

Chapter 5

STOCHASTIC REGRESSORS IN LINEAR MODELS

This chapter introduces the conditional Gauss-Markov Theorem, asymptotic theory, Monte Carlo, and Bootstrap as tools to evaluate estimators and tests. These tools are illustrated in the form that is convenient for most applications of structural econometrics for linear time series models in this chapter although they will be useful for nonlinear models as explained in later chapters.

In most applications in macroeconomics, regressors are stochastic, and the Gauss Markov Theorem for nonstochastic regressors do not apply. It is still possible to use the conditional Gauss Markov Theorem in some applications if a strict version of the exogeneity assumption (which will be called the strict exogeneity assumption) can be made to show that the OLS estimator is unbiased and efficient conditional on the realization of the regressors. If a normality assumption is added, it the estimator's exact small sample distributions can be obtained.

In some applications such as those of dynamic cointegrating regression explained in Chapter 14, the strict exogeneity assumption is typically made. So the conditional Gauss Markov Theorem can be used. However, in many other time series applications,

the strict exogeneity assumption is not attractive. If lagged dependent variables are included in regressors, the assumption cannot be made because it causes logical inconsistency. If the strict exogeneity assumption does not apply, then estimators are biased.

In rational expectations models, stringent distributional assumptions, such as an assumption that the disturbances are normally distributed, are unattractive. Without such assumptions, however, it is not possible to obtain the exact distributions of estimators in finite samples. For this reason, asymptotic theory describes the properties of estimators as the sample size goes to infinity.

Many researchers use asymptotic theory at initial stages of an empirical research project. Given the difficulties of obtaining the exact small sample distributions of estimators in many applications, this utilization seems to be a sound strategy. If the sample size is “large”, then asymptotic theory must be a good approximation of the true properties of estimators. The problem is that no one knows how large the sample size should be, because the answer depends on the nature of each application. After the importance of a research project is established, small sample properties of the estimators used in the project are often studied. For this purpose, Monte Carlo experiments can be used.

When asymptotic theory gives poor approximations in small sample, Bootstrap methods can be very useful. Bootstrap methods often give more accurate approximations of the exact small sample properties than asymptotic theory in applications to cross sectional data. In time series applications, there are some difficult issues that Bootstrap methods can have. This chapter explains such a difficulty that applied researchers should be aware of.

5.1 The Conditional Gauss Markov Theorem

In regressions (5.4) and (5.7), \mathbf{X}_t is *strictly exogenous in the time series sense* if $E(e_t | \dots, \mathbf{X}_{t+2}, \mathbf{X}_{t+1}, \mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots) = 0$. This is a very restrictive assumption that does not hold in all applications of cointegration discussed in Chapter 13. For example, $E(e_t | \mathbf{X}_t, \mathbf{X}_{t-1}, \mathbf{X}_{t-2}, \dots) = 0$ in some applications because e_t is a forecast error. However, the forecast error is usually correlated with future values of \mathbf{X}_t . Hence the strict exogeneity assumption is violated. Nevertheless, as Choi and Ogaki (1999) argue, it is useful to observe that the Gauss Markov theorem applies to cointegrating regressions in order to understand small sample properties of various estimators for cointegrating vectors. Moreover, this observation leads to a Generalized Least Squares (GLS) correction to spurious regressions.

Let $\sigma(\mathbf{X})$ be the smallest σ -field with respect to which the random variables in \mathbf{X} are measurable. We use the notation $E[Z | \sigma(\mathbf{X})]$ to denote the usual conditional expectation of Z conditional on \mathbf{X} as defined by Billingsley (1986) for a random variable Z . $E[Z | \sigma(\mathbf{X})]$ is a random variable, and $E[Z | \sigma(\mathbf{X})](s)$ denotes the value of the random variable at s in S (what is s ?). It should be noted that the definition is given under the condition that Z is integrable, namely $E(|Z|) < \infty$.

This condition can be too restrictive when we define the conditional expectation of the OLS estimator in some applications as we discuss later. ¹

For this reason, we will also use a different concept of expectation conditional on \mathbf{X} that can be used when Z and $vec(\mathbf{X})$ have probability density functions $f_Z(z)$

¹Loeve (1978) slightly relaxes this restriction by defining the conditional expectation for any random variable whose expectation exists (but may not be finite) with an extension of the Radon-Nikodym theorem. This definition can be used for $E(\cdot | \sigma(X))$, but this slight relaxation does not solve our problem which we describe later.

Masao
needs to
check this!

Masao
needs to
check this!

Masao
needs to
check this!

Masao
needs to
check this!

and $f_X(\text{vec}(\mathbf{x}))$, respectively. In this case, if $f_X(\text{vec}(\mathbf{x}))$ is positive, we define the expectation of Z conditional on $\mathbf{X}(s) = \mathbf{x}$ as

$$(5.1) \quad E[Z|\mathbf{X}(s) = \mathbf{x}] = \int_{-\infty}^{\infty} \frac{f_Z(z)}{f_X(\text{vec}(\mathbf{x}))} dz.$$

For this definition, we use the notation $E[Z|\mathbf{X}(s) = \mathbf{x}]$. This definition can only be used when the probability density functions exist and $f_X(\text{vec}(\mathbf{x}))$ is positive, but the advantage of this definition for our purpose is that the conditional expectation can be defined even when $E(Z)$ does not exist. For example let $Z = \frac{Y}{X}$ where Y and X are independent random variables with a standard normal distribution. Then Z has the Cauchy distribution, and $E(Z)$ does not exist. Thus, $E[Z|\sigma(X)]$ cannot be defined.² However, we can define $E[Z|X(s) = x]$ for all s in the probability space because the density function of X is always positive.

In the special case in which both types of conditional expectations can be defined, they coincide. More precisely, suppose that Z and $\text{vec}(\mathbf{X})$ have probability density functions, that the probability density function of $\text{vec}(\mathbf{X})$ is always positive, and that Z is integrable. Then $E[Z|\sigma(\mathbf{X})](s) = E[Z|\mathbf{X}(s)]$ with probability one.

Let $\mathbf{y} = (y_1, y_2, \dots, y_T)'$ be a $T \times 1$ vector of random variables, and $\mathbf{e} = (e_1, e_2, \dots, e_T)'$ be a $T \times 1$ vector of random variables. We are concerned with a linear model of the form:

Assumption 5.1 $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$,

where \mathbf{b}_0 is a $K \times 1$ vector of real numbers. We assume that the expectation of \mathbf{e} conditional on \mathbf{X} is zero:

²It should be noted that we cannot argue that $E(Z) = E(E(\frac{Y}{X}|\sigma(X))) = E(\frac{E(Y|\sigma(X))}{X}) = 0$ even though $\frac{1}{X}$ is measurable in $\sigma(X)$ because $E(\frac{Y}{X}|\sigma(X))$ is not defined.

Assumption 5.2 $E[\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{0}$.

Since $E[\mathbf{e}|\sigma(\mathbf{X})]$ is only defined when each element of \mathbf{e} is integrable, Assumption 5.2 implicitly assumes that $E(\mathbf{e})$ exists and is finite. It also implies $E(\mathbf{e}) = \mathbf{0}$ because of the law of iterated expectations. Given $E(\mathbf{e}) = \mathbf{0}$, a sufficient condition for Assumption 5.2 is that \mathbf{X} is statistically independent of \mathbf{e} . Since Assumption 5.2 does not imply that \mathbf{X} is statistically independent of \mathbf{e} , Assumption 5.2 is weaker than the assumption of the independent stochastic regressors. With the next assumption, we assume that \mathbf{e} is conditionally homoskedastic and e_t is not serially correlated:

Assumption 5.3 $E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})] = \sigma^2\mathbf{I}_T$.

Let $G = \{s \text{ in } S : \mathbf{X}(s)'\mathbf{X}(s) \text{ is nonsingular}\}$. Since the determinant of a matrix is a continuous function of the elements of a matrix, G is a member of the σ -field $\mathcal{F}?????$.

For any s in G , the OLS estimator is

$$(5.2) \quad \mathbf{b}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

From Assumption 5.1, $\mathbf{b}_T = \mathbf{b}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$. Hence the conditional Gauss-Markov theorem can be proved when the expectation of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ can be defined. For this purpose, we consider the following two alternative assumptions:

Assumption 5.4 $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]$ exists and is finite.

Assumption 5.4' \mathbf{e} and $\text{vec}(\mathbf{X})$ have probability density functions, and the probability density functions of $\text{vec}(\mathbf{X})$ are positive for all s in G .

Masao
needs to
check this!

A sufficient condition for Assumption 5.4 is that the distributions of \mathbf{X} and \mathbf{e} have finite supports. Under Assumption 5.4, $E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}]$ also exists and is finite. Hence $E(\mathbf{b}_T|\sigma(\mathbf{X}))$ can be defined. From Assumptions 5.1-5.3, $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0 + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{b}_0$ for s in G with probability $Pr(G)$. Under Assumptions 5.1-5.4, $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})]$ can be defined, and $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\sigma(\mathbf{X})] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ for s in G with probability $Pr(G)$. The problem with Assumption 5.4 is that it is not easy to verify Assumption 5.4 for many distributions of \mathbf{X} and \mathbf{e}_t that are often used in applications and Monte Carlo studies.

Under Assumptions 5.1-5.3 and 5.4', $E[\mathbf{b}_T|\mathbf{X}(s)] = \mathbf{b}_0$ and $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\mathbf{X}(s)] = \sigma^2(\mathbf{X}(s)'\mathbf{X}(s))^{-1}$ for any s in G .

Corresponding with Assumption 5.4 and 5.4', we consider two definitions of the conditional version of the Best Linear Unbiased Estimator (BLUE). Given a set H in the σ -field \mathcal{F} , the *Best Linear Unbiased Estimator (BLUE) conditional on $\sigma(\mathbf{X})$ in H* is defined as follows. An estimator \mathbf{b}_T for \mathbf{b}_0 is the BLUE conditional on $\sigma(\mathbf{X})$ in H if (1) \mathbf{b}_T is linear conditional on $\sigma(\mathbf{X})$, namely, \mathbf{b}_T can be written as $\mathbf{b}_T = \mathbf{A}\mathbf{y}$ where \mathbf{A} is a $K \times T$ matrix, and each element of \mathbf{A} is measurable $\sigma(\mathbf{X})$; (2) \mathbf{b}_T is unbiased conditional on $\sigma(\mathbf{X})$ in G , namely, $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0$ for s in H with probability $Pr(H)$; (3) for any linear unbiased estimator \mathbf{b}^* conditional on $\mathbf{X}(s) = \mathbf{x}$ for which $E(\mathbf{b}^*\mathbf{b}^{*'})$ exists and is finite, $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}]$ in H with probability $Pr(H)$, namely, $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X}(s) = \mathbf{x}]$ is a positive semidefinite matrix with probability one for s in H with probability $Pr(H)$.

An estimator \mathbf{b}_T for \mathbf{b}_0 is the BLUE conditional on $\mathbf{X}(s) = \mathbf{x}$ in H if (1) \mathbf{b}_T

is linear conditional on $\mathbf{X}(s)$ in H , namely, \mathbf{b}_T can be written as $\mathbf{b}_T = \mathbf{A}\mathbf{y}$ where \mathbf{A} is a $K \times T$ matrix, and each element of \mathbf{A} is measurable $\sigma(\mathbf{X})$; (2) \mathbf{b}_T is unbiased conditional on $\mathbf{X}(s) = \mathbf{x}$ in H , namely, $E(\mathbf{b}_T | \mathbf{X}(s) = \mathbf{x}) = \mathbf{b}_0$ for any s in H ; (3) for any linear unbiased estimator \mathbf{b}^* conditional on $\mathbf{X}(s) = \mathbf{x}$ for which $E(\mathbf{b}^* \mathbf{b}^{*'} | \mathbf{X}(s) = \mathbf{x})$ exists and is finite, $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}]$ in H , namely, $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)' | \mathbf{X}(s) = \mathbf{x}]$ is a positive semidefinite matrix for any s in H .

With these preparations, the following theorem can be stated:

Theorem 5.1 (*The Conditional Gauss-Markov Theorem*) Under Assumptions 5.1-5.4, the OLS estimator is the BLUE conditional on $\sigma(\mathbf{X})$ in G . Under Assumptions 5.1-5.3 and 5.4', the OLS estimator is the BLUE conditional on $\mathbf{X}(s) = \mathbf{x}$ in G . ■

The theorem can be proved by applying any of the standard proofs of the (unconditional) Gauss-Markov theorem by replacing the unconditional expectation with the appropriate conditional expectation.

