var}(X)}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$$

■

The probability bound defined by Chebychev's Inequality is general but may not be particularly tight. For example, suppose $X \sim N(0, 1)$ and let $\varepsilon = 1$. Then Chebychev's Inequality states that

$$\Pr(|X| \geq 1) = \Pr(X > 1) + \Pr(X < -1) \leq 1,$$

which is not very informative. Here, the exact probability is

$$\Pr(X > 1) + \Pr(X < -1) = 2 \times \Pr(X < -1) = 0.3173.$$

To prove Chebychev's LLN we apply Chebychev's inequality to the random variable $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ giving

$$\Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\text{var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

It trivially follows that

$$\lim_{n\to\infty} \Pr(|\bar{X} - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

and so $\bar{X} \xrightarrow{p} \mu$.
Remarks

1. The proof of Chebychev's LLN relies on the concept of convergence in mean square. That is,

$$\text{MSE}(\bar{X}, \mu) = E[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n} \to 0 \text{ as } n \to \infty$$

   In general, if $\text{MSE}(\bar{X}, \mu) \to 0$ then $\bar{X} \xrightarrow{p} \mu$. In words, convergence in mean square implies convergence in probability.

2. Convergence in probability, however, does not imply convergence in mean square. That is, it may be the case that $\bar{X} \xrightarrow{p} \mu$ but $\text{MSE}(\bar{X}, \mu) \nrightarrow 0$. This would occur, for example, if $\text{var}(X)$ does not exist.

2.1.2 Kolmogorov's LLN

The LLN with the weakest set of conditions on the random sample $X_1, \ldots, X_n$ is due to Kolmogorov.

Theorem 5 *Kolmogorov's LLN (aka Khinchine's LLN)*

Let $X_1, \ldots, X_n$ be iid random variables with $E[|X_i|] < \infty$ and $E[X_i] = \mu$. Then

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} E[X_i] = \mu$$

■

Remarks

1. If X_i is a continuous random variable with density $f(x_i, \boldsymbol{\theta})$, then

$$E[|X_i|] = \int |x_i| f(x_i, \theta) dx_i < \infty.$$

   This condition controls the tail behavior of $f(x_i, \theta)dx_i$. The tails cannot be too fat such that $E[|X_i|] = \infty$. However, the tails may be sufficiently fat such that $E[|X_i|] < \infty$ but $E[X_i^2] = \infty$.

2. Kolmogorov's LLN does not require $\text{var}(X_i)$ to exist. Only the mean needs to exist. That is, this LLN covers random variables with fat-tailed distributions (e.g., Student's t with 2 degrees of freedom).

2.1.3 Markov's LLN

Chebychev's LLN and Kolmogorov's LLN assume independent and identically distributed (iid) observations. In some situations (e.g. random sampling with cross-sectional data), the independence assumption may hold but the identical distribution assumption does not. For example, the $X_i's$ may have different means and/or variances for each i. If we retain the independence assumption but relax the identical distribution assumption, then we can still get convergence of the sample mean. In fact, we can go further and even relax the independence assumption and only require that the observations be uncorrelated and still get convergence of the sample mean. However, further assumptions are required on the sequence of random variables $X_1, \ldots, X_n$. For example, a LLN for independent but not identically distributed random variables that is particularly useful is due to Markov.

Theorem 6 *Markov's LLN*

Let $X_1, \ldots, X_n$ be a sample of uncorrelated random variables with finite means $E[X_i] = \mu_i < \infty$ and uniformly bounded variances $\text{var}(X_i) = \sigma_i^2 \leq M < \infty$ for $i = 1, \ldots, n$. Then

$$\bar{X} - \bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mu_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{p} 0$$

Equivalently,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

■

Remarks:

1. Markov's LLN does not require the observations to be independent or identically distributed - just uncorrelated with bounded means and variances.
2. The proof of Markov's LLN follows directly from Chebychev's inequality.
3. Sometimes the uniformly bounded variance assumption, $\text{var}(X_i) = \sigma_i^2 \leq M < \infty$ for $i = 1, \ldots, n$, is stated as

   $$\sup_i \ \sigma_i^2 < \infty$$

   where sup denotes the *supremum* or least upper bound.
4. Notice that when the iid assumption is relaxed, stronger restrictions need to be placed on the variances of each of the random variables. That is, we cannot get a LLN like Kolmogorov's LLN that does not require a finite variance. This is a general principle with LLNs. If some assumptions are weakened then other assumptions must be strengthened. Think of this as an instance of the "no free lunch" principle applied to probability theory.

2.1.4 LLNs for Serially Correlated Random Variables

In time series settings, random variables are typically serially correlated (i.e., correlated over time). If we go further and relax the uncorrelated assumption, then we can still get a LLN result. However, we must control the dependence among the random variables. In particular, if $\sigma_{ij} = \text{cov}(X_i, X_j)$ exists for all i, j and are close to zero for $|i - j|$ large; e.g., if

$$\sigma_{ij} \leq M \cdot \rho^{|i-j|}, \ 0 < \rho < 1 \text{ and } M < \infty$$

then it can be shown that

$$\Pr(|\bar{X} - \bar{\mu}| > \varepsilon) \leq \frac{M}{n\varepsilon^2} \to 0$$

as $n \to \infty$. Further details on LLNs for serially correlated random variables will be discussed in the following section on time series concepts.

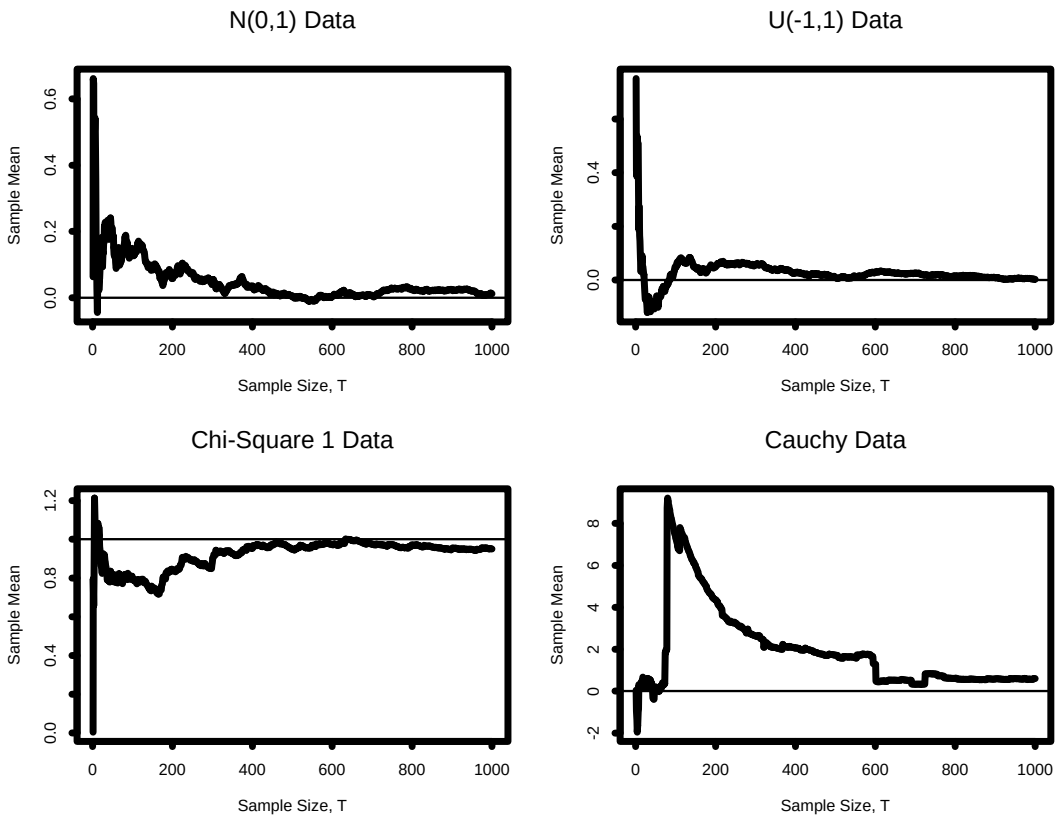

Figure 1: One realization of $Y_n = \bar{X}$ for $n = 1, \ldots, 1000$.

