Chapter 2

STOCHASTIC PROCESSES

In most macroeconomic models, expectations conditional on information sets are used to model the forecasting conducted by economic agents. Economic agents typically observe stochastic processes of random variables (collections of random variables indexed by time) to form their information sets. This chapter defines the concepts of conditional expectations and information sets for the case of a finite number of elements in the probability space.[1]

2.1 Review of Probability Theory

Since the probability statements made in asymptotic theory involve infinitely many random variables instead of just one random variable, it is important to understand basic concepts in probability theory. Thus, we first review those basic concepts.

Imagine that we are interested in making probability statements about a set of the states of the world (or a probability space), which we denote by S. For the purpose of understanding concepts, nothing is lost by assuming that there is a finite number of states of the world. Hence we adopt the simplifying assumption that S

[1]For the general probability space, these concepts are defined with measure theory (see Appendix 2.A). For our purpose, it is not necessary for the reader to understand measure theory.

consists of N possible states: $\mathrm{S} = \{s_1, \cdots, s_N\}$. We assign a probability $\pi_i = Pr(s_i)$ to s_i, depending on how likely s_i is to occur. It is assumed that $\sum_{i=1}^{N} \pi_i = 1$ and $0 \leq \pi_i \leq 1$ for all i. Note that we can now assign a probability to all subsets of S. For example, let Λ be $\{s_1, s_2\}$. Then the probability that the true s is in Λ is denoted by $Pr(s \in \Lambda)$, where $Pr(s \in \Lambda) = \pi_1 + \pi_2$.

Example 2.1 The state of the world consists of s_1: it rains tomorrow, and s_2: it does not rain tomorrow. According to a weather forecast, $\pi_1 = 0.8$ and $\pi_2 = 0.2$. ∎

A *random variable* assigns a real value to each element s in S (that is, it is a real valued function on S). Let $X(s)$ be a random variable (we will often omit the arguments s). For a real value x, the *distribution function*, $F(x)$, of the random variable is defined by $F(x) = Pr\{s : X(s) \leq x\}$. A random variable is assigned an expected value or mean value

$$E(X) = \sum_{i=1}^{N} X(s_i)\pi_i. \tag{2.1}$$

Example 2.2 Continuing Example 2.1, let $X(s)$ be the profit of an umbrella seller in terms of dollars with $X(s_1) = 100$ and $X(s_2) = 10$. Then $E(X) = 100 \times 0.8 + 10 \times 0.2 = 82$. The distribution function $F(x)$ is given by $F(x) = 0$ for $x < 10$, $F(x) = 0.2$ for $10 \leq x < 100$, and $F(x) = 1$ for $x \geq 100$. ∎

A random vector is a vector of random variables defined on the set of states. For a k-dimensional random vector $\mathbf{X}(s) = (X_1(s), \cdots, X_k(s))'$, the joint distribution function F is defined by

$$F(x_1, \cdots, x_k) = Pr[X_1 \leq x_1, \cdots, X_k \leq x_k]. \tag{2.2}$$

2.2 Stochastic Processes

A collection of random variables indexed by time is called a *stochastic process* or a *time series*. Let $X_t(s)$ be a random variable, then a collection $\{X_t : X_0(s), X_1(s), X_2(s), \cdots\}$ is a univariate stochastic process. It is sometimes more convenient to consider a stochastic process that starts from the infinite past, $\{\cdots, X_{-2}(s), X_{-1}(s), X_0(s), X_1(s), X_2(s), \cdots\}$. In general, $\{X_t(s) : t \in \mathrm{A}\}$ for any set A is a stochastic process. If A is a set of integers, then time is *discrete*. It is also possible to consider a *continuous* time stochastic process for which the time index takes any real value. For example, $\{X_t(s) : t$ is a nonnegative real number$\}$. Here, if we take X_t as a random vector rather than a random variable, then it is a vector stochastic process. When we observe a sample of size T of a random variable X or a random vector $\mathbf{X} : \{X_1, \cdots, X_T\}$, it is considered a particular realization of a part of the stochastic process.

Note that once s is determined, the complete history of the stochastic process becomes known. For asymptotic theory, it is usually easier to think about the stochastic nature of economic variables this way rather than the alternative, which is to consider a probability space for each period based on independent disturbances.

In a sense, the stochastic process modeled in this manner is deterministic because everything is determined at the beginning of the world when s is determined. However, this does not mean that there is no uncertainty to economic agents because they do not learn s until the end of the world. In order to illustrate this, let us consider the following example:

Example 2.3 Imagine an economy with three periods and six states of the world. The world begins in period 0. We observe two variables, aggregate output (Y_t) and

the interest rate (i_t), in period 1 and period 2. The world ends in period 2. In each period, Y_t can take two values, 150 and 300, and i_t can take two values, 5 and 10. We assume that i_2 is equal to i_1 in all states of the world, and that the $i_1 = 5$ in all states in which $Y_1 = 150$. The six states of the world can be described by the triplet, $[Y_1, i_1, Y_2]$.

The six states of the world are, $s_1 = [300, 10, 300]$, $s_2 = [300, 10, 150]$, $s_3 = [300, 5, 300]$, $s_4 = [300, 5, 150]$, $s_5 = [150, 5, 300]$, and $s_6 = [150, 5, 150]$. To illustrate, s_1 means the economy is in a boom (higher output level) with a high interest rate in period 1, and is in a boom in period 2. In period 0, the economic agents assign a probability to each state: $\pi_1 = 0.20$, $\pi_2 = 0.10$, $\pi_3 = 0.15$, $\pi_4 = 0.05$, $\pi_5 = 0.15$, and $\pi_6 = 0.35$. Unconditional expected values are taken with these probabilities. ∎

In this example, let $\mathbf{X}_t(s) = [Y_t(s), i_t(s)]$. Then $[\mathbf{X}_1(s), \mathbf{X}_2(s)]$ is a stochastic process. The whole history of the process is determined at the beginning of the world when s is chosen, and the agents learn which state of the world they are in at the end of the world in period 2. In period 1, however, the agents only have partial information as to which state of the world is true. For example, if $Y_1 = 300$ and $i_1 = 5$, the agents learn that they are in either s_3 or s_4, but cannot tell which one they are in until they observe Y_2 in period 2.

2.3 Conditional Expectations

Economic agents use available information to learn the true state of the world and make forecasts of future economic variables. This forecasting process can be modeled using conditional expectations.

Information can be modeled as a partition of S into mutually exclusive subsets:

$\mathcal{F} = \{\Lambda_1, \cdots, \Lambda_M\}$ where $\Lambda_1 \cup \cdots \cup \Lambda_M = \mathrm{S}$, and $\Lambda_j \cap \Lambda_k = \emptyset$ if $j \neq k$. For example, information $\mathcal{F}$ consists of two subsets: $\mathcal{F} = \{\Lambda_1, \Lambda_2\}$. Here $\Lambda_1 = \{s_1, \cdots, s_M\}$, and $\Lambda_2 = \{s_{M+1}, \cdots, s_N\}$. The information represented by $\mathcal{F}$ tells us which Λ contains the true s, but no further information is given by $\mathcal{F}$.