Under Assumptions 5.1-5.4, the unconditional expectation and the unconditional covariance matrix of \mathbf{b}_T can be defined. With an additional assumption that $Pr(G) = 1$ or

Assumption 5.5 $\mathbf{X}'\mathbf{X}$ is nonsingular with probability one,

we obtain the following corollary of the theorem:

Proposition 5.1 Under Assumptions 5.1-5.5, the OLS estimator is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on $\sigma(\mathbf{X})$.

Proof Using the law of iterated expectations, $E(\mathbf{b}_T) = E\{E[\mathbf{b}_T|\sigma(\mathbf{X})]\} = E(\mathbf{b}_0) = \mathbf{b}_0$, and $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'] = E\{E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\sigma(\mathbf{X})]\} = \sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$. For the minimum covariance matrix part, let \mathbf{b}^* be another linear unbiased estimator conditional on $\sigma(\mathbf{X})$. Then

$$(5.3) \quad E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\sigma(\mathbf{X})] = E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\sigma(\mathbf{X})] + \Delta,$$

where Δ is a positive semidefinite matrix with probability one. Then $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'] = [E(\mathbf{b}^*\mathbf{b}^{*'}) - \mathbf{b}_0\mathbf{b}_0'] - [E(\mathbf{b}_T\mathbf{b}_T') - \mathbf{b}_0\mathbf{b}_0'] = E[E(\mathbf{b}^*\mathbf{b}^{*'}|\sigma(\mathbf{X})) - E[E(\mathbf{b}_T\mathbf{b}_T'|\sigma(\mathbf{X}))]] = E(\Delta)$ is a positive semidefinite matrix. (?????) ■

Masao
needs to
check this!

A few remarks for this proposition are in order:

Remark 5.1 Assumption 5.4 cannot be replaced by Assumption 5.4' for this proposition. Under Assumption 5.4', $E(\mathbf{b}_T)$ and $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)']$ may not exist. ■

Remark 5.2 In this proposition, the covariance matrix of \mathbf{b}_T is $\sigma^2 E[(\mathbf{X}'\mathbf{X})^{-1}]$, which is different from $\sigma^2 [E(\mathbf{X}'\mathbf{X})]^{-1}$. This result may seem to contradict the standard asymptotic theory, but it does not. Asymptotically, $\frac{1}{T}\mathbf{X}'\mathbf{X}$ converges almost surely to $E[X_t X_t']$ if X_t is stationary and ergodic. Hence the limit of the covariance matrix of $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0)$, $\sigma^2 E[\{\frac{1}{T}(\mathbf{X}'\mathbf{X})\}^{-1}]$, is equal to the asymptotic covariance matrix, $\sigma^2 [E(X_t X_t')]^{-1}$. ■

5.2 Unconditional Distributions of Test Statistics

In order to study distributions of the t ratios and F test statistics, we need an additional assumption:

Assumption 5.6 Conditional on \mathbf{X} , \mathbf{e} follows a multivariate normal distribution.

Given a $1 \times K$ vector of real numbers \mathbf{R} , consider a random variable

$$(5.4) \quad N_R = \frac{\mathbf{R}(\mathbf{b}_T - \mathbf{b}_0)}{\sigma[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}]^{\frac{1}{2}}}$$

and the usual t ratio for $\mathbf{R}\mathbf{b}_0$

$$(5.5) \quad t_R = \frac{\mathbf{R}(\mathbf{b}_T - \mathbf{b}_0)}{\hat{\sigma}[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{\frac{1}{2}}}.$$

Here $\hat{\sigma}$ is the positive square root of $\hat{\sigma}^2 = \frac{1}{T-K}(\mathbf{y} - \mathbf{X}\mathbf{b}_T)'(\mathbf{y} - \mathbf{X}\mathbf{b}_T)$. With the standard argument, N_R and t_R can be shown to follow the standard normal distribution and Student's t distribution with $T - K$ degrees of freedom conditional on \mathbf{X} , respectively, under either Assumptions 5.1-5.6 or Assumptions 5.1-5.3, 5.4', and 5.5-5.6. The following proposition is useful in order to derive the unconditional distributions of these statistics.

Proposition 5.2 If the probability density function of a random variable Z conditional on a random vector \mathbf{Q} does not depend on the values of \mathbf{Q} , then the marginal probability density function of Z is equal to the probability density function of Z conditional on \mathbf{Q} . ■

This proposition is obtained by integrating the probability density function conditional on \mathbf{Q} over all possible values of the random variables in \mathbf{Q} . Since N_R and t_R follow the standard normal and the Student's t distribution conditional on \mathbf{X} , respectively, Proposition 5.2 implies the following proposition:

Proposition 5.3 Under the Assumptions 5.1-5.6, or under the Assumptions 5.1-5.3, 5.4', and 5.5-5.6, N_R is the standard normal random variable and t_R is the Student's t random variable with $T - K$ degrees of freedom. ■

Similarly, the usual F test statistics also follow (unconditional) F distributions. These results are sometimes not well understood by econometricians. For example, a standard textbook, Judge et al. (1985, p.164), states that "our usual test statistics

do not hold in finite samples” on the grounds that \mathbf{b}_T 's (unconditional) distribution is not normal. It is true that \mathbf{b}_T is a nonlinear function of \mathbf{X} and \mathbf{e} , so it does not follow a normal distribution even if \mathbf{X} and \mathbf{e} are both normally distributed. However, the usual t and F test statistics have usual (unconditional) distributions as a result of Proposition 5.2.

5.3 The Law of Large Numbers

If an estimator \mathbf{b}_T converges almost surely to a vector of parameters \mathbf{b}_0 , then \mathbf{b}_T is *strongly consistent* for \mathbf{b}_0 . If an estimator \mathbf{b}_T converges in probability to a vector of parameters \mathbf{b}_0 , then \mathbf{b}_T is *weakly consistent* for \mathbf{b}_0 .

Consider a univariate stationary stochastic process $\{X_t\}$. When X_t is stationary, $E(X_t)$ does not depend on date t . Therefore, we often write $E(X)$ instead of $E(X_t)$. Assume that $E(|X|)$ is finite, and consider a sequence of random variables $[Y_T : T \geq 1]$, where $Y_T = \frac{1}{T} \sum_{t=1}^T X_t$ is the sample mean of X computed from a sample of size T . In general, the sample mean does not converge to its unconditional expected value, but converges almost surely to an expectation of X conditional on an information set. For the sample mean to converge almost surely to its unconditional mean, we require the series to be ergodic. A stationary process $\{X_t\}$ is said to be *ergodic* if, for any bounded functions $f : R^{i+1} \mapsto R$ and $g : R^{j+1} \mapsto R$,

$$(5.6) \quad \begin{aligned} & \lim_{T \rightarrow \infty} |E[f(X_t, \dots, X_{t+i})g(X_{t+T}, \dots, X_{t+T+j})]| \\ &= |E[f(X_t, \dots, X_{t+i})]| |E[g(X_t, \dots, X_{t+j})]|. \end{aligned}$$

Heuristically, a stationary process is ergodic if it is asymptotically independent: that is, if (X_t, \dots, X_{t+i}) and $(X_{t+T}, \dots, X_{t+T+j})$ are approximately independent for large enough T .

Proposition 5.4 (*The strong law of large numbers*) If a stochastic process $[X_t : t \geq 1]$ is stationary and ergodic, and if $E(|X|)$ is finite, then $\frac{1}{T} \sum_{t=1}^T X_t \rightarrow E(X)$ almost surely. ■

5.4 Convergence in Distribution and Central Limit Theorem

This section explains a definition of convergence in distribution and presents some central limit theorems. These central limit theorems are based on martingale difference sequences, and are useful in many applications of rational expectations models.

Central limit theorems establish that the sample mean scaled by T converges in distribution to a normal distribution³ under various regularity conditions. The following central limit theorem by Billingsley (1961) is useful for many applications because we can apply it when economic models imply that a variable is a martingale difference sequence.

Proposition 5.5 (*Billingsley's Central Limit Theorem*) Suppose that e_t is a stationary and ergodic martingale difference sequence adapted to I_t , and that $E(|e|^2) < \infty$. Assume that $I_{t-1} \subset I_t$ for all t . Then

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T e_t \xrightarrow{D} N(0, E(e^2)).$$

If e_t is an i.i.d. white noise, then it is a stationary and ergodic martingale difference sequence adapted to I_t which is generated from $\{e_t, e_{t-1}, \dots\}$. Hence Billingsley's

³In some central limit theorems, the limiting distribution is not normal.

Central Limit Theorem is more general than the central limit theorems for i.i.d. processes such as the Lindeberg- Levy theorem, which is usually explained in econometric text books. However, Billingsley's Central Limit Theorem cannot be applied to any serially correlated series.

A generalization of the theorem to serially correlated series is due to Gordin (1969):

Proposition 5.6 (*Gordin's Central Limit Theorem*) Suppose that e_t is a univariate stationary and ergodic process with mean zero and $E(|e|^2) < \infty$, that $E(e_t|e_{t-j}, e_{t-j-1}, \dots)$ converges in mean square to 0 as $j \rightarrow \infty$, and that

$$(5.7) \quad \sum_{j=0}^{\infty} [E(r_{tj}^2)]^{\frac{1}{2}} < \infty,$$

where

$$(5.8) \quad r_{tj} = E(e_t | \mathbf{I}_{t-j}) - E(e_t | \mathbf{I}_{t-j-1}),$$

where \mathbf{I}_t is the information set generated from $\{e_t, e_{t-1}, \dots\}$. Then e_t 's autocovariances are absolutely summable, and

$$(5.9) \quad \frac{1}{\sqrt{T}} \sum_{t=1}^T e_t \xrightarrow{D} N(0, \Omega),$$

where

$$(5.10) \quad \Omega = \lim_{T \rightarrow \infty} \sum_{j=-T+1}^{T-1} E(e_t e_{t-j}).$$

■

When e_t is serially correlated, the sample mean scaled by T still converges to a normal distribution, but the variance of the limiting normal distribution is affected by serial correlation as in (5.10).

In (5.10), Ω is called a *long-run variance* of e_t . Intuition behind the long-run variance can be obtained by observing

$$(5.11) \quad E\left[\left(\frac{1}{\sqrt{T}} \sum_{t=1}^T e_t\right)^2\right] = \sum_{j=-T+1}^{T-1} \frac{T-|j|}{T} E(e_t e_{t-j})$$

and that the right hand side (5.11) is the Cesaro sum of $\sum_{j=-T+1}^{T-1} E(e_t e_{t-j})$. Thus when $\sum_{j=-T+1}^{T-1} E(e_t e_{t-j})$ converges, its limit is equal to the limit of the right hand side of (5.11) (Apostol, 1974).

Another expression for the long-run variance can be obtained from an MA representation of e_t . Let $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \dots$ be an MA representation. Then $E(e_t e_{t-j}) = (\Psi_j \Psi_0 + \Psi_{j+1} \Psi_1 + \Psi_{j+2} \Psi_2 + \dots) E(u_t^2)$, and $\Omega = \{(\Psi_0^2 + \Psi_1^2 + \Psi_2^2 + \dots) + 2(\Psi_1 \Psi_0 + \Psi_2 \Psi_1 + \Psi_3 \Psi_2 + \dots) + 2(\Psi_2 \Psi_0 + \Psi_3 \Psi_1 + \Psi_4 \Psi_2 + \dots) + \dots\} E(u_t^2) = (\Psi_0 + \Psi_1 + \Psi_2 + \dots)^2 E(u_t^2)$. Hence

$$(5.12) \quad \Omega = \Psi(1)^2 E(u_t^2).$$

In the next example, we consider a multi-period forecasting model. For this model, it is easy to show that Gordin's Theorem is applicable to the serially correlated forecast error.

Example 5.1 (*The Multi-Period Forecasting Model*) Suppose that I_t is an information set generated by $\{\mathbf{Y}_t, \mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \dots\}$, where \mathbf{Y}_t is a stationary and ergodic vector stochastic process. In typical applications, economic agents are assumed to use current and past values of \mathbf{Y}_t to generate their information set. Let X_t be a stationary and ergodic random variable in the information set I_t with $E(|X_t|^2) < \infty$. We consider an s -period ahead forecast of X_t , $E(X_{t+s}|I_t)$, and the forecast error, $e_t = X_{t+s} - E(X_{t+s}|I_t)$.