2.1.5 Examples

We can illustrate some of the LLNs using computer simulations. For example, Figure 1 shows one simulated path of $Y_n = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ for $n = 1, \ldots, 1000$ based on random sampling from a standard normal distribution (top left), a uniform distribution over [-1,1] (top right), a chi-square distribution with 1 degree of freedom (bottom left), and a Cauchy distribution (bottom right). For the normal and uniform random variables $E[X_i] = 0$; for the chi-square $E[X_i] = 1$; and for the Cauchy $E[X_i]$ does not exist. For the normal, uniform and chi-square simulations the realized value of the sequence Y_n appears to converge to the population expectation as n gets large. However, the sequence from the Cauchy does not appear to converge. Figure 2 shows 100 simulated paths of Y_n from the same distributions used for figure 1. Here we see the variation in Y_n across different realizations. Again, for the normal, uniform and chi-square distribution the sequences appear to converge, whereas for the Cauchy the sequences do not appear to converge.

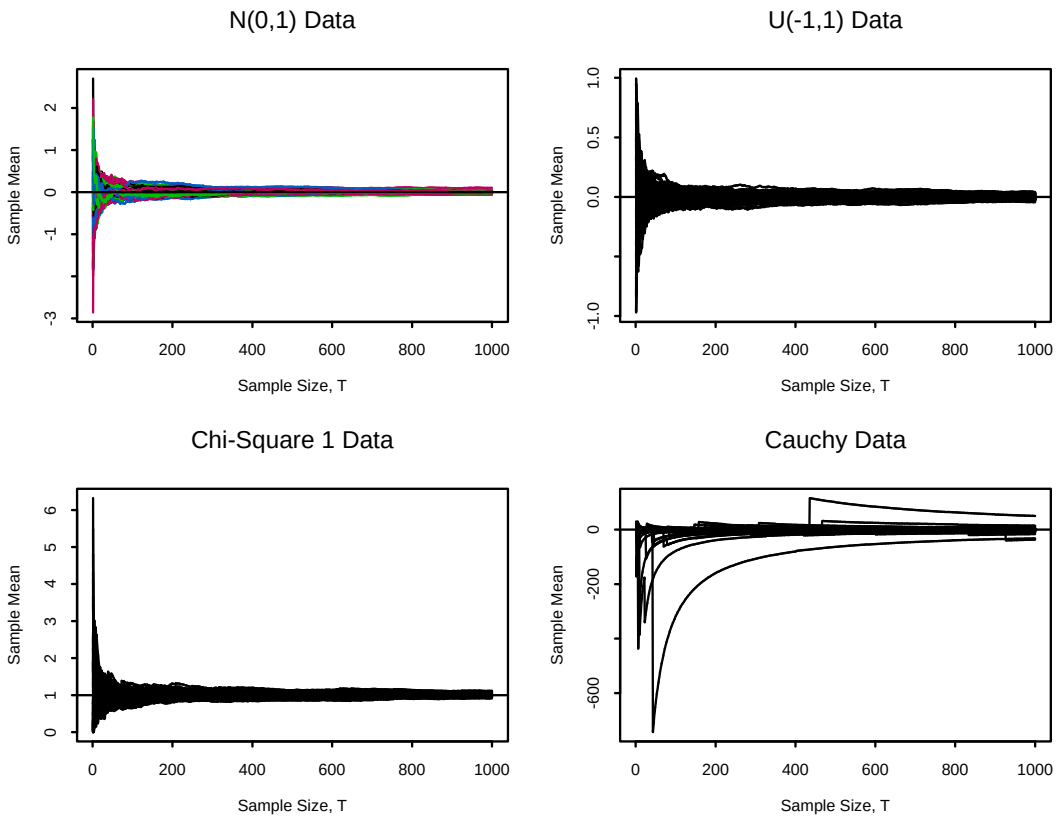

Figure 2: 100 realizations of $Y_n = \bar{X}$ for $n = 1, \ldots, 1000$.

2.2 Results for the Manipulation of Probability Limits

LLN's are useful for studying the limit behavior of averages of random variables. However, when proving asymptotic results in econometrics we typically have to study the behavior of simple functions of random sequences that converge in probability. The following Theorem due to Slutsky is particularly useful.

Theorem 7 *Slutsky's Theorem 1*

Let $\{Y_n\}$ and $\{Z_n\}$ be a sequences of random variables and let b, c and d be constants.

1. If $Y_n \xrightarrow{p} c$ then $bY_n \xrightarrow{p} bc$
2. If $Y_n \xrightarrow{p} c$ and $Z_n \xrightarrow{p} d$ then $Y_n + Z_n \xrightarrow{p} c + d$
3. If $Y_n \xrightarrow{p} c$ and $Z_n \xrightarrow{p} d$ then $\frac{Y_n}{Z_n} \xrightarrow{p} \frac{c}{d}$, provided $d \neq 0; Y_n Z_n \xrightarrow{p} cd$
4. If $Y_n \xrightarrow{p} c$ and $h(\cdot)$ is a continuous function then $h(Y_n) \xrightarrow{p} h(c)$

■

Example 8 *Convergence of the sample variance and standard deviation*

Let $X_1, \ldots, X_n$ be iid random variables with $E[X_1] = \mu$ and $\text{var}(X_1) = \sigma^2 < \infty$. Then the sample variance, given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

is a consistent estimator for σ^2; i.e. $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. The most natural way to prove this result is to write

$$X_i - \bar{X} = X_i - \mu + \mu - \bar{X}, \tag{6}$$

so that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + 2\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)(\mu - \bar{X}) + \frac{1}{n}\sum_{i=1}^{n}(\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 + 2\frac{1}{n}(\mu - \bar{X})\sum_{i=1}^{n}(X_i - \mu) + (\mu - \bar{X})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - (\mu - \bar{X})^2.
\end{aligned}$$

Now, by Chebychev's LLN $\bar{X} \xrightarrow{p} \mu$ so that second term vanishes as $n \to \infty$ using Slutsky's Theorem. To see what happens to the first term, let $W_i = (X_i - \mu)^2$ so that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n} W_i.$$

The random variables $W_1, \ldots, W_n$ are iid with $E[W_i] = E[(X_i - \mu)^2] = \sigma^2 < \infty$. Therefore, by Kolmogorov's LLN

$$\frac{1}{n}\sum_{i=1}^{n} W_i \xrightarrow{p} E[W_i] = \sigma^2,$$

which gives the desired result.

Remarks:

1. Notice that in order to prove the consistency of the sample variance we used the fact that the sample mean is consistent for the population mean. If $\bar{X}$ was not consistent for μ, then $\hat{\sigma}^2$ would not be consistent for σ^2.

2. By using the trivial identity (6), we may write

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\mu-\bar{X})^2 \\ &\quad \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2+o_p(1),\end{aligned}$$

   where $-(\mu-\bar{X})^2=o_p(1)$ denotes a sequence of random variables that converge in probability to zero. That is, if $Y_n=o_p(1)$ then $Y_n \xrightarrow{p} 0$.This short-hand notation often simplifies the exposition of certain derivations involving probability limits.

3. Given that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ and the square root function is continuous, it follows from Slutsky's Theorem that the sample standard deviation, $\hat{\sigma}$, is consistent for σ.