In this situation, once agents obtain the information represented by $\mathcal{F}$, then the agents know which subset contains the true s, and they can assign a probability of zero to all elements in the other subset. There is no reason to change the ratios of probabilities assigned to the elements in the subset containing the true s. Nonetheless, the absolute level of each probability should be increased, so that the probabilities add up to one. The probability conditional on the information that the true s is in Λ_j is denoted by $Pr\{s_i | s \in \Lambda_j\}$. The considerations given above lead to the following definition of conditional probability:

$$Pr\{s_i | s \in \Lambda_j\} = \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}}, \tag{2.3}$$

when s_i is in Λ_j. Here each probability is scaled by the probability of the subset containing the true s, so that the probabilities add up to one.

We use conditional probability to define the *conditional expectation*. The expectation of a random variable Y conditional on the information that the true s is in Λ_j is

$$E(Y | s \in \Lambda_j) = \sum_{s \in \Lambda_j} Y(s) \frac{Pr\{s_i\}}{Pr\{s \in \Lambda_j\}}, \tag{2.4}$$

where the summation is taken over all s in Λ_j.

It is convenient to view the conditional expectation as a random variable. For this purpose, the conditional expectation needs to be defined over all s in S, not just for s in a particular Λ_j. Given each s, we first find out which Λ contains s.

When Λ_j contains s, the expected value of Y conditional on $\mathcal{F}$ for s is given by $E(Y|\mathcal{F})(s) = E(Y|s \in \Lambda_j)$.

Instead of a partition, we can use a random variable or a random vector to describe information. Consider information represented by a partition $\mathcal{F} = \{\Lambda_1, \cdots, \Lambda_M\}$. Consider the set I, which consists of all random variables that take the same value for all elements in each Λ_j : I $= \{X(s) : X(s_i) = X(s_k)$ if $s_i \in \Lambda_j$ and $s_k \in \Lambda_j$ for all $i, j, k\}$. Then the information set I represents the same information as $\mathcal{F}$ does. A random variable X is said to be in this information set, when $X(s_i) = X(s_k)$ if both s_i and s_k are in the same Λ_j.[2] A random vector $\mathbf{X}$ is said to be in this information set when each element of $\mathbf{X}$ is in the information set.

If X is in the information set I, and if X takes on different values for all different Λ ($X(s_i) \neq X(s_k)$ when s_i and s_k are not in the same Λ), then we say that the random variable X generates the information set I. If a random vector $\mathbf{X}$ is in I, and if at least one element of $\mathbf{X}$ takes on different values for different Λ, then the random vector $\mathbf{X}$ is said to generate the information set I. When a random variable X or a random vector $\mathbf{X}$ generates the information set I, which represents the same information as a partition $\mathcal{F}$, we define $E(Y|\text{I})$ as $E(Y|\mathcal{F})$. If I is generated by X, we define $E(Y|X) = E(Y|\text{I})$; and if I is generated by a random vector $\mathbf{X}$, we define $E(Y|\mathbf{X}) = E(Y|\text{I})$. It should be noted that $E(Y|\text{I})$ is in the information set I.

Example 2.4 Continuing Example 2.3, let I be the information set generated by $X_1 = (Y_1, i_1)$, and let $\mathcal{F}$ be the partition that represents the same information as I. Then $\mathcal{F} = \{\Lambda_1, \Lambda_2, \Lambda_3\}$, where $\Lambda_1 = \{s_1, s_2\}$, $\Lambda_2 = \{s_3, s_4\}$, and $\Lambda_3 = \{s_5, s_6\}$.

[2] In the terminology of probability theory, we consider a set of all possible unions of Λ's in $\mathcal{F}$ plus the null set. This set of subsets of S is called a σ-field, and used to describe information. When a random variable X is in the information set I, we say that the random variable is measurable with respect to this σ-field.

Using (2.3), $Pr(s_1|s \in \Lambda_1) = \frac{0.20}{0.20+0.10} = \frac{2}{3}$ and $Pr(s_2|s \in \Lambda_1) = \frac{0.10}{0.20+0.10} = \frac{1}{3}$. Hence $E(Y_2|s \in \Lambda_1) = 300 \times \frac{2}{3} + 150 \times \frac{1}{3} = 250$. Similarly, $Pr(s_3|s \in \Lambda_2) = \frac{3}{4}$, $Pr(s_4|s \in \Lambda_2) = \frac{1}{4}$, $Pr(s_5|s \in \Lambda_3) = \frac{3}{10}$, $Pr(s_6|s \in \Lambda_3) = \frac{7}{10}$, $E(Y_2|s \in \Lambda_2) = 262.5$, and $E(Y_2|s \in \Lambda_3) = 195$. Hence the random variable $E(Y_2|\mathrm{I})$ is given by

$$E(Y_2|\mathrm{I})(s) = \begin{cases} 250 & \text{if} \quad s \in \Lambda_1 \\ 262.5 & \text{if} \quad s \in \Lambda_2 \\ 195 & \text{if} \quad s \in \Lambda_3 \end{cases} . \tag{2.5}$$

∎

Example 2.5 Continuing Example 2.4, consider the information set J which is generated by Y_1. Then J is a smaller information set than I in the sense that J $\subset$ I. Similar computations as those in Example 2.4 yield

$$E(Y_2|\mathrm{J})(s) = \begin{cases} 255 & \text{if} \quad s \in \{s_1, s_2, s_3, s_4\} \\ 195 & \text{if} \quad s \in \{s_5, s_6\} \end{cases} . \tag{2.6}$$

∎

Two properties of conditional expectations are very important in macroeconomics.

Proposition 2.1 (Properties of Conditional Expectations)

(a) If a random variable Z is in the information set I, then

$$E(ZY|\mathrm{I}) = ZE(Y|\mathrm{I}) \tag{2.7}$$

for any random variables Y with finite $E(|Y|)$, assuming that $E(|ZY|)$ is finite.

(b) *The Law of Iterated Expectations:* If the information set J is smaller than the information set I (J $\subset$ I), then

$$E(Y|\mathrm{J}) = E[E(Y|\mathrm{I})|\mathrm{J}] \tag{2.8}$$

for any random variable Y with finite $E(|Y|)$. ∎

Expectation can be viewed as a special case of conditional expectation in which the information set consists of constants. Since a constant is a random variable which takes the same value for all states of the world, any information set includes all constants. Therefore, the Law of Iterated Expectations implies

$$E(Y) = E[E(Y|\mathrm{I})]. \tag{2.9}$$

When we wish to emphasize the difference between expectations and conditional expectations, expectations are called *unconditional expectations*. Relation (2.9) states that an unconditional expected value of a random variable Y can be computed as an unconditional expected value of the expectation of the random variable conditional on any information set. For a proof of Proposition 2.1 in the general case, see, e.g., Billingsley (1986, Theorem 34.3 and Theorem 34.4).