It is easy to verify that all the conditions for Gordin's Theorem are satisfied for e_t . Moreover, because $E(e_t|\mathbf{I}_t) = 0$ and e_t is in the information set \mathbf{I}_{t+s} , $E(e_t e_{t-j}) = E(E(e_t e_{t-j}|\mathbf{I}_t)) = E(e_{t-j} E(e_t|\mathbf{I}_t)) = 0$ for $j \geq s$. Hence $\Omega = \lim_{j \rightarrow \infty} \sum_{-j}^j E(e_t e_{t-j}) = \sum_{j=-s+1}^{s-1} E(e_t e_{t-j})$. ■

Hansen (1985) generalized Gordin's Central Limit Theorem to vector processes. In this book, we call the generalized theorem Gordin and Hansen's Central Limit Theorem.

Proposition 5.7 (*Gordin and Hansen's Central Limit Theorem*) Suppose that \mathbf{e}_t is a vector stationary and ergodic process with mean zero and finite second moments, that $E(\mathbf{e}_t | \mathbf{e}_{t-j}, \mathbf{e}_{t-j-1}, \dots)$ converges in mean square to 0 as $j \rightarrow \infty$, and that

$$(5.13) \quad \sum_{j=0}^{\infty} [E(\mathbf{r}'_{tj} \mathbf{r}_{tj})]^{\frac{1}{2}} < \infty,$$

where

$$(5.14) \quad \mathbf{r}_{tj} = E(\mathbf{e}_t | \mathbf{I}_{t-j}) - E(\mathbf{e}_t | \mathbf{I}_{t-j-1}),$$

where \mathbf{I}_t is the information set generated from $\{\mathbf{e}_t, \mathbf{e}_{t-1}, \dots\}$. Then \mathbf{e}_t 's autocovariances are absolutely summable, and

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \mathbf{e}_t \xrightarrow{D} N(\mathbf{0}, \Omega)$$

where

$$(5.15) \quad \Omega = \lim_{T \rightarrow \infty} \sum_{j=-T+1}^{T-1} E(\mathbf{e}_t \mathbf{e}'_{t-j}).$$

■

The matrix Ω in Equation (5.15) is called the *long-run covariance matrix* of \mathbf{e}_t .

As in the univariate case, another expression for the long-run covariance can be obtained from an MA representation of \mathbf{e}_t . Let $\mathbf{e}_t = \Psi(L)\mathbf{u}_t = \Psi_0\mathbf{u}_t + \Psi_1\mathbf{u}_{t-1} + \dots$ be an MA representation. Then $E(\mathbf{e}_t\mathbf{e}'_{t-j}) = (\Psi_j + \Psi_{j+1} + \Psi_{j+2} + \dots)E(\mathbf{u}_t\mathbf{u}'_t)(\Psi_0 + \Psi_1 + \Psi_2 + \dots)'$, and $\Omega = (\Psi_0 + \Psi_1 + \Psi_2 + \dots)E(\mathbf{u}_t\mathbf{u}'_t)(\Psi_0 + \Psi_1 + \Psi_2 + \dots)'$. Hence

$$(5.16) \quad \Omega = \Psi(1)E(\mathbf{u}_t\mathbf{u}'_t)\Psi(1)'$$

In the next example, Gordin and Hansen's Central Limit Theorem is applied to a serially correlated vector process:

Example 5.2 Continuing Example 5.1, let \mathbf{Z}_t be a random vector with finite second moments in the information set \mathcal{I}_t . Define $\mathbf{f}_t = \mathbf{Z}_t e_t$. Then $E(\mathbf{f}_t|\mathcal{I}_t) = E(\mathbf{Z}_t e_t|\mathcal{I}_t) = E(\mathbf{Z}_t E(e_t|\mathcal{I}_t)) = \mathbf{0}$. In empirical work, it is often necessary to apply a central limit theorem to a random vector such as \mathbf{f}_t . It is easy to verify that all conditions for Gordin and Hansen's Theorem are satisfied for \mathbf{f}_t . Moreover, $E(\mathbf{f}_t|\mathcal{I}_t) = \mathbf{0}$ and \mathbf{f}_t is in the information set \mathcal{I}_{t+s} , thus $E(\mathbf{f}_t\mathbf{f}'_{t-j}) = E(E(\mathbf{f}_t\mathbf{f}'_{t-j}|\mathcal{I}_t)) = E(E(\mathbf{f}_t|\mathcal{I}_t)\mathbf{f}'_{t-j}) = \mathbf{0}$ for $j \geq s$. Hence $\Omega = \lim_{j \rightarrow \infty} \sum_{-j}^j E(\mathbf{f}_t\mathbf{f}'_{t-j}) = \sum_{j=-s+1}^{s-1} E(\mathbf{f}_t\mathbf{f}'_{t-j})$. ■

We assumed that the process is stationary and ergodic for the law of large numbers and central limit theorems. In most applications, this ergodic stationarity assumption is general enough. However, in some applications, such an assumption may not be convenient. For example, suppose that data of a process of interest shows an initial rapid growth and then stabilizes. It is not attractive to assume ergodic stationarity because the expected value of the process seems initially rising. In such cases, we can use an alternative assumption that the process is mixing. Mixing can be regarded as an asymptotic independence. For stationary and ergodic processes,

we used the concept of martingale difference sequence for central limit theorems. For mixing processes, the corresponding concept is mixingale processes. The concepts of mixing and mixingale are explained in Appendix A.

5.5 Consistency and Asymptotic Distributions of OLS Estimators

Consider a linear model,

$$(5.17) \quad y_t = \mathbf{x}'_t \mathbf{b}_0 + e_t,$$

where y_t and e_t are stationary and ergodic random variables, and \mathbf{x}_t is a p -dimensional stationary and ergodic random vector. We assume that the orthogonality conditions

$$(5.18) \quad E(\mathbf{x}_t e_t) = \mathbf{0}$$

are satisfied, and that $E(\mathbf{x}_t \mathbf{x}'_t)$ is nonsingular.⁴ Imagine that we observe a sample of (y_t, \mathbf{x}'_t) of size T . Proposition 5.4 shows that $\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t$ converges to $E(\mathbf{x}_t \mathbf{x}'_t)$ almost surely. Hence with probability one, $\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t(s)$ is nonsingular for large enough T , and the Ordinary Least Squares (OLS) estimator for (5.17) can be written as

$$(5.19) \quad \mathbf{b}_T = \left(\sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\sum_{t=1}^T \mathbf{x}_t y_t \right).$$

In order to apply the Law of Large Numbers to show that the OLS estimator is strongly consistent, rewrite (5.19) from (5.17) after scaling each element of the right side by T :

$$(5.20) \quad \mathbf{b}_T - \mathbf{b}_0 = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T (\mathbf{x}_t e_t) \right).$$

⁴Appendix 3.A explains why these types of conditions are called orthogonality conditions.

Applying Proposition 5.4, we obtain

$$(5.21) \quad \mathbf{b}_T - \mathbf{b}_0 \rightarrow [E(\mathbf{x}_t \mathbf{x}_t')]^{-1} (E(\mathbf{x}_t e_t)) = \mathbf{0} \quad \text{almost surely.}$$

Hence the OLS estimator, \mathbf{b}_T , is a strongly consistent estimator. In order to obtain the asymptotic distribution of the OLS estimator, we make an additional assumption that a central limit theorem applies to $\mathbf{x}_t e_t$. In particular, assuming that Gordin and Hansen's Martingale Approximation Central Limit Theorem is applicable, we multiply both sides of (5.20) by the square root of T :

$$(5.22) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\frac{1}{\sqrt{T}} \sum_{t=1}^T (\mathbf{x}_t e_t) \right).$$

Therefore,

$$(5.23) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, [E(\mathbf{x}_t \mathbf{x}_t')]^{-1} \boldsymbol{\Omega} [E(\mathbf{x}_t \mathbf{x}_t')]^{-1})$$

where $\boldsymbol{\Omega}$ is the long-run covariance matrix of $\mathbf{x}_t e_t$:

$$(5.24) \quad \boldsymbol{\Omega} = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} \mathbf{x}_t \mathbf{x}_{t-j}').$$

5.6 Consistency and Asymptotic Distributions of IV Estimators

Consider the linear model (5.17) for which the orthogonality conditions (5.18) are not satisfied. In this case, we try to find a p -dimensional stationary and ergodic random vector \mathbf{z}_t , which satisfies two types of conditions: the orthogonality condition

$$(5.25) \quad E(\mathbf{z}_t e_t) = \mathbf{0},$$

and the relevance condition that $E(\mathbf{z}_t \mathbf{x}_t')$ is nonsingular. We define the *Linear Instrumental Variable* (IV) estimator as

$$(5.26) \quad \mathbf{b}_T = \left(\sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^T \mathbf{z}_t y_t.$$

Then

$$(5.27) \quad \mathbf{b}_T - \mathbf{b}_0 = \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \mathbf{z}_t e_t \right).$$

Applying Proposition 5.4, we obtain

$$(5.28) \quad \mathbf{b}_T - \mathbf{b}_0 \rightarrow [E(\mathbf{z}_t \mathbf{x}_t')]^{-1} (E(\mathbf{z}_t e_t)) = \mathbf{0} \quad \text{almost surely.}$$

Hence the linear IV estimator, \mathbf{b}_T , is a strongly consistent estimator. Assuming that the Vector Martingale Approximation Central Limit Theorem is applicable to $\mathbf{z}_t e_t$,

$$(5.29) \quad \sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} N(\mathbf{0}, [E(\mathbf{z}_t \mathbf{x}_t')]^{-1} \boldsymbol{\Omega} [E(\mathbf{z}_t \mathbf{x}_t')]^{-1})$$

where $\boldsymbol{\Omega}$ is the long-run covariance matrix of $\mathbf{z}_t e_t$:

$$(5.30) \quad \boldsymbol{\Omega} = \sum_{j=-\infty}^{\infty} E(e_t e_{t-j} \mathbf{z}_t \mathbf{z}_{t-j}').$$

5.7 Nonlinear Functions of Estimators

In many applications of linear models, we are interested in nonlinear functions of \mathbf{b}_0 , say $\mathbf{a}(\mathbf{b}_0)$. This section explains the delta method, which is a convenient method to derive asymptotic properties of $\mathbf{a}(\mathbf{b}_T)$ as an estimator for \mathbf{b}_0 where \mathbf{b}_T is a weakly consistent estimator for \mathbf{b}_0 . In many applications, \mathbf{b}_T is an OLS estimator or a linear

Masao
needs to
check this!

IV estimator. Later????? in this book we will use the proof of the delta method to prove the asymptotic normality of the GMM estimator. (????? \mathbf{a} not bold, f is

Masao
needs to
check this!

better?)

Proposition 5.8 Suppose that $\{\mathbf{b}_T\}$ is a sequence of p -dimensional random vectors such that $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0) \xrightarrow{D} \mathbf{z}$ for a random vector \mathbf{z} . If $\mathbf{a}(\cdot) : R^p \mapsto R^r$ is continuously differentiable at \mathbf{b} , then

$$\sqrt{T}[\mathbf{a}(\mathbf{b}_T) - \mathbf{a}(\mathbf{b}_0)] \xrightarrow{D} \mathbf{d}(\mathbf{b}_0)\mathbf{z},$$

where $\mathbf{d}(\mathbf{b}_0) = \frac{\partial \mathbf{a}(\mathbf{b})}{\partial \mathbf{b}'} \Big|_{\mathbf{b}=\mathbf{b}_0}$ denotes the $r \times p$ matrix of first derivatives evaluated at \mathbf{b}_0 . In particular, if $\mathbf{z} \sim N(\mathbf{0}, \Sigma)$, then

$$\sqrt{T}[\mathbf{a}(\mathbf{b}_T) - \mathbf{a}(\mathbf{b}_0)] \xrightarrow{D} N(\mathbf{0}, \mathbf{d}(\mathbf{b}_0)\Sigma\mathbf{d}(\mathbf{b}_0)').$$

Proof ????????????

■ Masao
needs to
check this!

5.8 Remarks on Asymptotic Theory

When we use asymptotic theory, we do not have to make restrictive assumptions that the disturbances are normally distributed. Serial correlation and conditional heteroskedasticity can be easily taken into account as long as we can estimate the long-run covariance matrix (which is the topic of the next chapter).