3 Convergence in Distribution and the Central Limit Theorem

Let $Y_1,\ldots,Y_n$ be a sequence of random variables. For example, let $X_1,\ldots,X_n$ be an iid sample with $E[X_i]=\mu$ and $\text{var}(X_i)=\sigma^2$ and define $Y_n=\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$. We say that Y_n *converges in distribution* to a random variable W and write

$$Y_n \xrightarrow{d} W$$

if

$$F_{Y_n}(y)=\Pr(Y_n\leq y)\rightarrow F_W(y)=\Pr(W\leq y) \text{ as } n\rightarrow\infty$$

for every continuity point of the *cumulative distribution function* (CDF) of W.

Remarks

1. In most applications, W is either a normal or chi-square distributed random variable.

2. Convergence in distribution is usually established through *Central Limit Theorems* (CLTs). The proofs of CLTs show the convergence of $F_{Y_n}(y)$ to $F_W(y)$. The early proofs of CLTs are based on the convergence of the moment generating function (MGF) or characteristic function (CF) of Y_n to the MGF or CF of W.

3. If n is large, we can use the convergence in distribution results to justify using the distribution of W as an approximating distribution for Y_n. That is, for n large enough we use the approximation

$$\Pr(Y_n \in A) \approx \Pr(W \in A)$$

for any set $A \subset \mathbb{R}$.

4. Let $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nk})'$ be a multivariate sequence of random variables. Then

$$\mathbf{Y}_n \xrightarrow{d} \mathbf{W}$$

if and only if

$$\boldsymbol{\lambda}'\mathbf{Y}_n \xrightarrow{d} \boldsymbol{\lambda}'\mathbf{W}$$

for any $\boldsymbol{\lambda} \in \mathbb{R}^k$.

3.1 Central Limit Theorems

Probably the most famous CLT is due to Lindeberg and Levy.

Theorem 9 *Lindeberg-Levy CLT (Greene, 2003 p. 909)*

Let $X_1, \ldots, X_n$ be an iid sample with $E[X_i] = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Then

$$Y_n = \sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \xrightarrow{d} Z \sim N(0,1) \text{ as } n \to \infty$$

That is, for all $y \in \mathbb{R}$,

$$\Pr(Y_n \leq y) \to \Phi(y) \text{ as } n \to \infty$$

where

$$\begin{aligned} \Phi(y) &= \int_{-\infty}^{y} \phi(z)dz \\ \phi(z) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \end{aligned}$$

Remark

1. The CLT suggests that we may approximate the distribution of $Y_n = \sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$ by a standard normal distribution. This, in turn, suggests approximating the distribution of the sample average $\bar{X}$ by a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

Theorem 10 *Multivariate Lindeberg-Levy CLT (Greene, 2003 p. 912)*

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be $k-$dimensional iid random vectors with $E[\mathbf{X}_i] = \boldsymbol{\mu}$ and $\text{var}(\mathbf{X}_i) = E[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'] = \boldsymbol{\Sigma}$, where Σ is nonsingular. Let $\Sigma^{-1} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2\prime}$. Then

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}_k),$$

where $N(\mathbf{0}, \mathbf{I}_k)$ denotes a multivariate normal distribution with mean zero and identity covariance matrix. That is,

$$f(\mathbf{z}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\mathbf{z}'\mathbf{z}\right\}.$$

Equivalently, we may write

$$\begin{aligned}\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) &\overset{A}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}) \\ \bar{\mathbf{X}} &\overset{A}{\sim} N(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}).\end{aligned}$$

This result implies that

$$\text{avar}(\bar{\mathbf{X}}) = n^{-1}\boldsymbol{\Sigma}.$$

■

Remark

1. If the $k \times 1$ vector $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}\right)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

The Lindeberg-Levy CLT is restricted to iid random variables, which limits its usefulness. In particular, it is not applicable to the least squares estimator in the linear regression model with fixed regressors. To see this, consider the simple linear model with a single fixed regressor

$$y_i = x_i\beta + \varepsilon_i,$$

where x_i is fixed and ε_i is iid $(0, \sigma^2)$. The least squares estimator is

$$\begin{aligned}\hat{\beta} &= \left(\sum_{i=1}^n x_i^2\right)^{-1} \sum_{i=1}^n x_i y_i \\ &= \beta + \left(\sum_{i=1}^n x_i^2\right)^{-1} \sum_{i=1}^n x_i \varepsilon_i,\end{aligned}$$

and

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^n x_i^2\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \varepsilon_i.$$

The CLT needs to be applied to the random variable $w_i = x_i\varepsilon_i$. However, even though ε_i is iid, w_i is not iid since $\text{var}(w_i) = x_i^2\sigma^2$ and, thus, varies with x_i.

The Lindeberg-Feller CLT is applicable for the linear regression model with fixed regressors.

Theorem 11 *Lindeberg-Feller CLT (Greene, 2003 p. 901)*

Let $X_1, \ldots, X_n$ be independent (but not necessarily identically distributed) random variables with $E[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2 < \infty$. Define $\bar{\mu}_n = n^{-1}\sum_{i=1}^n \mu_i$ and $\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. Suppose

$$\begin{aligned} \lim_{n\to\infty} \max_i \frac{\sigma_i^2}{n\bar{\sigma}_n^2} &= 0, \\ \lim_{n\to\infty} \bar{\sigma}_n^2 &= \bar{\sigma}^2 < \infty, \end{aligned}$$

Then

$$\sqrt{n}\left(\frac{\bar{X} - \bar{\mu}_n}{\bar{\sigma}_n}\right) \xrightarrow{d} Z \sim N(0,1).$$

Equivalently,

$$\sqrt{n}\left(\bar{X} - \bar{\mu}_n\right) \xrightarrow{d} N(0, \bar{\sigma}^2).$$

■

A CLT result that is equivalent to the Lindeberg-Feller CLT but with conditions that are easier to understand and verify is due to Liapounov.

Theorem 12 *Liapounov's CLT (Greene, 2003 p. 912)*

Let $X_1, \ldots, X_n$ be independent (but not necessarily identically distributed) random variables with $E[X_i] = \mu_i$ and $\text{var}(X_i) = \sigma_i^2 < \infty$. Suppose further that

$$E[|X_i - \mu_i|^{2+\delta}] \leq M < \infty,$$

for some $\delta > 0$. If $\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$ is positive and finite for all n sufficiently large, then

$$\sqrt{n}\left(\frac{\bar{X} - \bar{\mu}_n}{\bar{\sigma}_n}\right) \xrightarrow{d} Z \sim N(0,1).$$

Equivalently,

$$\sqrt{n}\left(\bar{X} - \bar{\mu}_n\right) \xrightarrow{d} N(0, \bar{\sigma}^2),$$

where $\lim_{n\to\infty} \bar{\sigma}_n^2 = \bar{\sigma}^2 < \infty$.

Remark

1. There is a multivariate version of the Lindeberg-Feller CLT (See Greene, 2003 p. 913) that can be used to prove that the OLS estimator in the multiple regression model with fixed regressors converges to a normal random variable. For our purposes, we will use a different multivariate CLT that is applicable in the time series context. Details will be given in the section of time series concepts.

3.2 Asymptotic Normality

Definition 13 *Asymptotic normality*

A consistent estimator $\hat{\boldsymbol{\theta}}$ is *asymptotically normally distributed* (asymptotically normal) if

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(0, \boldsymbol{\Sigma}),$$

or

$$\hat{\boldsymbol{\theta}} \overset{A}{\sim} N(\boldsymbol{\theta}, n^{-1}\boldsymbol{\Sigma}).$$

■

3.3 Results for the Manipulation of CLTs

Theorem 14 *Slutsky's Theorem 2 (Extension of Slutsky's Theorem to convergence in distribution)*

Let Y_n and Z_n be sequences of random variables such that

$$Y_n \xrightarrow{d} W, Z_n \xrightarrow{p} c,$$

where W is a random variable and c is a constant. Then the following results hold as $n \to \infty$:

1. $Z_n Y_n \xrightarrow{d} cW$
2. $Y_n / Z_n \xrightarrow{d} W/c$ provided $c \neq 0$
3. $Y_n + Z_n \xrightarrow{d} W + c$

■

Remark

1. Suppose Y_n and Z_n are sequences of random variables such that

   $$Y_n \xrightarrow{d} W, Z_n \xrightarrow{d} Z,$$

   where W and Z are (possibly correlated) random variables. Then it is not necessarily true that

   $$Y_n + Z_n \xrightarrow{d} W + Z.$$

   We have to worry about the dependence between Y_n and Z_n.