2.4 Stationary Stochastic Processes

A stochastic process $\{\cdots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \cdots\}$ is *strictly stationary* if the joint distribution function of $(\mathbf{X}_t, \mathbf{X}_{t+1}, \cdots, \mathbf{X}_{t+h})$ is the same for all $t = 0, \pm 1, \pm 2, \cdots$ and all $h = 0, 1, 2, \cdots$. A stochastic process $\{\cdots, \mathbf{X}_{-1}, \mathbf{X}_0, \mathbf{X}_1, \cdots\}$ is *covariance stationary* (or *weakly stationary*) if $\mathbf{X}_t$ has finite second moments ($E(\mathbf{X}_t\mathbf{X}_t') < \infty$) and if $E(\mathbf{X}_t)$ and $E(\mathbf{X}_t\mathbf{X}_{t-h}')$ do not depend on the date t for all $t = 0, \pm 1, \pm 2, \cdots$ and all $h = 0, 1, 2, \cdots$.

Because all moments are computed from distribution functions, if $\mathbf{X}_t$ is strictly stationary and has finite second moments, then it is also covariance stationary. If $\mathbf{X}_t$ is covariance stationary, then its mean $E(\mathbf{X}_t)$ and its h-th *autocovariance* $\mathbf{\Phi}(h) = E[(\mathbf{X}_t - E(\mathbf{X}_t))(\mathbf{X}_{t-h} - E(\mathbf{X}_{t-h})')] = E(\mathbf{X}_t\mathbf{X}_{t-h}') - E(\mathbf{X}_t)E(\mathbf{X}_{t-h}')$ does not depend on date t.

Proposition 2.2 If a k-dimensional vector stochastic process $\mathbf{X}_t$ is strictly stationary, and if a continuous function $f(\cdot) : \mathrm{R}^k \longmapsto \mathrm{R}^p$ does not depend on date t, then $f(\mathbf{X}_t)$ is also strictly stationary.[3] ∎

This follows from the fact that the distribution function of $f(\mathbf{X}_t)$, $f(\mathbf{X}_{t+1})$, $\cdots$, $f(\mathbf{X}_{t+h})$ is determined by f and the joint distributions of $\mathbf{X}_t$, $\mathbf{X}_{t+1}$, $\cdots$, $\mathbf{X}_{t+h}$ (see Appendix 2.A). Proposition 2.2 will be used frequently to derive the cointegrating properties of economic variables from economic models in Chapter 15.

The next proposition is for covariance stationary processes.

Proposition 2.3 If a k-dimensional vector stochastic process $\mathbf{X}_t$ is covariance stationary, and if a linear function $f(\cdot) : \mathrm{R}^k \longmapsto \mathrm{R}^p$ does not depend on date t, then $f(\mathbf{X}_t)$ is also covariance stationary. ∎

This proposition is true because $f(\mathbf{X}_t)$ has finite second moments, and the first and second moments of $f(\mathbf{X}_t)$ do not depend on date t. However, unlike Proposition 2.2 for strictly stationary processes, a nonlinear function of a covariance stationary process may not be covariance stationary. For example, suppose that X_t is covariance stationary. Imagine that X_t's variance is finite but $E(|X_t|^4) = \infty$. Consider $Z_t = f(X_t) = (X_t)^2$. Then Z_t's variance is not finite, and hence Z_t is not covariance stationary.

In order to model strictly stationary and covariance stationary processes, it is convenient to consider white noise processes. A univariate stochastic process $\{e_t : t =$

[3]This proposition holds for any measurable function $f(\cdot) : \mathrm{R}^k \longmapsto \mathrm{R}^p$ (see Appendix 2.A). The term "measurable" is avoided because this book does not require knowledge of measure theory. All continuous functions are measurable but not vice versa. Thus the continuity condition in Proposition 2.2 is more stringent than necessary. This is not a problem for the purpose of this book because continuous functions are used in all applications of this proposition.

$\cdots, -1, 0, 1, \cdots\}$ is *white noise* if $E(e_t) = 0$, and

$$E(e_t e_j) = \begin{cases} \sigma^2 & \text{if} \quad t = j \\ 0 & \text{if} \quad t \neq j \end{cases}, \tag{2.10}$$

where σ is a constant. For a vector white noise, we require

$$E(\mathbf{e}_t \mathbf{e}_j') = \begin{cases} \Sigma & \text{if} \quad t = j \\ 0 & \text{if} \quad t \neq j \end{cases}, \tag{2.11}$$

where Σ is a matrix of constants. A white noise process is covariance stationary.

If a process is independent and identically distributed (i.i.d.), then it is strictly stationary. The simplest example of an i.i.d. process is an i.i.d. white noise. A *Gaussian white noise process* $\{e_t : -\infty < t < \infty\}$ is an i.i.d. white noise process for which e_t is normally distributed with zero mean. In these definitions, e_t can be a vector white noise process.

All linear functions of white noise random variables are covariance stationary because of Proposition 2.3. In addition, by Proposition 2.2, all functions of i.i.d. white noise random variables are strictly stationary. A simple example of this case is:

Example 2.6 Let $X_t = \delta + e_t$, where e_t is a white noise process, and δ is a constant. Then $E(X_t) = \delta$, and X_t is covariance stationary. If e_t is an i.i.d. white noise process, then X_t is strictly stationary. ∎

If X_t is strictly stationary with finite second moments, X_t is covariance stationary. Therefore, X_t's first and second moments cannot depend on date t. In empirical work, the easiest case to see that an observed variable is *not* strictly stationary is when a variable's mean shifts upward or downward over time. A simple example of this case is:

Example 2.7 Let $X_t = \delta + \theta t + e_t$, where e_t is an i.i.d. white noise random variable and δ and $\theta \neq 0$ are constants. Then X_t is *not* stationary because $E(X_t) = \delta + \theta t$ depends on time.[4] ∎

Strictly stationary and covariance stationary processes can be serially correlated, that is, their h-th order autocovariances can be nonzero for $h \neq 0$ as in the next two examples.

Example 2.8 *(The first order Moving Average Process)* Let $X_t = \delta + e_t + Be_{t-1}$, where e_t is a white noise which satisfies (2.10), and δ and B are constant. This is a moving average process of order 1 (see Chapter 4). Then X_t is covariance stationary for any B because of Proposition 2.3.[5] $E(X_t) = \delta$, and its h-th autocovariance is

$$\phi_h = E[(X_t - \delta)(X_{t-h} - \delta)] = \begin{cases} \sigma^2(1 + B^2) & \text{if } h = 0 \\ \sigma^2 & \text{if } |h| = 1 \\ 0 & \text{if } |h| > 1 \end{cases} . \tag{2.12}$$

In this example, if e_t is an i.i.d. white noise, then X_t is strictly stationary. ∎

Example 2.9 *(The first order Autoregressive Process)* Consider a process X_t which is generated from an initial random variable X_0, where

$$X_t = AX_{t-1} + e_t \qquad \text{for } t \geq 1, \tag{2.13}$$

where e_t is a Gaussian white noise random variable, and A is a constant. This is an autoregressive process of order 1 (see Chapter 4). If $|A| < 1$ and X_0 is a normally distributed random variable with mean zero and variance of $\frac{Var(e_t)}{1-A^2}$, then X_t is strictly

[4] Because X_t is stationary after removing a deterministic trend in this example, we say that X_t is trend stationary as we will discuss in Chapter 13. Trend stationarity is a way to model nonstationarity.