It is a common mistake to think that the linearity of the formula for the long-run covariance matrix means a linearity assumption for the process of $\mathbf{x}_t e_t$ (for the OLS estimator) or $\mathbf{z}_t e_t$ (for the IV estimator). It should be noted that we did not assume that $\mathbf{x}_t e_t$ or $\mathbf{z}_t e_t$ was generated by linear functions (i.e., a moving average process in the terminology of Chapter 4) of independent white noise processes. Even when $\mathbf{x}_t e_t$ or $\mathbf{z}_t e_t$ is generated from nonlinear functions of independent white noise processes, the distributions based on the long-run covariance matrices give the correct limiting distributions. This point is related to the Wold representation for nonlinear processes discussed in Chapter 4. Even when $\mathbf{z}_t e_t$ is generated as a nonlinear process, as long as it is a linearly regular and covariance stationary process, it has the Wold representation: $\mathbf{z}_t e_t = \Psi(L)\mathbf{u}_t$, and its long-run covariance matrix is given by (5.30).

5.9 Monte Carlo Methods

This section gives an introduction to Monte Carlo methods. An important advanced Monte Carlo method called the Markov Chain Monte Carlo (MCMC)⁵ will be explained in Chapter 12 for the Bayesian Approach. The MCMC method is a very powerful numerical integration method that can be used for both the Bayesian statistics and the Classical statistics even though most applications of the method so far have been in the Bayesian statistics. Asymptotic theory is used to obtain approximations of the exact finite sample properties of estimators and test statistics. In many time series applications, the exact finite sample properties cannot be obtained. For example, in a regression with lagged dependent variables, we can assume neither that the regressor is nonrandom nor that the error term is strictly exogenous in the time series sense. In many applications with financial variables, the assumption that the error term in a regression is normal is inappropriate because many authors have found evidence against normality for several financial variables. Asymptotic theory gives accurate approximations when the sample size is “large,” but exactly how “large” is enough depends on each application. One method to study the quality of asymptotic approximations is the Monte Carlo simulations.

5.9.1 Random Number Generators

In relatively simple Monte Carlo studies, data are generated with computer programs called *pseudo-random number generators*. These programs generate sequences of values that appear to be draws from a specified probability distribution. Modern pseudo-random generators are accurate enough that we can ignore the fact that

⁵It is important to the extent that its emergence is called the MCMC revolution.

numbers generated are not exactly independent draws from a specified probability distribution for most purposes.⁶ Hence in the rest of this appendix, phrases such as “values that appear to be” are often suppressed.

Recall that when a probability space Ω is given, the whole history of a stochastic process $\{e_t(s)\}_{t=1}^N$ is determined when a point in the probability space s is given. For a random number generator, we use a number called the starting *seed* to determine s . Then the random number generator automatically updates the seed each time a number is generated. It should be noted that the same sequence of numbers is generated whenever the same starting seed is given to a random number generator.

Generated random numbers are used to generate samples. From actual data, we obtain only one sample, but in Monte Carlo studies, we can obtain many samples from generated random numbers. Each time a sample is generated, we compute estimators or test statistics of interest. After replicating many samples, we can estimate small sample properties of the estimators or test statistics by studying the generated distributions of these variables and compare them with predictions of asymptotic theory.

Most programs offer random number generators for the uniform distribution and the standard normal distribution. One can produce random numbers with other distributions by transforming generated random numbers. See the Appendix for more explanations.

⁶One exception is that a pseudo-random number generator ultimately cycles back to the initial value generated and repeats the sequence when too many numbers are generated. Most modern pseudo-random number generators cycle back after millions of values are drawn, and this tendency is not a problem for most Monte Carlo studies. However, in some studies in which millions or billions of values are needed, there can be a serious problem.

5.9.2 Estimators

When a researcher applies an estimator to actual data without the normality assumption, asymptotic theory is used as a guide of small sample properties of the estimator. In some cases, asymptotic theory does not give a good approximation of the exact finite sample properties. A Monte Carlo study can be used to estimate the true finite sample properties. For example, the mean, median, and standard deviation of the realized values of the estimator over generated samples can be computed and reported as estimates of the true values of these statistics. For example, N independent samples are created and an estimate b_i ($i \geq 1$) for a parameter b_0 is calculated for the i -th sample. Then the expected value of the estimator $E(b_i)$ can be estimated by its mean over the samples: $\frac{1}{N} \sum_{i=1}^N b_i$. By the strong law of large numbers, the mean converges almost surely to the expected value.

Other properties can also be reported, depending on the purpose of the study. For example, Nelson and Startz (1990) report estimated 1%, 5%, 10%, 50%, 90%, and 99% fractiles for an IV estimator and compared them with fractiles implied by the asymptotic distribution. This influential paper uses Monte Carlo simulations to study the small sample properties of IV estimator and its t -ratio when instruments are poor in the sense that the relevance condition is barely satisfied.

When the deviation from the normal distribution is of interest, the skewness and kurtosis are often estimated and reported. The *skewness* of a variable Y with mean μ is

$$(5.31) \quad \frac{E(Y - \mu)^3}{[Var(Y)]^{\frac{3}{2}}}.$$

A variable with negative skewness is more likely to be far below the mean than it is

to be far above, and conversely a variable with positive skewness is more likely to be far above the mean than it is to be below. If Y has a symmetric distribution such as a normal distribution, then the skewness is zero. The *kurtosis* of Y is

$$(5.32) \quad \frac{E(Y - \mu)^4}{[Var(Y)]^2}.$$

If Y is normally distributed, the kurtosis is 3. If the kurtosis of Y exceeds 3, then its distribution has more mass in the tails than the normal distribution with the same variance.

5.9.3 Tests

When a researcher applies a test to actual data without the normality assumption, asymptotic theory is typically used. For example, the critical value of 1.96 is used for a test statistic with the asymptotic normal distribution for the significance level of 5%. The significance level and critical value based on the asymptotic distribution are called the *nominal significance level* and the *nominal critical value*, respectively. The probability of rejecting the null hypothesis when it is true is called the *size* of the test. Since the asymptotic distribution is not exactly equal to the exact distribution of the test statistic, the true size of the test based on the nominal critical value is usually either larger or smaller than the nominal significance level. This property is called the *size distortion*. If the true size is larger than the nominal significance level, the test *overrejects* the null hypothesis and is said to be *liberal*. If the true size is smaller than the nominal significance level, the test *underrejects* the null hypothesis and is said to be *conservative*. Using the distribution of the test statistic produced by a Monte Carlo simulation, one can estimate the true critical value.

The *power* of the test is the probability of rejecting the null hypothesis when

the alternative hypothesis is true. In Monte Carlo studies, two versions of the power can be reported for each point of the alternative hypothesis: the power based on the nominal critical value and the power based on the estimated true critical value. The latter is called the *size corrected power*. The power based on the nominal critical value is also of interest because it is the probability of rejecting the null hypothesis in practice if asymptotic theory is used. On the other hand, the size corrected power is more appropriate for the purpose of comparing tests. For example, a liberal test tends to have a higher power based on the nominal critical value than a conservative test. However, we cannot conclude the liberal test is better from this observation because the probability of Type I error is not equal for the two tests.

5.10 Bootstrap

When the asymptotic distribution of a random variable such as a parameter estimate and test statistic is unknown or unreliable, an estimation method called the *bootstrap* is used as an alternative to the asymptotic theory. The bootstrap estimates the unknown underlying probability distribution of interest using a known distribution function generated by a random sampling procedure. In this sense, the bootstrap distribution treats the random sample as if it is a good representation of the population. Under mild regularity conditions and with the sample sizes typically encountered in applied work, this method can provide as accurate an approximation as that obtained from the asymptotic theory. Moreover, often in cross-sectional applications, the bootstrap approximations can achieve the level of accuracy comparable to higher-order asymptotic approximations. When the bootstrap improves upon first-order asymptotic approximations, it is said to benefit from *asymptotic refinements*.

Asymptotic refinements are an important feature of the bootstrap in reducing or eliminating finite-sample bias of an estimator or finite-sample errors in the rejection probabilities of statistical tests. For these reasons, since its introduction by Efron (1979), the bootstrap has become a practical and increasingly popular tool in applied econometrics.⁷

To illustrate how the bootstrap is implemented in a simple setting, suppose you have a random sample $\{x_1, x_2, \dots, x_T\}$ of an i.i.d. random variable x with cumulative distribution function (CDF) F_0 . Let $Q_T = Q_T(x_1, x_2, \dots, x_T)$ denote the statistic of interest, and $G_T(q, F_0) \equiv Pr(Q_T \leq q)$ the exact, finite-sample CDF of Q_T . Because F_0 is usually unknown in applications, the bootstrap method replaces F_0 with its estimator F_T , and approximates $G_T(q, F_0)$ by the bootstrap distribution $G_T(q, F_T)$ based on which you can make inferences about Q_T .

There are two possible specifications of F_T . The *nonparametric bootstrap* uses the empirical distribution function of the data as F_T . The other approach, the *parametric bootstrap*, uses a parametric estimator of F_0 as F_T . For instance, if x is assumed to be normally distributed with mean μ and variance σ^2 , then F_T is defined as $N(\hat{\mu}, \hat{\sigma}^2)$ where $\hat{\mu}$ and $\hat{\sigma}^2$ are consistent estimates of μ and σ^2 , respectively.

In most applications, $G_T(q, F_T)$ cannot be evaluated analytically, but is approximated using a Monte Carlo simulation. The steps for this procedure are as follows.

1. Draw a bootstrap sample of size T , $X^* = \{x_1^*, x_2^*, \dots, x_T^*\}$, from the distribution corresponding to F_T randomly. For the nonparametric bootstrap, the observations are resampled from the original data set with replacement, with each point in the sample having the equal probability $1/T$ of being drawn. Clearly,

⁷See, e.g., Jeong and Maddala (1993) and Horowitz (2001) for survey and details of different bootstrap methods and their theoretical justification.

some of the original data points may be included in X^* once or more than once, while others may not be included at all. For the parametric bootstrap, X^* is generated using a random number generator.

2. Using X^* , compute the bootstrap statistic $Q_{T,1}^* \equiv Q_T(x_1^*, x_2^*, \dots, x_T^*)$.
3. Repeat steps 1 and 2 B times to obtain observations $\{Q_{T,1}^*, \dots, Q_{T,B}^*\}$.
4. The bootstrap distribution $G_T(q, F_T)$ is estimated by $G_T^*(q, F_T) = Pr(Q_T^* \leq q)$ putting mass $1/B$ at each point of $\{Q_{T,1}^*, \dots, Q_{T,B}^*\}$.

The resulting bootstrap distribution $G_T^*(q, F_T)$ is then used to compute p -values or confidence intervals, and make inferences about Q_T which is computed in conventional ways.

In order for the bootstrap distribution $G_T(\cdot, F_T)$ to be an adequate estimator of $G_T(\cdot, F_0)$, it must be consistent. That is, $G_T(\cdot, F_T)$ must converge in probability to the asymptotic CDF of Q_T , $G_\infty(\cdot, F_0)$, as $T \rightarrow \infty$. Essentially, the conditions for the consistency of $G_T(\cdot, F_T)$ require that F_T is a consistent estimator of F_0 , and $G_T(\cdot, F)$ is continuous in F in an appropriate sense. It then follows that $G_T(\cdot, F_T)$ approaches $G_T(\cdot, F_0)$ for a sufficiently large sample size.⁸

Although these conditions are likely to be satisfied in many cases of interest in econometrics, they can be violated in some applications. For instance, for the heavy-tailed distributions of Athreya (1987) or the unit root AR(1) model of Basawa, Mallik, McCormick, Reeves, and Taylor (1991), the standard bootstrap method results in poor approximations to the asymptotic distribution of interest. Thus, although the

⁸For a precise definition of consistency, see Appendix A. For conditions for consistency and a detailed discussion on consistency, see Section 2.1 of Horowitz (2001).

bootstrap methods serve as an attractive alternative to the asymptotic theory in many applications, it must be borne in mind that, just as with any econometric methods, they, too, cannot be used blindly.