Example 15 *Convergence in distribution of standardized sequence*

Suppose $X_1, \ldots, X_n$ are iid with $E[X_1] = \mu$ and $\text{var}(X_1) = \sigma^2$. In most practical situations, we don't know σ^2 and we must estimate it from the data. Let

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Earlier we showed that

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2 \text{ and } \hat{\sigma} \xrightarrow{p} \sigma.$$

Now consider

$$Y_n = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)\left(\frac{\sigma}{\hat{\sigma}}\right).$$

From Slutsky's Theorem and the Lindeberg-Levy CLT

$$\begin{aligned}\frac{1}{\hat{\sigma}} &\xrightarrow{p} \frac{1}{\sigma},\\ \frac{\sigma}{\hat{\sigma}} &\xrightarrow{p} \frac{\sigma}{\sigma} = 1,\\ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} &\xrightarrow{d} Z \sim N(0,1),\end{aligned}$$

so that

$$\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)\left(\frac{\sigma}{\hat{\sigma}}\right) \xrightarrow{d} Z \cdot 1.$$

Example 16 *Asymptotic normality of least squares estimator with fixed regressors*

Continuing with the linear regression example, the average variance of $w_i = x_i\varepsilon_i$ is

$$\bar{\sigma}_n^2 = n^{-1}\sum_{i=1}^{n}\text{var}(w_i) = \sigma^2 n^{-1}\sum_{i=1}^{n}x_i^2.$$

If we assume that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}x_i^2 \to S_{xx} > 0,$$

is finite then

$$\lim_{n\to\infty}\bar{\sigma}_n^2 = \bar{\sigma}^2 = \sigma^2 S_{xx} < \infty.$$

Further assume that

$$E[|\varepsilon_i|^{2+\delta}] < \infty.$$

Then it follows from Slutsky's Theorem 2 and the Liapounov CLT that

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i\\ &\xrightarrow{d} S_{xx}^{-1}\cdot N(0,\sigma^2 S_{xx}) \sim N(0,\sigma^2 S_{xx}^{-1}).\end{aligned}$$

Equivalently,

$$\hat{\beta} \overset{A}{\sim} N(\beta, \sigma^2 n^{-1} S_{xx}^{-1}),$$

and the asymptotic variance of $\hat{\beta}$ is

$$\text{avar}(\hat{\beta}) = \sigma^2 n^{-1} S_{xx}^{-1}.$$

Using

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\hat{\beta})^2 \xrightarrow{p} \sigma^2, \\ \frac{1}{n}\sum_{i=1}^{n} x_i^2 &= \frac{1}{n}\mathbf{x}'\mathbf{x} \to S_{xx}, \end{aligned}$$

gives the consistent estimate

$$\widehat{\text{avar}}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}.$$

Then, a practically useful asymptotic distribution is

$$\hat{\beta} \overset{A}{\sim} N(\beta, \hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}).$$

Theorem 17 *Continuous Mapping Theorem (CMT)*

Suppose $h(\cdot) : \mathbb{R} \to \mathbb{R}$ is continuous everywhere and $Y_n \xrightarrow{d} W$ as $n \to \infty$. Then

$$h(Y_n) \xrightarrow{d} h(W) \text{ as } n \to \infty$$

■

The CMT is typically used to derive the asymptotic distributions of test statistics; e.g., Wald, Lagrange multiplier (LM) and likelihood ratio (LR) statistics.

Example 18 *Asymptotic distribution of squared normalized mean*

Suppose $X_1, \ldots, X_n$ are iid with $E[X_1] = \mu$ and $\text{var}(X_1) = \sigma^2$. By the CLT we have that

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Z \sim N(0, 1)$$

Now set $h(x) = x^2$ and apply the CMT to give

$$h\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{d} h(Z) = Z^2 \sim \chi^2(1)$$

3.3.1 The Delta Method

Suppose we have an asymptotically normal estimator $\hat{\theta}$ for the scalar parameter θ; i.e.,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} W \sim N(0, \sigma^2).$$

Often we are interested in some function of θ, say $\eta = g(\theta)$. Suppose $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is continuous and differentiable at θ and that $g' = \frac{dg}{d\theta}$ is continuous. Then the *delta method* result is

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} W^* \sim N(0, g'(\theta)^2\sigma^2).$$

Equivalently,

$$g(\hat{\theta}) \overset{A}{\sim} N\left(g(\theta), \frac{g'(\theta)^2\sigma^2}{n}\right).$$

■

Example 19 *Asymptotic distribution of* $\bar{X}^{-1}$

Let $X_1, \ldots, X_n$ be iid with $E[X_1] = \mu$ and $\text{var}(X_1) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} Z \sim N(0, \sigma^2).$$

Let $\eta = g(\mu) = 1/\mu = \mu^{-1}$ for $\mu \neq 0$. Then

$$g'(\mu) = -\mu^{-2},\ g'(\mu)^2 = \mu^{-4}.$$

Then by the delta method

$$\sqrt{n}(\hat{\eta} - \eta) = \sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \xrightarrow{d} N(0, \mu^{-4}\sigma^2),$$
$$\frac{1}{\bar{X}} \overset{A}{\sim} N\left(\frac{1}{\mu}, \frac{1}{n}\frac{\sigma^2}{\mu^4}\right),$$

provided $\mu \neq 0$. The above result is not practically useful since μ and σ^2 are unknown. A practically useful result substitutes consistent estimates for unknown quantities:

$$\frac{1}{\bar{X}} \overset{A}{\sim} N\left(\frac{1}{\mu}, \frac{\hat{\sigma}^2}{n\hat{\mu}^4}\right),$$

where $\hat{\mu} \xrightarrow{p} \mu$ and $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. For example, we could use $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$. An asymptotically valid 95% confidence interval for $\frac{1}{\mu}$ has the form

$$\frac{1}{\bar{X}} \pm 1.96 \cdot \sqrt{\frac{\hat{\sigma}^2}{n\hat{\mu}^4}}.$$

Proof. The delta method gets its name from the use of a first-order Taylor series expansion. Consider a first-order Taylor series expansion of $g(\hat{\theta})$ at $\hat{\theta} = \theta$

$$\begin{aligned} g(\hat{\theta}) &= g(\theta) + g'(\tilde{\theta})(\hat{\theta} - \theta), \\ \tilde{\theta} &= \lambda\hat{\theta} + (1-\lambda)\theta,\ 0 \leq \lambda \leq 1. \end{aligned}$$

Multiplying both sides by $\sqrt{n}$ and re-arranging gives

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) = g'(\tilde{\theta})\sqrt{n}(\hat{\theta} - \theta).$$

Since $\tilde{\theta}$ is between $\hat{\theta}$ and θ and since $\hat{\theta} \xrightarrow{p} \theta$ we have that $\tilde{\theta} \xrightarrow{p} \theta$. Further since g' is continuous, by Slutsky's Theorem $g'(\tilde{\theta}) \xrightarrow{p} g'(\theta)$. It follows from the convergence in distribution results that

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} g'(\theta) \cdot N(0, \sigma^2) \sim N(0, g'(\theta)^2\sigma^2).$$

■ ■

Now suppose $\boldsymbol{\theta} \in \mathbb{R}^k$ and we have an asymptotically normal estimator

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \hat{\boldsymbol{\theta}} \overset{A}{\sim} N(\boldsymbol{\theta}, n^{-1}\boldsymbol{\Sigma}).$$