[5] Even though X_t is stationary for any B, it is often convenient to impose a restriction $|B| \leq 1$ as explained in Chapter 4.

stationary (see Exercise 2.3). The methods explained in Chapter 4 can be used to show that X_t is not strictly stationary when X_0's distribution is different from the one given above. ∎

2.5 Conditional Heteroskedasticity

Using conditional expectations, we can define variance and covariance conditional on an information set just as we use unconditional expectations to define (unconditional) variance and covariance. The variance of Y conditional on an information set I is

$$Var(Y|\mathrm{I}) = E[(Y - E(Y|\mathrm{I}))^2|\mathrm{I}], \tag{2.14}$$

and the covariance of X and Y conditional on an information set I is

$$Cov(X, Y|\mathrm{I}) = E[(X - E(X|\mathrm{I}))(Y - E(Y|\mathrm{I}))|\mathrm{I}]. \tag{2.15}$$

Consider a stochastic process $[Y_t : t \geq 1]$. If the unconditional variance of Y_t, $Var(Y_t)$, depends on date t, then the Y_t is said to be *heteroskedastic*; if not, it is *homoskedastic*. If Y_t's variance conditional on an information set I_t, $Var(Y_t|\mathrm{I}_t))$, is constant and does not depend on the information set, then Y_t is said to be *conditionally homoskedastic*; if not, it is *conditionally heteroskedastic*.

Example 2.10 Let $Y_t = \delta + h_t e_t$, where e_t is an i.i.d. white noise with unit variance ($E(e_t^2) = 1$), and $\{h_t : -\infty < t < \infty\}$ is a sequence of real numbers. Then the (unconditional) variance of Y_t is h_t, and Y_t is heteroskedastic as long as $h_t \neq h_j$ for some t and j. ∎

A heteroskedastic process is not strictly stationary because its variance depends on date t. It should be noted, however, that a strictly stationary random variable can

be conditionally heteroskedastic. This fact is important because many of the financial time series have been found to be conditionally heteroskedastic. For example, the growth rates of asset prices and foreign exchange rates can be reasonably modeled as strictly stationary processes. However, the volatility of such a growth rate at a point in time tends to be high if it has been high in the recent past. Therefore, such a growth rate is often modeled as a conditionally heteroskedastic process. A popular method to model conditional heteroskedasticity, introduced by Engle (1982), is an *autoregressive conditional heteroskedastic* (ARCH) process. The following is a simple example of an ARCH process.

Example 2.11 *(An ARCH Process)* Let I_t be an information set, and e_t be a univariate stochastic process such that e_t is in I_t, and $E(e_t|\mathrm{I}_{t-1}) = 0$. Assume that

$$e_t^2 = \eta + \alpha e_{t-1}^2 + w_t, \tag{2.16}$$

where $\eta > 0$, w_t is another white noise process in I_t with $E(w_t|\mathrm{I}_{t-1}) = 0$ and

$$E(w_k w_j|\mathrm{I}_t) = \begin{cases} \lambda^2 & \text{if} \quad k = j \\ 0 & \text{if} \quad k \neq j \end{cases}, \tag{2.17}$$

where λ is a constant. Relation (2.16) implies that e_t's conditional variance depends on I_t:

$$E(e_t^2|\mathrm{I}_{t-1}) = \eta + \alpha e_{t-1}^2, \tag{2.18}$$

and thus e_t is conditionally heteroskedastic.

In order to see whether or not e_t's unconditional variance is constant over time, take expectations of both sides of (2.18) to obtain

$$E(e_t^2) = \eta + \alpha E(e_{t-1}^2). \tag{2.19}$$

Hence if the variance of e_t is a constant σ^2, then $\sigma^2 = \eta + \alpha\sigma^2$, and $\sigma^2 = \frac{\eta}{1-\alpha}$. Because σ^2 is positive, this equation implies that $\alpha < 1$. When $\alpha < 1$, an ARCH process can be covariance stationary and strictly stationary. ∎

2.6 Martingales and Random Walks

Consider a stochastic process $[Y_t : -\infty < t < \infty]$, and a sequence of information sets $[\mathrm{I}_t : -\infty < t < \infty]$ that is increasing ($\mathrm{I}_t \subset \mathrm{I}_{t+1}$). If Y_t is in I_t and if

$$E(Y_{t+1}|\mathrm{I}_t) = Y_t, \tag{2.20}$$

then Y_t is a *martingale* adapted to I_t. Rational expectations often imply that an economic variable is a martingale (see Section 3.2). If Y_t is a martingale adapted to I_t and if its conditional variance, $E((Y_{t+1} - Y_t)^2|\mathrm{I}_t)$, is constant (that is, Y_t is conditionally homoskedastic), then Y_t is a *random walk*.

As we will discuss later in this book, most of the rational expectations models imply that certain variables are martingales. The models typically do not imply that the variables are conditionally homoskedastic, and hence do not imply that they are random walks. However, if the data for the variable does not show signs of conditional heteroskedasticity, then we may test whether or not a variable is a random walk. It is often easier to test whether or not the variable is a random walk than to test whether or not it is a martingale.

Consider a stochastic process $[e_t : -\infty < t < \infty]$, and a sequence of information sets $[\mathrm{I}_t : -\infty < t < \infty]$ which is increasing ($\mathrm{I}_t \subset \mathrm{I}_{t+1}$). If e_t is in I_t and if

$$E(e_{t+1}|\mathrm{I}_t) = 0, \tag{2.21}$$

then e_t is a *martingale difference sequence* adapted to I_t. If Y_t is a martingale adapted

to I_t, then $e_t = Y_t - Y_{t-1}$ is a martingale difference sequence (see Exercise 2.4). A covariance stationary martingale difference sequence is a white noise process (see Exercise 2.5). However, a white noise process may not be a martingale difference sequence for any sequence of information sets. An i.i.d. white noise process is a martingale difference sequence (see Exercise 2.6).

In these definitions, a martingale or a martingale difference sequence can be a vector stochastic process.

Appendix

2.A A Review of Measure Theory

Let S be an arbitrary nonempty set of points s. An event is a subset of S. A set of subsets is called a class. A class $\mathcal{F}$ of subsets of S is called a *field* if

(i) $S \in \mathcal{F}$;

(ii) $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, where A^c is the complement of A;

(iii) $A, B \in \mathcal{F}$ implies $A \cup B \in \mathcal{F}$.