The following example illustrates an application of the bootstrap to an autoregressive (AR) model, and shows why it requires a non-standard procedure. Consider the AR process of order 1 with an intercept and time trend,

$$x_t = \theta + \mu t + \alpha x_{t-1} + \epsilon_t \quad \text{for } 1 \leq t \leq T,$$

where ϵ_t is i.i.d., $|\alpha| < 1$, and x_0 is a random variable with a stationary distribution so that x_t is stationary. Let $\hat{\alpha}$ be the ordinary least square (OLS) estimator of the autoregressive root α . The usual asymptotic theory indicates that $T^{1/2}(\hat{\alpha} - \alpha)$ converges in distribution to a normal random variable with zero mean. On the contrary, the OLS estimator is significantly downward biased, and the exact, finite-sample distribution of α is asymmetric and has fatter tails than the normal distribution.⁹ In this case, if ϵ_t is i.i.d. and normally distributed, then the exact, finite-sample CDF of $\hat{\alpha}$ only depends on α , and can be computed numerically using Andrews' (1993) procedure without relying on Monte Carlo or bootstrap simulations. The deviations from the prediction of the asymptotic theory are considerable especially when α is close to one. For example, for the sample size of 60, the OLS estimator has downward median-biases of 0.08, 0.09, and 0.15 when α is 0.7, 0.85, and 0.99, respectively. Clearly, using the asymptotic distribution leads to an inaccurate approximation to the exact, finite-sample distribution of $\hat{\alpha}$ and hence results in misleading inferences.¹⁰

⁹It should be noted that the strict exogeneity assumption is violated because of the lagged dependent variable. Hence the argument for the conditional Gauss-Markov theorem cannot be applied.

¹⁰An alternative asymptotic theory called the local-to-unity asymptotic theory can be applied in this case as in Chan and Wei (1987) and Phillips (1987)

If ϵ_t is not i.i.d. or normally distributed, the exact, finite-sample distribution is estimated using bootstrap methods. Tables of the 0.05, 0.5, and 0.95 quantiles of $\hat{\alpha}$ can be found in Andrews (1993) for different sample sizes, AR specifications, and distributions of ϵ_t .¹¹

An important characteristic of the AR models with a near unit root is that the asymptotic distribution of and hence quantile functions for the test statistic depend on α . Nevertheless, the conventional bootstrap approximates quantile functions by evaluating them at the point estimate $\hat{\alpha}$ and thereby making an implicit assumption that these functions are constant, which is false in the AR models. Consequently, the standard bootstrap confidence intervals fail to provide asymptotically correct coverage probabilities.

Table 1 summarizes the 0.05, 0.5, and 0.95 true quantiles of the nonstudentized test statistic $S_T(\alpha) = \hat{\alpha} - \alpha$ for the sample sizes of 40 and 150 over the values of α from 0.70 to 1, assuming that the errors are i.i.d. and normally distributed.¹²

Table 1

α	T=40			T=150		
	$q_{0.05}$	$q_{0.5}$	$q_{0.95}$	$q_{0.05}$	$q_{0.5}$	$q_{0.95}$
0.70	-0.390	-0.118	0.071	-0.144	-0.029	0.062
0.80	-0.403	-0.135	0.038	-0.137	-0.031	0.045
0.85	-0.412	-0.146	0.019	-0.133	-0.033	0.035
0.90	-0.425	-0.160	-0.002	-0.129	-0.036	0.023
0.93	-0.436	-0.172	-0.017	-0.127	-0.038	0.015
0.97	-0.457	-0.194	-0.040	-0.127	-0.045	0.000
0.99	-0.472	-0.209	-0.055	-0.133	-0.052	-0.010
1.00	-0.481	-0.218	-0.065	-0.140	-0.060	-0.018

¹¹For a probability p , the p quantile of a random variable X is the minimum value of x for which $Pr(X \leq x) = p$ is satisfied.

¹²Following Andrews (1993), we restrict the parameter space to be $\alpha \in (-1, 1]$. This assumption is made in order to avoid the dependence of the distribution of the OLS estimator on the initial condition.

These values are computed from table 3 in Andrews (1993) by subtracting the true value of α from the quantile values in the corresponding row. It is clear from the above table that the quantile functions are varying for different values of α . An appropriate bootstrap quantile function must therefore be a function of α rather than $\hat{\alpha}$:

$$q_{0.05}^*(\alpha) \leq S_T(\alpha) \leq q_{0.95}^*(\alpha),$$

such that

$$Pr(q_{0.05}^*(\alpha) \leq S_T(\alpha) \leq q_{0.95}^*(\alpha)) = 0.90.$$

The above statement is *exact* in the sense that once we know the exact finite distribution of the quantiles for a given α , then this set has the correct coverage probability. The upper and lower bounds are thus given by

$$-q_{0.95}^*(\alpha) + \hat{\alpha} \leq \alpha \leq -q_{0.05}^*(\alpha) + \hat{\alpha}.$$

Table 1 can be used to compute the median-unbiased estimator and the two-sided 90% and one-sided 95% confidence intervals for α . Because the grid of α values is finite, interpolation may be necessary for the values of α in between those reported. To see how the table can be used in applications, suppose you have the OLS estimate $\hat{\alpha}$ of 0.781 and the sample size T of 40. The median-unbiased estimate of α is the intersection of $S_T(\alpha)$ and $q_{0.5}^*(\alpha)$. That is, α is such that $\hat{\alpha} - \alpha = q_{0.5}^*(\alpha)$. According to table 1, this occurs when $\alpha = 0.99$ ($0.781 - 0.99 = -0.209$). The lower and upper bounds of the 90% confidence interval can be found in the same way. For the lower bound, the endpoint is the value of α such that $\hat{\alpha} - \alpha = q_{0.95}^*(\alpha)$. You see that $\alpha + q_{0.95}^*(\alpha) = 0.771$ for $\alpha = 0.7$, and $\alpha + q_{0.95}^*(\alpha) = 0.838$ for $\alpha = 0.8$. Because $\hat{\alpha} = 0.781$, the lower bound must lie between 0.7 and 0.8. By interpolation, this is 0.715. The upper bound can be found by $\hat{\alpha} - \alpha = q_{0.05}^*(\alpha)$. Because the parameter

space is restricted to be $\alpha \in (-1, 1]$, any value of $\hat{\alpha}$ that is above 0.519 for $T=40$ and 0.860 for $T=150$ corresponds to the upper bound of 1. Thus, in this example, $\hat{\alpha} > 0.519$, and hence the upper bound of the confidence interval is 1.

This interval is equivalent to Hansen's (1999) *grid bootstrap* for the case of the i.i.d. Gaussian errors. He proposes a nonparametric bootstrap method for constructing confidence intervals for α from bootstrap quantile functions of α , and reports that it has improved performance over the standard bootstrap method when α is close to one.

The condition under which the grid bootstrap confidence interval is first-order accurate only requires that the nuisance parameters are consistently estimated, and no restriction is imposed on the estimate of the parameter of interest. On the other hand, the consistency of the standard bootstrap confidence interval requires that the parameters are consistently estimated and the test statistic of the hypothesis has an asymptotic distribution, where the convergence to the asymptotic distribution is locally uniform in the parameter space. Thus, the conditions for the grip bootstrap are strictly less restrictive than those for the latter in the sense of first-order asymptotic coverage, suggesting that the grid bootstrap is more broadly applicable.

Appendix

5.A Weakly dependence process

Weakly dependence process is a stochastic process where serial dependence exists, but it is restricted suitably so that the limit theorems, such as LLN, CLT, and FCLT, can be applied. There are many different types of weakly dependence processes

depending on its degree of serial dependence. In this section, we review some of the most commonly used ones in the nonstationary econometrics.

The reason why we study weakly dependence process for the nonstationary econometrics is that the nonstationary econometrics is also time-series econometrics, and in time-series econometrics, serial dependence exists in almost all applications. Therefore, we want our asymptotic theories for the nonstationary econometrics can also be applied to the data that has serial dependence.

5.A.1 Independent Process

Definition 5.A.1 *A stochastic process $\{X_t\}_{-\infty}^{\infty}$ is said to be independent if $P(A \cap B) = P(A)P(B)$ for a pair of $A \in \mathcal{F}_{-\infty}^t$ and $B \in \mathcal{F}_{t+m}^{\infty}$ for all t and m .*

Independence implies that there is no relationship between X_t and $X_{t'}$ for any $t \neq t'$, therefore each observation can be treated as an observation from a random sample. From the time series econometrics perspective, independence is the most stringent restriction on the behavior of a stochastic process. It is difficult to find a case where independence assumption is appropriate. However, it can be used as a benchmark against which asymptotic theories of other dependent processes might be compared.

5.A.2 Mixing Process

The idea of independence that there is no relationship between any pair of X_t and $X_{t'}$ is rather special, especially for time series data. However, it might be reasonable to expect that the degree of dependence between X_t and $X_{t'}$ is decreasing as the time t and t' are getting farther separated from each other. We formalize this idea by introducing the concept of mixing.

Definition 5.A.2 A stochastic process $\{X_t\}_{-\infty}^{\infty}$ is said to be mixing (or regular) if, for every $B \in \mathcal{F}$,

$$\sup_{A \in \mathcal{F}_{-\infty}^t} |P(A \cap B) - P(A)P(B)| \rightarrow 0 \text{ as } t \rightarrow -\infty.$$

Mixing can be regarded as an asymptotic independence. Note that an independent process is also mixing. An alternative definition of mixing can be described in terms of remote event. Remote event is defined as an event contained in the remote σ -field, $\mathcal{F}_{-\infty} = \bigcap_t \mathcal{F}_{-\infty}^t$.

Definition 5.A.3 A stochastic process $\{X_t\}_{-\infty}^{\infty}$ is said to be mixing (or regular) if every remote event has probability 0 or 1.

Since mixing is defined by remote events as in Definition 5.A.3, it can hardly provide us with useful description of dependence between events that are widely separated in time, but not in the remote events. Therefore, for a workable theory we need the concepts of mixing coefficients. In this section, we introduce only two most important mixing coefficients, α -mixing and ϕ -mixing although there are several other different versions available. Let \mathcal{G} and \mathcal{H} be σ -subfields of \mathcal{F} . The α -mixing (strong mixing) coefficient is defined by

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G)P(H)|,$$

the uniform mixing coefficient is defined by

$$\phi(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}; P(G) > 0} |P(H|G) - P(H)|$$

Then, the sequence $\{X_t\}_{-\infty}^{\infty}$ is said to be α -mixing (or strong mixing) if

$$\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^{\infty}) \rightarrow 0 \text{ as } m \rightarrow \infty,$$

similary, it is said to be ϕ -mixing (or uniform mixing) if

$$\phi_m = \sup_t \phi(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+m}^\infty) \rightarrow 0 \text{ as } m \rightarrow \infty$$

Note that if $\alpha_m = 0$ for all m , the sequence becomes independent. Measure of the dependence can be based on the rate of convergence at which the mixing coefficients tend to zero. The rate of convergence is quantified by that for some number $\varphi > 0$, $\alpha_m(\phi_m) \rightarrow 0$ sufficiently fast that

$$\sum_{m=1}^{\infty} \alpha_m^{\frac{1}{\varphi}} < \infty \text{ or } \sum_{m=1}^{\infty} \phi_m^{\frac{1}{\varphi}} < \infty.$$

A sequence is said to be α -mixing (ϕ -mixing) of size $-\varphi_0$ if $\alpha_m = O(m^{-\varphi})$ ($\phi_m = O(m^{-\varphi})$) for some $\varphi > \varphi_0$.

5.A.3 Martingale Difference Process

Independence and mixing are conditions for every event in \mathcal{F} . Since sup is taken over all the events in \mathcal{F} , usually it is the most peculiar event that determines the properties. However, in many case, those peculiar event that determine the properties of a stochastic process may not be our main interest. Therefore, sometimes it is more useful if we confine our attention to more restricted measure of dependence, and admit more stochastic process into consideration. Martingale difference and mixingale are two key concepts.

Definition 5.A.4 *A stochastic process $\{X_t\}_{-\infty}^{\infty}$ is said to be a martingale difference (m.d.) sequence if X_t is integrable and*

$$E(X_t | \mathcal{F}_{-\infty}^{t-1}) = 0 \text{ a.s.}$$

Table 5.1: Dependence between X_t and X_{t+m}

	For all m		As $m \rightarrow \infty$
Every events	Independent	\Rightarrow	Mixing
	\Downarrow		\Downarrow
1-period ahead predictability	Martingale Difference	\Rightarrow	Mixingale

This implies that $\{\dots, X_{t-1}\}$ have no impact on the prediction of X_t . It can be thought that X_t 's are independent each other in terms of one-period ahead predictability.