Let $\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\theta}) : \mathbb{R}^k \to \mathbb{R}^j$; i.e.,

$$\boldsymbol{\eta} = \mathbf{g}(\boldsymbol{\theta}) = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \\ \vdots \\ g_j(\boldsymbol{\theta}) \end{pmatrix},$$

denote the parameter of interest where $\boldsymbol{\eta} \in \mathbb{R}^j$ and $j \leq k$. Assume that $\mathbf{g}(\boldsymbol{\theta})$ is continuous with continuous first derivatives

$$\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_1(\boldsymbol{\theta})}{\partial \theta_k} \\ \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_2(\boldsymbol{\theta})}{\partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_j(\boldsymbol{\theta})}{\partial \theta_1} & \frac{\partial g_j(\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \frac{\partial g_j(\boldsymbol{\theta})}{\partial \theta_k} \end{pmatrix}.$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})) \xrightarrow{d} N\left(\mathbf{0}, \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) \boldsymbol{\Sigma} \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right)'\right).$$

If $\hat{\Sigma} \to \Sigma$ then a practically useful result is

$$g(\hat{\boldsymbol{\theta}}) \overset{A}{\sim} N\left(g(\boldsymbol{\theta}), n^{-1}\left(\frac{\partial \mathbf{g}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}\right) \hat{\boldsymbol{\Sigma}} \left(\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right)'\right).$$

Example 20 *Estimation of Generalized Learning Curve*

Consider the generalized learning curve (see Berndt, 1992, chapter 3)

$$C_t = C_1 N_t^{\alpha_c/R} Y_t^{(1-R)/R} \exp(u_t),$$

where C_t denotes real unit cost at time t, N_t denotes cumulative production up to time t, Y_t is production in time t, and u_t is an iid $(0, \sigma^2)$ error term. The parameter α_c is the elasticity of unit cost with respect to cumulative production or learning curve parameter. It is typically negative if there are learning curve effects. The parameter R is a returns to scale parameter such that: $R = 1$ gives constant returns to scale; $R < 1$ gives decreasing returns to scale; $R > 1$ gives increasing returns to scale. The intuition behind the model is as follows. Learning is proxied by cumulative production. If the learning curve effect is present, then as cumulative production (learning) increases real unit costs should fall. If production technology exhibits constant returns to scale, then real unit costs should not vary with the level of production. If returns to scale are increasing, then real unit costs should decline as the level of production increases.

The generalized learning curve may be converted to a linear regression model by taking logs:

$$\begin{aligned} \ln C_t &= \ln C_1 + \left(\frac{\alpha_c}{R}\right) \ln N_t + \left(\frac{1-R}{R}\right) \ln Y_t + u_t \\ &= \beta_0 + \beta_1 \ln N_t + \beta_2 \ln Y_t + u_t \\ &= \mathbf{x}_t' \boldsymbol{\beta} + u_t, \end{aligned} \tag{7}$$

where $\beta_0 = \ln C_1$, $\beta_1 = \alpha_c/R$, $\beta_2 = (1-R)/R$, and $\mathbf{x}_t = (1, \ln N_t, \ln Y_t)'$. The parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ may be estimated by least squares. Note that the learning curve parameters may be recovered using

$$\begin{aligned} \alpha_c &= \frac{\beta_1}{1+\beta_2} = g_1(\boldsymbol{\beta}), \\ R &= \frac{1}{1+\beta_2} = g_2(\boldsymbol{\beta}). \end{aligned}$$

Least squares applied to (7) gives the consistent estimates

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \xrightarrow{p} \boldsymbol{\beta}, \\ \hat{\sigma}^2 &= n^{-1}\sum_{t=1}^{n}(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2 \xrightarrow{p} \sigma^2, \end{aligned}$$

and

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Then from Slutsky's Theorem

$$\begin{aligned}\hat{\alpha}_c &= \frac{\hat{\beta}_1}{1+\hat{\beta}_2} \xrightarrow{p} \frac{\beta_1}{1+\beta_2} = \alpha_c, \\ \hat{R} &= \frac{1}{1+\hat{\beta}_2} \xrightarrow{p} \frac{1}{1+\beta_2} = R,\end{aligned}$$

provided $\beta_2 \neq -1$. We can use the delta method to get the asymptotic distribution of $\hat{\eta} = (\hat{\alpha}_c, \hat{R})'$:

$$\begin{pmatrix} \hat{\alpha}_c \\ \hat{R} \end{pmatrix} = \begin{pmatrix} g_1(\hat{\boldsymbol{\beta}}) \\ g_2(\hat{\boldsymbol{\beta}}) \end{pmatrix} \overset{A}{\sim} N\left(\mathbf{g}(\boldsymbol{\beta}), \left(\frac{\partial \mathbf{g}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'}\right) \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1} \left(\frac{\partial \mathbf{g}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}'}\right)'\right),$$

where

$$\begin{aligned}\frac{\partial \mathbf{g}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} &= \begin{pmatrix} \frac{\partial g_1(\boldsymbol{\beta})}{\partial \beta_0} & \frac{\partial g_1(\boldsymbol{\beta})}{\partial \beta_1} & \frac{\partial g_1(\boldsymbol{\beta})}{\partial \beta_2} \\ \frac{\partial g_2(\boldsymbol{\beta})}{\partial \beta_0} & \frac{\partial g_2(\boldsymbol{\beta})}{\partial \beta_1} & \frac{\partial g_2(\boldsymbol{\beta})}{\partial \beta_2} \end{pmatrix} \\ &= \begin{pmatrix} 0 & \frac{1}{1+\beta_2} & \frac{-\beta_1}{(1+\beta_2)^2} \\ 0 & 0 & \frac{-1}{(1+\beta_2)^2} \end{pmatrix}.\end{aligned}$$

4 Time Series Concepts

Definition 21 *Stochastic process*

A *stochastic process* $\{Y_t\}_{t=1}^{\infty}$ is a sequence of random variables indexed by time t

■

A realization of a stochastic process is the sequence of observed data $\{y_t\}_{t=1}^{\infty}$. We are interested in the conditions under which we can treat the stochastic process like a random sample, as the sample size goes to infinity. Under such conditions, at any point in time t_0, the *ensemble average*

$$\frac{1}{N}\sum_{k=1}^{N} Y_{t_0}^{(k)},$$

will converge to the sample *time average*

$$\frac{1}{T}\sum_{t=1}^{T} Y_t,$$

as N and T go to infinity. If this result occurs then the stochastic process is called *ergodic*.

4.1 Stationary Stochastic Processes

We start with the definition of strict stationarity.

Definition 22 *Strict stationarity*

A stochastic process $\{Y_t\}_{t=1}^{\infty}$ is *strictly stationary* if, for any given finite integer r and for any set of subscripts $t_1, t_2, \ldots, t_r$ the joint distribution of $(Y_t, Y_{t_1}, Y_{t_2}, \ldots, Y_{t_r})$ depends only on $t_1 - t, t_1 - t, \ldots, t_r - t$ but not on t.

■

Remarks

1. For example, the distribution of (Y_1, Y_5) is the same as the distribution of (Y_{12}, Y_{16}).
2. For a strictly stationary process, Y_t has the same mean, variance (moments) for all t.
3. Any transformation $g(\cdot)$ of a strictly stationary process, $\{g(Y_t)\}$ is also strictly stationary.

Example 23 *iid sequence*

If $\{Y_t\}$ is an iid sequence, then it is strictly stationary. Let $\{Y_t\}$ be an iid sequence and let $X \sim N(0,1)$ independent of $\{Y_t\}$. Let $Z_t = Y_t + X$. Then the sequence $\{Z_t\}$ is strictly stationary.