A class $\mathcal{F}$ is a σ-field if it is a field and if

(iv) $A_1, A_2, \cdots \in \mathcal{F}$ implies $A_1 \cup A_2 \cup \cdots \in \mathcal{F}$.

A *set function* is a real-valued function defined on some class of subsets of S. A set function Pr on a field $\mathcal{F}$ is a *probability measure* if it satisfies these conditions:

(i) $0 \leq Pr(A) \leq 1$ for $A \in \mathcal{F}$;

(ii) $Pr(0) = 0, Pr(S) = 1$;

(iii) if $\mathrm{A}_1, \mathrm{A}_2, \cdots$ is a disjoint sequence of $\mathcal{F}$-sets and if $\bigcup_{k=1}^{\infty} \mathrm{A}_k \in \mathcal{F}$, then $Pr(\bigcup_{k=1}^{\infty} \mathrm{A}_k) = \sum_{k=1}^{\infty} Pr(\mathrm{A}_k)$.

If $\mathcal{F}$ is a σ-field in S and Pr is a probability measure on $\mathcal{F}$, the triple $(\mathrm{S}, \mathcal{F}, Pr)$ is called a *probability space*. Given a class $\mathcal{A}$, consider the class which is the intersection of all σ-fields containing $\mathcal{A}$. This class is the smallest σ-field which contains $\mathcal{A}$, and is called the *σ-field generated by $\mathcal{A}$* and is denoted by $\sigma(\mathcal{A})$.

Proposition 2.A.1 A probability measure on a field has a unique extension to the generated σ-field. ∎

In Euclidean k-space R^k, consider the class of the bounded rectangles

$$[\mathbf{x} = (x_1, \cdots, x_k) : a_i \leq x \leq b_i, i = 1, \cdots, k].$$

The σ-field generated from this class is called the *k-dimensional Borel sets*, and denoted by $\mathcal{R}^k$.

Let $\mathcal{F}$ be a σ-field of subsets of S and $\mathcal{F}'$ be a σ-field of subsets of S$'$. For a mapping $T : \mathrm{S} \longmapsto \mathrm{S}'$, consider the inverse images $T^{-1}(\mathrm{A}') = [s \in \mathrm{S} : T(s) \in \mathrm{A}']$. The mapping T is measurable $\mathcal{F}/\mathcal{F}'$ if $T^{-1}(\mathrm{A}') \in \mathcal{F}$ for each $\mathrm{A}' \in \mathcal{F}'$.

For a real-valued function f, the image space S$'$ is the line R^1, and in this case $\mathcal{R}^1$ is always tacitly understood to play the role of $\mathcal{F}'$. A real-valued function on S is measurable $\mathcal{F}$ (or simply measurable when it is clear from the context what $\mathcal{F}$ is involved) if it is measurable $\mathcal{F}/\mathcal{R}^1$. If $(\mathrm{S}, \mathcal{F}, Pr)$ is a probability space, then a real-valued measurable function is called a *random variable*. For a random variable X, we can assign a probability to the event that $X(s)$ belongs to a Borel set $\mathcal{B}$ by $Pr(X^{-1}(\mathcal{B}))$.

For a mapping $f : \mathrm{S} \longmapsto \mathrm{R}^k$, $\mathcal{R}^k$ is always understood to be the σ-field in the image space. If $(\mathrm{S}, \mathcal{F}, Pr)$ is a probability space, then a measurable mapping $X : \mathrm{S} \longmapsto \mathrm{R}^k$ is called a *random vector*. It is known that X is a random vector if and only if each component of X is a random variable.

A mapping $f : \mathrm{R}^i \longmapsto \mathrm{R}^k$ is defined to be measurable if it is measurable $\mathcal{R}^i/\mathcal{R}^k$. Such functions are called *Borel functions*.

Proposition 2.A.2 If $f : \mathrm{R}^i \longmapsto \mathrm{R}^k$ is continuous, then it is measurable. ∎

If X is a j-dimensional random vector, and $g : \mathrm{R}^j \longmapsto \mathrm{R}^i$ is measurable, then $g(X)$ is an i-dimensional random vector. If the distribution of X is μ, the distribution of $g(X)$ is μg^{-1}. Proposition 2.2 can be proven by taking $X = [Y_t', \cdots, Y_{t+k}']'$.

We now introduce two definitions of conditional expectation. One definition is standard in measure theory. The other definition is given because it is convenient for the purpose of stating a version of the conditional Gauss-Markov theorem used in this book. Intuitively, the conditional Gauss-Markov theorem is obtained by stating all assumptions and results of the Gauss-Markov theorem conditional on the stochastic regressors. Formally, it is necessary to make sure that the conditional expectations of the relevant variables are well defined.

Let S be a probability space, $\mathcal{F}$ be a σ-field of S, and Pr be a probability measure defined on $\mathcal{F}$. The random variables we will consider in this section are defined on this probability space. Let $\mathbf{X} = (X_1, X_2, ..., X_T)'$ be a $T \times K$ matrix of random variables, which will be the regressor matrix of the regression to be considered. Let $\mathbf{y} = (y_1, y_2, ..., y_T)$ and $\mathbf{e} = (e_1, e_2, ..., e_T)$ be $T \times 1$ vectors of random variables. We are concerned with a linear model of the form: $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$, where $\mathbf{b}_0$ is a $K \times 1$ vector of real numbers.

For s such that $\mathbf{X}(s)'\mathbf{X}(s)$ is nonsingular, the OLS estimator is

$$\mathbf{b}_T = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2.A.1}$$

In order to apply a conditional version of the Gauss-Markov Theorem, it is necessary to define the expectation and variance of $\mathbf{b}_T$ conditional on $\mathbf{X}$.

Let Z be an integrable random variable (namely, $E(|Z|) < \infty$), and $\sigma(\mathbf{X})$ be the smallest σ-field with respect to which of the random variables in $\mathbf{X}$ are measurable. The standard definition of the expectation of Z given $\mathbf{X}$ is obtained by applying the Radon-Nikodym theorem (see, e.g., Billingsley, 1986). Throughout this paper, 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.[6] $E[Z|\sigma(\mathbf{X})]$ is a random variable, and $E[Z|\sigma(\mathbf{X})]_s$ denotes the value of the random variable at s in S. It satisfies the following two properties:

(i) $E(Z|\sigma(\mathbf{X}))$ is measurable and integrable given $\sigma(\mathbf{X})$.

(ii) $E(Z|\sigma(\mathbf{X}))$ satisfies the functional equation:

$$\int_{\mathrm{G}} E(Z|\sigma(\mathbf{X}))\,dPr = \int_{\mathrm{G}} Z\,dPr, \quad \mathrm{G} \in \sigma(\mathbf{X}). \tag{2.A.2}$$

There will in general be many such random variables which satisfy these two properties; any one of them is called a version of $E(Z|\sigma(\mathbf{X}))$. Any two versions are equal with probability 1.