5.A.4 Mixingale Process

Although martingale difference restrict our attention to more restricted measure of dependence, namely predictability, it is still rather special in time series setting that X_t has no prediction power on X_{t+m} at all. Similarly in mixing, it might be more natural to expect that the degree of dependence between between X_t and X_{t+m} in term of predictability is getting smaller as the time m increases. Mixingale captures this idea.

Definition 5.A.5 $\{X_t\}_{-\infty}^{\infty}$ is said to be an L_p -mixingale if

$$\|E(X_t | \mathcal{F}_{-\infty}^{t-m})\|_p \leq \zeta_m \rightarrow 0 \text{ as } m \rightarrow \infty$$

This is the most general dependence concept for that most of asymptotic theories go through.

It can be said that mixingales are to mixing as martingale differences are to independent. Table 1 summarize the relationship among these dependence concepts.

5.A.5 Near-Epoch Dependent (NED) Process

Definition 5.A.6 Let $\{V_t\}_{-\infty}^{\infty}$ be a stochastic process on a probability space $(\mathcal{S}, \mathcal{F}, P)$. Define σ -subfields $\mathcal{F}_s^t = \sigma(V_s, \dots, V_t)$. A stochastic process $\{X_t\}_{-\infty}^{\infty}$ is said to be L_p -NED on $\{V_t\}_{-\infty}^{\infty}$ for $p > 0$, if for $m \geq 0$,

$$\|X_t - E(X_t | \mathcal{F}_{t-m}^{t+m})\|_p \leq d_t \nu(m),$$

where d_t is a sequence of positive constants, and $\nu(m) \rightarrow 0$ as $m \rightarrow \infty$.

We say that X_t is NED of size $-\lambda$ on the process V_t if $\nu(m) = O(m^{-\lambda-\varepsilon})$ for some $\varepsilon > 0$. In the application, V_t usually is a mixing process.

The near-epoch dependence concept is most useful due to the following theorem

Theorem 5.2 Let $\{V_t\}_{-\infty}^{\infty}$ be α -mixing of size $-a$. If $\{X_t\}_{-\infty}^{\infty}$ is an L_r -bounded zero-mean sequence and L_p -NED of size $-b$ on V_t with constant $\{d_t\}$ for $r > p \geq 1$, then $\{X_t, \mathcal{F}_{-\infty}^t\}$ is an L_p -mixingale of size $-\min\left[b, a\left(\frac{1}{p} - \frac{1}{r}\right)\right]$ with constant $c_t \ll \max\{\|X_t\|_r, d_t\}$.

Theorem 5.3 Let $\{V_t\}_{-\infty}^{\infty}$ be ϕ -mixing of size $-a$. If $\{X_t\}_{-\infty}^{\infty}$ is an L_r -bounded zero-mean sequence and L_p -NED of size $-b$ on V_t with constant $\{d_t\}$ for $r > p \geq 1$, then $\{X_t, \mathcal{F}_{-\infty}^t\}$ is an L_p -mixingale of size $-\min\left[b, a\left(1 - \frac{1}{r}\right)\right]$ with constant $c_t \ll \max\{\|X_t\|_r, d_t\}$.

5.B Functional Central Limit Theorem

The functional central limit theorem (FCLT) is a generalization of the central limit theorem (CLT) to a stochastic process; in the CLT, a sequence of distributions of

random variables converges to its limit, meanwhile, in the FCLT, a sequence of distributions of stochastic processes converges to its limit.

To see the difference, consider a sequence of stationary random variables u_t where $E(u_t) = 0$ and $E(u_t^2) = \sigma^2$:

$$u_1, u_2, \dots, u_n.$$

From them, we can construct the following sequence of random variables:

$$X_n = \frac{1}{\sqrt{n}} \sum_{t=1}^n u_t.$$

Note that for every n , X_n is a well-defined random variable, therefore it has a distribution denoted by $F_n(x)$. In the CLT, we are concerned about the limit of the sequence of the distributions. What the CLT imply is that for every x where $F_\infty(x)$ is continuous, as $n \rightarrow \infty$,

$$F_1(x), F_2(x), \dots, F_n(x), \dots \rightarrow F_\infty(x)$$

where $F_\infty(x)$ is a normal distribution.

From the sequence of u_t 's, we can also construct the following sequence of random function of $r \in [0, 1]$:

$$X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t.$$

Although it is not a simple task to define the distributions of the random functions, by abuse of notation, we can define $F_n(x)$ be a distribution of $X_n(r)$. Then, what we are concerned about with the FCLT is limit of the sequence of the distributions, $F_n(x)$. What the FCLT imply is that for every x where $F_\infty(x)$ is continuous, as $n \rightarrow \infty$,

$$F_1(x), F_2(x), \dots, F_n(x), \dots \rightarrow F_\infty(x).$$

where $F_\infty(x)$ is the distribution of the Wiener process. Formal definitions and theorems are given in the subsequent sub-sections.

For the notational convenience, we introduce a triangular stochastic array. Array notation is especially convenient when the points of a sample are subjected to scale transformations, depending on the whole sample. A typical example is $\{\{X_{nt}\}_{t=1}^n\}_{n=1}^\infty$ where $X_{nt} = \frac{X_t}{n}$. A triangular stochastic array is a doubly-indexed collection of random variables,

$$\begin{pmatrix} X_{11} & X_{21} & X_{31} & \dots \\ X_{12} & X_{22} & X_{32} & \dots \\ \vdots & \vdots & \vdots & \\ X_{1,k_1} & \vdots & \vdots & \\ & X_{2,k_2} & \vdots & \\ & & X_{3,k_3} & \\ & & & \ddots \end{pmatrix},$$

which is compactly written as $\{\{X_{mn}\}_{m=1}^{k_n}\}_{n=1}^\infty$, where k_n is an increasing integer sequence.

5.B.1 Central Limit Theorem

Since the FCLT is a generalization of the CLT, we can understand the FCLT through the comparison with the CLT. Therefore, we review the CLT first. In below, we present two versions of the CLT: one for the martingale difference sequence, and the other for NED functions of strong mixing processes.

Theorem 5.4 *Let $\{U_{nt}, \mathcal{F}_{nt}\}$ be a martingale difference array with finite unconditional variances $\{\sigma_{nt}^2\}$, and $\sum_{t=1}^n \sigma_{nt}^2 = 1$. Define $X_n = \sum_{t=1}^n U_{nt}$. If the following assumptions holds:*

1. $\sum_{t=1}^n U_{nt}^2 \xrightarrow{p} 1$

$$2. \max_{1 \leq t \leq n} |U_{nt}| \xrightarrow{p} 0$$

then, $X_n \xrightarrow{d} N(0, 1)$.

It is instructive to apply the above theorem to the i.i.d. data, which is the simplest case. Let $u_1, u_2, \dots, u_t, \dots$ be a i.i.d. sequence with $E(u_t) = 0$ and $E(u_t^2) = \sigma^2$. Also, define $U_{nt} = \frac{u_t}{\sigma\sqrt{n}}$, and $\mathcal{F}_{nt} = \sigma(u_t, u_{t-1}, \dots)$. Then, U_{nt} has the finite unconditional variance

$$\sigma_{nt}^2 = E(U_{nt}^2) = E\left(\frac{u_t^2}{\sigma^2 n}\right) = \frac{1}{n} < \infty,$$

and its sum is equal to one

$$\sum_{t=1}^n \sigma_{nt}^2 = \sum_{t=1}^n \frac{1}{n} = 1$$

Also, it can be shown that two conditions are satisfied:

1. $\sum_{t=1}^n U_{nt}^2 = \sum_{t=1}^n \frac{u_t^2}{\sigma^2 n} = \frac{1}{n} \sum_{t=1}^n \left(\frac{u_t}{\sigma}\right)^2 \xrightarrow{p} 1$ by the LLN.
2. $\max_{1 \leq t \leq n} |U_{nt}| = \max_{1 \leq t \leq n} \left| \frac{u_t}{\sigma\sqrt{n}} \right| = \left| \frac{u_t}{\sigma\sqrt{n}} \right| \xrightarrow{p} 0$. Note that the last equality holds because u_t is identically distributed, and it converges to zero because any random variable is finite.

Therefore, $X_n = \sum_{t=1}^n U_{nt} = \frac{1}{\sqrt{n}} \sum_{t=1}^n \frac{u_t}{\sigma} \xrightarrow{d} N(0, 1)$.

Theorem 5.5 Let $\{\{U_{nt}\}_{t=1}^n\}_{n=1}^\infty$ be a triangular stochastic array, let $\{\{\mathbf{V}_{nt}\}_{t=-\infty}^\infty\}_{n=1}^\infty$ be a stochastic array, and let $\mathcal{F}_{n,t-m}^{t+m} = \sigma(\mathbf{V}_{n,s}, t-m \leq s \leq t+m)$. Define $X_n = \sum_{t=1}^n U_{nt}$. If the following assumptions holds:

1. U_{nt} is $\mathcal{F}_{n,-\infty}^t/\mathcal{B}$ -measurable, with $E(U_{nt}) = 0$ and $E(X_n^2) = 1$

2. There exists a positive constant array $\{c_{nt}\}$ such that $\sup_{n,t} \|U_{nt}/c_{nt}\|_r < \infty$ for $r > 2$
3. U_{nt} is L_2 -NED of size -1 on $\{\mathbf{V}_{nt}\}$, which is α -mixing of size $-r/(r-2)$
4. $\sup_n nM_n^2 < \infty$, where $M_n = \max_{1 \leq t \leq n} \{c_{nt}\}$

then, $X_n \xrightarrow{d} N(0, 1)$

5.B.2 Functional Central Limit Theorem

In the FCLT, a sequence of distributions of stochastic processes converges to the limit. In below, we present two versions of the FCLT: one for the martingale difference sequence, and the other for NED functions of mixing processes.

Theorem 5.6 *Let $\{U_{nt}, \mathcal{F}_{nt}\}$ be a martingale difference array with finite unconditional variances $\{\sigma_{nt}^2\}$, and $\sum_{t=1}^n \sigma_{nt}^2 = 1$. Define $X_n(r) = \sum_{t=1}^{[nr]} U_{nt}$ for $r \in [0, 1]$. If the following assumptions holds:*

1. $\sum_{t=1}^n U_{nt}^2 \xrightarrow{p} 1$
2. $\max_{1 \leq t \leq n} |U_{nt}| \xrightarrow{p} 0$
3. $\lim_{n \rightarrow \infty} \sum_{t=1}^{[nr]} \sigma_{nt}^2 = r$ for all $r \in [0, 1]$

then, $X_n \Rightarrow W(r)$

Theorem 5.7 *Let $\{\{U_{nt}\}_{t=1}^{K_n}\}_{n=1}^\infty$ be a zero-mean stochastic array, $\{\{c_{nt}\}_{t=1}^{K_n}\}_{n=1}^\infty$ be an array of positive constants, and $\{K_n(r)\}_{n=1}^\infty$ be a sequence of integer-valued, right-continuous and increasing function of $r \in [0, 1]$ with $K_n(0) = 0$ for all n and $K_n(r) - K_n(s) \rightarrow \infty$ as $n \rightarrow \infty$ if $r > s$. Define $X_n^K(r) = \sum_{t=1}^{K_n(r)} U_{nt}$. If the following assumptions hold:*

1. $\sup_{n,t} \left\| \frac{U_{nt}}{c_{nt}} \right\|_r < \infty$ for $r > 2$
2. U_{nt} is L_2 -NED of size $-\gamma \in [-1, -\frac{1}{2}]$ with respect to the constants c_{nt} on an array $\{\mathbf{V}_{nt}\}$ which is α -mixing of size $-r/(r-2)$
3. $\sup_{r \in [0,1], \delta \in (0,1-r]} \left\{ \limsup_{n \rightarrow \infty} \frac{v_n^2(r, \delta)}{\delta} \right\} < \infty$, where $v_n^2(r, \delta) = \sum_{t=K_n(r)+1}^{K_n(r+\delta)} c_{nt}^2$
4. $\max_{1 \leq i \leq K_n(1)} c_{nt} = O(K_n(1)^{\gamma-1})$, where γ is defined in (2)
5. $E(X_n^K(r)^2) \rightarrow r$ as $n \rightarrow \infty$, for each $r \in [0, 1]$

then, $X_n^K(r) \Rightarrow W(r)$

In Theorem 5.7, we use a general increasing function of r , $K_n(r)$. It is instructive to consider the standard case where $K_n(r) = [nr]$ and $X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$. This case is presented in the following theorem

Theorem 5.8 *Let $\{u_t\}$ be a stochastic process with $E(u_t) = 0$, uniformly L_r -bounded, and L_2 -NED of size $-\frac{1}{2}$ on an α -mixing process of size $-r/(r-2)$ for $r > 2$. Define $X_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nr]} u_t$. If the following assumption holds:*

$$E \left(\frac{1}{\sqrt{n}} \sum_{t=1}^n u_t \right)^2 \rightarrow \sigma^2 < \infty$$

then, $X_n(r) \Rightarrow W(r)$

5.C Consistency of Bootstrap

Definition: Let P_T denote the joint probability distribution of the sample $\{x_1, x_2, \dots, x_T\}$.