Definition 24 *Covariance (Weak) stationarity*

A stochastic process $\{Y_t\}_{t=1}^{\infty}$ is *covariance stationary* (weakly stationary) if

1. $E[Y_t] = \mu$ does not depend on t
2. $\text{cov}(Y_t.Y_{t-j}) = \gamma_j$ exists, is finite, and depends only on j but not on t for $j = 0, 1, 2, \ldots$

■

Remark:

1. A strictly stationary process is covariance stationary if the mean and variance exist and the covariances are finite.

For a weakly stationary process $\{Y_t\}_{t=1}^{\infty}$ define the following:

$$\begin{aligned} \gamma_j &= \text{cov}(Y_t, Y_{t-j}) = \text{ j}^{th} \text{ order autocovariance} \\ \gamma_0 &= \text{var}(Y_t) = \text{ variance} \\ \rho_j &= \gamma_j/\gamma_0 = \text{ j}^{th} \text{ order autocorrelation} \end{aligned}$$

Definition 25 *Ergodicity*

Loosely speaking, a stochastic process $\{Y_t\}_{t=1}^{\infty}$ is *ergodic* if any two collections of random variables partitioned far apart in the sequence are almost independently distributed. The formal definition of ergodicity is highly technical (see Hayashi 2000, p. 101 and note typo from errata).

■

Proposition 26 *Hamilton (1994) page 47.*

Let $\{Y_t\}$ be a covariance stationary process with mean $E[Y_t] = \mu$ and autocovariances $\gamma_j = \text{cov}(Y_t, Y_{t-j})$. If

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty,$$

then $\{Y_t\}$ is *ergodic for the mean*. That is, $\bar{Y} \xrightarrow{p} E[Y_t] = \mu$.■

Example 27 *MA(1)*

Let

$$\begin{aligned} Y_t &= \mu + \varepsilon_t + \theta\varepsilon_{t-1},\ |\theta| < 1 \\ &\quad \varepsilon_t \sim \text{iid } (0, \sigma^2) \end{aligned}$$

Then

$$\begin{aligned} E[Y_t] &= \mu \\ \gamma_0 &= E[(Y_t - \mu)^2] = \sigma^2(1 + \theta^2) \\ \gamma_1 &= E[(Y_t - \mu)(Y_{t-1} - \mu)] = \sigma^2\theta \\ \gamma_k &= 0,\ k > 1. \end{aligned}$$

Clearly,

$$\sum_{j=0}^{\infty} |\gamma_j| = \sigma^2(1 + \theta^2) + \sigma^2|\theta| < \infty,$$

so that $\{Y_t\}$ is ergodic.

Theorem 28 *Ergodic Theorem*

Let $\{Y_t\}$ be stationary and ergodic with $E[Y_i] = \mu$. Then

$$\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t \xrightarrow{p} E[Y_t] = \mu.$$

■

Remarks

1. The ergodic theorem says that for a stationary and ergodic sequence $\{Y_t\}$ the time average converges to the ensemble average as the sample size gets large. That is, the ergodic theorem is a LLN.

2. The ergodic theorem is a substantial generalization of Kolmogorov's LLN because it allows for serial dependence in the time series.

3. Any transformation $g(\cdot)$ of a stationary and ergodic process $\{Y_t\}$ is also stationary and ergodic. That is, $\{g(Y_t)\}$ is stationary and ergodic. Therefore, if $E[g(Y_t)]$ exists then the ergodic theorem gives

$$\bar{g} = \frac{1}{T}\sum_{t=1}^{T} g(Y_t) \xrightarrow{p} E[g(Y_t)].$$

This is a very useful result. For example, we may use it to prove that the sample autocovariances

$$\gamma_j = \frac{1}{T}\sum_{t=j+1}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}),$$

converge in probability to the population autocovariances $\gamma_j = E[(Y_t - \mu)(Y_{t-j} - \mu)] = \text{cov}(Y_t, Y_{t-j})$.

Example 29 *Stationary but not ergodic process (White, 1984)*

Let $\{Y_t\}$ be an iid sequence with $E[Y_t] = \mu$, $\text{var}(Y_t) = \sigma^2$ and let $X \sim N(0,1)$ independent of $\{Y_t\}$. Let $Z_t = Y_t + X$. Note that $E[Z_t] = \mu$. Z_t is stationary but not ergodic. To see this, note that

$$\text{cov}(Z_t, Z_{t-j}) = \text{cov}(Y_t + X, Y_{t-j} + X) = \text{var}(X) = 1,$$

so that $\text{cov}(Z_t, Z_{t-j}) \nrightarrow 0$ as $j \to \infty$. Now consider the sample average of Z_t

$$\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t = \frac{1}{T}\sum_{t=1}^{T}(Y_t + X) = \bar{Y} + X.$$

By Chebychev's LLN $\bar{Y} \xrightarrow{p} \mu$ and so

$$\bar{Z} \xrightarrow{p} \mu + X \neq E[Z_t] = \mu.$$

Because $\text{cov}(Z_t, Z_{t-j}) \nrightarrow 0$ as $j \to \infty$, the sample average does not converge to the population average.

4.2 Martingales and Martingale Difference Sequences

Let $\{Y_t\}$ be a sequence of random variables and let $\{I_t\}$ be a sequence of information sets ($\sigma-$fields) with $I_t \subset I$ for all t and I the universal information set. For example,

$$\begin{aligned} I_t &= \{Y_1, Y_2, \ldots, Y_t\} = \text{ past history of } Y_t \\ I_t &= \{(Y_s, Z_s)_{s=1}^t\},\ \{Z_t\} = \text{ auxiliary variables} \end{aligned}$$

Definition 30 *Conditional Expectation*

Let Y_t be a random variable with conditional pdf $f(y_t|I_s)$, where I_s is an information set with $s < t$. Then

$$E[Y_t|I_s] = \int_{-\infty}^{\infty} y_t f(y_t|I_s) dy_t.$$

■

Proposition 31 *Law of Iterated Expectation*

Let I_1 and I_2 be information sets such that $I_1 \subseteq I_2$, and let Y be a random variable such that $E[Y|I_1]$ and $E[Y|I_2]$ are defined. Then

$$E[Y|I_1] = E[E[Y|I_2]|I_1] \text{ (smaller set wins).}$$

If $I_1 = \varnothing$ (empty set) then

$$\begin{aligned} E[Y|I_1] &= E[Y] \text{ (unconditional expectation)} \\ E[Y] &= E[E[Y|I_2]|\varnothing] = E[E[Y|I_2]]. \end{aligned}$$

■

Definition 32 *Martingale*

The pair (Y_t, I_t) is a *martingale* (MG) if

1. $I_t \subset I_{t+1}$ (increasing sequence of information sets - a *filtration*)
2. $Y_t \subset I_t$ (Y_t is *adapted* to I_t; i.e., Y_t is an event in I_t)
3. $E[|Y_t|] < \infty$
4. $E[Y_t|I_{t-1}] = Y_{t-1}$ (MG property)

■

Example 33 *Random walk*

Let $Y_t = Y_{t-1} + u_t$ where $\{u_t\}$ is an iid sequence with mean zero and variance σ^2. Let $I_t = \{Y_1, Y_2, \ldots, Y_1\}$. Then

$$E[Y_t|I_{t-1}] = Y_{t-1}.$$

Example 34 *Heteroskedastic random walk*

Let $Y_t = Y_{t-1} + u_t/t = Y_{t-1} + v_t$ where $\{u_t\}$ is an iid sequence with mean zero and variance σ^2 and $v_t = u_t/t$. Note that $\text{var}(v_t) = \sigma^2/t$. Let $I_t = \{Y_1, Y_2, \ldots, Y_1\}$. Then

$$E[Y_t|I_{t-1}] = Y_{t-1}.$$

If (Y_t, I_t) is a MG, then

$$E[Y_{t+m}|I_t] = Y_t \text{ for all } t \geq 1.$$

To see this, let $m = 2$ and note that by iterated expectations

$$E[Y_{t+2}|I_t] = E[E[Y_{t+2}|I_{t+1}]|I_t].$$

By the MG property

$$E[Y_{t+2}|I_{t+1}] = Y_{t+1},$$

which leave us with

$$E[Y_{t+2}|I_t] = E[Y_{t+1}|I_t] = Y_t.$$

Definition 35 *Martingale Difference Sequence (MDS)*

The pair (Y_t, I_t) is a *martingale difference sequence* (MDS) if (Y_t, I_t) is an adapted sequence and

$$E[Y_t|I_{t-1}] = 0.$$

Remarks

1. If (Y_t, I_t) is a MG and we define

   $$u_t = Y_t - E[Y_t|I_{t-1}],$$

   we have, by virtual construction,

   $$E[u_t|I_{t-1}] = 0,$$

   so that (u_t, I_t) is a MDS.