It should be noted that this definition is given under the condition that Z is integrable, namely $E(|Z|) < \infty$. This condition is too restrictive when we define

[6] If $\mathbf{z}$ is a vector, the conditional expectation is taken for each element in $\mathbf{z}$.

the conditional expectation and variance of the OLS estimator in many applications[7] because the moments of $(\mathbf{X}'\mathbf{X})^{-1}$ may not be finite even when $\mathbf{X}$ has many finite moments. For this reason, it is difficult to confirm that $E(\mathbf{b}_T|\sigma(\mathbf{X}))$ can be defined in each application even if $\mathbf{X}$ is normally distributed. Thus, Judge et al. (1985) conclude that the Gauss-Markov theorem based on $E(\cdot|\sigma(\mathbf{X}))$ is not very useful.

We avoid this problem by adopting a different definition of conditional expectation based on conditional distribution. For this purpose, we first define conditional probabilities following Billingsley (1986). Given A in $\mathcal{F}$, define a finite measure v on $\sigma(\mathbf{X})$ by $v(\mathrm{G}) = \Pr(\mathrm{A} \cap \mathrm{G})$ for G in $\sigma(\mathbf{X})$. Then $\Pr(\mathrm{G}) = 0$ *implies that* $v(\mathrm{G}) = 0$. The Radon-Nikodym theorem can be applied to the measures v and Pr, and there exists a random variable f that is measurable and integrable with respect to Pr, such that $\Pr(\mathrm{A} \cap \mathrm{G}) = \int_{\mathrm{G}} f d\Pr$ for all G in $\sigma(\mathbf{X})$. Denote this random variable by $\Pr(\mathrm{A}|\sigma(\mathrm{G}))$. This random variable satisfies these two properties:

(i) $\Pr(\mathrm{A}|\sigma(\mathbf{X}))$ is measurable and integrable given $\sigma(\mathbf{X})$.

(ii) $\Pr(\mathrm{A}|\sigma(\mathbf{X}))$ satisfies the functional equation

$$\int_{\mathrm{G}} \Pr(\mathrm{A}|\sigma(\mathbf{X}))\, dPr = \Pr(\mathrm{A} \cap \mathrm{G}), \quad \mathrm{G} \in \sigma(\mathbf{X}). \tag{2.A.3}$$

There will in general be many such random variables, but any two of them are equal with probability 1. A specific such random variable is called a version of the conditional probability.

Given a random variable Z, which may not be integrable, we define a conditional distribution $\mu(\cdot, s)$ given $\mathbf{X}$ for each s in S. Let $\mathcal{R}^1$ be the σ-field of the Borel sets

[7] Loeve (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 $E(\cdot|\sigma(\mathbf{X}))$, but this slight relaxation does not solve our problem.

in R^1. By Theorem 33.3 in Billingsley (1986, p.460), there exists a function $\mu(\mathrm{H}, s)$, defined for H in $\mathcal{R}^1$ and s in S, with these two properties:

(i) For each s in S, $\mu(\mathrm{H}, \mathrm{s})$ is, as a function of H, a probability measure on $\mathcal{R}^1$.

(ii) For each H in $\mathcal{R}^1$, $\mu(\mathrm{H}, s)$ is, as a function of s, a version of $Pr(Z \in \mathrm{H}|\sigma(\mathbf{X}))_s$.

For each s in S, we define $E(Z|\mathbf{X})_s$ to be $\int_{R^1} z\mu(dz, s)$. It should be noted that $E(Z|\mathbf{X})_s$ does not necessarily satisfy the usual properties of conditional expectation such as the law of iterated expectations. In general, $E(Z|\mathbf{X})_s$ may not even exist for some s. If $\int_{R^1} |z|\mu(dz, s)$ is finite, then, $E(Z|\mathbf{X})_s$ is said to exist and be finite.

Given a $T \times K$ matrix of real numbers x, $E(Z|\mathbf{X})_s$ is identical for all s in $\mathbf{X}^{-1}(x)$. Therefore, we define $E(Z|\mathbf{X} = x)$ as $E(Z|\mathbf{X})_s$ for s in $\mathbf{X}^{-1}(x)$. This is the definition of the conditional expectation of Z given $\mathbf{X} = x$ in this paper.

We are concerned with a linear model of the form:

Assumption 2.A.1 $\mathbf{y} = \mathbf{X}\mathbf{b}_0 + \mathbf{e}$

where $\mathbf{b}_0$ is a $K \times 1$ vector of real numbers. Given a $T \times K$ matrix of real numbers x, we assume that the conditional expectation of $\mathbf{e}$ given $\mathbf{X} = x$ is zero:

Assumption 2.A.2 $E[\mathbf{e}|\mathbf{X} = x] = 0$.

Next, we assume that $\mathbf{e}$ is homoskedastic and e_t is not serially correlated given $\mathbf{X} = x$:

Assumption 2.A.3 $E[\mathbf{e}\mathbf{e}'|\mathbf{X} = x)] = \sigma^2\mathbf{I}_T$.

The OLS estimator can be expressed by (2.A.1) for all s in $\mathbf{X}^{-1}(x)$ when the next assumption is satisfied:

Assumption 2.A.4 $x'x$ is nonsingular.

Under Assumptions 2.A.1–2.A.4, $E[\mathbf{b}_T|\mathbf{X} = x] = \mathbf{b}_0$ and $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\mathbf{X} = x] = \sigma^2(x'x)^{-1}$. The conditional version of the Best Linear Unbiased Estimator (BLUE) given $\mathbf{X} = x$ can be defined as follows: An estimator $\mathbf{b}_T$ for $\mathbf{b}_0$ is BLUE conditional on $\mathbf{X} = x$ if (1) $\mathbf{b}_T$ is linear conditional on $\mathbf{X} = x$, namely, $\mathbf{b}_T$ can be written as $\mathbf{b}_T = \mathbf{A}\mathbf{y}$ for all s in $\mathbf{X}^{-1}(x)$ where $\mathbf{A}$ is a $K \times T$ matrix of real numbers; (2) $\mathbf{b}_T$ is unbiased conditional on $\mathbf{X} = x$, namely, $E(\mathbf{b}_T|\mathbf{X} = x) = \mathbf{b}$; (3) for any linear unbiased estimator $\mathbf{b}^*$ conditional on $\mathbf{X} = x$, $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X} = x] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X} = x]$, namely, $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\mathbf{X}(s) = x] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\mathbf{X}(s) = x]$ is a positive semidefinite matrix.

With these preparations, the following theorem can be stated:

Theorem 2.A.1 (The Conditional Gauss-Markov Theorem) Under Assumptions 2.A.1–2.A.4, the OLS estimator is BLUE conditional on $\mathbf{X} = x$. ∎

Applying any of the standard proofs of the (unconditional) Gauss-Markov theorem can prove this theorem by replacing the unconditional expectation with $E(\cdot|\mathbf{X} = x)$.