Let Φ denote the space of permitted distribution functions. The bootstrap estimator

$G_T(\cdot, F_T)$ is consistent if for $\varepsilon > 0$ and $F_0 \in \Phi$,

$$\lim_{T \rightarrow \infty} P_T \left(\sup_q |G_T(q, F_T) - G_\infty(q, F_0)| > \varepsilon \right) = 0$$

5.D Hansen's (1999) Grid Bootstrap

A sample X_T of size n is generated from a distribution $G_T(x|\alpha, \nu) = P(X_T \leq x|\alpha, \nu)$ which depends on a parameter of interest $\alpha \in R$ and a nuisance parameter $\nu \in \Xi$. Denote by $\hat{\alpha}$ an estimate of α with standard error $s(\hat{\alpha})$. We assume that, for each α , there is some estimator $\hat{\nu} \in \Xi$ of the nuisance parameter ν , which may or may not depend on α . Let $S(\alpha)$ be a test statistic of the hypothesis $H_0 : \alpha_0 = \alpha$, and $F_T(x|\alpha, \nu) = P(S_T(\alpha) \leq x|\alpha, \nu)$ be a distribution function of $S(\alpha)$. Examples of $S(\alpha)$ include the nonstudentized estimate $b(\alpha) = \hat{\alpha} - \alpha$ and the t -statistic $t(\alpha) = (\hat{\alpha} - \alpha)/s(\hat{\alpha})$. The quantile function $q_T(\theta|\alpha, \nu)$ is the θ quantile of the distribution of $S_T(\alpha)$, and satisfies

$$F_T(q_T(\theta|\alpha, \nu)|\alpha, \nu) = \theta.$$

$q_T(\theta|\alpha, \nu)$ is approximated by the *bootstrap quantile function* $q_T^*(\theta|\alpha) = q_T(\theta|\alpha, \hat{\nu}(\alpha))$, which is evaluated at the estimate $\hat{\nu}(\alpha)$ and is thus random. The β -level grid-bootstrap confidence interval for α is defined as the set

$$C_g = \{\alpha \in R : q_T^*(\theta_1|\alpha) \leq S_T(\alpha) \leq q_T^*(\theta_2|\alpha)\}$$

where $\theta_1 = 1 - (1 - \beta)/2$ and $\theta_2 = (1 - \beta)/2$; so $\beta = \theta_2 - \theta_1$.

In order to calculate C_g , we need to estimate the bootstrap quantile functions $q_T^*(\theta|\alpha)$, which are generally unknown, by simulation as follows. For a given α , let $G_T^*(x|\alpha) = G_T(x|\alpha, \hat{\nu}(\alpha))$ be the bootstrap distribution of the sample.

1. Generate random samples X_T^* from $G_T^*(x|\alpha)$ by simulation.
2. Using X_T^* , calculate the test statistic $S_T^*(\alpha)$.
3. Repeat steps 1 and 2 B times.

4. Sort the B simulated test statistics $S_T^*(\alpha)$. The $100\theta\%$ order statistic $\hat{q}_T^*(\theta|\alpha)$ is the simulation estimate of $q_T^*(\theta|\alpha)$ as a function of α .
5. Pick a grid $A_G = [\alpha_1, \dots, \alpha_G]$, and calculate $\hat{q}_T^*(\theta|\alpha)$ at each $\alpha \in A_G$ by simulation.
6. For a given α , smooth the estimated function $\hat{q}_T^*(\theta|\alpha)$ using the kernel estimate:

$$\tilde{q}_n^*(\theta|\alpha) = \frac{\sum_{j=1}^G K\left(\frac{\alpha-\alpha_j}{\gamma}\right) \hat{q}_n^*(\theta|\alpha_j)}{\sum_{j=1}^G K\left(\frac{\alpha-\alpha_j}{\gamma}\right)}$$

where $K(z)$ is the Epanechnikov kernel $K(z) = \frac{3}{4}(1-z^2)I(|z| \leq 1)$, and γ is a bandwidth chosen by least-square cross-validation.

5.E Monte Carlo Methods with GAUSS

This appendix explains how Monte Carlo methods explained in this chapter are implemented with GAUSS, that is explained in Appendix A. The concepts and programs are similar in other computer languages such as MATLAB.

5.E.1 Random Number Generators

Most programs offer random number generators for the uniform distribution and the standard normal distribution. For example,

```
y=RNDN(r,c);
```

in GAUSS generates $r \times c$ values that appear to be a realization of independent standard normal random variables that will be stored in an $r \times c$ matrix. The starting seed for RNDN can be given by a statement

```
RNDSEED n;
```

where the value of the seed n must be in the range $0 < n < 2^{31} - 1$.

One can produce random numbers with other distributions by transforming generated random numbers. The following examples are some of the transformations that are often used.

Example 5.E.1 A χ^2 random variable with d degrees of freedom can be created from d independent random variables with the standard normal distribution. If $e_i \sim N(0, 1)$, and if e_i is independent from e_j for $j \neq i$, then $\sum_{i=1}^d e_i^2$ follows the χ^2 distribution with d degrees of freedom. ■

For example, in GAUSS one can generate a $T \times 1$ vector with values that appear to be a realization of an i.i.d. $\{x\}_{t=1}^T$ of random variables with the χ^2 distribution with d degrees of freedom by the following program:

```
e=RNDN(T,d);
x=sumc((e^2)');
```

Example 5.E.2 A random variable that follows the Student's t distribution with d degrees of freedom can be generated from $d + 1$ independent random variables with the standard normal distribution. If $e_i \sim N(0, 1)$, and if e_i is independent from e_j for $j \neq i$, then $x = e_1 / \sqrt{\sum_{i=2}^{d+1} e_i^2 / d}$ follows the t distribution with d degrees of freedom. ■

For example, in GAUSS one can generate a $T \times 1$ vector with values that appear to be a realization of an i.i.d. $\{x\}_{t=1}^T$ of random variables with the t distribution with d degrees of freedom by the following program:

```
e=RNDN(T,d+1);
c=sumc((e[.,2:d+1]^2)');
x=e[.,1]./sqrt(c/d);
```

Example 5.E.3 A K -dimensional random vector which follows $N(\mathbf{0}, \Psi)$ for any positive definite covariance matrix Ψ can be generated from K independent random variables with the standard normal distribution. Let $\Psi = \mathbf{P}\mathbf{P}'$ be the Cholesky decomposition of Ψ , in which \mathbf{P} is a lower triangular matrix. If $e_i \sim N(0, 1)$, and if e_i is independent from e_j for $j \neq i$, then $\mathbf{X} = \mathbf{P}\mathbf{e} \sim N(\mathbf{0}, \Psi)$ where $\mathbf{e} = (e_1, e_2, \dots, e_K)'$. ■

For example, in GAUSS one can generate a $T \times K$ matrix with values that appear to be a realization of an i.i.d. $\{\mathbf{X}_t\}_{t=1}^T$ of K -dimensional random vectors with the $N(0, C)$ distribution with the following program provided that the matrix C is already defined.

```
e=RNDN(T,K);
P=chol(C)';
x=eP;
```

Note that the Cholesky decomposition in GAUSS gives an upper triangular matrix. Thus, the above program transposes the matrix to a lower triangular matrix.

5.E.2 Estimators

5.E.3 A Pitfall in Monte Carlo Simulations

Common mistakes are made by many graduate students when they first use Monte Carlo simulations. These mistakes happen when they repeatedly use a random number generator to conduct simulations. These mistakes are caused by updating seeds arbitrarily in the middle of a simulation. Recall that once the starting seed is given, a random number generator automatically updates the seed whenever it creates a number. The way the seed is updated depends on the program.

The following example illustrates common mistakes in a simple form:

Example 5.E.4 *A Monte Carlo Program with a Common Mistake (I)*

```
ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss;
    y=RNDN(50,1);
    m=meanc(y);
    vecm[i]=m;
    i=i+1;
endo;
```

In this example, the programmer is trying to create 100 samples of the sample mean of a standard normal random variable y when the sample size is 50. However, exactly the same data are generated 100 times because the same starting seed is given for each replication inside the do-loop. This mistake is innocuous because it is easy to detect. The following program contains a mistake which is harder to detect:

Example 5.E.5 *A Monte Carlo Program with a Common Mistake (II)*

```
ss=3937841;
i=1;
vecm=zeros(100,1);
do until i>100;
    RNDSEED ss+i;
    y=RNDN(50,1);
    m=meanc(y);
    vecm[i]=m;
    i=i+1;
endo;
```

The problem is that the seed is updated in an arbitrary way in each sample by giving a different starting seed. There is no guarantee that one sample is independent from the others. A correct program would put the RNDSEED statement before the do loop. For example, the RNDSEED statement inside the do loop should be removed and the statement

```
RNDSEED ss;
```

can be added after the first line.

In Monte Carlo simulations, it is also important to control the starting seed so that the simulation results can be replicated. When you publish Monte Carlo results, it is advisable to put enough information in the publication so that others can exactly replicate the results.¹³ At the very least, a record of the information should be kept. If no RNDSEED statement is given before the RNDN command is used, GAUSS will take the starting seed from the computer clock. Then there is no way to exactly replicate these Monte Carlo results.

5.E.4 An Example Program

This section describes basic Monte Carlo methods with an example program. In the following example, the sample mean is calculated as an estimator for the expected value of X_t , $E(X_t)$, where $X_t = \mu + e_t$ and e_t is drawn from the t distribution with 3 degrees of freedom. The t distribution with 3 degrees of freedom has thick tails and large????? , outlying values have high probability. Hence the t distribution is often considered a better distribution to describe some financial variables. Because X_t is not normally distributed, the standard theory for the exact finite sample properties cannot be applied. The example is concerned with the t test of the null hypothesis that $\mu = 0$. Because a random variable with the t distribution with 3 degrees of freedom has zero mean and a finite second moment, asymptotic theory predicts that the t test statistic of the sample mean divided by the estimated standard error approximately follows the standard normal distribution.

Masao
needs to
check this!

¹³This information is also relevant because different computer specifications and different versions of the program (such as GAUSS) can produce different results.

Example 5.E.6 *The program.*

```

@MCMEAN.PRG @ Monte Carlo Program for the sample mean@
@This example program is a GAUSS program to calculate
the empirical size and power of the t-test for  $H_0: E(X)=0$ ,
where X follows t-distribution with 3 degrees of freedom.
The power is calculate for the case when  $E(X)=0.2$ . @

RNDSEED 382974;
output file=mc.out reset;
tend=25; @the sample size@
nor=1000; @the number of replications@
df=3; @ d.f. for the t-distribution of X@
i=1;
tn=zeros(nor,1); @used to store t-values under H0@
ta=zeros(nor,1); @used to store t-values under H1@
do until i>nor;
  nrv=RNDN(tend,df+1); @normal r.v.'s@
  crv=nrv[.,2:df+1]^2; @chi square r.v.'s@
  x0=nrv[.,1]/sqrt(sumc(crv)/df); @t distribution: used under H0@
  x1=x0+0.2; @used for H1@
  mx0=meanc(x0);
  mx1=meanc(x1);
  sighat0=sqrt((x0-mx0)'(x0-mx0)/(tend-1)); @simgahat under H0@
  sighat1=sqrt((x1-mx1)'(x1-mx1)/(tend-1)); @sigmahat under H1@
  tn[i]=meanc(x0)*sqrt(tend)/sighat0; @t-value under H0@
  ta[i]=meanc(x1)*sqrt(tend)/sighat1; @t-value under H1@
  i=i+1;
endo;
? "***** When H0 is true *****";
? "The estimated size with the nominal critical value";
? meanc(abs(tn).>1.96);
? "The estimated true 5-percent critical value";
sorttn=sortc(abs(tn),1);
etcv=sorttn[int(nor*0.95)];
? etcv;
? "***** When H1 is true *****";
? "The estimated power with the nominal critical value";
? meanc(abs(ta).>1.96);
? "The estimated size corrected power";
? meanc(abs(ta).>etcv);

```

```
output off;
```

Some features of the example are important. Before the do-loop of the replications, the program set up an output file by

```
output file=mc.out;
```

Then to avoid the common mistake explained in 5.E.3, it makes the RNDNSEED statement before the do-loop.