2. The sequence $\{u_t\}$ is sometime referred to as a sequence of nonlinear innovations. The term arises because if Z_t is any function of the past history of Y_t, and thus $Z_t \subset I_t$, we have by iterated expectations

   $$\begin{aligned} E[u_t Z_{t-1}] &= E[E[u_t Z_{t-1}|I_{t-1}]] \\ &= E[Z_{t-1}E[u_t|I_{t-1}]] \\ &= 0, \end{aligned}$$

   so that u_t is orthogonal to any function of the past history of Y_t.

3. If (u_t, I_t) is a MDS then

   $$E[u_{t+m}|I_t] = 0.$$

Example 36 *ARCH process*

Consider the first order *autoregressive conditional heteroskedasticity* (ARCH(1)) process

$$\begin{aligned} u_t &= Z_t\sigma_t \\ & \quad Z_t \sim \text{iid } N(0,1) \\ \sigma_t^2 &= \omega + \alpha u_{t-1}^2,\ 0 < \alpha < 1,\ \omega > 0 \end{aligned}$$

The process was proposed by Nobel Laureate Robert Engle to describe and predict time varying volatility in macroeconomic and financial time series. If $I_t = \{u_t, u_{t-1}, \ldots, u_1\}$ then (u_t, I_t) is a stationary and ergodic conditionally heteroskedastic MDS. The unconditional moments of u_t are:

$$\begin{aligned} E[u_t] &= E[E[Z_t\sigma_t|I_{t-1}]] = E[\sigma_t E[Z_t|I_{t-1}] = 0, \\ \text{var}(u_t) &= E[u_t^2] = E[E[Z_t^2\sigma_t^2|I_{t-1}]] = E[\sigma_t^2 E[Z_t^2|I_{t-1}]] = E[\sigma_t^2]. \end{aligned}$$

Furthermore,

$$\begin{aligned} E[\sigma_t^2] &= E[\omega + \alpha u_{t-1}^2] \\ &= \omega + \alpha E[u_{t-1}^2] = \omega + E[\sigma_{t-1}^2] \\ &= \omega + E[\sigma_t^2] \text{ (assuming stationarity)} \\ \Longrightarrow \quad & E[\sigma_t^2] = \frac{\omega}{1-\alpha} > 0. \end{aligned}$$

Next, for $k \geq 1$

$$E[u_t u_{t-k}] = E[E[u_t u_{t-k}|I_{t-1}]] = 0.$$

Finally,

$$\begin{aligned} E[u_t^4] &= E[E[Z_t^4 \sigma_t^4 | I_{t-1}]] = E[\sigma_t^4 E[Z_t^4 | I_{t-1}]] \\ &= 3 \cdot E[\sigma_t^4] \geq 3 \cdot E[\sigma_t^2]^2 = 3 \cdot E[u_t^2]^2 \\ \Longrightarrow \quad & \frac{E[u_t^4]}{E[u_t^2]^2} \geq 3. \end{aligned}$$

The inequality in the second line above comes from Jensen's inequality.

The conditional moments of u_t are

$$\begin{aligned} E[u_t|I_{t-1}] &= 0, \\ E[u_t^2|I_{t-1}] &= \sigma_t^2. \end{aligned}$$

Interestingly, even though u_t is serially uncorrelated it is clearly not an independent process. In fact, u_t^2 has an AR(1) representation. To see this, add u_t^2 to both sides of the expression for σ_t^2 to give

$$\begin{aligned} \sigma_t^2 + u_t^2 &= \omega + \alpha u_{t-1}^2 + u_t^2 \\ \Longrightarrow \quad & u_t^2 = \omega + \alpha u_{t-1}^2 + v_t, \end{aligned}$$

where $v_t = u_t^2 - \sigma_t^2$ is MDS.

Theorem 37 *Multivariate CLT for stationary and ergodic MDS (Billingsley, 1961)*

Let $(\mathbf{u}_t, I_t)$ be a vector MDS that is stationary and ergodic with $k \times k$ covariance matrix $E[\mathbf{u}_t \mathbf{u}_t'] = \mathbf{\Sigma}$. Let

$$\bar{\mathbf{u}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{u}_t.$$

Then

$$\sqrt{T}\bar{\mathbf{u}} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \mathbf{u}_t \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma})$$

■

4.3 Large Sample Distribution of Least Squares Estimator

As an application of the previous results, consider estimation and inference for the linear regression model

$$y_t = \underset{(1\times k)(k\times 1)}{\mathbf{x}_t'\boldsymbol{\beta}} + \varepsilon_t,\ t = 1, \ldots, T, \tag{8}$$

under the following assumptions:
Assumption 1 (Linear regression model with stochastic regressors)

1. $\{\mathbf{x}_t, \varepsilon_t\}$ is jointly stationary and ergodic
2. $E[\mathbf{x}_t\mathbf{x}_t'] = \boldsymbol{\Sigma}_{xx}$ is positive definite (full rank k)
3. $E[x_{it}\varepsilon_t] = 0$ for all t, k
4. The process $\{\mathbf{g}_t\} = \{\mathbf{x}_t\varepsilon_t\}$ is a MDS with $E[\mathbf{g}_t\mathbf{g}_t'] = E[\mathbf{x}_t\mathbf{x}_t'\varepsilon_t^2] = \mathbf{S}$ nonsingular.

Note: Part 1 implies that ε_t is stationary so that $E[\varepsilon_t^2] = \sigma^2$ is the unconditional variance. However, Part 4 allows for general conditional heteroskedasticity; e.g. $\text{var}(\varepsilon_t|\mathbf{x}_t) = f(\mathbf{x}_t)$. In this case,

$$\begin{aligned} E[\mathbf{x}_t\mathbf{x}_t'\varepsilon_t^2] &= E[E[\mathbf{x}_t\mathbf{x}_t'\varepsilon_t^2|\mathbf{x}_t]] \\ &= E[\mathbf{x}_t\mathbf{x}_t'E[\varepsilon_t^2|\mathbf{x}_t]] = E[\mathbf{x}_t\mathbf{x}_t'f(\mathbf{x}_t)] = \mathbf{S}. \end{aligned}$$

If the errors are conditionally homoskedastic then $\text{var}(\varepsilon_t|x_t) = \sigma^2$ and

$$\begin{aligned} E[\mathbf{x}_t\mathbf{x}_t'\varepsilon_t^2] &= E[\mathbf{x}_t\mathbf{x}_t'E[\varepsilon_t^2|\mathbf{x}_t]] \\ &= \sigma^2 E[\mathbf{x}_t\mathbf{x}_t'] = \sigma^2\boldsymbol{\Sigma}_{xx} = \mathbf{S}. \end{aligned}$$

The least squares estimator of $\boldsymbol{\beta}$ in (8) is

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\sum_{t=1}^T \mathbf{x}_t\mathbf{y}_t \\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\sum_{t=1}^T \mathbf{x}_t\varepsilon_t. \end{aligned} \tag{9}$$

Proposition 38 *Consistency and asymptotic normality of the least squares estimator*