Modifying some assumptions and adding another yields the textbook version of the conditional Gauss-Markov theorem based on $E(\cdot|\sigma(\mathbf{X}))$.

Assumption 2.A.2′ $E[\mathbf{e}|\sigma(\mathbf{X})] = 0$.

Since $E[\mathbf{e}|\sigma(\mathbf{X})]$ is defined only when each element of $\mathbf{e}$ is integrable, Assumption 2.A.2′ implicitly assumes that $E(\mathbf{e})$ exists and is finite. It also implies $E(\mathbf{e}) = 0$ because of the law of iterated expectations. Given $E(\mathbf{e}) = 0$, a sufficient condition for Assumption 2.A.2′ is that $\mathbf{X}$ is statistically independent of $\mathbf{e}$. Since Assumption 2.A.2′ does not imply that $\mathbf{X}$ is statistically independent of $\mathbf{e}$, Assumption 2.A.2′

is weaker than the assumption of independent stochastic regressors. With the next assumption, we assume that $\mathbf{e}$ is conditionally homoskedastic and e_t is not serially correlated:

Assumption 2.A.3′ $E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})] = \sigma^2\mathbf{I}_T$.

The next assumption replaces Assumption 2.A.4.

Assumption 2.A.4′ $\mathbf{X}'\mathbf{X}$ is nonsingular with probability one.

From Assumption 2.A.1, $\mathbf{b}_T = \mathbf{b}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$. Hence we can prove a version of the conditional Gauss-Markov theorem based on $E(\cdot|\sigma(\mathbf{X}))$ when the expectations of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ exist and are finite. For this purpose, we consider the following assumption:

Assumption 2.A.5 $E[trace((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1})]$ exists and is finite.

The problem with Assumption 2.A.5 is that it is not easy to verify the assumption for many distributions of $\mathbf{X}$ and $\mathbf{e}$ that are often used in applications and Monte Carlo studies. However, a sufficient condition for Assumption 2.A.5 is that the distributions of $\mathbf{X}$ and $\mathbf{e}$ have finite supports.

Under Assumptions 2.A.1, 2.A.2′–2.A.4′, and 2.A.5,

$$E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0 + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}|\sigma(\mathbf{X})] = \mathbf{b}_0.$$

Moreover, $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})]$ can be defined, and $E[(\mathbf{b}_T - \mathbf{b}_0)'(\mathbf{b}_T - \mathbf{b}_0)|\sigma(\mathbf{X})] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}|\sigma(\mathbf{X})] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{e}\mathbf{e}'|\sigma(\mathbf{X})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

We now consider a different definition of the conditional version of the Best Linear Unbiased Estimator (BLUE). The *Best Linear Unbiased Estimator (BLUE)*

conditional on $\sigma(\mathbf{X})$ is defined as follows. An estimator $\mathbf{b}_T$ for $\mathbf{b}_0$ is 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 given $\sigma(\mathbf{X})$; (2) $\mathbf{b}_T$ is unbiased conditional on $\sigma(\mathbf{X})$ in G, equivalently, $E(\mathbf{b}_T|\sigma(\mathbf{X})) = \mathbf{b}_0$, (3) for any linear unbiased estimator $\mathbf{b}^*$ conditional on $\sigma(\mathbf{X})$ for which $E(\mathbf{b}^*\mathbf{b}^{*\prime})$ exists and is finite, $E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\sigma(\mathbf{X})] \leq E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\sigma(\mathbf{X})]$ with probability 1, namely, $E[(\mathbf{b}^* - \mathbf{b}_0)(\mathbf{b}^* - \mathbf{b}_0)'|\sigma(\mathbf{X})] - E[(\mathbf{b}_T - \mathbf{b}_0)(\mathbf{b}_T - \mathbf{b}_0)'|\sigma(\mathbf{X})]$ is a positive semidefinite matrix with probability 1.

Proposition 2.A.3 Under Assumptions 2.A.1, 2.A.2′–2.A.4′, and 2.A.5, the OLS estimator is BLUE conditional on $\sigma(\mathbf{X})$. Moreover, it is unconditionally unbiased and has the minimum unconditional covariance matrix among all linear unbiased estimators conditional on $\sigma(\mathbf{X})$. ∎

Proof The proof of this proposition is given in Greene (1997, Section 6.7).

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 property may seem to contradict the standard asymptotic theory, but it does not. Asymptotically, $(1/T)\mathbf{X}'\mathbf{X}$ converges almost surely to $E[\mathbf{X}_t'\mathbf{X}_t]$ if $\mathbf{X}_t$ is stationary and ergodic. Hence the limit of the covariance matrix of $\sqrt{T}(\mathbf{b}_T - \mathbf{b}_0), \sigma^2 E[\{(1/T)(\mathbf{X}'\mathbf{X})\}^{-1}]$, is equal to the asymptotic covariance matrix, $\sigma^2[E(\mathbf{X}_t'\mathbf{X}_t)]^{-1}$.

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

Assumption 2.A.6 Conditional on $\mathbf{X}$, $\mathbf{e}$ follows a multivariate normal distribution.

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

$$N_R = \frac{R(\mathbf{b}_T - \mathbf{b}_0)}{\sigma[R(\mathbf{X}'\mathbf{X})^{-1}R]^{1/2}} \tag{2.A.4}$$

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

$$t_R = \frac{R(\mathbf{b}_T - \mathbf{b}_0)}{\hat{\sigma}[R(\mathbf{X}'\mathbf{X})^{-1}R]^{1/2}}. \tag{2.A.5}$$

Here $\hat{\sigma}$ is the positive square root of $\hat{\sigma}^2 = (\mathbf{y} - \mathbf{X}\mathbf{b}_T)'(\mathbf{y} - \mathbf{X}\mathbf{b}_T)/(T - K)$. 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 with appropriate conditioning, respectively, under either Assumptions 2.A.1–2.A.6 or Assumptions 2.A.1, 2.A.2′, 2.A.3′, and 2.A.5–2.A.6. The following proposition is useful in order to derive the unconditional distributions of these statistics.

Proposition 2.A.4 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 a standard normal distribution and a t distribution conditional on $\mathbf{X}$, respectively, Proposition 2.A.4 implies the following proposition:

Proposition 2.A.5 Suppose that Assumptions 2.A.1, 2.A.5, and 2.A.6 are satisfied and that Assumptions 2.A.2 and 2.A.3 are satisfied for all x in a set H such that $Pr(\mathbf{X}^{-1}(\mathrm{H})) = 1$. Then N_R is a standard normal random variable and t_R is a t random variable with $T - K$ degrees of freedom. ∎

Alternatively, the assumptions for Proposition 2.A.3 with Assumption 2.A.6 can be used to obtain a similar result:

Proposition 2.A.5′ Suppose that Assumptions 2.A.1, 2.A.2′–2.A.3′, 2.A.5, and 2.A.6 are satisfied for s and that Assumptions 2.A.2 and 2.A.3 are satisfied for all x in a set H such that $Pr(\mathbf{X}^{-1}(\mathrm{H})) = 1$. Then N_R is a standard normal random variable and t_R is a 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 ground that the (unconditional) distribution of $\mathbf{b}_T's$ 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 the usual (unconditional) distributions as a result of Proposition 2.A.4.