It is a good idea to minimize the content inside the do-loop to speed up replications. Everything that can be done outside the do-loop should be done there. For example, the program defines variables to store the test results:

```
tn=zeros(nor,1);
ta=zeros(nor,1);
```

In GAUSS, the do-loop can be set up as follows:

```
i=1;
do until i>250;
... (Program for each replication)
i=i+1;
endo;
```

After the do-loop, the program calculates characteristics of the generated distributions of test statistics under the null hypothesis and the alternative hypothesis such as the frequency of rejecting the null with the nominal critical value.

Exercises

5.1 Show that all conditions of Gordin's Central Limit Theorem are satisfied for e_t in Example 5.1.

5.2 Show that all conditions of Gordin and Hansen's Central Limit Theorem are satisfied for \mathbf{f}_t in Example 5.2.

5.3 Let $e_t = \Psi(L)u_t = \Psi_0 u_t + \Psi_1 u_{t-1} + \dots$ be an MA representation. What is the long-run variance of $f_t = (1 - L)e_t$?

5.4 Explain what it means to say that “a test under-rejects in small samples” (or “a test is conservative”). When a test is conservative, which is greater, the true critical value or the nominal critical value?

5.5 Consider the linear model

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + e_t.$$

where \mathbf{x}_t is a k -dimensional vector.

Let \mathbf{z}_t be a $k \times 1$ vector of instrumental variables. We will adopt the following set of assumptions:

(A1) $(e_t, \mathbf{x}_t, \mathbf{z}_t)_{t=1}^{\infty}$ is a stationary and ergodic stochastic process.

(A2) $\mathbf{z}_t e_t$ have finite second moments.

(A3) $E(e_t^2 | \mathbf{z}_t) = \sigma^2$, where σ is a constant.

(A4) $E(e_t | \mathbf{I}_t) = 0$ for a sequence of information sets $(\mathbf{I}_t)_{t=1}^{\infty}$ which is increasing (i.e., $\mathbf{I}_t \subset \mathbf{I}_{t+1}$), \mathbf{z}_t and \mathbf{x}_t are in \mathbf{I}_t , and y_t is in \mathbf{I}_{t+1} .

(A5) $E(\mathbf{z}_t \mathbf{x}_t')$ is nonsingular.

Note that $E(e_t) = 0$ is implied by (A4) if \mathbf{z}_t includes a constant.

Note that many rational expectations models imply **(A4)**. For the following problems, prove the asymptotic properties of the instrumental variable (IV) estimator, \mathbf{b}_{IV} , for $\boldsymbol{\beta}$ under **(A1)**-**(A5)**. Use a central limit theorem and a strong law of large numbers given in this chapter, and indicate which ones you are using and where you are using them in your proof.

- (a) Express the IV estimator \mathbf{b}_{IV} in terms of $\mathbf{z}_t, \mathbf{x}_t$, and $y_t (t = 1, \dots, T)$ when $\sum_{t=1}^T \mathbf{z}_t \mathbf{x}_t'$ is nonsingular.
- (b) Let $\mathbf{g}_t = \mathbf{z}_t e_t$. Prove that \mathbf{g}_t is a martingale difference sequence.
- (c) Prove that the IV estimator is consistent under **(A1)**-**(A5)**.
- (d) Prove that the IV estimator is asymptotically normally distributed. Derive the formula of the covariance matrix of the asymptotic distribution.
- (e) Explain what happens if y_t is in I_{t+2} in **(A4)**.

5.6 Consider the linear model

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + \epsilon_t,$$

where \mathbf{x}_t is a k -dimensional vector. Following Hayashi (2000), suppose that this model satisfies the classical linear regression model assumptions for any sample size (n) as follows:

- (A1)** Linearity: $y_t = \mathbf{x}_t' \boldsymbol{\beta} + e_t$.
- (A2)** Ergodic stationarity: $\{y_t, \mathbf{x}_t\}$ is jointly stationary and ergodic.
- (A3)** Predetermined regressors: $E(e_t \mathbf{x}_t) = \mathbf{0}$.

(A4) Rank condition: $E(\mathbf{x}_t \mathbf{x}_t')$ is nonsingular (and hence finite).

(A5) $\mathbf{x}_t e_t$ is a martingale difference sequence with finite second moments.

(A6) Finite fourth moments for regressors: $E[(x_{it} x_{jt})^2]$ exists and finite for all i, j ($= 1, 2, \dots, k$).

(A7) Conditional homoskedasticity: $E(e_t^2 | \mathbf{x}_t) = \sigma^2 > 0$.

Further, assume that e_t is normally distributed conditional on \mathbf{X} , where \mathbf{X} is an $n \times k$ matrix with \mathbf{x}_t' in its t -th row. Let

$$t_k = \frac{b_k - \bar{\beta}_k}{SE(b_k)} = \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 [(\mathbf{X}'\mathbf{X})^{-1}]_{kk}}}$$

be the t statistic for the null hypothesis $\beta_k = \bar{\beta}_k$.

(a) Prove that t_k converges in distribution to the standard normal distribution as the sample size goes to infinity. You do not have to prove that s^2 is consistent σ^2 for this question. You can assume that s^2 is consistent.

(b) Based on the asymptotic result in (a), suppose that you set the nominal size to be 5 percent and reject the null hypothesis when $|t_k|$ is greater than 1.96. Does this test overreject or underreject. How do you know? Suppose that $k = 3$. Is the actual size larger than 10 percent when $n = 4$. What if $n = 8, 9, 10, 11$? Explain.

5.7 Consider the linear model

$$(5.E.1) \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Let $k \times 1$ matrix \mathbf{x}'_t be the t -th row of the regressor matrix \mathbf{X} . The model (5.E.1) can be written as

$$(5.E.2) \quad y_t = \mathbf{x}'_t \boldsymbol{\beta} + e_t$$

We will adopt the following set of assumptions:

(A1) $(e_t, \mathbf{x}_t)_{i=t}^{\infty}$ are independent and identically distributed (i.i.d.) random vectors.

(A2) \mathbf{x}_t and e_t have finite second moments.

(A3) $E(e_t^2 | \mathbf{x}_t) = \sigma^2$ which is a constant.

(A4) $E(\mathbf{x}_t e_t) = 0$ for all $t \geq 1$

(A5) $E(\mathbf{x}_t \mathbf{x}'_t)$ is nonsingular.

Note that $E(e_t) = 0$ is implied by (A4) if \mathbf{x}_t includes a constant.

Consider the model (5.E.1) with $k = 1$. Assume that x_t follows $N(0,1)$. Assume that x_t and e_t are independent. Under the null hypothesis H_0 , the true value of β is 0, so that $y_t = e_t$.

Consider the standard t statistic,

$$(5.E.3) \quad t_1 = \frac{b - \beta}{\hat{\sigma}_1 \sqrt{\mathbf{X}'\mathbf{X}}^{-1}}$$

where $\hat{\sigma}_1^2 = (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b)/(n - k)$. Consider another version of the t statistic

$$(5.E.4) \quad t_2 = \frac{b - \beta}{\hat{\sigma}_2 \sqrt{\mathbf{X}'\mathbf{X}}^{-1}}$$

where $\hat{\sigma}_2^2 = (\mathbf{Y} - \mathbf{X}b)'(\mathbf{Y} - \mathbf{X}b)/n$. Note that both t_1 and t_2 converge in distribution to a random variable with the standard normal distribution.

Consider two alternative assumptions for e_t .

(A6) e_t follows the t distribution with 4 degrees of freedom.

(A6') e_t follows the standard normal distribution.

Note that Assumptions 1.1 - 1.5 are satisfied with **(A6')**, so that t_1 has the exact t distribution with $n - k$ degrees of freedom.

Using GAUSS, conduct a Monte Carlo simulation with the sample size of 26 and 500 replications under Assumption **(A6)**.

- (a) Use the t_1 in (5.E.3) to estimate
- (i) the true size of the t test for $H_0 : \beta = 0$ based on the nominal significance level of 5% and the nominal critical value based on the standard normal distribution are used.
 - (ii) the true size of the t test for $H_0 : \beta = 0$ based on the nominal significance level of 5% and the nominal critical value based on the t distribution with 25 degrees of freedom.
 - (iii) the true critical value of the t test for the 5% significance level,
 - (iv) the power of the test at $\beta = 0.15$ based on the nominal critical value,
 - (v) the size corrected power of the test.
- (b) Use the t_2 in (5.E.4) and repeat the exercises (a) – (e).

For the starting seed, use 3648xxxx, where xxxx is your birth date, such as 0912 for September 12. Submit your program and output. For each t ratio, discuss whether it is better to use the standard distribution or the t distribution critical values for this application. Also discuss whether t_1 or t_2 is better for this application.

References

- ANDREWS, D. W. K. (1993): "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models," *Econometrica*, 61(1), 139–165.
- APOSTOL, T. M. (1974): *Mathematical Analysis*. Addison-Wesley, Reading, Massachusetts.

- ATHREYA, K. B. (1987): “Bootstrap of the Mean in the Infinite Variance Case,” *Annals of Statistics*, 15(2), 724–731.
- BASAWA, I. V., A. K. MALLIK, W. P. MCCORMICK, J. H. REEVES, AND R. L. TAYLOR (1991): “Bootstrapping Unstable Autoregressive Processes,” *Annals of Statistics*, 19(2), 1098–1101.
- BILLINGSLEY, P. (1961): “The Lindeberg-Levy Theorem for Martingales,” *Proceeding of the American Mathematical Society*, 12, 788–792.
- (1986): *Probability and Measure*. Wiley, New York.
- CHAN, N. H., AND C. Z. WEI (1987): “Asymptotic Inference for Nearly Nonstationary AR(1) Processes,” *Annals of Statistics*, 15(3), 1050–1063.
- CHOI, C.-Y., AND M. OGAKI (1999): “The Gauss-Markov Theorem for Cointegrating and Spurious Regressions,” Manuscript.
- EFRON, B. (1979): “Bootstrap Methods: Another Look at the Jackknife,” *Annals of Statistics*, 7(1), 1–26.
- GORDIN, M. I. (1969): “The Central Limit Theorem for Stationary Processes,” *Soviet Mathematics-Doklady*, 10, 1174–1176.
- HANSEN, B. E. (1999): “The Grid Bootstrap and the Autoregressive Model,” *Review of Economics and Statistics*, 81(4), 594–607.
- HANSEN, L. P. (1985): “A Method for Calculating Bounds on the Asymptotic Covariance Matrices of Generalized Method of Moments Estimators,” *Journal of Econometrics*, 30, 203–238.
- HAYASHI, F. (2000): *Econometrics*. Princeton University Press, Princeton.
- HOROWITZ, J. L. (2001): “The Bootstrap in Econometrics,” in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. V, chap. 52, pp. 3159–3228. Elsevier Science Publishers.
- JEONG, J., AND G. S. MADDALA (1993): “A Perspective on Application of Bootstrap Methods in Econometrics,” in *Handbook of Statistics*, ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod, vol. 11, pp. 573–610. Elsevier Science Publishers.
- JUDGE, G. G., W. E. GRIFFITHS, R. C. HILL, H. LÜTKEPOHL, AND T. LEE (1985): *The Theory and Practice of Econometrics*. Wiley, New York, 2nd edn.
- LOEVE, M. (1978): *Probability Theory*. Springer-Verlag, New York, 4th edn.
- NELSON, C. R., AND R. STARTZ (1990): “The Distribution of the Instrumental Variables Estimator and Its t-Ratio When the Instrument is a Poor One,” *Journal of Business*, 63, S125–S140.
- PHILLIPS, P. C. B. (1987): “Towards a Unified Asymptotic Theory for Autoregression,” *Biometrika*, 74(3), 535–547.