Under Assumption 1, as $T \to \infty$

1. $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$

2. $\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1})$, $\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, T^{-1}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{xx}^{-1})$ where $\hat{\boldsymbol{\Sigma}}_{xx} \xrightarrow{p} \boldsymbol{\Sigma}_{xx}$and $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$

Proof. For part 1, first write (9) as

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t$$

Since $\{x_t\}$ is stationary and ergodic, by the ergodic theorem

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} E[\mathbf{x}_t\mathbf{x}_t'] = \Sigma_{xx}$$

and by Slutsky's theorem

$$\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}$$

Similarly, since $\{\mathbf{g}_t\} = \{\mathbf{x}_t\varepsilon_t\}$ is stationary and ergodic by the ergodic theorem

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t = \frac{1}{T}\sum_{t=1}^{T}\mathbf{g}_t \xrightarrow{p} E[\mathbf{g}_t] = E[\mathbf{x}_t\varepsilon_t] = \mathbf{0}$$

As a result,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t \xrightarrow{p} \Sigma_{xx}^{-1}\cdot\mathbf{0} = \mathbf{0}$$

so that

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

For part 2, write (9) as

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t$$

Previously, we deduced $\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1} \xrightarrow{p} \boldsymbol{\Sigma}_{xx}^{-1}$. Next, since $\{\mathbf{g}_t\} = \{\mathbf{x}_t\varepsilon_t\}$ is a stationary and ergodic MDS with $E[\mathbf{g}_t\mathbf{g}_t'] = \mathbf{S}$ by the MDS CLT we have

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{g}_t \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$$

Therefore

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t \xrightarrow{d} \boldsymbol{\Sigma}_{xx}^{-1}\cdot N(\mathbf{0},\mathbf{S}) \\ &\sim N(\mathbf{0},\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1})\end{aligned}$$

From the above result, we see that the asymptotic variance of β is given by

$$\text{avar}(\hat{\boldsymbol{\beta}}) = T^{-1}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}$$

Equivalently,

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, T^{-1}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{xx}^{-1})$$

where $\hat{\boldsymbol{\Sigma}}_{xx} \xrightarrow{p} \boldsymbol{\Sigma}_{xx}$and $\hat{\mathbf{S}} \xrightarrow{p} \mathbf{S}$ and

$$\widehat{\text{avar}}(\hat{\boldsymbol{\beta}}) = T^{-1}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\mathbf{S}}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}$$

Remark

1. If the errors are conditionally homoskedastic, then $var(\varepsilon_t^2|x_t) = \sigma^2$, $\mathbf{S} = \sigma^2\boldsymbol{\Sigma}_{xx}$ and $\text{avar}(\hat{\boldsymbol{\beta}})$ simplifies to

   $$\text{avar}(\hat{\boldsymbol{\beta}}) = T^{-1}\sigma^2\boldsymbol{\Sigma}_{xx}^{-1}$$

   Using $\mathbf{S}_{xx} = T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = T^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \boldsymbol{\Sigma}_{xx}$ and $\hat{\sigma}^2 = \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2 \xrightarrow{p} \sigma^2$ then gives

   $$\widehat{\text{avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Proposition 39 *Consistent estimation of* $\mathbf{S} = \mathbf{E}[\varepsilon_t^2\mathbf{x}_t\mathbf{x}_t']$

Assume $E[(x_{ik}x_{ij})^2]$ exists and is finite for all k, j $(i = 1, 2, \ldots, k)$ Then as $T \to \infty$

$$\hat{\mathbf{S}} = T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t^2\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{S}$$

where $\hat{\varepsilon}_t = y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}$.

Sketch of Proof. Consider the simple case in which $k = 1$. Then it is assumed that $E[x_t^4] < \infty$. Write

$$\begin{aligned}\hat{\varepsilon}_t &= y_t - x_t\hat{\beta} = y_t - x_t\beta + x_t\beta - x_t\hat{\beta} \\ &= \varepsilon_t - x_t(\hat{\beta}-\beta)\end{aligned}$$

so that

$$\begin{aligned}\hat{\varepsilon}_t^2 &= \Big(\varepsilon_t - x_t(\hat{\beta}-\beta)\Big)^2 \\ &= \varepsilon_t^2 - 2x_t\varepsilon_t(\hat{\beta}-\beta) + x_t^2(\hat{\beta}-\beta)^2\end{aligned}$$

Then

$$\begin{aligned}\hat{S} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 x_t^2 = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 x_t^2 - 2(\hat{\beta}-\beta)\frac{1}{T}\sum_{t=1}^{T}x_t^3\varepsilon_t + (\hat{\beta}-\beta)^2\frac{1}{T}\sum_{t=1}^{T}x_t^4 \\ &= \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 x_t^2 + o_p(1) \xrightarrow{p} E[\varepsilon_t^2 x_t^2] = S\end{aligned}$$

In the above, the following results are used

$$\hat{\beta}-\beta \xrightarrow{p} 0$$

$$\frac{1}{T}\sum_{t=1}^{T}x_t^3\varepsilon_t \xrightarrow{p} E[x_t^3\varepsilon_t] < \infty$$

$$\frac{1}{T}\sum_{t=1}^{T}x_t^4 \xrightarrow{p} E[x_t^4] < \infty$$

The third line follows from the ergodic theorem and the assumption that $E[x_t^4] < \infty$. The second line follows from the ergodic theorem and the Cauchy-Schwarz inequality (see Hayashi 2000, analytic exercise 4, p. 169)

$$E[|f \cdot h|] \leq \left(E[f^2]E[h^2]\right)^{1/2}$$

with $f = x_t\varepsilon_t$ and $h_t = x_t^2$ so that

$$E[|x_t^3\varepsilon_t|] \leq \left(E[x_t^2\varepsilon_t^2]E[x_t^4]\right)^{1/2} < \infty$$

Using

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{xx} &= T^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = T^{-1}\mathbf{X}'\mathbf{X} \\ \hat{\mathbf{S}} &= T^{-1}\sum_{t=1}^{T}\hat{\varepsilon}_t^2\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{S}\end{aligned}$$

it follows that

$$\hat{\boldsymbol{\beta}} \overset{A}{\sim} N(\boldsymbol{\beta}, T\cdot(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{S}}(\mathbf{X}'\mathbf{X})^{-1})$$

so that

$$\widehat{\text{avar}}(\hat{\boldsymbol{\beta}}) = T\cdot(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{S}}(\mathbf{X}'\mathbf{X})^{-1} \tag{10}$$

The estimator (10) is often referred to as the "White" or heteroskedasticity consistent (HC) estimator. The square root of the diagonal elements of (10) are known as the White or HC standard errors for $\hat{\beta}_i$:

$$\widehat{\text{SE}}_{\text{HC}}(\hat{\beta}_i) = \sqrt{\left[T\cdot(\mathbf{X}'\mathbf{X})^{-1}\hat{\mathbf{S}}(\mathbf{X}'\mathbf{X})^{-1}\right]_{ii}},\ i = 1,\ldots,k$$

Remark:

1. Davidson and MacKinnon (1993) highly recommend using the following degrees of freedom corrected estimate for $\mathbf{S}$

$$\hat{\mathbf{S}} = (T-k)^{-1} \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \mathbf{x}_t \mathbf{x}_t' \xrightarrow{p} \mathbf{S}$$

   They show that the HC standard errors based on this estimator have better finite sample properties than the HC standard errors based on $\hat{\mathbf{S}}$ that doesn't use a degrees of freedom correction.

References

[1] Davidson, R. and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. Oxford University Press, Oxford.

[2] Greene, W. (2003). *Econometric Analysis, Fifth Edition*. Prentice Hall, Upper Saddle River.

[3] Hayashi, F. (2000). *Econometrics*. Princeton University Press, Princeton.

[4] White, H. (1984). *Asymptotic Theory for Econometricians*. Academic Press, San Diego.