2.B Convergence in Probability

Let $c_1, c_2, \cdots, c_T, \cdots$ be a sequence of real numbers and c be a real number. The sequence is said to *converge* to c if for any ε, there exists an N such that $|c_T - c| < \varepsilon$ for all $T \geq N$. We write $c_T \to c$ or $\lim_{T\to\infty} c_T = c$. This definition is extended to a sequence of vectors of real numbers $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_T, \cdots\}$ by interpreting $|\mathbf{c}_T - \mathbf{c}|$ as the Euclidean distance $(\mathbf{c}_T - \mathbf{c})'(\mathbf{c}_T - \mathbf{c})$.

Consider a univariate stochastic process $[X_T : T \geq 1]$, and a random variable X. Fix s, and then $[X_T(s) : T \geq 1]$ is a sequence of real numbers and $X(s)$ is a real

number. For each s, verify whether or not $X_T(s) \to X(s)$. Then collect s such that $X_T(s) \to X(s)$, and calculate the probability that $X_T(s) \to X(s)$. If the probability is one, we say the sequence of random variables, $[X_T : T \geq 1]$, converges to X *almost surely* or *with probability one*. We write $X_T \to X$ almost surely. This definition is extended to a sequence of random vectors by using convergence for a sequence of vectors for each s. In general, if a property holds for all s except for a set of s with probability zero, we say that the property holds *almost surely* or *with probability one*.

If Ω has finite elements, almost sure convergence is the same thing as convergence of $X_T(s)$ to $X(s)$ in all states of the world. In general, however, almost sure convergence does not imply convergence in all states.

The sequence of random variables $[X_T : T \geq 1]$ *converges in probability* to the random variable X_T if, for all $\varepsilon > 0$, $\lim_{T\to\infty} Prob(|X_T - X| > \varepsilon) = 0$. This is expressed by writing $X_T \xrightarrow{P} c$ or $\text{plim}_{T\to\infty} X_T = X$. This extension to the vector case is done by using the Euclidean distance. Almost sure convergence implies convergence in probability.

Slutsky's Theorem is important for working with probability limits. It states that, if $\text{plim}\mathbf{X}_T = \mathbf{X}$ and if $f(\cdot)$ is a continuous function, then $\text{plim}(f(\mathbf{X}_T)) = f(\text{plim}(\mathbf{X}_T))$.

2.B.1 Convergence in Distribution

Consider a univariate stochastic process $[X_T : T \geq 1]$, and a random variable X with respective distribution functions F_T and F. If $F_T(x) \to F(x)$ for every continuity point x of F, then X_T is said to *converge in distribution* to X; this is expressed by writing $X_T \xrightarrow{D} X$. The distribution F is called the *asymptotic distribution* or the

limiting distribution of X_T.

2.B.2 Propositions 2.2 and 2.3 for Infinite Numbers of R.V.'s (Incomplete)

In Propositions 2.2 and 2.3, we only allow for a finite number of random variables. In many applications, we are often interested in infinite sums of covariance or strictly stationary random variables. We need the convergence concepts explained in Appendix 2.B. A sequence of real numbers $\{a_j\}_{j=0}^{\infty}$ is *square summable* if $\sum_{j=0}^{\infty} a_j^2$ is finite. A sufficient condition for $\{a_j\}_{j=0}^{\infty}$ is that it is *absolutely summable*, that is, $\sum_{j=0}^{\infty} |a_j|$ is finite. In the following propositions, the infinite sum $\sum_{j=0}^{\infty} a_j X_{t-j}$ means the convergence in mean square of $\sum_{j=0}^{T} a_j X_{t-j}$ as T goes to infinity.

Proposition 2.B.1 If X_t is a scalar covariance stationary process, and if $\{a_j\}_{j=0}^{\infty}$ is square summable, then $X = \sum_{j=0}^{\infty} a_j X_{t-j}$ is covariance stationary. ∎

The vector version of this proposition is:

Proposition 2.B.2 If $\mathbf{X_t}$ is a k-dimensional vector covariance stationary process, and if the absolute value of the i-th row of a sequence of a $k \times k$ matrix of real numbers $\{\mathbf{A}_j\}_{j=0}^{\infty}$ is square summable for $i = 1, \cdots, k$, then $\mathbf{X}_t = \sum_{j=0}^{\infty} \mathbf{A}_j \mathbf{X}_{t-j}$ is covariance stationary. ∎

Exercises

2.1 In Example 2.3, assume that $\pi_1 = 0.15$, $\pi_2 = 0.05$, $\pi_3 = 0.20$, $\pi_4 = 0.30$, $\pi_5 = 0.10$, and $\pi_6 = 0.20$. As in Example 2.4, compute $E(Y_2|\mathrm{I})(s)$ and $E(Y_2|\mathrm{J})(s)$. Then compute $E(E(Y_2|\mathrm{I})|\mathrm{J})(s)$. Verify that $E(Y_2|\mathrm{J})(s) = E(E(Y_2|\mathrm{I})|\mathrm{J})(s)$ for all $s \in \mathrm{S}$.

2.2 In example 2.9, assume that $|A| < 1$. This condition does not ensure that Y_t is strictly stationary. In order to see this, suppose that $Y_0 = 0$. Then compute the expected values of Y_1 and Y_2 and the variance of Y_1 and Y_2, and show that Y_t is not strictly stationary if $A \neq 0$.

2.3 In example 2.9, assume that $|A| < 1$ and that Y_0 is $N(0, \frac{\sigma^2}{1-A^2})$. Then compute the expected values of Y_1 and Y_2, the variance of Y_1 and Y_2, and the k-th autocovariance of Y. Prove that Y_t is strictly stationary in this case. (Hint: Remember that first and second moments completely determine the joint distribution of jointly normally distributed random variables.)

2.4 Let Y_t be a martingale adapted to I_t. Then prove that $e_t = Y_t - Y_{t-1}$ is a martingale difference sequence.

2.5 Prove that a covariance stationary martingale difference sequence is a white noise process.

2.6 Prove that an i.i.d. white noise process is a martingale difference sequence.

References

Billingsley, P. (1986): *Probability and Measure*. Wiley, New York.

Engle, R. F. (1982): "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation," *Econometrica*, 50(4), 987–1008.

Greene, W. H. (1997): *Econometric Analysis*. Prentice-Hall, 3rd edn.

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 II*. Springer-Verlag, New York, 4th edn.